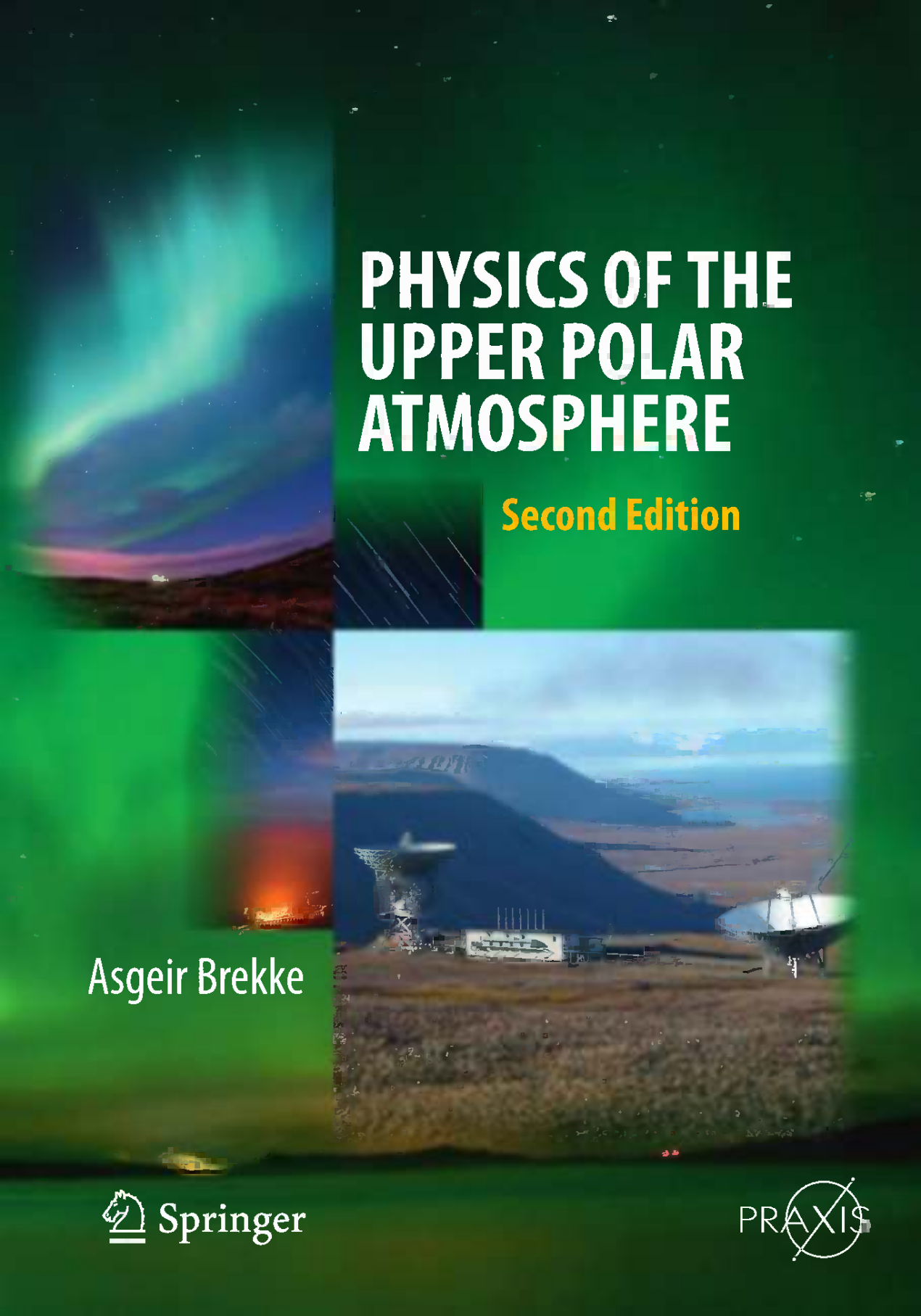

PHYSICS OF THE UPPER POLAR ATMOSPHERE
Second Edition
Asgeir Brekke
Springer
PRAXIS

Physics of the Upper Polar Atmosphere
(Second Edition)

Asgeir Brekke

Physics of the Upper Polar Atmosphere

(Second Edition)

Springer

Published in association with
Praxis Publishing
Chichester, UK

PRAXIS

Professor Asgeir Brekke
Department of Physics and Technology
University of Tromsø
Tromsø
Norway

SPRINGER–PRAXIS BOOKS IN ENVIRONMENTAL SCIENCES

ISSN 2194-5217 ISSN 2194-5225 (electronic)
ISBN 978-3-642-27400-8 ISBN 978-3-642-27401-5 (eBook)
DOI 10.1007/978-3-642-27401-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012931415

First Edition published 1997

Cover design: Jim Wilkie
Copy editor: Mike Shardlow
Typesetting: OPS Ltd., Gt. Yarmouth, Norfolk, U.K.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Linnea and Bjørn

Contents

Preface

PREFACE TO THE FIRST EDITION

The upper polar atmosphere sets the scene for one of the nature's most beautiful celestial phenomena, the aurora borealis or the northern lights. The colourful, dynamical and airy forms are the end product and the most spectacular of a long chain of plasma processes initiated by particle eruptions on the Sun. Such plasma processes are thought to be of fundamental importance all over the universe. Therefore the polar atmosphere is a natural laboratory in which we can study physical processes that give us insight into the understanding of similar light phenomena at other planets and celestial bodies. In fact, the polar atmosphere is the nearest laboratory in space from which we can expand our knowledge into the most remote places in our environment.

This book tries to follow the chain of processes which take place when the stream of particles (the solar wind) leaves the Sun, travels through interplanetary space and ends up as energetic particle beams producing the spectacular auroral forms in the polar sky.

This book has been made possible through encouragements from colleagues and students, and I will in particular thank Professor Leroy C. Cogger and Professor Nobuo Matuura who made it possible for me to spend extended visits at the universities of Calgary and Nagoya during which I was able to work on this book. I am also grateful to valuable discussions with Professor Yoshuke Kamide, Dr. Satonori Nozawa, and my Norwegian colleagues Professor Alv Egeland, Dr. Chris Hall, and Dr. Jøran Moen. Special thanks go to my students Trygve Sparr and Mårten Blixt. My secretary Liv Larssen who has typed the manuscript over and over again and who has supplied many of the illustrations in the book is greatly acknowledged for her patience.

Without the support of my wife who took all the extra loads during my many trips away from home this book would never have been a reality.

Tromsø, August 1996

PREFACE TO THE SECOND EDITION

The motivation for revising this book, *Physics of the Upper Polar Atmosphere*, which was first published in 1997, is first of all to meet modern syllabus demands in university courses at the master of science and doctor of science levels in upper-atmospheric physics. Therefore, large sections dealing with detailed descriptions of the motion of electrically charged particles in magnetic fields and the Størmer calculations are left out as these matters can be found in today's more fundamental textbooks.

Another motivation has been to correct the many errors that have been disclosed through the years of active use. In this respect I am very grateful to many of my students and colleagues who have given me advice and proposed improvements, and I owe a special debt of gratitude to Professor Takashi Okuzawa and Dr. Susume Saito for their very conscientious and careful reading of the book.

I am also grateful to Professor Takeiko Aso of the National Institute of Polar Research (NIPR) in Japan and Professor Ryoichi Fujii of the Solar Terrestrial Environment Laboratory (STEL) at Nagoya University for inviting me to spend extended visits to their institutes which allowed me to devote myself to this work during my sabbatical leave of absence from my home institution at the University of Tromsø.

Last but not least I am extremely grateful to my secretary Liv Larssen who again has helped me extensively in rewriting the text as well as redrawing some of the figures.

The picture on the cover was kindly presented to me by Professor Cesar La Hoz who took the photograph.

Tromsø, October 2011
Asgeir Brekke

Figures

Tables

Abbreviations and acronyms

1N	1 Negative
CIRA	COSPAR International Reference Atmosphere
EISCAT	European Incoherent SCATter Radar
EUV	Extreme UltraViolet
GCR	Galactic Cosmic Ray
HF	High Frequency
IMP-1	Interplanetary Monitoring Platform
IGRF	International Geomagnetic Reference Field
IGY	International Geophysical Year
IMF	Interplanetary Magnetic Field
ISIS-I	International Satellites for Ionospheric Studies
MF	Middle Frequency
NOAA	National Oceanic and Atmospheric Administration
PMSE	Polar Mesosphere Summer Echo
STP	Standard Temperature and Pressure
SC	Sudden Commencement
TMA	Trimethyl aluminum
USAF	United States Air Force
UV	UltraViolet

1

The Sun and the solar wind

1.1 GENERAL POINTS CONCERNING THE SUN

The mean distance from the Sun to the Earth is one astronomical unit (1 AU) and is about 1.5×10^{11} m. As the solar radius $R_\odot$ or the radius of the visible disk (Figure 1.1) is close to 7×10^8 m, the distance between the Sun and the Earth is about 215 $R_\odot$. Seen from the Earth the Sun covers approximately 1,920 arcseconds or about half a degree. Since the speed of light (c) is close to 3×10^8 m/s, it will take light about 500 seconds or close to 8 minutes to pass from the Sun to the Earth.

The mass of the Sun ($M_\odot$) is close to 2×10^{30} kg, and for the given radius of the Sun it leaves the medium mass density (ρ) at about 1.4×10^3 kg/m^3. The acceleration ($g_\odot$) due to gravity at the solar surface is about 2.7×10^2 m/s^2 and the escape velocity ($v_{\odot c}$) is 6.2×10^5 m/s.

The temperature at the solar center is assumed to be as high as 1.5×10^7 K. At this high temperature protons will be converted to helium nuclei by thermonuclear reactions such as the proton–proton and carbon-cycle chains. The proton–proton chain can be illustrated as follows:

$$^1\mathrm{H} + {}^1\mathrm{H} \rightarrow {}^2\mathrm{H} + e^+ + \nu + 0.42\ \mathrm{MeV}$$

$$^1\mathrm{H} + {}^2\mathrm{He} \rightarrow {}^3\mathrm{H} + \gamma + 5.5\ \mathrm{MeV}$$

$$^3\mathrm{He} + {}^3\mathrm{He} \rightarrow {}^4\mathrm{He} + 2\,{}^1\mathrm{H} + 12.8\ \mathrm{MeV}$$

where e^+, ν, and γ represent a positron, a neutrino, and a gamma ray quantum, respectively. We notice that four protons are converted into one helium nucleus while an energy of 24.64 MeV is released. In addition, energy will be released as the positron is annihilated by an electron, and the kinetic energy of the neutrino must also be included.

A. Brekke, *Physics of the Upper Polar Atmosphere*, Springer Praxis Books,
DOI 10.1007/978-3-642-27401-5_1, © Springer-Verlag Berlin Heidelberg 2013

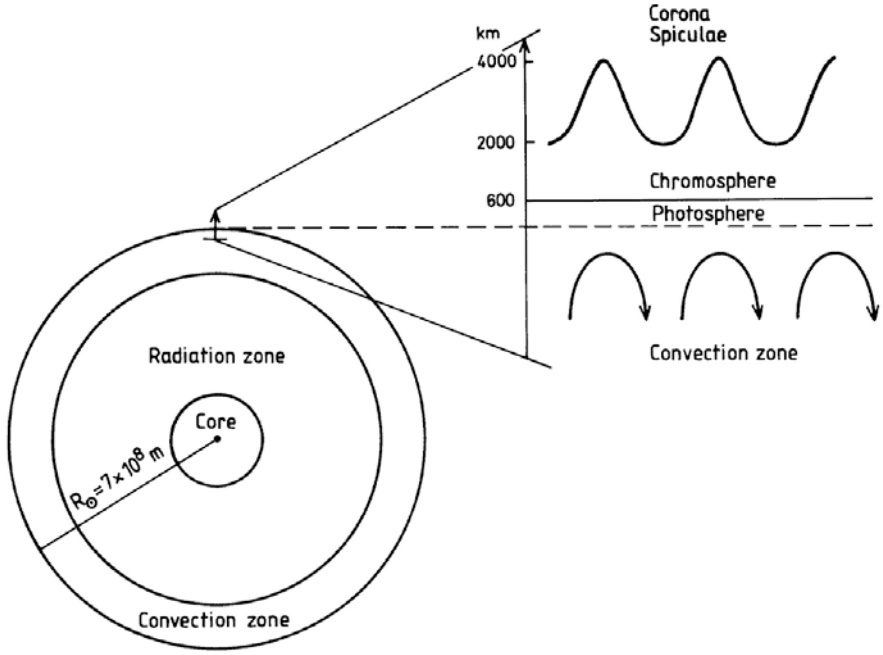


Figure 1.1. A schematic diagram showing the different zones of the Sun: the core, the radiation zone, and the convection zone. Also indicated is the structure of the solar atmosphere above the convection zone. The innermost region reaching out to about 600 km above the top of the convection zone is the *photosphere*, between 600 and 2,000 km is the *chromosphere*, and outside this the *corona*. So-called spiculae can reach out to 3,000–4000 km above the base of the photosphere.

In the central portions of the Sun it is expected that the temperature can be high enough for the carbon cycle to take place. This will result in a helium nucleus according to the following scheme:

$$^{12}\mathrm{C} + {}^{1}\mathrm{H} \rightarrow {}^{13}\mathrm{N} + \gamma \text{ (unstable)}$$

$$^{13}\mathrm{N} \rightarrow {}^{13}\mathrm{C} + e^{+} + \nu$$

$$^{13}\mathrm{C} + {}^{1}\mathrm{H} \rightarrow {}^{14}\mathrm{N} + \gamma$$

Again four protons are converted into one helium nucleus. The remaining product is ^{12}C, which can start the process all over again. The kinetic energy that is released in this manner will rapidly be converted to local thermal energy which will be transported out of the central regions.

The Sun rotates with a synodic angular velocity (relative to the Earth) which depends on the heliocentric latitude θ_s. Observations of sunspots have led to the following empirical formula for the synodic rotation period:

$$T_s = 26.9 + 5.4 \cdot \sin^2 \theta_s \text{ days}$$

1.2 THE SOLAR ATMOSPHERE

The heat located in the interior of the Sun is transported outwards in the solar atmosphere by radiation and convection. Photons and energetic particles emitted from the interior will be absorbed in the layers immediately above. These layers will emit new photons, and by absorption and emission at successive layers the radiation energy will be transported outwards in the solar atmosphere.

We discriminate between three regions in the solar atmosphere without any sharp line of demarcation (Figure 1.1). The innermost region where most of the visible light comes from is called the *photosphere*. This is supposed to be about 500–600 km thick. Negative hydrogen ions, which strongly absorb visible radiation, are present in the photospheric atmosphere, thus visible radiation emitted from a layer deeper than 400 km below the photospheric surface cannot be observed. In the photosphere, temperature (T) as well as density (ρ) decreases. In the inner part of the layer (Figure 1.2), the temperature is about 7,500 K and it decreases to about 4,200 K at the top of the photosphere.

Situated above the photosphere is the *chromosphere* which is transparent to the visible part of the spectrum except for the strong absorption lines that are

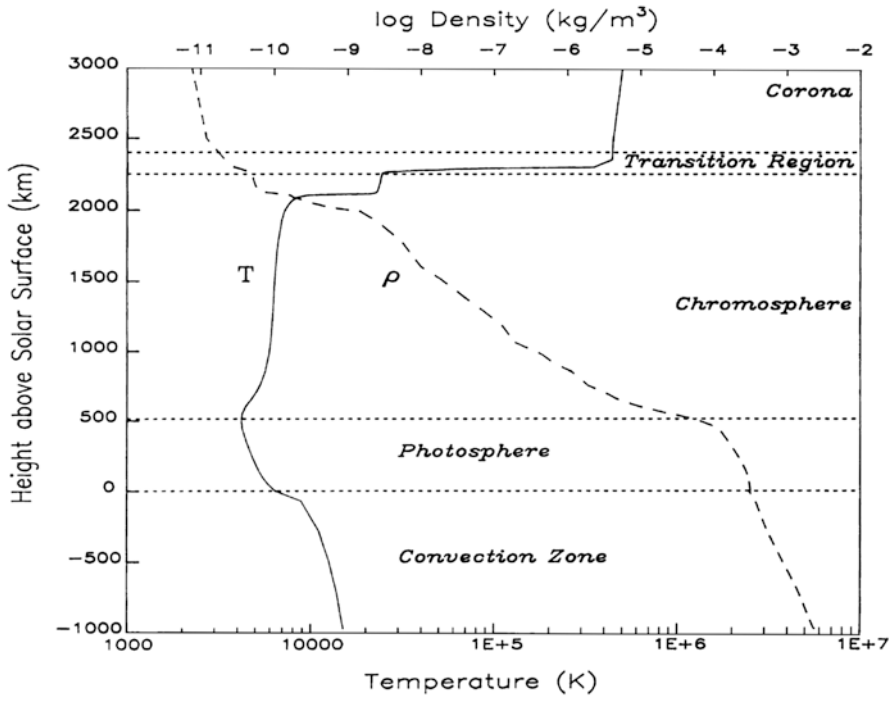


Figure 1.2. Temperature and density profiles in the solar atmosphere showing the steep temperature increase in the transition region. The zero level refers to about one solar radius from the center of the solar disk. (From Lean, 1991.)

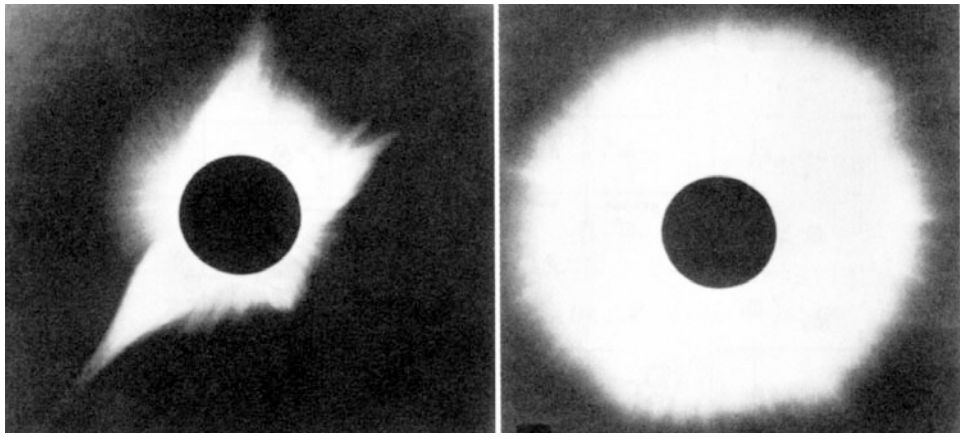


Figure 1.3. Photographs of the solar corona at solar maximum (1947, right panel) and solar minimum (1952, left panel). (From Tohmatsu, 1990.)

formed here. Density decreases continuously by height in the chromosphere, but temperature which reaches a minimum at the top of the photosphere increases slowly and reaches a value of 9,000 K at 2,000 km or so. Within 500 km above the top of the chromosphere, in the so-called transition region, the temperature increases tremendously and reaches a value of 4×10^5 K at 2,500 km above the base of the photosphere.

Outside the transition region we find the hottest part of the solar atmosphere, the *corona*, where the temperature is fairly uniform and close to 10^6 K. In this region the solar atmosphere is fully ionized. This is also true in the interplanetary space at the distance of the Earth (1 AU). The temperature here, however, decreases to a few times 10^5 K.

During solar eclipses the corona can be photographed (Figure 1.3) at distances out to 4–5 $R_\odot$ while some extreme rays can be seen to reach out to 15 $R_\odot$. From scintillation studies of radio stars the corona can be traced almost half the distance to the Earth.

The corona is by no means uniform as it has a very complicated fine structure, with rays, plumes, fans, and arcs. These structures are closely connected to the shape of the magnetic field in the corona which changes in a partly systematic way from one solar cycle to the next.

The thermal conductivity in the corona is very high, where hydrogen as well as helium are almost totally ionized. The corona is therefore more or less isothermal at a temperature of close to 10^6 K.

1.3 ELECTROMAGNETIC RADIATION FROM THE SUN

The observed radiation spectrum from the Sun as measured from the Earth's surface in the wavelength region 0.2–3.2 μm (2000–32,000 Å) is shown in Figure

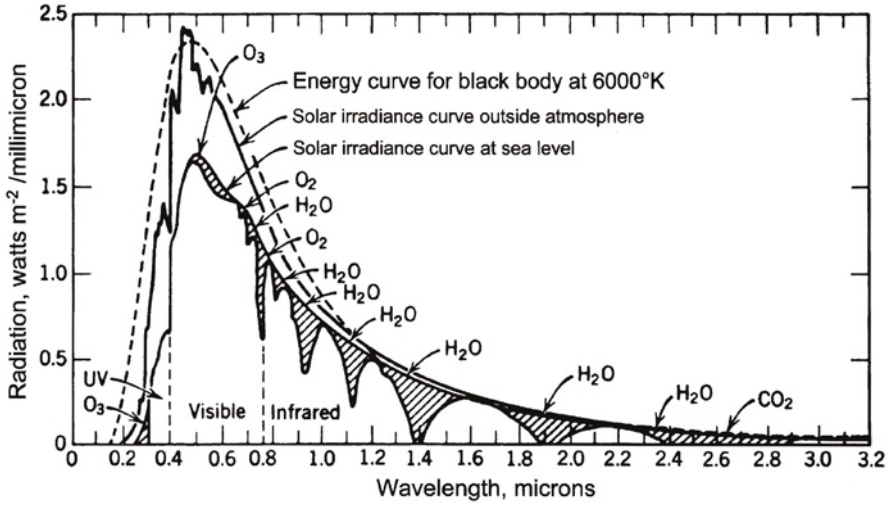


Figure 1.4. The Sun's spectral irradiance per unit wavelength of solar radiation observed at sea level compared with at blackbody spectrum at 6,000 K and the spectrum observed outside the Earth's atmosphere. The hatched areas indicate absorption windows due to H_2O, CO_2, O_2, and O_3 in the Earth's atmosphere. (From Pettit, 1951.)

1.4. The black areas in the spectrum represent absorption due to water vapor (H_2O), carbon dioxide (CO_2), oxygen (O_2), and ozone (O_3) in the Earth's atmosphere. The solar spectrum, as observed outside the Earth's atmosphere, is also illustrated together with the spectrum from a blackbody at a temperature of 6,000 K. The peak intensity of the solar spectrum is close to 0.15 W/m^2 Å.

Ultraviolet (UV) and X-ray radiation shorter than 0.2 μm are strongly absorbed by air. The UV radiation region is usually divided into two parts: (1) the far ultraviolet region between 0.1 and 0.2 μm, and (2) the extreme ultraviolet region (EUV) between 0.01 and 0.1 μm.

The continuum spectrum in the far ultraviolet region originates from the uppermost layer of the photosphere, therefore the effective blackbody temperature for this spectral region decreases to about 4,500 K as compared with 6,000 K for the visible regime. Solar radiation in this wavelength region is completely absorbed by molecular oxygen at altitudes between 80 and 120 km in the terrestrial atmosphere. It is therefore an important heat source of the upper atmosphere (*mesosphere* and *thermosphere*) and also for ionization and dissociation of molecular oxygen in these regions.

In the EUV regime there is a number of strong emission lines such as the hydrogen Balmer lines at 121.6 nm (Ly_α), 102.5 nm (Ly_β), etc., the helium line at 58.4 nm, and the oxygen line at 130.4 nm. Another group of emission lines are the ion lines such as the He^+ line at 30.4 nm and the emission lines from multi-ionized Si, Fe, Mg, N, and O.

EUV radiation originates in the chromosphere, which is the region extending above the photosphere in the solar atmosphere. It is almost transparent in the visible spectral region.

The strong atomic emission lines all originate in the low-temperature region of the chromosphere while the ionic lines are emitted from the high-temperature region near the coronal base. Since the temperature in the chromosphere varies considerably from approximately 4,500 K at the bottom to some hundred thousand kelvins at the top, the emission spectra from different layers in the chromosphere show different features corresponding to the degree of ionization, temperature, and density there.

In order to understand the behavior of the upper atmosphere of the Earth and its ionized part, the *ionosphere*, it is important to have precise knowledge of solar EUV radiation flux since it is this radiation which is responsible for the formation of the ionosphere through photoionization processes.

While the solar spectrum between 0.4 and 3.2 μm is rather constant during a solar cycle, the UV lines and X-ray emissions are more variable and often enhanced during so-called solar flare events. Emissions in the EUV and X-ray regimes (<50 nm) are often connected to sunspots and other visible phenomena on the solar surface related to solar activity. The estimated peak intensity of these emissions is much less than the peak in the visible part of the spectrum. This variability of the EUV and X-ray radiation intensity is also illustrated in Figure 1.5 where the Sun's spectral irradiance typical of solar minimum conditions is presented in the upper panel. The approximate amplitude of the Sun's spectral irradiance variation from the maximum to the minimum phase relative to the amplitude at minimum for an 11-year solar cycle is presented in the lower panel of Figure 1.5. Clearly, there is much larger variability in solar radiation below 300 nm than in any other part of the spectrum.

1.4 PLANCK'S RADIATION LAW

Let us assume that the Sun is radiating its energy approximately as a blackbody with temperature T, then the spectral brightness, B_ν, per frequency band, $d\nu$, according to Planck's radiation law will be:

$$B_\nu = \frac{2h\nu^3}{c^2} \frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1} \tag{1.1}$$

where c is the speed of light ($= 3 \times 10^8$ m/s), h is Planck's constant ($= 6.63 \times 10^{-34}$ Js), k is the Boltzmann constant ($= 1.38 \times 10^{-23}$ J/K), and T is the temperature in kelvins. B_ν is then given in $\mathrm{W\,Hz^{-1}\,sr^{-1}\,m^{-2}}$ in the frequency range $\nu + d\nu$ to ν. We notice that B_ν only depends on the temperature and not the geometry of the body. We can also express the spectral brightness B_λ per unit

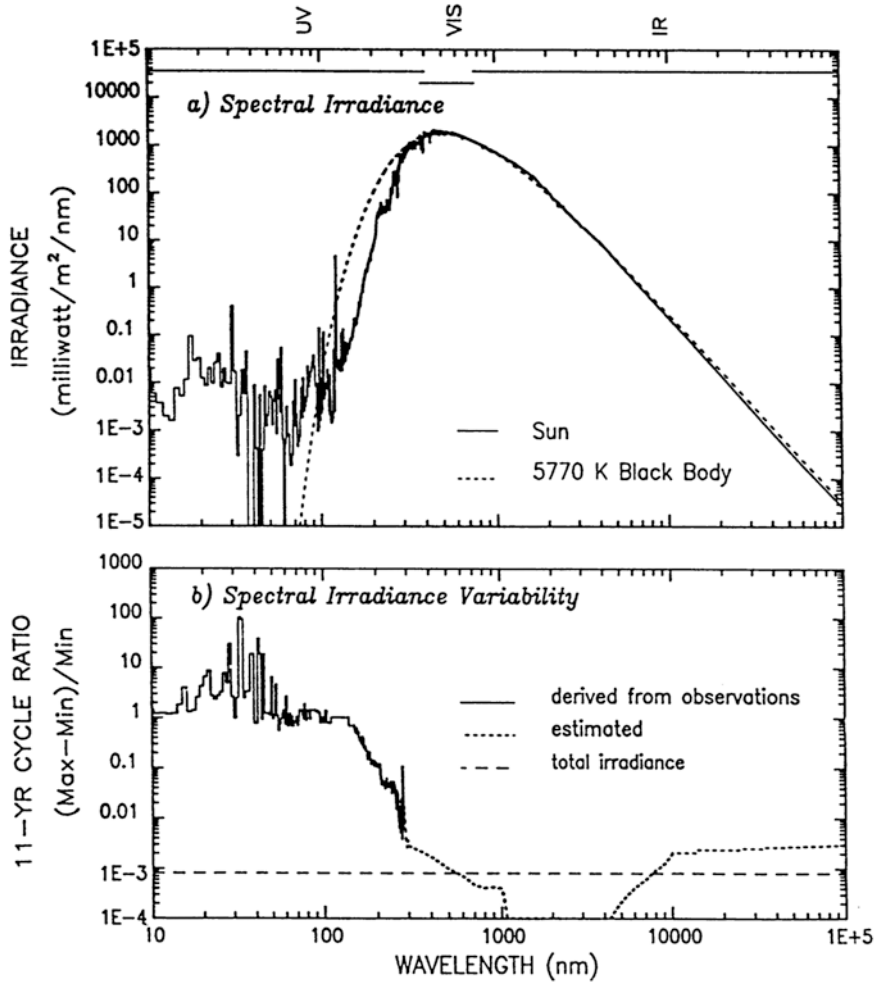


Figure 1.5. (a) The Sun's spectral irradiance or energy distribution typical for a solar minimum condition compared with a blackbody spectrum at 5,770 K. The spectral bands indicated are the ultraviolet (UV), visible (VIS), and infrared (IR). (b) The approximate amplitude of the Sun's spectral irradiance variation from the maximum to the minimum of the 11-year solar cycle with respect to the intensity at minimum. The variability for wavelengths shorter than 300 nm is outstanding. (From Lean, 1991.)

wavelength band between $\lambda + d\lambda$ and λ in the following manner when noting that $d\nu = -(c/\lambda^2)d\lambda$ and $B_\nu d\nu = B_\lambda d\lambda$:

$$B_\lambda = \frac{2hc^2}{\lambda^5} \frac{1}{\exp\left(\dfrac{hc}{k\lambda T}\right) - 1} \tag{1.2}$$

We find that the maximum in B_λ occurs at the wavelength λ_m which satisfies the following transcendental equation:

$$5(1 - \exp(-x_m)) = x_m \tag{1.3}$$

where

$$x_m = \frac{hc}{k\lambda_m T}$$

Equation (1.3) has the solution

$$x_m = 4.9651\ldots$$

and finally:

$$\lambda_m \cdot T = 2.898 \times 10^{-3}\ \text{m K} \tag{1.4}$$

which is the Wien displacement law expressing that the wavelength at maximum radiation for a blackbody is inversely proportional to the temperature. Figure 1.6 demonstrates the Wien displacement law where B_λ is presented for different temperatures. From Figure 1.4 we notice that there is a maximum in the solar spectrum close to 500.0 nm which, according to equation (1.4), corresponds to a temperature of

$$T = 5{,}780\ \text{K}$$

or very close to 6,000 K which represents the theoretical spectrum introduced in Figure 1.4.

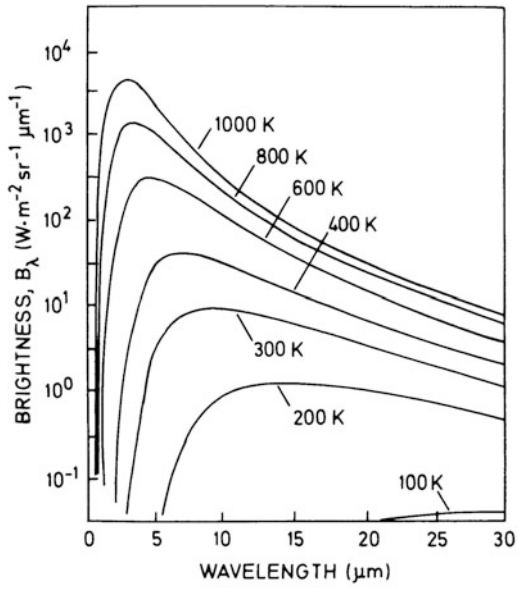


Figure 1.6. Curves showing blackbody radiation functions for bodies with different temperatures.

According to the Stephan–Boltzmann law, which is the integrated Planck's equation over all wavelengths and solid angles, we have the total radiated power Q given by:

$$Q = q \cdot S = \sigma \cdot T^4 \cdot S \tag{1.5}$$

where S is the surface of the radiating body, and q is radiated power per unit area. The Stephan–Boltzmann constant, σ, is given by:

$$\sigma = \frac{2\pi^5 \cdot k^4}{15c^2h^3} = 5.67 \times 10^{-8}\ \mathrm{W\,m^{-2}\,K^{-4}}$$

For the Sun the radiated power per unit area will be given by:

$$q_\odot = \sigma \cdot {T_\odot}^4$$

where $T_\odot$ is the average solar radiative temperature. For a blackbody temperature of the Sun $T_\odot = 5780\,\mathrm{K}$ the radiated power per unit area will be:

$$q_\odot = 6.3 \times 10^7\ \mathrm{W/m^2}$$

Since the solar radius is $R_\odot = 7 \times 10^8$ m, the total radiated power from the Sun according to equation (1.5) will be:

$$Q_\odot = 4\pi {R_\odot}^2 q_\odot = 3.9 \times 10^{26}\ \mathrm{W} \tag{1.6}$$

From Einstein's energy–mass relation we have:

$$E = mc^2$$

The total radiated power from the Sun then corresponds to a mass loss per unit time:

$$\dot{m} = \frac{Q_\odot}{c^2} = 4.3 \times 10^9\ \mathrm{kg/s} \tag{1.7}$$

As the solar mass $M_\odot = 2 \times 10^{30}$ kg it would take $T = M_\odot/\dot{m} = 1.5 \times 10^{13}$ years before the Sun had burned out. The power $Q_\odot$ given by (1.6) will radiate in all directions. At a distance from the Sun corresponding to the average position of the Earth ($r = 1$ AU) we find by assuming conservation of the power flux:

$$4\pi {R_\odot}^2 q_\odot = 4\pi r^2 E_e$$

E_e, often called the solar constant, is the radiation power per unit area reaching the Earth at a distance of 1 AU. It is given from (1.7) by:

$$E_e = q_\odot \left(\frac{R_\odot}{r}\right)^2 = 1380\ \mathrm{W/m^2} \tag{1.8}$$

The Earth does not, however, move around the Sun at a constant distance, since perihelion (r_p) is 1.46×10^{11} m and aphelion (r_a) is 1.52×10^{11} m. Therefore, although the radiation power density from the Sun may be constant at a fixed

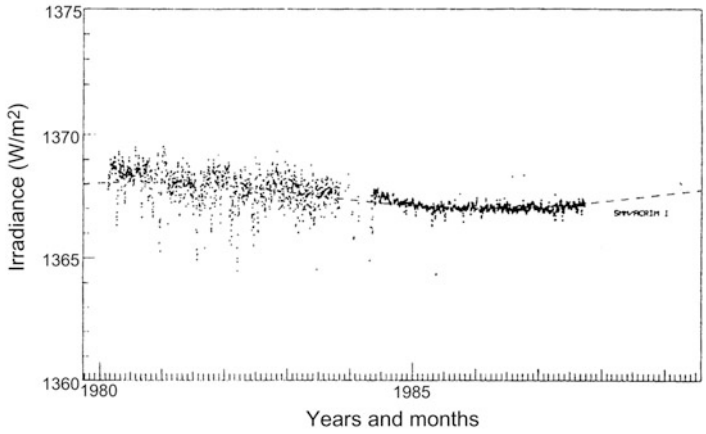


Figure 1.7. Solar total irradiance variations in watts/m^2 between 1980 and 1990. (From Willson and Hudson, 1988.)

distance, it will vary at the position of the Earth by:

$$\left(\frac{r_a}{r_p}\right)^2 = 1.07$$

or 7% during one year.

The radiation from the Sun in the visible and infrared part of the spectrum is rather constant and changes only by a fraction of one percent in the course of a year, as seen in Figure 1.7. The short-wavelength part of the spectrum of the EUV and the X-ray regime is, as we have seen, highly variable, and so also is the long-wavelength region, the radio emissions, from the Sun.

1.5 THE GREENHOUSE EFFECT

It is the heat radiating from the Sun that is the main heat source for the Earth and its atmosphere. To hold a stable temperature in the atmosphere and to stop the Earth from being heated above all limits there has to be comparable heat radiation away from the Earth. We have seen that the Sun radiates approximately as a blackbody at a temperature close to 6,000 K.

According to the Stephan–Boltzmann law (1.5) the Earth as a blackbody will radiate a total power given by:

$$Q_e = 4\pi \cdot R_e^2 \cdot \sigma \cdot T_e^4 \tag{1.9}$$

where R_e is the Earth's radius ($= 6.371 \times 10^6$ m), and T_e is the mean radiative temperature of the Earth. The Earth will only absorb a fraction of the radiation from the Sun:

$$Q_{Se} = (1 - \alpha) \cdot E_e \cdot \pi \cdot R_e^2 \tag{1.10}$$

where α is the albedo or reflection coefficient (≈ 0.3), and E_e is the solar constant according to (1.8), set equal to 1,380 W/m^2 at 1 AU.

For thermal equilibrium (heat in = heat out) we have:

$$Q_{Se} = Q_e \tag{1.11}$$

and inserted from (1.9) and (1.10):

$$(1-\alpha)\cdot E_e \cdot \pi \cdot R_e^2 = \sigma \cdot 4\pi R_e^2 T_e^4 \tag{1.12}$$

T_e is given by:

$$T_e = \left(\frac{(1-\alpha)\cdot E_e}{4\cdot\sigma}\right)^{1/4}$$

By inserting the appropriate numbers

$$T_e = 255\,\mathrm{K}$$

This is a temperature that on average is almost 35 K lower than the present-day average global radiative temperature. We notice that even if we assume that $\alpha = 0$ (i.e., all solar energy is absorbed by the Earth), the temperature of the Earth would still be only 279 K or 10 K lower than the global average.

The albedo of the Earth is a difficult parameter to estimate as it depends on vegetation, ice coverage, clouds, etc. Table 1.1 gives a list of the albedo for different surfaces. In Figure 1.8 the variation with latitude of the planetary albedo

Table 1.1. Albedo of various surfaces

Surface	*%*
Fresh snow, high Sun	80–85
Fresh snow, low Sun	90–95
Old snow	50–60
Sand	20–30
Grass	20–25
Dry earth	15–25
Wet earth	10
Forest	5–10
Water (Sun near horizon)	50–80
Water (Sun near zenith)	3-5
Thick cloud	70–80
Thin cloud	25–50
Planetary albedo	30

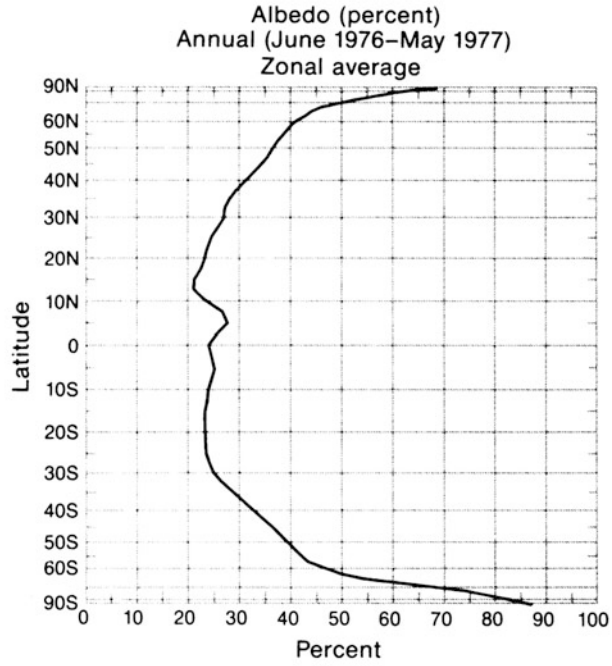


Figure 1.8. Variation with latitude of the planetary albedo averaged around the Earth along parallels of latitude for the period from June 1976 to May 1977 according to measurements made by NOAA polar-orbiting satellites. (From *Earth–Atmosphere Radiation Budget Analyses Derived from NOAA Satellite Data June 1974–February 1978*, Washington, D.C., NOAA–NESS, 1979.)

averaged around the Earth along parallels of latitude is shown. The high values in polar latitudes are due to snow cover there. In middle latitudes intermediate values correspond to the relatively large amounts of cloudiness in the region of storminess associated with the polar front. The low values in low latitudes are due to the infrequency of clouds in the belt of subtropical high pressure. The slightly increased values just north of the equator correspond to the greater cloudiness in the intertropical convergence zone where the trade winds from north and south merge.

Independent of what albedo we choose we need a "blanket" around the Earth to keep us from freezing to death.

This comes about because solar radiation, which is mainly distributed in the visible region of the spectrum, passes through the Earth's atmosphere with little absorption. The Earth, on the other hand, radiates as a blackbody at an average temperature of 288 K. According to the Wien displacement law (1.4) this corresponds to a wavelength for maximum radiation:

$$\lambda_m = 10\ \mu\text{m}$$

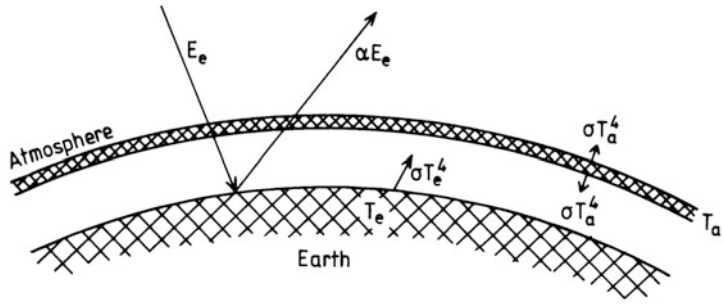


Figure 1.9. A simple illustration of the greenhouse effect in the Earth's atmosphere showing incoming solar radiation E_e and the reflected part αE_e. The blackbody temperature of the Earth is represented by T_e. An atmospheric layer is indicated having a blackbody temperature T_a. The system is supposed to be in radiative thermal equilibrium.

which is in the infrared. It is then very fortunate for life on Earth that H_2O, CO_2, and O_3 among other gases in the atmosphere absorb the infrared emission from the ground and reradiate part of this heat back to the Earth.

Let us therefore assume that the atmosphere forms a thin layer encircling the Earth at a height h which is much smaller than the Earth's radius. Let this layer be transparent to solar radiation, but opaque to infrared radiation from the ground (Figure 1.9). If this atmosphere has a temperature T_a, it will radiate heat both outward into space and inward towards the Earth. The total heat radiated per unit time from the atmosphere according to (1.5) will be:

$$Q_a = 2 \cdot \sigma \cdot 4\pi \cdot R_e^2 \cdot T_a^4 \tag{1.13}$$

where we have neglected the height as well as the thickness of the atmosphere compared with the Earth radius. If the atmosphere is only heated by radiation from the Earth, the balance in heat radiation gives

$$Q_e = Q_a$$

and by inserting (1.9) for Q_e:

$$\sigma \cdot 4\pi \cdot R_e^2 \cdot T_e^4 = 2 \cdot \sigma \cdot 4\pi \cdot R_e^2 \cdot T_a^4$$

and solving for T_a we find:

$$T_a^4 = \tfrac{1}{2} T_e^4 \tag{1.14}$$

For the Earth receiving heat radiation from the Sun as well as from the atmosphere the energy balance is now expressed by:

$$Q_{Se} + \tfrac{1}{2} Q_a = Q_e$$

and by inserting from (1.10), (1.13), and (1.9):

$$(1 - \alpha) E_e \cdot \pi \cdot R_e^2 + \sigma \cdot 4\pi R_e^2 \cdot T_a^4 = \sigma \cdot 4 \cdot \pi \cdot R_e^2 T_e^4$$

By introducing (1.14) for T_a we can solve for T_e:

$$T_e = \left[\frac{(1-\alpha)E_e}{2\sigma}\right]^{1/4}$$

Inserting the proper values we find

$$T_e = 304\,\mathrm{K}$$

and the temperature increases by almost 50°C compared with a situation without any atmosphere acting as a blanket around the globe. The temperature would, however, in this simplified model, be about 16°C too high, but we can definitely conclude that without a greenhouse effect in the atmosphere the globe would not support the life we know.

To treat the atmosphere in a more realistic way is of course far more difficult than this simple illustration indicates.

To give a full treatment of the problem the radiation through the atmosphere has to be computed by the equation of radiation transfer.

1.6 RADIOWAVE EMISSIONS FROM THE SUN

Radio emissions from a quiet Sun have been studied over a wide range of wavelengths from the mm- to the m-region. For radio frequencies corresponding to this wavelength regime (10^4–10^2 MHz) we notice that the energy related to them is typically (for $\nu = 10^3$ MHz):

$$h\nu = 6.6 \times 10^{-34}\,\mathrm{J \cdot s} \cdot 10^3 \cdot 10^6\,\mathrm{s^{-1}} = 6.6 \times 10^{-25}\,\mathrm{J}$$

where h is Planck's constant. This is much less than the thermal energy ($kT_\odot$) for a reasonable solar temperature ($T_\odot = 6{,}000\,\mathrm{K}$) which corresponds to

$$kT_\odot = 8.4 \times 10^{-20}\,\mathrm{J}$$

Therefore, in the radio wavelength regime we can use the Rayleigh–Jeans approximation of Planck's radiation law ($h\nu \ll kT$):

$$B_\nu = \frac{2h\nu^3}{c^2}\frac{1}{1 + h\dfrac{\nu}{kT} + \cdots - 1} = \frac{2kT\nu^2}{c^2} \tag{1.15}$$

From Figure 1.10 we notice that the intensity of radio emissions for a quiet Sun decreases by more than two orders of magnitude in the wavelength regime 1 cm to 10 m. Solar irradiance at the Earth's distance can be found from spectral brightness, B_ν, by calculating the radiation flux passing through a unit area of the Earth perpendicular to the solar rays during unit time.

In Figure 1.11 we notice that the radiation energy emitted from an infinitesimal area $d\sigma$ perpendicular to the Sun–Earth line at point P on the Sun and hitting

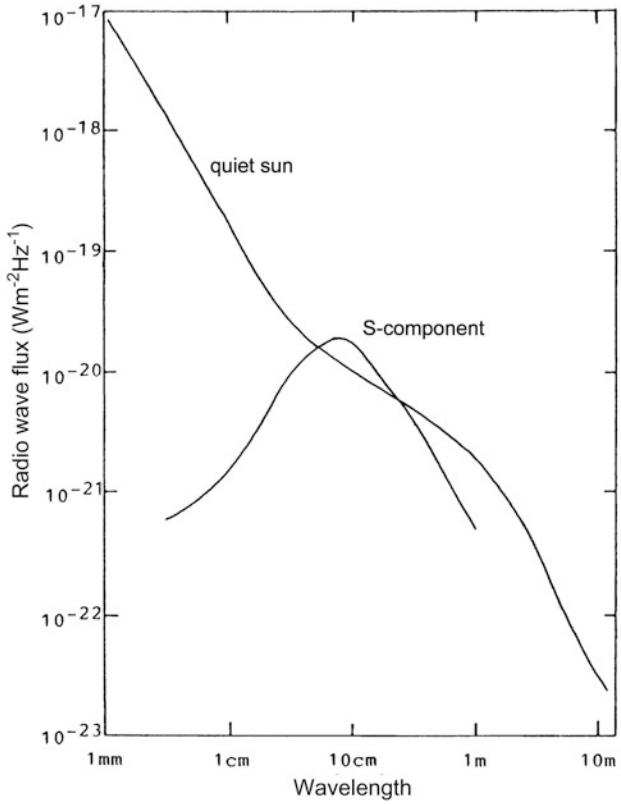


Figure 1.10. Schematic diagram of the solar radiowave radiation spectrum for a quiet Sun as well as for the slowly varying component (the *S*-component). (From Tohmatsu, 1990.)

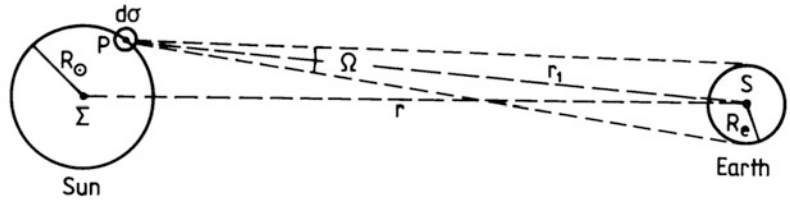


Figure 1.11. The Sun–Earth geometry where *S* is the projection of the Earth's surface to a plane perpendicular to the Sun–Earth line, *r* is equal to 1 AU, and $d\sigma$ is the projection of an infinitesimal area on to a plane perpendicular to the line from a point *P* on the Sun to the Earth's center. Since $r \gg R_\odot$, where $R_\odot$ is the solar radius, r_1 can be considered equal to *r*.

the surface of the Earth is:

$$dI_\nu = B_\nu d\sigma \cdot \Omega = B_\nu d\sigma \frac{S}{r_1^2}$$

where *S* is the projection of the surface of the Earth facing the Sun on a plane perpendicular to the line from *P* to the Earth's center, r_1 is the distance from *P* to

the Earth ($r_1 \gg R_e$), and $\Omega = S/r_1^2$ is the solid angle corresponding to the surface S as seen from P. Assuming now that the brightness, B_ν, is isotropic all over the solar surface, and that $r \gg R_\odot$ so that $r = r_1$, then the total radiation flux per unit area reaching the Earth is:

$$\phi_\nu = \frac{I_\nu}{S} = \int_\Sigma B_\nu \frac{d\sigma}{r^2} = \pi \left(\frac{R_\odot}{r}\right)^2 B_\nu$$

where the integration is made over the effective surface Σ of the Sun facing the Earth. Introducing B_ν from (1.15) we have:

$$\phi_\nu = \pi \left(\frac{R_\odot}{r}\right)^2 \cdot \frac{2k}{c^2}\nu^2 T = \pi \left(\frac{R_\odot}{r}\right)^2 \frac{2kT}{\lambda^2}$$

and applying the proper constants we find:

$$\phi_\nu = 2.09 \times 10^{-44} \nu^2 T \ [\mathrm{W\,m^{-2}\,Hz^{-1}}] \tag{1.16}$$

Superposed on these thermal background solar radiowave emissions there is a component varying with phenomena on the Sun such as sunspots and plages. It is predominant in the cm- and dm-regions and its flux varies on a time scale of a few days, therefore it is often referred to as the slowly varying component or the *S*-component (see Figure 1.10). It is rather well correlated with the fluxes of EUV and X-ray radiation and appears to be emitted from areas on the solar surface that have extremely high temperatures. The flux of this component is frequently referred to in the context of ionospheric physics due to its high correlation with EUV radiation (Figure 1.12) and because it can be monitored continuously. The observed 10.7 cm (2.8 GHz) radiowave has been commonly used and represented as $F_{10.7}$ or S_a in units of $10^{-22}\,\mathrm{W\,m^{-2}\,Hz^{-1}}$ reduced to 1 AU, as illustrated in Figure 1.13 for the years between 1986 and 1997.

For the 10.7 cm radio flux we find from (1.16) that the radiation flux is given by ($F_\lambda = \phi_\nu$):

$$F_{10.7} = 16.4 \times 10^{-26} \cdot T \quad \mathrm{W\,m^{-2}\,Hz^{-1}}$$

and the index is therefore a direct indicator of the effective temperature.

Before we discuss the observations of radio bursts from solar flares we will introduce the electron plasma frequency given by:

$$f_{pe} = \frac{1}{2\pi}\left(\frac{e^2 n_e}{\varepsilon_0 m_e}\right)^{1/2} = K n_e^{1/2} \tag{1.17}$$

where

$$K = \frac{1}{2\pi}\left(\frac{e^2}{\varepsilon_0 m_e}\right)^{1/2}$$

e is the electronic charge ($= 1.6 \times 10^{-19}$ C), n_e is the electron density ($\mathrm{m^{-3}}$), m_e is the electron mass ($= 9.1 \times 10^{-31}$ kg), and ε_0 is the permittivity of a vacuum ($= 8.854 \times 10^{-12}$ F/m).

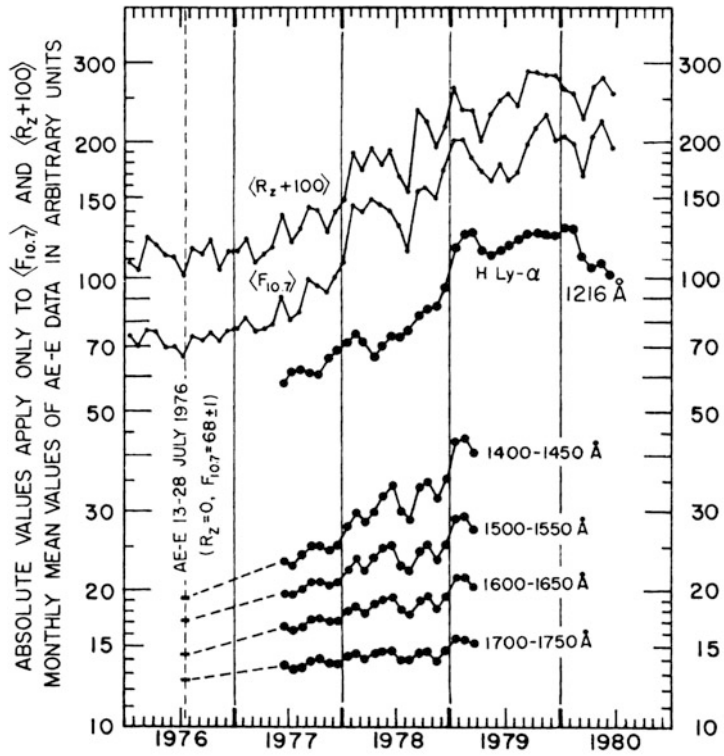


Figure 1.12. Variations in the monthly mean sunspot number (R_z) compared with the mean radiowave flux $F_{10.7}$ and the irradiance at wavelengths from 121.6 nm to 175.0 nm (1216–1750 Å). (From Hinteregger, 1980.)

From a given level of the solar atmosphere characterized by a specific electron density of n_e only radiation with a frequency $\nu > f_{pe}$ will escape if the influence of the magnetic field is neglected. Most of the radiation will actually originate from layers where f_{pe} is close to ν. In Figure 1.14 a schematic illustration of variation in electron density is presented as a function of distance from the solar center. As can be noticed, the electron density decreases strongly from close to $10^{15}\,\mathrm{m}^{-3}$ at $1\,R_\odot$ to close to $10^{10}\,\mathrm{m}^{-3}$ at $10\,R_\odot$. Electron density is on average lower during solar minimum conditions.

According to Figures 1.10 and 1.14 and our understanding of the plasma frequency we now realize that by observing radio emissions from the Sun at increasing frequencies (or decreasing wavelengths) we are probing deeper and deeper into the solar atmosphere. It is found that while m-waves are formed far out in the corona, cm-waves are generated deep in the solar atmosphere and mm-waves penetrate almost all the way from the photosphere.

An interesting aspect of radiowave emissions from the Sun is the apparent magnitude of the solar disk when viewed at different wavelengths. If we compare

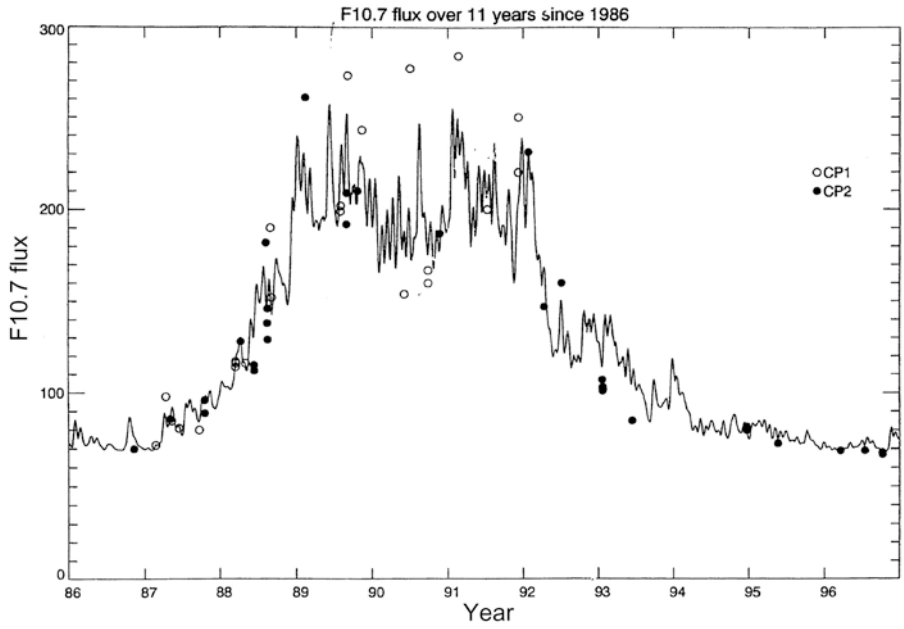


Figure 1.13. Annual variation of the $F_{10.7}$ flux. (From Nozawa and Brekke, 1999.)

the size of the solar disk as observed in the radio wavelength regime with the size of the visual Sun (as is done in Figure 1.15), we notice that at a wavelength of 25 cm the Sun appears as a sharp ring at the rim of the visual Sun. At 5 m wavelength, however, the Sun has strong emission in the center while it gets weaker toward the rim. It can, however, be seen at twice the distance from the solar center compared with the visual Sun.

Although the spectrum of electromagnetic radiation from the Sun extends from X-rays with wavelengths of 10^{-8} m or less to radiowaves of wavelength 100 m or more, 99% of the energy is concentrated in the range 276 to 4,960 nm and 99.9% in the range 217 to 10,940 nm. In other words, all but 0.1% of the energy radiating from the Sun is found in the visible, infrared, and ultraviolet parts of the spectrum. As this integrated part of solar radiative energy is fairly constant by time it is often argued that the Sun has very little effect on global climatic variations on Earth. The effects of the very variable ultraviolet particle radiation from the Sun are then overlooked. There is also a lack of knowledge about detailed variations of the infrared part of the spectrum during a solar cycle.

1.7 SUNSPOTS AND THE SOLAR CYCLE

Sunspots are one of the most fascinating facets of the Sun; they have been observed by the naked eye for thousands of years (Figure 1.16). They have often

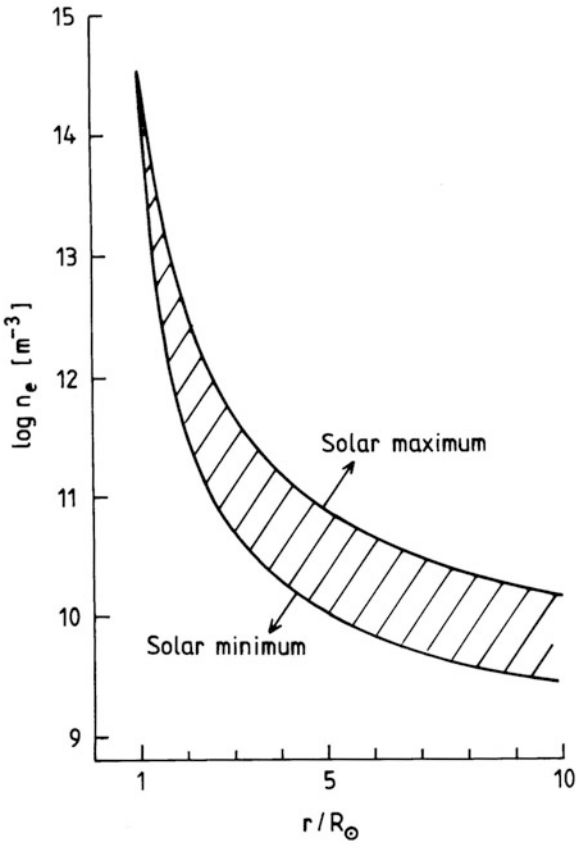


Figure 1.14. Variations in electron density in the solar atmosphere from the base of the photosphere to a distance of 10 $R_\odot$ from the center of the solar disk under different solar activity conditions.

caused fear and been observed as omens of famine, war, fire, etc. But they have also been deeply admired and inspired people to speculate on their cause. It is only the very strongest of the sunspots, which under certain favorable circumstances, can be observed by the naked eye. They were first observed by telescope at the beginning of the 17th century by Galileo (in 1616).

The fact that sunspots come and go in cycles (Figure 1.17) was not appreciated until 1843, when Schwabe (1844) published a short paper based on observations between 1826 and 1843. Schwabe found the cycle length to be about 10 years. Wolf devised a quantitative definition for a sunspot number and studied sunspot data for the years 1700 to 1848 and indeed confirmed Schwabe's discovery of a cycle, but Wolf found a period of 11.1 years.

In its present form the Wolf sunspot number R_z is defined as:

$$R_z = k(10g + f)$$

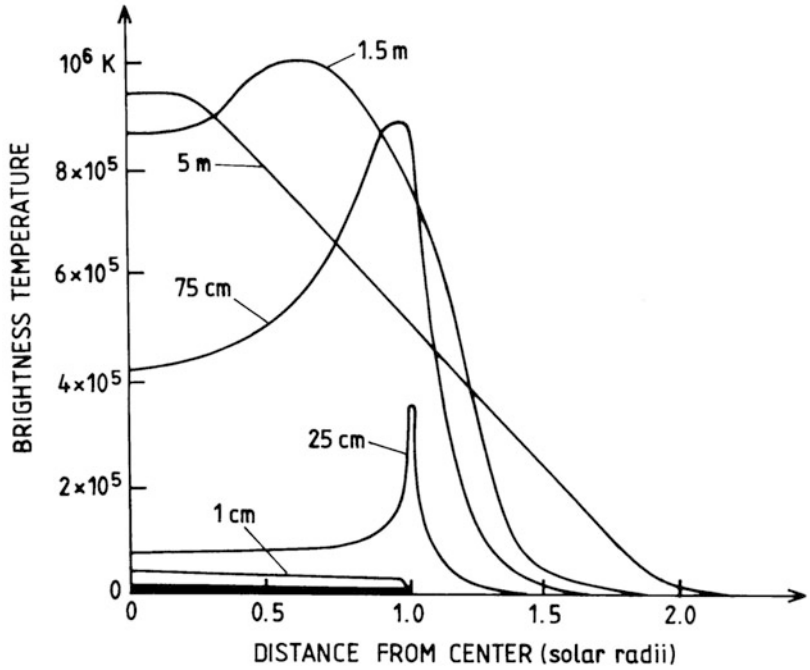


Figure 1.15. The variation of brightness temperature with distance from the center of the solar disk at various wavelengths of observation. The optical disk is indicated by the heavy black bar. (From Akasofu and Chapman, 1972.)

where f is the total number of sunpots regardless of size, g is the number of sunspot groups, and k is a normalization constant accounting for the different observatories. A typical plot of the yearly mean sunspot number is given in Figure 1.17. The most prominent feature is an apparent periodicity of the order of 11 years, but longer periods may also be identified. The period between two successive minima can vary from 8.5 to 14 years and there may be from 7.3 to 17 years between successive maxima. There is a general tendency for a more active cycle to have its maximum at an earlier phase of the cycle and for a less active cycle to have its maximum at a later phase (as shown in Figure 1.18). It is also noticeable that in general there is a rapid increase from a minimum to the following maximum and a slower decrease from the maximum to the following minimum.

From spectral analysis of variations in sunspot numbers it is found that the predominant period often depends on the length of the time series analyzed, provided that the time series is long enough to enclose a reasonable number of complete cycles. The time series is therefore not stationary as the different frequency components probably change phase in the course of time. At the moment, six periods appear to stand out from this analysis. These are listed as follows (in

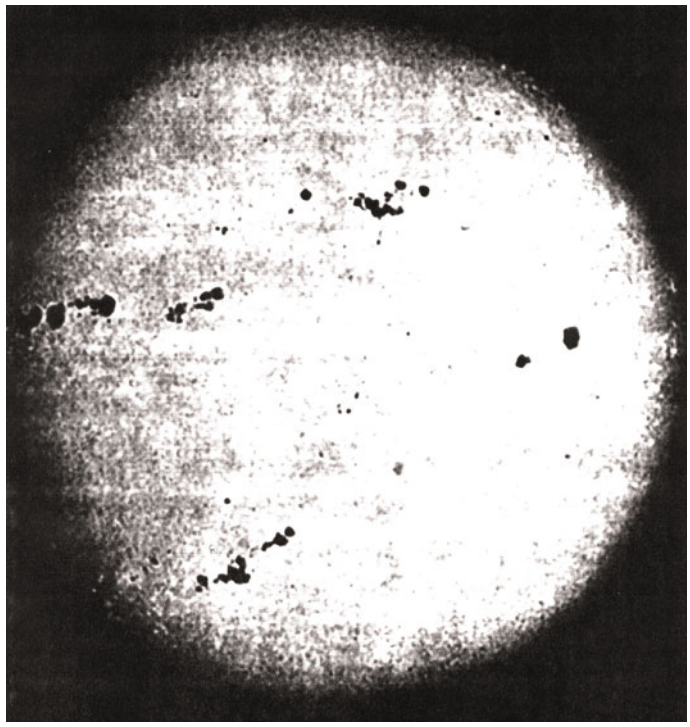


Figure 1.16. A photograph of the Sun near maximum sunspot activity (December 1957). The sunspot zones are clearly outlined. (From Zirin, 1966.)

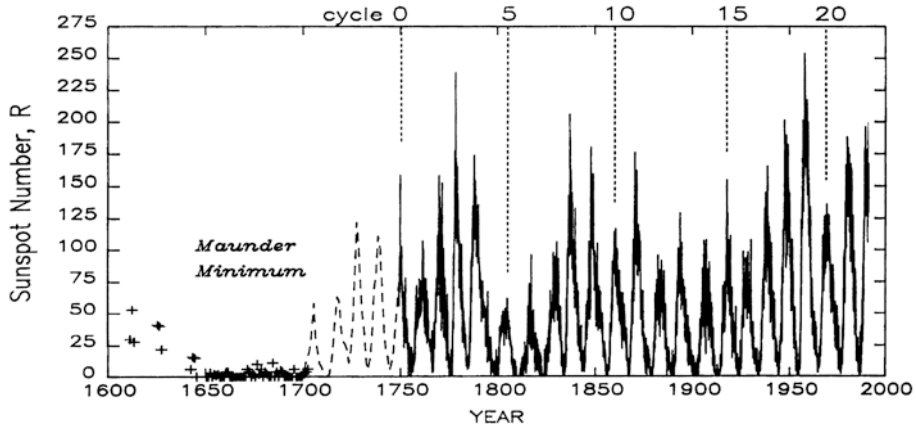


Figure 1.17. The sunspot number represented by monthly means since 1750 and yearly means from 1610 to 1750. The numbers indicated above represent the sunspot number starting from zero at the cycle maximizing in 1750. Cycle 1 is therefore the one starting at 1755. (From Lean, 1991.)

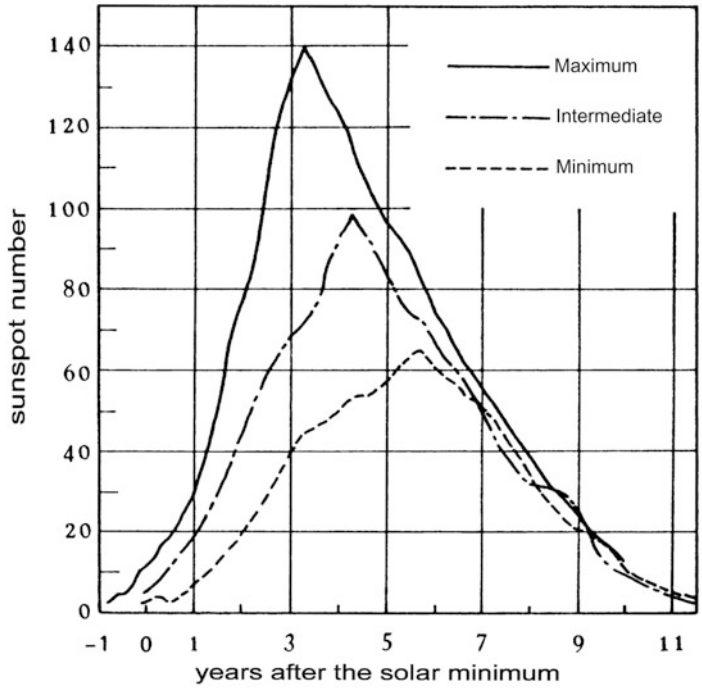


Figure 1.18. Differences in amplitude and phase of a solar cycle showing that stronger cycles reach maximum faster than weaker cycles. (From Waldmeier, 1955.)

years) (from Herman and Goldberg, 1978):

$$\begin{array}{ccc} & 5.5 & \\ 8.1 & 9.7 & 11.2 \\ & 100 & \\ & 180 & \end{array}$$

The three components 8.1, 9.7, and 11.2 years are believed to be a triplet giving rise to the commonly known 11-year solar cycle.

It has also been indicated that some of the periods actually result from beating between the more fundamental ones.

If T_1 and T_2 are two fundamental periods, we know from the coupling between waves that two other periods (T_3 and T_4) may occur as a result of beating, and these are given by the following relations:

$$\frac{1}{T_3} = \frac{1}{T_1} + \frac{1}{T_2}$$

$$\frac{1}{T_4} = \frac{1}{T_1} - \frac{1}{T_2}$$

The Sun itself rotates around the center of mass of the solar system with a period of 179.2 years or very close to the 180 years found in existing data. If we also consider the motion of the planets around the Sun and take into account the relative masses of the planets and their relative distances from the Sun, we can show that the gravitational vector force exerted by the planets on a fixed spot on the Sun varies cyclically with a period of 11.1 years.

If we take these two fundamental periods in the solar system and let $T_1 = 11.1$ years and $T_2 = 179.2$ years, we find the following beat periods:

$$T_3 = 10.4 \text{ years}$$

$$T_4 = 11.8 \text{ years}$$

If we then take into account these two beat periods and the first harmonic periods of T_1 and T_2 which are 5.5 and 89.6 years, respectively, we obtain the following set of periods:

$$\begin{array}{ccc} & 5.5 & \\ 10.4 & 11.1 & 11.8 \\ & 89.6 & \\ & 179.2 & \end{array}$$

These are reasonably close to the periods deduced from the time series of the sunspot number to be within the accuracy of observations.

We should remember that the sunspot number is an empirical quantity which gives us no direct link to the mechanism causing sunspots, and therefore we should not expect complete matching between theoretical values and deduced periods from sunspot series.

Many attempts have been made to study the solar cycle before 1700, and in Figure 1.19 a well-known graph by Eddy (1976) is reproduced. The graph indicates that the 11-year solar cycle started abruptly at about 1700 which is an artifact due to the lack of high-quality data before this time. According to Eddy (1976) the sunspot numbers (Wolf numbers) from 1818 to 1847 are good, from 1749 to 1817 are questionable, from 1700 to 1748 are poor, and before 1700 they are of course worse than poor. It is noticeable, however, that a long periodic variation appears to be interfering with the 11-year solar cycle leading to a strong reduction in sunspot numbers in the so-called Spörer minimum (1460–1550) and the Maunder minimum (1645–1715). To a large extent these data lean on studies of C^{14} concentration in tree rings. The exact relationship between behavior in the C^{14} concentration and sunspots is not well known. This is also clear for the period between 1800 and 1820 where the sunspot data are good and show two very weak cycles while the C^{14} results indicate a minor decrease and a minimum about 10 years later than the sunspot numbers.

It has become common practice to refer to each 11-year solar cycle measured from minimum to minimum by a number. The period from 1755 to 1766 has been assigned cycle 1. Figure 1.20 presents the smoothed sunspot numbers in cycle 22 and previous cycles.

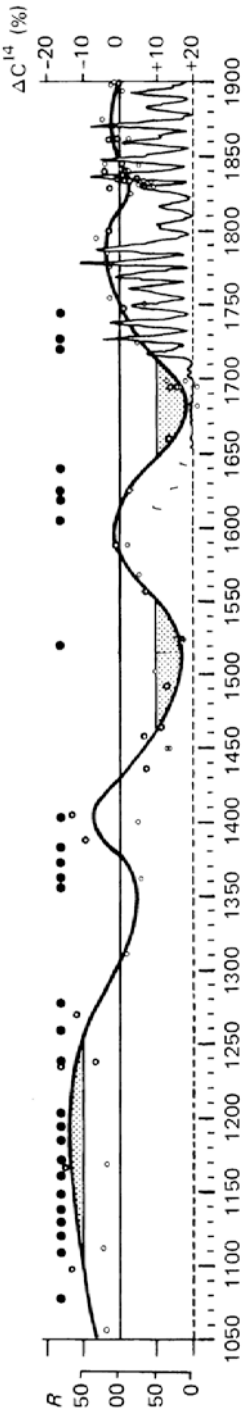


Figure 1.19. Historical sunspot variability. The heavy solid line indicates the variation in C^{14} concentration in tree rings. The thin curve refers to the Wolf sunspot numbers (R_z). Shaded portions of the C^{14} curve are the Spörer (1460–1550) and Maunder (1645–1715) periods. Circles are from historical oriental sunspot records and tree ring analyses. (From Eddy, 1976.)

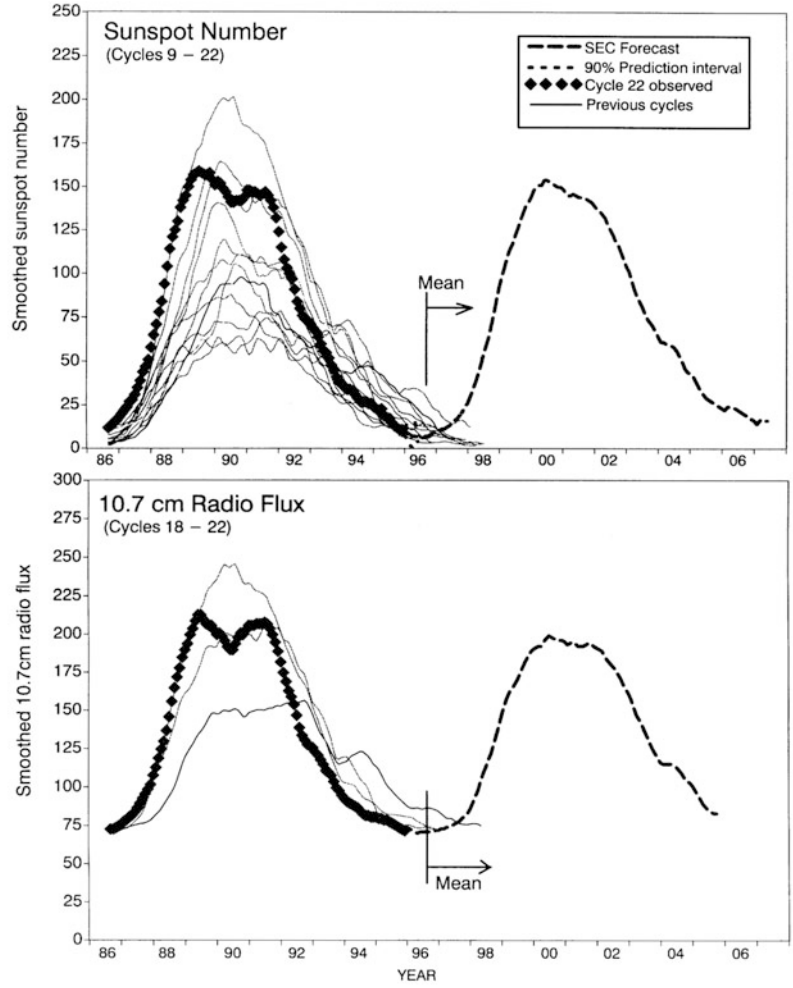


Figure 1.20. (Upper panel) The smoothed sunspot numbers (R) for cycle 22 beginning in September 1986, compared with the same for the preceding cycles. Also indicated are the mean sunspot numbers for cycles 9–22 as well as predictions for the following cycle. (Lower panel) The smoothed variations of the 10.7 cm radio flux ($F_{10.7}$) for cycle 22 compared with the same for cycles 18–22. (From Joint NOAA–USAF Space Weather Operations, 1996.)

The reason there is such a strong interest in sunspots is that conditions on the Earth appear to vary in concordance with the sunspot cycle. Many attempts have been made to show correlations between the solar cycle and different variable phenomena on the Earth, especially the weather. So far no conclusive results appear to have been obtained, and the discussion will continue with respect to this fundamental problem for life on Earth in the years to come. For scientists working with problems in the Earth's upper atmosphere (the ionosphere), solar

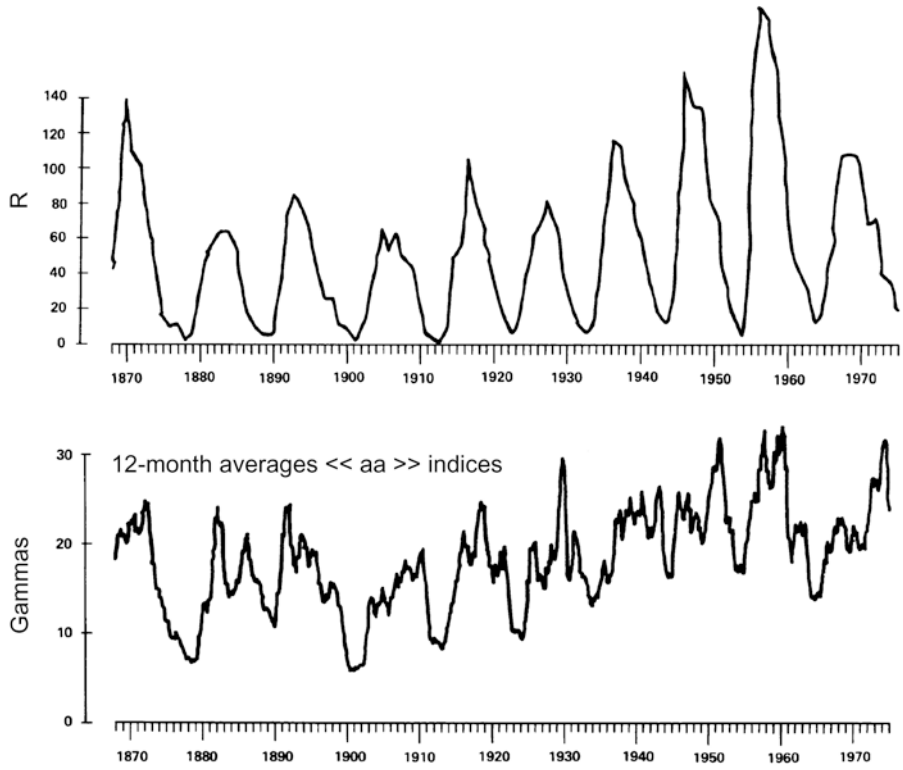


Figure 1.21. Comparison between solar sunspot numbers between 1850 and 1975 and the *aa* index for the same period. The *aa* index is the average amplitude in γ over 3-hour intervals of irregular variations of the horizontal components of the Earth's magnetic field measured at two antipodal sites (Europe and Australia) at about 50 degrees latitude. (From Giraud and Petit, 1978.)

cycle variations stand out very clearly. For example, as shown in Figure 1.21, a close correlation exists between variations in sunspot numbers and in variations of the Earth's geomagnetic field as evidenced by the so-called *aa* index. These variations in the Earth's field are due to currents in the ionosphere, which are modulated by radiation from the Sun. We will return to these currents in Chapter 6.

The spots first appear as a small dark pore which gradually grows bigger and develops after a day or so into the characteristic structure with an inner dark umbra surrounded by the penumbra (Figure 1.22). Around 500 nm the umbra intensity is only 3–5% of the photospheric intensity, while in the penumbra it is about 80%. The effective temperature in the umbra of a large spot is of the order of 4,500 K. That is to say, the spot appears cooler than its surroundings, which are closer to 6,000 K. The lifetime of the spots is highly variable, but on average there is a positive correlation with the maximum area that a spot covers. Half of all spots have lifetimes less than 2 days while only 10% survive for more than 11

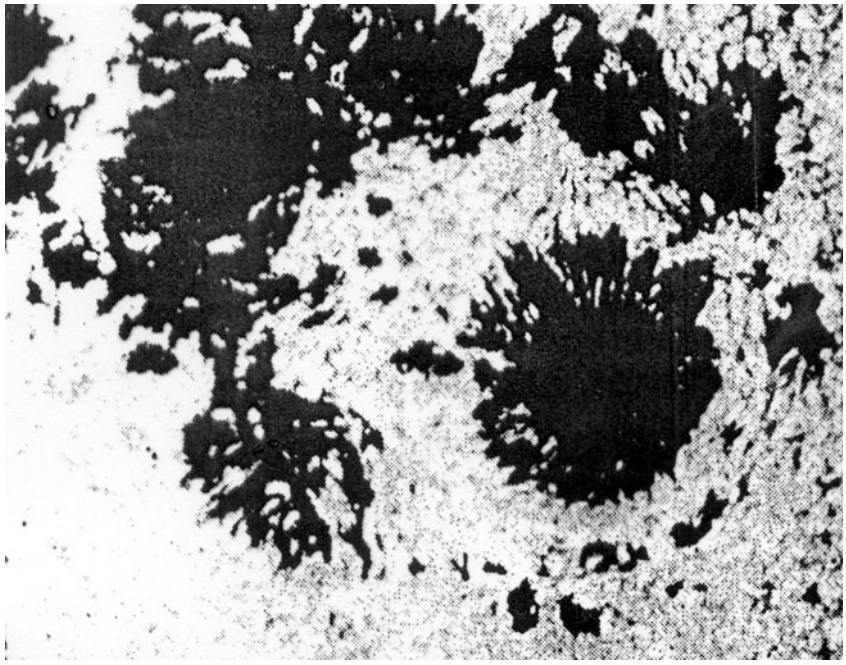


Figure 1.22. A photograph of a regular sunspot and a complex sunspot group. (From Danielson, 1961.)

days. Only exceptionally has a spot been observed for more than five solar rotations.

Comparison between variations in sunspot numbers and cosmic ray flux as observed on the ground at Apatity in the Kola peninsula (as presented in Figure 1.23) demonstrates a strong anticorrelation between the two data series. As it is believed that cosmic rays have an impact on the level of ionization in the troposphere with subsequent effects on the formation of water vapor, it is possible that solar activity may play an indirect role in global climate due to the solar wind modulation of cosmic ray flux.

Spots often appear in groups where spots of one magnetic polarity cluster in the preceding part, while a group with opposite polarity clusters in the trailing part. Such a set of spots is called a bipolar sunspot group. More than 90% of all spots appear in this category. Unipolar groups only account for 9% of the spots.

Complex groups with many small spots and no apparent clear distribution of magnetic polarity represent less than 0.5% of the total number of spots. This group is very important geophysically, however, as the probability of flare and disturbance activity increases markedly with the complexity of the group.

The distribution of spots over the solar surface changes with phase in the solar cycle (as illustrated by the "butterfly diagram" in Figure 1.24). Shortly before a solar minimum is reached new spots appear at heliocentric latitudes

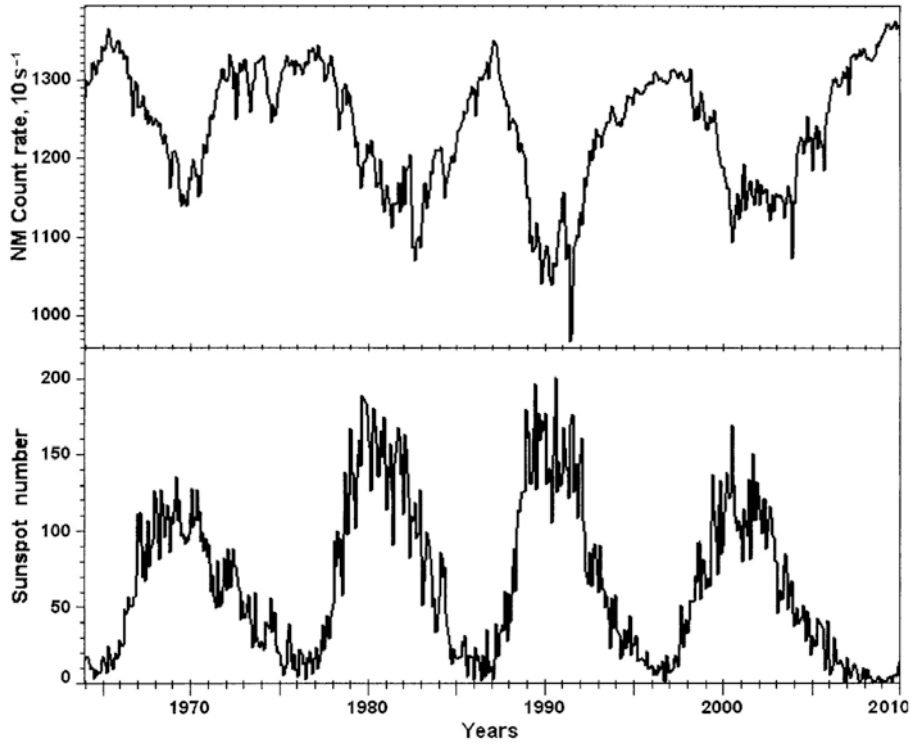


Figure 1.23. Comparison between variations in sunspot numbers and cosmic ray flux as measured at the ground in the Kola peninsula. (Polar Geophysical Institute, Kola Branch, private communication.)

around $\pm 30^{\circ}$. As the number of spots increases, the spots appear at lower latitudes and at solar maximum they are at an average latitude of $\pm 12^{\circ}$. The average latitude decreases further and reaches about $\pm 5^{\circ}$ at sunspot minimum.

The magnetic polarity of spots changes from cycle to cycle. It appears that within the same cycle the leading spots have the same polarity in the same hemisphere but opposite in the other hemisphere (Figure 1.25). At minimum the polarity changes, and sometimes new spots at high latitudes with a changed polarity may be seen together with spots from the old cycle at low latitude.

Because of this change in polarity the solar cycle has a characteristic double period close to 22 years (as shown in Figure 1.26).

1.8 ELECTROMAGNETIC RADIATION FROM A DISTURBED SUN

The electromagnetic radiation from a quiet Sun is to be explained as thermal radiation from the corona and the solar atmosphere. The variations in these background emissions from X-rays to radiowaves are rather smooth and follow

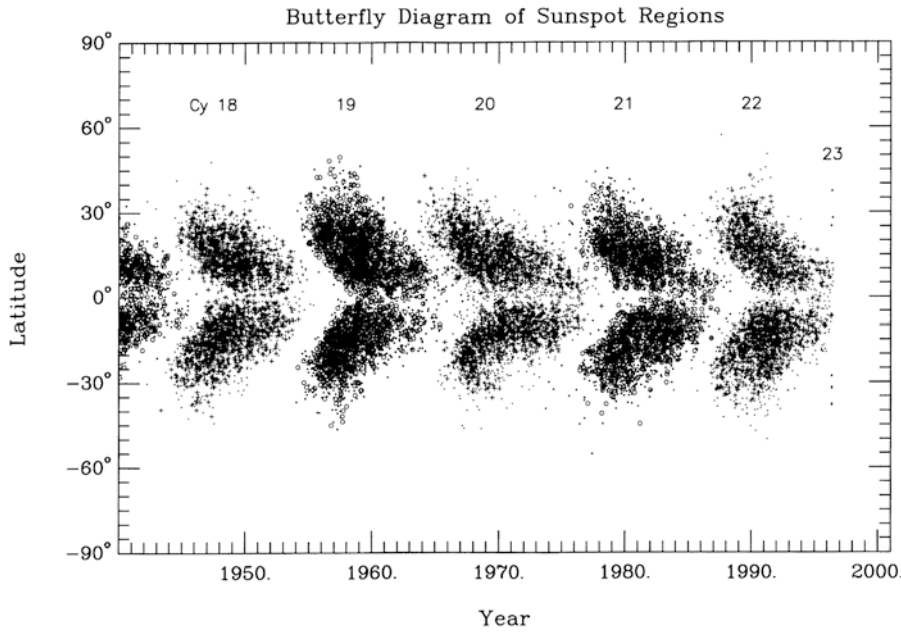


Figure 1.24. Butterfly diagrams from 1940 until 1996 showing the latitudinal positions of sunspots marked for solar cycles 18–23. This illustrates the equatorward movement of the active latitude band over each solar cycle. (From Joint NOAA–USAF Space Weather Operations, 1996.)

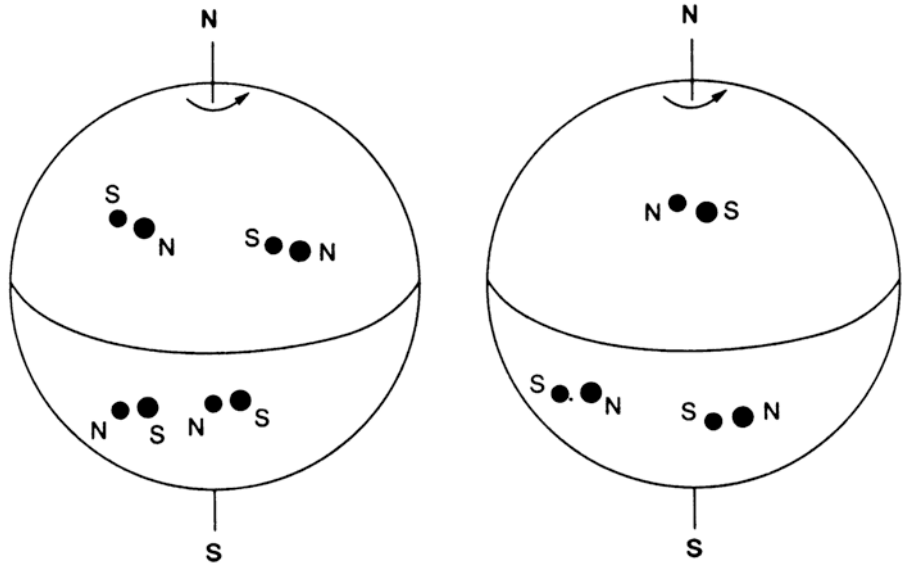


Figure 1.25. A schematic diagram showing the relationship between magnetic polarity and sunspots for consecutive solar cycles. (From Egeland *et al.*, 1990.)

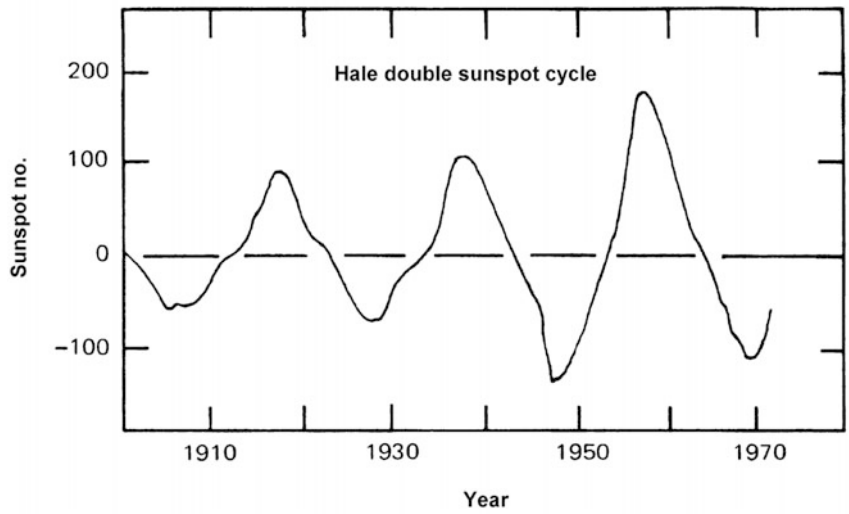


Figure 1.26. The double sunspot cycle with alternate maxima plotted with opposite signs to indicate the 22-year period related to the reversal of magnetic polarity from one cycle to the next. (From Roberts, 1975.)

the 27-day recurrence tendency corresponding to the rotation period of the Sun as seen from Earth. Figure 1.26 illustrates this behavior rather well where the emission at X-rays between 0.8 and 1.2 nm ($E(8, 12)$) are compared with radiowave emissions at 10.7 cm (2,800 MHz). Also noticeable in Figure 1.27 is the close correlation with the solar sunspot number indicated by R_z as well as the plage index ($\Sigma A \times I$) with these emissions. Superimposed on these smooth variations, however, are considerable fluctuations with short time scales especially in X-ray emissions. Similar rapid fluctuations are also found in UV emissions from the Sun.

The most short-lived phenomenon observed from Earth on the solar surface is the flare event which usually occurs within a region encompassed by a large magnetically complex bipolar sunspot group. It lasts usually from a few minutes to a few hours. Its occurrence is defined by a sudden brightening of the H Ly_α line, and it is usually accompanied by enhanced X-ray emissions (Figure 1.28).

The X-ray and UV energy from flares arrive at the Earth in about 8 min and are known to produce intense ionization in the upper atmosphere which is manifested in radiowave fade-out and sudden cosmic noise absorption.

The occurrence frequency of small flares is positively correlated with the 11-year solar cycle, and although the frequency of major flares also tends to peak at sunspot maxima, they can occur at any time. There is some evidence for a double peak in major flare occurrence, the second one appearing a few years after sunspot maximum.

Radio emissions from solar flares have little energy compared with X-ray and EUV emissions, but as indicators of fast particles streaming out of the solar atmosphere they are of great importance.

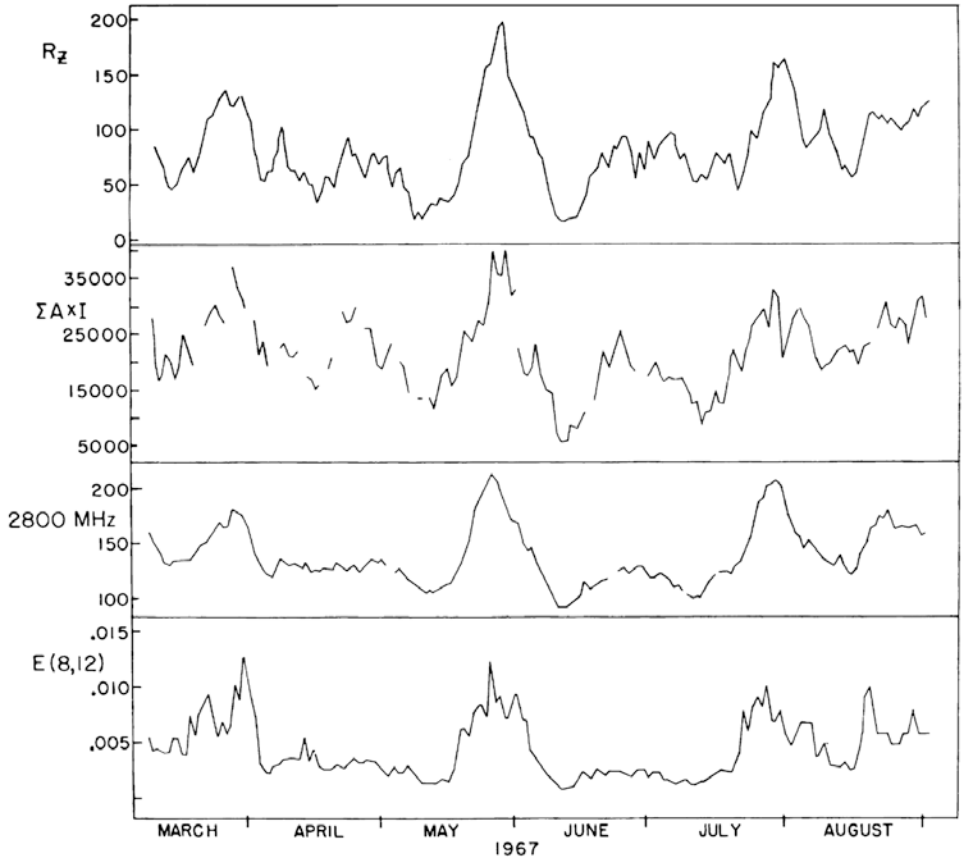


Figure 1.27. Intercomparison between time variations in X-ray flux between 0.8 and 1.2 nm ($E(8, 12)$), 10.7 cm flux density, the plage index ($\Sigma A \times I$), and solar sunspot number R_z showing the 27-day recurrence period. (From Teske, 1969.)

In a time frequency diagram where intensity is included as a third dimension (as is done in Figure 1.28), so-called radio bursts from the Sun are illustrated and classified as types from I to V. Since electron density decreases outward in the solar atmosphere, as does plasma frequency, we may conclude that if a burst is due to a plasma wave decreasing in frequency with time, the source of the burst moves outward through the corona. The velocity of the wave can then actually be derived from the rate of change of the observed burst frequency.

Let us assume by inspecting Figure 1.14 that electron density to a first approximation decays exponentially by distance from the solar surface, then the electron density $n_e(z)$ at a distance z from the solar surface is given by:

$$n_e(z) = n_{e0} \exp\left(-\frac{z - z_0}{H}\right)$$

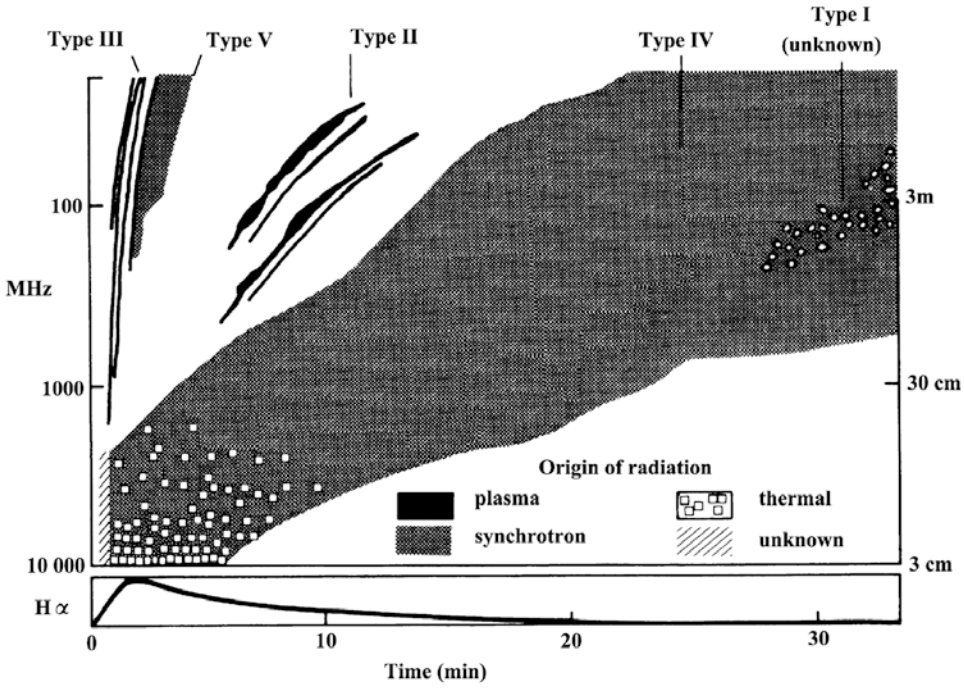


Figure 1.28. Idealized sketch showing the dynamic spectrum of solar radiowave bursts. (From Wild *et al.*, 1963.)

where n_{e0} is the electron density at a reference height z_0, and H is the e-folding scale height. From the expression of plasma frequency (1.17) we find that

$$\frac{df_{pe}}{dt} = K \cdot \tfrac{1}{2}[n_e(z)]^{-1/2} \frac{dn_e(z)}{dz}\frac{dz}{dt} = -\frac{f_{pe}}{2H}\frac{dz}{dt}$$

and the propagation velocity v_{RS} of the radio source emitting a radio frequency f equal to the instantaneous plasma frequency f_{pe} is then given by:

$$v_{RS} = \frac{dz}{dt} = -2H \cdot \frac{d}{dt}(\ln f\,)$$

By observing the rate of change of the received frequency by time and for a given model of the electron density profile in the solar atmosphere, the speed of the radio emission source can be found.

During the first minutes of development after the flare flash has occurred, groups of very narrow bursts appear that show a fast drift from high to low frequency. These bursts, called Type-III bursts, are supposed to be carried by relativistic electrons moving upward in the corona at speeds between 0.2 and 0.9 times the speed of light. These bursts, assumed to be due to synchrotron radiation, probably originate between 50 and 100 $R_\odot$ from the Sun according to their original frequencies (100–300 MHz). A continuum, appearing in the frequency–

time diagram as a diffuse extended region called Type-V bursts, is sometimes observed 1–3 min after the Type-III burst.

Another type of burst showing much slower drift velocities than the Type-III burst is called the Type-II burst which usually starts between 75 and 100 MHz. They can last from 5 to 10 minutes and typically lag 5 minutes behind Type-III bursts. The velocity in the source region through the corona for Type-II bursts is between 500 and 1,500 km/s. Since the speed of sound in this part of the corona is typically 175 km/s, these waves are supersonic and usually interpreted as electromagnetic radiation caused by shock waves.

The last type of burst accompanying flares is the Type-IV burst. They have a broad bandwidth (continuum) and last from 10 min up to hours after the flare. It is generally assumed that this radiation is due to relativistic particles trapped in the magnetic field of the active region.

Type-I bursts may last for hours and days and are not related to flares. They consist of an enhanced background continuum with short-lived (0.5 s) bursts superimposed. These noise bursts (Type I) have their origin close to large sunspot groups. It is found that most Type-I events are preceded by flare events; however, there is no clear connection between storm characteristics and flare importance. The way Type-I events are created is uncertain.

1.9 PARTICLE EMISSIONS FROM THE SUN

A quiet Sun radiates not only electromagnetic waves, but also particles. A solar wind is always blowing. This wind is not a steady wind, but rather gusty as it varies quite strongly in velocity as well as density at a distance from the Sun corresponding to the Earth's orbit.

Under quiet conditions the wind has a medium velocity of 400 km/s but can, however, vary between 200 and 700 km/s (Figure 1.29). Note that with a speed (v) of 435 km/s the solar wind takes 4 days to reach the Earth from the Sun. The solar wind particle density (n) can vary between 10^6 and 2×10^7 m^{-3}. The average energy of the protons is of the order of 1 keV while the electrons have energies of the order of 1 eV. The average particle flux from the Sun can then be estimated to be:

$$\phi = n \cdot v \approx 2 \times 10^{12} \text{ m}^{-2} \text{ s}^{-1}$$

Total average particle loss per unit time from the Sun is therefore

$$\dot{N} = 4\pi R_{\odot}^2 \cdot \phi \approx 1.23 \times 10^{31} \text{ s}^{-1}$$

where $R_{\odot}$ is the solar radius. By neglecting the mass of the electrons in the solar wind, this particle loss per unit time corresponds to a mass loss per unit time given by:

$$\dot{m} = \dot{N} m_p = 2.21 \times 10^5 \text{ kg/s}$$

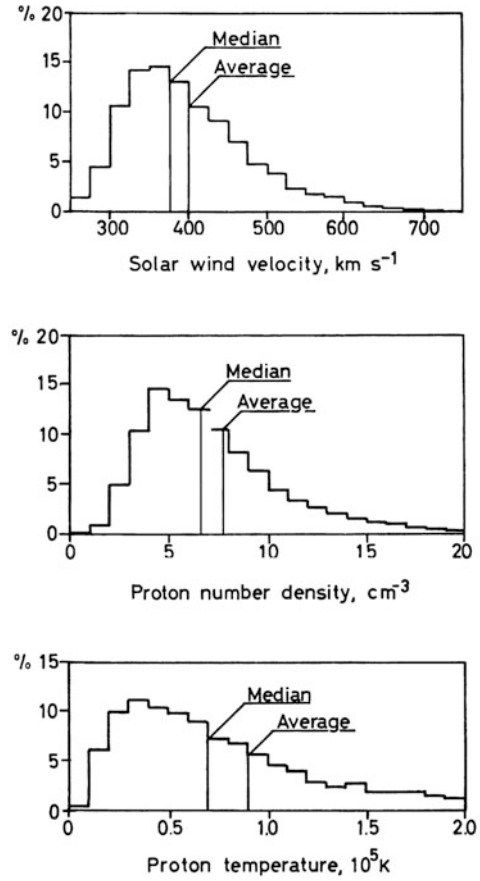


Figure 1.29. Histograms of occurrence frequency for the values of solar wind velocity, proton number density, and proton temperature in interplanetary space. (From Hundhausen *et al.*, 1970.)

where $m_p = 1.672 \times 10^{-26}$ kg is the proton mass. This can be compared with the mass loss per unit time due to the blackbody radiation from the Sun and equal to 4.3×10^9 kg/s as derived in equation (1.7).

The proton temperature in the solar wind is $10^4 - 2 \times 10^5$ K while the electron temperature is a factor of 3–4 larger under quiet average conditions. The magnetic field strength associated with the solar wind varies between 1 and 15 γ ($1\,\gamma = 10^{-9}$ tesla) (Figure 1.30). Temperature is always observed higher along the magnetic field ($T_\parallel$) than perpendicular to it ($T_\perp$). On average $T_\parallel \approx 2T_\perp$.

Of the characteristic parameters we have given for the solar wind so far, we notice that the speed of sound of the solar wind gas is approximately:

$$c_s = \sqrt{\gamma \frac{kT_\odot}{m_p}} = 1.17 \times 10^4 \text{ m/s}$$

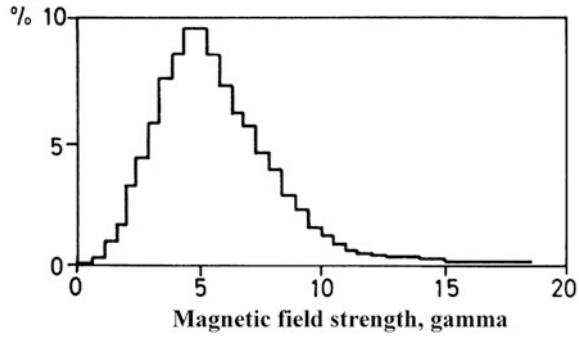


Figure 1.30. Histogram of occurrence frequency for magnetic field strength values in interplanetary space. (From Ness, 1969; e.g., Fälthammar, 1973.)

where γ is the adiabatic constant equal to 5/3 for a monoatomic gas and $T_\odot$ is set equal to 10^4 K.

Since the solar wind speed is more than 30 times higher on average, the solar wind is supersonic.

This high-flow wind is produced by the pressure difference between the hot dense gas of the solar atmosphere and the cold tenuous gas in the background interstellar medium. This pressure difference easily overcomes the solar gravitational pull on the plasma. Also the high electrical conductivity of the plasma prevents motions across the magnetic field lines and these lines are therefore frozen into the plasma as a "glue" and force the plasma to act collectively as a hydrodynamic fluid. Since these field lines are anchored in the solar surface while the wind plasma moves radially outward, the lines of force will form Archimedean spirals making an angle of about 45° with the radius vector to the Sun at the Earth's orbit.

The density of the solar wind gas, as it expands spherically outwards from the Sun, must decrease as $1/r^2$ where r is the distance from the Sun. At a heliocentric distance r_h the pressure in the solar wind must balance the pressure of the interstellar medium, and the solar wind must again be subsonic. The change from a supersonic plasma to a subsonic plasma is expected to be related to a shock wave, and at this shock the interplanetary magnetic field terminates, forming the boundary of the heliosphere, the heliopause. It is at the moment impossible to determine the distance to this shock, but it is believed to be situated between 60 and 100 AU.

1.10 FLUID FLOW IN A NOZZLE

To get an idea of the mechanism behind the acceleration of the solar wind, we will first discuss fluid flow through a nozzle.

Let an incompressible gas with density ρ stream through a tube with a varying cross-section A (Figure 1.31). Since mass flux through any cross-section of the

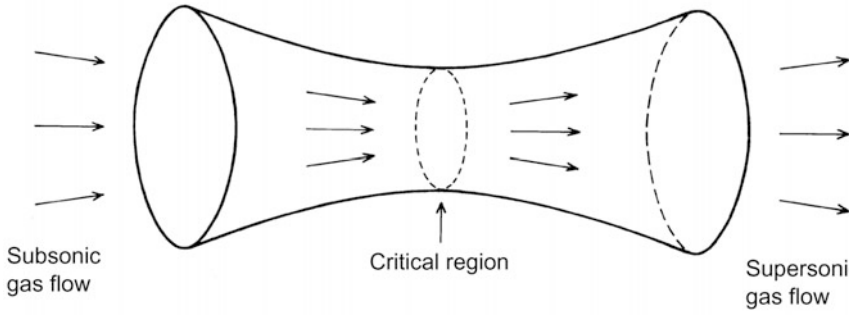


Figure 1.31. Mass flow through a nozzle with a minimum cross-section to explain the presence of a critical region in the mass flow in order for the flow speed to become supersonic.

tube must be constant, we have:

$$\phi_m = A \cdot \rho \cdot v = \text{const.} \tag{1.18}$$

where v is the velocity of the gas through the cross-section.

The pressure gradient will be balanced by the inertia force when no other forces are acting in the equation of motion. Considering only one dimension along r:

$$\frac{dp}{dr} = -\rho \frac{dv}{dt} = -\rho \frac{dv}{dr} \cdot v$$

and by dividing by ρ and rearranging:

$$\frac{dp}{\rho} = -v\, dv$$

Rewriting dp by another partial derivative we have:

$$\frac{dp}{\rho} = \frac{dp}{d\rho} \cdot \frac{d\rho}{\rho} = -v \cdot dv \tag{1.19}$$

and

$$\frac{d\rho}{\rho} = -v\, dv \left(\frac{dp}{d\rho}\right)^{-1} \tag{1.20}$$

If we now assume that this process of the gas flow is adiabatic, we have:

$$p \cdot \rho^{-\gamma} = \text{const.}$$

where γ is the adiabatic constant. Then by differentiating:

$$\frac{dp}{d\rho} = \gamma \frac{p}{\rho} = c_s^2 \tag{1.21}$$

where c_s is the speed of sound in the gas. By inserting (1.21) into (1.20)

$$\frac{d\rho}{\rho} = -\frac{v}{c_s^2} dv \tag{1.22}$$

According to the conservation of mass flux (1.18) we can form:

$$\frac{d\phi_m}{\phi_m} = \frac{dA}{A} + \frac{d\rho}{\rho} + \frac{dv}{v} = 0 \tag{1.23}$$

By inserting for $d\rho/\rho$ from (1.22) we get:

$$\frac{dA}{A} - \frac{v}{c_s^2}dv + \frac{dv}{v} = 0$$

and

$$\frac{dA}{A} = \left(\frac{v}{c_s^2} - \frac{1}{v}\right)dv = \left(\frac{v^2}{c_s^2} - 1\right)\frac{dv}{v} \tag{1.24}$$

If v is supersonic ($v > c_s$) from the start, v has to decrease as long as A decreases, and v will become subsonic.

If, on the other hand, v is subsonic at the start it can only increase and become supersonic if $v = c_s$ when $dA = 0$ and the cross-section increases again. This is illustrated in Figure 1.31. If, however, the gas is streaming so slowly that it does not reach the speed of sound at the narrowest part of the tube, the speed of the gas will then continue to decrease as it passes through the throttle.

1.11 THE SOLAR WIND EQUATION

We will now assume that the coronal gas is an ideal and incompressible gas and that we can treat it as a hydrodynamic fluid. This means that the collisions are frequent enough that local thermodynamic equilibrium is maintained. We will in the following treatment neglect viscosity and magnetism and limit ourselves only to pressure, internal forces, and gravity. Then (1.19) takes the form

$$\frac{dp}{\rho} = \frac{dp}{d\rho}\frac{d\rho}{\rho} = -v\,dv - \frac{GM_\odot}{r^2}dr \tag{1.25}$$

where r is the radial distance from the Sun, G the constant of gravity ($= 6.67 \times 10^{-11}\ \mathrm{N\,m^2\,kg^{-2}}$), and $M_\odot = 2 \times 10^{30}$ kg is the solar mass. By inserting $dp/d\rho = c_s^2$ from (1.21), (1.25) takes the form

$$c_s^2\frac{d\rho}{\rho} = -v\,dv - \frac{GM_\odot}{r^2}dr$$

Dividing by c_s^2 gives:

$$\frac{d\rho}{\rho} = -\frac{v}{c_s^2}dv - \frac{GM_\odot}{c_s^2}\frac{dr}{r^2} \tag{1.26}$$

From (1.23) we have

$$\frac{d\rho}{\rho} = -\left(\frac{dA}{A} + \frac{dv}{v}\right) \tag{1.27}$$

Inserting for $d\rho/\rho$ from (1.27) into (1.26) and solving for dA/A we derive:

$$\frac{dA}{A} = \left(\frac{v^2}{c_s^2} - 1\right)\frac{dv}{v} + \frac{GM_\odot}{c_s^2}\frac{dr}{r^2}$$

The solar wind is expanding spherically and symmetrically so that:

$$\frac{dA}{A} = \frac{2}{r}dr$$

This leads us to the following relationship between v and r:

$$\left(2 - \frac{GM_\odot}{c_s^2}\frac{1}{r}\right)\frac{dr}{r} = \left(\frac{v^2}{c_s^2} - 1\right)\frac{dv}{v}$$

and finally

$$2\left(1 - \frac{r_c}{r}\right)\frac{dr}{r} = \left(\frac{v^2}{c_s^2} - 1\right)\frac{dv}{v} \tag{1.28}$$

where $r_c = GM_\odot/2c_s^2$ is the so-called critical distance. There are several solutions to this equation (Figure 1.32).

1. If the wind starts out from the Sun with an increasing velocity which initially is smaller than c_s, then both sides in (1.28) are negative. When r equals the critical distance r_c, the left-hand side is zero and therefore the right-hand side has to be zero also (i.e., $v = c_s$). If $v = c_s$ at the critical distance, then v can continue to accelerate to a supersonic speed.
2. The velocity is lower than c_s at the surface of the Sun, but is less than c_s at the critical distance (it is not accelerated fast enough). The left-hand side of (1.28) will then be positive past the critical point, and the right-hand side can only become positive here if the velocity decreases. The wind ends up as a breeze.

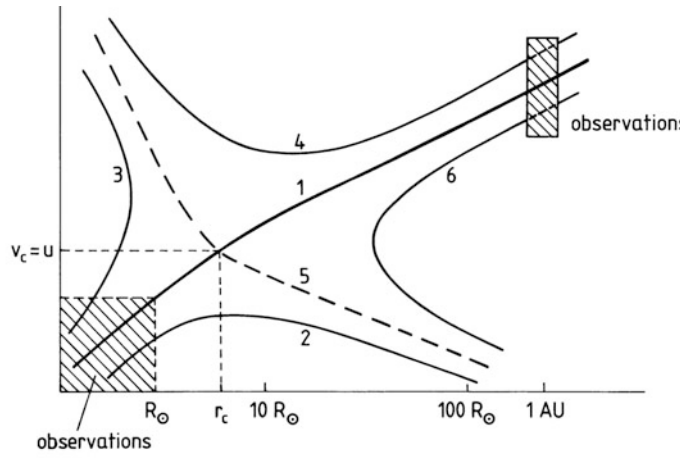


Figure 1.32. Alternative solutions of the solar wind equation explained in the text.

3. The velocity is subsonic at the Sun's surface, but is accelerated so fast that it is supersonic before it reaches the critical distance. The right-hand side of (1.28) then becomes positive, and the left-hand side can only be positive if r decreases. The wind blows back to the Sun supersonically.
4. The wind blows out from the Sun at supersonic speed. The left-hand side of (1.28) is negative, and the only way the right-hand side can be negative is for the velocity to decrease. At the critical distance the left-hand side of the equation changes sign. The only way the right-hand side can change sign, as long as v is supersonic, is when the velocity increases again.
5. If the wind starts supersonically at the Sun, the left-hand side of (1.28) is negative, and in order for the right-hand side to become negative, the velocity must decrease. At the critical distance the left-hand side changes sign and becomes positive. If at the same instance the velocity has decreased to c_s, it must continue to decrease in order for the right-hand side to become positive.
6. Finally, there is a solution for inward flow in the corona. If the flow starts at infinity blowing subsonically toward the Sun, the left-hand side of (1.28) is negative, and the right-hand side can only become negative if the velocity increases. When the velocity reaches c_s at some distance from the Sun, larger than r_c, the right-hand side changes sign, and for the left-hand side to become positive, r must increase again.

The one solution that connects the observations at the Sun and at the Earth's distance from the Sun is solution 1 which is the supersonic solar wind.

It is worth noticing that the solar temperature is just within the range which makes the supersonic solar wind possible. If the solar atmosphere is too cool, it will remain static. If it is too hot, the corona will expand subsonically as a solar breeze.

We notice from (1.26) that, as long as the velocity is small, the first term on the right-hand side can be neglected compared with the gravity term and then:

$$\frac{1}{\rho}dp = -\frac{GM_\odot}{r^2}dr$$

The atmosphere is in static equilibrium. On the other hand, if the gravity term can be neglected, then from (1.28) we have:

$$2\frac{dr}{r} = \left(\frac{v^2}{c_s^2} - 1\right)\frac{dv}{v}$$

For outward motion, dr and the left-hand side are positive. If the wind starts out subsonically from the Sun, it will decrease outward in order to maintain the same signs in the equation, and the solar wind will be a "breeze".

1.12 THE "FROZEN-IN" FIELD CONCEPT

Observed features that support the existence of solar magnetic fields are the loop prominences observed in the H_α line (Figure 1.33). Tentative measurements give magnetic fields of the order of 50 gauss (1 gauss $= 1\,\Gamma = 10^{-4}$ tesla). It is believed that the solar magnetic field is rather complicated and that local features can change dramatically within minutes. Loop prominences, for example, can be seen to last typically 20 minutes.

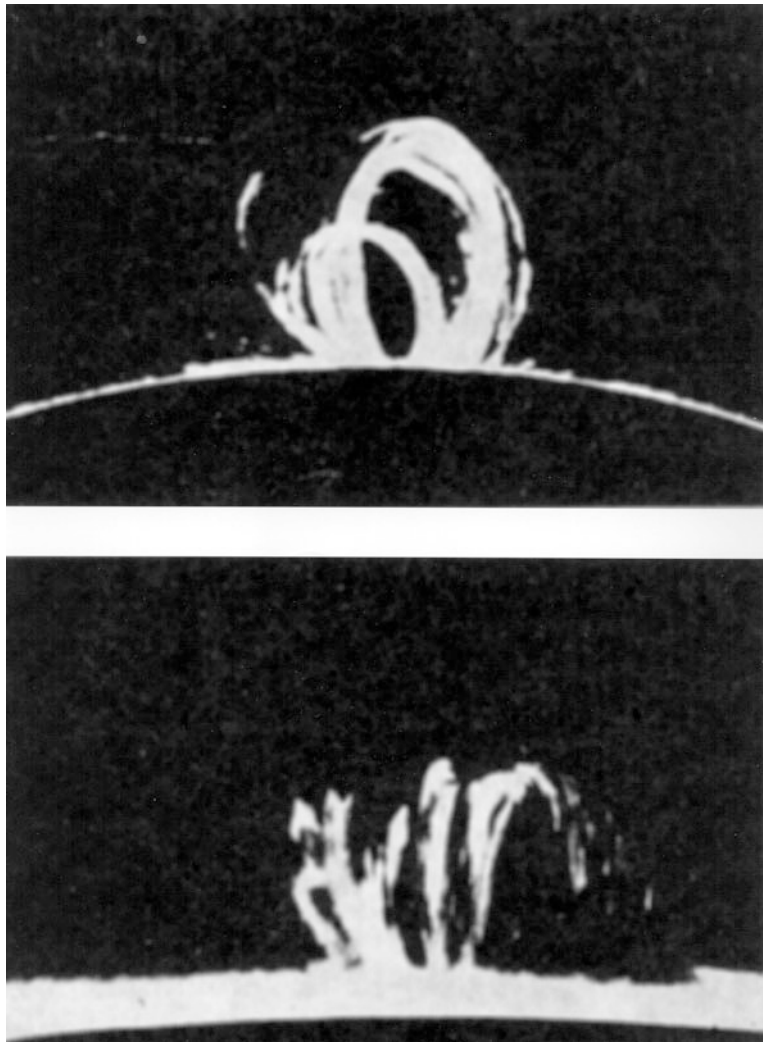


Figure 1.33. Loop prominences on the Sun photographed in the H_α emission. (From Zirin, 1966.)

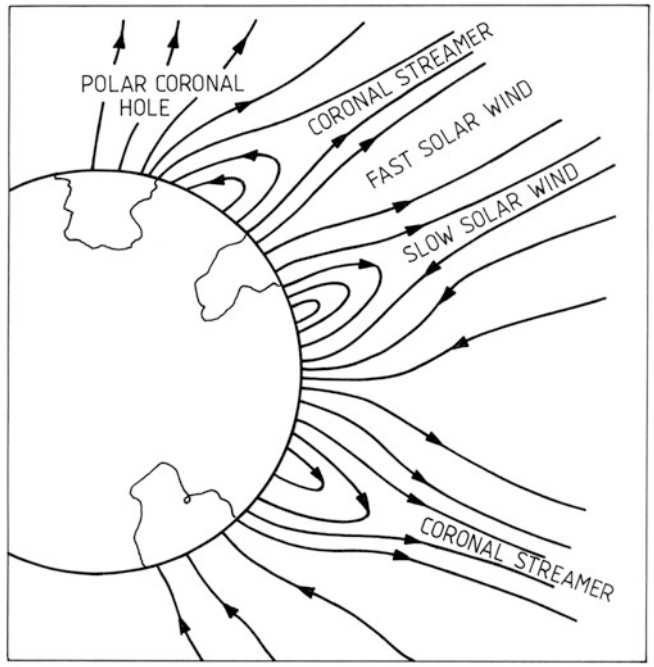


Figure 1.34. A schematic illustration of the different magnetic field structures believed to be present at the Sun. Some of the field lines are pulled into space by the solar wind, while the boundary between neighboring inward- and outward-pointing field lines must be penetrated by concentrated layers of electric currents. (After Dryer, 1987.)

An attempt to draw the gross field configuration of the Sun is shown in Figure 1.34. What is important to notice here is the separated regions where the magnetic field points toward or away from the solar surface. In the region between inward- and outward-pointing magnetic field lines there must be concentrated layers of electrical currents. At other regions a slow solar wind may occur or coronal streamers may be formed.

As already mentioned, the plasma or gas of the solar wind has high electrical conductivity, and therefore it carries the magnetic field along as a "glue" in the plasma. We will now explain the phenomenon of "frozen-in" magnetic fields in more detail. We will do this because the frozen-in magnetic field concept turns out to be a useful tool in understanding plasma flow in the solar wind, magnetosphere, and upper ionosphere. For a true conducting fluid the current density is given by $\mathbf{j}' = \sigma \mathbf{E}'$ where $\mathbf{j}'$ and $\mathbf{E}'$ are currents and electric fields in the frame moving with the plasma. Since the plasma is highly conductive, conductivity takes on this very simple scalar form, σ, and not the usual tensor form. From another frame of reference where the plasma is moving with velocity $\mathbf{v}$, the electric field is $\mathbf{E} = \mathbf{E}' - \mathbf{v} \times \mathbf{B}$, the current, however, is conserved, $\mathbf{j} = \mathbf{j}'$, and so is the magnetic field, $\mathbf{B} = \mathbf{B}'$, if the velocity is small enough ($v \ll c$). Thus, from the generalized

Ohm's law:

$$\mathbf{j} = \sigma \mathbf{E}' = \sigma(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \tag{1.29}$$

From this equation we find:

$$\mathbf{E} + \mathbf{v} \times \mathbf{B} = \frac{\mathbf{j}}{\sigma}$$

and notice that when $\sigma \to \infty$

$$\mathbf{E} = -\mathbf{v} \times \mathbf{B} \tag{1.30}$$

which is the so-called frozen-in condition implying that no current is flowing in the medium.

We now want to investigate further what other physical consequences the frozen-in conditions imply. From Maxwell's equations we have:

$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E}$$

By introducing (1.30) we derive for an infinite σ:

$$\frac{\partial \mathbf{B}}{\partial t} = \nabla \times (\mathbf{v} \times \mathbf{B}) \tag{1.31}$$

If we now consider the time variation of the magnetic flux through a surface moving with velocity v, given by:

$$\frac{d\phi}{dt} = \iint_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{s} + \oint_L \mathbf{B} \cdot (\mathbf{v} \times d\ell) \tag{1.32}$$

where S is the surface, and L is a closed loop encircling this surface. The first integral yields the change in ϕ due to the time variation of $\mathbf{B}$, while the second integral yields the change in ϕ due to the motion of the surface across $\mathbf{B}$.

The second term in (1.32) can be rewritten:

$$\oint_L \mathbf{B} \cdot (\mathbf{v} \times d\ell) = -\oint_L (\mathbf{v} \times \mathbf{B})\, d\ell = -\iint_S \nabla \times (\mathbf{v} \times \mathbf{B})\, d\mathbf{s}$$

and therefore, by inserting this into (1.32):

$$\frac{d\phi}{dt} = \iint_S \left[\frac{\partial \mathbf{B}}{\partial t} - \nabla \times (\mathbf{v} \times \mathbf{B}) \right] ds \tag{1.33}$$

We have, however, just shown in (1.31) that for an infinite value of σ, the expression in the integral vanishes so that $d\phi/dt = 0$ and ϕ is a constant. The flux through the surface is kept, and the magnetic field is therefore frozen-in. Remember, however, that strictly speaking this is correct only when $v \ll c$.

The concept of the frozen-in magnetic field is a very useful one. If a plasma sticks to a magnetic field line, then it will continue to do so as long as the frozen-in condition is valid. This allows visualization of complex plasma flow if we know the magnetic field configuration, and conversely if the motion of the plasma is known, the field geometry can be deduced. On the other hand, if a highly conduct-

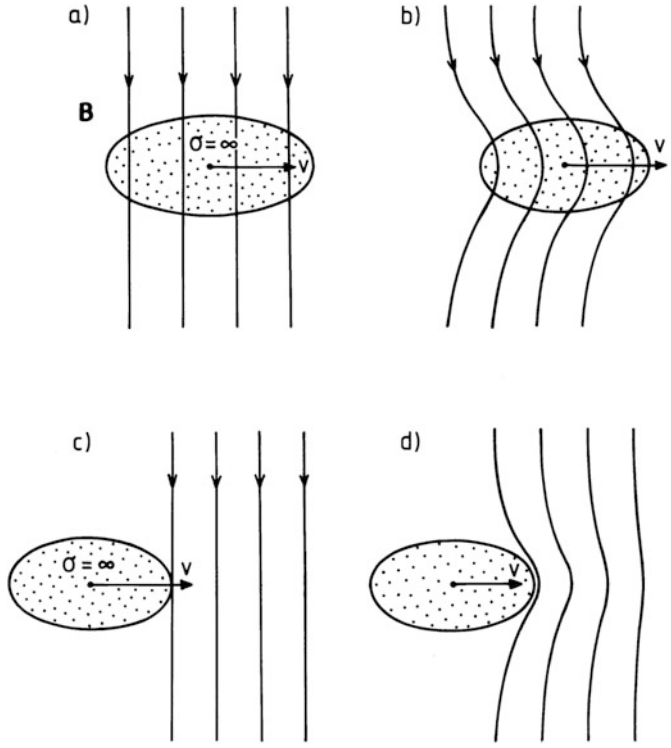


Figure 1.35. An illustration of the frozen-in field concept. (a) A magnetic field **B** is assumed to be penetrating a region of highly conducting plasma. (b) When the plasma starts to move, the magnetic field lines will be frozen-in and follow the motion of the plasma. (c) A highly conducting plasma is approaching an area of the magnetic field. (d) Due to the high conductivity the field cannot penetrate the plasma and is pushed ahead of the plasma blob.

ing plasma is approaching a volume with a magnetic field, the field cannot penetrate the plasma, and the plasma pushes the field ahead. This is illustrated in Figure 1.35.

1.13 THE GARDEN HOSE EFFECT

We have now seen that the magnetic field can be carried along in the solar wind from the Sun and through interplanetary space in a frozen-in manner. As the plasma is streaming radially out from the rotating Sun, it will move in spirals like water from a spinning nozzle of a garden hose.

Let us assume that the plasma leaves the solar surface in the equatorial plane at a distance r_0 from the solar center. Then, after a time, t, the position of the

plasma in the same plane can be described as

$$r = v \cdot t + r_0$$

$$\phi = \Omega \cdot t + \phi_0$$

where ϕ_0 is the longitude on the Sun from which the plasma emerges, Ω is the angular velocity of the Sun, and v is the solar wind velocity.

Eliminating t gives:

$$r = v \cdot \frac{\phi - \phi_0}{\Omega} + r_0$$

which is the equation for an Archimedean spiral (illustrated in Figure 1.36).

As the plasma will now drag the magnetic field along, it will follow the same trajectory and form similar spirals. Introducing polar coordinates and still confining ourselves to the equatorial plane, we have:

$$\mathbf{v} = (v_r, v_\phi)$$

$$\mathbf{B} = (B_r, B_\phi)$$

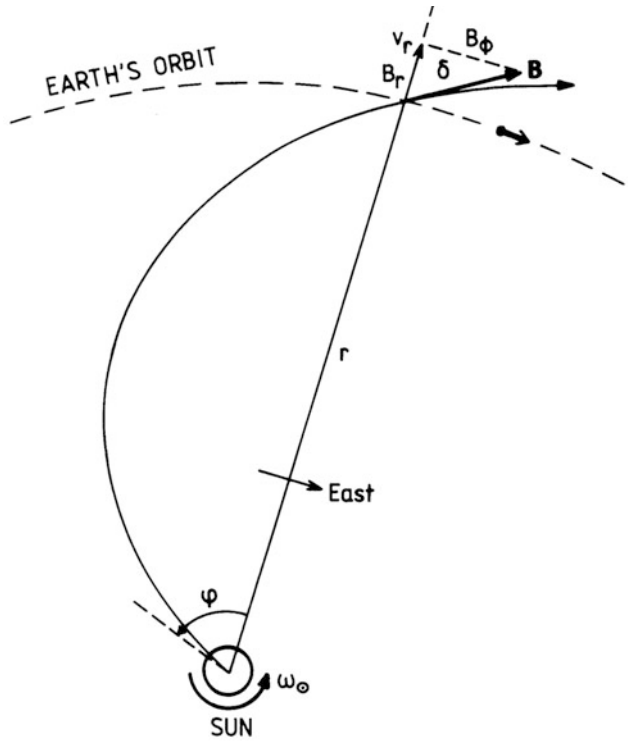


Figure 1.36. Solar wind plasma streams radially out from a rotating Sun, and its motion can be described as an Archimedean spiral (garden hose). At the position of the Earth the angle (δ) between the plasma velocity and the Sun–Earth line is close to 45°. The Earth's orbit is indicated.

We will assume that

$$B = B(r)$$

and

$$v = v(r)$$

(i.e., the magnitude of the solar wind velocity and the magnetic field depend only on the radial distance from the Sun). We then have in spherical coordinates:

$$\nabla \cdot \mathbf{B} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 B_r) = 0$$

or

$$r^2 B_r = r_0^2 B_0 = \text{const.}$$

The magnetic flux through spherical shells is conserved. The radial component of the magnetic field therefore decreases as

$$B_r = B_0\left(\frac{r_0}{r}\right)^2 \tag{1.34}$$

where B_0 is the magnetic field in the reference point at the solar surface.

Since $\partial B/\partial t = 0$ and the frozen-in concept apply we have from (1.31):

$$\nabla \times (v \times \mathbf{B}) = 0$$

and in cylindrical coordinates

$$\frac{1}{r}\frac{\partial}{\partial r}(r(v_\phi B_r - v_r B_\phi)) = 0$$

$$r(v_\phi B_r - v_r B_\phi)) = \text{const}$$

If we assume that $\mathbf{B}$ is radial at the reference point, then $B_{\phi_0} = 0$ and $B_{r_0} = B_0$ there:

$$r_0 v_{\phi_0} B_0 = r v_\phi B_r - r v_r B_\phi \tag{1.35}$$

Since the gas is rotating with the rotation speed of the Sun, we find at the surface of the Sun:

$$v_{\phi_0} = r_0 \Omega$$

By inserting this into (1.35) we have

$$r_0^2 \Omega B_0 = r v_\phi B_r - r v_r B_\phi$$

and solving for B_ϕ:

$$B_\phi = \frac{r v_\phi B_r - r_0^2 \Omega B_0}{r v_r} = \frac{v_\phi B_r - r\Omega\left(\frac{r_0}{r}\right)^2 B_0}{v_r} = \frac{v_\phi - r\Omega}{v_r} B_r$$

For very large r, $r\Omega > v_\phi$; then:

$$B_\phi = -\frac{r\Omega}{v_r} B_r = -\frac{r_0^2 \Omega}{r v_r} B_0$$

The azimuthal component therefore decreases with distance from the Sun as $1/r$ (i.e., more slowly than the radial component as expressed by (1.34)).

The angle the magnetic field will form with the radius vector to the Sun is given by (see Figure 1.35):

$$\tan\delta = \left|\frac{B_\phi}{B_r}\right| = \frac{r\Omega}{v}$$

for large distances ($r\Omega > v_\phi$).

At the Earth's orbit ($r = 1\,\mathrm{AU}$) and $\Omega = 2\pi/T$ and $T = 24.7$ days for the rotation time at the solar equator, we find when assuming a solar wind speed equal to 400 km/s:

$$\tan\delta \approx 1$$

At the Earth's orbit the "garden hose" angle is therefore close to 45°, in good agreement with observations. From observations of the motion of charged particles it has been inferred that those leaving the Sun at the western heliospheric hemisphere are more often observed reaching the Earth than particles leaving the eastern hemisphere as if the particles are guided toward the Earth along Archimedean spirals.

As the solar wind plasma emerges from the Sun irrespective of whether the solar magnetic field points toward or away from the Sun, it will carry a magnetic field with different polarity into interplanetary space. Therefore, we find that the interplanetary field is divided up into sectors where the field points predominantly away from or toward the Sun. This is illustrated in Figure 1.37 which shows a Cartesian coordinate system often used to describe the interplanetary magnetic field (IMF). The z-axis is vertical to the ecliptic plane in the northward direction (i.e., parallel to the Earth's rotation axis). x is directed towards the Sun parallel to the midday–midnight meridian plane, and y is directed eastward in order to complete the right-handed orthogonal system.

Because of the spiral structure of the interplanetary field we notice that when the field is directed away from the Sun, the y-component is positive, and for a toward-directed IMF the B_y component is negative.

We have in our discussion neglected any component of the magnetic field perpendicular to the equatorial plane of the Sun. This is of course an oversimplification. The component vertical to the equatorial plane, or rather to the ecliptic plane, has a profound influence on the shape of the magnetosphere of the Earth.

Furthermore, the magnetic field lines emanate not only from the ecliptic plane but at any latitude of the Sun. Therefore, the sector structure has a rather complicated shape in three dimensions, and the border sheets between the different structure regions will wobble around the Sun like the skirt of a ballerina. The Earth will move in and out of the pleats in this skirt (Figure 1.38). During one solar revolution the Earth will usually encounter two sector reversals, but quite often four and sometimes more. During the crossing of such sector boundaries disturbances in the Earth's atmosphere are likely to change character. Remember, however, that solar wind effects on Earth will usually be delayed by about 4 days from a sector crossing at the central meridian of the Sun.

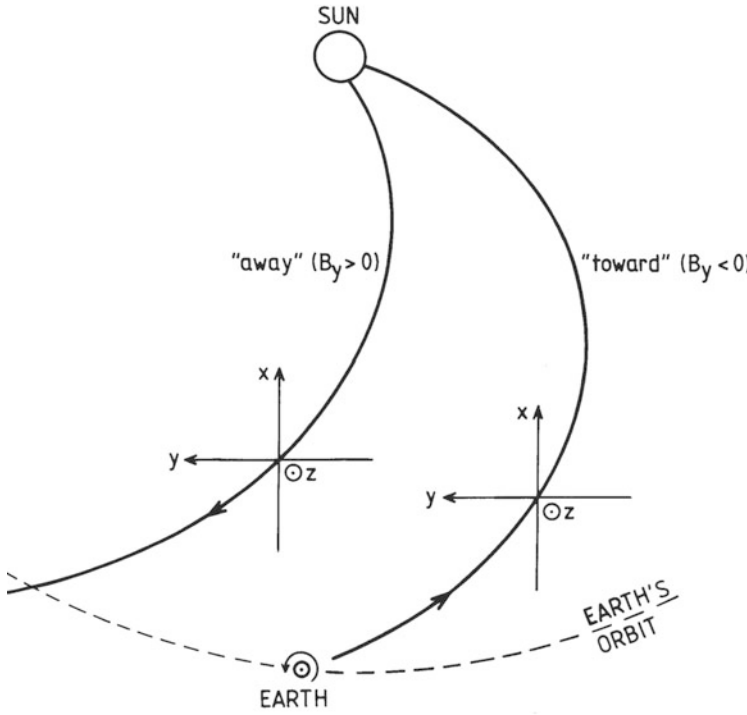


Figure 1.37. A sketch to illustrate the relationship between "toward" and "away" sectors of the interplanetary field and the positive and negative values of the *y*-component. The *z*-axis is perpendicular to the ecliptic plane. The *x*-axis is in the noon–midnight meridional plane pointing positive toward the Sun. The *y*-axis is pointing positive toward the east.

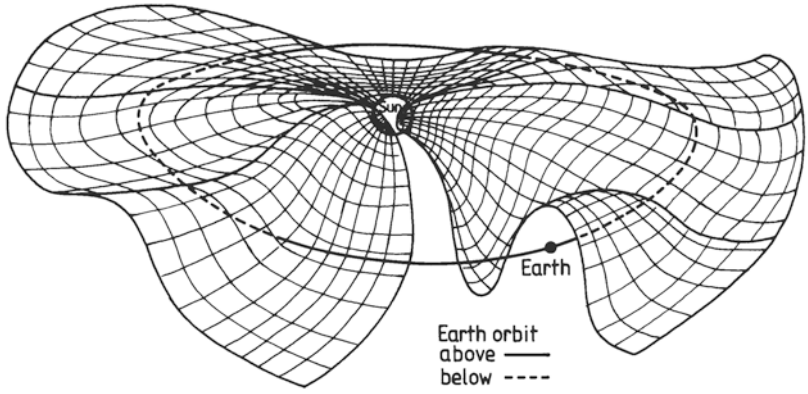


Figure 1.38. A three-dimensional sketch of the solar equatorial current sheet associated with the magnetic field line configuration in interplanetary space. The current sheet is here portrayed as lying near the solar equator and containing folds or flutes. When the Sun rotates, an observer near the ecliptic, such as on the Earth, will alternately lie above and below this current sheet and will see a changing sector pattern. (After Kelley, 1989.)

1.14 EXERCISES

1. Show that

$$q = \int_0^{2\pi} \cos\theta \, d\Omega \int_0^{\infty} B_\lambda \, d\lambda = \sigma T^4$$

where B_λ is given by (1.2) and determine σ. (Hint: $\int_0^{\infty} [x^3/(\exp(x) - 1)] \, dx = \pi^4/15$.)

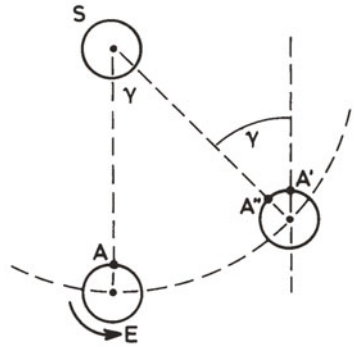


Figure 1.39.

2. Figure 1.39 illustrates the portion of the Earth's (E) orbit around the Sun (S) covered during one "day". When the point A has reached A' one *sidereal day* has passed, while when A has reached A'' one *solar day* has been completed. Show that the difference between one solar day and one sidereal day is approximately equal to 240 s.

3. Assume that the Earth's atmosphere in Figure 1.9 has a reflection coefficient α_{ar} and an absorption coefficient α_{aa}.

   (a) Assume radiative equilibrium and express the energy balance equations for the atmosphere and the Earth.

   (b) Find an expression for the radiative temperature of the Earth.

4. Assume that the electron density in the solar atmosphere and the corona can be expressed as a function of distance from the solar center by:

$$n_e(r) = n_\odot \left(\frac{R_\odot}{r} \right)^\beta$$

where $n_\odot$ is the electron density at the reference distance 1 $R_\odot$ from the solar center.

We have observed a Type-III radio burst associated with a flare and measured a dispersion of the frequency (df/dt). Let the burst be due to a plasma wave whose source is moving upwards in the atmosphere.

(a) Find an expression for the velocity of this source expressed by df/dt.

Assume that the electron density at $1\,R_\odot$ is equal to $10^{16}\,\mathrm{m}^{-3}$ and at $10\,R_\odot$ it is equal to $10^{9}\,\mathrm{m}^{-3}$. Let the burst first appear at 90 MHz and assume a frequency dispersion of 8 MHz/s.

(b) At what height did the burst start?

(c) What was the speed of the source at the onset of the burst?

5. Assume that the mass loss per unit time from the Sun due to blackbody radiation would be valid forever. How long would it take for the total mass of the Sun to be converted to radiative energy?

6. Show that for a plasma with a very high conductivity, σ, and moving with a velocity $v \ll c$ the magnetic flux through an area following the plasma can be expressed as:

$$\frac{d\phi}{dt} = -\iint_S \left(\nabla \times \frac{\mathbf{j}}{\sigma}\right) \cdot d\mathbf{s}$$

where S is the surface and $\mathbf{j}$ is the current density.

7. Assume that the magnetic flux at the Sun is $B_\odot = 10^{-4}$ tesla at the solar surface ($r = 1\,R_\odot$). What will be the strength of this field at the trajectory of the Earth? (Assume that the solar wind has a radial velocity of 400 km/s and that the rotation speed of the Sun is 24.7 days.)

8. (a) Find a solution for the equation

$$-AkT^{5/2}\frac{dT}{dr} = \text{const.}$$

where $A = 4\pi r^2$. This satisfies the conditions $T = T_1$ when $r = r_1$ and $T = 0$ when $r \to \infty$.

(b) Assume that $T = 10^6$ K when $r = 1\,R_\odot$. Determine the temperature in the solar wind at 1 AU.

Assume that the energy conservation equation for the solar wind plasma can be expressed as

$$A\rho v\left[\frac{v^2}{2} + \frac{5}{2}\frac{p}{\rho} - \frac{GM_\odot}{r}\right] + A \cdot \phi = \mathcal{E}$$

where $\mathcal{E} =$ const. Far out in the solar wind the heat conduction flux ϕ can be expressed by

$$\phi = \alpha nkTv$$

where α is a constant and n is the particle number density.

(c) Show that by introducing (1.18) and (1.25) the temperature can be expressed as:

$$T = T_2 \left(\frac{n}{n_2} \right)^{2/(\alpha+3)}$$

when $T = T_2$ and $n = n_2$ at $r = r_2$. Remember that $p = 2nkT$.

2

The atmosphere of the Earth

2.1 NOMENCLATURE

The atmosphere of the Earth is an ocean of gas encircling the globe. It stretches out far from the surface; how far out is a question of definition.

In Figure 2.1 we have illustrated schematically the variation of some characteristic parameters up to an altitude of 450 km. There are rather strong variations in the upper atmosphere above 100 km during a solar cycle, and this is illustrated in the figure by curves representing average solar maximum and minimum conditions. The cross-hatched areas between these curves illustrate the variability of the parameters. The number density (n) of the atmosphere decreases monotonically with height from $10^{25}\,\mathrm{m}^{-3}$ at ground level to $10^{14}\,\mathrm{m}^{-3}$ at 400 km. The atomic mass number is constant and close to 30 at the ground and up to about 100 km; above that region it decreases gradually toward 15 at 400 km. This corresponds to an air density of $1.2\,\mathrm{kg/m}^3$ at ground level decreasing to about $3.4 \times 10^{-6}\,\mathrm{kg/m}^3$ at 90 km, $2.4 \times 10^{-8}\,\mathrm{kg/m}^3$ at 120 km, and $2.8 \times 10^{-10}\,\mathrm{kg/m}^3$ at 200 km, respectively. According to variations in its composition the atmosphere is divided into two main regions, the *homosphere* and the *heterosphere*, indicating that below 100 km the gas constituents are fully mixed into a homogeneous gas. Above this height, however, the different constituents behave independently, and the atmosphere is heterogeneous.

The temperature (T) of the atmosphere has a more complicated behavior with height. It starts out by decreasing in the *troposphere* from about 290 K at the ground and reaching a minimum close at 215 K at 1,520 km, called the *tropopause*.

Above the tropopause is the *stratosphere*, and here the temperature increases up to a maximum of close to 280 K, called the *stratopause*, usually situated close to 50 km. Above the stratopause the temperature decreases again in the *mesosphere* and reaches the lowest temperature in the atmosphere in the *mesopause*, usually situated at about 70–90 km. The temperature in the mesopause may be as

A. Brekke, *Physics of the Upper Polar Atmosphere*, Springer Praxis Books,
DOI 10.1007/978-3-642-27401-5_2, © Springer-Verlag Berlin Heidelberg 2013

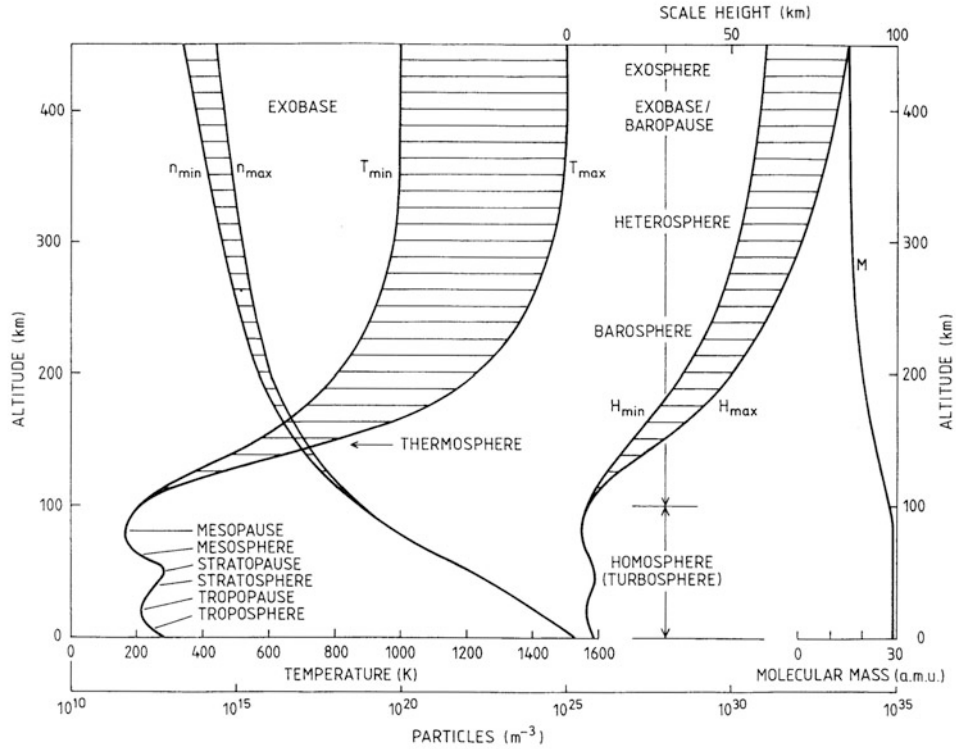


Figure 2.1. Model height profiles of the temperature T, density n, molecular mass M, and scale height H distributions in the Earth's atmosphere below 450 km. The different regions are also indicated by their characteristic names according to temperature or composition. Variability in the different parameters with respect to solar activity is indicated by the hatched areas.

low as 160 K or even lower at occasions. As a curiosity, however, the mesopause temperature in polar regions is higher in winter than in summer.

Above the mesopause the temperature increases dramatically in the thermosphere where temperatures of more than 1,000 K can be found above the *exobase* indicated in Figure 2.1 at 400 km. Above this region the temperature is fairly constant by height, but may vary considerably by time.

The nomenclature used for characterizing different areas in the atmosphere refers to either temperature, composition, or dynamics. At the ground we find that the atmosphere is composed of close to 80% N_2 and 20% O_2, while the contribution from other gases is less than 1%. This mixture holds all the way up to about 100 km. The region below 100 km is therefore called the *homosphere* or the *turbosphere*. The latter reflects the fact that turbulence causes the mixture. Above 100 km, molecules start to dissociate and become more independent of each other; they are more heterogeneous. This region is therefore called the *heterosphere* or the *barosphere*; the latter reflecting the fact that the different species have different scale heights or barometric heights.

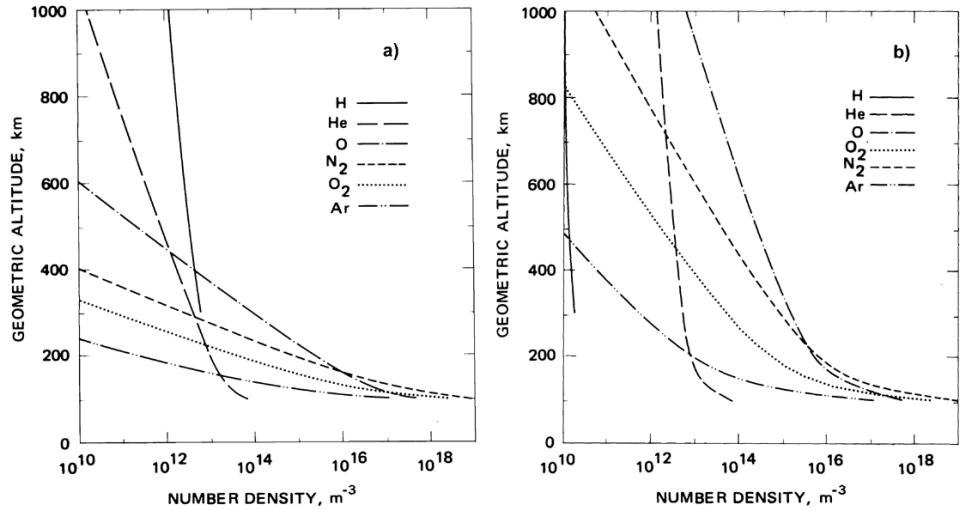


Figure 2.2. Composition changes in the atmosphere with respect to (a) solar minimum and (b) solar maximum conditions showing the dominance of heavier constituents at larger altitudes during solar maximum. (From U.S. Standard Atmosphere, 1976.)

In Figure 2.2 we show a more detailed presentation of the altitude profiles of different atmospheric species between 1,000 km and the ground under solar minimum and maximum conditions. While atomic oxygen dominates the upper atmosphere above 250 km under solar maximum conditions, hydrogen is more dominant above 400 km under solar minimum conditions. Below 200 km, however, molecular oxygen and nitrogen together with argon are the dominant species independent of solar activity.

Due to the composition at ground level (80% N_2 and 20% O_2) the average molecular mass will be 28.8 a.m.u. Since molecules start to dissociate above about 100 km and nitrogen molecules dissociate faster than oxygen molecules, the molecular weight will decrease and oxygen atoms will be the dominant species from 400 km to above 1,000 km, at least during solar cycle maximum (see Figure 2.2 for references).

At 600 km where we might have 84% O and 16% He, the molecular mass is 14 a.m.u. While the number density at the ground is about $2.5 \times 10^{25}\ m^{-3}$, it is about $10^{19}\ m^{-3}$ at 100 km, and at 200 km it has been reduced to about $10^{16}\ m^{-3}$. The density therefore has decreased by more than 10^9 from the ground up to typical rocket trajectories.

2.2 TEMPERATURE STRUCTURE OF THE ATMOSPHERE

The temperature in the atmosphere decreases quite monotonically up to the tropopause. This is due to the fact that infrared radiation from the ground, which

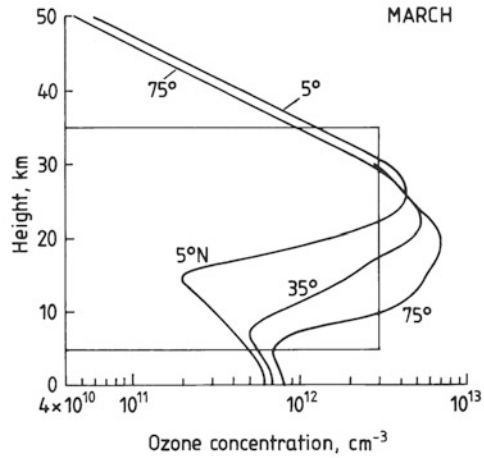


Figure 2.3. Ozone density profiles in the atmosphere at different latitudes in the northern hemisphere. The concentration is given in molecules per cm^3. (After Shimazaki, 1987.)

is absorbed in the atmosphere, is rather constant because the surface temperature of the globe is constant, and therefore the heat expands out in the atmosphere in radial directions. The heat will then be distributed into larger and larger volumes and therefore the temperature must decrease.

Due to the ozone layer situated between 5 and 40 km above the ground depending on latitude (Figure 2.3) which absorbs a large portion of the solar radiation between 200 and 300 nm, the atmosphere becomes heated in the stratosphere and the temperature increases. Above the stratopause, however, the heat balance results in excess outward radiation again, and the temperature decreases rapidly in the mesosphere until the sharp minimum occurs at the mesopause.

By observing the ozone density profile at different latitudes, however, the peak altitude is found to decrease and the peak magnitude to increase with increasing latitude. More ozone is therefore present at lower altitudes (<25 km) in the polar region (Figure 2.3).

This is even more clearly brought out in Figure 2.4 where the latitudinal average of the ozone content at different altitudes below 50 km is presented versus the latitude for the average February month. The winter pole has rather high ozone content at low altitudes (<20 km) while the ozone density above the equator is much smaller and maximizes above 20 km. This latter effect cannot be explained by the simple photochemical equilibrium models of the atmosphere since from these we would expect the ozone content to be high at lower heights in the equatorial region where the solar radiation and the radiative dissociation of O_2 is highest. Transport processes must therefore be of fundamental importance for global ozone distribution.

In Figure 2.5 the latitudinal distribution of total ozone content for different seasons of the year is presented. Again there are clear maxima above the polar regions, but it is seen that these maxima occur in spring. The maximum in the

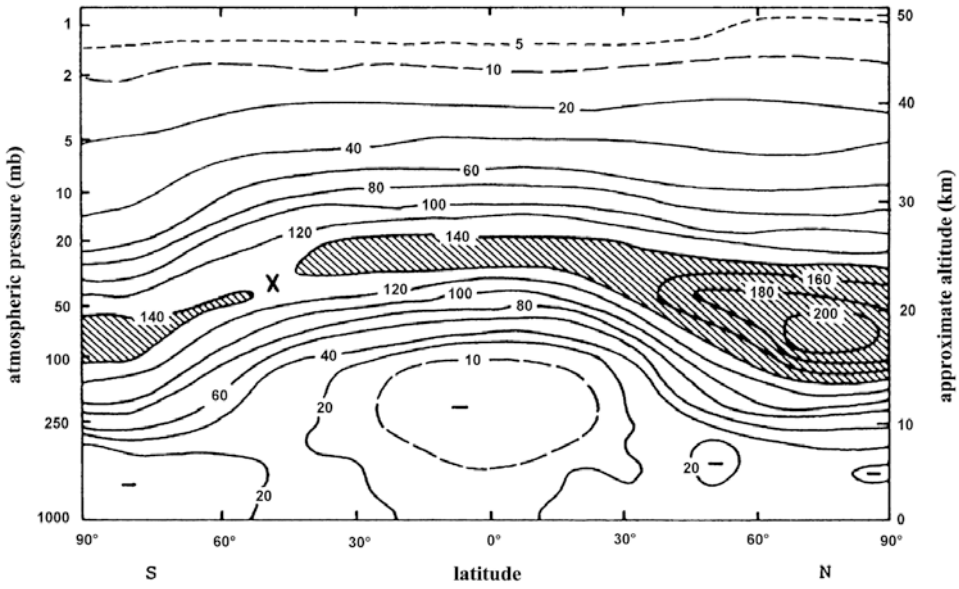


Figure 2.4. Latitudinally averaged ozone distribution for February shown as a function of height and latitude. A strong maximum is observed below 20 km in the Arctic region. (From Dütsch, 1978.)

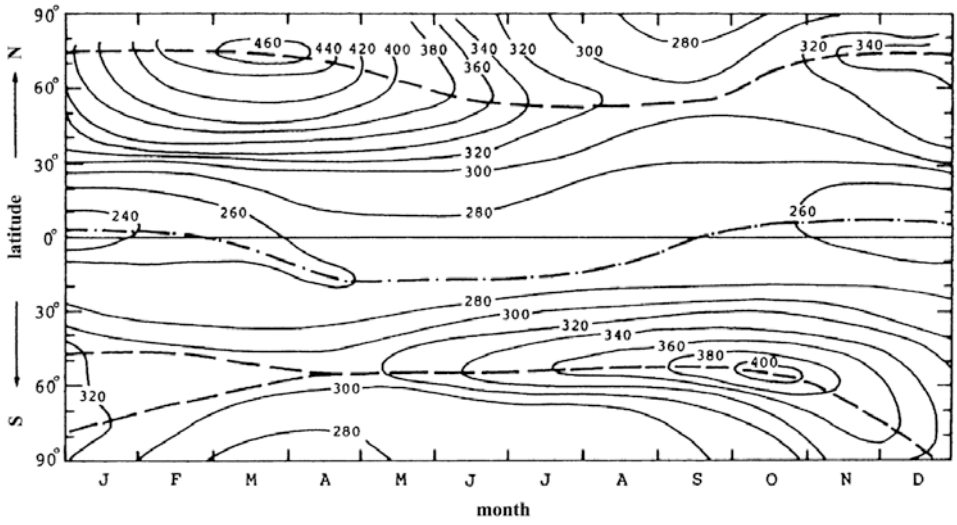


Figure 2.5. Latitudinal and seasonal variation of the total ozone content. The labels added to the isolines are given in Dobson units. (From London, 1985.)

northern hemisphere in March is about 15% higher than the corresponding maximum in the southern hemisphere in October.

In the northern hemisphere the amplitudes of seasonal variation are also larger, varying from a maximum of close to 460 Dobson units in April at times to about 280 Dobson units in November. That is a variation of more than 20% around the mean value.

The total ozone content in a vertical column above a ground-based observer has been measured for more than 50 years at different places on Earth.

The term "total ozone", sometimes referred to as total layer thickness, thus represents the height integral of the column density and is often abbreviated as atm-cm or cm (STP). This corresponds to the thickness of the layer if the pressure and density were reduced to standard atmospheric values throughout the layer. The Dobson unit is derived in a similar manner and abbreviated as m-atm-cm which equals 10^{-3} atm-cm. Therefore, 300 Dobson units represent an ozone content which, when reduced to standard temperature and pressure throughout the layer, would correspond to a column of 3 mm.

One of the early findings in the research of ozone was the smaller amount of ozone in the Antarctic spring than in the Arctic spring. Another early result of such measurements was the occasional decrease observed in total ozone during the early springtime.

As is well known, the ozone layer shields the Earth from UV radiation which can be a health hazard to some people. The shielding, however, is most effective in the polar region where the ozone layer is at its thickest and the Sun has a large angle to the zenith, while the layer is more shallow at lower latitudes where the Sun is close to overhead. This is illustrated in Figure 2.6.

For a station at 70° latitude, for example, the Sun can never make an angle α with the zenith less than 46.5°. Therefore, the distance that the solar UV ray must pass through the ozone layer will be (Figure 2.7)

$$l = \frac{h}{\cos \alpha}$$

if h is the thickness of the layer. In the tropics where the Sun can be at zenith and $\alpha = 0$, the ray path through the layer is equal to the layer thickness. In order for the UV intensity observed at 70° latitude to be equal to the intensity observed at the equator the layer thickness at 70° is to a first approximation given by:

$$h' = h_e \cos \alpha$$

Since the present thickness of the ozone layer at high latitudes is about $1.5h_e$ we notice that the layer thickness must be reduced to about half its present value:

$$\frac{h'}{h} = \frac{h_e}{1.5h_e} \cos 46.5 \approx 0.5$$

Above the mesopause solar radiation in the UV band is strongly absorbed owing to dissociation of molecules such as O_2, N_2, and NO and ionization of

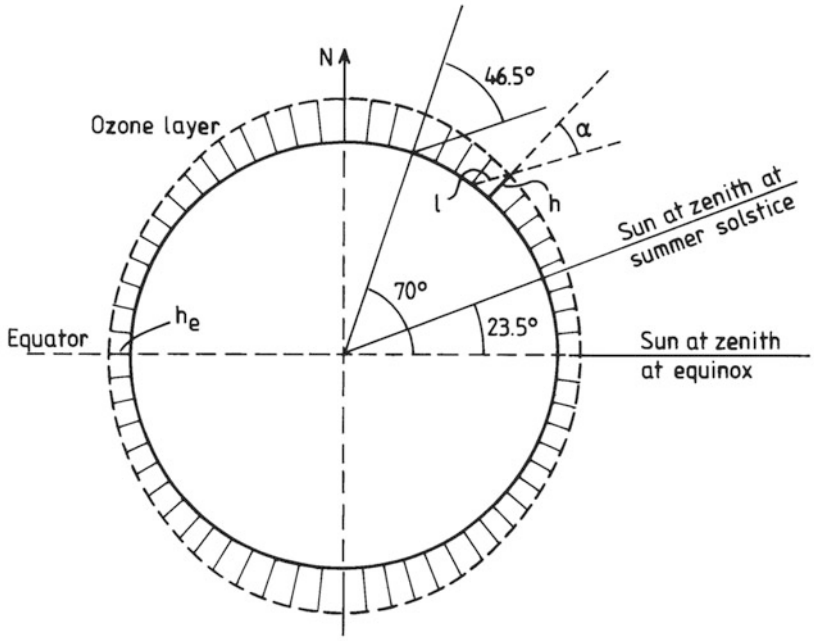


Figure 2.6. A schematic illustration of the global ozone distribution showing a thinning of the layer at lower latitudes where also solar irradiation can have a vertical impact on the atmosphere. The height of the maximum in the layer, however, is larger at equatorial regions than at the poles.

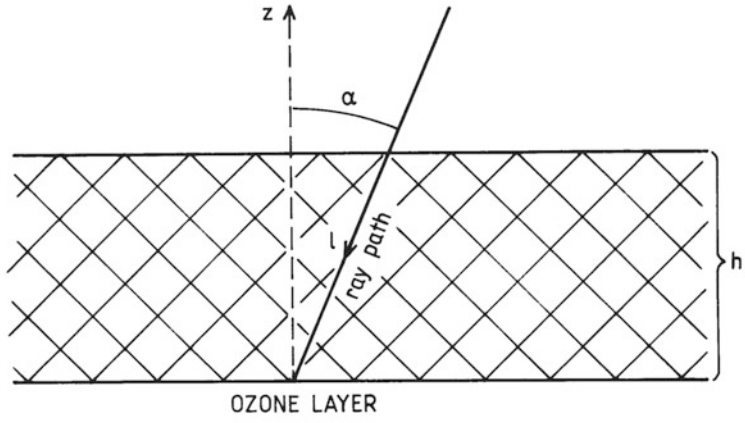


Figure 2.7. Illustration of variation in the length of the ray path through the ozone layer with a varying solar zenith angle.

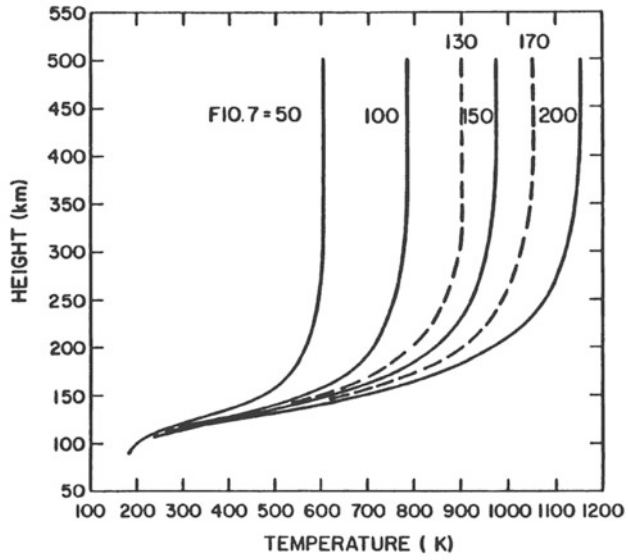


Figure 2.8. Variability in the thermospheric temperature for different values of the solar radio flux index ($F_{10.7}$) in units of 10^{-22} W m^{-2} Hz^{-1} reduced to 1 AU. Under average solar minimum and maximum conditions the $F_{10.7}$ index is 50 and 200, respectively. (From Roble, 1987.)

atomic oxygen and other molecules and atoms. This leads to a new temperature increase which is particularly strong up to about 400 km.

The very strong absorption of UV radiation in the thermosphere, however, is accompanied by strong variability in the temperature of this region as illustrated in Figure 2.1. This is because solar UV radiation itself is so variable. As the temperature above 400–600 km appears fairly constant with height as a function of altitude, it is often referred to as the exospheric temperature (T_∞). Figure 2.8, which is an alternative presentation of Figure 2.1 above 100 km, shows the range of variability in thermospheric temperature at different altitudes in the course of a solar cycle due to variations in the exospheric temperature. The intensity index of solar radio emission at $F_{10.7}$ is used as a reference parameter to the solar cycle. The value 50 represents solar minimum conditions while the value 200 represents solar maximum conditions.

The exospheric temperature can change by 600 K or more during a solar cycle according to similar model calculations (as presented in Figure 2.8).

The thermospheric temperature, however, is not constant all over the globe and exhibits a seasonal variation. The thermospheric temperature at 300 km as a function of latitude for solstice and equinox conditions is presented in Figure 2.9. Also shown in Figure 2.9 is the mean molecular mass (a.m.u.) for the corresponding conditions. At solstice especially, latitudinal variations are as large as the temperature changes from close to 1,400 K at the summer pole to slightly above 900 K at the winter pole, and the molecular mass changes from 21 a.m.u. to 17 a.m.u. in the same region.

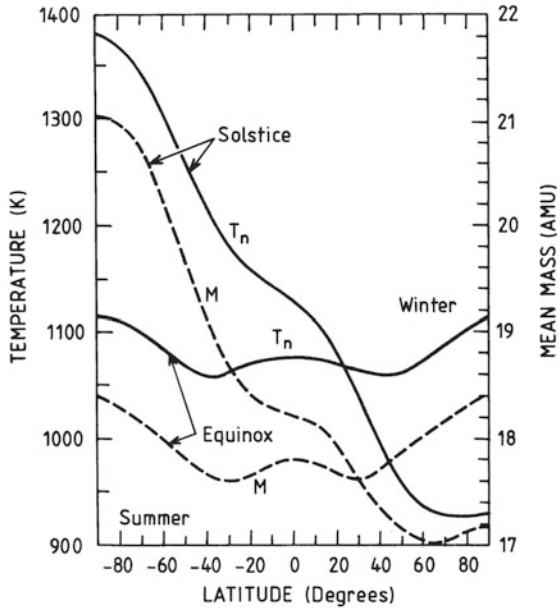


Figure 2.9. Latitudinal distribution of the neutral temperature T_n and mean molecular mass M at 300 km under equinox and solstice conditions. (After Roble, 1987.)

The consequence of this is that heavier molecules are brought up to higher altitudes from below in the summer hemisphere and downward in the winter hemisphere (Figure 2.10). The summer thermosphere at 300 km and above is therefore dominated by N_2 molecules, while the winter thermosphere in the same height region has a large contribution of helium atoms.

The thermospheric temperature also responds to solar variations on a shorter time scale than solar cycles or seasonal periods. Figure 2.11 presents a comparison of variations in exospheric temperature T_∞, atmospheric mass density ρ, solar radio emission flux S (= $F_{10.7}$), and a geomagnetic index Ap. The latter represents variations in the Earth's magnetic field. presumably due to ionospheric currents.

The 27-day period in solar radio emission, S ($F_{10.7}$), is clearly reflected in T_∞ as well as ρ, as if the atmosphere is expanding and contracting as the solar flux increases and decreases.

2.3 ATMOSPHERIC DRAG ON SATELLITES

Figure 2.12 gives a schematic presentation of the variability of atmospheric density between 100 and 1,000 km under solar maximum and minimum conditions as well as under extreme solar maximum daytime conditions. Density variations of two orders of magnitude can in fact take place during a solar cycle above 400 km altitude.

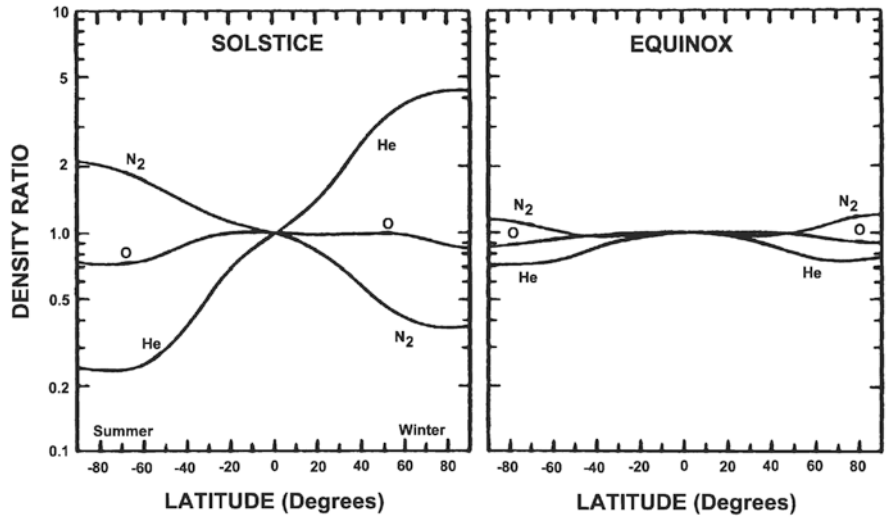


Figure 2.10. The latitudinal distribution of molecular nitrogen (N_2), atomic oxygen (O_2), and helium (He) under solstice and equinox conditions. (From Roble, 1987.)

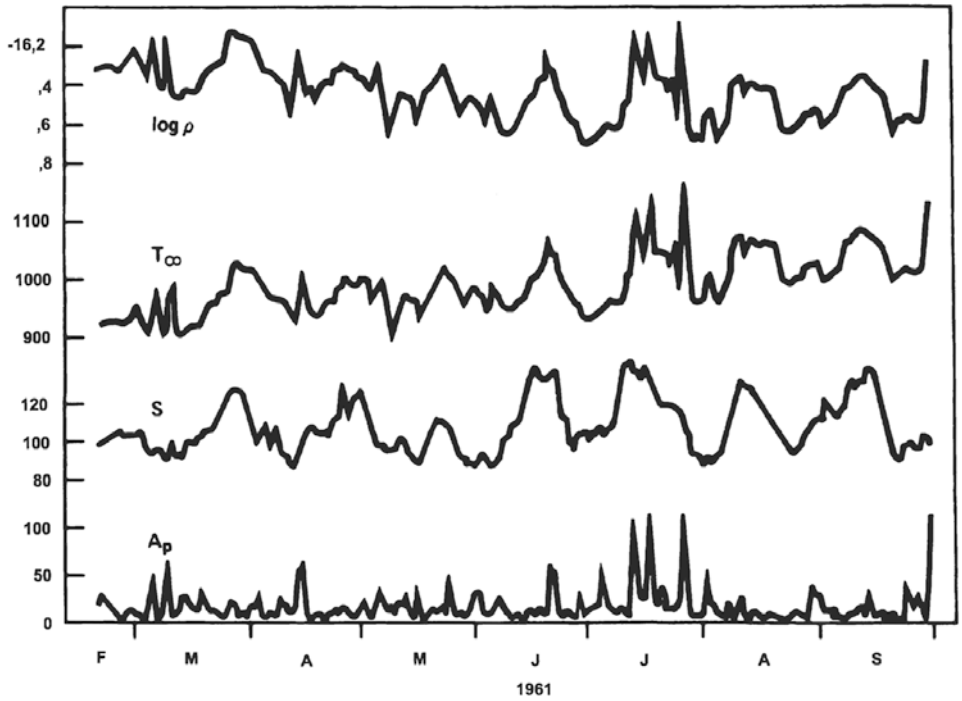


Figure 2.11. A comparison between thermospheric mass density ρ, exospheric temperature T_∞, the solar 10.7 cm radio flux index S ($F_{10.7}$), and the Ap index as a function of time for 1961. (From Giraud and Petit, 1978.)

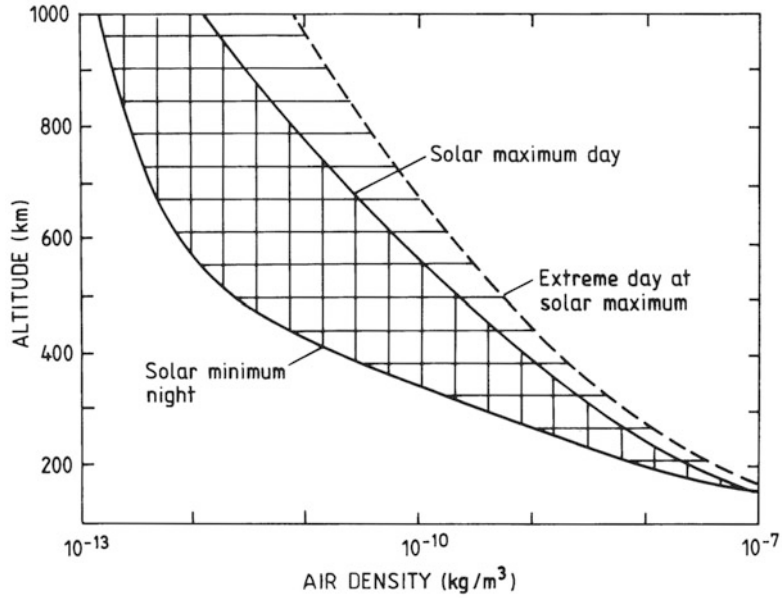


Figure 2.12. Illustration showing the extreme variability in the neutral density of the thermosphere under solar minimum and solar maximum conditions.

Satellites at these altitudes will experience great differences in the friction force due to atmospheric drag in response to these density variations, and during severe disturbances the satellite trajectories can be altered significantly. It is therefore of great importance to know the behavior of the upper atmosphere when planning space missions, because an unexpected atmospheric disturbance can reduce the lifetime of a satellite by several months, if not destroy the mission completely.

We also notice from Figure 2.11 that the geomagnetic activity due to enhanced ionospheric currents can give rise to abrupt changes in exospheric temperature as well as density; again, this is one of the main reasons for the interest in geomagnetic disturbances as it has practical importance for satellite trajectories. These disturbances are also the most difficult to predict because geomagnetic disturbances occur in a very erratic manner especially at high latitudes.

Let us therefore consider the effect of variations in atmospheric density on a satellite orbit. For a sphere with mass m moving with velocity v with respect to atmospheric gas, there will be a drag force F_D acting on the sphere due to collisions with atmospheric particles, which can be expressed in the following way:

$$F_D = \tfrac{1}{2}\rho v^2 C_D$$

where ρ is atmospheric gas density, and C_D is what is called the ballistic coefficient. It is proportional to the cross-section of the sphere, and depends on surface conditions of the sphere's material.

For a satellite moving in a circular orbit with radius r around the Earth the total energy is:

$$E = \tfrac{1}{2}mv^2 - \frac{GM_e m}{r}$$

where G is the constant of gravity, and M_e the mass of the Earth.

If the atmospheric drag is small, the circular orbit will be maintained and then for circular orbits:

$$\frac{v^2}{r} = \frac{GM_e}{r^2} \tag{2.1}$$

The total energy of the satellite moving in a circular orbit in a central force field therefore is given by:

$$E = -\frac{GM_e m}{2r} \tag{2.2}$$

The rate of change of energy for the satellite due to atmospheric drag can be expressed as:

$$\frac{dE}{dt} = -F_D v = -\tfrac{1}{2}\rho v^3 C_D \tag{2.3}$$

By deriving dE/dt from (2.2) and equating it to (2.3) we have:

$$\frac{dE}{dt} = \frac{GM_e m}{2r^2}\frac{dr}{dt} = -\tfrac{1}{2}\rho v^3 C_D$$

and when applying (2.1) the rate of change of the radius of the orbit is

$$\frac{dr}{dt} = -\frac{\rho v C_D \cdot r}{m} \tag{2.4}$$

By observing the rate of change of the orbit's radius one could now derive the atmospheric density or, vice versa, when the atmospheric density is known the expected rate of change of the orbital radius could be obtained. Since variations in r from orbit to orbit are very small, it is not so practical to use (2.4) to study the effects on satellite orbits from atmospheric drag. It turns out that the orbital period ($T = 2\pi r/v$) is a better parameter for this.

According to Kepler's third law we have:

$$T^2 = \frac{4\pi^2 r^3}{GM_e}$$

By taking the time derivative

$$2T\frac{dT}{dt} = \frac{12\pi^2 r^2}{GM_e}\frac{dr}{dt} = \frac{12\pi^2 r}{v^2}\frac{dr}{dt}$$

and solving for dT/dt

$$\frac{dT}{dt} = \frac{6\pi^2 r}{Tv^2}\frac{dr}{dt}$$

we find by inserting (2.4) and solving for $dT/dt \equiv \dot{T}$:

$$\dot{T} \equiv \frac{dT}{dt} = -\frac{3\pi C_D r}{m}\rho$$

The period therefore decreases faster as the density increases. This can appear as an acceleration of the satellite, but since the energy is constant, the gain in kinetic energy must be lost by a reduction in potential energy due to the reduction in orbital radius by increasing neutral density.

Measuring the amount of change of the rotation period of the satellite is much easier than measuring the rate of change of the radius, and it can actually be used to derive the atmospheric density at the satellite altitude.

In Figure 2.13 the variation in the rotation periods of the satellites Explorer IV and Vanguard I is compared for a few months in 1958–1959. A close correlation in the variations is shown indicating that the effect is not local but global. Furthermore, it is demonstrated in the lower panel of Figure 2.13 that these variations in the case of Sputnik III are correlated with variations in the solar sunspot number. These observations are interpreted as due to variations in atmospheric density above altitudes of 300 km due to expansion and contraction caused by variations in solar heat input.

Figure 2.14 illustrates the rate of changes in seconds per day of the orbital period as a function of solar heat input for satellites at different altitudes. The 10.7 cm radio emission flux ($F_{10.7}$) is used as a parameter for solar heat input. At lower altitudes the rate of change becomes more severe for higher solar fluxes.

2.4 THE ATMOSPHERE AS AN IDEAL GAS

Consider the forces on a small mass element dm of air (Figure 2.15) at a height z above the ground. Let the mass element take the form of a small cylinder with horizontal cross-section A and height dz. This air mass will be acted upon by gravity, and this force can be expressed as

$$df = -nmgA\,dz$$

where g is the acceleration of gravity, and n is the number density of the molecules with mass m. In static equilibrium the gravity force must be balanced by the net pressure force in the following way:

$$[p - (p + dp)]A - nmgA\,dz = 0$$

where p is the pressure, which gives:

$$\frac{dp}{dz} = -nmg = -\rho g \tag{2.5}$$

Here $\rho = nm$ is the mass density. Equation (2.5) is called the barometric law. Assuming that the atmosphere is an ideal gas which is a very good assumption at

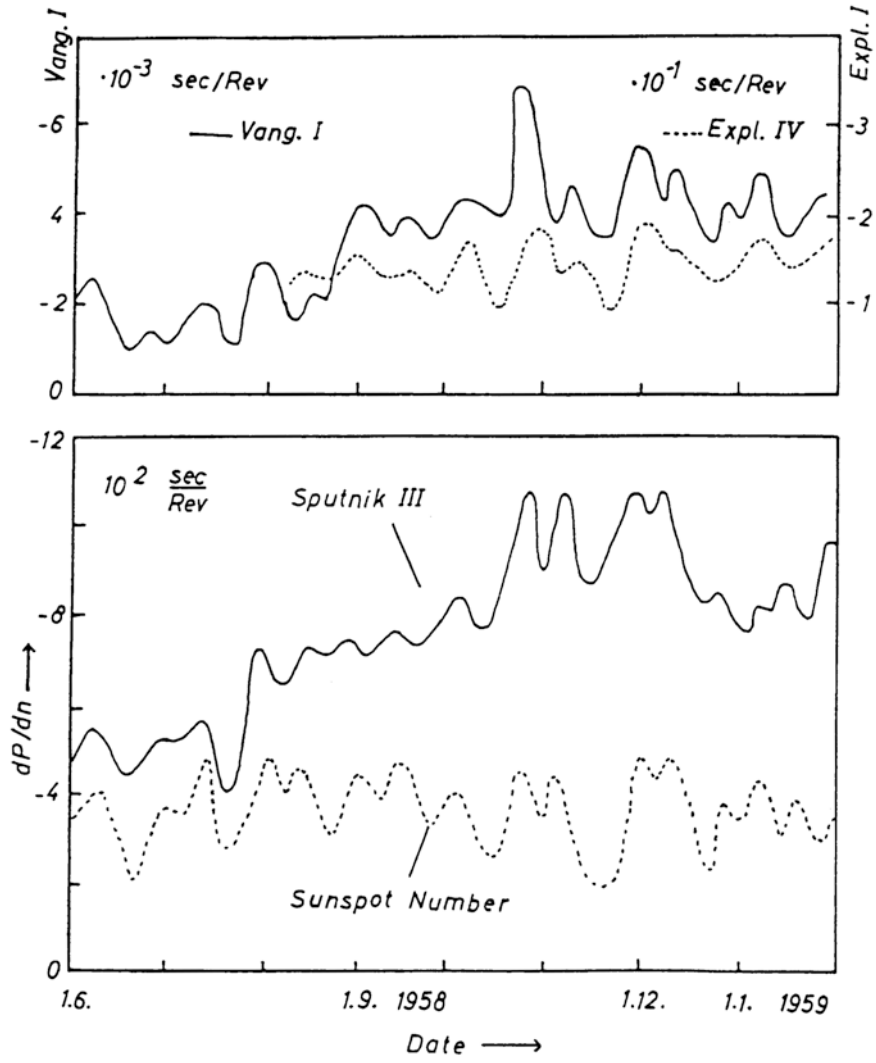


Figure 2.13. (Upper panel) The rate of change in orbital period per period for Explorer IV and Vanguard I satellites between June 1958 and February 1959. (Lower panel) A correlation between the variations in orbital periods for Sputnik III and sunspot numbers for the same period as above. (From Paetzold and Zschörner, 1960.)

least for the lower parts of the atmosphere, then we can apply the ideal gas law $p = nkT$ and derive from (2.5)

$$\frac{1}{p}\frac{dp}{dz} = -\frac{nmg}{nkT} = -\frac{mg}{kT} = -\frac{1}{H} \tag{2.6}$$

The parameter $H = kT/mg$ is called the scale height.

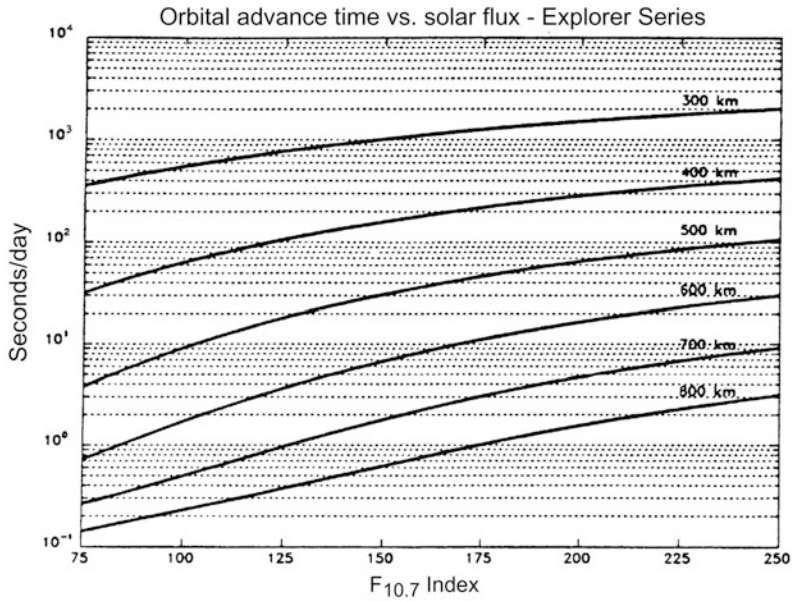


Figure 2.14. The rate of change in orbital period per day as a function of solar flux at 10.7 cm radio emission ($F_{10.7}$) for satellites at different heights between 300 and 800 km. (From Walterscheid, 1989.)

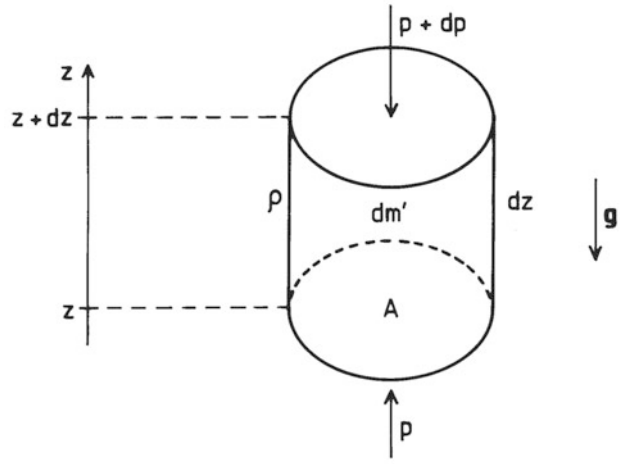


Figure 2.15. A volume element of air used to illustrate the balance between pressure and gravity forces.

We recall from kinetic gas theory that according to the equipartition principle each degree of freedom gives rise to a kinetic energy equal to $\frac{1}{2}kT$. We then notice that if all this kinetic energy in the vertical direction is converted to potential energy, the molecule must be lifted a height h given by:

$$\left.\begin{aligned} mgh &= \tfrac{1}{2}mv_z^2 = \tfrac{1}{2}kT \\ h &= \frac{1}{2}\frac{kT}{mg} = \tfrac{1}{2}H \end{aligned}\right\} \tag{2.7}$$

where v_z is the vertical velocity. This height then corresponds to half the scale height. Whether a particle will reach this height or not depends on its velocity and rate of collision. All in all, in thermal equilibrium there will always be a balance between the kinetic and potential energies of the particles so that the velocity distribution is kept constant at all heights. Only density will decrease in an isothermal atmosphere as will be shown below.

The pressure at any height can now be found by integrating (2.6):

$$\ln p = -\frac{z}{H} + \text{const.}$$

If $p = p_0$ at a reference height z_0, we have:

$$p = p_0 \exp\left(-\frac{z - z_0}{H}\right) \tag{2.8}$$

For a constant temperature and a constant molecular mass the scale height is constant as long as the acceleration of gravity is constant. Close to Earth, where $T = 288\,\text{K}$, $M' = 28.8$ is the mean molecular mass number, and $m = M' \cdot m_0$ is the mean mass of the molecules, where $m_0 = 1\,\text{a.m.u.} = 1.660 \times 10^{-27}\,\text{kg}$, $H = 8.47\,\text{km}$. Since the temperature between the ground and 100 km altitude vary quite markedly, the scale height will vary between 5.0 and 8.3 km (see Figure 2.1) in this region. Above 100 km, however, the temperature increases drastically, and the molecules dissociate so that the molecular mass decreases. The scale height increases and the pressure therefore does not decrease as rapidly above 100 km as below. At 300 km, for example, at a temperature of 980 K and with molecular mass number of 16 we find $H \approx 50\,\text{km}$.

Since $p = nkT$, we find an expression for the mass density by:

$$\rho = n \cdot m = \frac{m}{kT} \cdot p \tag{2.9}$$

We see that for a constant temperature and mean mass of the molecules, the mass density will decrease exponentially at the same rate as the pressure.

Let us consider for a while the number density when the atmosphere is isothermal. From the ideal gas law and the expression of the pressure (2.8) we have

$$p = nkT = p_0 \exp\left[-\frac{z - z_0}{H}\right] = n_0 k T_0 \exp\left[-\frac{z - z_0}{H}\right] \tag{2.10}$$

where n_0 and T_0 are the number density and temperature at the reference height

z_0, respectively. Since the atmosphere is assumed isothermal, we see that

$$n = n_0 \exp\left[-\frac{z - z_0}{H}\right] \tag{2.11}$$

When the reference height is set at the ground level $z_0 = 0$, the total sum of all particles from the ground to infinity above a unit area on Earth is given by:

$$\mathcal{N} = \int_0^\infty n\, dz = n_0 \int_0^\infty \exp\left(-\frac{z}{H}\right) dz = n_0 H \tag{2.12}$$

We therefore see that the scale height is equivalent to the height the atmosphere would have if the atmosphere encircling the Earth had a constant density by height. It would then reach only 8.4 km above our heads, and outside there would be a vacuum. This is the reason people in the old days believed that the atmosphere was 8–10 km thick. They knew that the air pressure balanced a water column of about 10 m, and by knowing the ratio between the density of water (10^3 kg/m^3) and the density of air at ground level (1.2 kg/m^3), the height of the atmospheric "lid" would be

$$h_a = \frac{10^3 \text{ kg/m}^3}{1.2 \text{ kg/m}^3} \cdot 10 \text{ m} = 8.3 \text{ km}$$

Realizing that the atmosphere was not finite really opened up the universe to human beings.

For an isothermal atmosphere with constant molecular mass the mass density is given by:

$$\rho = \rho_0 \exp(-z/H) \tag{2.13}$$

2.5 THE EXOSPHERE

The ordinary gas law can only be applied to atmospheric gas as long as the molecules make enough collisions to establish statistical equilibrium with their surroundings. Let us assume that a molecule traveling vertically at the most probable velocity (see Exercise 2) $v_{mp} = \sqrt{2kT/m}$ without experiencing any collisions, then we have according to (2.7):

$$mgh_e = \tfrac{1}{2} m v_{mp}^2 = kT$$

and the molecule will reach a height h_e given by:

$$h_e = \frac{kT}{mg} = H$$

where H is the scale height. If this distance h_e was less than the mean free path, l, which is the distance between two collisions, the molecule could on average be considered to move freely. The part of the atmosphere where the scale height is of the same order as, or less than, the mean free path is called the exosphere.

The mean free path is given approximately as

$$l = \frac{1}{\sigma \cdot n}$$

where σ is the cross-section for collisions, and n is the number density of the atmosphere. We then notice at the exobase, which is the bottom of the exosphere and where $H = l$, that:

$$H \cdot n = \frac{1}{\sigma}$$

From (2.12) we find that:

$$H \cdot n = \mathcal{N}$$

where $\mathcal{N}$ is the total number of particles per unit area above the height where the number density is n. We therefore have at the exobase that:

$$\mathcal{N} \cdot \sigma = 1$$

One such particle will therefore experience exactly one collision on its way up to the exobase. If $H < l$, the particle will not on average experience such a collision at all.

Most particles entering the exosphere from below travel in gravity-controlled orbits (ballistic motion) without making collisions until they either escape or return back to the atmosphere below. If a particle escapes from the Earth's gravitational field, its kinetic energy must be larger than its potential energy at the height of escape. Therefore

$$\tfrac{1}{2} m v^2 > m g_r r$$

where g_r is the acceleration of gravity at the distance of escape r, measured from the center of the Earth.

$$v > \sqrt{2 g_r \cdot r} = \sqrt{2 g_0 R_e^2 \frac{1}{r}} = v_{\text{esc}}$$

where g_0 is the acceleration of gravity at the Earth surface ($= 9.80\,\text{m/s}^2$). At the Earth's surface $v_{\text{esc}} = 11.2\,\text{km/s}$, while for larger distances v_{esc} becomes smaller. At about 2,000 km $v_{\text{esc}} \approx 9.7\,\text{km/s}$.

2.6 HEIGHT-DEPENDENT TEMPERATURE

We have shown in Figures 2.1 and 2.3 that the temperature in the lower atmosphere (below, say, 100 km) varies by height, and therefore it is strictly not legitimate to assume that T is constant. Let us therefore express T as a linearly varying function with height

$$T = T_0 + \alpha \cdot z$$

α is often called the linear "lapse rate". α can be positive as in the stratosphere and thermosphere and negative as in the troposphere and the mesosphere.

Introducing T in (2.6), but still assuming an ideal gas, we find:

$$\frac{dp}{p} = -\frac{mg}{kT} dz = -\frac{T_0}{H_0} \frac{dz}{T_0 + \alpha z}$$

where $H_0 = kT_0/mg$ is the scale height referring to $z = 0$. Solving this equation for p we find

$$\ln\frac{p}{p_0} = -\frac{T_0}{H_0}\int_0^z \frac{dz}{T_0 + \alpha z} = -\frac{T_0}{H_0\alpha}\ln\left(\frac{T}{T_0}\right) = -\frac{mg}{k\alpha}\ln\frac{T}{T_0}$$

Therefore

$$\frac{p}{p_0} = \left(\frac{T}{T_0}\right)^{-mg/k\alpha} \tag{2.14}$$

By inserting (2.14) into (2.9) we get:

$$\rho = \frac{m}{kT}\cdot p_0\left(\frac{T}{T_0}\right)^{-mg/k\alpha} = \rho_0\left(\frac{T}{T_0}\right)^{-1-(mg/k\alpha)}$$

Therefore

$$\frac{\rho}{\rho_0} \neq \frac{p}{p_0}$$

Pressure and density for a non-isothermal atmosphere will not have the same variation profile by altitude.

Another complication to be mentioned here is variation in the acceleration of gravity

$$g(z) = g_0\left(\frac{R_e}{R_e + z}\right)^2$$

where R_e is the Earth radius and g_0 is the acceleration of gravity at the Earth's surface. For $z < 100\,\text{km}$ this variation will contribute to a variation in H by only 3%. For $z = 1{,}000\,\text{km}$, however, the error will be significant when neglecting the variation in g ($\approx 25\%$).

2.7 THE ADIABATIC LAPSE RATE

Assume that when a volume element in the atmosphere moves in altitude, the motion will occur without any exchange of heat with the surrounding atmosphere. This can happen if the motion is rapid enough. We then have the following adiabatic relationship for p and T:

$$Tp^{(1-\gamma)/\gamma} = \text{const.} \tag{2.15}$$

where γ ($= c_p/c_v$) is the adiabatic constant, and $c_v = 712\,\text{J/kg K}$ and $c_p = 996\,\text{J/kg K}$ are the specific heat for air at constant volume and constant pressure, respectively. By differentiating (2.15) with respect to z we find

$$\frac{\partial T}{\partial z} = +\frac{\gamma - 1}{\gamma}\frac{T}{p}\frac{\partial p}{\partial z}$$

By introducing (2.6)

$$\frac{\partial T}{\partial z} = -\frac{\gamma - 1}{\gamma}\frac{mg}{k} = \alpha^* \tag{2.16}$$

where α^* is called the adiabatic "lapse rate".

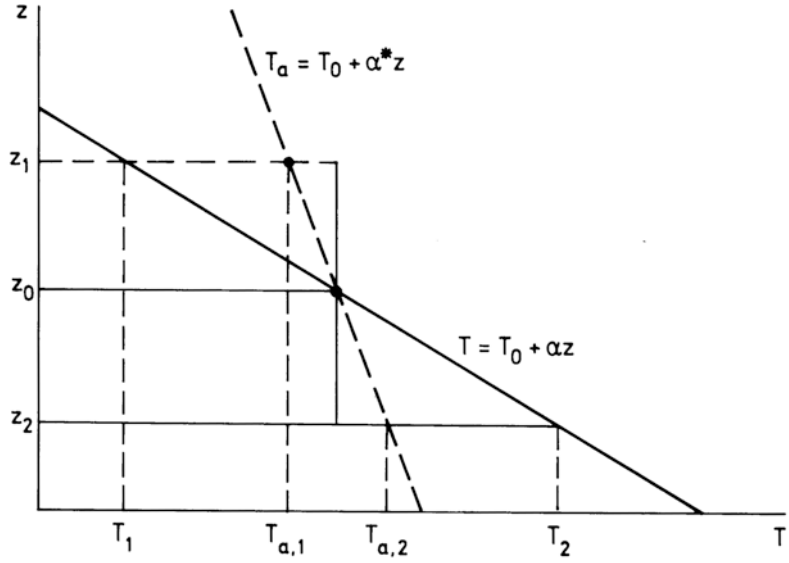


Figure 2.16. Atmospheric temperature T as a function of height with a lapse rate α as compared with temperature representing the adiabatic lapse rate α^* for an unstable atmosphere.

At the Earth's surface $\gamma = 1.4$ and therefore

$$\alpha^* = -9.8\ \mathrm{K/km}$$

Temperature decreases by almost 1 K per 100 m elevation at ground level.

Let us now assume that we have an atmosphere where the temperature within a certain height region decreases more rapidly than the adiabatic lapse rate (Figure 2.16).

If we now imagine a small bubble of air ascending from height z_0 to height z_1 without heat exchange with the environment, then the temperature of the bubble will follow the adiabatic temperature illustrated by T_a to the temperature $T_{a,1}$ which is above the temperature T_1 in the atmosphere itself at height z_1. Therefore, the air bubble will be lighter than the surrounding air and the bubble will continue to ascend.

If, on the other hand, the bubble at z_0 starts to descend to z_2 without heat exchange with the surroundings, the temperature in the bubble will be $T_{a,2}$ according to the adiabatic temperature. The temperature in the bubble will therefore be less than in the surrounding air, and the bubble becomes heavier and continues to sink. In a situation where the temperature of the air decreases more rapidly than the adiabatic lapse rate, the air is unstable.

For the opposite sense, when the temperature in the atmosphere decreases more slowly than the adiabatic lapse rate as illustrated in Figure 2.17, the atmosphere becomes stable.

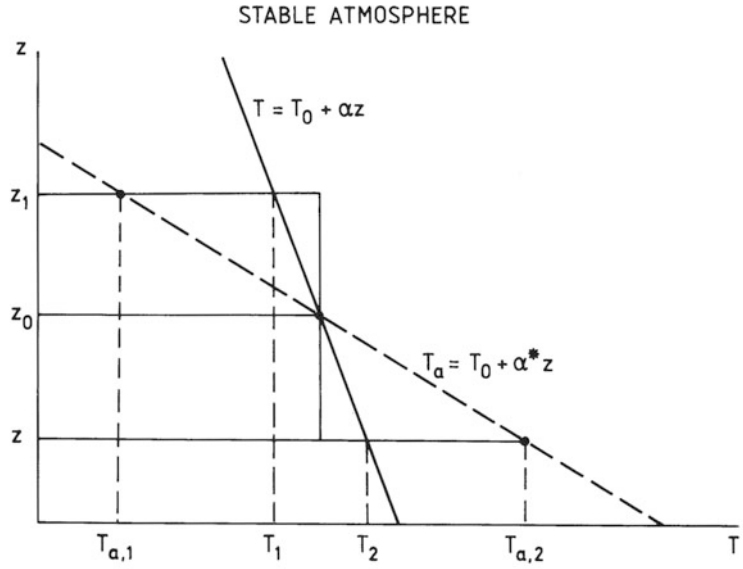


Figure 2.17. The same as in Figure 2.16 except the atmosphere is stable.

Small disturbances in the atmosphere are usually adiabatic. The most likely areas for an unstable atmosphere, however, are the regions of the atmosphere where the temperature decreases (i.e., in the troposphere and the mesosphere which are also the regions where turbulence is most prominent). In the stratosphere, where the temperature increases, the atmosphere is stable.

2.8 DIFFUSION

We have seen from Figure 2.2 that above 100 km the densities of different species decay according to individual scale heights. This comes about because above the mesopause the temperature increases and the atmosphere is stabilized; turbulence does not take place, and the different species are no longer homogeneously mixed. They will therefore distribute themselves according to the barometric law (2.5) with individual scale heights as if they were the only gas.

Sometimes, however, it is of interest to study the situation where the main gas, the major constituent, is distributed according to its specific scale height while another gas, a minor constituent, represented by a smaller density is moving through the majority one at a velocity that is determined by diffusion.

Assume first that there is no gravitational field and the majority gas is at rest and uniformly distributed. Let the minority gas with mean mass per molecule m be distributed by its density n, where there is a gradient dn/dx along the x-axis. Because of this gradient in the gas density the molecules will move down the

gradient at a speed v, and the particle flux $\phi = nv$ will be proportional to this gradient (Fick's law):

$$\phi = nv = -D \cdot \frac{\partial n}{\partial x} \tag{2.17}$$

where D is what we will call the diffusion coefficient with dimension m^2/s.

From the equation of continuity for a gas we have:

$$\frac{\partial n}{\partial t} + \frac{\partial}{\partial x}(nv) = 0$$

or when inserting (2.17)

$$\frac{\partial n}{\partial t} = -\frac{\partial}{\partial x}(nv) = \frac{\partial}{\partial x}\left(D\frac{\partial n}{\partial x}\right) \tag{2.18}$$

For a constant diffusion coefficient with respect to x we get:

$$\frac{\partial n}{\partial t} = D\frac{\partial^2 n}{\partial x^2}$$

which is the rate of change of the gas density at a given point in space. This is the diffusion equation for the gas density n.

Because of the gradient in the minority gas density, there will be a pressure force acting on the gas although the temperature is constant, and this is given by:

$$F_p = -\frac{\partial p}{\partial x} = -kT\frac{\partial n}{\partial x} \tag{2.19}$$

where it is assumed an ideal gas. If now each of the minority particles experiences ν collisions per unit time with the majority gas which is at rest, there will be a restoring force:

$$F_\nu = -nm\nu v$$

which, when no other force is acting, must balance the pressure force (2.19). Thus

$$-kT\frac{\partial n}{\partial x} - nm\nu v = 0$$

and

$$nv = -\frac{kT}{m\nu}\frac{\partial n}{\partial x} = -D\frac{\partial n}{\partial x}$$

A simple expression for the diffusion coefficient now emerges:

$$D = \frac{kT}{m\nu} \tag{2.20}$$

Since the collision frequency between the minority and the majority gas is proportional to the density of the majority gas, n_M, and the square root of the temperature which is the same for the two gases:

$$\nu \propto n_M T^{1/2}$$

the diffusion coefficient obeys the following proportionality

$$D \propto T^{1/2} n_M^{-1}$$

Diffusion therefore increases with temperature and decreases with density n_M.

Let now the space coordinate be vertical and assume that the gravity force mg is acting downward on each volume of the minority constituent, then the collisions must balance the sum of the pressure and gravity forces as follows:

$$-\frac{\partial p}{\partial z} - nmg = nm\nu w$$

where w is the vertical velocity. Then again for an ideal gas:

$$-kT\frac{\partial n}{\partial z} - nmg = nm\nu w$$

and solving for the vertical flux we get

$$nw = -\frac{kT}{m\nu}\left(\frac{\partial n}{\partial z} + \frac{gm}{kT}n\right) = -D\left(\frac{\partial n}{\partial z} + \frac{n}{H_m}\right) \tag{2.21}$$

D is given by (2.20), and $H_m = kT/mg$ is the scale height of the minority constituent. We notice that H_m enters the equation as a constant and does not need to be equal to the distribution height $[-(1/n)(\partial n/\partial z)]^{-1}$ of the minority gas.

Now, when applying (2.18) for vertical motion and inserting (2.21) for nw:

$$\frac{\partial n}{\partial t} = -\frac{\partial}{\partial z}(nw) = \frac{\partial}{\partial z}\left\{D\left(\frac{\partial n}{\partial z} + \frac{n}{H_m}\right)\right\} \tag{2.22}$$

From (2.20) we have that $D = kT/m\nu$, and since ν must be proportional to the density n_M of the majority constituent which is distributed according to the scale height H_M, we get:

$$\nu \propto n_M = n_{M_0} \exp(-z/H_M)$$

The diffusion coefficient can then be expressed as:

$$D = D_0 \exp(z/H_M) \tag{2.23}$$

It increases exponentially with height in an isothermal atmosphere. By inserting this expression for D into (2.22) we get:

$$\begin{aligned}\frac{\partial n}{\partial t} &= \left(\frac{\partial n}{\partial z} + \frac{n}{H_m}\right)\frac{\partial D}{\partial z} + D\left(\frac{\partial^2 n}{\partial z^2} + \frac{1}{H_m}\frac{\partial n}{\partial z}\right) \\ &= D\left\{\frac{\partial^2 n}{\partial z^2} + \left(\frac{1}{H_M} + \frac{1}{H_m}\right)\frac{\partial n}{\partial z} + \frac{n}{H_M H_m}\right\}\end{aligned}$$

If it is assumed that at a particular height z the density of the minority constituent is distributed according to the exponential function

$$n = n_0 \exp(-z/\delta) \tag{2.24}$$

where δ is the local scale height; then at this height z the density will change by time at a rate:

$$\begin{aligned}\frac{\partial n}{\partial t} &= D\left\{+\frac{1}{\delta^2}-\left(\frac{1}{H_m}+\frac{1}{H_M}\right)\frac{1}{\delta}+\frac{1}{H_M H_m}\right\}n \\ &= D\left\{\left(\frac{1}{\delta}-\frac{1}{H_m}\right)\left(\frac{1}{\delta}-\frac{1}{H_M}\right)\right\}n \\ &= \Gamma \cdot n\end{aligned}$$

As long as the density of the minority constituent remains approximately exponentially distributed by distribution height δ, its concentration will vary in time at altitude z by the rate $\Gamma \cdot n$ where Γ is given by:

$$\Gamma = D\left\{\left(\frac{1}{\delta}-\frac{1}{H_m}\right)\left(\frac{1}{\delta}-\frac{1}{H_M}\right)\right\}$$

This result, however, deserves some comments. In Table 2.1 we show the different values that δ can take in relation to H_M and H_m and the following values of Γ/D.

In most cases of δ, H_m and H_M are roughly of the same order of magnitude (Ratcliffe, 1972). There are two situations, however, where $\partial n/\partial t = 0$ or $\Gamma = 0$ (i.e., steady state). First, when $\delta = H_m$ or the minority constituent is distributed according to its natural scale height in (2.24). There will then according to (2.21) be no vertical motion because

$$w = -D\left(-\frac{1}{H_m}+\frac{1}{H_m}\right) = 0$$

Table 2.1. Time constants for loss by diffusion. (From Ratcliffe, 1972.)

Value of δ	*Approximate value of* Γ/D
$\delta > H_m$ and $\delta > H_M$	$1/H_M \cdot H_m$
$\delta = H_m$	0
$\delta > H_M$ and $\delta < H_m$	$-1/\delta \cdot H_M$
$\delta < H_M$ and $\delta > H_m$	$-1/\delta \cdot H_m$
$\delta = H_M$	0
$H_m > \delta$ and $H_M > \delta$	$1/\delta^2$
$\delta < 0$	$> 1/H_M \cdot H_m$

everywhere. The second case appears when $\delta = H_M$ in (2.24); the minority gas is distributed according to the scale height of the majority gas. Then, according to (2.21)

$$w = +D\left(+\frac{1}{H_M} - \frac{1}{H_m}\right) \neq 0$$

and w is positive if $H_m > H_M$, otherwise it is negative. The number of particles crossing a unit area per unit time will be according to (2.21):

$$nw = +nD\left(+\frac{1}{H_M} - \frac{1}{H_m}\right) = n_0 D_0\left(\frac{1}{H_M} - \frac{1}{H_m}\right)$$

where we have inferred D from (2.23) and n from (2.24) when $\delta = H_M$. The upward decrease in n ($\approx \exp(-z/H_M)$) is just equal to the upward increase in D ($\approx \exp(z/H_M)$) so that there is a steady flow of gas upwards or downwards depending on whether H_m is greater or smaller than H_M. In spite of the fact that $\partial n/\partial t = 0$ the situation is therefore not one of dynamic equilibrium.

2.9 THE EQUATION OF MOTION OF THE NEUTRAL GAS

Since we are going to discuss the motion of the atmosphere of a rotating planet, it is convenient to express the kinetic equations in a reference frame rotating with the planet at an angular velocity $\boldsymbol{\Omega}$. We start out with the general equation of motion for a neutral gas

$$\frac{d\mathbf{u}}{dt} = -\frac{1}{\rho}\nabla p + \mathbf{g} + \mathbf{f} \tag{2.25}$$

Here we have split up the external force into the pressure force, the force of gravity, $\mathbf{g}$, and any other accelerating force, $\mathbf{f}$, per unit mass.

We now denote by $\mathbf{u}_f$ the velocity observed in an inertial frame of reference (e.g., one fixed to the center of the Earth). The velocity $\mathbf{u}_r$, observed in a reference frame rotating with an angular velocity $\boldsymbol{\Omega}$ around an axis through the center of the Earth and at a distance $\mathbf{r}$ from this center, is according to the laws of mechanics related to $\mathbf{u}_f$ by (Figure 2.18):

$$\left(\frac{d\mathbf{r}}{dt}\right)_f = \left(\frac{d\mathbf{r}}{dt}\right)_r + \Omega \times \mathbf{r}$$

and

$$\mathbf{u}_f = \mathbf{u}_r + \Omega \times r$$

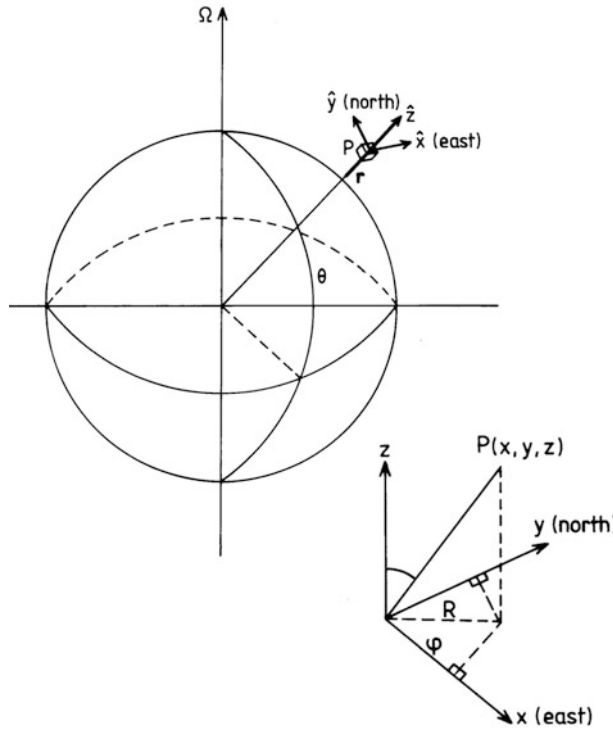


Figure 2.18. The coordinate system used to describe the motion of an air parcel at a distance r from the Earth's center and rotating with the Earth at an angular velocity Ω.

The acceleration of this air parcel in the fixed frame of reference is now, since $\mathbf{\Omega}$ is constant, given by:

$$\begin{aligned}
\mathbf{a}_f &= \left(\frac{d\mathbf{u}_f}{dt}\right)_f = \left(\frac{d}{dt}(\mathbf{u}_r + \mathbf{\Omega} \times \mathbf{r})\right)_f \\
&= \left(\frac{d}{dt}\mathbf{u}_r\right)_f + \left(\frac{d}{dt}(\mathbf{\Omega} \times \mathbf{r})\right)_f \\
&= \left(\frac{d\mathbf{u}_r}{dt}\right)_r + \mathbf{\Omega} \times \mathbf{u}_r + \mathbf{\Omega} \times \left(\frac{d\mathbf{r}}{dt}\right)_f \\
&= \mathbf{a}_r + \mathbf{\Omega} \times \mathbf{u}_r + \mathbf{\Omega} \times \mathbf{u}_r + \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}) \\
&= \mathbf{a}_r + 2\mathbf{\Omega} \times \mathbf{u}_r + \mathbf{\Omega} \times (\Omega \times \mathbf{r})
\end{aligned}$$

where $\mathbf{a}_r$ is acceleration in the rotating reference frame. This acceleration can be expressed as

$$\mathbf{a}_r = \mathbf{a}_f - 2\mathbf{\Omega} \times \mathbf{u}_r - \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r})$$

Since the equation of motion in the fixed frame of reference is given by

$$\mathbf{a}_f = \left(\frac{d\mathbf{u}}{dt}\right)_f = -\frac{1}{\rho}\nabla p + \mathbf{g} + \mathbf{f}$$

the equation of motion in a reference system following the Earth's rotation becomes

$$\mathbf{a} = \frac{d\mathbf{u}}{dt} = -\frac{1}{\rho}\nabla p + \mathbf{g} + \mathbf{f} - 2\boldsymbol{\Omega} \times \mathbf{u} - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r})$$

when deleting the index r. Expanding the final vector product we have

$$\frac{d\mathbf{u}}{dt} = -\frac{1}{\rho}\nabla p + \mathbf{g} + \mathbf{f} - 2\boldsymbol{\Omega} \times \mathbf{u} - (\boldsymbol{\Omega} \cdot \mathbf{r})\boldsymbol{\Omega} + \Omega^2 \mathbf{r}$$

It is then customary to introduce the potential function

$$\chi = -\mathbf{g} \cdot \mathbf{r} - \tfrac{1}{2}\Omega^2 r^2 \cos^2\theta$$

where θ is the latitude angle and r the radial distance to the point P as seen from the Earth's center (Figure 2.18).

Finally, the equation of motion takes the form:

$$\frac{d\mathbf{u}}{dt} + 2\boldsymbol{\Omega} \times \mathbf{u} + \frac{1}{\rho}\nabla p + \nabla\chi = \mathbf{f} \tag{2.26}$$

where $2(\boldsymbol{\Omega} \times \mathbf{u})$ is called the Coriolis acceleration, while $\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r})$ is the centrifugal acceleration.

For an observer fixed on the Earth we have the rotation frequency $\Omega = 7.27 \times 10^{-5}\ \mathrm{s}^{-1}$ and, since the Earth radius is $R_e = 6.37 \times 10^6$ m, we find $\Omega^2 R_e = 3.37 \times 10^{-2}\ \mathrm{m/s}^2$ which represents the maximum centrifugal acceleration (at the equator). This can be neglected as it is much smaller than $g = 9.81\ \mathrm{m/s}^2$.

We note that in the northern hemisphere the Coriolis force will tend to deflect motion to the right of the direction of the velocity vector. The Coriolis force is quite weak since for a velocity of 1,000 m/s, which is very high for neutral air velocity, the maximum acceleration corresponds to $1.49 \times 10^{-2}\, g$. For motions persisting over long periods, however, the deflection of velocity is quite important and creates, in meteorological terms, cyclones in the northern hemisphere. These are important elements in the weather system. For practical reasons we will now choose a Cartesian coordinate system (x, y, z) fixed on the Earth's surface as indicated in Figure 2.18 and denote the velocity components (u, v, w) along the different axes. u, v, and w are counted positive eastward, northward, and upward, respectively. Since we have neglected centrifugal force, the acceleration of gravity is given by

$$\mathbf{g} = -g\hat{\mathbf{z}}$$

The angular velocity vector is given by:

$$\boldsymbol{\Omega} = \Omega \sin\theta\, \hat{\mathbf{z}} + \Omega \cos\theta\, \hat{\mathbf{y}}$$

From (2.26) we find for each component of the acceleration:

$$\frac{du}{dt} - 2\Omega v \sin\theta + 2\Omega w \cos\theta + \frac{1}{\rho}\frac{\partial p}{\partial x} = f_x \tag{2.27a}$$

$$\frac{dv}{dt} + 2\Omega u \sin\theta + \frac{1}{\rho}\frac{\partial p}{\partial y} = f_y \tag{2.27b}$$

$$\frac{dw}{dt} - 2\Omega u \cos\theta + \frac{1}{\rho}\frac{\partial p}{\partial z} + g = f_z \tag{2.27c}$$

2.10 GEOSTROPHIC AND THERMAL WINDS

We will now study some of the steady-state solutions ($d/dt = 0$) by assuming first of all that there is no vertical motion ($w = 0$), then that all other external forces except gravity are negligible and finally that the barometric law (2.5) applies; that is:

$$\frac{\partial p}{\partial z} = -\rho g = -nmg$$

where $\rho = n \cdot m$. For horizontal motion we obtain from (2.27) the so-called geostrophic wind solution:

$$u = -\frac{1}{2\rho\Omega \sin\theta}\frac{\partial p}{\partial y} \tag{2.28a}$$

$$v = \frac{1}{2\rho\Omega \sin\theta}\frac{\partial p}{\partial x} \tag{2.28b}$$

We noticee that, since $\sin\theta = 0$ at the equator, there can be no geostrophic wind there. u and v change sign, however, across the equator. The geostrophic equations describe the motion of an air parcel that is initially moving from a high- to a low-pressure area. This motion is slow enough for the Coriolis force to deflect the air motion so that it finally becomes parallel to the isobars. We notice this because

$$\mathbf{v} \cdot \nabla p = u\frac{\partial p}{\partial x} + v\frac{\partial p}{\partial y} = 0$$

and v is perpendicular to the pressure gradients which by definition are perpendicular to the isobars.

By introducing the ideal gas law ($p = nkT$) in the geostrophic wind equations (2.28) we find the so-called thermal wind equation:

$$\frac{u}{T} = -\frac{k}{2m\Omega \sin\theta}\frac{\partial}{\partial y}(\ln p)$$

$$\frac{v}{T} = +\frac{k}{2m\Omega \sin\theta}\frac{\partial}{\partial x}(\ln p)$$

By now differentiating with respect to height we obtain, when $p = p_0 \exp(-z/H)$ from (2.8) and $z_0 = 0$:

$$\frac{\partial}{\partial z}\left(\frac{u}{T}\right) = -\frac{g}{2\Omega T^2 \sin\theta}\frac{\partial T}{\partial y}$$

$$\frac{\partial}{\partial z}\left(\frac{v}{T}\right) = +\frac{g}{2\Omega T^2 \sin\theta}\frac{\partial T}{\partial x}$$

These equations relate vertical wind shears in the geostrophic wind to horizontal gradients in the temperature.

Consider the northern hemisphere close to ground where it is hot at the equator and cold at the poles. Then, $\partial T/\partial y$ will be less than zero there so that $(\partial/\partial z)(u/T)$ becomes larger than zero. Up toward the tropopause therefore the meridional wind will increase eastward. Even if the wind is westward at the ground, it may become eastward at some height in the troposphere. In the southern hemisphere, because both $\partial T/\partial y$ as well as $\sin\theta$ change sign there, the wind will also increase towards the east as one ascends in the troposphere.

In the stratosphere the situation can be different. The summer pole is continually heated at a higher temperature than at the equator which in turn is hotter than the totally unilluminated winter pole in the stratosphere. Thus, thermal winds in the stratosphere will have a seasonal dependence.

The wind pattern and temperature distribution up to the mesosphere is fairly well known and, except for the lowest 1,000 meters of the atmosphere, the zonal winds appear to be adequately described by the thermal wind equation (Murgatroyd, 1957). It is at about 65 km that the summer and winter temperatures are roughly equal. Above this height in the mesosphere the winter temperatures are higher.

There are also meridional prevailing winds leading to a net latitudinal flow of air which must be balanced by return flows at other heights. This will set up large meridional wind cells.

2.11 THE WIND SYSTEMS OF THE UPPER ATMOSPHERE

We will now discuss the wind system in the mesosphere and thermosphere. In Figure 2.19 average zonal winds for January are shown below 130 km. We recognize the strong jet streams at the tropopause. Note that by meteorological usage E and W stands for easterly and westerly, which actually means winds coming from the east and west, respectively. This is the opposite sense to the convention in ionospheric physics where E and W means eastward- and westward-blowing winds, respectively.

In the mesosphere with its center around 60 km, however, there is a strong eastward-blowing wind in the winter hemisphere and a similar but weaker westward-blowing wind in the summer hemisphere. The maximum wind speed is observed close to 40° of latitude in both hemispheres. This wind pattern again

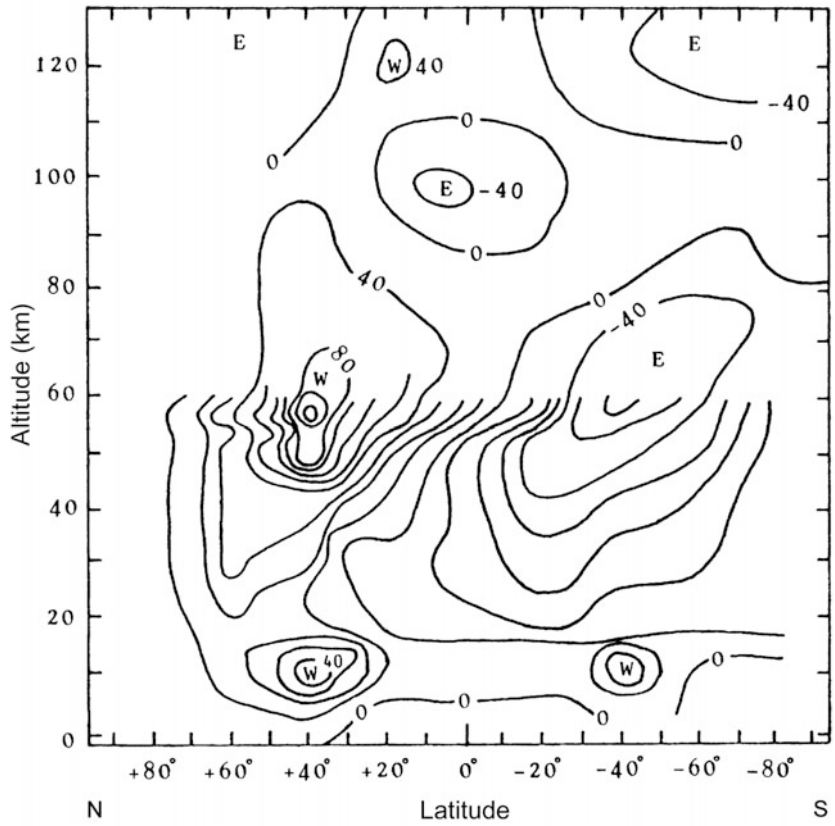


Figure 2.19. Zonal average east–west wind in January. Numerals on the contours are wind speed in units of $m\,s^{-1}$ below 130 km altitude. Positive numbers represent eastward winds. Notice the meteorological convention of W representing westerly (i.e., winds coming from the west). (From Tohmatsu, 1990.)

appears to be well explained by the thermal wind equation. As can be seen from the temperature distribution presented in Figure 2.20, the temperatures are high in the stratosphere in the summer hemisphere, while they are low in the winter hemisphere. The difference is about 50 K at about 40 km. In the mesosphere, however, the situation is the opposite where the temperature in the polar hemisphere at about 80 km is about 230 K, while it is less than 180 K in the summer hemisphere (as illustrated in Figure 2.21 where the vertical temperature distribution as observed from Fort Churchill (59° N) in summer and winter are compared). In contrast to average zonal winds in the troposphere, which are eastward in both hemispheres, the average zonal winds in the mesosphere have opposite directions.

Meridional winds in the mesosphere appear to have an average net flow from the summer hemisphere to the winter hemisphere. Large variability in this height

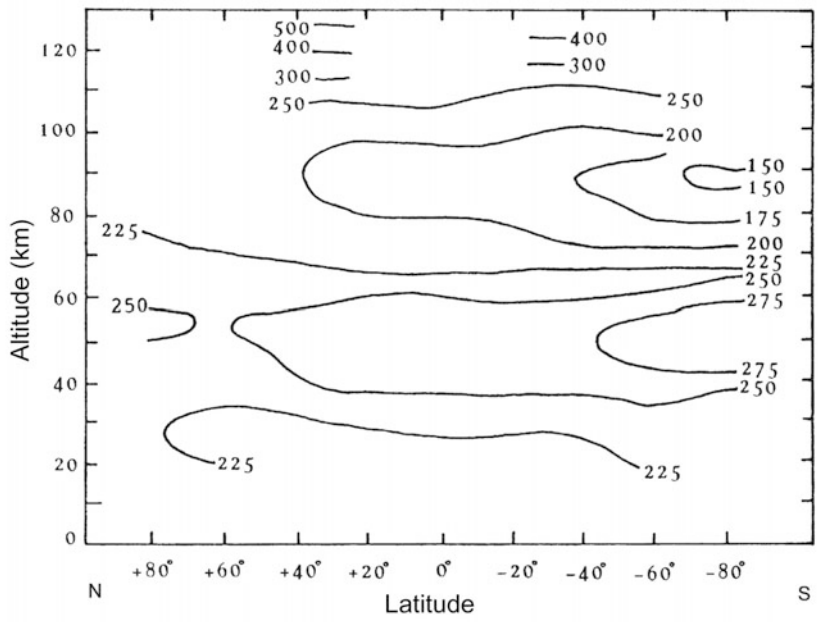


Figure 2.20. Meridional distribution of the mean atmospheric temperature in January. Numerals labeled on the contours are in kelvins. Notice that the mesosphere is warmer in wintertime than in summertime. (From Tohmatsu, 1990.)

region seems to be present as can be observed by comparing the schematic wind diagrams for 80 and 100 km, respectively, in Figure 2.22.

2.12 OBSERVATIONS OF THE NEUTRAL WIND

While winds at the ground can be sensed by the human body and relatively easily measured, the wind at higher altitudes is a rather elusive parameter to observe. In the stratosphere and lower mesosphere our knowledge of the neutral wind has been derived from balloon and meteorological rocket measurements. One method quite often used has been observations of anomalous propagation of sound waves from small grenade detonations released by rockets. Higher up in the mesosphere and lower thermosphere meteor trail tracking by radars has produced considerable insight in the neutral wind system at these height regions. Observations of the motion of noctilucent clouds have also been used to deduce neutral winds in the mesopause region.

At higher altitudes in the lower thermosphere information about the neutral wind has, in addition to meteor trail observations, been derived from chemiluminescent cloud releases from rockets at twilight. Solar radiation that is resonance-scattered by alkali metal atoms can easily be observed from ground in

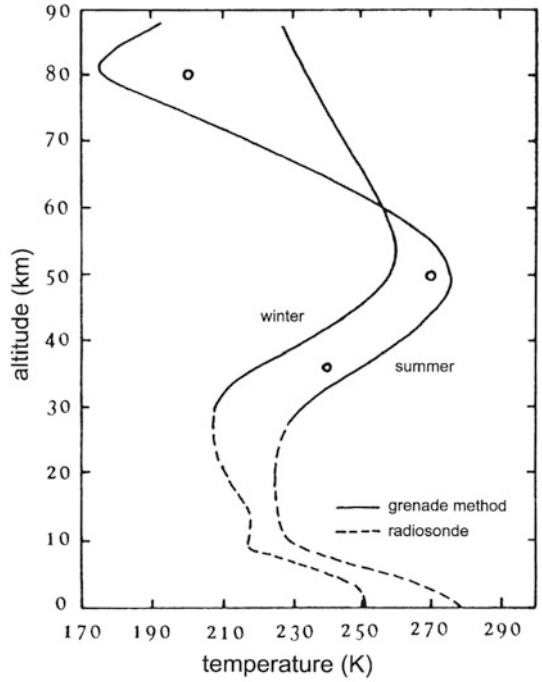


Figure 2.21. Vertical distributions of the mean atmospheric temperature in summer and winter measured at Fort Churchill (59° N). The solid and dashed curves are those measured by the grenade method and radiosondes, respectively. Open circles represent the average values at White Sands (23° N). (From Stroud *et al.*, 1959.)

clear weather. The extended height profiles of the neutral wind in the lower thermosphere can also be obtained by chemical trimethyl aluminum (TMA) releases from rockets. As these chemical releases have to be made at twilight, limited time coverage is available for these methods.

For F-region and partly also E-region neutral winds at high-latitude optical observations of the Doppler shift in auroral emission lines from oxygen can be used at high latitudes. But again the ability of the method to give extended time coverage of the winds is limited to the period of dark and clear polar sky. Optical Doppler measurements of airglow emissions have also been used to study F-region neutral winds at middle and lower latitudes.

The incoherent scatter method for deriving F-region and E-region neutral winds is a method that can give continuous coverage. This method, as we will see, is rather indirect and is still related to large uncertainties in the choice of model atmospheres especially at high latitudes.

In the upper thermosphere observations of atmospheric drag on satellites have been used to derive the temperature distribution in these higher regions, and from these again neutral wind patterns have been derived by applying these temperature

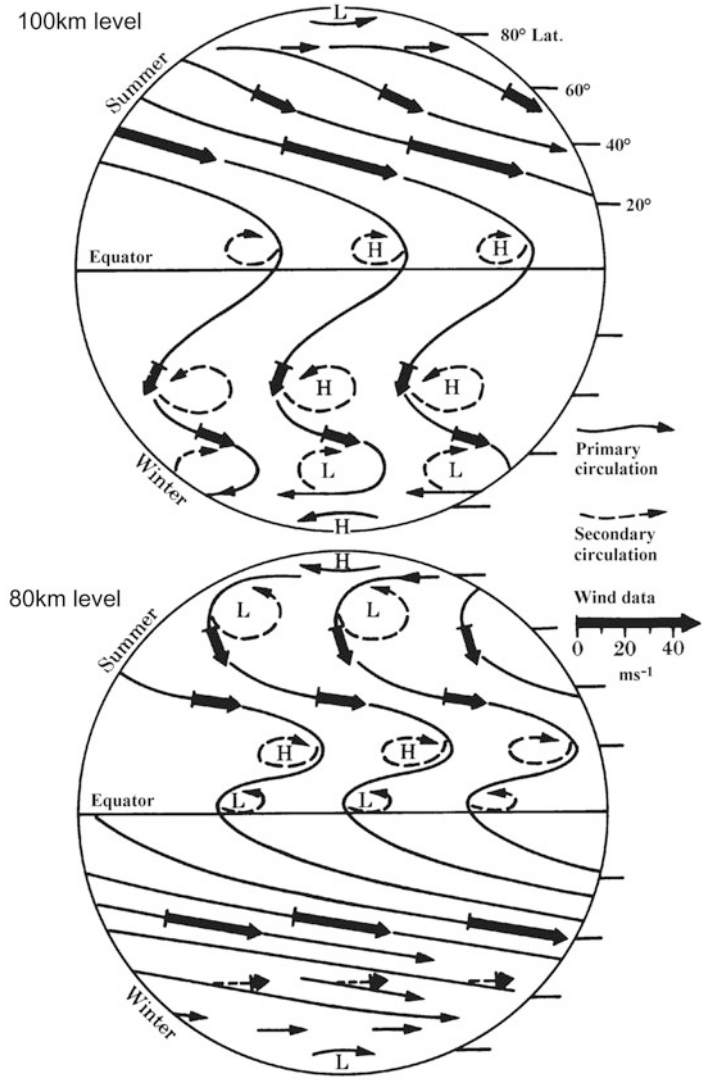


Figure 2.22. Schematic models of the planetary neutral air circulation at 80 and 100 km. (From Kochanski, 1963.)

observations as an input to the pressure term in the mobility equation of the neutral gas.

For E-region neutral winds, probably the most important basis for our understanding has been the analysis of geomagnetic variations on the ground. These have a very long history which has resulted in a larger amount of work attempting to resolve the different factors contributing to magnetic variations—one of the most important of these being the E-region neutral wind setup by tidal forces.

2.13 COLLISIONS BETWEEN PARTICLES

Let us assume a gas where we have two kinds of gas species, m and n, that are colliding with each other. Each collision must be related to a mutual force. Assume, for simplicity, that this force per unit mass acting on m-type particles due to the collision with n-type particles is proportional to the velocity difference $\mathbf{u}_m - \mathbf{u}_n$ between the two kinds such that

$$\mathbf{f}_{m,n} = -\nu_{mn}(\mathbf{u}_m - \mathbf{u}_n)$$

is the force reducing $\mathbf{u}_m$ when $\mathbf{u}_m > \mathbf{u}_n$ and vice versa. ν_{mn} is called the collision frequency for momentum transfer. The acceleration of m-type particles due to collisions with n-type particles is now:

$$\frac{d\mathbf{u}_m}{dt} = \mathbf{f}_{mn} = -\nu_{mn}(\mathbf{u}_m - \mathbf{u}_n) \tag{2.29}$$

Since momentum must be conserved during a collision, we have

$$\rho_m \mathbf{f}_{m,n} + \rho_n \mathbf{f}_{n,m} = 0$$

since $\mathbf{f}_{m,n}$ and $\mathbf{f}_{n,m}$ are forces per unit mass, and $\mathbf{f}_{n,m}$ is the force on particles of type n due to the collision with particles of type m. ρ_m and ρ_n are the mass density of species m and n, respectively. Inserting for $\mathbf{f}_{m,n}$ and $\mathbf{f}_{n,m}$ gives:

$$\rho_m \nu_{mn}(\mathbf{u}_m - \mathbf{u}_n) + \rho_n \nu_{nm}(\mathbf{u}_n - \mathbf{u}_m) = 0$$

$$(\rho_m \nu_{mn} - \rho_n \nu_{nm})(\mathbf{u}_m - \mathbf{u}_n) = 0$$

For any $\mathbf{u}_m$ and $\mathbf{u}_n$ therefore

$$\rho_m \nu_{mn} = \rho_n \nu_{nm} \tag{2.30}$$

For simplicity we only consider one-dimensional velocities and that there is only one kind of particle type n, then the equation of motion for the m-type particle can be written:

$$\frac{du_m}{u_m - u_n} = -\nu_{mn}\, dt$$

and the solution for u_m is:

$$u_m = u_m^0 \exp(-\nu_{mn} t) + u_n(1 - \exp(-\nu_{mn} t))$$

if $u_m = u_m^0$ at $t = 0$. The equilibrium solution ($t \to \infty$) for u_m is:

$$u_m = u_n$$

The characteristic relaxation time is defined as:

$$\tau_{m,n} = \frac{1}{\nu_{mn}}$$

In the opposite case the velocity of the n-type particle is given from (2.29) by:

$$u_n = u_n^0 \exp(-\nu_{nm} t) + u_m(1 - \exp(-\nu_{nm} t))$$

and the relaxation time is:

$$\tau_{n,m} = \frac{1}{\nu_{nm}} = \frac{\rho_n}{\rho_m \nu_{mn}} = \frac{\rho_n}{\rho_m} \tau_{mn}$$

when (2.30) is applied. We therefore notice that, if $\rho_m \gg \rho_n$, the n-type particle will reach the equilibrium velocity $\mathbf{u}_m$ much more quickly than the m-type particle will reach its equilibrium velocity $\mathbf{u}_n$. In such a situation n-type particles will feel the presence of m-type particles much more strongly than the latter type feels the former. This is in simple terms the effect of drag from one kind of gas particle to another.

The mean velocity is defined as:

$$\mathbf{u} = \frac{\rho_m \mathbf{u}_m + \rho_n \mathbf{u}_n}{\rho}$$

where

$$\rho = \rho_m + \rho_n$$

Now, since $d\rho/dt = 0$ we obtain

$$\rho \frac{d\mathbf{u}}{dt} = -(\rho_m \nu_{mn}(\mathbf{u}_m - \mathbf{u}_n) + \rho_n \nu_{nm}(\mathbf{u}_n - \mathbf{u}_m)) = 0$$

Therefore, the mean velocity $\mathbf{u}$ is constant and

$$\mathbf{u}(\infty) = \mathbf{u}(t) = \mathbf{u}(0)$$

The collision term being proportional to the relative velocities between each type of particle contributes to a force which tends to drive each individual gas velocity toward a total mean velocity which is constant (i.e., an equalization in velocity takes place).

For the collisional interaction between ions and neutrals in the ionosphere it is of interest to compare the time constants with which the ions approach the neutral motion (τ_{in}) and vice versa (τ_{ni}). Such a comparison is presented for the E-region in Figure 2.23 for different days at an auroral zone station (Tromsø). We notice that while τ_{in} is always less than 10^{-1} s, τ_{ni} is hardly less than 10^4 s. A constant ion velocity acting for several hours is therefore needed in order to bring the neutrals into a motion along with the ions. On the other hand, the ions will almost immediately adjust their velocities to the neutral motion. Therefore, the ion motion can be used as a good tracer of the neutral motion.

2.14 COLLISIONS IN GASES WITH DIFFERENT TEMPERATURES

It is customary to treat the gas of the upper atmosphere as a plasma or fluid, assuming the collisions between individual particles are frequent enough for an MHD description to be valid.

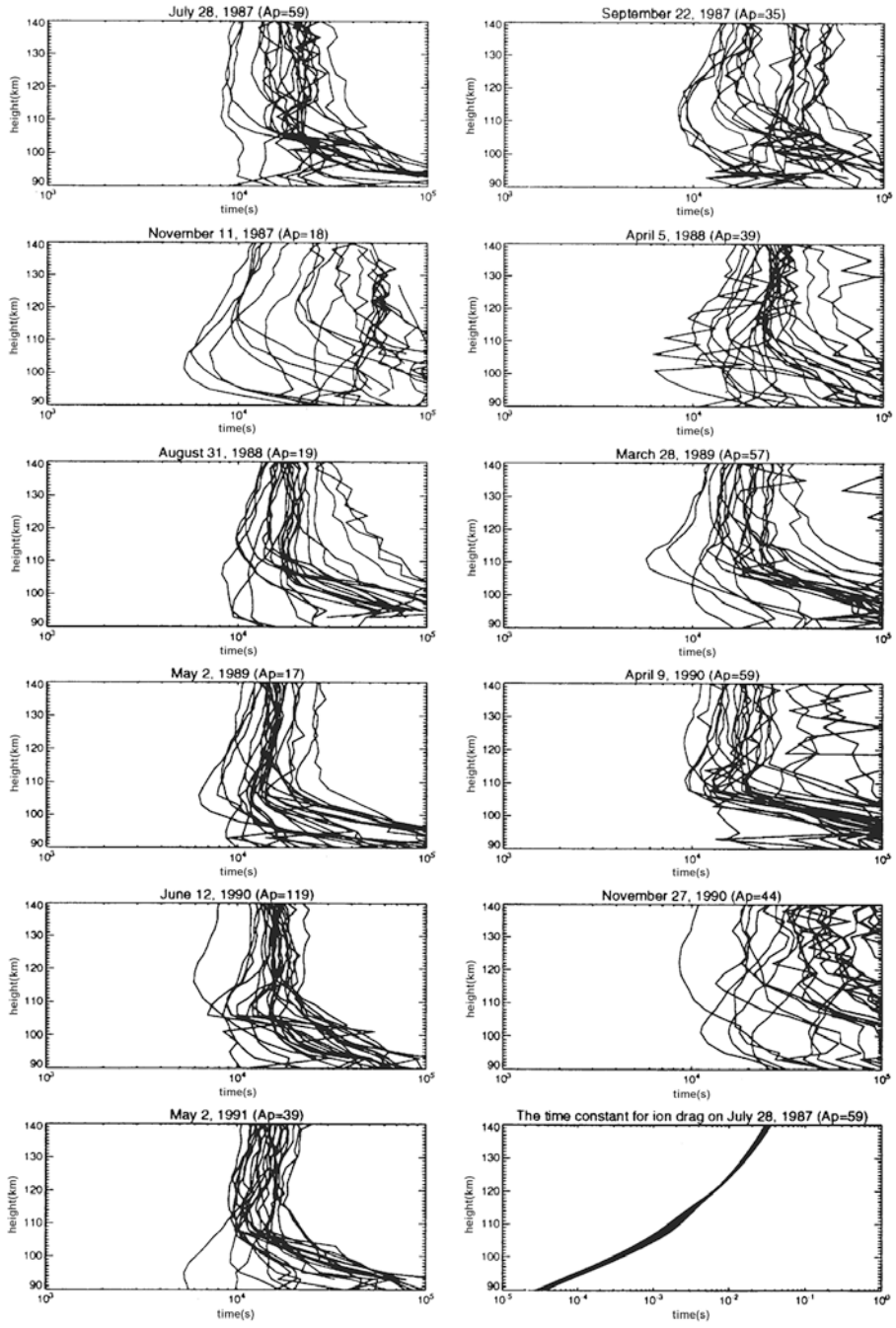


Figure 2.23. A comparison of the time constant τ_{in} for the ions to approach the neutral velocity (the lowest right panel) and τ_{ni} for the neutrals to approach the ion velocity for different days in the E-region above Tromsø (the 11 other panels). (From Nozawa and Brekke, 1995.)

To discuss the collisional interactions between individual kinds of particles m and n in the ionospheric plasma is not a straightforward problem, and it is not the purpose of this book to outline this in detail here.

By performing a statistical analysis where the macroscopic behavior of the plasma is considered to be due to a mean behavior of individual particles, one introduces the distribution function $f_m(\mathbf{r}, \mathbf{u}, t)$. This describes the distribution of particles of kind m having a velocity $\mathbf{u}$ at position $\mathbf{r}$ at time t. The Boltzmann equation describing the variation of this function in space and time can be written as:

$$\frac{\partial f_m}{\partial t} + v_i \frac{\partial f_m}{\partial x_i} + a_i \frac{\partial f_m}{\partial u_i} = \left(\frac{\partial f_m}{\partial t}\right)_c$$

where the index i indicates three individual coordinates, $i = 1, 2, 3$, and the term to the right is the collision term which is crucial for determining the shape of f_m.

By using this equation, however, and forming statistical averages, it is possible to find conservation laws for most parameters related to the behavior of particles of type m.

Boltzmann introduced the following term:

$$\left(\frac{\partial f_m}{\partial t}\right)_c = \sum_n \int (f'_m f'_{n1} - f_m f_{n1}) g b \, db \, d\varepsilon \, d^3 v_1$$

where f_m and f_{n1} are the distribution functions of particles of kind m and n before the collision, respectively, and f'_m and f'_{n1} are the corresponding parameters after the collision, g is the relative velocity between the two kinds of particles, b is the impact parameter, ε is an angle to account for all directions of impact, and $d^3 v_1$ is the velocity space for n-type particles.

For so-called Maxwellian gas particles, where the reaction force between the particles can be expressed as

$$F_R = \frac{\kappa}{r^5}$$

where κ is an arbitrary constant, it can be shown that for such particles in thermodynamic equilibrium the collision term in the energy equation results in a heat transfer rate

$$\frac{dQ_m}{dt} = \frac{\rho_m \nu_{mn}}{m_m + m_n} [3k(T_n - T_m) + m_n(\mathbf{u}_n - \mathbf{u}_m)^2]$$

where k is Boltzmann's constant, T_n and T_m are the temperatures of particles of kind n and m, respectively, and m_n and m_m are their respective masses. This equation shows that as long as the temperatures or the velocities of the two kinds of particles are different, heat transfer between them will occur.

In the ionosphere where the two kinds of particles are neutrals and ions, the last term in the equation above is often referred to as the Joule heating term due to ohmic losses by ionospheric currents. Assuming one type of ions (i) and one

type of neutrals (n) only, we obtain for the heating rate:

$$\frac{dQ_i}{dt} = \frac{\rho_i \nu_{in}}{m_i + m_n}[3k(T_n - T_i) + m_n(\mathbf{u}_n - \mathbf{v}_i)^2] \tag{2.31}$$

where we have introduced $\mathbf{v}_i$ and $\mathbf{u}_n$ as the ion and neutral velocities, respectively.

2.15 DRAG EFFECTS

Let us consider the ionosphere where we have one kind of ion and one kind of neutral with velocities $\mathbf{v}_i$ and $\mathbf{u}_n$, respectively. Assume an initial situation where $\mathbf{u}_n = 0$ and that an electric field $\mathbf{E}_\perp$ perpendicular to $\mathbf{B}$ induces an ion velocity $\mathbf{v}_{i0}$ according to the "frozen-in" concept:

$$\mathbf{v}_{i0} = \mathbf{E}_\perp \times \mathbf{B}/B^2$$

If $\mathbf{E}_\perp$ is directed eastward (Figure 2.24(a)), then v_{i0} will be upward and northward in the magnetic meridional plane. There will then be a horizontal ion velocity in this plane. Since the neutral gas is considered to be in hydrostatic equilibrium, we can neglect any motion of the neutrals in the vertical direction. The horizontal ion velocity component will then drag the neutrals along after some time. Since the neutrals effectively drag the ions with them, the ions will also obtain a component $\mathbf{v}_{i\parallel}$ parallel to the magnetic field. A stationary state will be obtained when the ions and the neutrals are moving horizontally at the same speed. This occurs when

$$u_n = \frac{v_{i0}}{\sin I} = \frac{1}{\sin I}\frac{E}{B}$$

where I is the inclination angle of $\mathbf{B}$. The time constant for this to happen in the ionosphere is very long due to the small ρ_i/ρ_n ratio. Assume instead the ions are initially at rest and that a neutral wind $\mathbf{u}_{n0}$ is blowing horizontally in the negative y-direction (south) in the northern hemisphere (Figure 2.24(b)). The component of the wind parallel to the magnetic field will drag the ions until they reach a steady-state velocity $\mathbf{v}_i$ along the field which is given by

$$v_i = u_{n0} \cos I$$

The ions therefore attain a vertical velocity given by

$$v_{iV} = u_{n0} \cos I \sin I$$

and a horizontal velocity

$$v_{iH} = u_{n0} \cos^2 I$$

Let us again assume that $\mathbf{u}_n = 0$ initially and that the plasma is diffusing downward at a velocity v_{iD} (Figure 2.24(c)) due to gravity and pressure gradients. Since a magnetic field is present, ions are forced to follow field lines. The ion motion will therefore attain a horizontal component given by:

$$v_{iH} = v_{iD} \cos I$$

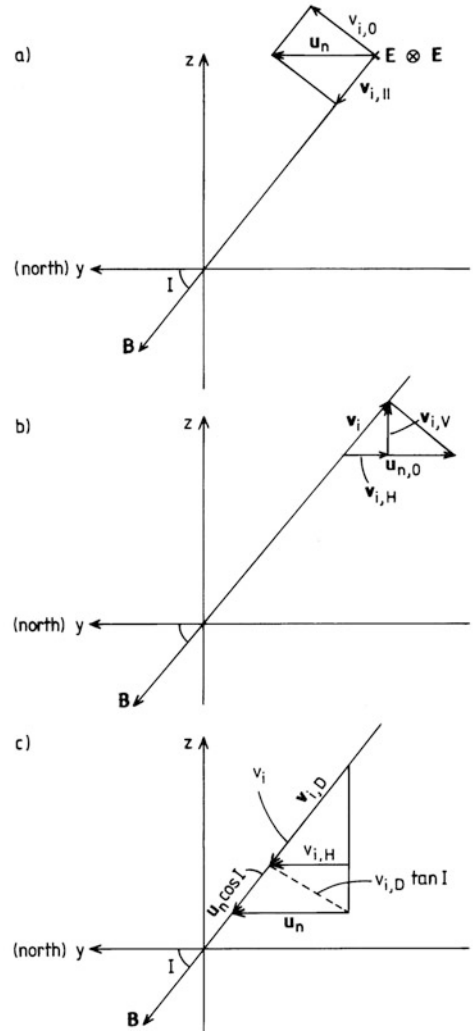


Figure 2.24. Diagrams illustrating the different drag effects taking place between the neutral and the ionized gas species in the upper atmosphere. (a) The neutral gas is initially at rest. An electric field perpendicular to **B** is associated with an ion drift $\mathbf{v}_{i0}$. Since the neutrals are restricted to moving mainly horizontally, there will be a horizontal neutral velocity $\mathbf{u}_n$. This neutral velocity will drag the ions along the **B**-field, and the ions will obtain a velocity component $\mathbf{v}_{i\parallel}$ parallel to **B**. A steady-state situation occurs when the ions and the neutrals are moving horizontally at the same speed. (b) The ions are initially at rest and the neutral gas moves horizontally at velocity $\mathbf{u}_{n0}$. The component of the neutral velocity along **B** drags the ions along until they reach a steady-state velocity $\mathbf{v}_i$ equal to this component. (c) The neutral gas is initially at rest while the plasma diffuses downward along **B** at a velocity $\mathbf{v}_{iD}$. The horizontal component of this velocity will drag the neutrals along until they reach a velocity $\mathbf{u}_n$. This horizontal neutral motion will again drag the ions down along the **B**-field, and a steady state occurs when the horizontal neutral and ion velocities are equal.

This will drag the neutrals along until they eventually obtain a horizontal velocity $\mathbf{u}_n$. Then the neutrals will also drag the ions along the B-field by the component $u_n \cos I$. Equilibrium is obtained when the ions and neutrals move horizontally at the same speed. This happens when

$$(v_{iD} + u_n \cos I) \cos I = u_n$$

$$u_n = \frac{v_{iD} \cos I}{\sin^2 I}$$

and then the plasma velocity along the field becomes:

$$\mathbf{v}_i = \mathbf{v}_{iD} / \sin^2 I$$

2.16 THERMOSPHERIC NEUTRAL WINDS

Atmospheric drag on satellites has been extensively used to deduce the temperature distribution in the upper thermosphere, especially at the thermopause. Figure 2.25 is a representation of the inferred isotherms at the thermopause on the basis of a large number of satellite passes. The temperature distribution is rather symmetric around the equator but is shifted about one hour to the east of the

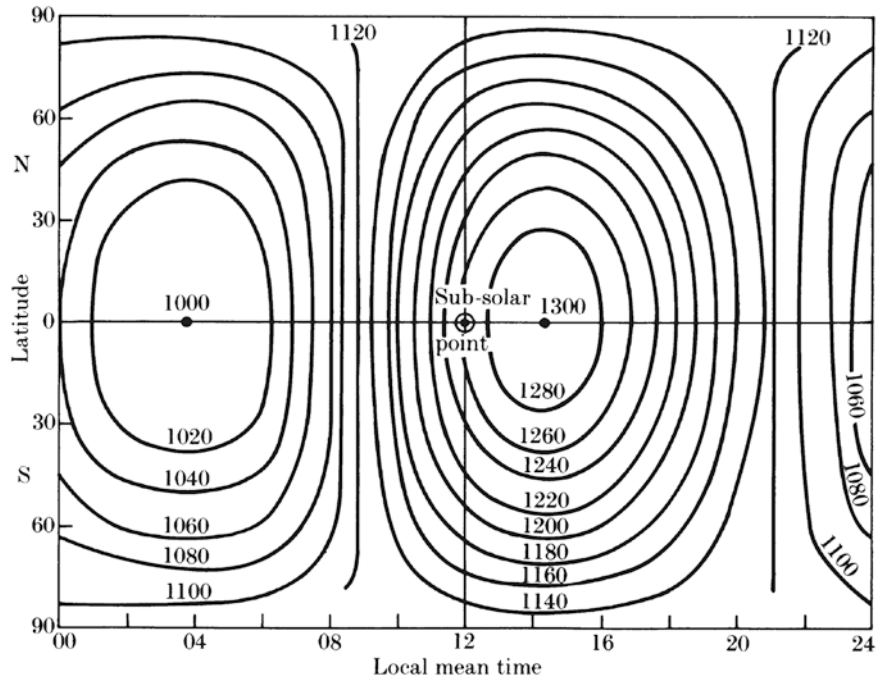


Figure 2.25. The inferred isotherms at the thermopause. The numbers labeled on the contours are in kelvins. (From Jacchia, 1965.)

subsolar point, probably due to sluggishness in the neutral gas. For this special diagram a maximum temperature of about 1,300 K is observed at the equator at about 14:00 LT. This maximum temperature, however, is sensitive to solar cycle variations and to magnetospheric disturbances. In fact, it does appear that the temperature at the thermopause is more sensitive to magnetic disturbances than to variations in solar EUV flux.

The daily global temperature distribution as shown in Figure 2.25 leads to a daytime expansion of the atmosphere which is called the "diurnal bulge". The horizontal pressure gradients around this "diurnal bulge" provide the driving force for thermospheric winds. From this type of temperature distribution it is then possible to derive the large-scale pressure gradients and from those again to calculate the resulting neutral wind when applying an appropriate equation of motion for the neutrals. Figure 2.26 gives an example of such calculations where we notice that the velocities are directed from the hot dayside across the pole to the nightside (i.e., the wind is very close to being perpendicular to the isotherms). The wind speeds are of the order of 40 m/s.

Let us for simplicity neglect all other terms than the pressure gradient force and **f** in the equation of motion for the neutral gas (2.25). Then for a steady-state solution of the horizontal winds (as Figure 2.26 represents):

$$\mathbf{f} = \frac{1}{\rho}\nabla p$$

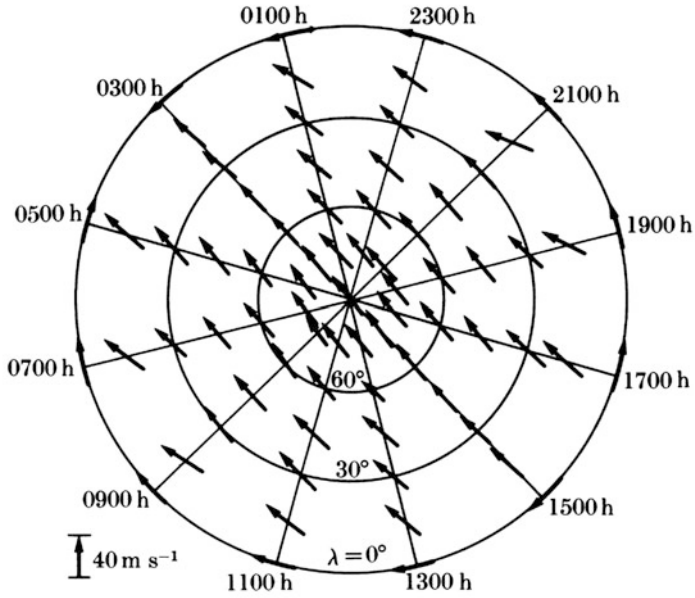


Figure 2.26. The inferred wind system at an altitude of 300 km based on a temperature distribution of the thermopause similar to the one shown in Figure 2.25. (From Kohl and King, 1967.)

If now **f** is due only to collisions $(-\nu_{ni}(\mathbf{u}_n - \mathbf{v}_i))$ between the neutral gas and the ions in the thermosphere and assuming that the ions are stationary, then:

$$u_n = -\frac{1}{\rho \cdot \nu_{ni}} \nabla p = -\frac{k}{m\nu_{ni}} \nabla T$$

where we have assumed an ideal gas and neglected horizontal variations in the neutral density. ν_{ni} is the neutral–ion collision frequency. In this simplified situation we therefore find that the neutral wind should blow down the temperature gradients (i.e., perpendicularly to the isotherms shown in Figure 2.25). We also notice, however, that only for very high collision frequencies can it be legitimate to neglect all other terms in the equation of motion except the pressure force and the collision term. For smaller collision frequencies the situation will be more complex.

A more realistic treatment makes the equation of motion for the neutrals far more difficult to solve, especially when we notice that the ions are not at all stationary at times at high latitudes when they are acted on by electric fields propagating from the magnetosphere. Furthermore, neutral air motion at one height can carry the ions along, which will in turn set up polarization fields that can propagate to other heights and other latitudes along magnetic field lines where they can act on the ion motion. The collisions between the ions and the neutrals will always be present, forcing the different species to drag each other.

Returning again to the equation of motion (2.25) and dividing the force term **f** up into a potential term, a viscosity term, and a collision term, we can write:

$$\frac{\partial \mathbf{u}_n}{\partial t} = -(\mathbf{u}_n \cdot \nabla)\mathbf{u}_n - 2\boldsymbol{\Omega} \times \mathbf{u}_n + \mathbf{g} - \frac{1}{\rho}\nabla p - \nabla\psi + \frac{\mu}{\rho}\nabla^2 \mathbf{u}_n - \nu_{ni}(\mathbf{u}_n - \mathbf{v}_i) \quad (2.32)$$

where ψ is a potential due to the centrifugal force and tides, μ the coefficient of viscosity, and $\mathbf{v}_i$ the ion velocity. In order to get a better understanding of the neutral wind behavior we need to know more about the different terms. Especially since we observe that the velocity of the ion gas enters the equation of motion for the neutrals, the equation of motion for the ions should also be solved simultaneously to give a self-consistent picture. Let us therefore introduce the equation of motion for ions:

$$m_i \frac{\partial \mathbf{v}_i}{\partial t} = -m_i(\mathbf{v}_i \cdot \nabla)\mathbf{v}_i + q\mathbf{E} + q\mathbf{v}_i \times \mathbf{B} - m_i\nu_{in}(\mathbf{v}_i - \mathbf{u}_n) \quad (2.33)$$

where m_i is the ion mass, q the ion charge, and **E** the electric field. We have neglected the Coriolis force, gravity, potential, and pressure forces together with viscosity since these are considered to be small in the thermosphere compared with the electric field, Lorentz force, and collision terms. We now observe that since the electric field and Lorentz force enter into the equation of motion for the ions, these may also affect the motion of the neutrals through the collision term between ions and neutrals. In a situation where **E** is very large, the ion velocity may dominate and the neutrals act as a drag on ions. In the opposite case when the **E**-field is negligible, the ions may act as a drag on neutrals. Since the **E**-field is

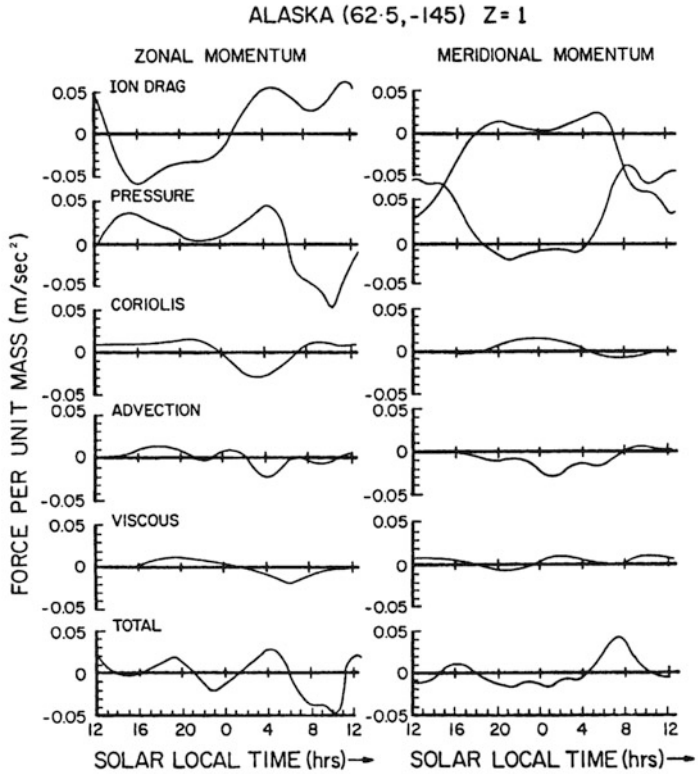


Figure 2.27. Analysis of forces for the neutral gas over the Alaska station (62.5° N, −145° E) calculated under conditions corresponding to December 4, 1981, for the F-region height at about 300 km. (From Killeen and Roble, 1986.)

so variable, especially at high latitudes, it is realized that the neutral gas motion can be a strongly modified and dynamic feature of the thermosphere.

Since there are so many dynamic parameters affecting the neutral gas motion, it is an almost insurmountable problem to obtain a comprehensive view of the situation. We have to make some assumptions and hope that the results we derive have a validity general enough that we can use them as a rule of thumb.

In Figure 2.27 calculations obtained for a high-latitude station (Chatanika, Alaska, 62.5° N, −145° E) pertinent to different terms in the equation of motion for the neutrals are illustrated. The potential term $\nabla\psi$ in equation (2.32), however, is neglected.

We notice from Figure 2.27, which applies for an altitude close to 300 km, that the ion drag and the pressure terms are dominant for both meridional and zonal components.

In the auroral zone and high latitudes where the ion velocity may become much larger than the neutral velocity, it is not legitimate to neglect the ion velocity in the collision term in equation (2.32). In a steady state to a first approximation

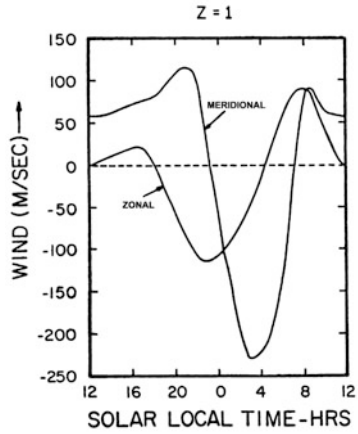


Figure 2.28. F-region neutral wind predictions for a station in Alaska corresponding to conditions as of December 4, 1981, at F-region heights (300 km). Of particular interest is the strong equatorward surge in the meridional component after midnight: a common feature observed at this station is apparently a consequence of the pressure and ion drag terms acting together on the neutral gas. (From Killeen and Roble, 1986.)

the neutral wind will therefore be:

$$\mathbf{u}_n = -\frac{1}{\rho\nu_{ni}}\nabla p + \mathbf{v}_i$$

This implies that the ion velocity can have a strong effect on the neutral velocity. This is illustrated in Figure 2.28 where a large equatorward surge in the neutral velocity occurs at night-time having a maximum of close to 200 m/s. This is apparently related to an enhancing effect of the pressure term and the ion drag term in the mobility equation. They both happen to force the neutrals in the same direction at this time over Alaska. A southward motion would be expected from the pressure gradients alone, but the ion motion, which for this particular day, December 4, 1981, was forced by a cross-polar cap electric field corresponding to about 60 kV, enhanced the equatorward motion in the morning hours.

At higher latitudes, within the polar cap where it is possible to observe the Doppler-shifted O I (6,300 Å) line throughout a 24-hour period in midwinter, the neutral wind is found to be nearly cross-polar from day to night (Figure 2.29). In this diagram the circle marks the latitude of observation (Longyearbyen, Svalbard, 78° N), and the center of the circle is the geographic north pole. The diagram does not represent a snapshot of the wind vectors since the observations are collected over a 24-hour period. Under the assumption, however, that the wind pattern does not change during this period of observations, it can be interpreted as an instantaneous pattern.

In Figure 2.30(a) more detailed model calculations of the global F-region neutral wind circulation field are demonstrated as a function of longitude and latitude toward a background of isothermal contours. This can be compared with

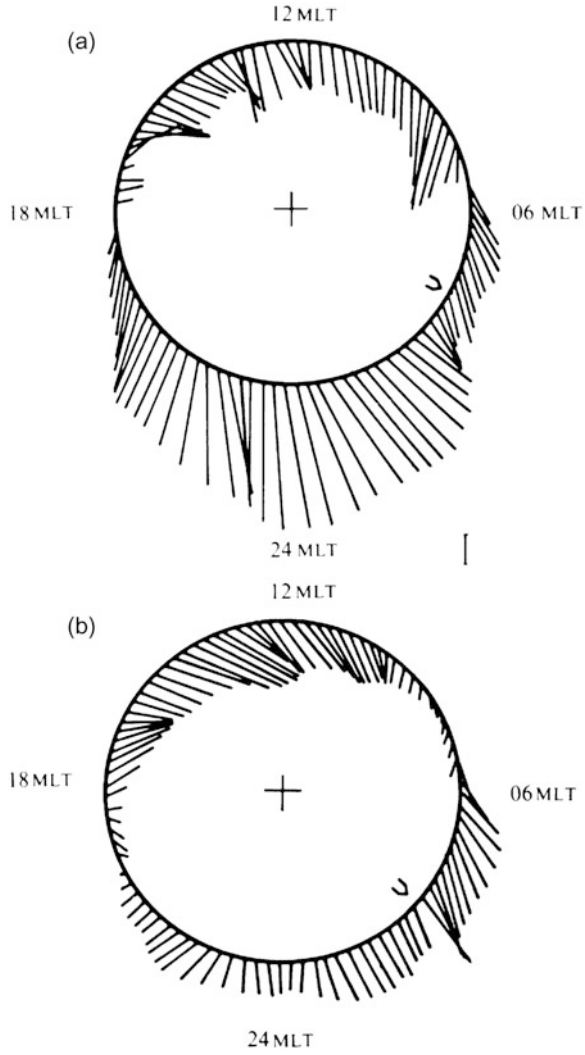


Figure 2.29. Neutral wind patterns in the polar thermosphere obtained for two days, January 21 and 29, 1979. The circle and the center represent the latitude of Longyearbyen, Svalbard (78° N) and the geographic north pole, respectively. An arrow pointing toward the center is a northward wind and tangentially toward the right from north is eastward. The scale bar corresponds to a wind velocity of 100 m/s. (From Smith and Sweeney, 1980.)

the one shown in Figure 2.26. In this model (Figure 2.30(a)) there is no potential drop across the polar cap such that ions act as a drag on neutrals. Again we notice that the main wind field is perpendicular to the isotherms from the hot subsolar point at about 16:00 LT toward the cooler nightside. The cross-polar cap neutral wind is rather outstanding, reaching velocities of the order of 100 m/s.

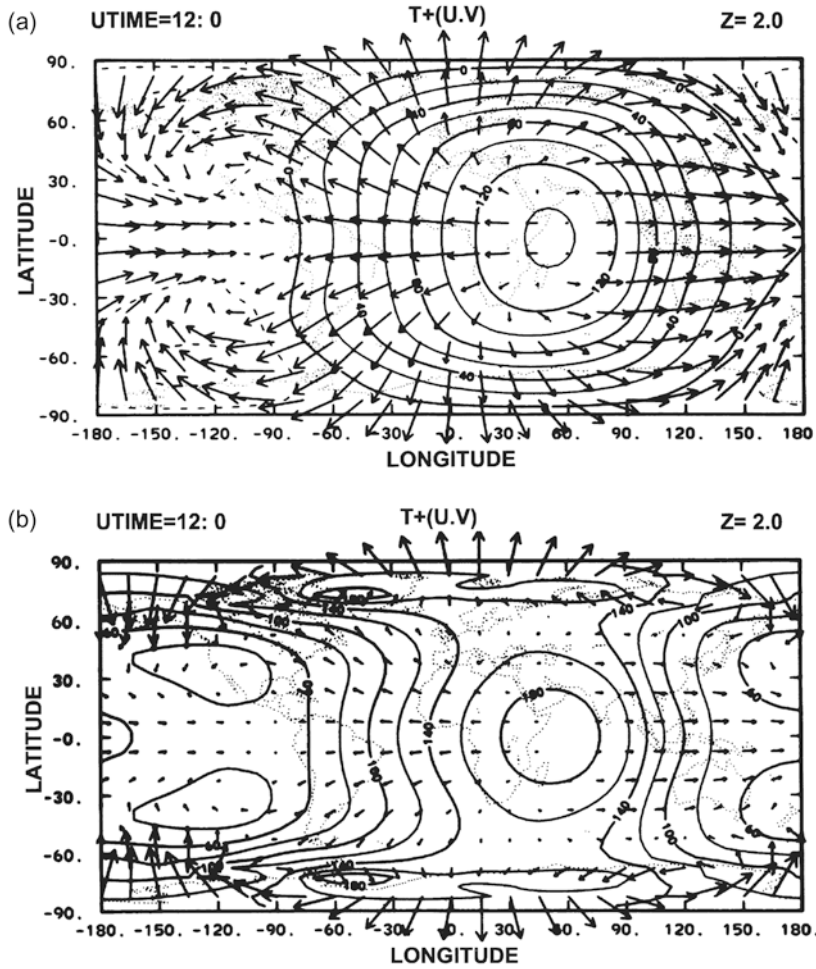


Figure 2.30. (a) Calculated global circulation and temperatures along a constant pressure surface corresponding to about 300 km for the case of solar heating as the only driving force mechanism. The contours describe the temperature perturbation in K and the arrows give wind directions. The length of the arrows gives the wind speed with a maximum of 100 m/s. (b) The same as in (a) except that a magnetospheric convection source with cross-tail potential corresponding to 60 kV is included. The wind speed represented by the length of arrows now corresponds to 336 m/s. (From Dickinson *et al.*, 1984.)

When, however, a cross-polar cap potential of about 60 kV is applied (Figure 2.30(b)), the wind field as well as the temperature distribution changes character; the latter is related to local heating due to enhanced collisions between the neutrals and ions (Joule heating). The velocities across the polar region are strongly enhanced (notice the change in scales by a factor of more than 3 on the velocity arrows). At night-time the high-latitude velocities are more strongly equatorward related to the surge, as noticed already in relation to Figure 2.28. In

the daytime, however, at middle latitudes the velocities are strongly reduced and the zonal component is reversed from eastward to westward. These effects are due to enhanced Joule heating at high latitudes caused by strong currents in the auroral zone.

That electromagnetic disturbances at high latitudes can have a significant effect at thermospheric heights is illustrated more clearly in Figure 2.31. In panel (a) the meridional circulation under quiet conditions is mainly forced by solar heating which produces a large cell with warm air rising at the equator, flowing towards the poles, and sinking down again to lower heights. Above the polar regions ($>70^\circ$ latitude), however, a small cell with the opposite flow direction is set up above 300 km altitude due to heat influx from the magnetosphere outside the plasmapause. In panel (b), however, in which a large magnetic storm takes place, the contrary cell expands in altitude (down to 120 km) as well as in latitude (down to the equator around 200 km of altitude). Air rich in molecular constituents rises in the thermosphere above the high-latitude regions and is being brought at high altitude toward the equator where the air descends to about 150 km. Ions with high velocities are observed flowing outward from the thermosphere along magnetic field lines during such stormy events.

2.17 E-REGION WINDS

By now moving down in altitude to study E-region winds, it is again useful to compare the different terms in the equation of motion (2.26). In Figure 2.32 the outcome of such a comparison for the same high-latitude station as shown in Figure 2.27 is presented.

Contrary to the F-region situation where the pressure and ion drag term appeared to be dominant forcing mechanisms to the neutral wind, it is the pressure and Coriolis term that rule over the E-region neutral air motion. A global model calculation of the E-region neutral wind at about 120 km of altitude is illustrated in Figure 2.33(a) against isothermal contours. It is noticed that in contrast to F-region winds which are mainly directed perpendicularly to the isotherms, the E-region neutral air motion more closely follows these contour lines. A geostrophic component is therefore present in E-region neutral winds. E-region wind speeds are also only a third of F-region speeds on average. A 60 kV cross-polar cap potential is enforced on the ion motion for the wind data shown in Figure 2.33(b). The high-latitude winds are now predominantly eastward and the cross-polar cap motion is reduced. At low latitudes, however, there do not appear to be marked changes in the E-region wind field.

2.18 OBSERVATIONS OF E-REGION NEUTRAL WINDS

Due to the strong coupling between neutrals and ions in the E-region, observations of the ion motion can sometimes be used as a tracer for the neutral motion.

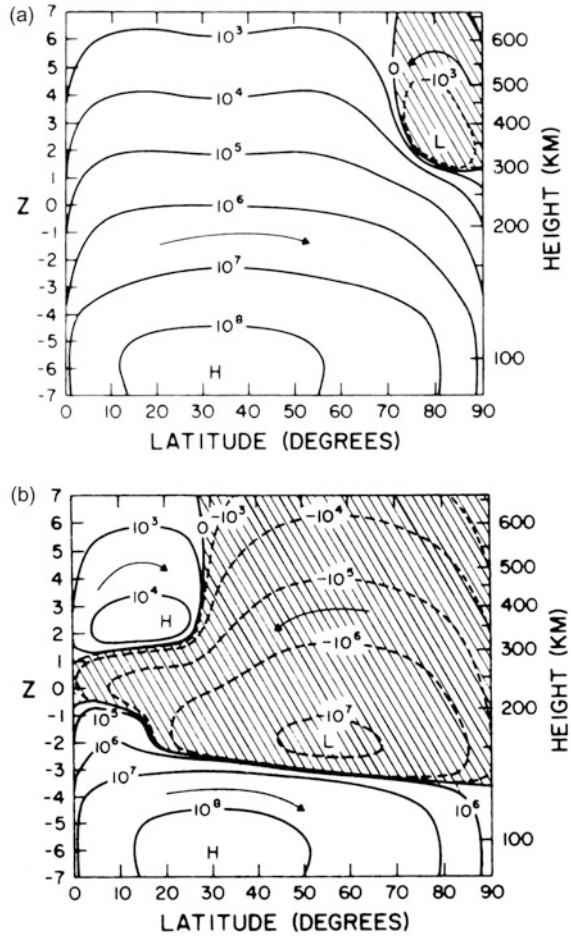


Figure 2.31. Calculated contours of a mass flow stream function. The numbers on the contours are in g/s. The difference between the contour values gives the global flux mass between the two contours. (a) For solar heating as the only driving source. (b) For solar heating, Joule heating and momentum source also included. (From Dickinson *et al.*, 1975.)

Let us now consider the equation of motion for the ions (2.33). In the steady state when $dv_i/dt = 0$, we can solve for neutral velocity in a simple manner:

$$\mathbf{u}_n = \mathbf{v}_i - \frac{q}{m_i \nu_{i,n}} \mathbf{v}_i \times \mathbf{B} - \frac{q}{m_i \nu_{i,n}} \mathbf{E} \tag{2.34}$$

At high latitude where the magnetic field is almost vertical, this equation applies to the horizontal component of $\mathbf{u}_n$ while the vertical component is assumed equal to the ion motion along the magnetic field line. We notice that when ν_{in} is very large, the neutral velocity is equal to the ion velocity. Observing the ion motion by some radio or radar technique therefore allows for deriving the neutral motion

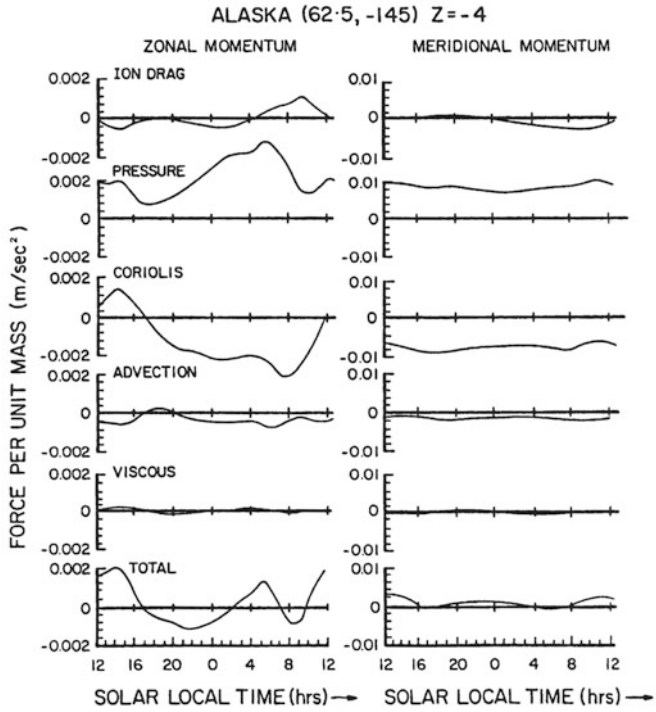


Figure 2.32. Similar to the data shown in Figure 2.27 but applied at 120 km approximately. (From Killeen and Roble, 1986.)

in a rather direct manner. For smaller values of ν_{in} the complete coupling between ions and neutrals as given by (2.34) has to be accounted for in order to derive $\mathbf{u}_n$ from measurements of $\mathbf{v}_i$. By means of an incoherent scatter radar such as EISCAT close to Tromsø, Norway (70° N) one can obtain almost simultaneous observations of E- and F-region velocities. In the F-region (300 km) where the collision term will be negligible to the ion motion compared with the E-field and Lorentz force term, the electric field can be derived for a steady-state solution (1.29):

$$\mathbf{E} = -\mathbf{v}_i^F \times \mathbf{B}$$

where $\mathbf{v}_i^F$ is the F-region ion velocity observed simultaneously with the E-region ion velocity. The neutral wind in the E-region can then be derived with an appropriate choice of collision frequency model.

Figure 2.34 illustrates neutral wind (x, y, z) components in the geographical frame of reference as a function of time at several E-region heights for quiet days ($Ap \leq 16$) in the four seasons. From such time series the mean velocity components as well as the average 24-, 12-, 8-, and 6-hour tidal components can be derived for each of the six heights. This has been carried out for all four seasons and the results are presented in Figure 2.35. The most outstanding feature of these

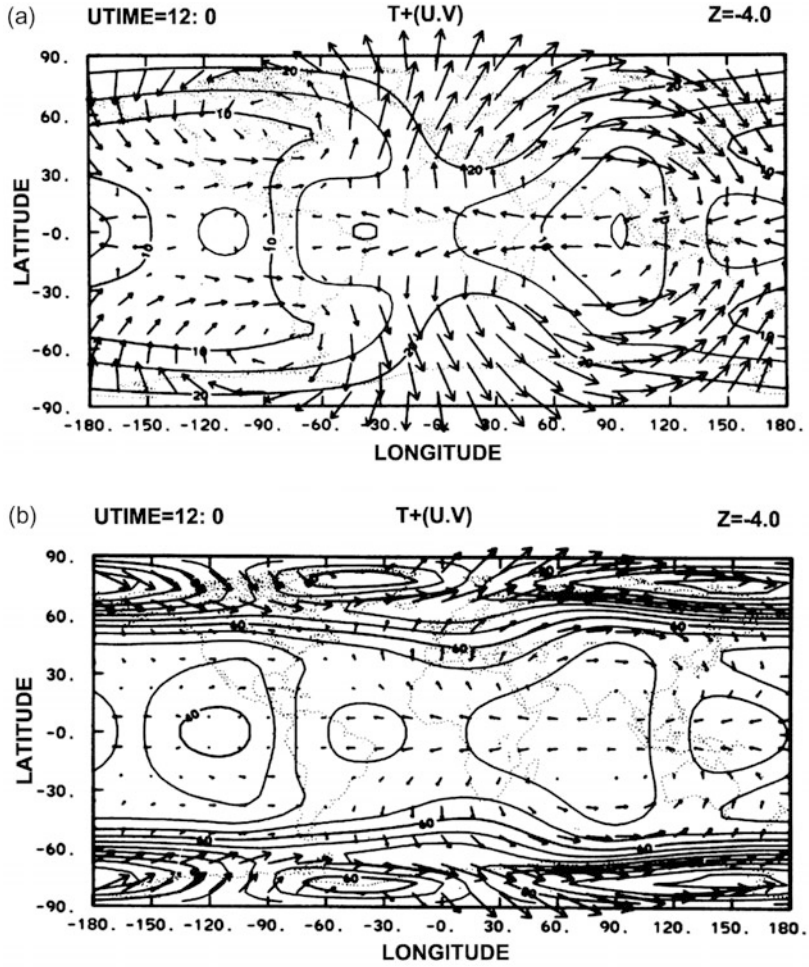


Figure 2.33. Same as for Figure 2.30 except that the constant pressure level corresponds to close to 120 km. (a) The maximum wind speed corresponds to 32 m/s. (b) The maximum wind speed corresponds to 79 m/s. (From Dickinson *et al.*, 1984.)

components is the strong mean eastward wind between 90 and 120 km having a maximum at about 110 km of 60 m/s in summer and 20 m/s in winter. The very low mean northward and vertical wind components in the same height region are also noticeable; these components are relatively less influenced by change of season.

The eastward component of the diurnal tide increases gradually by height and reaches an amplitude close to 100 m/s above 120 km. The northward component of this tide is fairly independent of height. For the other horizontal tidal components no systematic height dependence is found. In the vertical component, however, all tidal components except the 6-hour tide have a minimum amplitude at the height

region of the strong mean eastward wind. In Figure 2.36 wind components presented as a function of time for the six E-region heights for quiet autumn days in Figure 2.34 are shown as horizontal clock dial plots. We notice the outstanding daytime poleward wind at the four lower heights 96–117 km. Clearly, there are large variations by height in the neutral wind in the auroral E-region.

It is generally believed, in spite of the long time constants involved (as shown in Figure 2.23), that the E-region neutral velocities will be modified during strong auroral disturbances at the higher latitudes.

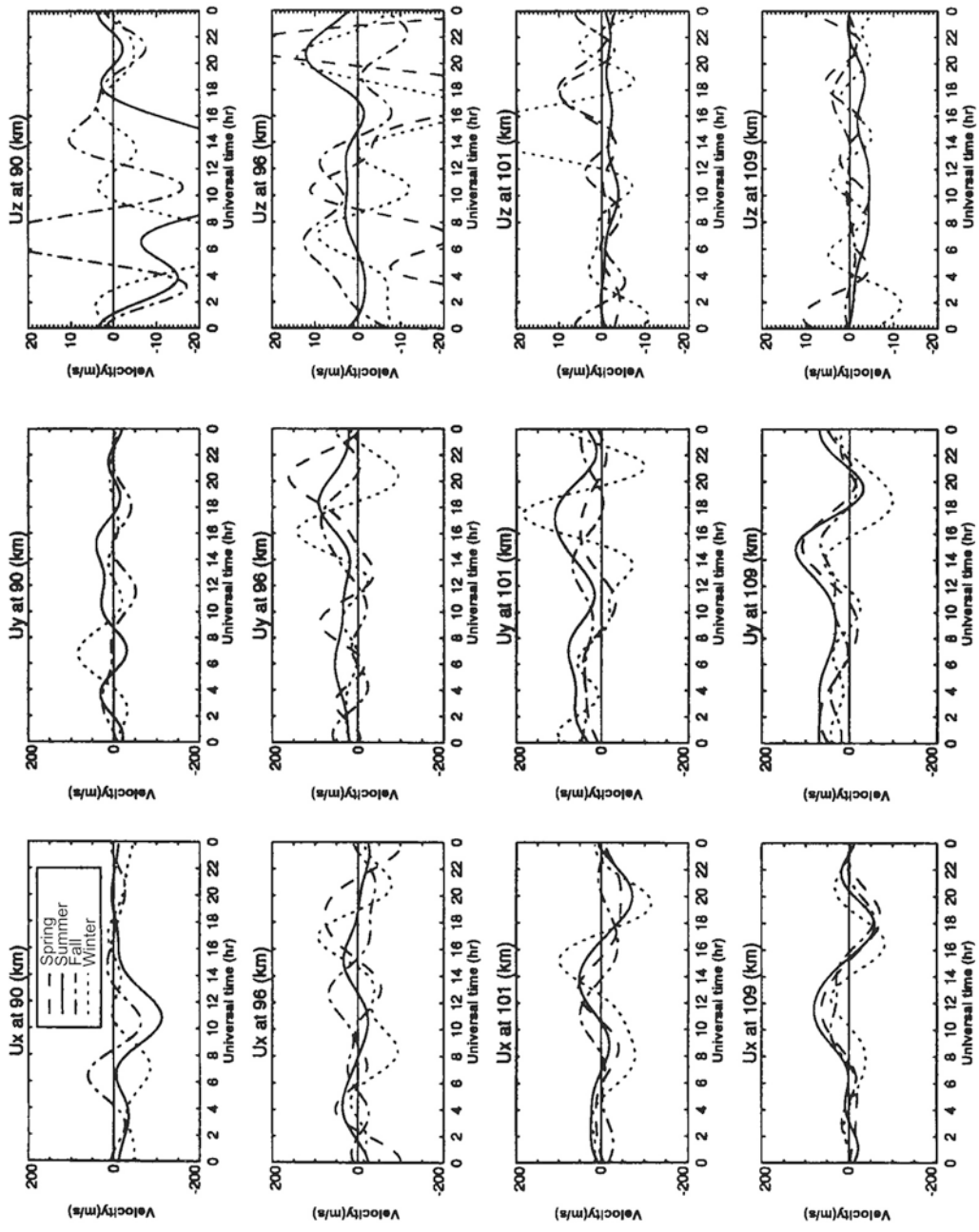
A similar averaging process of the average neutral wind at different E-region heights has been performed on data obtained by EISCAT under disturbed conditions ($Ap > 16$). The mean velocity components as well as the tidal components are shown as a function of height in Figure 2.37. Data obtained from 11 disturbed days are compared with corresponding data from 24 quiet days ($Ap \leq 16$). We notice that the strong eastward mean velocity is rather unaffected by the condition of disturbance, and is more dependent on season (as shown in Figure 2.35). An outstanding feature in the auroral E-region is therefore an eastward neutral jet stream or superrotation. The most outstanding difference between quiet and disturbed days appears to be the enhancement of the eastward diurnal tide above 110 km altitude.

By the EISCAT Svalbard Radar (ESR) it is also possible to derive the E-region neutral wind at polar latitudes. A comparison of the height profiles of the mean wind and diurnal and semidiurnal tides is shown in Figure 2.38 for observations in Tromsø and Longyearbyen, Svalbard (78.1° N, 15.3° E). It is noticed that the behavior of the different components is rather similar except that the mean eastward wind in Tromsø is weaker than in the polar cap latitude.

To study the importance of the electric field on the neutral wind at different heights, Figure 2.39 shows a correlation analysis of the horizontal neutral wind components and the electric field at different altitudes as derived from Tromsø and Svalbard. It is noticed that the correlation coefficient is very low (~ 0) below 110 km for all components and that the neutral wind appears to be less influenced by the E-field above 110 km in Longyearbyen than in Tromsø. The ion drag appears to be less efficient at higher latitudes. The eastward component of the neutral wind in Tromsø appears almost in antiphase with the electric field above 120 km while the northward component has a maximum correlation for both places at about 115 km.

2.19 VERTICAL MOTION

It is of special interest to study vertical winds, although these are usually very small compared with horizontal winds and also difficult to measure. They are, however, important as they play a significant role in the vertical transport of momentum and kinetic energy, and also of gravity potential energy. Furthermore, vertical winds control to a large extent the vertical distribution of the different constituents as we have seen in relation to diffusion. Let us assume that there are


Ux at 90 (km)
Uy at 90 (km)
Uz at 90 (km)
Ux at 96 (km)
Uy at 96 (km)
Uz at 96 (km)
Ux at 101 (km)
Uy at 101 (km)
Uz at 101 (km)
Ux at 109 (km)
Uy at 109 (km)
Uz at 109 (km)
Spring
Summer
Fall
Winter
Velocity(m/s)
Universal time (hr)

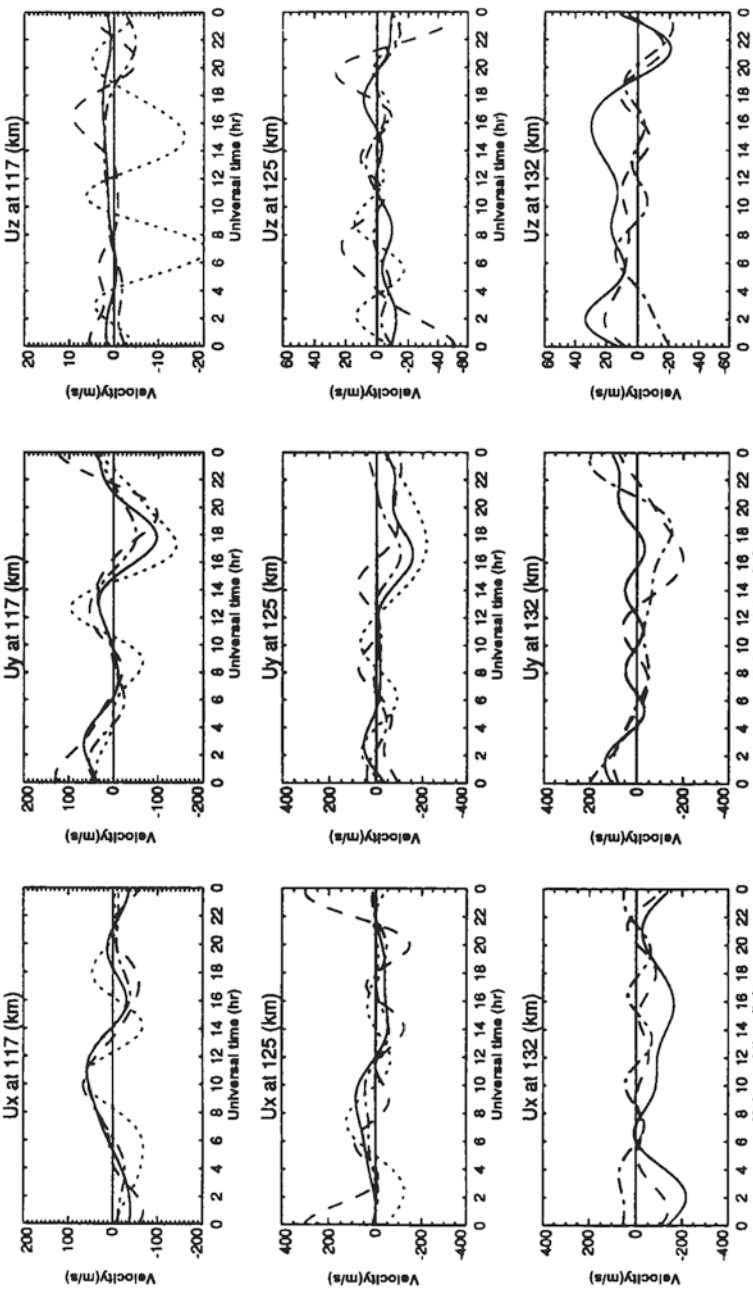
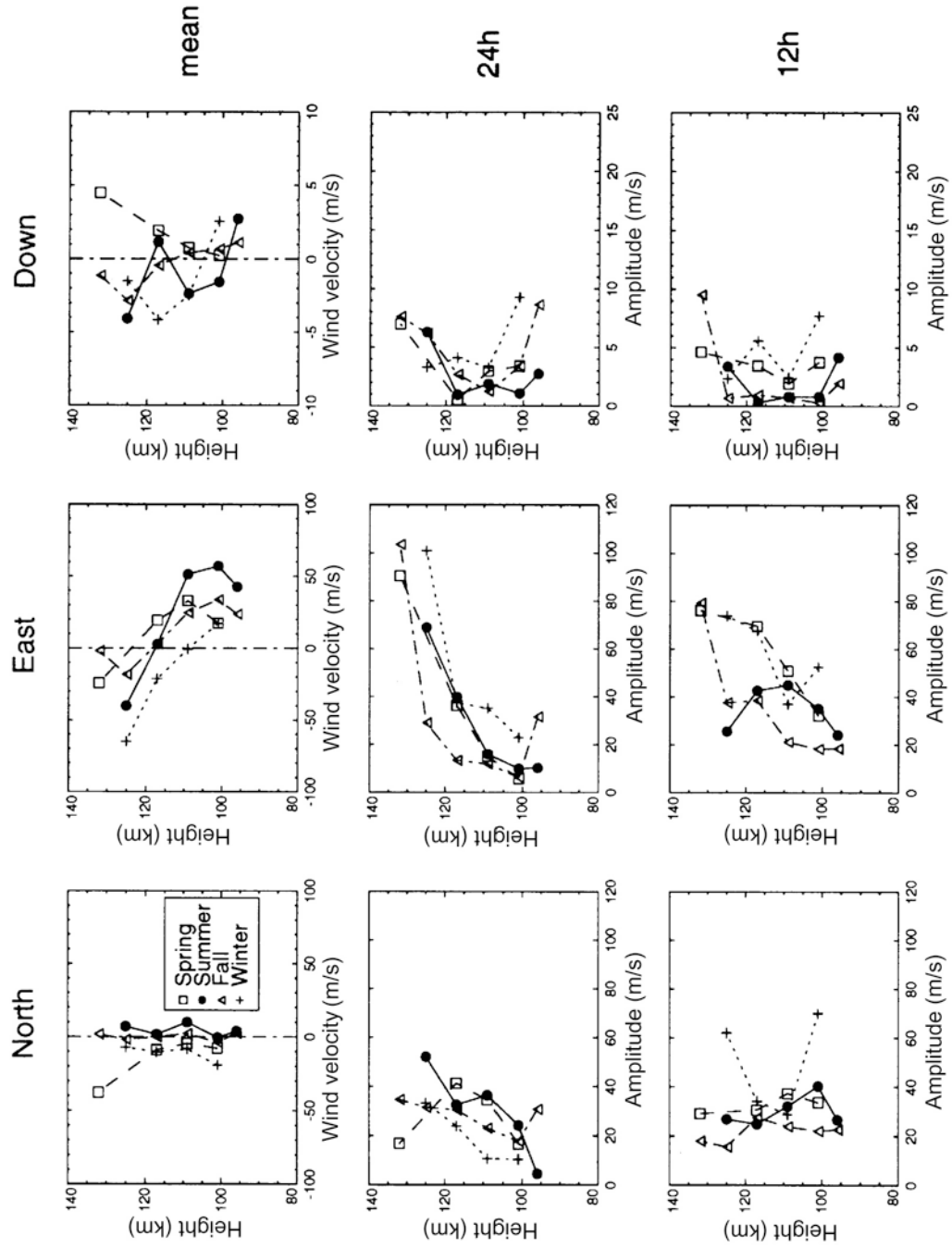


Figure 2.34. The average neutral wind components derived by the EISCAT incoherent scatter radar at Tromsø (69.66° N, 18.94° E) at seven E-region heights for quiet days ($Ap \leq 16$) in the four seasons. These average wind components are derived by adding the mean wind components together with the 24-, 12-, 8-, and 6-hour tidal components with appropriate phases. (Courtesy Nozawa, 1994.)


North
East
Down
mean
24h
12h
Spring
Summer
Fall
Winter
Height (km)
Wind velocity (m/s)
Amplitude (m/s)

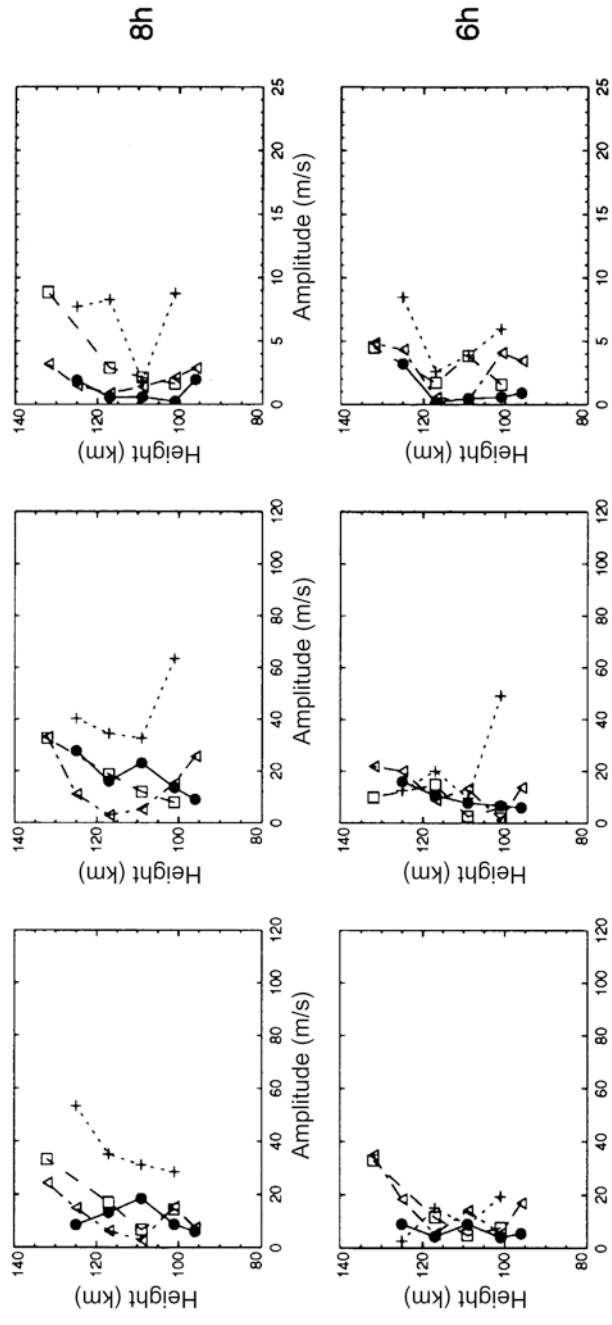


Figure 2.35. (Upper row of panels) The mean components as a function of height for the four seasons. (Four lower rows of panels) The 24-, 12-, 8-, and 6-hour tidal components as a function of height for the four seasons. (From Brekke *et al.*, 1994.)

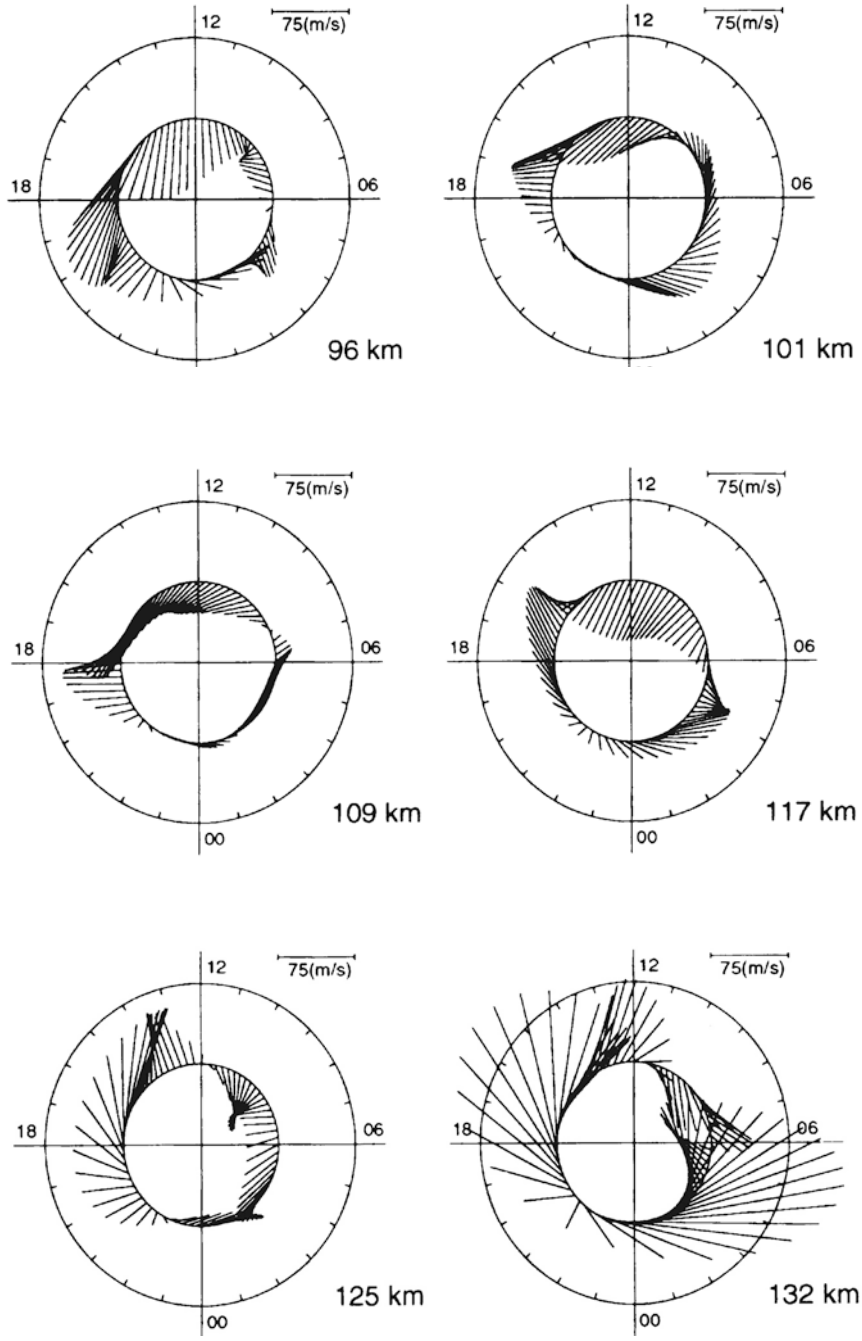


Figure 2.36. The average horizontal neutral wind velocity vectors at six different E-region heights for the average quiet autumn day. These vectors are obtained by adding together the appropriate north and east components in Figure 2.34. (From Brekke *et al.*, 1994.)

no horizontal gradients in density but that the atmosphere is vertically stratified ($\partial/\partial x = \partial/\partial y = 0$).

For a quasi-steady state ($\partial/\partial t = 0$) we derive from the continuity equation

$$\left.\begin{aligned} \frac{d\rho}{dt} + \rho\nabla\cdot\mathbf{u} &= 0 \\ \frac{\partial\rho}{\partial t} + (\mathbf{u}\cdot\nabla)\rho + \rho\nabla\cdot\mathbf{u} &= 0 \\ (\mathbf{u}\cdot\nabla)\rho + \rho\nabla\cdot\mathbf{u} &= 0 \end{aligned}\right\} \tag{2.35}$$

and

$$w\frac{\partial\rho}{\partial z} + \rho\nabla\cdot\mathbf{u} = 0$$

where we have used $\mathbf{u} = (u, v, w)$. By multiplying by p and dividing by ρ we get

$$p\nabla\cdot\mathbf{u} = -\frac{p}{\rho}\frac{\partial\rho}{\partial z}w \tag{2.36}$$

From the equation of state for an ideal gas

$$p = \rho RT \tag{2.37}$$

where $R = 8.3\,\mathrm{J/mol\,K}$ is the individual gas constant, we form

$$\frac{\partial p}{\partial z} = R\left(\rho\frac{\partial T}{\partial z} + T\frac{\partial\rho}{\partial z}\right)$$

and solving for $\partial\rho/\partial z$ gives:

$$\frac{\partial\rho}{\partial z} = \frac{1}{T}\left(\frac{1}{R}\frac{\partial p}{\partial z} - \rho\frac{\partial T}{\partial z}\right)$$

By inserting this into (2.36) we have

$$p\nabla\cdot\mathbf{u} = -\frac{p}{\rho}\frac{1}{T}\left(\frac{1}{R}\frac{\partial p}{\partial z} - \rho\frac{\partial T}{\partial z}\right)w = -\left(\frac{\partial p}{\partial z} - \frac{p}{T}\frac{\partial T}{\partial z}\right)w$$

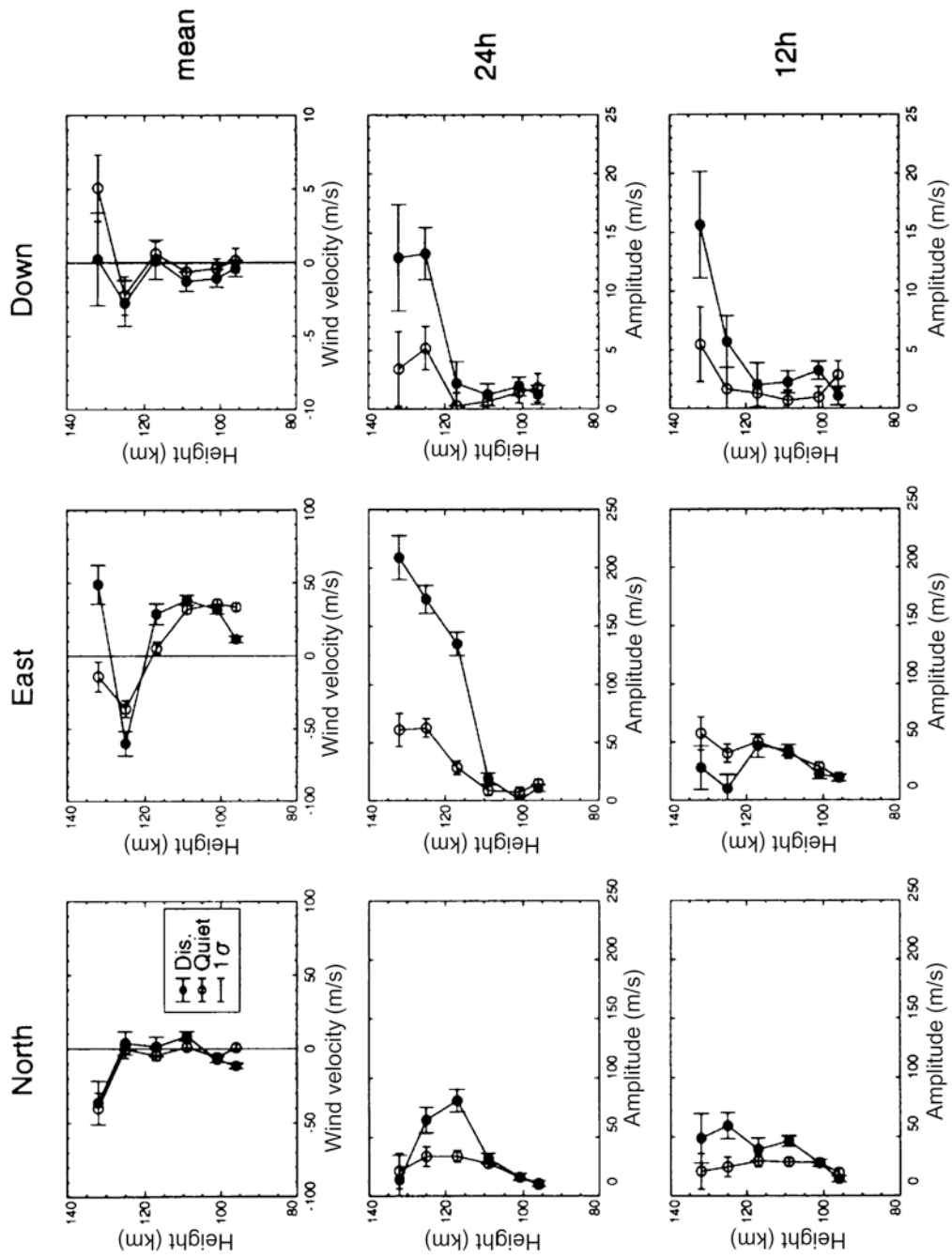
The vertical velocity is now given by

$$w = -\frac{p\nabla\cdot\mathbf{u}}{\dfrac{\partial p}{\partial z} - \dfrac{p}{T}\dfrac{\partial T}{\partial z}}$$

By introducing the partial derivative $\partial T/\partial p$ and the barometric law (2.5) we have:

$$w = \frac{p\nabla\cdot\mathbf{u}}{g\rho\left(1 - \dfrac{p}{T}\dfrac{\partial T}{\partial p}\right)} \tag{2.38}$$

The vertical velocity is now expressed by the divergence of $\mathbf{u}$ and the variation in temperature by pressure. By introducing the scale height from (2.6) into (2.38), we


North
East
Down
mean
24h
12h
Dis.
Quiet
1σ
Height (km)
Wind velocity (m/s)
Amplitude (m/s)

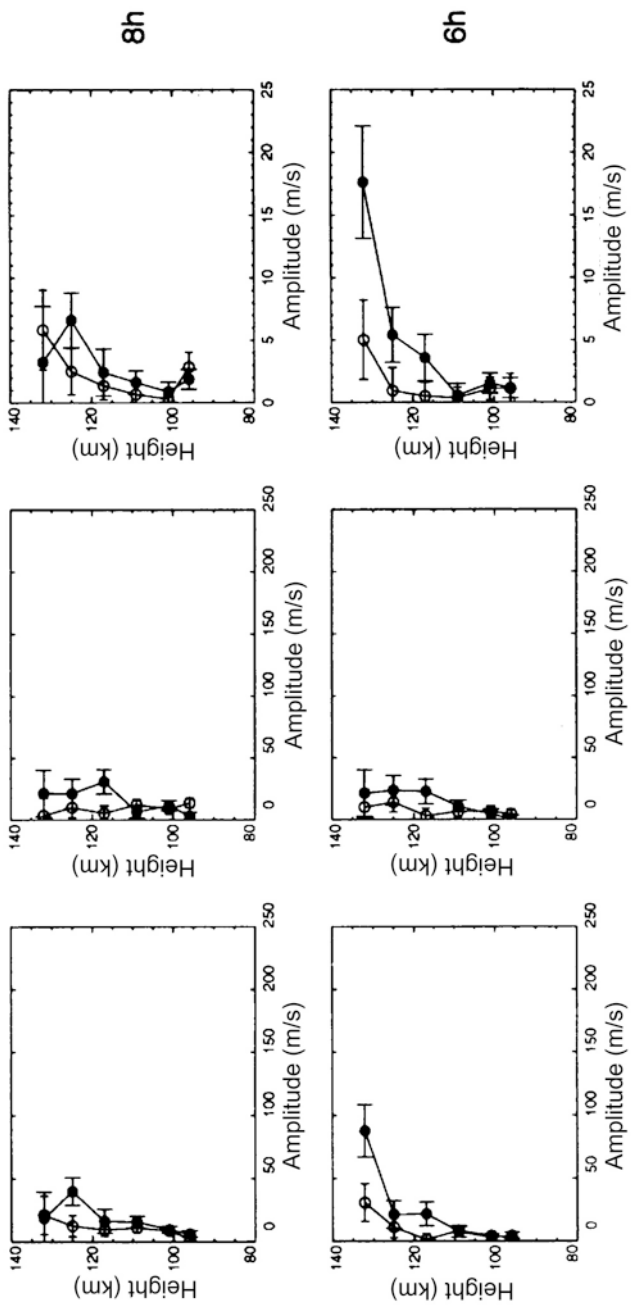
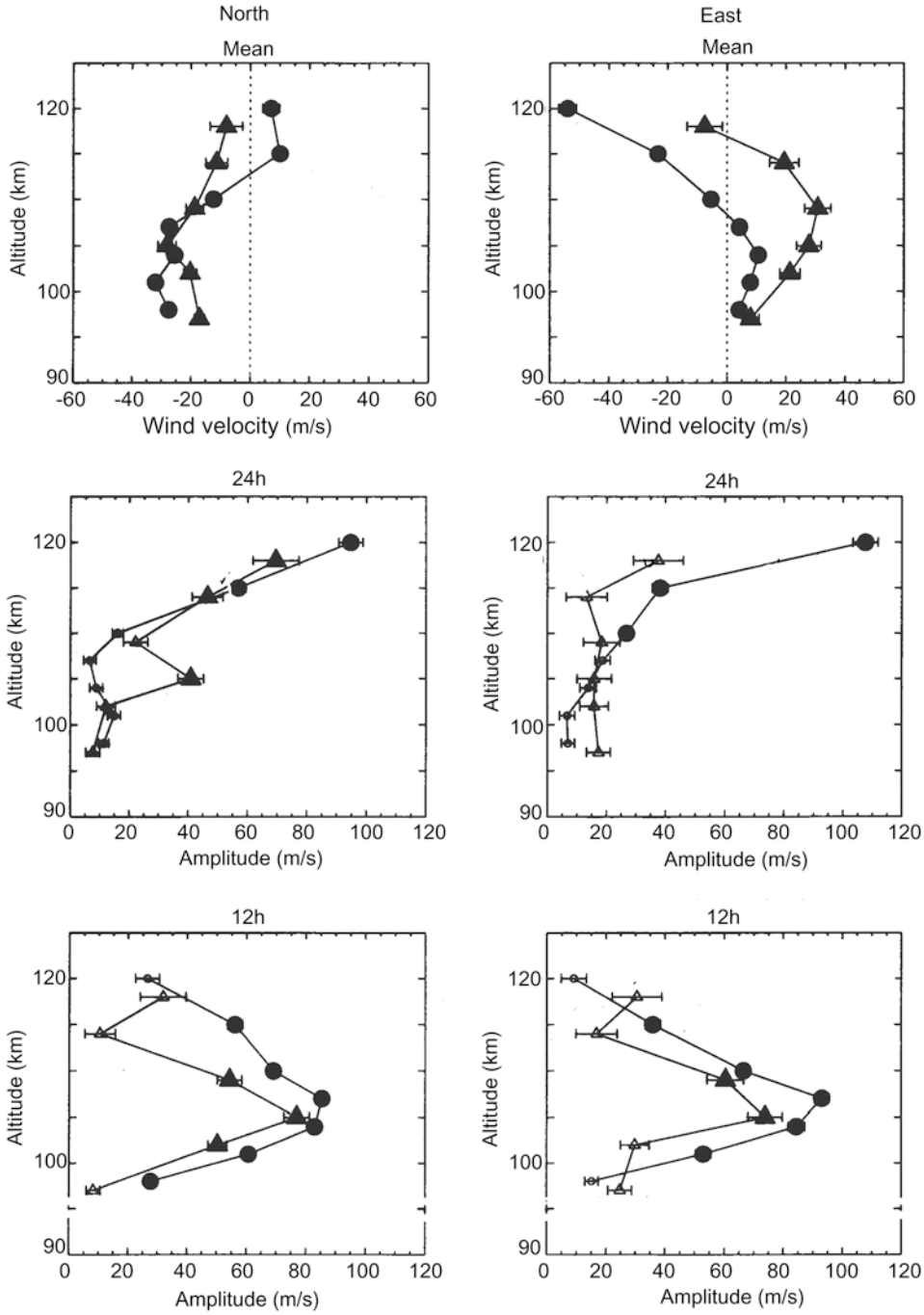


Figure 2.37. The variation with height of the northward, eastward, and downward components of the mean wind and the amplitudes of the 24-, 12-, 8-, and 6-hour tidal components. The different disturbance levels are indicated by quiet ($Ap \leq 16$) and disturbed ($Ap > 16$) conditions. The standard deviations are shown by vertical bars. (From Nozawa and Brekke, 1995.)

(a)


North
Mean
East
Mean
Altitude (km)
120
100
90
-60 -40 -20 0 20 40 60
Wind velocity (m/s)
24h
12h
0 20 40 60 80 100 120
Amplitude (m/s)

(b)

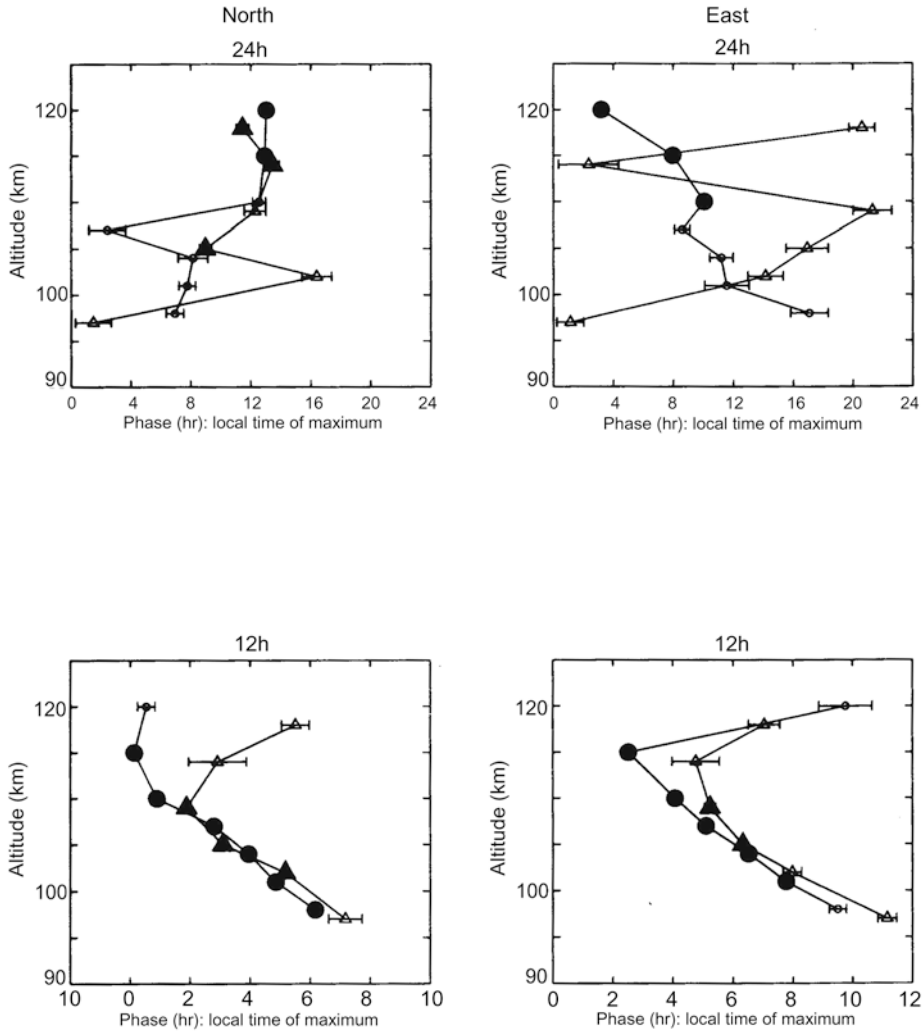


Figure 2.38. (a) Altitude profiles of mean wind (top), diurnal tidal amplitude (middle), and semidiurnal amplitude (bottom) observed at Tromsø (circle) and Longyearbyen (triangle) are presented for meridional (left) and zonal (right) components. For tidal components, larger filled symbols denote data values whose significance level is greater than 90%, smaller filled symbols whose significance level is between 90 and 50% and small open symbols denote data values whose significance level is less than 50%. Horizontal lines associated with each symbol denote error bars (1 standard deviation). (b) Corresponding phase values (in local time of maximum) are shown for diurnal (top) and semidiurnal (bottom) components for meridional (left) and zonal (right) components. (From Nozawa *et al.*, 2005.)

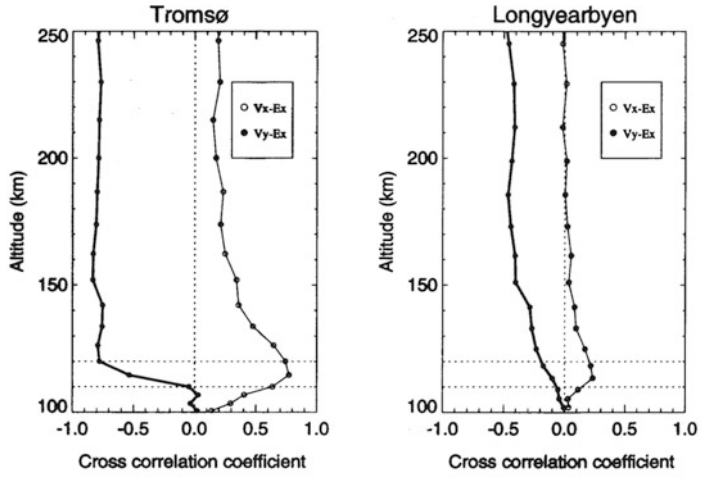


Figure 2.39. Altitude profiles of correlation coefficients between the E-region neutral wind components and the electric field as derived from Tromsø and Longyearbyen with incoherent scatter observations. (Nozawa, private communication.)

have:

$$w = \frac{H \nabla \cdot \mathbf{u}}{1 - \dfrac{p}{T}\dfrac{\partial T}{\partial p}} \tag{2.39}$$

For an isothermal atmosphere where $\partial T = 0$ we find

$$w = H \frac{\partial w}{\partial z}$$

and

$$w = w_0 \exp\left(\frac{z - z_0}{H}\right)$$

when $w = w_0$ at a reference height z_0. Vertical velocity therefore increases with height for an expansion in an isothermal atmosphere. This comes about since density decreases by height, and in order to keep the energy flux constant the velocity must increase.

For an adiabatic expansion we get from (2.15)

$$\frac{p}{T}\frac{\partial T}{\partial p} = \frac{\gamma - 1}{\gamma}$$

and find from (2.39)

$$w = \gamma H \frac{\partial w}{\partial z}$$

and

$$w = w_0^* \exp\left(\frac{z - z_0}{\gamma H}\right)$$

Since $\gamma > 1$, the vertical velocity will increase more slowly with height in the adiabatic situation than for an expansion in an isothermal atmosphere.

The term $p\nabla \cdot \mathbf{u}$ represents a rate of mechanical work against the pressure force due to expansion or compression. We now want to relate this to the change of energy in the system.

The first law of thermodynamics per unit mass can be expressed by:

$$dq = d\varepsilon + \frac{p}{m} dV \tag{2.40}$$

where q is the heat, $d\varepsilon = c_v\, dT$ is the change in the internal energy, and c_v $(= 712\,\mathrm{J\,K^{-1}\,kg^{-1}})$ is the heat capacity per unit mass at constant volume.

From the ideal gas law we have

$$p = nkT = \rho RT$$

and since

$$\rho = m/V$$

$$dV = -m\frac{d\rho}{\rho^2}$$

Inserting this into (2.40) and forming

$$\dot{q} = \frac{dq}{dt}$$

$$\dot{q} = \dot{\varepsilon} - \frac{RT}{\rho}\frac{d\rho}{dt}$$

Inserting from (2.35) for $(1/\rho)(d\rho/dt)$ we have

$$\dot{q} = \dot{\varepsilon} + RT\nabla \cdot \mathbf{u}$$

And finally solving for $\nabla \cdot u$

$$\nabla \cdot u = \frac{1}{RT}(\dot{q} - \dot{\varepsilon})$$

and multiplying by p we finally have

$$p\nabla \cdot \mathbf{u} = \frac{p}{RT}(\dot{q} - \dot{\varepsilon}) = \rho(\dot{q} - \dot{\varepsilon})$$

The left-hand side represents the rate at which work is done by the system if the right-hand side is positive, and on the system if the right-hand side is negative. The vertical velocity can now be derived from (2.38):

$$w = \frac{\dot{q} - \dot{\varepsilon}}{g\left(1 - \dfrac{p}{T}\dfrac{\partial T}{\partial p}\right)}$$

For an adiabatic process where $\dot{q} = 0$ we now have:

$$w = -\frac{\gamma \dot{\varepsilon}}{g} = -\frac{\gamma c_v}{g}\frac{dT}{dt}$$

If the internal temperature increases with time, then work must be done on the system by the surroundings and the air must contract downward. In the opposite sense, if the temperature decreases, the work is done by the system and the atmosphere expands upward. Since we have only allowed for vertical motion, this is equivalent to considering the motion of a column of air in a cylinder with a fixed base.

For external heat allowed to exchange with the system the situation can be far more dramatic, especially if $\dot{q}$ is large. This occurs quite frequently within the auroral region of the upper atmosphere and then strong vertical motion can be expected.

2.20 EXERCISES

1. Derive the total mass of an isothermal atmosphere and compare that with the mass of the Earth.
2. Let the normalized velocity distribution in a gas in thermal equilibrium be given by:

$$f(v) = 4\pi\left(\frac{m}{2\pi kT}\right)^{3/2} v^2 \exp\left(-\frac{mv^2}{2kT}\right)$$

   (a) Derive the average speed $\bar{v}$.

   (b) Show that the most probable speed

$$v_{\text{m.p.}} = \left(\frac{2kT}{m}\right)^{1/2}$$

   (c) Show that the root mean square speed

$$v_{\text{r.m.s.}} = \left(\frac{3kT}{m}\right)^{1/2}$$

3. Choose a place in the atmosphere at latitude θ. In a Cartesian coordinate system x is positive eastward, y northward, and z in the vertical direction.

   (a) Find the horizontal geostrophic wind components u and v in this coordinate system.

   Let the pressure gradient around the equator look like that in Figure 2.40.

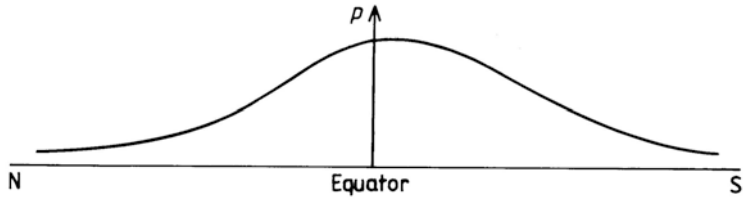


Figure 2.40

(b) Give the zonal wind directions in the northern and southern hemisphere.

(c) Does it make sense to talk about a geostrophic wind at the equator?

Let the pressure gradient around the equator look like that in Figure 2.41.

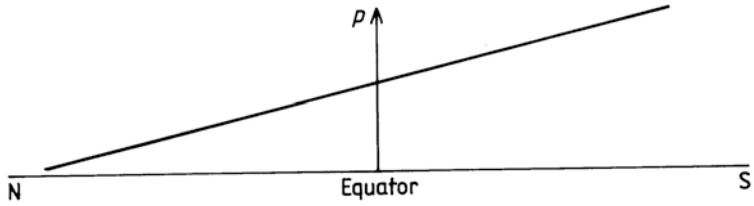


Figure 2.41

(d) What do the zonal wind directions look like in both hemispheres in this case?

3

The Earth's magnetic field and magnetosphere

3.1 AN HISTORICAL INTRODUCTION

It has been known for hundreds of years that the Earth possesses a magnetic field. The modern science of geomagnetism, however, can probably be dated back to the English physicist William Gilbert who in 1600 published his book *De Magnete* in which he stated that *Magnus magnetis ipse est globus terrestris* ("The Earth itself is a great magnet"). Not much happened in the field of geomagnetism from when his book was printed and until about 100 years later. Then the English scientist Halley compared observations of the magnetic field in London as made by Gilbert, a few others, and himself through a period of 100 years or more. He found that the direction of the magnetic needle had changed from 10° east in Gilbert's time to almost 10° west around 1700 (Figure 3.1). Halley immediately came up with an explanation for this and maintained that the Earth itself consists of two separate magnetic systems, one connected to a solid sphere in the internal regions of the Earth, and another one connected to a concentric spherical shell or the Earth's crust. If these two were rotating at different rotation speeds, a drift in the magnetic system could according to Halley be expected.

Halley and many others were fairly convinced that by more accurate mapping of the geomagnetic field, especially at sea, it should be possible to obtain a remedy by which one could safely sail irrespective of cloudiness, in contrast to navigation by the Sun and the stars. He then convinced the British government to equip several naval expeditions led by him to observe magnetic elements as often and as accurately as possible. He also encouraged captains in the merchant navy to do the same. From these observations Halley was able to construct the first map showing the deviation of the magnetic needle from true north at many positions in the Atlantic and on the mainland and drew the first magnetic chart with the so-called isogonic lines (Figure 3.2). From then on the study of the geomagnetic field and field mapping was an important issue in the scientific and marine com-

A. Brekke, *Physics of the Upper Polar Atmosphere*, Springer Praxis Books,
DOI 10.1007/978-3-642-27401-5_3, © Springer-Verlag Berlin Heidelberg 2013

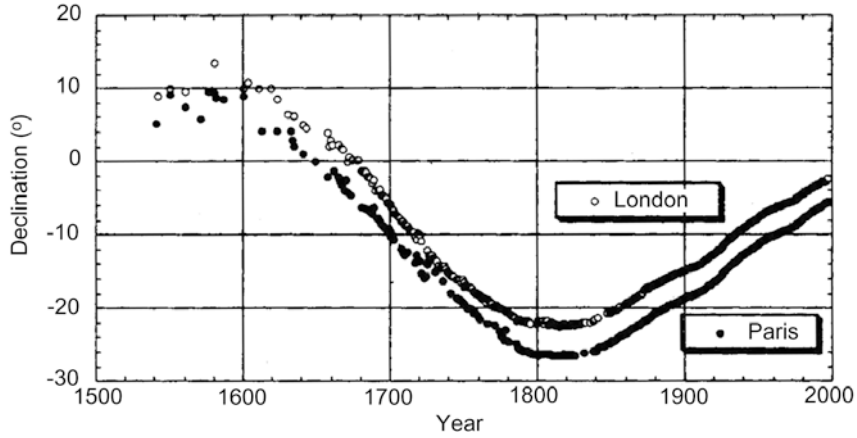


Figure 3.1. A graph showing reconstructed series of direct measurements of declination in Paris and London from the mid-16th century to present. (From Alexandrescu *et al.*, 1997.)

munities. It was not until 1838 that magnetic field observations were organized in such a manner that they could be continuous and referred to accurate absolute standards. Thanks to C. F. Gauss who urged different observatories to cooperate, the geomagnetic field has been measured with great accuracy and continuity ever since. Today about 300 observatories all over the globe monitor the Earth's magnetic field.

3.2 DESCRIPTION OF THE EARTH'S MAGNETIC FIELD

Within a few Earth radii the magnetic field of the Earth is similar to the field one would find if the Earth itself was a magnetized sphere (Figure 3.3) or if a gigantic rod magnet was situated somewhere close to the center of the sphere. This imaginary rod magnet, in order to represent the real field as accurately as possible, would have to be situated about 400 km from the center of the Earth with a magnetic moment (strength) of 7.74×10^{22} A m^2. The magnetic axis would make an angle of 10° with the Earth axis, and the extensions of this axis would penetrate the Earth at two points: the geomagnetic poles (80.0° N, 72.2° W and 80.0° S, 107.8° E) as of 2010. The first point is at the northwestern tip of Greenland and the other is in the Antarctic. Note that these two points are different from the magnetic poles, or the points where the magnetic needle will point vertically. The north magnetic pole is situated at 65.0° N and 132.6° W as of 2010. In contrast to the magnetic poles that have a practical significance, the geomagnetic poles are imaginary and serve as a model to simplify the description of the Earth's magnetic field. With such a simple model, however, the field is correct to within 30% of the real field up to a distance of 4 R_e. Closer to the Earth the error is within 10%. As a curiosity one should remember that, due to our convention of magnetic poles, it is the north

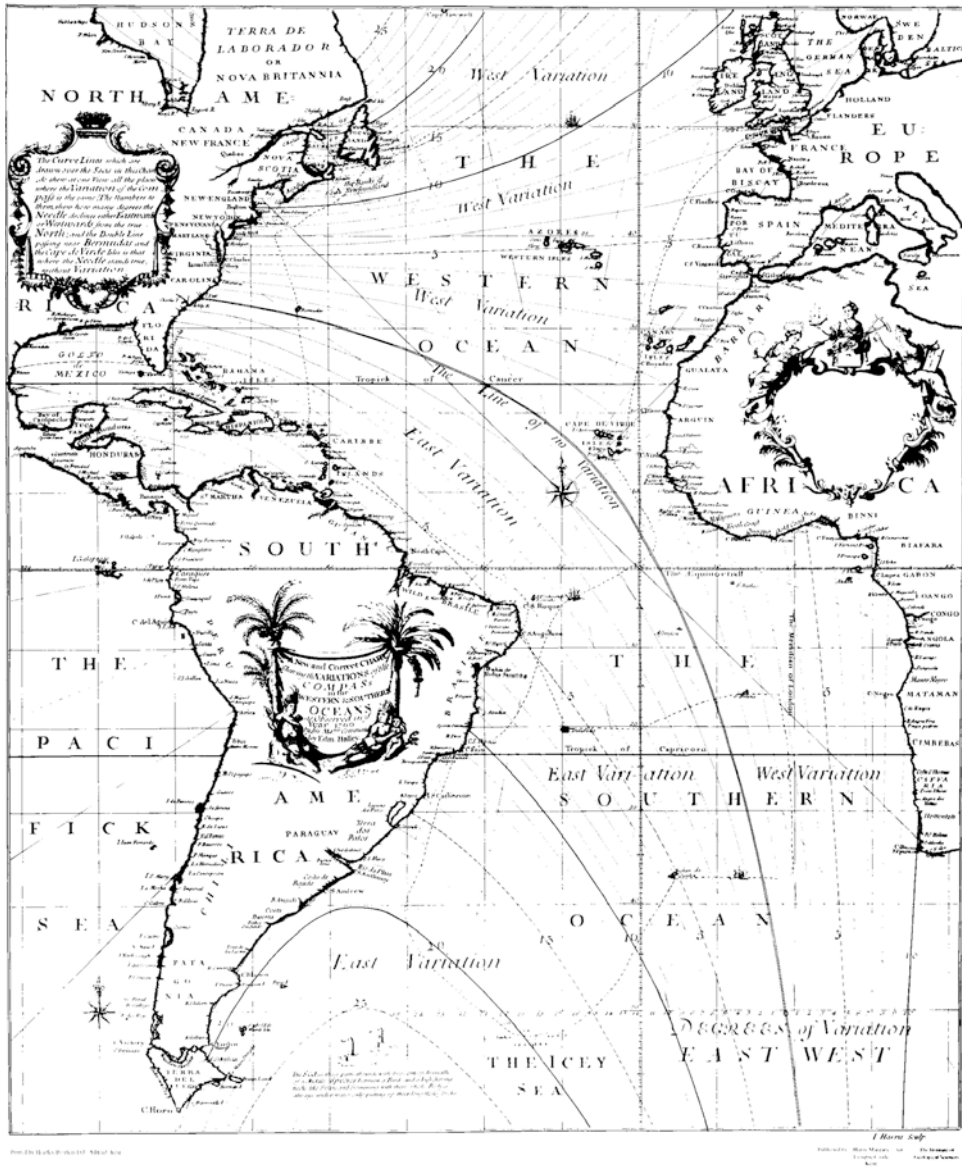


Figure 3.2. Sir Edmund Halley's first magnetic chart from 1702 showing the isogonic lines. (Courtesy The Institute of Geological Sciences, Edinburgh University.)

magnetic pole of the magnetic needle that points toward the north, therefore the magnetic north pole is in fact a south magnetic pole.

Neither the magnetic poles nor the geomagnetic poles remain in fixed positions in time. Figure 3.4 illustrates how these poles have drifted during the last century. While both geomagnetic poles have shifted position by less than

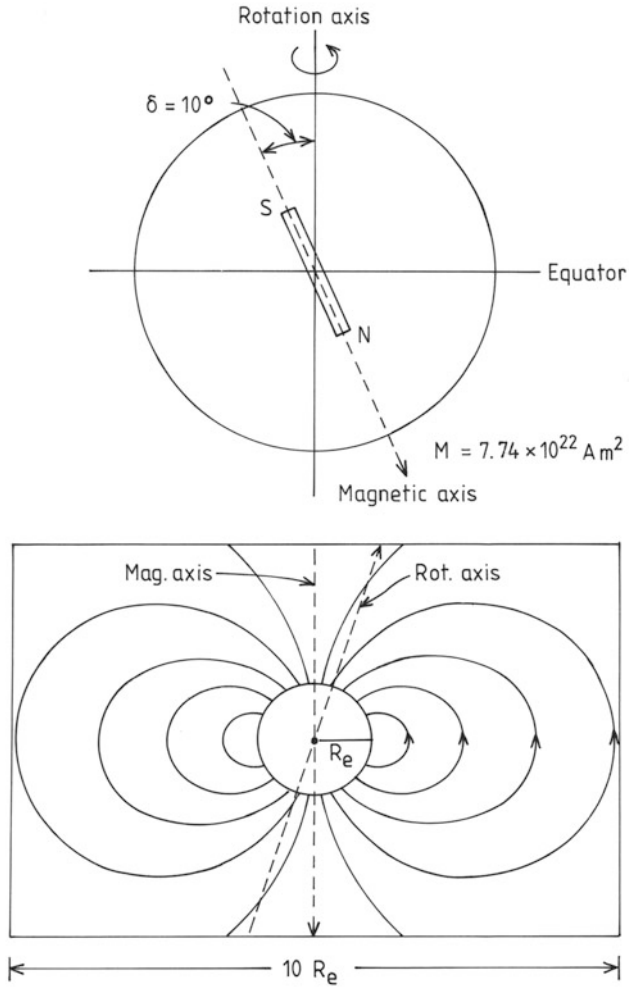


Figure 3.3. A sketch showing (below) the magnetic field line of the Earth as if it was a perfect dipole magnet in vacuum. The rotation axis is tilted with respect to the magnetic dipole axis. In order to represent the Earth's magnetic field in the best possible way by a simple dipole, it had to have a magnetic moment $M_o = 7.74 \times 10^{22}$ A m^2 to be situated close to the center of the Earth and tilted by 10° with respect to the Earth's rotation axis (above).

200 km during this period the north magnetic pole has moved by more than 1,500 km and the south magnetic pole by about 800 km during these 110 years.

The magnetic poles are defined as the points on the globe where the magnetic needles become vertical and are very dependent on local conditions in the Earth's crust, while the geomagnetic poles are hypothetical points corresponding to intersections of the Earth's surface and the axis of a hypothetical bar magnet situated close to the center of the Earth as a model representation of the source of global

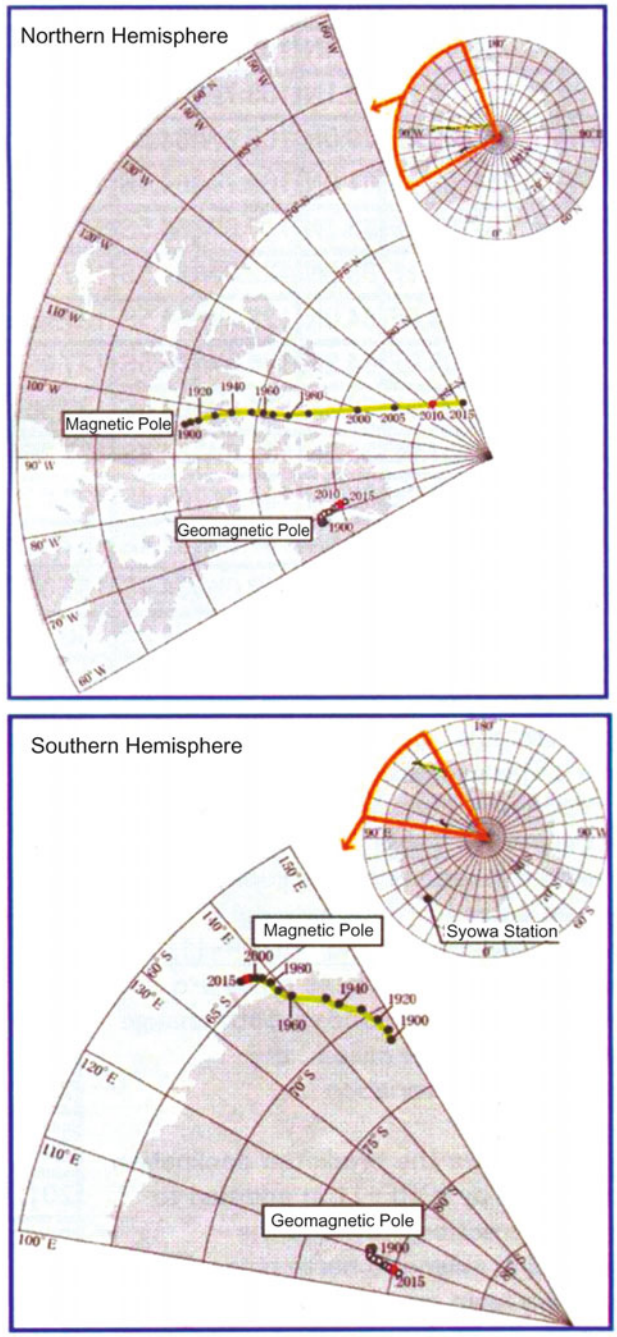


Figure 3.4. Maps of sectors in the Arctic (a) and the Antarctic (b) showing the area where the magnetic and geomagnetic poles have been situated during the last century. (From http://wdc.kugi.kyoto-u.ac.jp/poles/polesexp.html.)

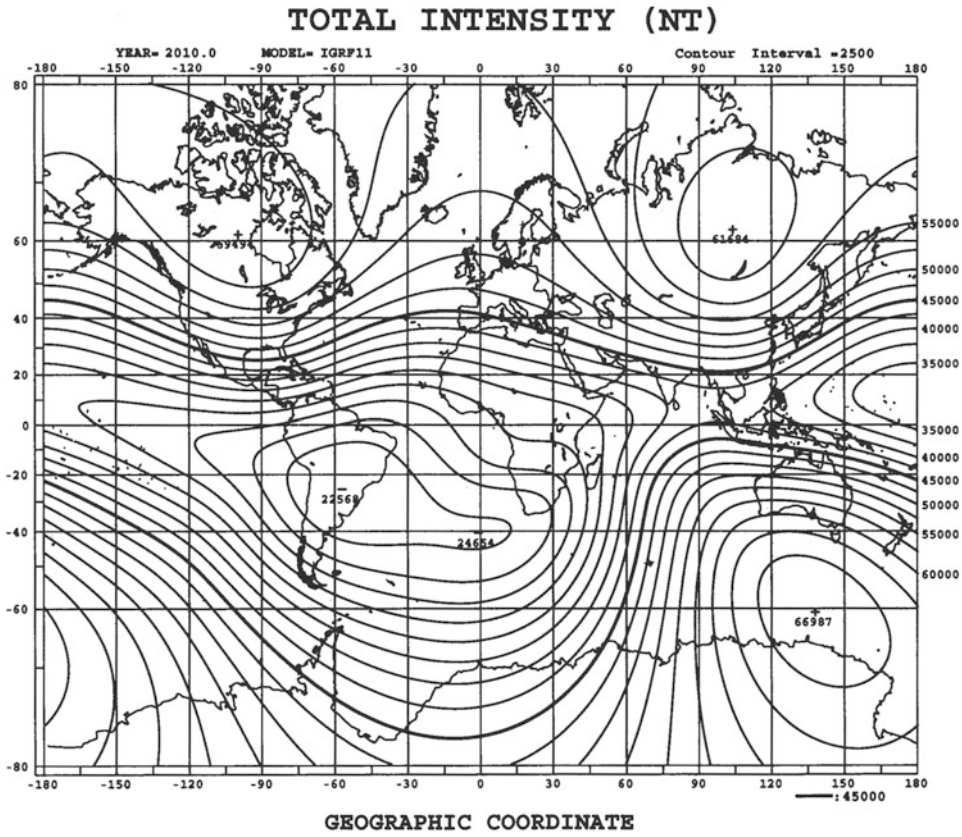


Figure 3.5. A chart showing the isointensity lines of the global magnetic field. The magnetic poles are represented by two antipodal maximum areas in northeast Canada and in the southeastern Antarctic Ocean. Another maximum is seen in Siberia and there is an elongated minimum in the Atlantic Ocean close to South America (the Atlantic Anomaly). (From http://wdc.kugi. kyoto-u.ac.jp/igrf.2010.)

magnetic field. The geomagnetic poles therefore depend on a network of magnetic field observations covering the whole globe.

Figure 3.5 shows a global map of the total magnetic field intensity of the Earth illustrated with the so-called isointensity lines. We notice the magnetic north and south poles west of Hudson Bay and at the northeastern coast of the Antarctic, respectively. Furthermore, there is an area of very strong magnetic field in Siberia. For many years this was thought to be one of the poles of another pair. The discussion of whether the globe had two or four poles was very vivid up towards the middle of the 20th century. Finally, there is an area of unusually low magnetic field intensity just off the Atlantic coast of South America, the so-called Atlantic Anomaly.

It is often very convenient to use the geomagnetic field system as an alternative reference system since many problems dealt with in geophysics are

influenced or controlled by the magnetic field of the Earth. One often-used elementary geomagnetic coordinate system is spherical with its origin at the Earth's center and with its axis along the geomagnetic axis.

Several geomagnetic coordinate systems, often called corrected geomagnetic coordinate systems, exist in the literature. Since the magnetic system is always changing, these corrected geomagnetic systems are regularly updated.

Observations of the geomagnetic field through decades indicate that the electric current flowing between the Earth's surface and space, although of considerable interest in themselves, are of minor importance with respect to the origin of the geomagnetic field. Outside the Earth we can assume in this regard that the current is negligible ($\mathbf{j} = 0$) and therefore:

$$\nabla \times \mathbf{B} = 0$$

outside the Earth. There must then exist a magnetic potential such that:

$$\mathbf{B} = -\nabla V_M \tag{3.1}$$

Since $\nabla \times \mathbf{B} = 0$ everywhere and especially at the Earth's surface, we find:

$$\nabla^2 V_M = 0 \tag{3.2}$$

Solving then for V_M the magnetic field could be derived at any point in space by the following formulas in spherical polar coordinates (Figure 3.6):

$$B_\theta = -\frac{1}{r}\frac{\partial V_M}{\partial \theta} \tag{3.3a}$$

$$B_\varphi = -\frac{1}{r \sin\theta}\frac{\partial V_M}{\partial \varphi} \tag{3.3b}$$

$$B_r = \frac{\partial V_M}{\partial r} \tag{3.3c}$$

The derived values of $(B_\theta, B_\varphi, B_r)$ could then be compared with the measured components of $\mathbf{B}$ at any point in space and especially with the measured elements (X, Y, Z) or (H, D, Z) (Figure 3.7) of the magnetic field on the ground. The degree of correspondence between these values could then be used to assess the value of the derived model V_M. To solve for V_M is, however, no straightforward process since the real B-field cannot be described by any analytical function.

When discussing the elements of $\mathbf{B}$, (X, Y, Z), or (H, D, Z) at a point on the Earth, it is customary to let the z-axis point downwards towards the center of the Earth (Figure 3.7) and the x- and y-axis towards geographic north and east, respectively. In the northern hemisphere $\mathbf{B}$ points downward slanted towards north. The (X, Y, Z) elements of $\mathbf{B}$ are then defined positive in the north, east, and downward directions, respectively. The angle D is the declination or the deviation with which the magnetic field is pointing with respect to true north. It is defined positive eastward. The angle I is called inclination and describes the angle the magnetic field makes with the horizontal plane, measured positive downward from this plane. H is the component of the B-field in the horizontal plane.

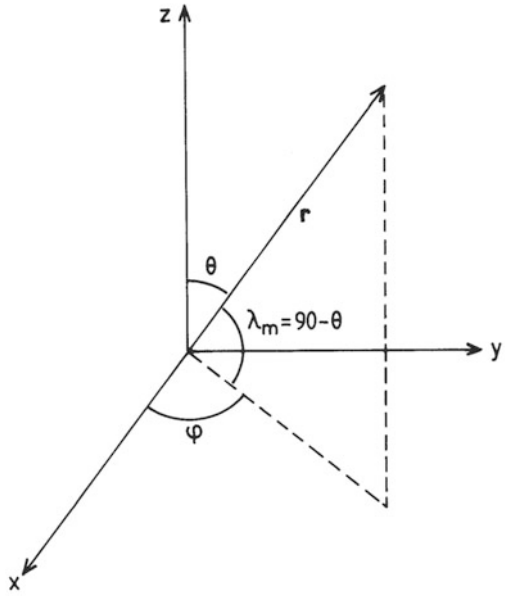


Figure 3.6. The polar coordinate system used in the text.

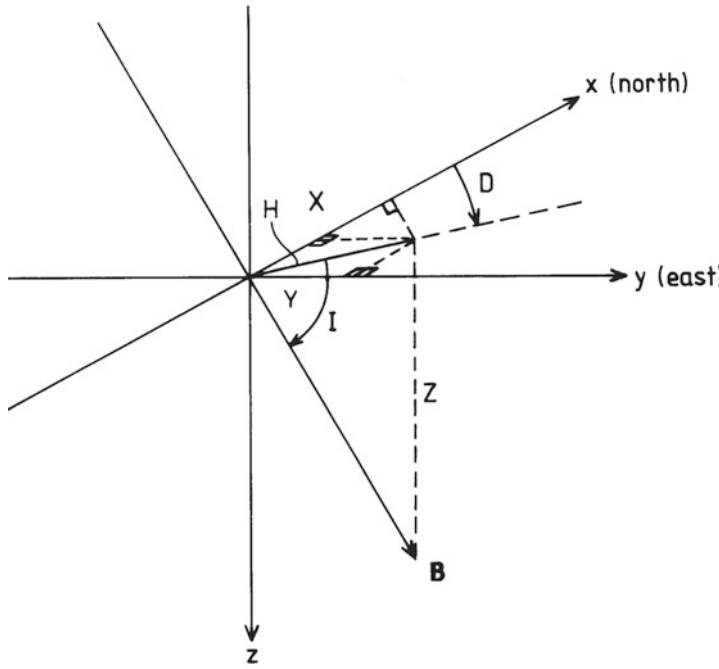


Figure 3.7. The magnetic elements used to describe the Earth's magnetic field (X, Y, Z), (H, D, Z), or (H, D, I).

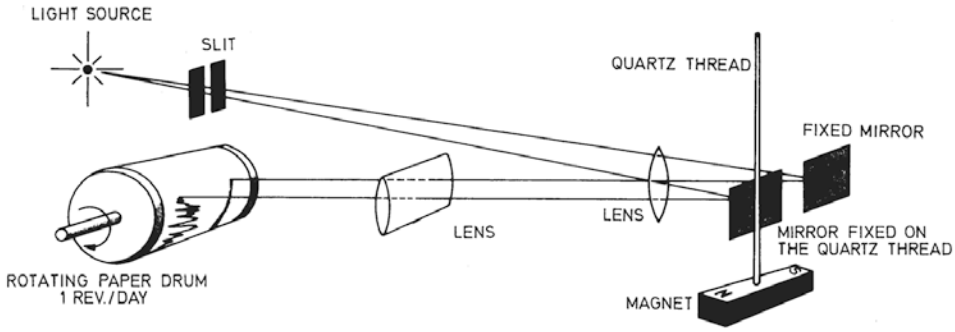


Figure 3.8. A schematic diagram showing the principles for magnetic field recording using a suspended magnet and photographic recording paper on a rotating drum—the classical variometer setup.

We notice that when (X, Y, Z) are measured, H, D, and I can be derived:

$$H = \sqrt{X^2 + Y^2}$$
$$D = \arctan(Y/X)$$
$$I = \arctan(Z/H)$$

which also gives the relationship between the two sets of elements depending on which is measured.

Historically, the Earth's magnetic field has been measured by the effect it has on small magnets suspended in torsion threads of quartz or balanced on tiny knife edges. A schematic illustration of the principle for a magnetometer recording on photographic paper is shown in Figure 3.8. Today, however, these magnetometers are being replaced by computerized electronic devices. At present about 300 magnetic observatories are in operation, but many of them are under constant threat due to the expansion of urbanization all over the globe. Due to the high sensitivity of the magnetometers, they are exposed to disturbances of many kinds.

3.3 MATHEMATICAL REPRESENTATION OF THE EARTH'S MAGNETIC SYSTEM

Since the magnetic field of the Earth is not derivable from a simple analytical function, one has to take the opposite approach to find V_M from a set of observations of **B** and make a fitting analysis of these measurements to the derivatives of a model potential. The customary way to handle this problem is to expand the magnetic potential in a series of spherical harmonics. Such a solution expressed in

the polar coordinates (r, θ, φ) (Figure 3.6) is given below:

$$V_M = R_e \sum_{n=0}^{\infty} \left(\frac{R_e}{r}\right)^{n+1} \sum_{m=0}^{n} P_n^m(\cos\theta) \cdot [g_n^m \cdot \cos(m\varphi) + h_n^m \cdot \sin(m\varphi)]$$

where R_e is the radius of the Earth, θ and φ are the geographic colatitude and east longitude, respectively, $P_n^m(\cos\theta)$ are normalized associated Legendre functions, g_n^m and h_n^m are the so-called Gaussian coefficients that one wishes to derive in order to find V_M, μ_0 $(= 4\pi \times 10^{-7}$ H/m) is the permeability in a vacuum, and V_M refers to internal sources. There is also a term related to external sources in the full expression for V_M, but, as already mentioned, these are negligible and contribute less than 1% to the total field at the Earth's surface and are therefore neglected in the following.

Since potential is not a measured quantity but the components of **B** at different places on the Earth's surface, a spherical harmonic analysis can be performed on a number of observed sets of the elements (components of **B**).

When such a large set of the elements (X, Y, Z) are collected from many stations all over the globe, these can be fitted by the method of least squares, and the coefficients g_n^m and h_n^m can be derived at the Earth's surface $(r = R_e)$. For practical reasons, however, the summation is truncated at a reasonable m. Table 3.1 lists the first eight coefficients derived by this method for the International Geomagnetic Reference Field (IGRF) every fifth year since 1965.

Table 3.1. Spherical harmonic expansion coefficients in nanoteslas (nT)

Model	g_1^0	g_1^1	h_1^1	g_2^0	g_2^1	g_2^2	h_2^1	h_2^2
IGRF* 10	−29,496	−1,586	4,945	−2,397	3,026	1,669	−2,708	−575
IGRF 05	−29,554	−1,669	5,077	−2,337	3,047	1,658	−2,594	−515
IGRF 00	−29,619	−1,728	5,186	−2,268	3,068	1,671	−2,482	−458
IGRF 95	−29,692	−1,784	5,306	−2,200	3,070	1,681	−2,336	−413
IGRF 90	−29,775	−1,848	5,406	−2,131	3,059	1,686	−2,279	−373
DGRF** 85	−29,873	−1,905	5,500	−2,072	3,044	1,687	−2,197	−306
DGRF 80	−29,992	−1,956	5,604	−1,997	3,027	1,663	−2,129	−200
DGRF 75	−30,100	−2,013	5,675	−1,902	3,010	1,632	−2,067	−68
DGRF 70	−30,220	−2,068	5,737	−1,781	3,000	1,611	−2,047	25
DGRF 65	−30,334	−2,119	5,776	−1,662	2,997	1,594	−2,016	114

*IGRF, International Geomagnetic Reference Field.
**DGRF, Definite Geomagnetic Reference Field.

The potential due to the first term g_1^0 is the main dipole term ($n = 1$) given by

$$V_1 = g_1^0 \cos\theta \left(\frac{R_e}{r}\right)^2 R_e \tag{3.4}$$

g_0^0 is zero since no monomagnetic pole ($n = 0$) exists. This dipole will then be associated with a magnetic moment given by:

$$M_1 = \frac{4\pi}{\mu_0} g_1^0 R_e^3$$

which is oriented along and antiparallel to the Earth's rotation axis since g_1^0 is negative (Figure 3.9). The next two terms g_1^1 and h_1^1 are also dipole terms in the plane of the geographic equator. $M_2 = (4\pi/\mu_0) g_1^1 R_e^3$ points towards 180°

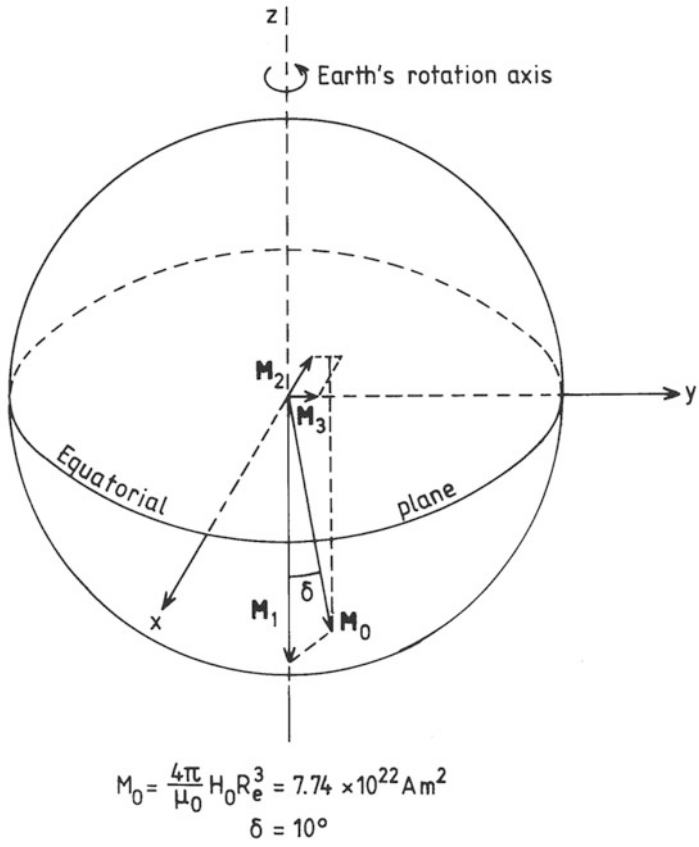


Figure 3.9. The first three dipole terms $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$ in the harmonic series expression of the Earth's magnetic field represented by the coefficients g_1^0, g_1^1, and h_1^1, respectively (see Table 3.1). The resultant magnetic moments $\mathbf{M}_0$ ($= 7.74 \times 10^{22}$ A m^2) make an angle δ ($= 10°$) with the Earth's rotation axis.

geographic longitude ($g_1^1 < 0$), and $M_3 = (4\pi/\mu_0)h_1^1 R_e^3$ points towards 90° E longitude. The resultant of these three dipoles is a dipole with a magnetic moment of

$$M_0 = \frac{4\pi}{\mu_0} H_0 R_e^3 \tag{3.5}$$

where H_0 is determined by

$$H_0 = [(g_1^0)^2 + (g_1^1)^2 + (h_1^1)^2]^{1/2}$$

Inserting from Table 3.1 (IGRF 10) we find

$$H_0 = 29{,}950 \text{ nanotesla} = 3.0 \times 10^{-5} \text{ tesla}$$

and since $R_e = 6.37 \times 10^6$ m

$$H_0 R_e^3 = 7.74 \times 10^{15} \text{ Wb m}$$

and from (3.5)

$$M_0 = 7.74 \times 10^{22} \text{ A m}^2$$

We also notice that M_0 makes an angle with the rotation axis of the Earth which is given by:

$$\tan \delta = [(g_1^1)^2 + (h_1^1)^2]^{1/2}/g_1^0 = 0.18$$

and

$$\delta = 10^\circ$$

which is the angle we mentioned at the beginning of this chapter.

Having derived the magnetic potential with the proper number of coefficients demanded for the problem of interest, the empirical components of **B** can be obtained at any point of space where the assumptions made apply and **B** is given by (3.1) in any coordinate system chosen for the problem to be solved.

3.4 SECULAR VARIATIONS IN THE EARTH'S MAGNETIC FIELD

As mentioned earlier the magnetic field system of the Earth is always changing and the series of spherical harmonics have to be updated from year to year (Table 3.1). For the last 150 years or so we have reasonably good data which enable us to study the variation in the Earth's dipole moment through this time. Figure 3.10 illustrates this variation where the dipole moment, M_0, derived between 1550 and 2010, is given in units of A m^2. We notice that there has been a decreasing trend in M_0 and that the decrease is in fact stronger now than it has been throughout this period. If this tendency continues, the magnetic field will disappear within about 2,000 years.

As already pointed out by Halley, the magnetic field system is drifting. The speed of the drift is not constant but decreases by latitude as shown in Figure 3.11, where the average westward drift per year is plotted versus latitude for some stations in the northern hemisphere. In Canada and Norway the drift is more than 0.5 degrees per year. From these analyses it is also possible to deduce where the geomagnetic pole would be at any time. In Figure 3.12 the geographic longitude

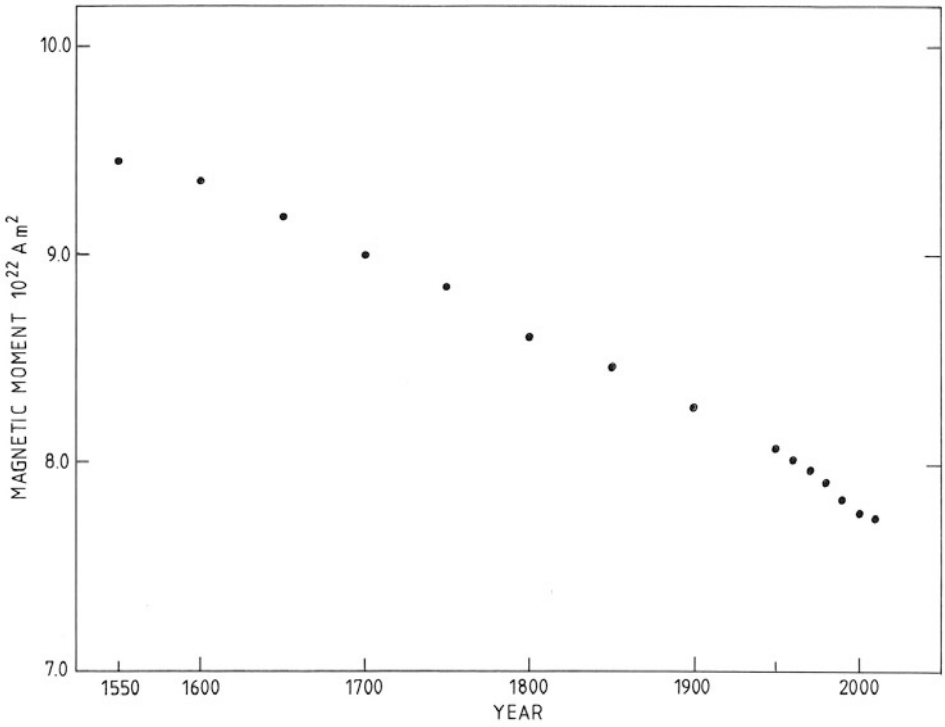


Figure 3.10. Variation in the Earth's dipole moment (in A m^2) as a function of time between 1550 and 2010 showing a gradual decline in strength.

and latitude of the geomagnetic north pole is shown for the period between AD 1600 and 1900. While it was at almost 83° N and 40° W at Gilbert's time, it is now closer to 80° N and 70° W.

It is in fact possible to study variations in the Earth's magnetic field further back than 400 years by so-called archeomagnetic studies. This research is based on magnetic material in pottery, burned clay, sediments, and lava. Without going into detail we reproduce a figure (Figure 3.13) showing the drift of the geomagnetic pole in the northern hemisphere as far back as the beginning of our era (i.e., since the time of Jesus Christ). At the present time the drift is about 2 arcminutes per year westwards.

This drift of the magnetic system has special importance for auroral morphology. As is well known from years of ground-based observations and especially revealed by satellite imaging from the DE-1 and Viking satellites, the aurora tends to appear in an oval-shaped pattern with its center close to the geomagnetic pole (Figure 3.14). Since this pole has changed position over the years, it is also likely that the auroral oval must have been situated differently in periods compared with the present average situation. This is illustrated in Figure 3.15 where the auroral oval is drawn for four different periods in historic time. It is especially noted that

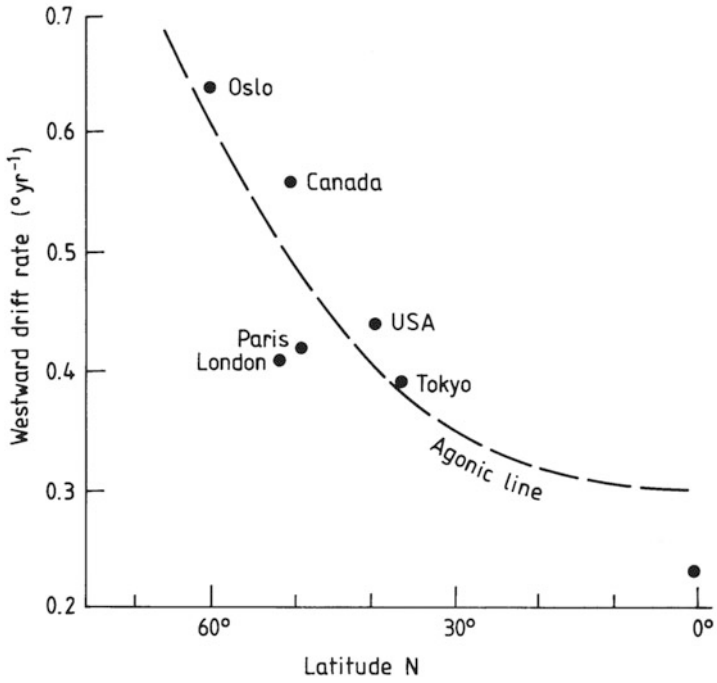


Figure 3.11. Westward drift of the global magnetic system indicated by the rate of change (°/year) of the declination angle at different latitudes. (After Yukutake, 1967.)

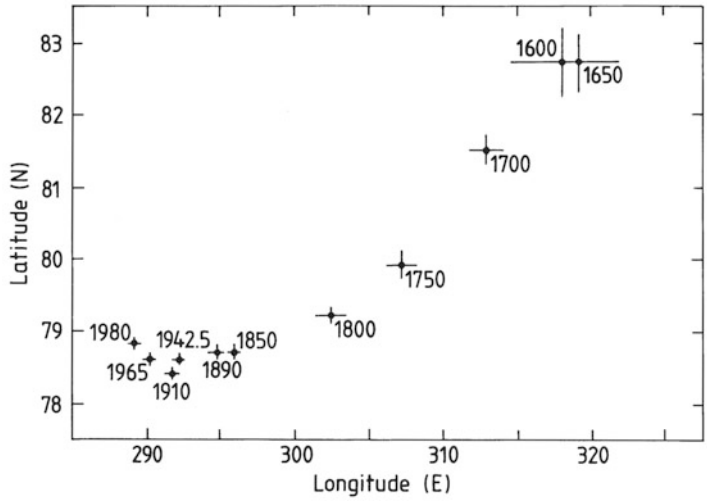


Figure 3.12. The variation in position of the north geomagnetic pole from AD 1600 to AD 1980 showing a southwestward drift. (After Barraclough, 1974.)

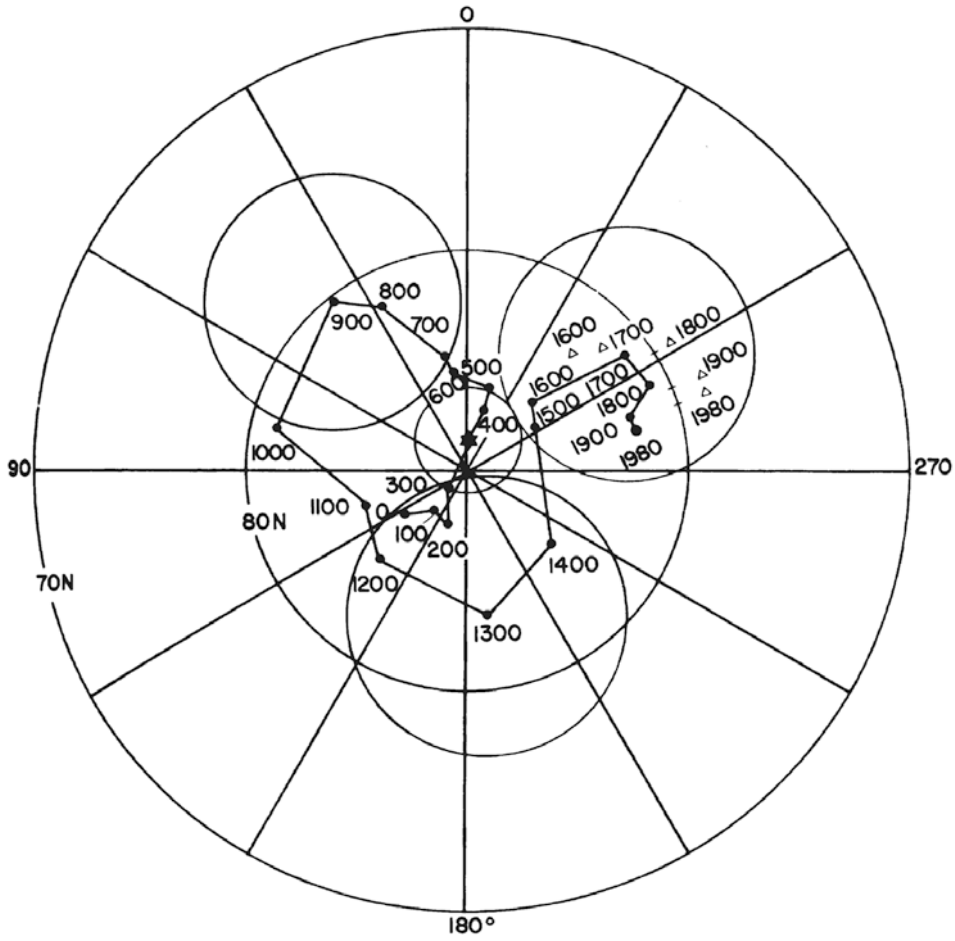


Figure 3.13. The position of the geomagnetic north pole for the last 2,000 years. (From Merrill and McElhinny, 1983.)

at about AD 1200 the oval was well north of Europe. The shift of the oval towards the south over northern Europe is also evident in the figure.

The fact that the magnetic field decreases will also have an impact on the geographical appearance of the aurora. The weaker the field, the lower in latitude the aurora will appear. For the last 3,000 years or so (Figure 3.16) the magnetic field strength has been decreasing. Therefore, more people, not only those living at extreme latitudes, can occasionally enjoy the displays of the aurora. 6,000 years ago the magnetic moment of the Earth was weaker than it is today, corresponding to a period when the aurora was occurring at lower latitudes. We have indications from this early time that the aurora was observed in China at about 30° of latitude.

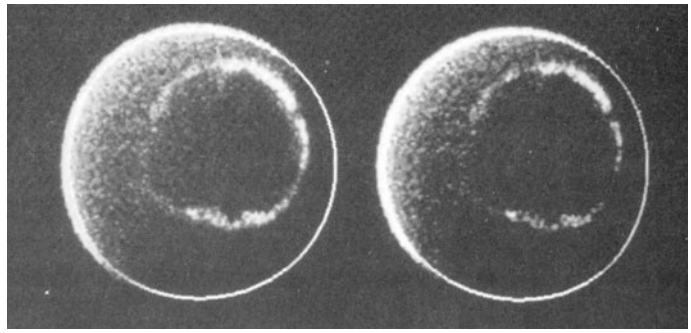


Figure 3.14. Modern pictures of the auroral oval as derived from imaging techniques onboard the DE-1 satellite. (From Frank, 1994.)

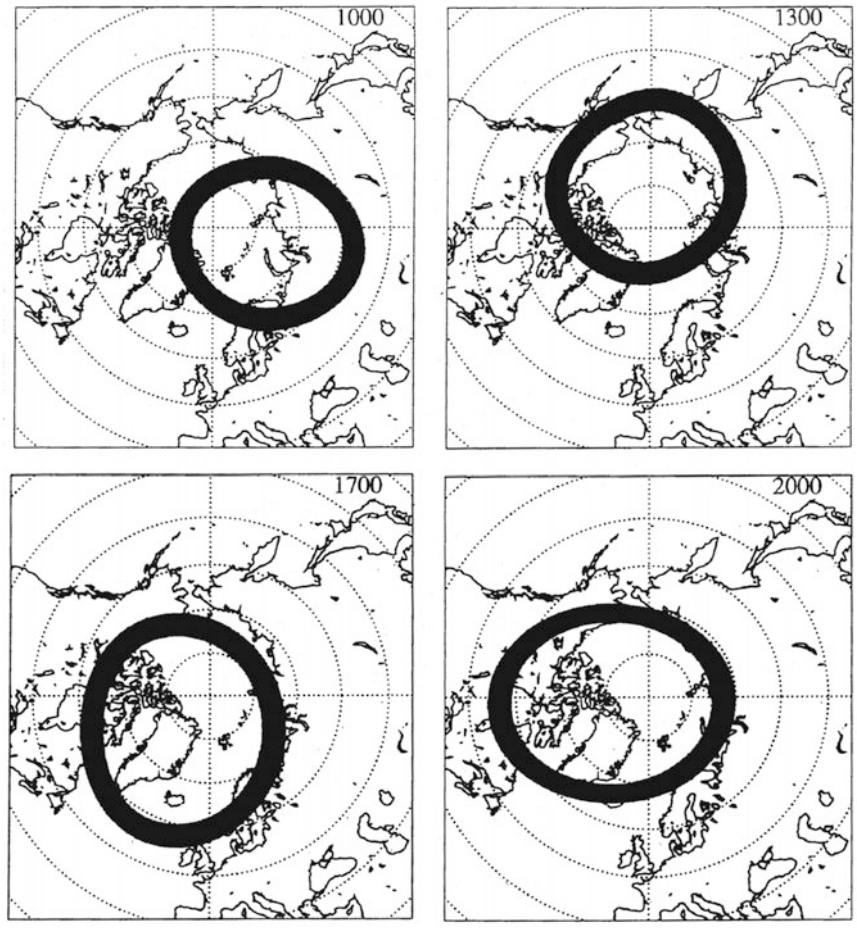


Figure 3.15. The position of the auroral oval for four periods in historic time.

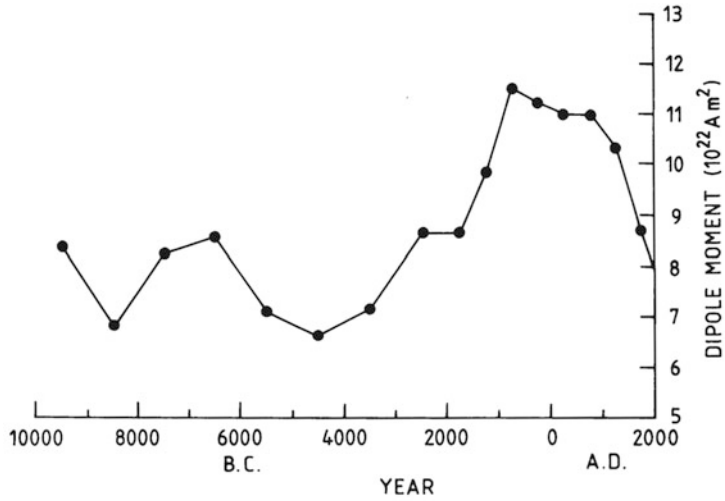


Figure 3.16. The change of the magnetic dipole moment as deduced from archeomagnetic data for the last 12,000 years. (After McElhinny and Senanayake, 1982.)

By taking this reduction into account and assuming a continuous drift of the geomagnetic axes, the position of the oval in the future can be predicted. One such prediction for the year AD 2300 is shown in Figure 3.17. According to this the oval will probably be crossing the southern part of Scandinavia in a few hundred years.

It was indicated earlier that if the decrease in the magnetic moment, as we observe today, continues with the same trend, it might disappear within 2000 years or so. The question is, however, whether it disappears or just changes direction. We can get an idea about this feature of the Earth's magnetic field by studying archeomagnetic data, some of which are shown in Figure 3.18. The results are presented in terms of geomagnetic declination referring to the angle δ between the Earth's rotation axis and the magnetic axis. At present and for the last 1 million years the polarity of the magnetic field of the Earth has probably been the same. About 1 million years ago, however, the magnetic moment flipped over by 180° in the course of 100,000 years or less. Similar events appear to have taken place several times in the last 4 million years. The question of how fast the magnetic moment changes polarity is still open, but observations have been found that can indicate a turnover time of the order of 10,000 years or less.

The magnetic field of the Earth acts as a shield against cosmic rays and energetic particles radiated from the Sun. At present it mainly shields the lower latitudes of the Earth from this radiation, while the particles can penetrate more freely along the magnetic field lines at higher latitudes.

Charged cosmic rays are affected by the Earth's magnetic field long before they enter the Earth's atmosphere. The magnetic field actually acts as a kind of momentum analyzer on cosmic rays. For each location on the Earth and for each

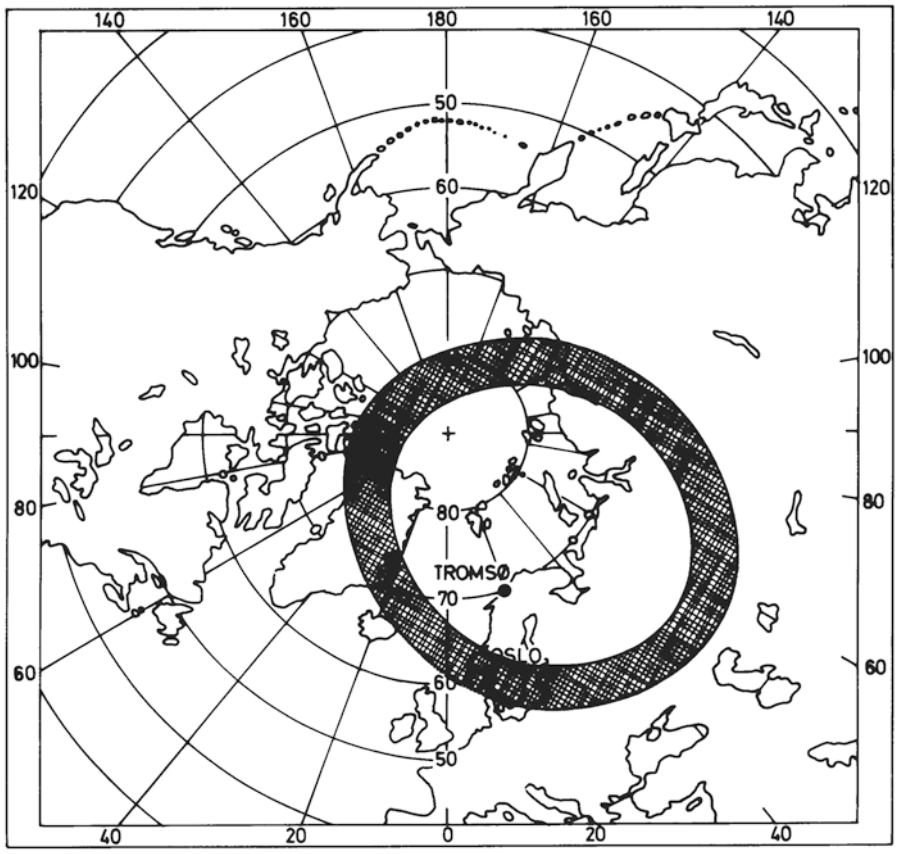


Figure 3.17. Prediction of the position of the auroral oval in AD 2300. (After Oguti, 1994.)

direction of incidence one can define a rigidity that is the ratio of momentum to charge of the cosmic ray particle:

$$R_c = pc/Z$$

where $p = mv$ is the momentum, Z the number of charges of the particle, and c is the speed of light. Only those particles that have a rigidity exceeding that of the threshold of the observation point may be detected at that position on the Earth's surface or at a distance away from the Earth. It can be demonstrated that this rigidity is proportional to the Earth's magnetic moment. The unit of R_c can be given in electronvolts, and in Figure 3.19 is shown iso-contour lines of effective cutoff rigidities on the Earth's surface for vertical incident cosmic rays. While particles with very small energies < 1 GeV can reach latitudes above 70° only particles with energies >14 GeV can reach the Earth's surface at the equator. These curves are calculated for the IGRF 1990 epoch model where the Earth's magnetic moment was approximately 7.84×10^{22} A m^2. As the magnetic moment decreases

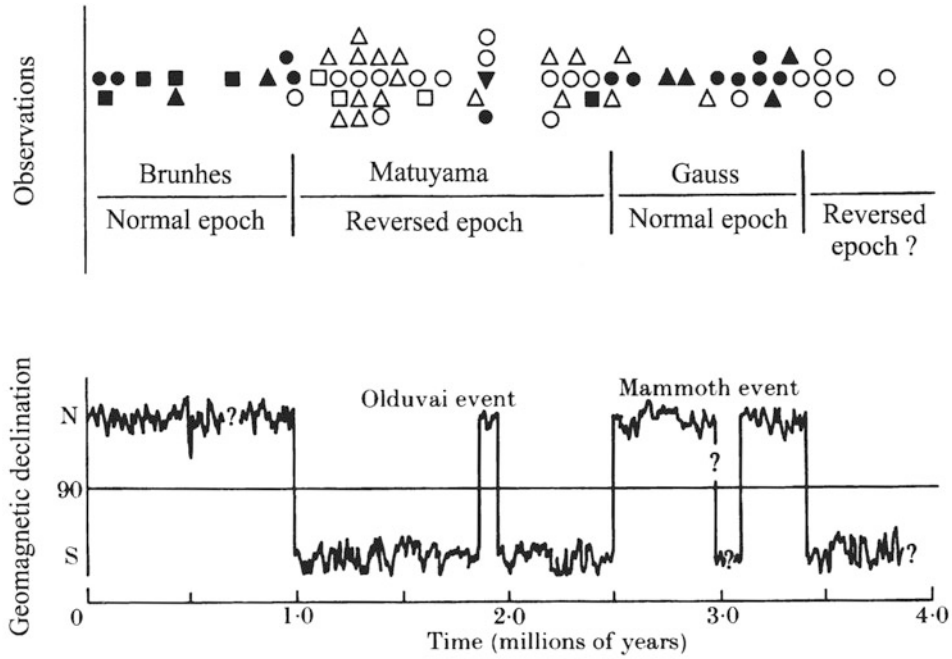


Figure 3.18. Magnetic polarities as deduced from 64 volcanic rocks taken from ○ North America, ▽ Africa, □ Europe, and △ Hawaii. The open and closed symbols represent different polarities. (From Cox *et al.*, 1967.)

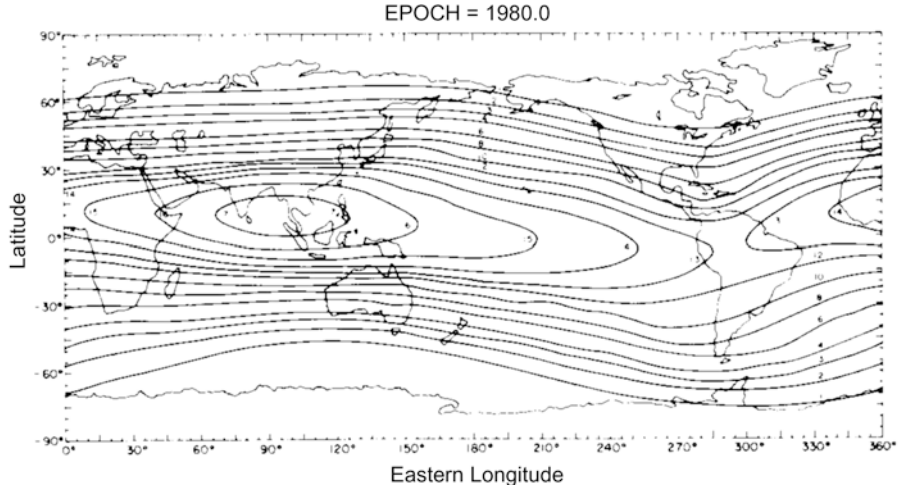


Figure 3.19. Iso-rigidity contours of vertical cosmic ray effective cutoff rigidities. The units of cutoff rigidity are in gigavolts. (From *Handbook of Geophysics and the Space Environment*, Air Force Geophysics Laboratory, 1985.)

at present the cutoff rigidities will decrease and cosmic rays with power energies will enter the atmosphere. This change will be especially noticeable at lower latitudes. It is believed that cosmic rays can play a role as condensation nuclei for water vapor in the troposphere. As the Earth's magnetic field now (2010) has decreased to $7.74 \times 10^{22}\,\mathrm{A\,m^2}$ the threshold energy also has decreased by a similar factor (~ 0.98). It is for this reason it is expected that cosmic rays will have an increasing climatic effect.

By coincidence the high latitudes at the present geological time are rather hostile areas, where only a few extant species are able to survive. Those who do so are probably adapted to any high-energy particle radiation that might hit the Earth. If the magnetic field disappeared for a while, all living species on the Earth would be exposed to a much higher radiation source. Since some of these species would probably not be adapted to such high radiation, they may well become extinct. It has been speculated that some dinosaur species actually succumbed to extraterrestrial radiation in periods when the Earth's magnetic field underwent a reversal. The time history for such reversals is shown in Figure 3.18.

3.5 TRACING MAGNETIC FIELD LINES

The magnetic field system can, to a good approximation (within 10%), be represented by a dipole with magnetic moment $\mathbf{M}_0$. Furthermore, we notice that the angle δ between $\mathbf{M}_0$ and the Earth's rotation axis is very small. In the following we will therefore neglect all terms other than the dipole moment $\mathbf{M}_0$ and assume that this is antiparallel to the Earth's rotation axis. By the choice of our coordinate system we then have

$$\mathbf{M}_0 = -M_0\hat{\mathbf{z}}$$

The dipole is in itself a solution of the Laplace equation (3.2) and the dipole potential is

$$V_M = -\frac{\mu_0}{4\pi}\mathbf{M}_0 \cdot \nabla\left(\frac{1}{r}\right) = \frac{\mu_0}{4\pi}\frac{\mathbf{M}_0 \cdot \mathbf{r}}{r^3} = -\frac{\mu_0 M_0}{2\pi}\frac{\cos\theta}{r^2}$$

The components of $\mathbf{B}$ can then be derived by taking the partial derivatives of V_M according to (3.3) and introducing (3.5)

$$B_r = -\frac{\partial V_M}{\partial r} = -\frac{\mu_0 M_0}{2\pi}\frac{\cos\theta}{r^3} = -\frac{\mu_0 M_0}{2\pi}\frac{\sin\lambda_m}{r^3} = -2H_0\left(\frac{R_e}{r}\right)^3 \sin\lambda_m \tag{3.6a}$$

$$B_\varphi = -\frac{1}{r\sin\theta}\frac{\partial V_M}{\partial \varphi} = 0 \tag{3.6b}$$

$$B_\theta = -\frac{1}{r}\frac{\partial V_M}{\partial \theta} = -\frac{\mu_0 M_0}{4\pi}\frac{\sin\theta}{r^3} = -\frac{\mu_0 M_0}{4\pi}\frac{\cos\lambda_m}{r^3} = -H_0\left(\frac{R_e}{r}\right)^3 \cos\lambda_m \tag{3.6c}$$

The magnitude of **B** is given by:

$$B(r, \lambda_m) = (B_r^2 + B_\varphi^2 + B_\theta^2)^{1/2} = \frac{\mu_0}{4\pi}\frac{M_0}{r^3}(1 + 3\sin^2 \lambda_m)^{1/2}$$

Introducing (3.5) for M_0 we get

$$B(r, \lambda_m) = H_0\left(\frac{R_e}{r}\right)^3 (1 + 3\sin^2 \lambda_m)^{1/2}$$

We notice that the pole where $\lambda_m = 90^\circ$ and $r = R_e$

$$B_p = 2H_0$$

and at the equator where $\lambda_m = 0$

$$B_e = H_0$$

Introducing the unit vector $\hat{\boldsymbol{\lambda}}_m$ where

$$\hat{\boldsymbol{\lambda}}_m = -\hat{\boldsymbol{\theta}}$$

into (3.6) we can form:

$$\mathbf{B} = B_r\hat{\mathbf{r}} + B_{\lambda_m}\hat{\boldsymbol{\lambda}}_m$$

For a position at the Earth's surface where $r = R_e$ and the magnetic latitude is λ_m, we can find an expression for **B** by inserting M_0 from (3.5) into (3.6):

$$\mathbf{B} = H_0(-2\sin \lambda_m\hat{\mathbf{r}} + \cos \lambda_m\hat{\boldsymbol{\lambda}}_m)$$

The inclination angle for the magnetic field at any point on the Earth would be (Figure 3.20)

$$\left.\begin{aligned} \tan I &= -\frac{B_r}{B_\lambda} = 2\tan \lambda_m \\ \tan \alpha &= \frac{r\, d\lambda_m}{dr} \\ \tan \alpha &= \frac{B_\lambda}{B_r} = -\frac{\cos \lambda_m}{2\sin \lambda_m} \end{aligned}\right\} \tag{3.7}$$

We notice from Figure 3.20 that since **B** is parallel to the field line drawn at any point, $\tan \alpha$, where α is the angle indicated, it can be expressed in two ways where $d\lambda_m$ and dr are increments in λ_m and r, respectively. There is therefore a simple relationship between dr and $d\lambda_m$ given by:

$$\frac{dr}{r} = -2\frac{\sin \lambda_m}{\cos \lambda_m}d\lambda_m$$

and the solution is

$$\ln r = 2\ln \cos \lambda_m + C$$

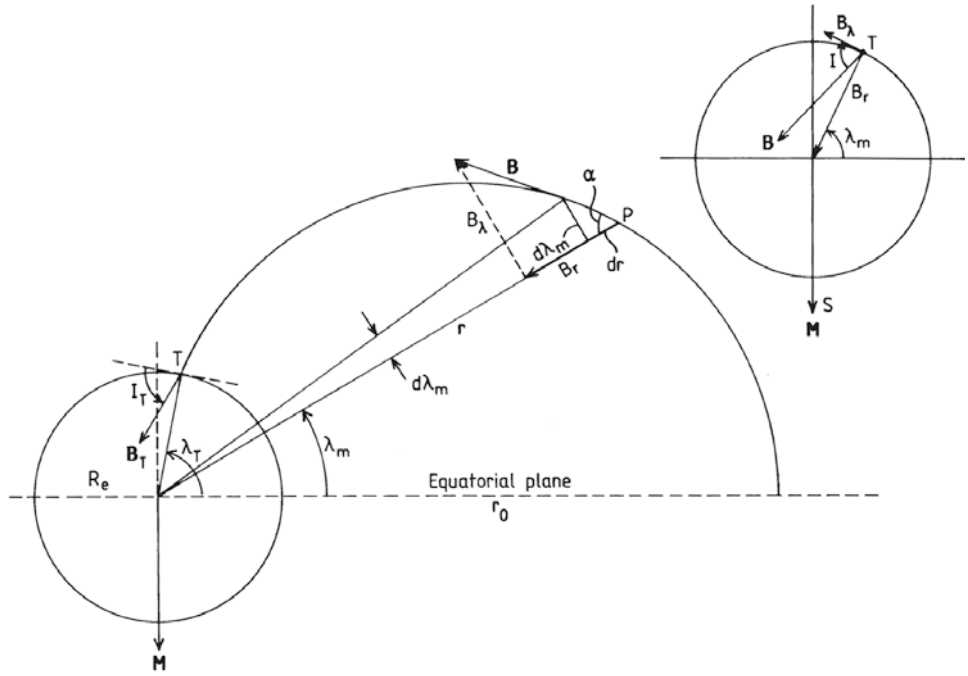


Figure 3.20. An illustration showing the geometry of the magnetic field line to assist in deriving the geometric formula for B.

If we now designate $r = r_0$ where the given field line hits the equatorial plane (i.e., where $\lambda_m = 0$), then:

$$C = \ln r_0$$

and

$$r = r_0 \cos^2 \lambda_m$$

For a place on Earth where $r = R_e$ we find that the magnetic field line through that place reaches the equatorial plane at a distance

$$r_0 = \frac{R_e}{\cos^2 \lambda_m} \tag{3.8}$$

Frequently, the ratio r_0/R_e is termed L and

$$L = \cos^{-2} \lambda_m \tag{3.9}$$

and the magnetic latitude or invariant latitude λ_m is given by:

$$\lambda_m = \arccos\sqrt{\frac{1}{L}}$$

3.6 E-FIELD MAPPING ALONG CONDUCTING MAGNETIC FIELD LINES

Let us now assume that the magnetic field lines are highly conducting. An electric potential that might have formed perpendicular to the magnetic field lines in the ionosphere due to polarization effects or induction by neutral winds blowing charges across, will then be propagated to the equatorial plane. Similar electric potentials formed between field lines in the equatorial plane or anywhere along the lines will propagate downward to the ionosphere.

Let us therefore consider two field lines (Figure 3.21) separated in azimuth by a distance $l_{i\varphi}$ along a constant geomagnetic latitude in the ionosphere. The distance between the same field lines in the equatorial plane will be denoted $L_{m\varphi}$. When neglecting the height of the ionosphere, we find:

$$l_{i\varphi} = R_e \cdot \cos \lambda_m \, d\varphi$$

where $d\varphi$ is the azimuthal angle between these two field lines. In the magnetic equatorial plane we find

$$L_{m\varphi} = r_0 \, d\varphi = \frac{R_e}{\cos^2 \lambda_m} d\varphi$$

where r_0 is given by (3.8). When conserving the electric potential between the equatorial plane and the ionosphere, we therefore have

$$V_\varphi = E_{i\varphi} l_{i\varphi} = E_{m\varphi} L_{m\varphi}$$

where $E_{i\varphi}$ and $E_{m\varphi}$ are the azimuthal components of the electric field in the ionosphere and the equatorial plane, respectively. The mapping ratio for the electric field in the azimuthal direction is then given by:

$$\frac{E_{i\varphi}}{E_{m\varphi}} = \frac{L_{m\varphi}}{L_{i\varphi}} = \cos^{-3} \lambda_m = L^{3/2}$$

when applying (3.9). The electric field in the ionosphere is therefore enhanced by a factor $L^{3/2}$ with respect to the electric field in the equatorial plane, and the enhancement factor increases for higher L values or higher geomagnetic dipole latitudes.

For two magnetic field lines separated by an angle $d\lambda_m$ in the meridional plane the distance between these lines in the ionosphere is $l_{i\lambda}$, and the corresponding distance in the equatorial plane is $L_{m\lambda}$. From Figure 3.21 we notice that $L_{m\lambda} = dr_0$ and, since r_0 is given by (3.8), we find that:

$$L_{m\lambda} = dr_0 = -\frac{2R_e}{\cos^3 \lambda_m} (-\sin \lambda_m) \, d\lambda_m$$

or

$$L_{m\lambda} = 2R_e \frac{\sin \lambda_m}{\cos^3 \lambda_m} d\lambda_m$$

The corresponding distance in the ionosphere must be the line marked $l_{i\lambda}$ in

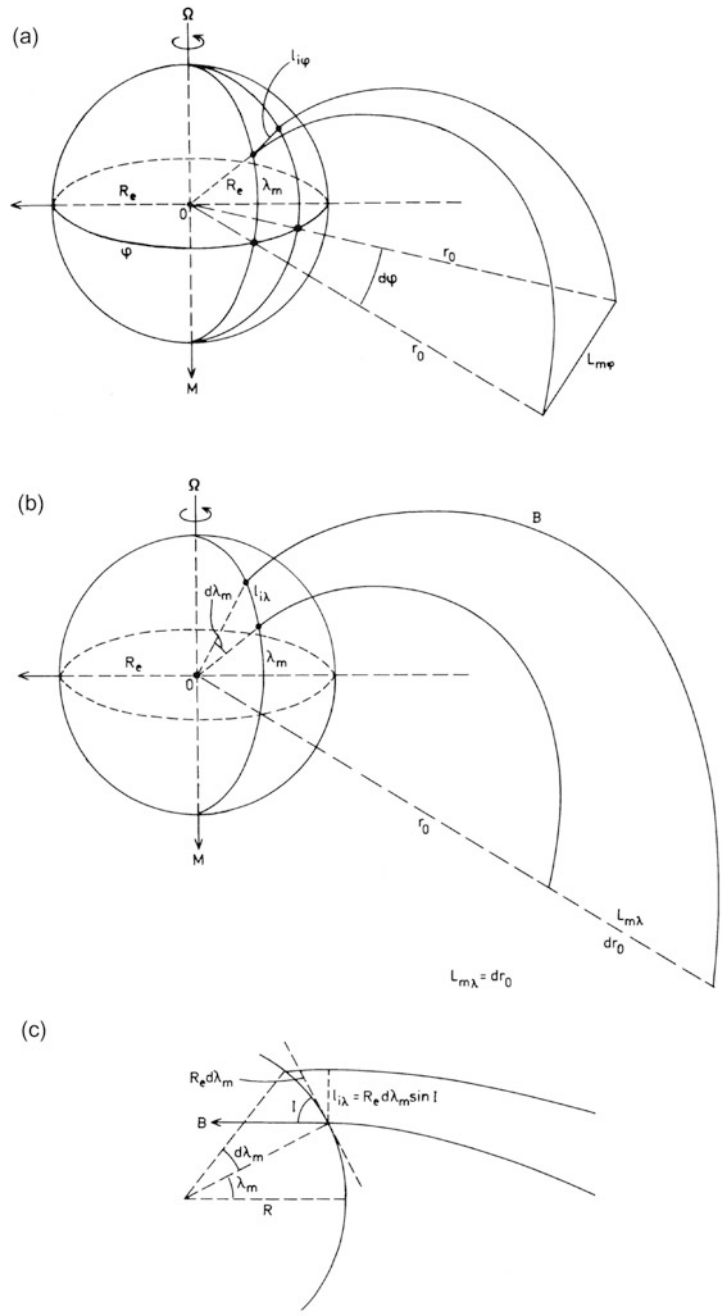


Figure 3.21. An illustration showing the geometry used to derive the mapping ratios between (a) the azimuthal distances $l_{i\varphi}$ and $L_{m\varphi}$ and (b) the latitudinal distances $l_{i\lambda}$ and $L_{m\lambda}$ for neighboring field lines in the ionosphere and equatorial plane, respectively. (c) Detail in the meridian plane.

Figure 3.21 (bottom part) which is perpendicular to the field line leaving the Earth at a geomagnetic latitude λ_m. From the figure we see that

$$l_{i\lambda} = R_e \, d\lambda_m \sin I$$

where I is the inclination angle of the B-field. For corresponding electric field components in the meridional direction $E_{i\lambda}$ and $E_{m\lambda}$ in the ionosphere and the equatorial plane, respectively, conservation of the electric potential between the two field lines gives:

$$V_\lambda = E_{i\lambda} \cdot l_{i\lambda} = E_{m\lambda} \cdot L_{m\lambda}$$

The mapping ratio is therefore

$$\frac{E_{i\lambda}}{E_{m\lambda}} = \frac{L_{m\lambda}}{l_{i\lambda}} = \frac{1}{\cos^2 \lambda_m} \cdot \frac{1}{\cos I} \tag{3.10}$$

where (3.7) has been introduced. From (3.7) we can derive:

$$\tan I = \sqrt{\frac{1 - \cos^2 I}{\cos^2 I}} = 2 \tan \lambda_m$$

and solving for $1/(\cos I)$ we get:

$$\frac{1}{\cos I} = \sqrt{4 \tan^2 \lambda_m + 1}$$

Inserting this into (3.10) we get:

$$\frac{E_{i\lambda}}{E_{m\lambda}} = \frac{1}{\cos^2 \lambda_m} \sqrt{4 \tan^2 \lambda_m + 1} = \frac{1}{\cos^3 \lambda_m} \sqrt{1 - \tfrac{3}{4} \cos^2 \lambda_m}$$

By finally introducing the L value from (3.9) we obtain:

$$\frac{E_{i\lambda}}{E_{m\lambda}} = 2 \cdot L \cdot \sqrt{L - \tfrac{3}{4}}$$

Notice that since E_λ and E_φ do not map identically, the field will be rotated when transplanted from the equatorial plane to the ionosphere or vice versa (Figure 3.22). Let the angle between the E-field and the meridional plane in the equatorial plane and in the ionosphere be δ_m and δ_i, respectively, then:

$$\tan \delta_m = \frac{E_{m\varphi}}{E_{m\lambda}}$$

$$\tan \delta_i = \frac{E_{i\varphi}}{E_{i\lambda}}$$

The relationship between $\tan \delta_i$ and $\tan \delta_m$ is then given by:

$$\tan \delta_i = \tan \delta_m \cdot \sqrt{\frac{L}{4L - 3}}$$

The electric field will be more meridional in the ionosphere than in the

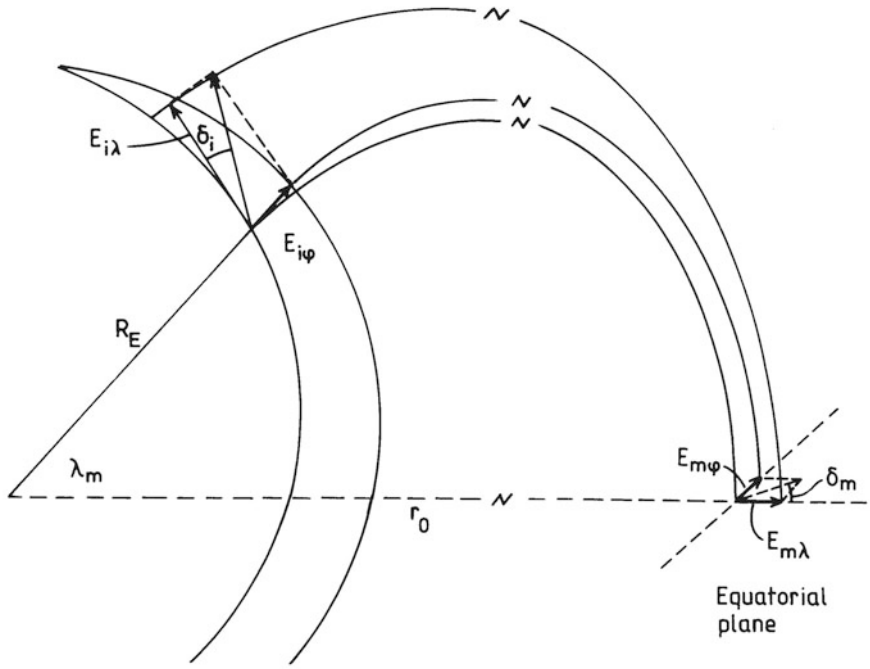


Figure 3.22. A diagram illustrating the rotation of an ionospheric *E*-field due to non-uniform mapping along field lines when projected out to the equatorial plane.

magnetosphere. Notice also that, while the azimuthal component maintains its direction between the equatorial plane and the ionosphere, the meridional component flips over from being outward in the equatorial plane to becoming poleward in the ionosphere (Figure 3.22).

3.7 THE SOURCE OF THE MAGNETIC FIELD OF THE EARTH

It is rather useless to speculate on the past and future behavior of the geomagnetic field unless we have a quite well-founded theory for the creation of the field itself to lean on. At present it is probably fair to claim that no existing theory about the source of the magnetic field can explain the dramatic episodes that appear to have taken place in the Earth's magnetic field. It is implicitly assumed in our discussions of the magnetic potential that the source of the field is in the Earth's interior. We notice from the harmonic expansion of the geomagnetic potential that the contribution to the magnetic field from a term of order n will decrease by distance from the center of the Earth by $r^{-(n+2)}$. The contribution from higher order terms will therefore vanish very rapidly, for example, at or just above the Earth's surface. Furthermore, we notice that the higher harmonic terms represent

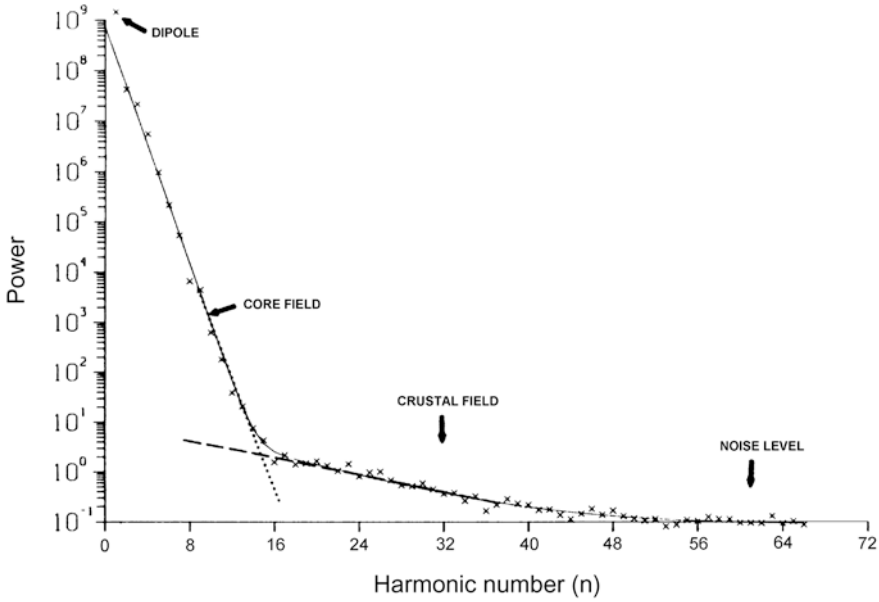


Figure 3.23. Spectrum of the geomagnetic field as seen by the Magsat satellite at 400 km altitude. The n values are those of the spherical harmonic modes; $n = 1$ is the dipole. The break in the spectrum represents the transition from the core field to that of crustal magnetic anomalies. The scale is in nanoteslas^2 and represents the energy density of the field at 400 km altitude. (From Cain, 1987.)

the smallest scale sizes in the field structure. For the term of order n

$$\varphi_n = \frac{2\pi}{n}$$

where φ_n is the angular dimension of the term of order n. Converted into Earth radii:

$$L_n = R_e \varphi_n = \frac{2\pi R_e}{n}$$

Now, the power of each sinusoidal component of the magnetic field can be measured outside the Earth (e.g., by a satellite). In Figure 3.23 the results from a satellite at 400 km are illustrated. The unit along the ordinate is in nanoteslas^2. Up to $n = 13$ which corresponds to distances of $\approx 0.48\, R_e$, the field terms decrease very rapidly with n. It is believed that these terms have their sources in the core region of the Earth. (The core has a radius of 3,485 km.) The terms between $n = 13$ and $n = 50$ have a very different slope from those at lower n, and these terms are believed to originate in crustal anomalies. For $n = 50$ the dimensions of these anomalies are ≈ 800 km.

As the magnetic field appears to originate from internal processes of the Earth, it is expected that the rotation of the Earth itself is associated with the

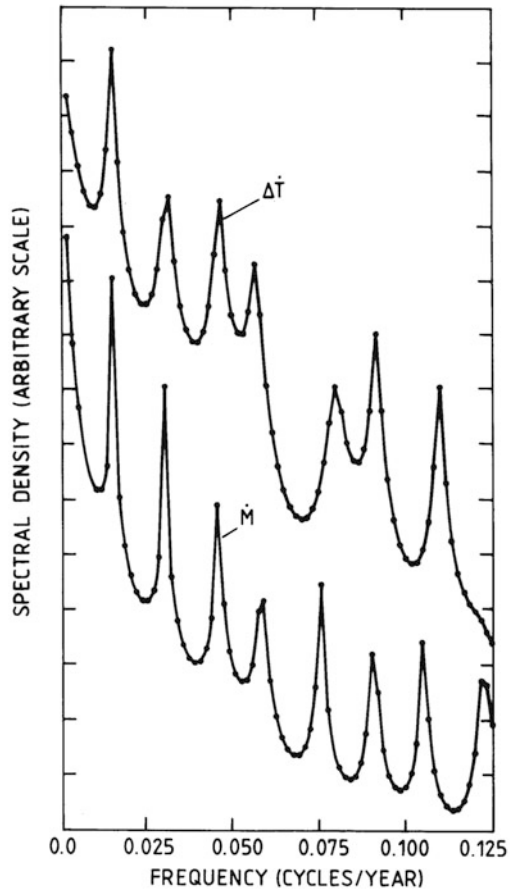


Figure 3.24. Power spectra of the time rate of change, $\Delta\dot{T}$, for the rotation period of the Earth based on data from 1865 to 1965. This is compared with the power spectrum of the variations by time $\dot{M}$ in the geomagnetic moment for the period between 1901 and 1969. The spectral densities are given in arbitrary scales. (After Jin and Thomas, 1977.)

creation of the magnetic field. A strong indication lending support to this assumption can be found in Figure 3.24 where the power spectra of the time rate of change $\Delta\dot{T}$ for the period 1865–1965 (T meaning the rotation period of the Earth), and of the variations in the dipole geomagnetic moment ($\dot{M}$) for the period 1901–1969, are shown. There is a striking similarity between these two spectra, as if they have a common course or are causally related to one another.

More observed evidence in support of a connection between the rotation of the globe and its magnetic moment is provided by Bode's law for the magnetism of celestial bodies. This is illustrated in Figure 3.25 where the magnetic moments (in $\mathrm{A\,m^2}$) for several planets in the solar system are shown versus their respective

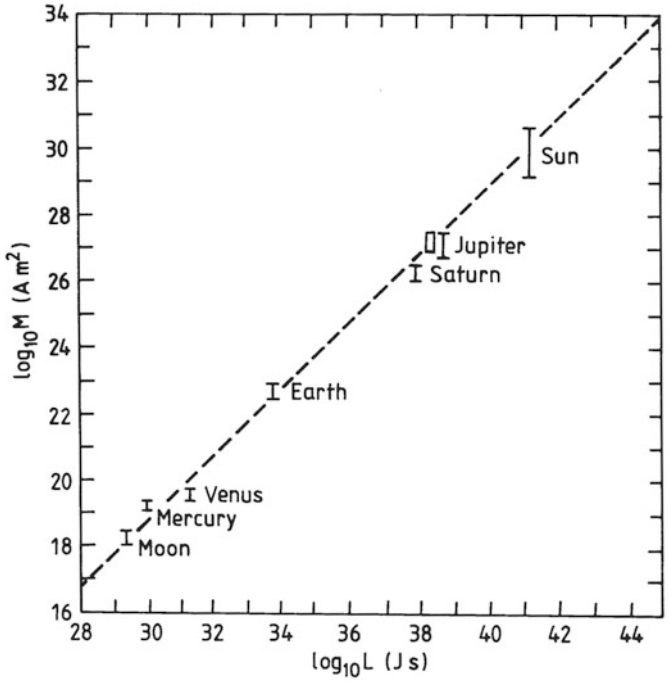


Figure 3.25. Magnetic moment for several celestial bodies as a function of the corresponding angular momentum.

angular moment L (in J s). The logarithms of the two parameters appear to be linearly related within 15 orders of magnitude.

Figure 3.26 shows the magnetic field systems for the outermost planets together with the geomagnetic system of the Earth. We notice that for the Earth $\mathbf{M}\cdot\mathbf{L} < 0$, while for the other planets $\mathbf{M}\cdot\mathbf{L} > 0$. The rotation direction and the magnetic moment therefore do not appear to be related in the same sense for all bodies. Furthermore, if the magnetic moment was strictly related to the direction of the angular momentum, the Earth would have had to reverse its rotation about the polar axis on several occasions in prehistoric time. There is no indication that such has happened.

Since the interior of the Earth is extremely hot and part of it is supposed to be a fluid of highly conducting plasma, the Earth's rotation is supposed to create rotation cells in the interior which will give rise to currents and magnetic fields.

3.8 THE UNIPOLAR INDUCTOR

One attempt to visualize the processes that can create the Earth's magnetic field is by studying the so-called unipolar inductor (Figure 3.27(a)). Rotating highly

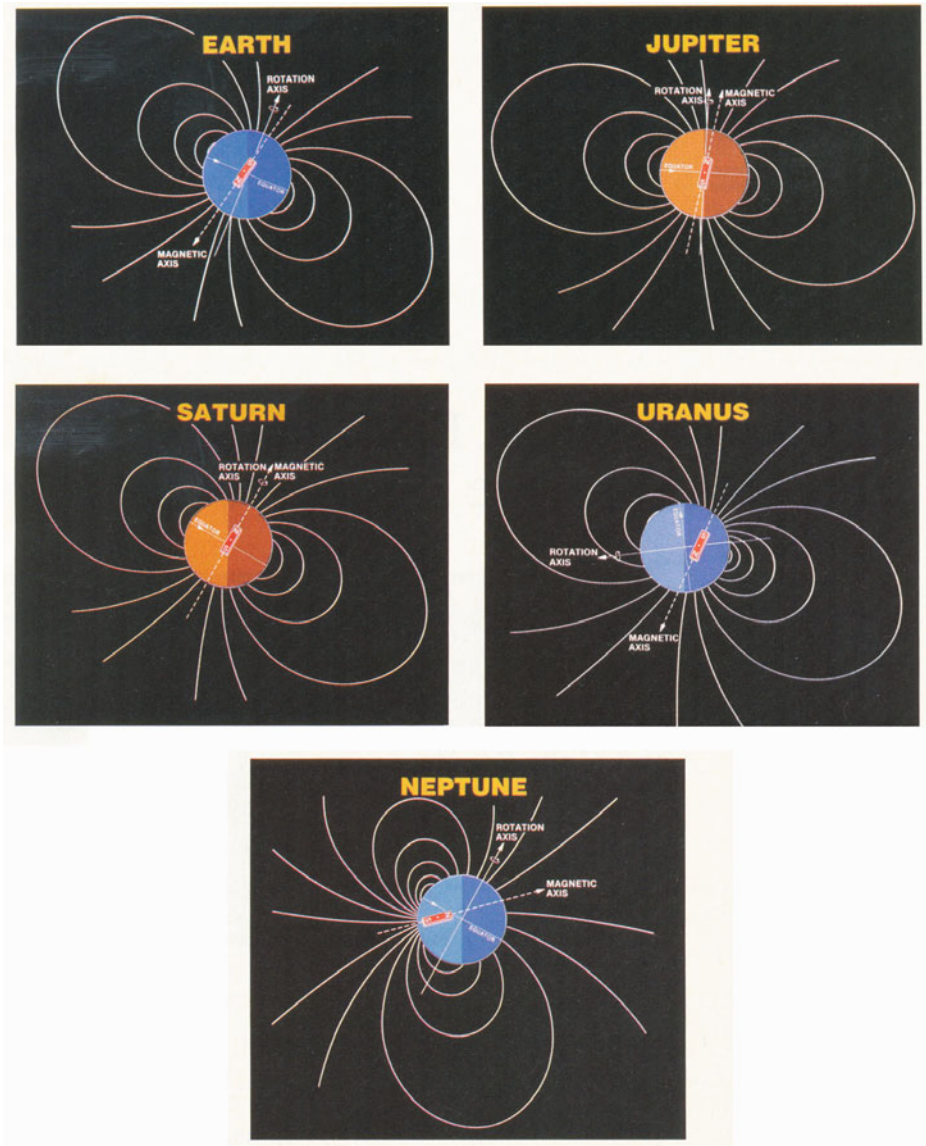


Figure 3.26. Models of the magnetic systems for the large planets indicating the relationship between the rotation axis and the magnetic axis for each planet. Also illustrated is the corresponding model for the Earth.

conducting plasma is represented by a disk spinning around an external magnetic field, say, with its origin in interplanetary space. A current feedback loop is illustrated which, when it encircles the rotation axis, can enhance or decrease the external field.

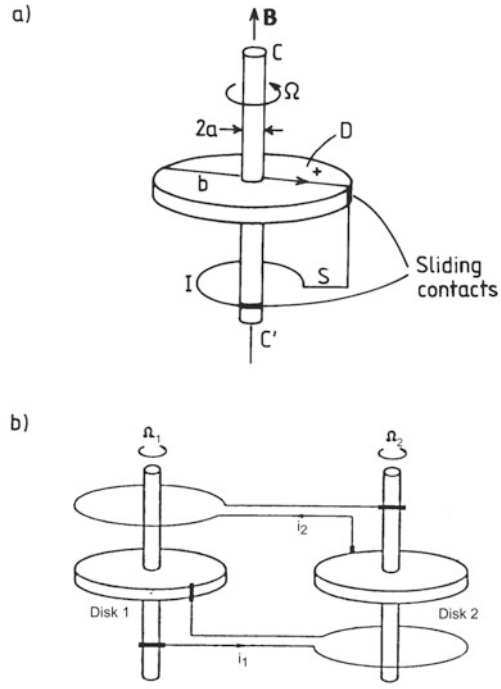


Figure 3.27. Illustrations of (a) one unipolar inductor rotating with angular velocity Ω in the presence of a magnetic field **B**. (b) A coupled pair of two unipolar inductors rotating with different angular velocities Ω_1 and Ω_2. (From Rikitake, 1958).

The current will be a function of the rotation speed of the disk and the external magnetic field. Such a model can only give rise to oscillations around a constant background magnetic field but not to a complete turn of its direction by 180°.

We notice that it is again impossible to turn the current around from plus to minus, but very large oscillations around a mean value can occur.

In order to achieve a complete turn of the current, one can link together two or more such inductors with different rotation speeds and feedback current loops (Figure 3.27(b)).

It is fair to say, however, that to establish a reliable theory for the geomagnetic field or any magnetic field associated with celestial bodies is among the most difficult problems of cosmic geophysics.

3.9 THE MAGNETIC FIELD AWAY FROM THE EARTH

The magnetic field strength in the equatorial plane observed along the direction toward the Sun between 4 and 12\,R_e from the center of the Earth is shown in

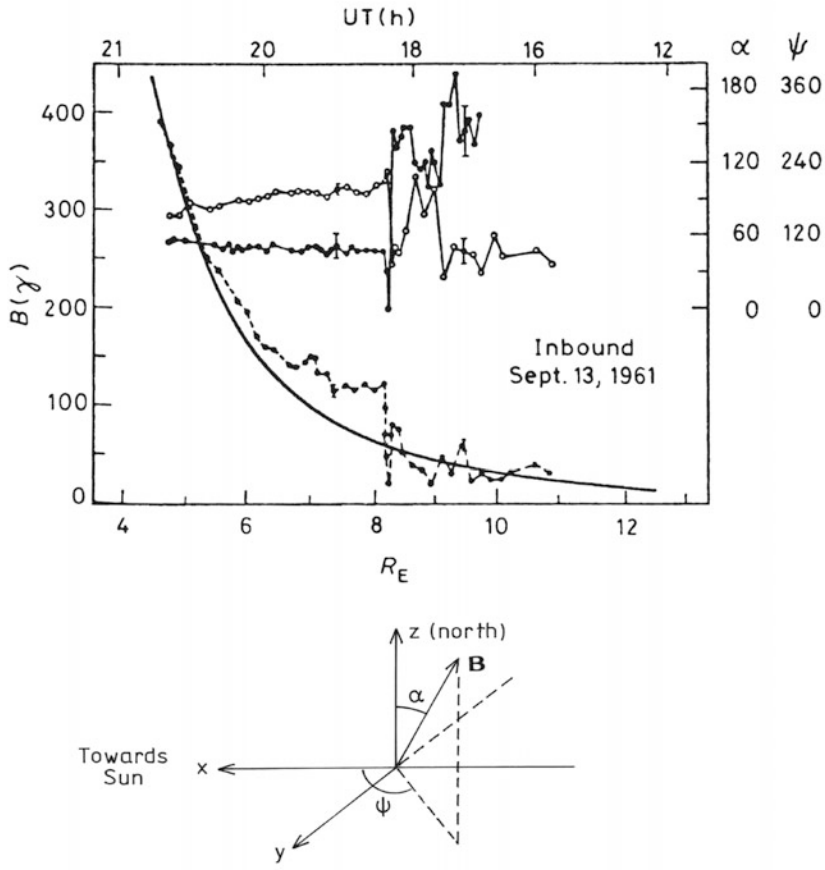


Figure 3.28. Observations of the Earth's magnetic field strength and direction in interplanetary space as derived from Explorer XII. (From Cahill and Amazeen, 1963.) Also shown at the bottom of the figure is the coordinate system to which the above data refer. The z-axis is perpendicular to the ecliptic plane and positive northward. The x-axis is directed toward the Sun and the y-axis completes the Cartesian coordinate system. ψ is the azimuthal angle of **B** in the equatorial plane counted positive toward east. α is the poloidal angle observed positive from the positive z-axis.

Figure 3.28. The field decreases by distance as r^{-3} until about 5 R_e. Between 5 R_e and 8 R_e the field is stronger than expected from the pure dipole field. At about 8 R_e a discontinuity is observed, and further outwards the field is irregular, turbulent, and weak. This is confirmed by observations of the direction of the field which changes abruptly between 8 and 10 R_e. The discontinuity is identified as the magnetopause. On the earthward side is the magnetosphere and on the sunward side is interplanetary space and solar wind. This would, however, be the situation only at the noon meridian in the equatorial plane. If the satellite moved away from the Earth in another meridional plane, it would find the position of the

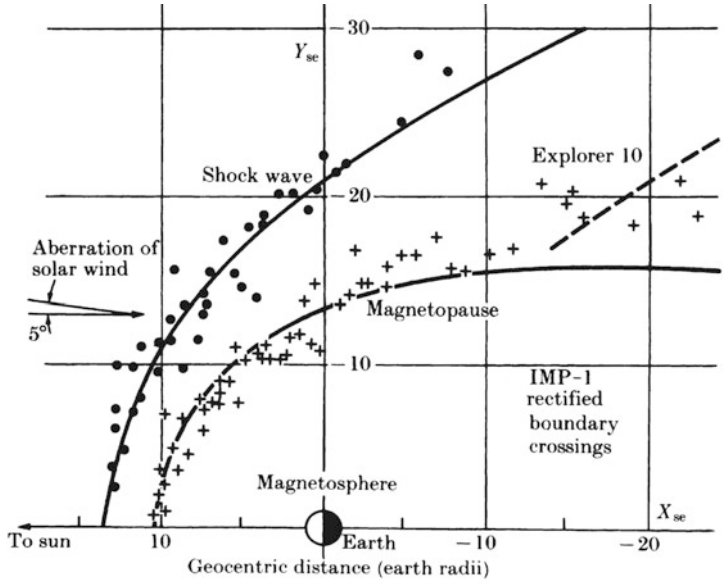


Figure 3.29. Comparison between observed and predicted positions of the magnetopause and shock wave in the ecliptic or x–y plane. (From Ness *et al.*, 1964.)

magnetopause at increasing distances from the Earth the larger the angle the satellite trajectory is making with the Sun–Earth line (Figure 3.29). The magnetopause would be found to take the approximate form of a parabola in the equatorial plane with the Earth close to its focal point.

Outside the magnetopause there is a new discontinuity where turbulence in the field comes to a halt. This is called the shock wave, and outside the shock wave the field is very weak but more regular. The shock wave forms another parabola-shaped curve in the equatorial plane. For other planes making an angle with the equatorial plane, the magnetopause and the shock wave will form similar conic sections, and in total the magnetopause makes an approximate paraboloid surface around the Sun–Earth axis with its focal point close to the Earth (Figure 3.30). The volume inside this paraboloid will be the magnetosphere. The volume between the two paraboloids representing the shock wave and the magnetopause is often called the "magnetosheet".

We also notice from Figure 3.28 that just inside the magnetopause the field is nearly twice as large as the field expected from a dipole field, and immediately outside the discontinuity the field is close to zero.

The same situation would arise if we had a magnetized sphere close to a conducting surface (Figure 3.31(b)). If the magnetized sphere with a dipole moment M were situated in free space (Figure 3.31(a)), the magnetic field at a distance d would be

$$B_d = \frac{\mu_0}{4\pi}\frac{M}{d^3} \tag{3.11}$$

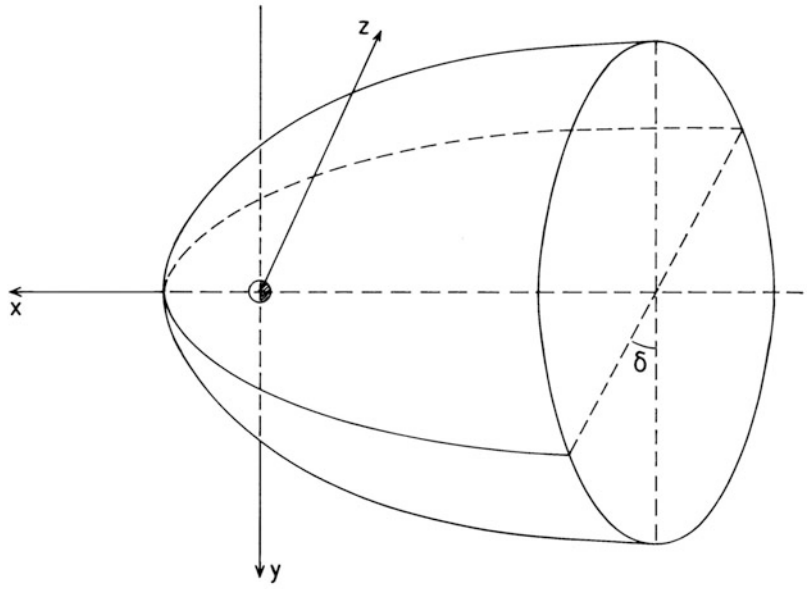


Figure 3.30. A schematic figure illustrating the shape of the magnetospheric paraboloid. Two cross-sections are shown, one in the ecliptic plane and one making an angle δ with the ecliptic plane. The x, y, and z axes are also indicated in agreement with Figure 3.29.

If instead a conducting plate was placed vertically on the magnetic equatorial plane at a distance d, then currents would arise in the plane which would prevent the field from entering behind the plane. This would then enhance the magnetic field on the sphere's side of the plane, and the field strength at the plane would be:

$$B' = \frac{\mu_0}{2\pi}\frac{M}{d^3} \tag{3.12}$$

It is also interesting to notice that from the outside of the plane where $B = 0$, it would appear as if there were two magnetic dipole moments M_0 in the center of the sphere with opposite polarities (see Figure 3.32(a)). From the sphere's side of the plane it would appear that a magnetic dipole moment parallel to M_0 was situated at a distance equal to $2d$ from the center of the sphere. Either point of view gives the same result.

The conducting plane must, according to these considerations, be associated with a current. We notice that due to the direction of M this current must be vertical out of the paper (Figure 3.31(b)) which will be from dawn towards dusk of the dayside of the Earth.

Let the magnetic field at the surface of the Earth due to the Earth's dipole be B_0 (Figure 3.33(b)) then B_d from (3.11) can be expressed as:

$$B_d = \frac{\mu_0}{4\pi}\frac{M_0}{d^3} = B_0\left(\frac{R_e}{d}\right)^3 \tag{3.13}$$

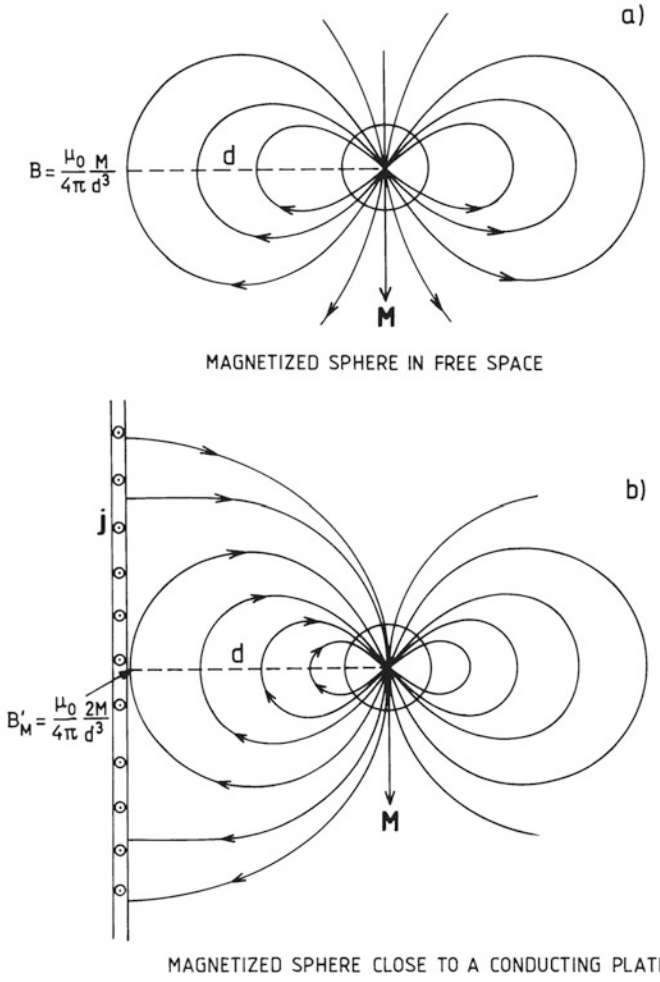


Figure 3.31. (a) A perfect magnetic dipole field of a magnetized sphere with magnetic moment **M** in free space. (b) The magnetic field configuration when a dipole with magnetic moment **M** is placed at a distance d from a conducting plane. The magnetic field will be perpendicular to the plane as if a current of density **j** is flowing in the plane. The magnetic field strength at distance d with a conducting plane present will be twice as great as if the plane was not there.

where

$$\mu_0 = \frac{4\pi}{\mu_0} R_e^3 B_0$$

Due to the current sheet, however, the field at d in (3.12) will be

$$B'_d = 2B_d = 2B_0 \left(\frac{R_e}{d}\right)^3 \tag{3.14}$$

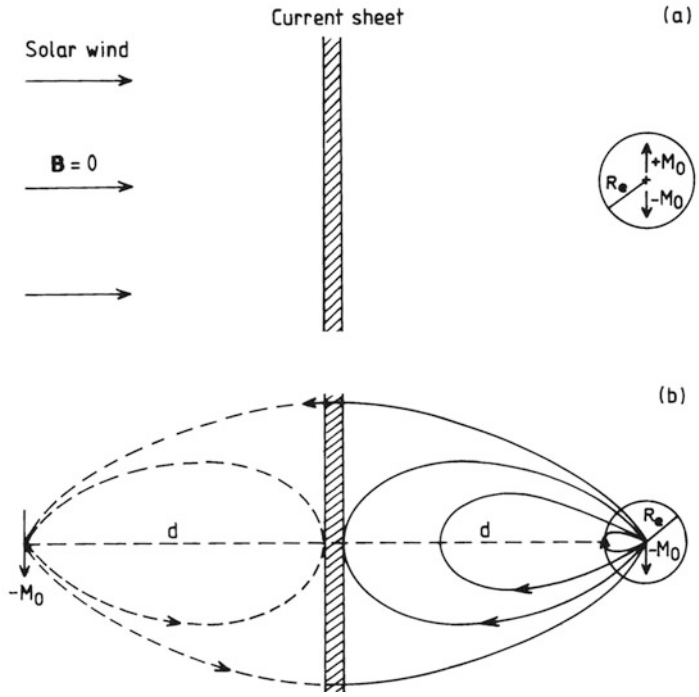


Figure 3.32. (a) The magnetic field of the Earth that would be produced if the magnetosphere was an infinite conducting plane in front of the Earth represented by a perfect dipole moment at its center ($-M_0$). Currents that flow in the plane will produce a field that, seen on the sunward side of the sheet, appears to be due to a dipole ($+M_0$) at the Earth's center, and the total field is zero. (b) On the earthward side, however, the field appears to originate at an image dipole ($-M_0$) at the Earth's distance d on the opposite side of the plane. The field on the sunward side of the plane is zero, while on the earthward side close to the plane it is twice the magnitude it would have without the conducting plane. (After Ratcliffe, 1972.)

Due to the same current sheet the field at the Earth's surface will also be enhanced and

$$B_0' = \frac{\mu_0}{4\pi}\left(\frac{M_0}{R_e^3} + \frac{M_0}{(2d - R_e)^3}\right)$$

where the last term is due to the image dipole (Figure 3.33(b)). The increase in B_0 at the Earth's surface is then

$$\Delta B_0 = B_0' - B_0 = \frac{\mu_0}{4\pi}\frac{M_0}{(2d - R_e)^3}$$

If we take $d = 8\ R_e$ as seen in Figure 3.28, we find that for $M_0 = 7.74 \times 10^{22}\ \mathrm{A\,m^2}$

$$\Delta B_0 \approx 9\ \gamma$$

which is a negligible increase compared with $B_0 \approx 30\,000\ \gamma = 3 \times 10^{-5}$ tesla. The

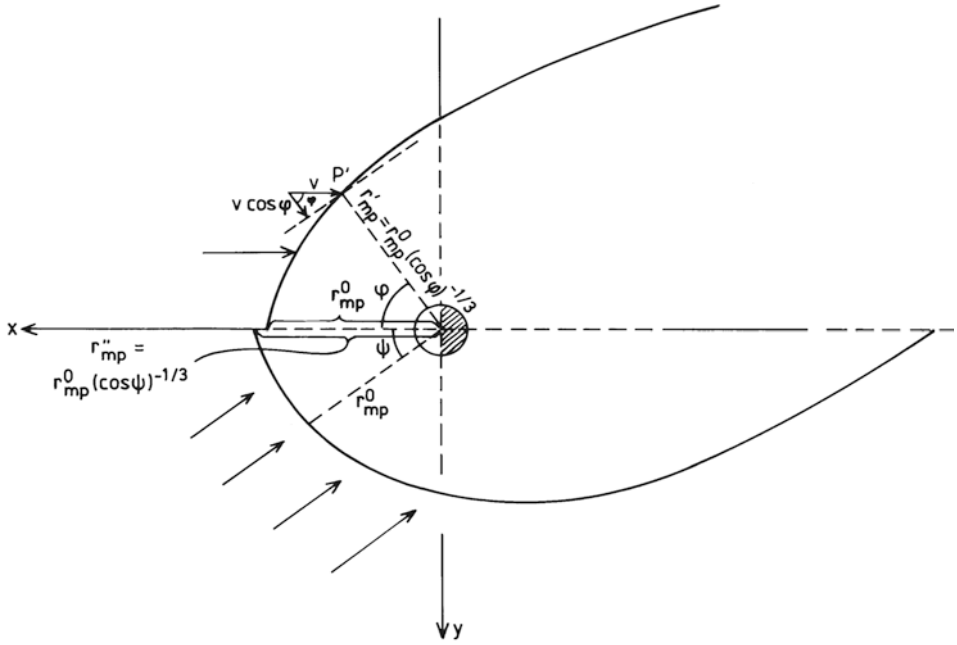


Figure 3.33. A diagram showing the geometry in the ecliptic plane of the magnetopause at the dayside of the Earth for different directions of the solar wind. The upper half shows the solar wind in a head-on situation at the midday meridian. The lower part shows the solar wind blowing at an angle ψ with respect to the midday meridian.

magnetic field strength $B'_d \approx 120\,\gamma$ at $d = 8\,R_e$. If the magnetosphere had occurred at $10\,R_e$ as it does in Figure 3.29, the increase in B_0 at the Earth's surface would be $\approx 4.5\,\gamma$, or almost exactly half the magnitude for the sheet at $8\,R_e$. The field strength at d, however, would be close to $60\,\gamma$ in this case. Under extraordinary conditions in the solar wind the magnetopause has been observed at $4\,R_e$. The expected increase in B_0 at the Earth's surface is then close to $90\,\gamma$.

If the solar wind did not exist, the outer ionosphere would consist of a proton–electron plasma with a concentration determined by diffusion along a magnetic field which could be represented rather well with a dipole field. The ionospheric plasma at the upper strata of the Earth's atmosphere would then merge with the interplanetary plasma at some great distance. Let us for simplicity assume that the Earth with its magnetic field is embedded in a stationary interplanetary plasma with the same density and temperature as the solar wind plasma. The border between the inner region which we could call the "magnetosphere" and the outer region which would be controlled by the interplanetary plasma, would be determined by the balance between the magnetic field pressure of the Earth's magnetic field and the thermal pressure of the interplanetary plasma. When we neglect any magnetic field in the interplanetary plasma and the plasma

density of the ionospheric plasma at this distance, then we have:

$$\frac{(B_m)^2}{2\mu_0} = nk(T_e + T_p)$$

where B_m is the magnetic field strength at this border, n is the plasma density equal for electrons and protons, and T_e and T_p are the electron and proton temperatures respectively, in the interplanetary medium. The magnetic field strength is then given by

$$B_m = [2\mu_0 nk(T_e + T_p)]^{1/2}$$

and since

$$B_m = B_0 \left(\frac{R_e}{r_m}\right)^3$$

the standoff distance r_m to the border between the two regimes is given by:

$$r_m = R_e \cdot B_0^{1/3} \cdot [2\mu_0 nk(T_e + T_p)]^{-1/6}$$

For $n = 5 \times 10^6\ \mathrm{m}^{-3}$, $T_e = 10^5$ K, and $T_p = 4 \times 10^4$ K we find for $B_0 = 3.0 \times 10^{-5}$ tesla

$$r_m = 53\ R_e$$

The extreme limit of a magnetosphere in static equilibrium with an interplanetary thermal plasma would therefore be close to 50 R_e.

Because the solar wind is a highly conducting plasma, the situation is strongly altered. This high conductivity prevents the Earth's magnetic field from penetrating the solar wind plasma, and the solar wind therefore pushes the magnetic field back towards the Earth. Currents are induced in the solar wind plasma as it moves by the Earth, and these currents give rise to new magnetic fields which add to the Earth's dipole field and increase the field at the Earth's surface. Moreover, the forces exerted by the Earth's magnetic field on the induced currents cause the wind to alter direction so that it avoids a region surrounding the Earth. The boundary between the region where the solar wind continues to flow and the region from which it is excluded, is then the magnetopause.

In the case of a head-on solar wind this pause must be situated at a distance where the kinetic pressure of the solar wind must be balanced by the pressure (energy density) of the magnetic field. Let the number density, molecular mass, and velocity of the solar wind be denoted by n, m, and v, respectively, then the kinetic pressure in the solar wind assuming elastic collision and reflections of the particles at the magnetopause is:

$$p_k = 2nmv^2$$

while the pressure of the magnetic field is given by

$$p_m = \frac{B_{mp}^2}{2\mu_0}$$

where B_{mp} is the magnetic field at the magnetopause. We now know that

$$B_{mp} = 2B_p = 2B_0\left(\frac{R_e}{r^0_{mp}}\right)^3 \tag{3.15}$$

and

$$B_p = B_0\left(\frac{R_e}{r^0_{mp}}\right)^3$$

where r^0_{mp} is the distance to the magnetopause along the midday meridian from the Earth's center. When these two pressures are equal ($p_k = p_m$) at the magnetopause, we will have:

$$\left.\begin{aligned} 2mnv^2 &= \frac{(B_{mp})^2}{2\mu_0} = \frac{2B_p^2}{\mu_0} = \frac{2B_0^2}{\mu_0}\left(\frac{R_e}{r^0_{mp}}\right)^6 \\ mnv^2 &= \frac{B_p^2}{\mu_0} \end{aligned}\right\} \tag{3.16}$$

and the distance in R_e to the magnetopause is given by:

$$r^0_{mp} = \left(\frac{B_0^2}{\mu_0 mnv^2}\right)^{1/6} R_e$$

For $v = 400\,\text{km/s}$, $m = m_p$, $n = 5 \times 10^6\ \text{m}^{-3}$, and $B_0 = 3 \times 10^{-5}$ tesla we find:

$$r^0_{mp} = 9 \cdot R_e$$

For $v = 340\,\text{km/s}$ and $570\,\text{km/s}$, r^0_{mp} is equal to $10\,R_e$ and $8\,R_e$ respectively. We notice that for a point P' at the magnetopause at an angle ϕ from the Sun–Earth line, the distance to the magnetopause from the center of the Earth is (see Figure 3.33)

$$r'_{mp} = r^0_{mp}(\cos\phi)^{-1/3}$$

which to a first approximation gives the position of the magnetopause for $\phi < 90°$. On the other hand, if the solar wind is making an angle ψ with the Sun–Earth line, the stagnation point will not be at the midday meridian but at the angle ψ. The distance to the magnetopause at the midday meridian will then be

$$r''_{mp} = r^0_{mp}(\cos\psi)^{-1/3}$$

Since the direction of the wind changes with time, so will the stagnation point.

As already mentioned, the discontinuity in the magnetic field at the magnetopause could be explained by a current sheet perpendicular to the equatorial plane in the dawn–dusk direction. We will consider this current sheet more closely.

Let us take an oversimplified model (Figure 3.34) of protons and electrons with equal concentrations and velocities moving in the negative x-direction towards a uniform magnetic field B perpendicular to their velocities in the positive z-direction. The volume with the magnetic field is divided from the volume

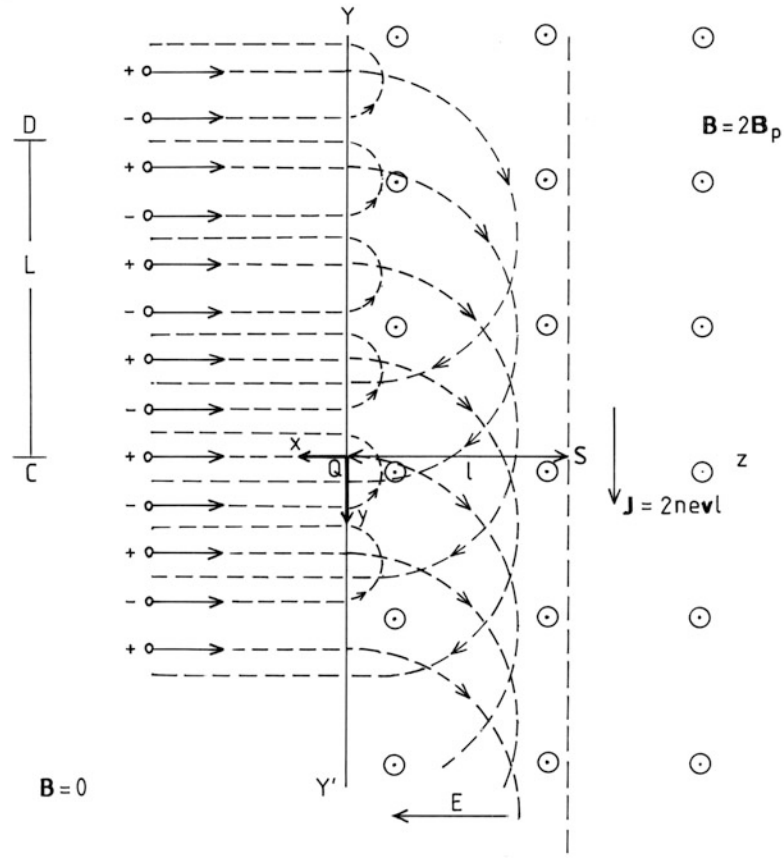


Figure 3.34. A schematic illustration of the trajectories of protons and electrons when they hit the magnetosheet, here represented by a conducting plane YY' in the y–z plane. On the sunward side of the conducting plane the magnetic field is assumed zero and the velocities of the protons and electrons are **v** and perpendicular to the plane. The protons due to their heavier mass will move in a semicircle with diameter L on the earthward side where the field strength B is twice the value represented by the Earth's dipole moment. The electrons, however, will move in smaller circles, and the motion will be in the opposite direction due to the Lorentz force. The protons will reach a distance of $l = L/2$ earthward of the current boundary YY'. The result is a current density in the positive y-direction given by $J = 2nevl$.

without such a field by a plane YY', perpendicular to the x–y plane. A current sheet must then flow in the direction of YY' to produce a magnetic field on the positive x-side which is equal but opposite to the field B_p that would be there without the current sheet present. On the negative x-side the sheet must produce a field that is exactly equal to B_p and the resultant field on this side will be

$$B = 2B_p$$

in the positive z-direction.

For an infinite current sheet producing a uniform magnetic field strength of B_p the current density has to be:

$$J = \frac{2B_p}{\mu_0} \tag{3.17}$$

If we now consider a charged particle that has crossed the boundary and entered the area with a magnetic field of magnitude B_p, it will follow a circular path with a radius given by

$$r = \frac{mv}{eB_p} \tag{3.18}$$

and after moving around half a circle, it will return up against the stream again. Since electrons have much smaller masses than protons, they will be reflected relatively shortly after they have penetrated into the magnetic field. Let us now consider an area perpendicular to the x–y plane at QS. All particles of mass m that originally are found in the solar wind with velocity v between D and C will cross the area QS. Again, because the protons have much larger mass, they move in the B-field with larger radius of curvature. Therefore, there will be far more protons than electrons penetrating the area QS. The flux of charges (protons) through the area QS per unit time is then

$$\phi_p = 2nevl$$

This is then the current density J that creates the extra B-field at the magnetopause, and therefore this current can be expressed as:

$$J = 2nevl = \frac{2nmv^2}{B_p}$$

where we have set l equal to r in (3.18). Equating J here derived with the one given by (3.17), we find that:

$$\frac{2B_p}{\mu_0} = \frac{2nmv^2}{B_p}$$

and

$$\frac{B_p^2}{\mu_0} = nmv^2$$

which gives us the balance between the magnetic and kinetic pressures we originally started out with (3.16).

We find that the thickness of the current sheet from this simple model is

$$l = \frac{mv}{eB_p}$$

For $m = m_p$, $v = 400\,\text{km/s}$ and $B_p = 117\,\gamma$ at $r_p = 8\,R_e$, the thickness of the sheet becomes

$$l \approx 35\,\text{km}$$

an extremely thin layer on the astronomical scale.

Because the electrons penetrate a much shorter distance into the magnetic field before they are reflected, there must be an electric field in the positive x-direction which will modify the simple model just described. This E-field will be associated with an $\mathbf{E} \times \mathbf{B}$ motion of the plasma which actually will be opposite to the proton motion but parallel to the electron motion.

3.10 THE MAGNETIC TAIL

If we now consider the magnetic field on the antisunward side of the Earth, we find the magnetopause to be situated at longer and longer distances from Earth the closer we get to the midnight meridian. Observations with the IMP-1 satellite (Figure 3.35) of the direction of the field indicate that it is directed in the antisunward direction under the equatorial plane, for $z < 0$. A model of the magnetic

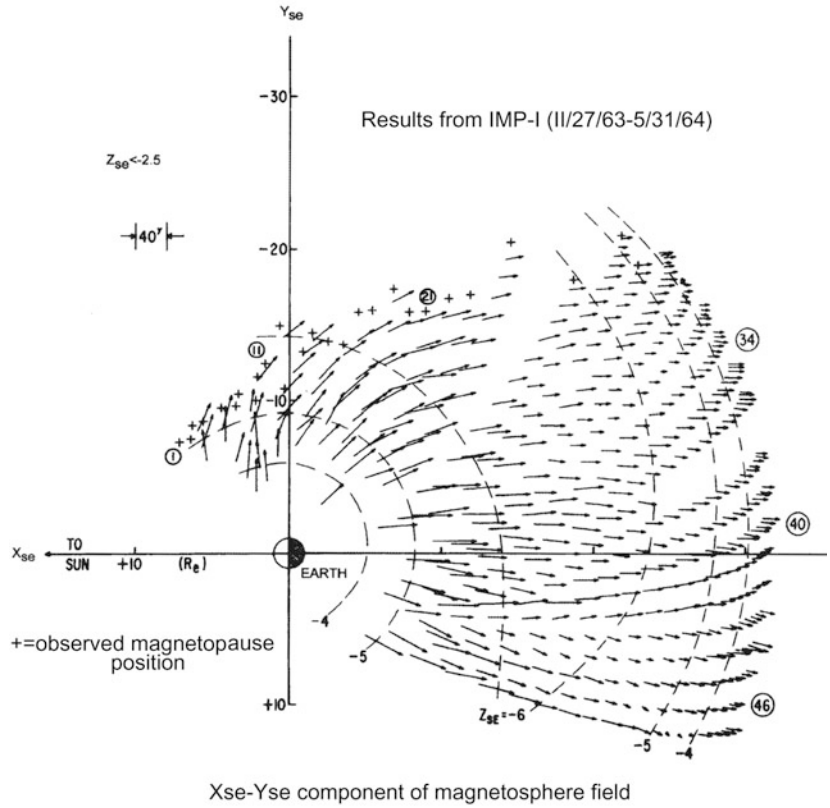


Figure 3.35. Magnetic field vectors in planes parallel to and below the ecliptic plane. The Z_{SE} value indicates the distance in Earth radii below the ecliptic plane. The arrows represent the magnetic field vectors. The scale of the vectors is indicated by the bar. (From Ness, 1965.)

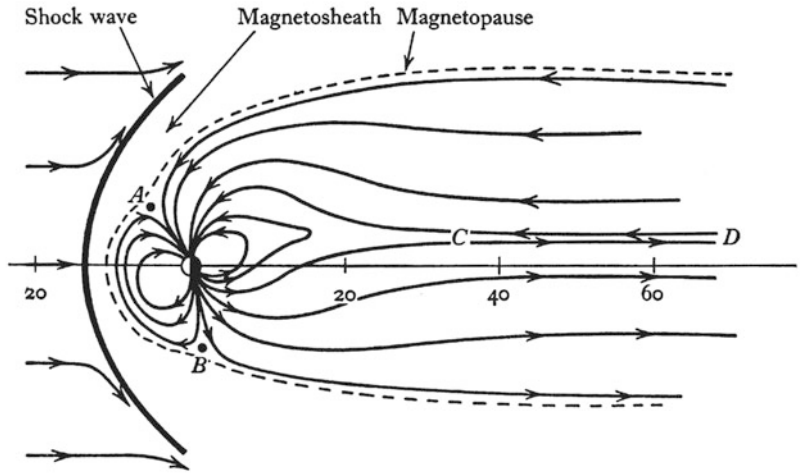


Figure 3.36. A midday–midnight cross-section of the magnetosphere with the magnetic field directions indicated. Also illustrated is the solar wind and the presence of the magnetosheet, magnetopause, and the magnetoshock wave. There are neutral points at A and B and a neutral sheet between C and D. The approximate geocentric distances in Earth radii are also indicated. (From Hess, 1967.)

field configuration in the midday–midnight meridian plane is shown in Figure 3.36. At distances larger than 20 R_e on the nightside the field is nearly parallel or antiparallel to the Sun–Earth line depending on whether the field line is below or above the equatorial plane. To sustain such a magnetic field configuration there must be a current in the equatorial plane directed from dawn to dusk across the plane (Figure 3.37(a)). For a field strength of about 20 γ as seen by IMP-1 at about 30 R_e, the current density in the sheet must be:

$$J = \frac{2B}{\mu_0} \approx 3 \times 10^{-2}\,\mathrm{A/m}$$

For a current sheet between C and D in Figure 3.36 approximately 10 R_e wide, the total current will be close to 10^6 A. There is therefore a current in the equatorial plane of the order of 10^6 A to sustain the field configuration close to the equatorial plane.

The magnetic field at the far nightside of the Earth is split into two parts as already mentioned, one above the equatorial plane where the field points towards the Earth and one below this plane where the field points away from the Earth (Figure 3.37(b)). This region of the magnetosphere inside the magnetopause is called the "magnetospheric tail". We notice that all field lines stretching out in the far tail below the equatorial plane emerge from the southern polar cap, and similarly the field lines directed parallel toward the Sun above the equatorial plane all penetrate the northern polar cap.

It is possible to estimate the cross-section of the magnetic tail when assuming a steady state in which the solar wind kinetic and magnetic pressures are balanced

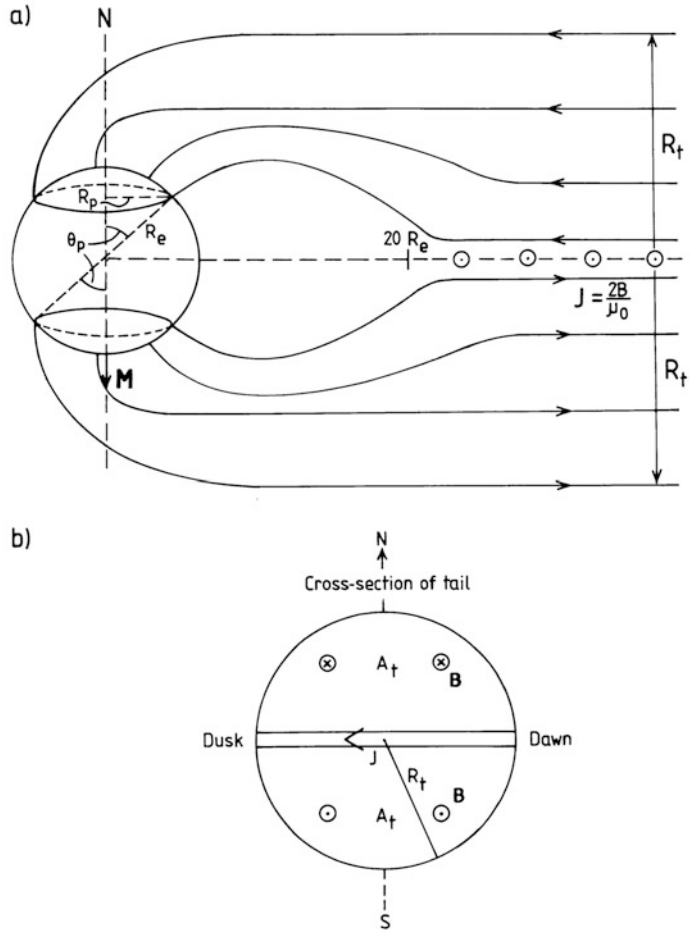


Figure 3.37. Simplified diagrams showing the geometry of the magnetic field lines: (a) along the day–night meridian in the tail and their connection to the polar cap. (b) The tail is assumed to be an approximate cylinder of radius R_t, and divided by a cross-tail current sheet. The distance from the Earth to the cylindrical tail is set to 20 R_e. The colatitude of the polar cap boundary is indicated by θ_p. The cross-section of the tail shows a cross-tail current from dawn to dusk of density $J = 2B/\mu_0$, where B is magnetic induction above and below this sheet current.

at the magnetopause. The plasma density in the magnetotail is small except for the regions of the cross-tail current and the plasma sheet—the kinetic pressure in the tail is therefore small. The pressure balance equation can then approximately be written as

$$\frac{B_t^2}{2\mu_0} = \frac{B_s^2}{2\mu_0} + n_s k(T_p + T_e)_s$$

and

$$B_t = [B_s^2 + 2\mu_0 n_s k(T_p + T_e)_s]^{1/2}$$

where B_t is the magnetic field in the tail at the magnetopause, B_s and n_s are the magnetic field and particle density in the solar wind, respectively, and $(T_e + T_p)_s$ is the sum of the electron and proton temperatures in the solar wind plasma. For solar wind temperatures, $T_e = 10^6$ K and $T_p = 4 \times 10^5$ K, magnetic field $B_s = 3$ nanotesla, and particle density $n_s = 5 \times 10^6$ m^{-3}, we find that the magnetic field in the magnetopause will be:

$$B_t = 6 \text{ nanotesla}$$

Assuming then that the magnetic field in the upper half of the magnetic tail is uniform and that all field lines merge in the polar cap region within a colatitude θ_p (Figure 3.37(a)), we find that for magnetic flux conservation

$$\phi = B_p \cdot A_p = B_t \cdot A_t$$

where B_p is the polar cap magnetic field strength assumed uniform and vertical on the polar cap area A_p, and B_t is the magnetotail magnetic field strength assumed uniform perpendicular to the upper cross-tail area A_t. If R_t is the radius of the tail cross-section, then

$$A_t = \tfrac{1}{2}\pi R_t^2$$

and

$$A_p = \pi R_e^2 \sin^2 \theta_p$$

The cross-tail radius is then given by:

$$R_t = (2B_p)^{1/2} \cdot (B_s^2 + 2\mu_0 n_s k (T_e + T_p)_s)^{-1/4} \sin \theta_p \cdot R_e$$

We notice that R_t will decrease when either the solar wind temperature or the solar wind magnetic field increases. With the values used so far for the solar wind parameters and assuming $\theta_p = 15^\circ$, we find for $B_p = 6.0 \times 10^{-5}$ tesla

$$R_t = 47\ R_e$$

the magnetotail radius is approximately of the order of 50 R_e.

3.11 MAGNETIC FIELD MERGING

The situation described here is obviously not a stationary one since the magnetic field lines are attached to the Earth at least by one end and are forced to rotate by the Earth. The field lines on the dayside which are compressed by the solar wind are stretched out on the nightside by the same wind, and actually torn apart. Before many of these field lines return to the dayside they have to reconnect or merge again.

Since the magnetic field lines are flowing in the opposite directions in the regions very close to the central current sheet, it is believed that magnetic merging can take place here through reconnection.

The plasma immediately above and below the current sheet is believed to be highly conductive. In the merging process the plasma will therefore be transported toward the current sheet from above and below (see Figure 3.38(a)). When the field lines finally reconnect, plasma is jetted partly towards the Earth and partly outward through the tail (Figure 3.38(b)).

The only way this merging can take place is when the ''frozen-in'' concept is invalidated in a small region around the neutral line. In Figure 3.38(b) the field lines entering the merging region are numbered 1 and 2 and the double arrows indicate the motion of the plasma attached to the field lines. At the crossing the field lines ''cut'' and reconnect in a crosswise fashion so that two field lines (indicated 3 and 4) appear. As these move away from the merging region earthward and tailward, respectively, they carry the plasma along. We notice that there is no magnetic field at the crossing point (the neutral line). It is expected that plasma attached here to field line 3 may be accelerated to considerable energies and transported in a subsequent flow along the magnetic field lines to the lower ionosphere. Similarly the plasma attached to field line 4 may be accelerated and jetted out into interplanetary space. In this process the energy comes from the current which ceases when the neutral sheet disappears.

By cutting field lines we only mean that they can no longer be identified with the plasma motion since the flux through any loop following the plasma motion is not conserved in the vicinity of the neutral line.

It is believed that charged particles forced out of the merging region along the magnetic field lines toward the poles are the source of the polar auroral displays (Figure 3.39(a)). Due to symmetry in the Earth's magnetic field, simultaneous auroral forms may occur in the two hemispheres, aurora borealis and aurora australis (Figures 3.39(a) and 3.40(a)). Due to the influence of the IMF these forms are not always symmetric.

The auroral displays are formed when the energetic plasma (mostly electrons) approaches the Earth and reacts, as a result of collisions, with the neutral particles of the upper atmosphere, mainly above 100 km. This light phenomenon is dominated by emissions from neutral species such as atomic oxygen and molecular nitrogen.

Another light phenomenon in the Sun–Earth interaction process that is strongly related to the aurora is the solar flare. This is often seen as two parallel luminous bands (Figure 3.40(b)) dominated by hydrogen emissions. These bands are often found to be situated between two magnetic poles of opposite polarities (Figure 3.39(b)). As in the Earth's magnetosphere, field line merging occurs in a region above the magnetic poles, and energetic plasma is forced along the magnetic field lines toward the solar surface at the magnetic polar regions. When this plasma reaches the solar atmosphere which is so rich in hydrogen, collisions occur and hydrogen emissions are created which can be observed as parallel bands from the ground. Due to these similarities between the polar flares and the auroras we can claim they are sister phenomena, but there is even a mother–daughter relationship.

Some of the plasma in the merging region is, however, forced outwards into

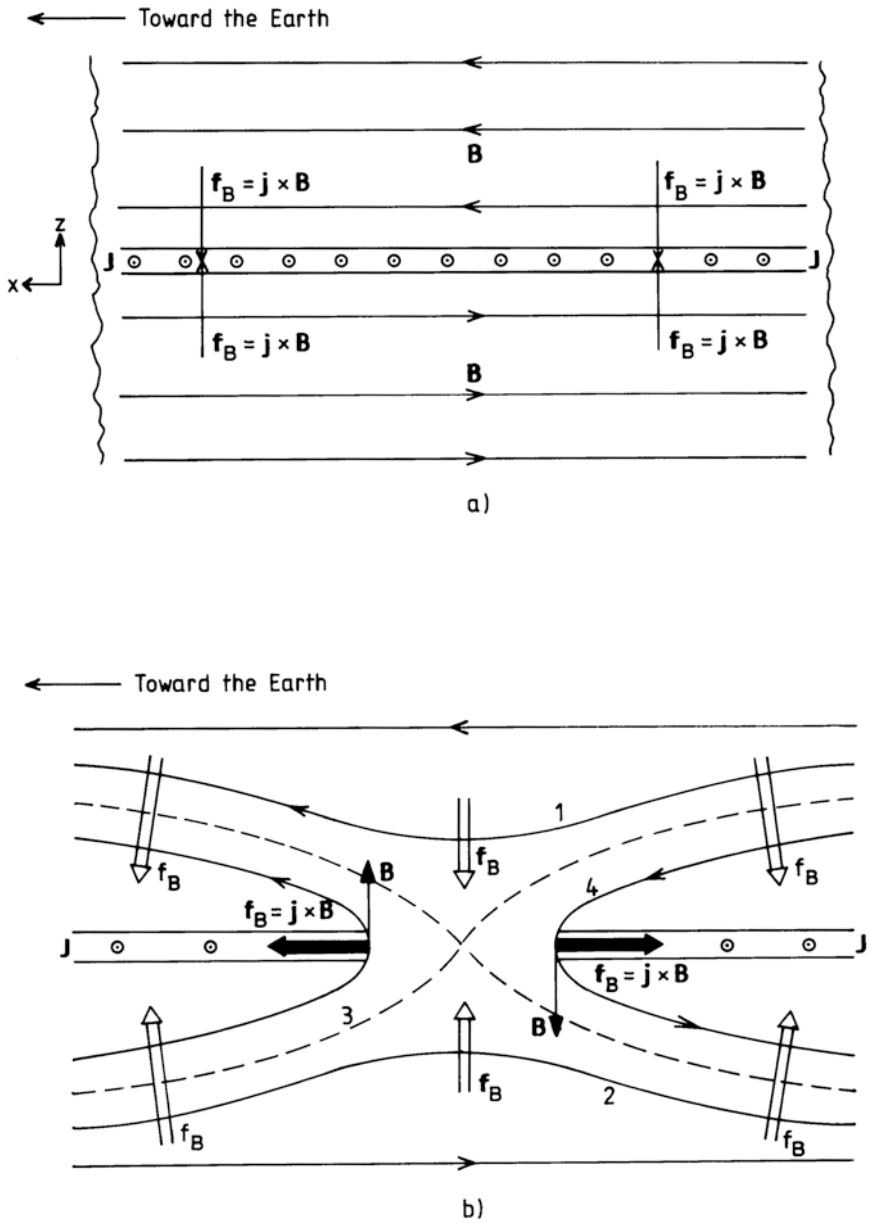


Figure 3.38. (a) A cross-section of part of the cylindrical magnetospheric tail in the noon–midnight meridian illustrating the necessary current sheet **j** in the ecliptic plane, to maintain the magnetic field lines antiparallel above and below this sheet. The magnetic force $\mathbf{f}_B = \mathbf{j} \times \mathbf{B}$ on the plasma is directed toward this sheet from above and below the ecliptic plane. (b) The magnetic field configuration together with the magnetic force is illustrated in relation to a magnetic reconnection (merging) region. The open arrows bring plasma towards, while the filled arrows take plasma from the merging region.

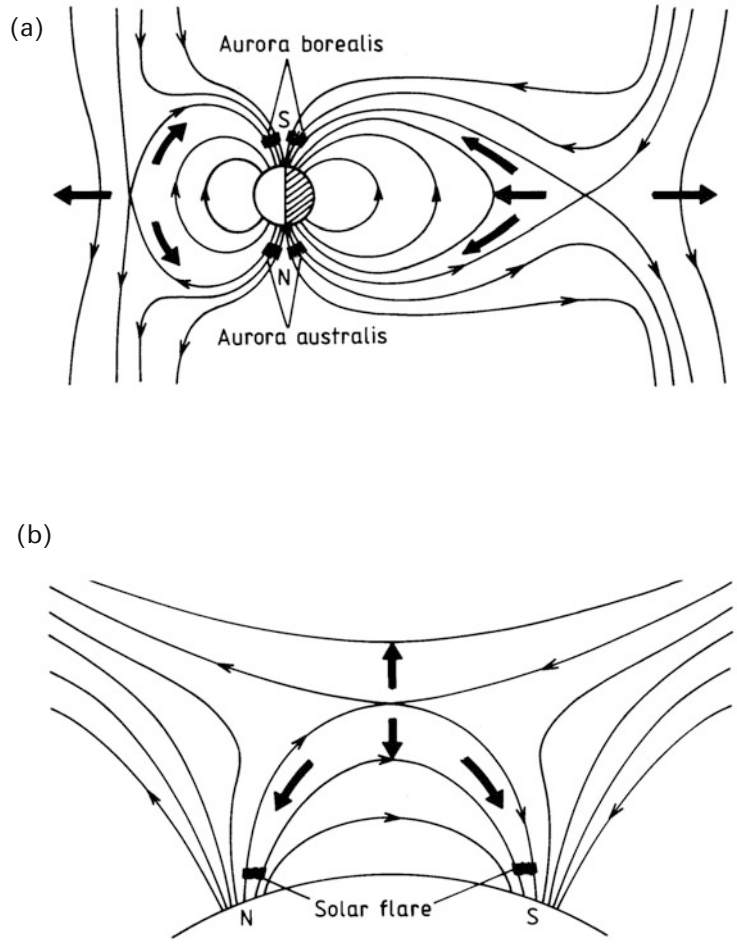


Figure 3.39. Schematic illustration of magnetic field merging processes taking place in the Earth's magnetosphere creating polar auroras (a) (after Dungey, 1961), and in the solar atmosphere creating solar flares (b). The thick arrows indicate the plasma flow away from the merging region. The plasma causing light emissions propagate along magnetic field lines towards the magnetic poles. Light emissions occur when the atmospheric density of the Earth or the Sun is high enough that frequent collisions occur between the streaming plasma and atmospheric particles.

interplanetary space, and part of it may reach the Earth's magnetosphere. Here it can evidently be forced into a new merging process creating the aurora. In this sense we notice that the plasma processes taking place in the near-Earth environment are almost universal. By learning to understand the merging process in the magnetosphere, we will also be able to understand similar processes at the Sun, the stars, and at other celestial bodies in the Universe.

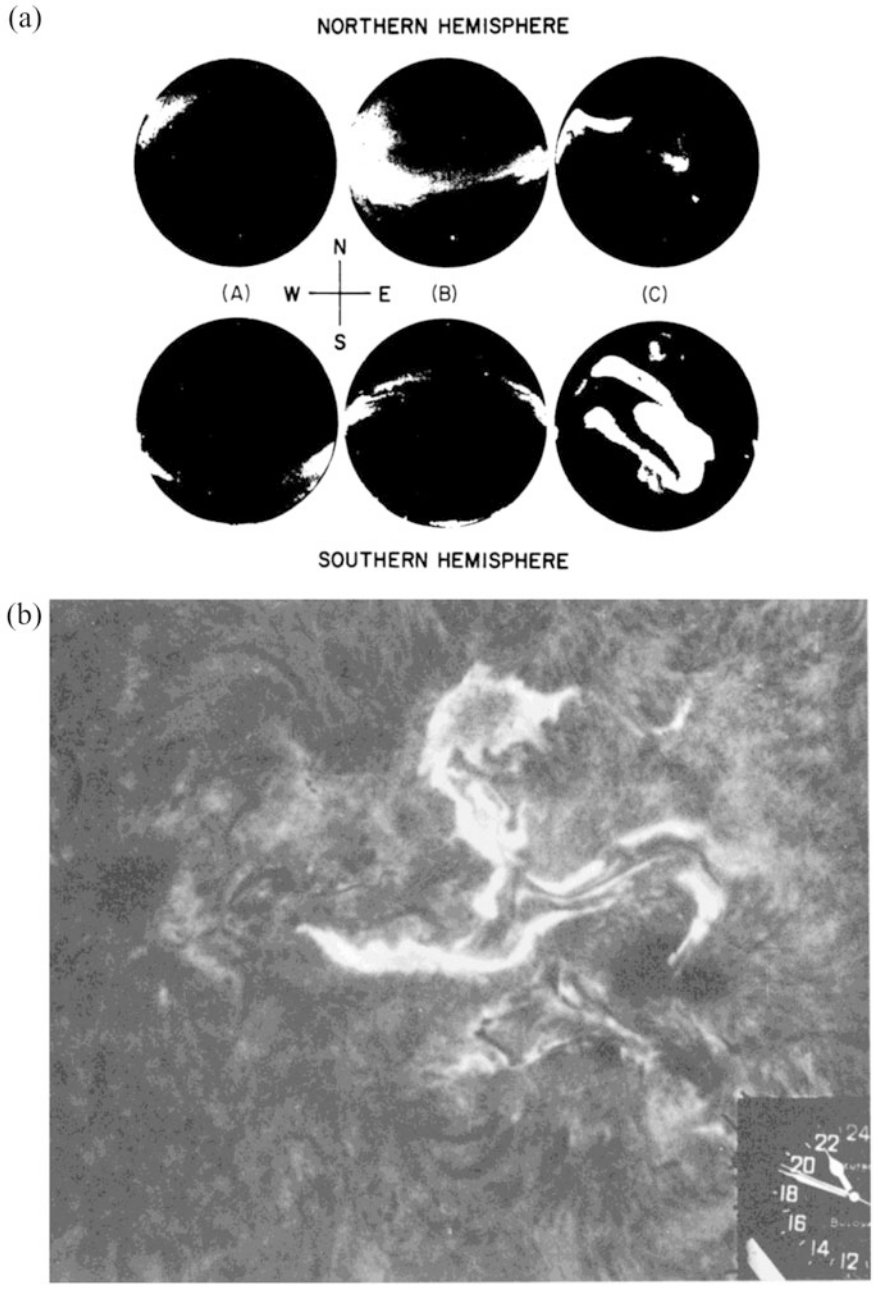


Figure 3.40. (a) Simultaneous observations of the aurora borealis and the aurora australis at approximately the same magnetic field lines, showing the strong similarity between these two polar light phenomena. (b) Solar flare event observed from ground as two approximate parallel luminous bands of H_α emissions. (From Akasofu and Chapman, 1972.)

3.12 EFFECTS OF THE MAGNETIC FORCE

To describe the magnetohydrodynamic behavior of plasma we need the continuity equation, the mobility equation, Maxwell's equations, and the generalized Ohm's law together with an equation of state.

In the following we will consider the mobility equation and the importance of the magnetic force. The mobility equation is basically the same as for hydrodynamics except that we must include forces that can act on the plasma due to the electric currents flowing there

$$\rho \frac{d\mathbf{v}}{dt} = -\nabla p + \mathbf{j} \times \mathbf{B} = \mathbf{f}_p + \mathbf{f}_B \tag{3.19}$$

where p is the pressure, $\mathbf{j}$ is the current density, and $\mathbf{B}$ the magnetic field. Here we have neglected gravity and viscosity.

The motion of the plasma is due either to the pressure force

$$\mathbf{f}_p = -\nabla p$$

or the magnetic force (Lorentz force)

$$\mathbf{f}_B = \mathbf{j} \times \mathbf{B}$$

We can eliminate $\mathbf{j}$ from (3.19) with help from Maxwell's equation:

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j} \tag{3.20}$$

when neglecting the displacement current. Then

$$\mathbf{f}_B = \frac{1}{\mu_0} (\nabla \times \mathbf{B}) \times \mathbf{B}$$

By using the vector identity

$$\nabla B^2 = 2(\mathbf{B} \cdot \nabla)\mathbf{B} + 2\mathbf{B} \times (\nabla \times \mathbf{B}) \tag{3.21}$$

$\mathbf{f}_B$ can be expressed as:

$$\mathbf{f}_B = \frac{1}{\mu_0} (\nabla \times \mathbf{B}) \times \mathbf{B} = -\nabla \frac{B^2}{2\mu_0} + \frac{1}{\mu_0} (\mathbf{B} \cdot \nabla)\mathbf{B}$$

The mobility equation (3.19) can now be written as:

$$\rho \frac{d\mathbf{v}}{dt} = -\nabla \left(p + \frac{B^2}{2\mu_0} \right) + \frac{1}{\mu_0} (\mathbf{B} \cdot \nabla)\mathbf{B} \tag{3.22}$$

We notice that in the mobility equation there are two terms related to the magnetic field $\mathbf{B}$ both of which, however, are due to the magnetic force $\mathbf{f}_B$. The first term $(1/2\mu_0)\nabla B^2$, which is included in the pressure term, is often analogously mentioned as the magnetic pressure, while the second term given by $(1/\mu_0)(\mathbf{B} \cdot \nabla)\mathbf{B}$ is called magnetic tension. Therefore, any deviation of $\mathbf{B}$ from the force-free case ($\mathbf{j} = 0$) must be balanced by fluid pressure when the fluid is in equilibrium. In a perfect dipole field (Figure 3.41(a)), where $\mathbf{j} = 0$ outside the Earth's

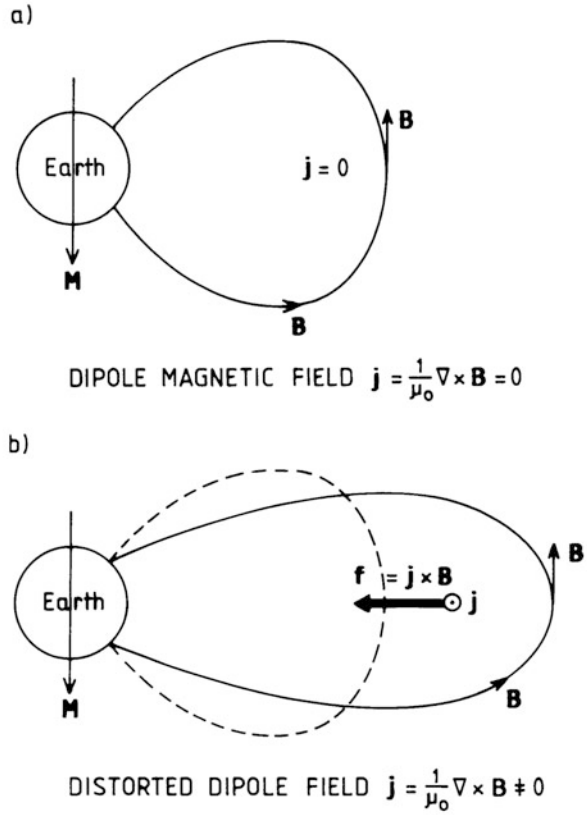


Figure 3.41. (a) An idealized magnetic field line from a dipole at the center of the Earth, when no current is flowing between the Earth and the field line. (b) A stretched magnetic field line when a current is present outside the Earth, creating a magnetic force toward the Earth, trying to restore the dipole configuration.

core, the magnetic forces balance the pressure force everywhere. In a distorted dipole field, such as will occur in the magnetic tail and elsewhere in the magnetosphere (Figures 3.41(a) and 3.38), there must be magnetic forces that will in the case of Figure 3.41(b) force plasma back towards the Earth. In order to stretch the magnetic field in such a way that $\nabla \times \mathbf{B} \neq 0$, there must be a current out of the plane in Figure 3.41(b) inside the magnetic field line. The force due to this current $\mathbf{f}_B = \mathbf{j} \times \mathbf{B}$ will be directed toward the Earth, since $\mathbf{B}$ is pointing vertically upward (northward). This force is equivalent to the magnetic pressure and tension forces, and therefore in a steady state there is a pressure balancing the magnetic forces and directed outward from the Earth due to a pressure gradient directed toward the Earth.

By inspecting the situation in the equatorial tail of the magnetosphere where magnetic field lines are antiparallel above and below the neutral sheet (Figure

3.38(a)), we notice that a current must flow across this sheet perpendicular to **B** and out of the paper. The magnetic force acting on the plasma ($\mathbf{f}_B = \mathbf{j} \times \mathbf{B}$) will be directed towards the sheet both from above and below. If merging now occurs in the current sheet, then the magnetic force will still be directed toward the sheet away from the merging region. Because the magnetic field will be perpendicular to the neutral sheet within the merging region (Figure 3.38(b)), however, the magnetic force will drive the plasma away from the merging region towards the Earth on the earthward side and into planetary space on the tailward side.

Since we are assuming that the magnetospheric and solar wind plasma is collisionless and that the "frozen-in" concept of the magnetic field applies due to the large conductivities along the field lines, one may ask how these regions can act as current sources for currents perpendicular to the B-field at all. In order to obtain a better insight into that, we must go back to the mobility equation (3.19) to find the current perpendicular to **B** and form the cross-product with **B**

$$\rho \frac{d\mathbf{v}}{dt} \times \mathbf{B} = -\nabla p \times \mathbf{B} + (\mathbf{j} \times \mathbf{B}) \times \mathbf{B}$$

$$= -\nabla p \times \mathbf{B} - B^2 \mathbf{j}_\perp$$

and

$$\mathbf{j}_\perp = \frac{1}{B^2} \left[\mathbf{B} \times \left(\rho \frac{d\mathbf{v}}{dt} + \nabla p \right) \right] \tag{3.23}$$

Currents in the magnetosphere and the solar wind can therefore occur perpendicularly to the magnetic field whenever there are spatial variations in the pressure or time variations in the plasma flow itself.

3.13 ENERGY FLUX INTO THE MAGNETOSPHERE

There is another important result that emerges from the magnetohydrodynamic theory, which is of great importance to the understanding of the physical processes in the magnetosphere, and that is the relationship between thermal and electromagnetic energy. For a magnetic field B in a vacuum there is always associated an energy density given by

$$u_B = \frac{1}{2\mu_0} B^2$$

or identical to the associated magnetic pressure. Within a volume τ there is therefore a total energy

$$U_B = \frac{1}{2\mu_0} \int_\tau B^2 \, d\tau$$

where $d\tau$ is an infinitesimal volume element. The time derivative of this energy is given by:

$$\frac{\partial U_B}{\partial t} = \frac{1}{\mu_0} \int_\tau \left(\mathbf{B} \cdot \frac{\partial \mathbf{B}}{\partial t} \right) d\tau$$

Now using Maxwell's equation:

$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E}$$

and the relation

$$\mathbf{B} = \mu_0 \mathbf{H}$$

where $\mathbf{H}$ is the magnetic intensity, we find:

$$\frac{\partial U_B}{\partial t} = -\int_\tau \mathbf{H} \cdot (\nabla \times \mathbf{E})\, d\tau$$

Introducing the vector identity:

$$\nabla \cdot (\mathbf{E} \times \mathbf{H}) = \mathbf{H} \cdot (\nabla \times \mathbf{E}) - \mathbf{E} \cdot (\nabla \times \mathbf{H})$$

we get

$$\frac{\partial U_B}{\partial t} = -\int_\tau \nabla \cdot (\mathbf{E} \times \mathbf{H})\, d\tau - \int_\tau \mathbf{E} \cdot (\nabla \times \mathbf{H})\, d\tau$$

Since

$$\int_\tau \nabla \cdot (\mathbf{E} \times \mathbf{H})\, d\tau = \int_S (\mathbf{E} \times \mathbf{H})\, d\mathbf{s}$$

and

$$\mathbf{j} = \nabla \times \mathbf{H}$$

where $\mathbf{s}$ is an area enclosing and pointing out from τ

$$\frac{\partial U_B}{\partial t} = -\int_S (\mathbf{E} \times \mathbf{H})\, d\mathbf{s} - \int_\tau \mathbf{j} \cdot \mathbf{E}\, d\tau$$

The time rate of the magnetic energy in a volume τ is equal to the energy flux across the surface into the volume as expressed by the Poynting vector $\mathbf{E} \times \mathbf{H}$ minus the electromechanical work $\mathbf{j} \cdot \mathbf{E}$ done in the volume.

This last term can be specified into two separate terms by using Ohm's law. Since

$$\mathbf{E} = \frac{1}{\sigma}\mathbf{j} - \mathbf{v} \times \mathbf{B}$$

$$\mathbf{j} \cdot \mathbf{E} = \frac{1}{\sigma} j^2 - (\mathbf{v} \times \mathbf{B}) \cdot \mathbf{j}$$

and

$$\frac{\partial U_B}{\partial t} = -\int_S (\mathbf{E} \times \mathbf{H})\, d\mathbf{s} - \int_\tau \frac{j^2}{\sigma} d\tau - \int_\tau \mathbf{v} \cdot (\mathbf{j} \times \mathbf{B})\, d\tau \tag{3.24}$$

The second term represents the resistive energy loss inside the volume (the electrical load or the Joule heating rate) while the last term represents the mechanical work done per unit time due to plasma flow against the magnetic force $\mathbf{j} \times \mathbf{B}$ inside the same volume.

3.14 SOME ASPECTS OF THE ENERGY BALANCE

If the Earth was situated in a vacuum, the magnetic field would reach out to infinity without being distorted. Outside the Earth in interplanetary space there is a rarefied plasma penetrated by a magnetic field of solar and cosmic origin and known as the interplanetary magnetic field (IMF). Depending on the magnitude and direction of this magnetic field and the dynamic behavior of the interplanetary plasma, the Earth's magnetic field will be distorted accordingly. Let us first assume that the interplanetary field, B_z, is directed northward (i.e., perpendicular to the ecliptic plane and parallel to the Earth's magnetic field—Figure 3.42(a)). In this situation the magnetic field of the Earth would be embedded in the interplanetary field to form a closed magnetospheric cavity indicated by the dashed line in Figure 3.42(a). The surface of this cavity would form an equipotential. Since **E** is everywhere perpendicular to this surface and **B** is parallel to it, the Poynting vector would be parallel to the surface everywhere and no energy would enter into the magnetospheric cavity except maybe for the two neutral points marked N and N' in Figure 3.42a.

If, on the other hand, the interplanetary magnetic field was directed southward, the magnetic field of the Earth would open up. A neutral line would form around the equator in a complete symmetric situation. **B** is perpendicular to this line, and the electric field would be directed along it. The Poynting flux would therefore point into the magnetosphere and energy could enter.

The situation is, however, much more complicated since the solar wind pushes toward the Earth's magnetic field and deforms it by compression on the dayside and by elongation into a tail on the nightside. Let us now consider a simplified cylindrical model of the surface of the tail (Figure 3.43) when there is a southward IMF ($\mathbf{B}_{\mathrm{IMF}}$). The magnetic field in the tail will be directed toward the Earth in the northern hemisphere and against the solar wind stream. The field will make a small angle α with the surface in the northern (upper) half of the tail and far away from the Earth, and also make a similar but negative angle α with the surface in the southern half of the tail in order to merge with the southward IMF. In the solar wind we know that conductivity is very high so the "frozen-in" concept applies

$$\mathbf{E} = -\mathbf{v}_S \times \mathbf{B}$$

where $\mathbf{v}_S$ is the solar wind velocity, and therefore the energy flux (Poynting vector) can be expressed as:

$$\mathbf{P} = \mathbf{E} \times \mathbf{H} = -(\mathbf{v}_S \times \mathbf{B}) \times \mathbf{H} = -(\mathbf{v}_S \cdot \mathbf{H})\mathbf{B} + (\mathbf{B}\cdot\mathbf{H})\mathbf{v}_S$$

The energy input to the magnetosphere is therefore given by:

$$\phi_E = -\int_A \mathbf{E} \times \mathbf{H}\, d\mathbf{s} = \int_A (\mathbf{v}_S \cdot \mathbf{H})\mathbf{B}\, d\mathbf{s} - \int_A (\mathbf{B}\cdot\mathbf{H})\mathbf{v}_S\, d\mathbf{s}$$

where A is the surface, and $d\mathbf{s}$ a surface element of the magnetospheric tail. Since the solar wind is along the magnetospheric surface, $\mathbf{v}_S$ is perpendicular to $d\mathbf{s}$

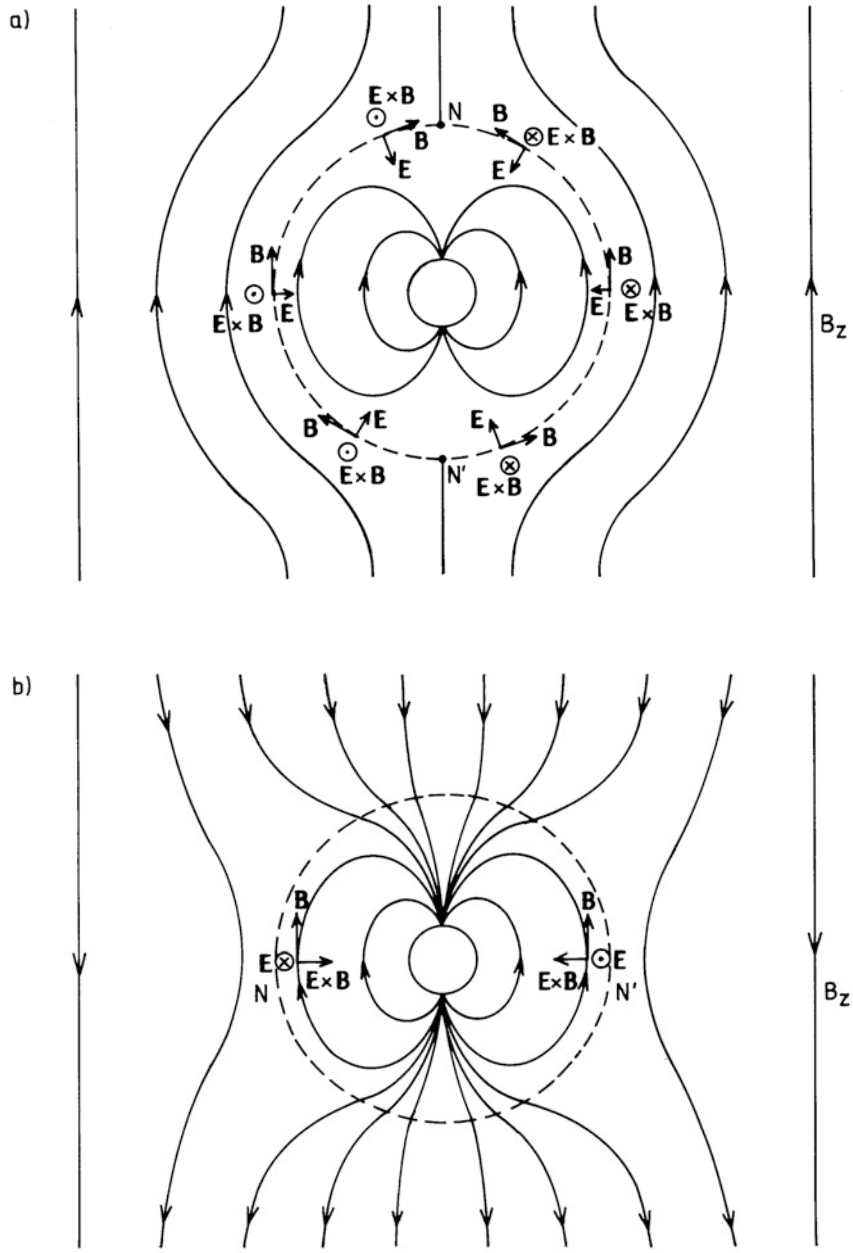


Figure 3.42. (a) The Earth as an idealized magnetic sphere in an area with a northward-directed interplanetary magnetic field ($B_z > 0$), making a closed magnetospheric cavity. The energy flux ($\mathbf{E} \times \mathbf{B}$) is along the cavity surface everywhere. No energy enters the magnetosphere from the outside. (b) The similar situation when there is a southward IMF ($B_z < 0$). The magnetosphere is opened up in the polar regions, and the energy flux is directed inward around the equatorial perimeter of the magnetospheric cavity.

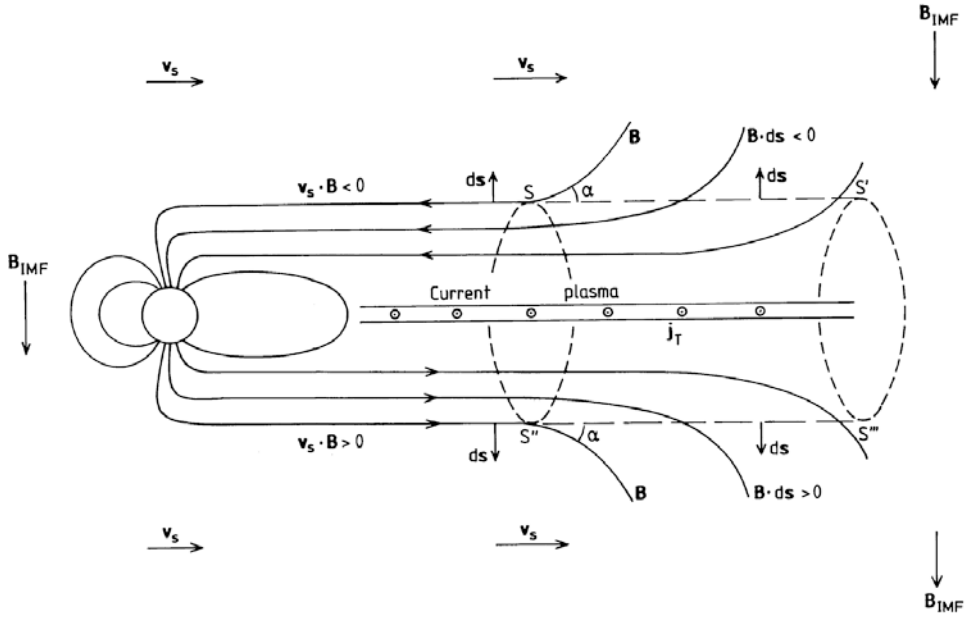


Figure 3.43. An illustration of the interaction between the solar wind and the magnetospheric tail when there is a negative IMF ($\mathbf{B}_{\mathrm{IMF}}$). The solar wind velocity $\mathbf{v}_S$ is parallel and antiparallel to **B** on the southern and northern part of the semicylindrical tail lobes, respectively. The surface element $d\mathbf{s}$ is parallel or antiparallel to **B** on the same tail lobes, respectively. This result is an energy flux inwards to the tail surface everywhere.

everywhere. The last term therefore disappears. Energy flux into the magnetosphere is then given by:

$$\phi_E = + \int_A (\mathbf{v}_S \cdot \mathbf{H}) \mathbf{B}\, d\mathbf{s} \tag{3.25}$$

We notice that on the northern half $\mathbf{v}_S$ and **B** are antiparallel so $\mathbf{v}_S \cdot \mathbf{H} < 0$ and, due to the inclination α the **B**-field makes with the surface SS', $\mathbf{B}\, d\mathbf{s} < 0$ and energy flux is positive (i.e., into the magnetosphere). On the southern half $\mathbf{v}_S$ is parallel to **B** so $\mathbf{v}_S \cdot \mathbf{H} > 0$ and, due to the negative inclination α, **B** makes with the surface $S''S'''$, $\mathbf{B}\, d\mathbf{s} > 0$. Energy flux is therefore directed into the magnetosphere at the southern tail lobe too.

When taking a cross-section of the tail and looking at it towards the Sun (as sketched in Figure 3.44 when the IMF is southward), we notice that the tail current $\mathbf{j}_T$ must flow from dawn to dusk in order to maintain the open tail. In order to complete the current loop, we let it flow around the surface of the cylinder with a current density $\mathbf{j}_T/2$.

In the solar wind we notice that the electric field is directed from dawn to dusk everywhere. The current at the surface of the cylinder is therefore antiparallel

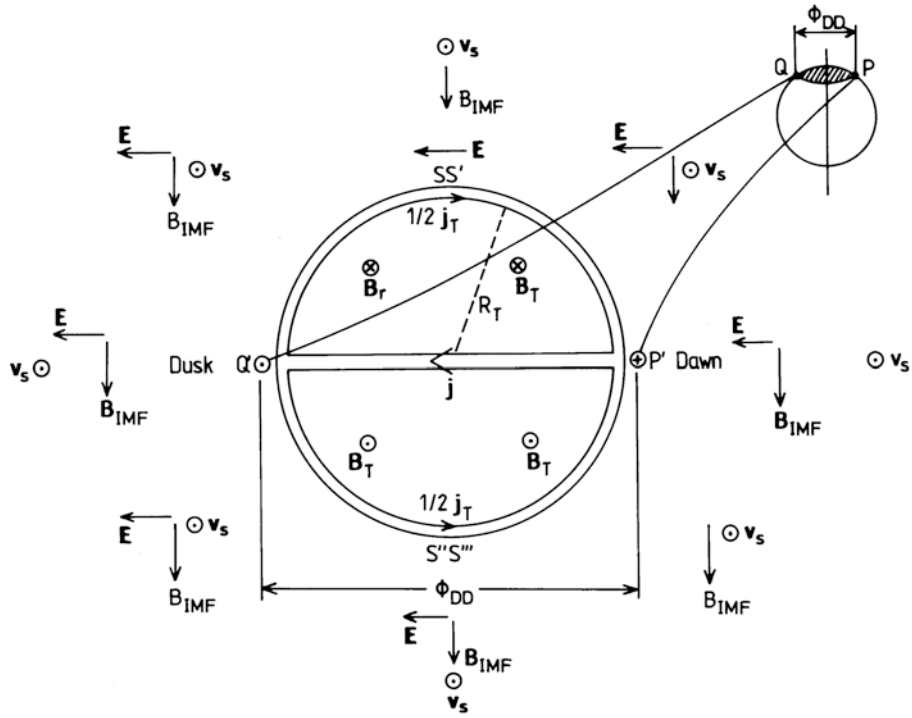


Figure 3.44. An idealized cross-section of the tail showing the central current sheet from dawn to dusk together with the return currents along the surface of the tail lobes. The electric field associated with the solar wind flow is parallel to the central tail current but antiparallel to the currents at the surface. Thus there is a magnetospheric generator at the surface of the tail driving the currents to the load in the central sheet. A potential ϕ_{DD} is formed between Q' and P' across the tail, which is mapped across the polar cap between Q and P.

to **E** and energy is created there. The surface acts as a generator or a dynamo. The cross-tail current, however, will be parallel to the field and acts as a load.

If the IMF turns around, however, the electric field also does, and the generator ceases to work. This is a major difference between southward-directed and northward-directed IMF, and it leads to a more closed magnetosphere for northward IMF, as already mentioned.

Applying reasonable magnitudes on $\mathbf{B}_{\mathrm{IMF}}$ ($\approx 5\,\gamma$) and the solar wind velocity (≈ 400 km/s), we can find the potential drop between dawn and dusk (marked P' and Q', respectively) if we assume that the radius of the tail is $R_T \approx 15\,R_e$. The potential is then given by

$$\phi_{DD} = -\int \mathbf{E}\, d\mathbf{l} = 2E \cdot R_T = 2v_S B R_T \approx 400\ \mathrm{kV}$$

To understand how the different regions in the magnetosphere are connected to the Earth's ionosphere via the geomagnetic field lines is one of the greatest

challenges in ionosphere–magnetosphere physics today. Let us for simplicity assume that this potential ϕ_{DD} between P' and Q' in Figure 3.44 is exactly mapped to the polar cap between the dawn and dusk points P and Q, then we can give an estimate of the dawn–dusk electric field across the polar cap. If we let the polar cap be above 75° latitude ($\theta_p = 15°$), we find:

$$E_{DD} = \frac{\phi_{DD}}{2R_e \sin\theta_p} = 8 \text{ mV/m}$$

Let us now go back to the force $\mathbf{f}_T$ on the magnetospheric surface due to the tail current $\mathbf{j}_T$ which is generated by the solar wind dynamo and given by:

$$\mathbf{f}_T = \mathbf{j}_T \times \mathbf{B} = \frac{1}{\mu_0}(\nabla \times \mathbf{B}) \times \mathbf{B}$$

This force must have a component along the surface (tangential) opposing the solar wind. By using the vector identity (3.21)

$$\mathbf{f}_T = \frac{1}{\mu_0}[(\mathbf{B} \cdot \nabla)\mathbf{B} - \tfrac{1}{2}\nabla B^2]$$

The total force on the volume τ of the magnetospheric cylinder is

$$\mathbf{F}_T = \int_\tau \mathbf{f}_T \, d\tau = \frac{1}{\mu_0}\int_\tau [(\mathbf{B} \cdot \nabla)\mathbf{B} - \tfrac{1}{2}\nabla B^2] \, d\tau$$

where $d\tau$ is a volume element. And finally

$$\mathbf{F}_T = \frac{1}{\mu_0}\int_S [(\mathbf{B}\cdot\hat{\mathbf{n}})\mathbf{B} - \tfrac{1}{2}B^2\hat{\mathbf{n}}] \, d\mathbf{s} \tag{3.26}$$

where S is the surface of the tail, and $d\mathbf{s}$ is a surface element along the unit vector $\hat{\mathbf{n}}$.

$$d\mathbf{s} = ds\,\hat{\mathbf{n}}$$

The last term in (3.26) expresses the magnetic pressure force. Since we are interested in the tangential force only, which acts to stretch the tail, we find when $\mathbf{B}$ is assumed constant across the surface

$$F_T^t = (F_T \cdot \hat{\mathbf{t}}) = \frac{1}{\mu_0}(\mathbf{B}\cdot\hat{\mathbf{n}})\cdot(\mathbf{B}\cdot\hat{\mathbf{t}})S$$

where t is a unit vector tangential to S.

$$F_T^t = \frac{1}{\mu_0}(B_n \cdot B_t)S$$

where B_n and B_t are components of the magnetic field normal and parallel to the surface S, respectively. In a stable situation this force balances the momentum force in the solar wind.

The rate at which the solar wind brings energy into the magnetosphere can now be found:

$$P_S = F_T^t \cdot v_S = \frac{v_S}{\mu_0}(B_n \cdot B_t)S$$

since v_S is parallel to F_T everywhere. This energy rate can be compared with the energy flux due to the Poynting vector given by (3.25). For reasonable values of v_s, B, and S we can find ϕ_E and P_S to be of the order of 10^{11} W.

Probably the most direct interaction taking place between the solar wind and near-Earth plasma is found above the polar cap regions. In order to gain insight into this, we again give an oversimplified illustration (Figure 3.45) where the magnetic field lines enter as parallel and vertical lines to the polar cap and perpendicular to the solar wind. The Earth is viewed from the nightside toward the Sun. In relation to Figure 3.45 the solar wind will generate an electric field

$$\mathbf{E}_S = -\mathbf{v}_S \times \mathbf{B}$$

which will be directed from dawn to dusk and, since the magnetic field lines are perfect conductors, it will penetrate the ionosphere and drive a cross-polar Pedersen current (J_{Pc}) in the same direction as the electric field because of the

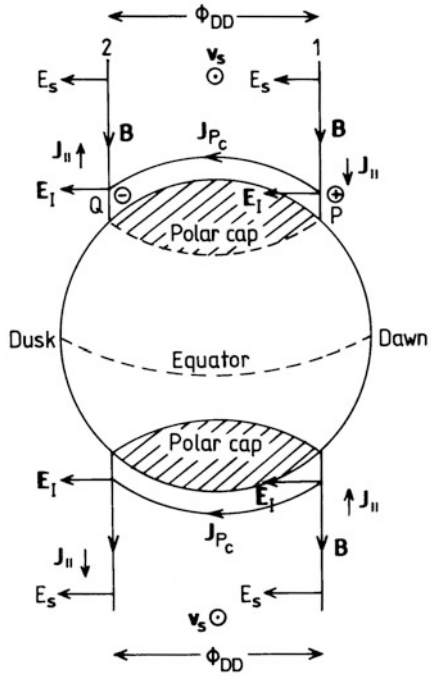


Figure 3.45. A schematic illustration of the parallel currents penetrating the ionosphere from the magnetosphere along magnetic field lines. They are deflected in a horizontal direction in a thin ionospheric layer between 90 and 150 km.

finite resistivity there. The ionospheric current will be limited to a height region h (90–140 km) in which the Pedersen and Hall conductivities can be denoted σ_P and σ_H, respectively. We assume for simplicity that these are independent of height. The cross-polar cap current J_{Pc} will then be:

$$J_{Pc} = \sigma_P h E_I = \Sigma_P E_I$$

where we have allowed for some difference in $\mathbf{E}_S$ and $\mathbf{E}_I$ due to finite conductivity in the ionosphere which affects the mapping. These currents have to be closed in some loops, and due to the high conductivity along the field lines, parallel currents are the likely candidates. How they close in the magnetosphere, magnetosheet, or the solar wind is another matter. We notice, however, that these currents enter the polar cap ionosphere on the dawn side and leave it in the dusk side. These currents are usually referred to as region-1 currents. Due to the complicated geometry involved (Figure 3.45) these currents will form curved sheet currents or currents flowing along cylinder-like surfaces.

3.15 MAGNETIC FIELD CONVECTION

The Earth's magnetic field can merge with the interplanetary field in such a way that it opens up or closes the magnetosphere. Furthermore, the solar wind pressure deforms the symmetric magnetosphere that could otherwise occur in a vacuum and stretches the field lines and tears them apart. Since the field lines are anchored in the Earth, they will take part in the Earth's rotation and sweep around in loops which will be stretched and possibly torn apart on the nightside, later to tie together again toward the daytime (as first outlined by Dungey, 1961). The magnetic field lines therefore undergo steady motion, sometimes disrupted by disturbances in the solar wind, and they will, due to the high conductivity, carry the magnetospheric plasma along. A schematic diagram of this field line in the midday–midnight meridional plane convection is shown in Figure 3.46 where the interplanetary magnetic field at point 0 starts to merge with the Earth's magnetic field at stage 1. The interplanetary field forced by the solar wind tears the Earth's magnetic field apart and carries it across the poles in stages 2, 3, and 4. In stage 5 the Earth's magnetic field starts on the reconnection process in the tail. At stage 6 reconnection occurs, and at stage 7 the Earth's field line is tied together again as it is on its way toward a new cycle. Although the picture is illustrated in the meridional plane, the complete process takes place as the field rotates around. At any one moment, however, a snapshot of the meridional plane will show different field lines at different stages.

The plasma transport related to this process is indicated by the thick arrow in Figure 3.46. Of special interest is the return flow toward the Earth in the equatorial plane.

This sunward convection in the equatorial plane is illustrated in more idealized detail in Figure 3.47. The plasma flow $\mathbf{v}_m$ is assumed rather uniform in the nightside and is deflected toward the morning and evening flanks to flow out

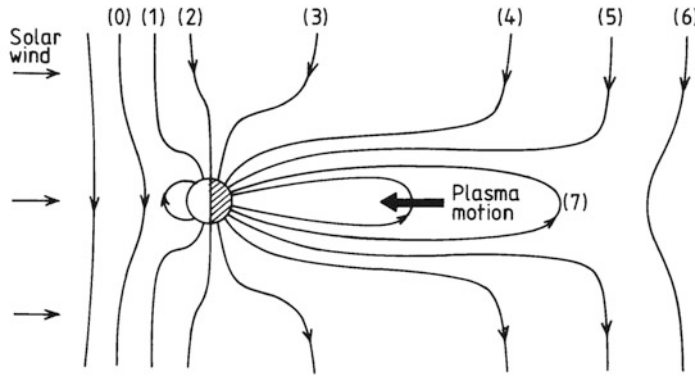


Figure 3.46. A schematic diagram showing the magnetic field convection in the noon–midnight meridian plane. At point 0 the interplanetary field line starts to merge with a field line from the Earth. At 1 the merging has occurred, and at 2–5 the Earth's field line is drawn back over the pole as the IMF is carried tailward by the solar wind plasma. At 6 the open field line starts to reconnect, and at 7 the Earth's magnetic field line is closed again. (After Dungey, 1961.)

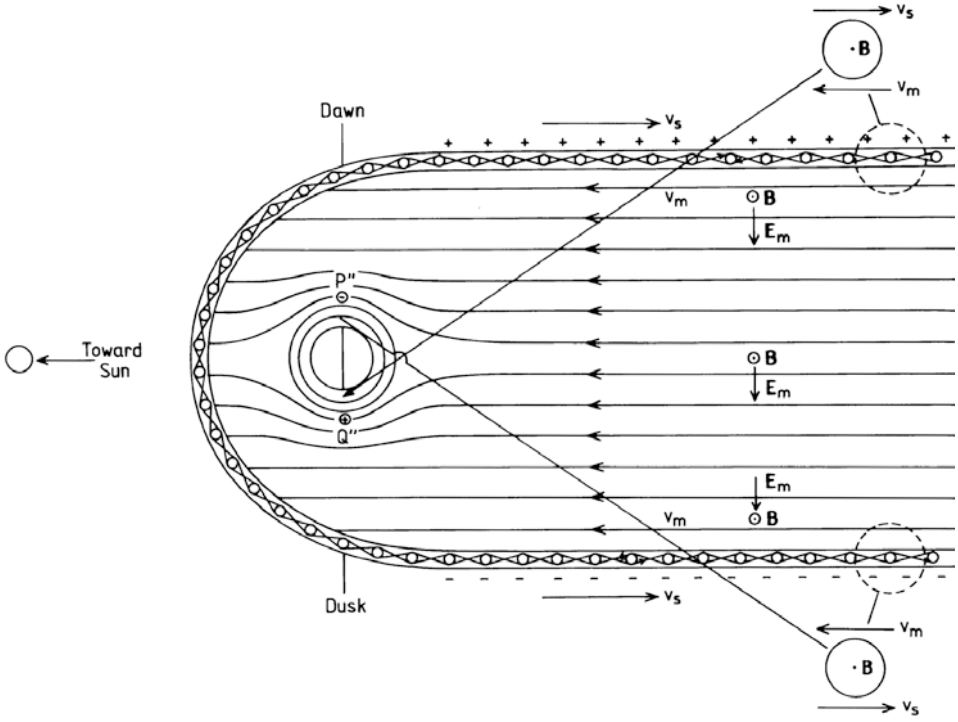


Figure 3.47. An illustration to show areas of strong shear velocities in the magnetospheric equatorial plane. The shear is such that it leads to vortices associated with positive and negative space charges on the dawn and dusk sides of the magnetosheet boundary and on the dusk and dawn side of the inner plasma sheet, respectively.

of the magnetosphere and to the magnetosheet on the dayside front of the equatorial plane. Plasma flow in the magnetosphere is of course far more complicated than depicted there, since it certainly is three dimensional.

We notice that the magnetic field in the equatorial plane is directed upward through the paper and that therefore the "frozen-in" concept must imply an electric field directed from dawn to dusk across the equatorial plane (as illustrated in the figure).

The magnetosheet acts as a shield between the solar wind and magnetospheric plasma. It is not very well understood how energy can actually be transferred into the magnetosphere and drive the convection there. There are several suggestions: one refers to viscous interaction between the solar wind and magnetospheric plasma; another is field line merging and reconnection. Independent of the mechanism the processes occurring at the boundary between the solar wind and magnetospheric plasma create large irregularities and wave-like features there. The interest in magnetospheric sources for wave- and vortex-like large-scale structures has been intensified lately due to recent auroral images obtained from satellites (Figure 3.48). The remarkable features to be observed are the distinct intensified regions (hot spots) with a fairly constant distance between them along the auroral oval. It is evident that these features must be related to similar characteristic struc-

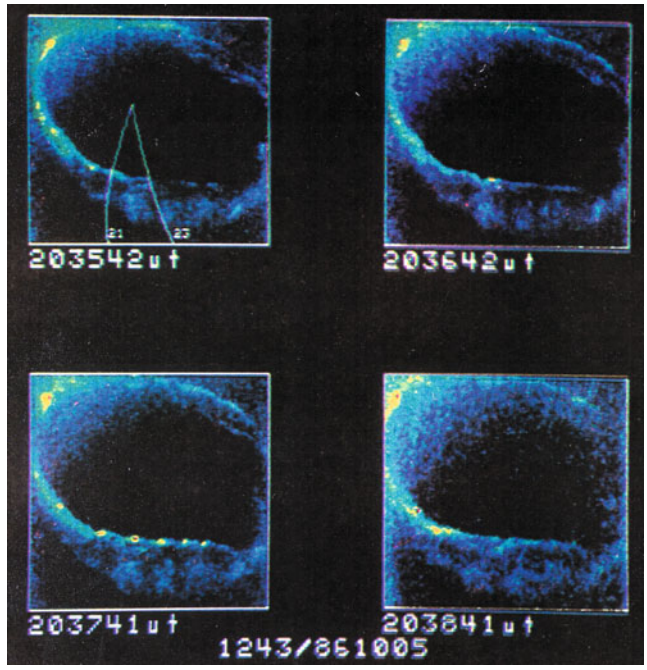


Figure 3.48. Auroral images from the Viking satellite showing the auroral oval aligned with approximately equidistant spots of strong auroral emissions. (From Hultqvist, 1988.)

tures in the magnetosphere, but exactly how and where is a question of mapping in a topology that no one can answer for certain.

The very light spots must also be related to concentrated particle precipitation carrying strong currents. The question is, therefore, where can charges build up and be sustained in such concentrated regions and for such a long time that these high-intensity features can exist?

We have seen that the magnetosheet in the front of the magnetosphere can be very narrow. Let us therefore assume that this sheet is also relatively narrow on the tailward side of the magnetosphere. Because the plasma flow reverses sign across this boundary from being antisunward in the solar wind to becoming sunward inside the magnetosphere, a large gradient will exist in the plasma flow perpendicular to the flow—a region known from hydrodynamics to be the source of vortices and turbulence. In Figure 3.47 this is indicated by a turbulent wavy structure.

Let us consider maintaining a positive space charge q in a highly conducting plasma creating an electric field perpendicular to a magnetic field (as illustrated in Figure 3.49(a)). In this situation the plasma would have a velocity given by

$$\mathbf{v} = \mathbf{E} \times \mathbf{B}/B^2$$

which will correspond to rotation in a clockwise direction around the positive **B**-direction.

On the other hand, if the space charge was negative, rotation would be in an anticlockwise direction around positive **B** (Figure 3.49(b)).

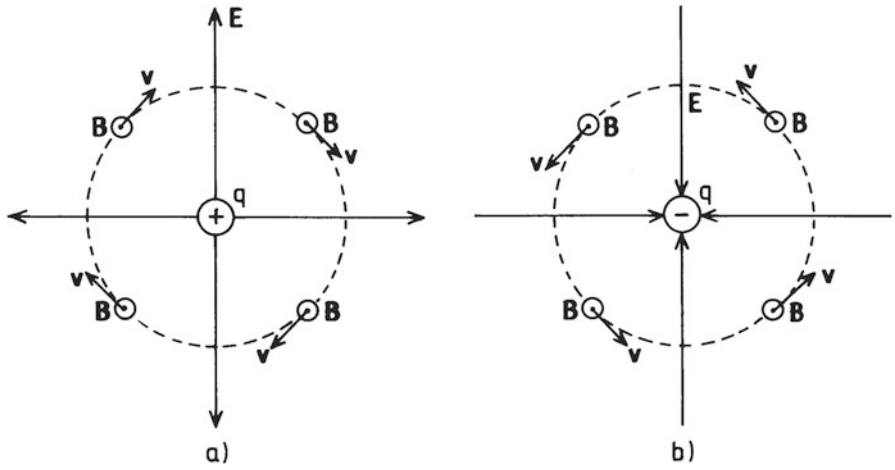


Figure 3.49. (a) An idealized picture to illustrate the direction of plasma rotation when a positive charge is situated in an area with a magnetic field. Rotation due to the $\mathbf{E} \times \mathbf{B}$ motion is clockwise observed toward the direction of the **B**-field. (b) When there is a low potential permeated by a magnetic field the rotation direction of the plasma becomes anticlockwise when viewed toward the direction of **B**.

If we now go back to Figure 3.47 and the rather strong velocity shears across the magnetosheet in the tail region, these shears could lead to vortices which on the dawn side would be in the clockwise direction and on the dusk side in the anticlockwise direction. This would then correspond to imaginary space charges, positive on the morning side and negative on the evening side as indicated by the plus and minus signs, respectively. We notice that this would be true, however, if a magnetic field directed out of the paper was prevailing as it is on closed field lines inside the magnetopause at least.

We can therefore associate these vortices with space charges. But space charges cannot exist for very long in the magnetosphere, as they will try to find ways of neutralizing each other. One possibility is, as was indicated for the situation above the polar cap, that the charges neutralize through currents via the ionosphere (Figure 3.45). The current would then flow into the ionosphere on the morning side and out of the ionosphere on the evening side (as illustrated in Figure 3.50), and this current would be in the same sense as the cross-polar cap current, as discussed earlier, which is again called a "region-1" current. In the magnetosphere this current could close as a part of the surface tail current.

There is another area closer to the Earth (as indicated in Figure 3.47) where

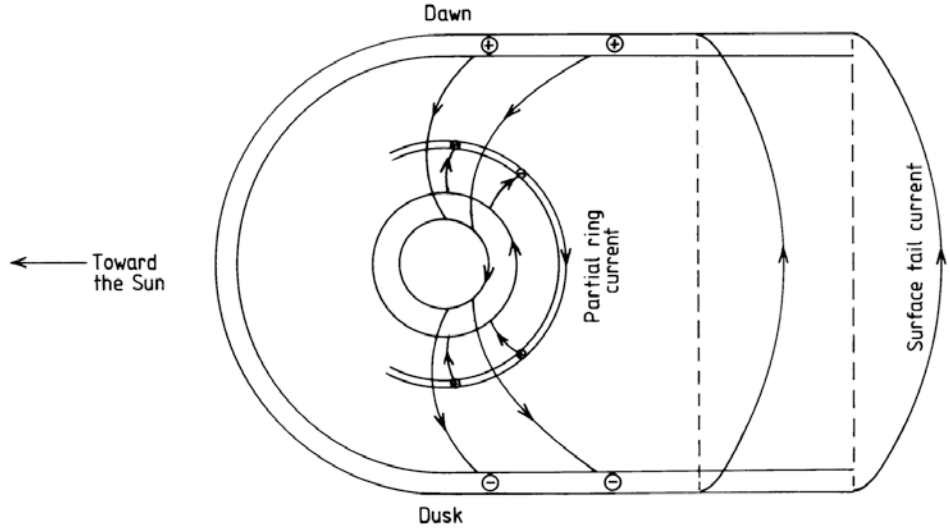


Figure 3.50. A possible closure of the magnetospheric–ionospheric current system. The space charges in the magnetosheet in the ecliptic plane drive a current along the magnetic field lines to the polar cap, downward on the dawn side and upward on the dusk side, respectively. The currents flow across the polar region along the poleward boundary of the auroral oval and are called "region-1" currents. These currents are probably closed on the magnetospheric surface in the tail as indicated. From the inner magnetosphere similar field-aligned currents are driven from the dusk side toward the dawn side through the lower latitudes of the auroral oval. These currents are denoted "region-2" currents. In the magnetosphere the current is probably closed in the partial ring current as indicated.

strong velocity shears are expected to occur. Close to the Earth the plasma will corotate ($r < 6$–$7\ R_e$) while the plasma from the magnetotail will be drifting by the cross-tail electric field. Velocity shears are such that a positive space charge is expected on the dusk side and a negative charge on the dawn side, opposite to what we found in the magnetosheet. These charges can then create currents through the ionosphere from dusk to dawn at a lower latitude than the region-1 current. These currents are related to the so-called region-2 currents and traverse the ionosphere at the equatorward boundary of the auroral oval. In the magnetosphere they are probably closed by the partial ring current.

3.16 HIGH-LATITUDE CONVECTION PATTERNS AND FIELD-ALIGNED CURRENTS

Due to the high conductivity along magnetic field lines, any potential that is created between them in the magnetosphere will map down to the ionosphere where they will give rise to an electric field which sets the ionospheric plasma into motion (as shown in Section 6.8). As already mentioned, in order to understand this mapping in full, we need detailed knowledge of the topology of the magnetosphere which we are not in possession of today.

An idealized representation of the plasma flow in the high-latitude and polar cap ionosphere is shown in Figure 3.51.

The dawn–dusk directed polar cap electric field $\mathbf{E}_{pc}$, which corresponds to the cross-tail electric field $\mathbf{E}_m$ or the solar wind electric field $\mathbf{E}_{DD}$, drives ionospheric plasma from the dayside across the polar cap to the nightside. In the nightside this plasma flow diverges eastwards and westwards along a narrow latitudinal belt associated with the auroral oval. On the morning side these two flow branches converge again to complete the flow circuit. Flow reversal takes place between 70° and 80° of latitude on the dawn and dusk side. Since rotation is clockwise in the dawn sector at P in the figure (Figure 3.51), a positive potential occurs here and the contrary takes place in the dusk sector at Q. The electric field, $\mathbf{E}_a$, in the auroral oval will therefore be directed poleward on the dusk side and equatorward on the dawn side. This picture has emerged from years of observations and study by means of ground-based and satellite techniques. The picture portrayed in Figure 3.51 is of course a considerable oversimplification, but it helps in getting the main frame of reference under control. Due to great variability in the solar wind, the interplanetary field, and their interaction with the magnetosphere, this two-cell convection model in the polar cap is highly distorted and may be broken up into a large number of local cells and turbulence.

Another result which has emerged from satellite measurements is the presence of field-aligned currents at any time. An overview of the main results from these observations is shown in Figure 3.52. The spiral-formed areas around the polar cap indicate the region for the average position of currents flowing into and out of the ionosphere. The inner ring is termed "region-1" current, while the outer

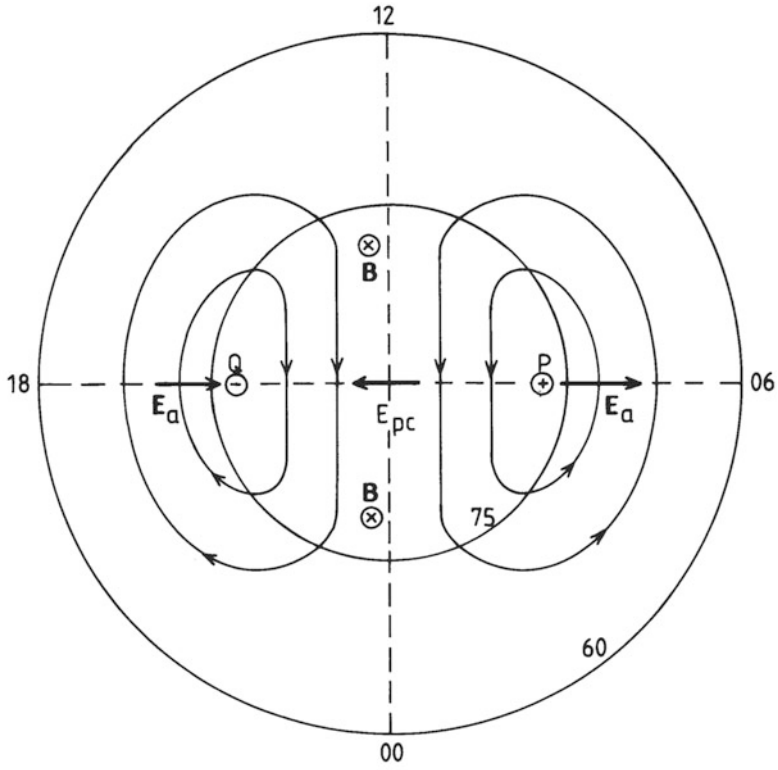


Figure 3.51. An oversimplified model of the high-latitude ionospheric convection pattern showing a symmetric two-cell system with clockwise rotation on the dusk side and anticlockwise rotation on the dawn side. Plasma convection is thought to be driven by the dawn–dusk electric field $\mathbf{E}_{pc}$ across the polar cap. The return flow takes place along narrow belts at lower latitudes in the auroral oval. The points P and Q are related to high and low potentials, respectively, thus a poleward electric field, $\mathbf{E}_a$, occurs on the dusk side and an equatorward electric field on the dawn side of the auroral oval, respectively.

ring is denoted "region-2" current. There is, however, an area of overlap around midnight. Similarly there is another area of confusion at midday. For the dawn and dusk sides the situation appears well behaved.

To synthesize our understanding of the dawn–dusk current system between the ionosphere and magnetosphere, we refer to the diagram sketched in Figure 3.53. Here the Earth is seen from the nightside with the midnight meridian marked as the diameter. The field-aligned currents in the dawn–dusk meridian are in accordance with the satellite observations illustrated in Figure 3.52. In the ionosphere we know, especially from balloon and radar observations, that the auroral zone electric field $\mathbf{E}_a$ is directed northward in the dusk sector and southward in the dawn sector (i.e., from dusk to dawn). Since the ionosphere has finite conductivity, the field-aligned currents will mainly shorten in the ionosphere as a result of

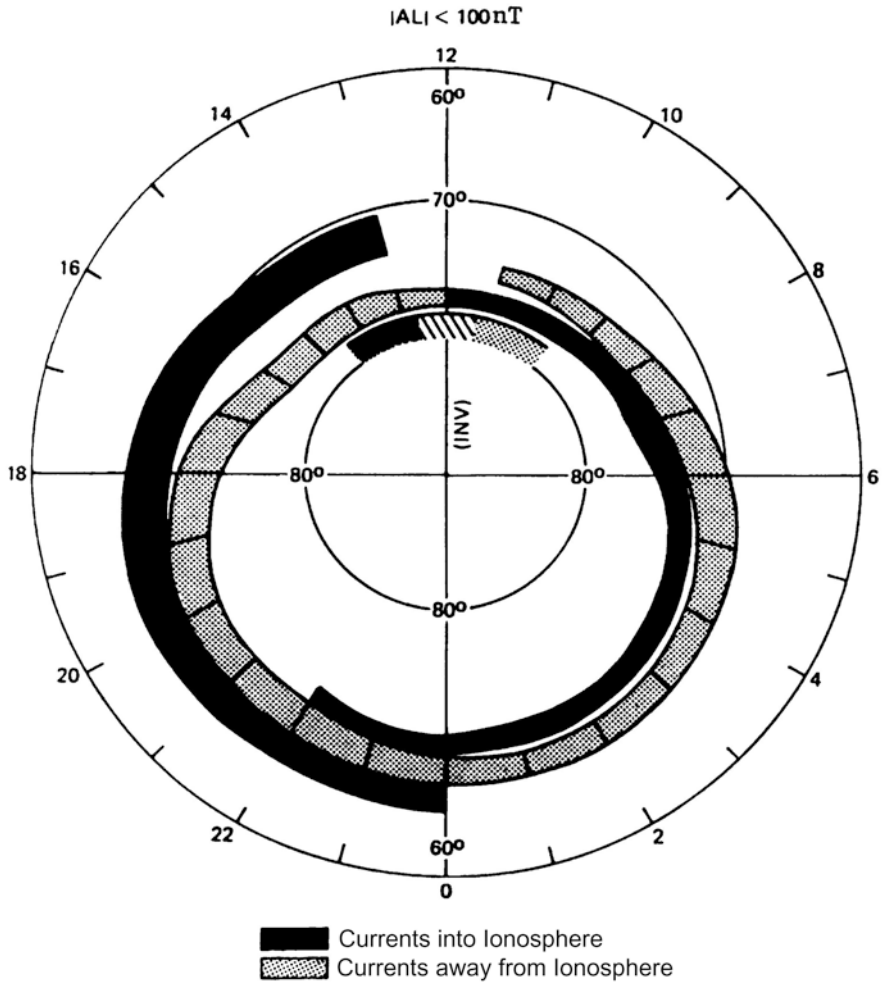


Figure 3.52. The average pattern of field-aligned currents in the high-latitude region when the IMF is southward. The currents into and out of the ionosphere are indicated by different symbols, thus showing currents out of the ionosphere in the evening and into the ionosphere in the morning associated with the region-1 current. The equatorial currents that enter the ionosphere in the evening and leave the ionosphere in the morning are associated with region-2 currents. The field-aligned currents at very high latitude at noon (cusp currents) are strongly dependent on the B_y-component of the interplanetary magnetic field (IMF). (From Ijiima and Potemra, 1976.)

Pedersen currents (i.e., along the E-fields), northward in the dusk sector and southward in the dawn sector, respectively. In the magnetosphere the closure of the loop is not well established, but we indicate here that it is radial in the dusk–dawn direction. This will correspond to currents closing between Q', Q'' and P', P'', respectively. We notice that, when the dusk–dawn auroral zone electric

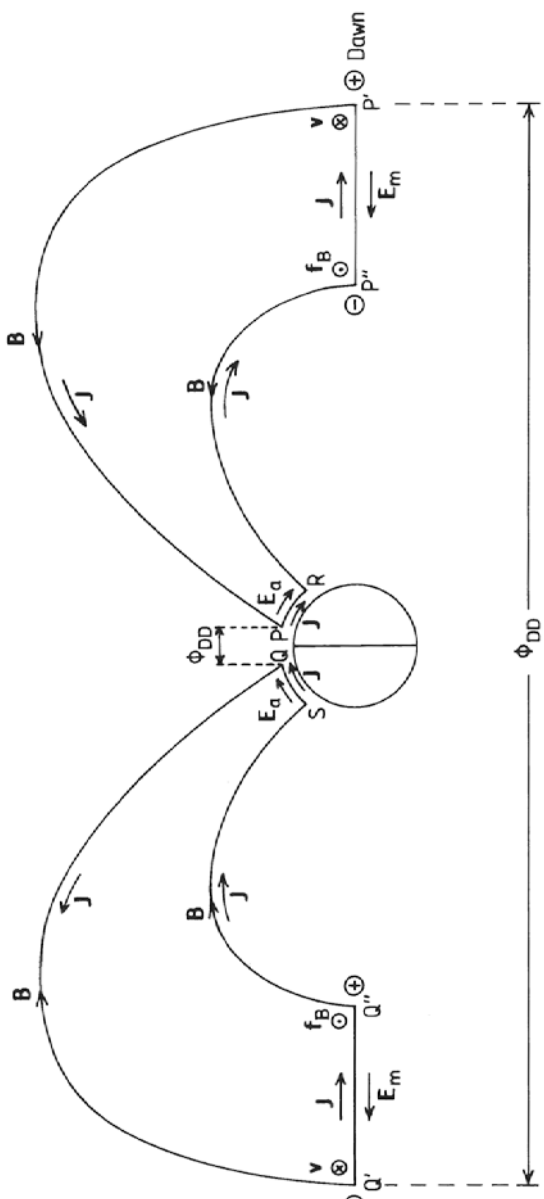


Figure 3.53. Possible synthesis of ground-based and satellite observations of the ionospheric–magnetospheric current system and electric fields. The cross-section is in the y–z plane along the dawn–dusk meridian seen from the tail side of the Earth (toward the Sun). The magnetospheric dynamo drives the currents in the dusk–dawn direction in the ecliptic plane. These currents continue to the ionosphere as currents parallel to B. In the ionosphere they are deflected in the horizontal direction as Pedersen currents along the auroral zone electric field. The magnetic force $\mathbf{f}_B$ in the ecliptic plane opposes the convection velocity $\mathbf{v}$. The electric potential ϕ_{DD} in the magnetosphere between the dawn and the dusk sector is projected along the magnetic field line across the polar cap between the points P and Q on the dawn and dusk side, respectively.

field is mapped out to the magnetosphere, they will be reversed and directed from dawn to dusk in agreement with the polar cap and cross-tail electric fields $\mathbf{E}_{pc}$ and $\mathbf{E}_m$, respectively (Figures 3.51 and 3.47, respectively). In the magnetosphere, therefore, the electric field and currents are antiparallel and thus form a generator, while they are parallel in the ionospheric load.

In the tail region of the magnetospheric equatorial plane convection is driven towards the Sun by the dawn–dusk electric field. This is indicated by the arrows $\mathbf{v}$ in Figure 3.53. Since the magnetic field is directed perpendicular to the equatorial plane, the magnetic force (Lorentz force)

$$\mathbf{f}_B = \mathbf{j} \times \mathbf{B}$$

due to the closure currents in the magnetospheric plane will be directed tailwards and against the plasma flow. If, therefore, the plasma flow for some reason is accelerated so that $d\mathbf{v}/dt > 0$, the term $\mathbf{B} \times d\mathbf{v}/dt$ in (3.23) will be directed from dusk to dawn and increase the current $\mathbf{j}$ in the loop. If, on the other hand, the plasma motion is decelerated, the current in the loop will decrease. This is the effect of the magnetospheric generator.

We will now leave the magnetosphere–ionosphere interaction and these rather speculative reflections concerning the current systems for a while. It should be clear that the magnetosphere with all its complexity is not only open for intrusion by the solar wind but also to the imagination of mankind.

We have indicated that polar cap and auroral zone convection in reality is quite different from the simplified picture in Figure 3.51 and that it will change according to the direction of the IMF. In Figure 3.54 a synthesis of the high-latitude daytime convection patterns as derived from satellite observations is presented for a negative B_z-component or a southward IMF, but for different directions and magnitudes of the B_y-component. When $B_y = 0$ we recognize the two-cell structure. When $B_y < 0$, or IMF is directed eastward, the eastern or dawn side convection cell shrinks and retreats to the polar cap, while the dusk side cell intrudes the eastern half of the dayside auroral oval. For a positive B_y or B_y directed westward, the dusk side convection cell covers the whole polar region while the dawn side cell remains in place but shrinks.

Figure 3.55 shows the high-latitude convection patterns under different conditions of B_y and B_z. The most characteristic difference between southward and northward IMF in these patterns is that several convection cells occur for $B_z > 0$. At $B_y \cong 0$ there are two closed cells inside the polar cap that are fairly symmetric around the noon meridian and situated between the more extended and permanent dawn and dusk cells. For $B_y < 0$ the dusk side polar cap cell increases while the opposite is true for $B_y > 0$, then the dawn side polar cap convection cell dominates.

When B_z is close to zero or weakly northward, a circumpolar convection cell covers the polar cap between the dawn and dusk cells. The flow also changes direction when the y-component changes sign from clockwise when $B_y > 0$ to anticlockwise when $B_y < 0$.

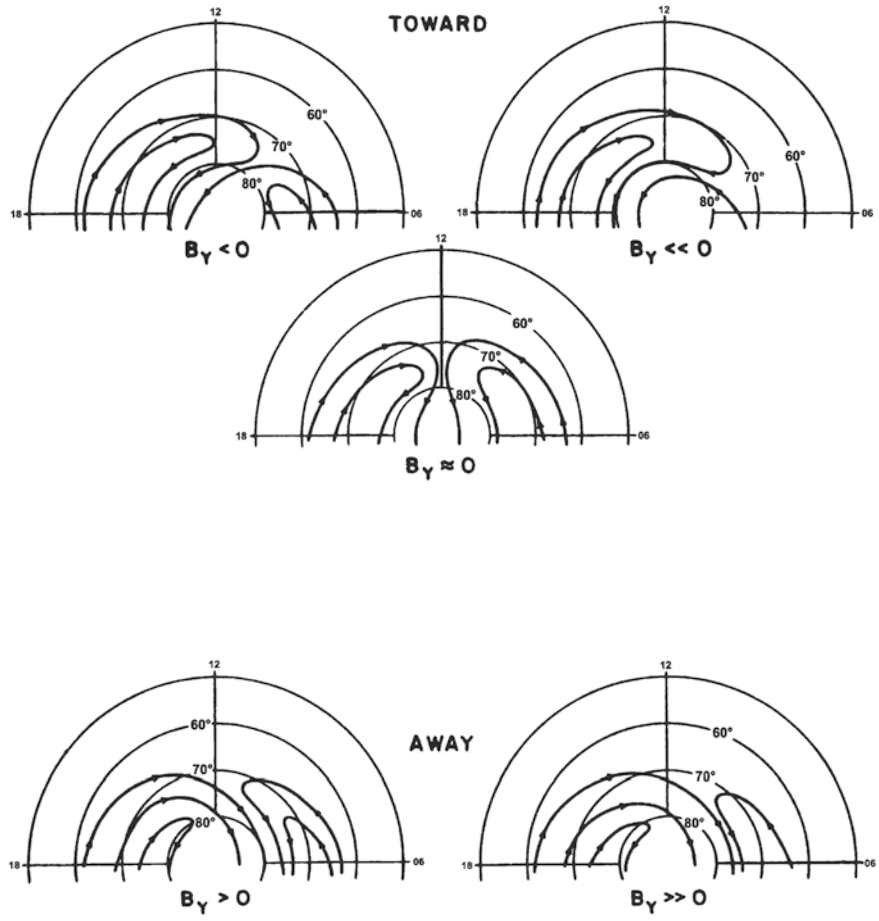


Figure 3.54. Schematic representation of the dayside high-latitude convection pattern showing its dependence on the IMF when B_z is southward. (From Heelis, 1984.)

When the IMF is southward, the convection pattern is dominated by an asymmetric two-cell system where the dawn cell is enhanced for $B_y < 0$ and for $B_y > 0$ the dusk side is largest.

For $B_y > 0$ there is also a narrow cell inside the dusk cell, which displays the same rotation as the main cell, and similarly there is a small cell inside the dawn cell when $B_y < 0$. It is not clear whether this is a separate cell or whether it simply represents a flow reversal. It is indicated in Figure 3.55 that the different cells have different origins. The labels V, M, and L represent viscous cells, merging cells, and tail lobe cells, respectively.

In general, we notice that the area taking part in high-latitude convection is larger for IMF southward than IMF northward, possibly a result of more efficient energy transfer from the solar wind to the magnetosphere when IMF < 0.

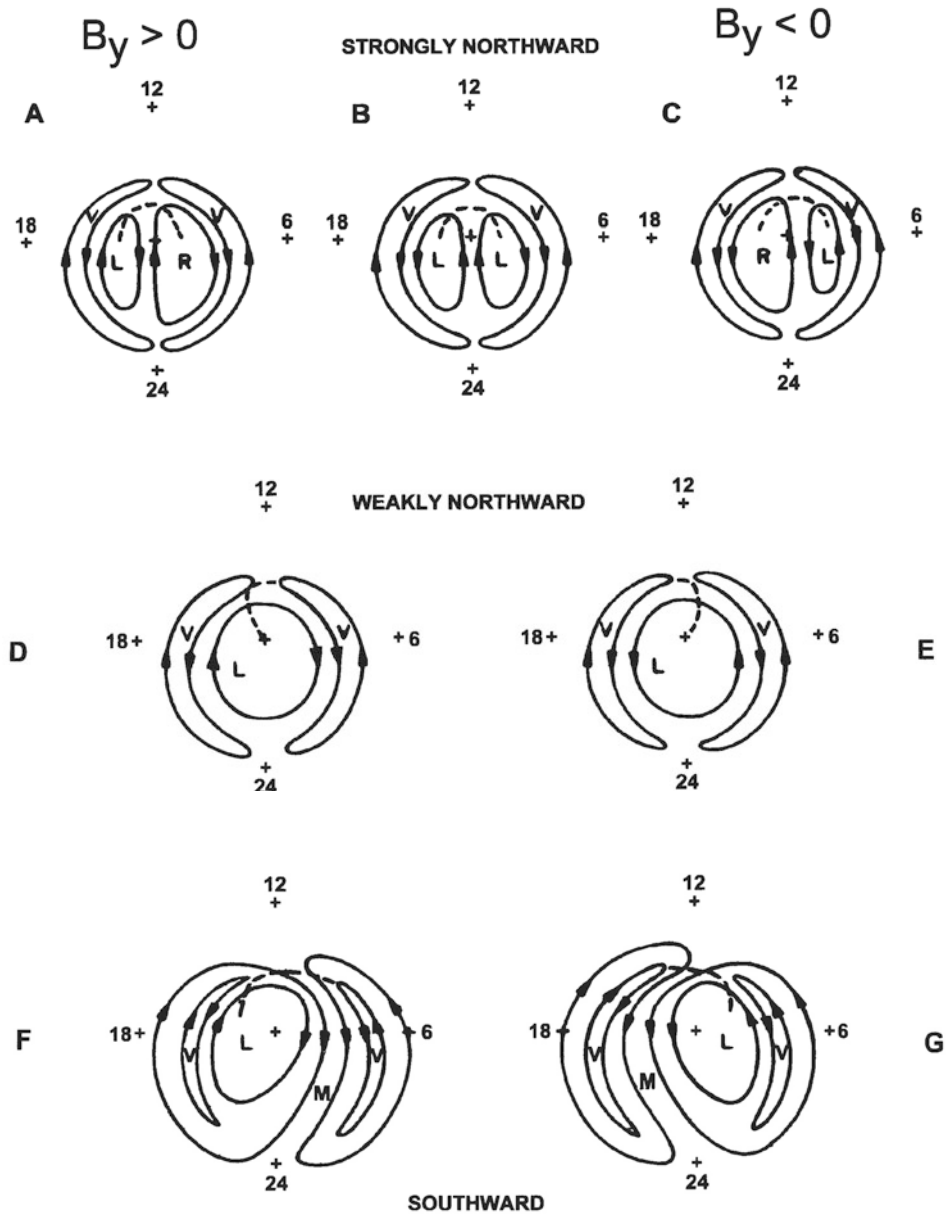


Figure 3.55. Schematic northern hemisphere polar cap convection patterns for various orientations of the IMF. The left column (i.e., A, D, and F) is for $B_y > 0$, the center column (i.e., B) has $B_y = 0$, and the right column (i.e., C, E, and G) is for $B_y < 0$. The top row (i.e., A, B, and C) is for strongly northward IMF ($B_z > 0$), the middle row (i.e., D and E) for weakly northward IMF, and the bottom row (i.e., F and G) for southward IMF. Viscous cells are marked with a "V", merging cells are marked with an "M", and lobe cells with an "L". (From Reiff and Burch, 1985.)

3.17 EXERCISES

1. (a) Prove that

$$\mathbf{M}_0 \cdot \nabla\left(\frac{1}{r}\right) = \frac{M_0 \cos\theta}{r^2}$$

when $\mathbf{M}_0 = -M_0\hat{z}$ in the coordinate system shown in Figure 3.6.

(b) Prove by using the same coordinate system the identity

$$-\nabla\left(\mathbf{M}_0 \cdot \nabla\left(\frac{1}{r}\right)\right) = 3(\mathbf{M}_0 \cdot \mathbf{r})\frac{\mathbf{r}}{r^5} - \frac{\mathbf{M}_0}{r^3}$$

Use this to derive equations (3.6a, b, c).

2. Assume the Earth has an ideal dipole (as outlined in Figure 3.3). The geomagnetic coordinates of Tromsø are 67.25° N, 116.0° E

(a) Find the magnetic field strength in Tromsø.

(b) Find the angle I in Tromsø.

(c) Find the value for r_0 for Tromsø.

(d) Derive the E-field mapping factors for Tromsø.

(e) Derive the relationship between δ_m and δ_i for Tromsø.

3. The auroral oval is at present, to a good approximation, equal to a circle around the geomagnetic dipole axis at 67° dipole latitude. Assume that the magnetic moment is reduced to half its present value and determine then the geomagnetic latitude of the auroral oval.

4. Assume that the Earth's magnetic moment is created by a current around the Earth's equator. What would the strength of this current be?

5. Assume different values for v in (3.16) and derive r^0_{mp}.

4

The ionosphere

4.1 THE PRODUCTION OF IONIZATION BY SOLAR RADIATION

The existence of an ionized layer in the upper atmosphere was probably already appreciated at the beginning of the last century when Marconi demonstrated that radiowaves could propagate across large distances beyond the free horizon, as if they were guided between a conducting layer and the ground. A conducting layer in the upper atmosphere was also inferred as early as 1880 when studies of diurnal oscillations in the Earth's magnetic field indicated that these variations might be caused by tidal oscillations in the upper atmosphere.

It was, however, not until 1924 that the existence of an ionized conducting layer in the upper atmosphere was fully proven, when the existence of the ionosphere was demonstrated by studies of the reflections of radiowaves from the atmosphere.

The ionosphere can be viewed as a variable shell of ionization (plasma) surrounding the Earth. The range of variability in vertical profiles of the electron density up to altitudes of 800 km are shown in Figure 4.1. Typically, there is a maximum around 300 km; this can, however, vary between 200 and 600 km. Below this maximum there often appear bumps in the profile, sometimes distinct enough to form secondary peaks such as during auroral events when they can form rather sharp layers (Figure 4.2).

Historically, the ionosphere has been divided into layers, the earliest detected was the E-layer, so named because of the reflections of electric fields. For alphabetical order we have since got D- and F-layers below and above the E-layer, respectively. Today it is more common to speak about regions as distinctions between layers are not all that clear, except maybe for some special events in the E-region such as sporadic and auroral E-layers.

A. Brekke, *Physics of the Upper Polar Atmosphere*, Springer Praxis Books,
DOI 10.1007/978-3-642-27401-5_4, © Springer-Verlag Berlin Heidelberg 2013

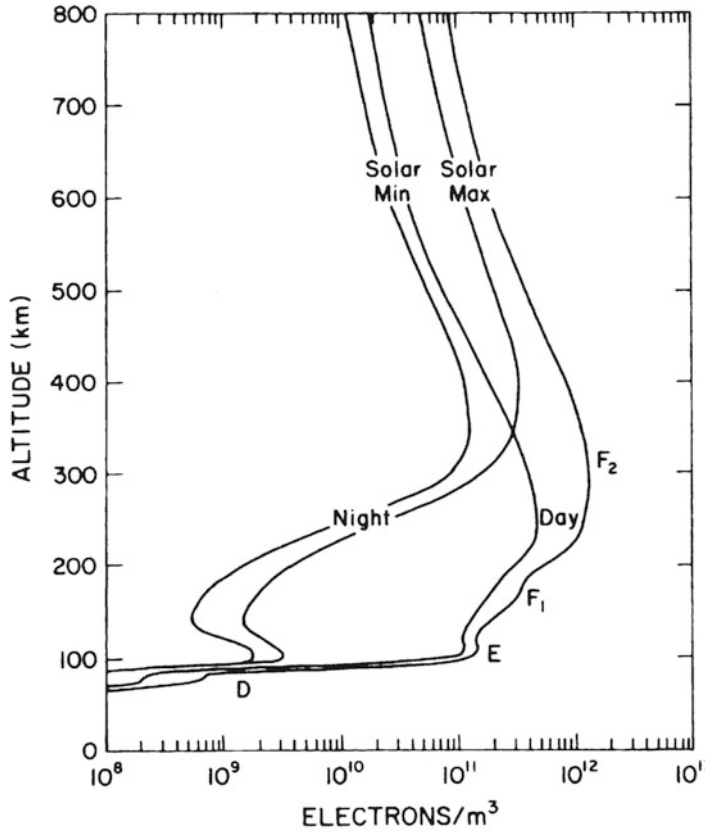


Figure 4.1. Typical mid-latitude ionospheric electron density profiles for sunspot maximum and minimum conditions at daytime and night-time. The different altitude regions in the ionosphere are labeled with the appropriate nomenclature. (From Richmond, 1987.)

Below the E-region, between 60 and 90 km, is the D-region which is very variable with lesser electron density. Sometimes the lower D-region may even be referred to as the C-region, alluding to its complicated chemistry.

The ion composition changes in accordance with the change of molecular constituents in the neutral atmosphere. As illustrated in Figure 4.3, NO^+ and O_2^+ are the dominant ions below, say, 150 km, above which atomic oxygen ions, O^+, are more abundant. Above 300 km, H^+ is a more abundant ion species than either O_2^+ or NO^+. O^+ may dominate even up to 600 km and higher. This will, however, depend strongly on magnetospheric and solar conditions. For strong disturbances in the auroral region heavy ions can dominate up to 200 km or above.

The temperature of the ionosphere is essentially controlled by the absorption of solar UV radiation in the thermosphere. Since this radiation is able to ionize molecules and atoms, free electrons as well as ions are formed which carry some excess energy, obtained during the ionization process, with them. Electrons have

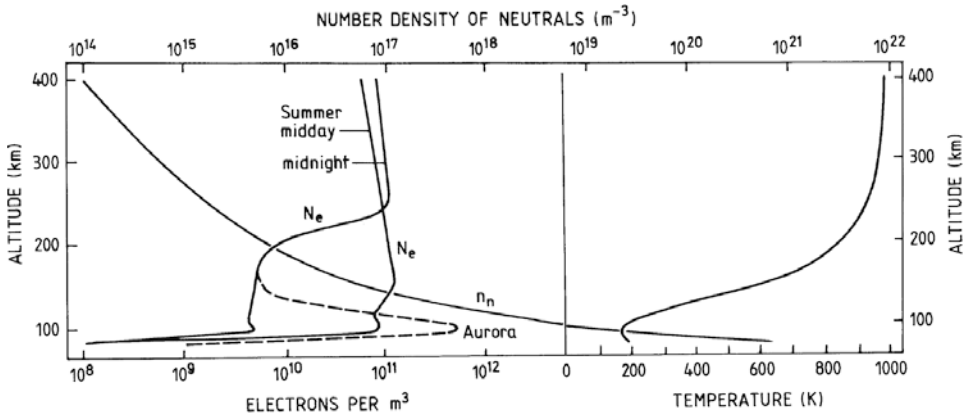


Figure 4.2. Electron density profiles representing average daytime and night-time conditions at high latitudes. Also indicated is a profile for auroral conditions. The background neutral density profile together with an average neutral atmosphere temperature profile are also schematically illustrated.

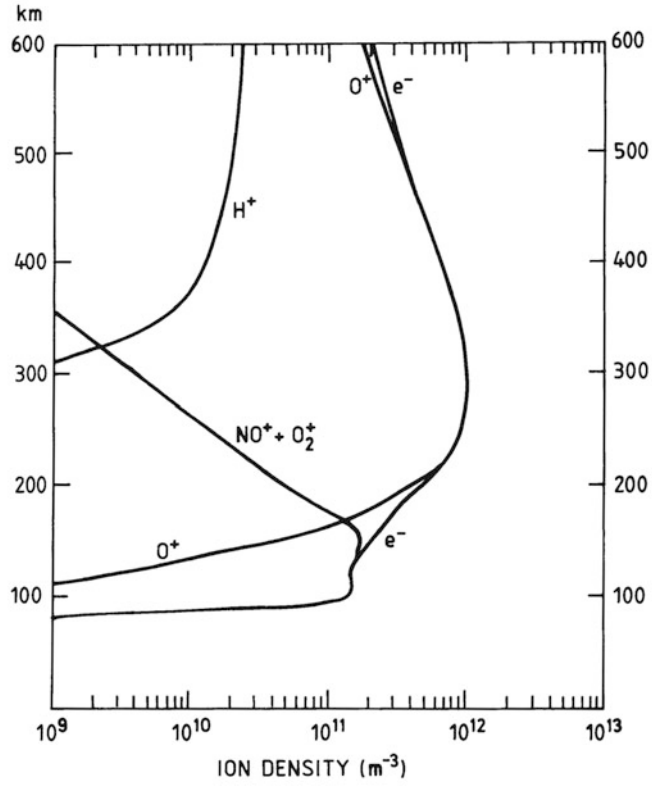


Figure 4.3. Altitude profiles of the most typical ion species in the ionosphere between 100 and 600 km, together with the corresponding electron density profile. (After Richmond, 1987.)

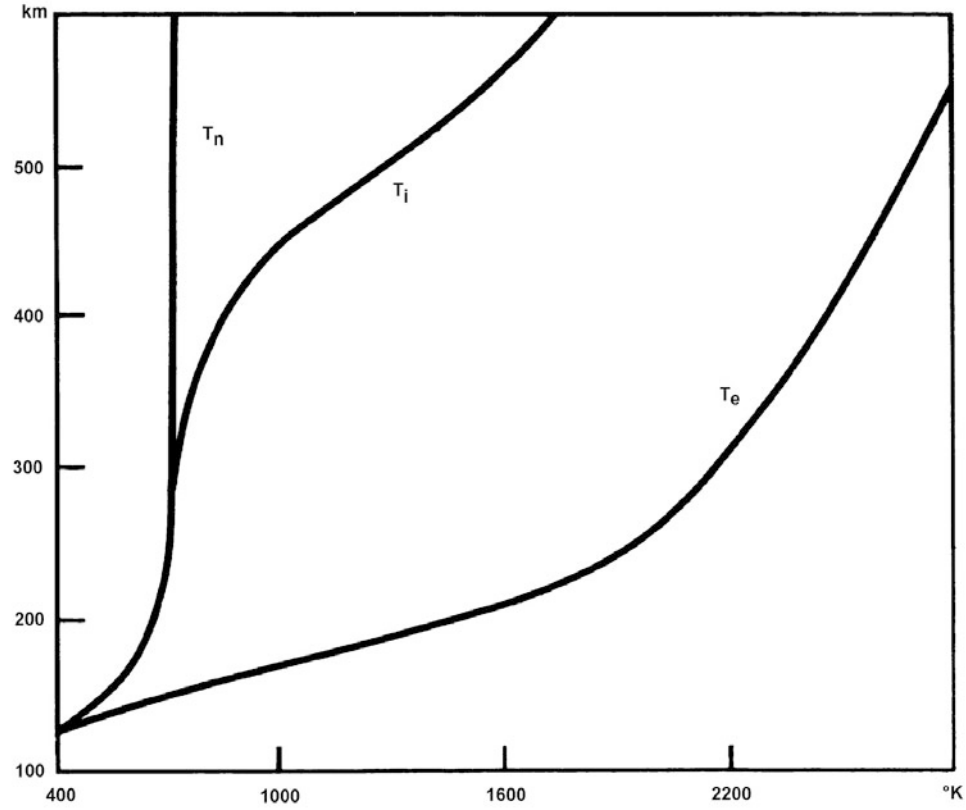


Figure 4.4. Representative altitude profiles of the ion, electron and neutral temperatures between 100 and 600 km. (After Giraud and Petit, 1978.)

greater mobility and heat conductivity so their temperature usually gets higher than that for ions. Ions are heavier and interact by collisions more strongly with the neutral gas. Because of the similar masses between the colliding species, much of the excess energy of ions is transferred to the neutral gas. The temperature of ions is therefore lower than for electrons, but higher than for neutrals. A typical model of temperature profiles for neutrals, ions, and electrons is shown in Figure 4.4.

Ion and electron temperature vary strongly with time of day, season, and level of disturbance. Figure 4.5 shows isotherm plots for electrons and ions according to various local times and heights as observed at mid-latitude stations. At all heights, from 200 km to 600 km, electron and ion temperatures increase abruptly at sunrise and decrease almost as rapidly at sunset. Variations in electron temperature, however, are more rapid than in ion temperature.

One of the most powerful tools to study the conditions of the ionosphere from ground is the incoherent scatter radar. One such radar system is situated in the auroral zone close to Tromsø, Norway. That is the European Incoherent Scatter

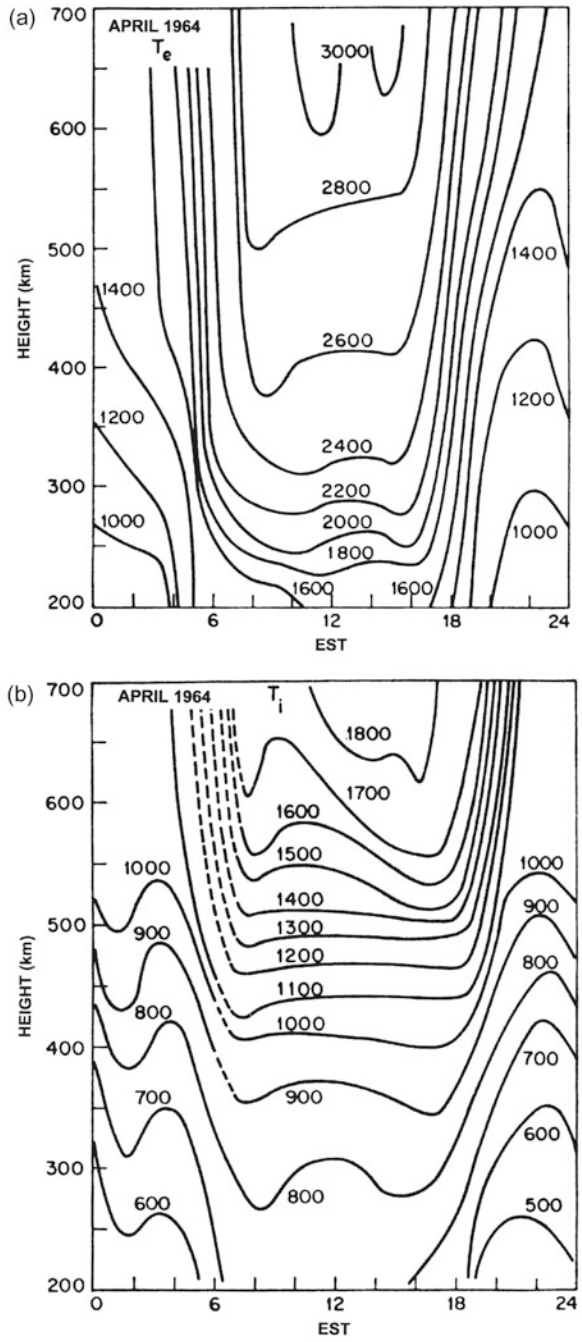


Figure 4.5. Diurnal variation of the electron (T_e) (left) and ion (T_i) (right) temperatures between 200 and 700 km observed at mid-latitude by an incoherent scatter radar. (From Evans, 1967.)

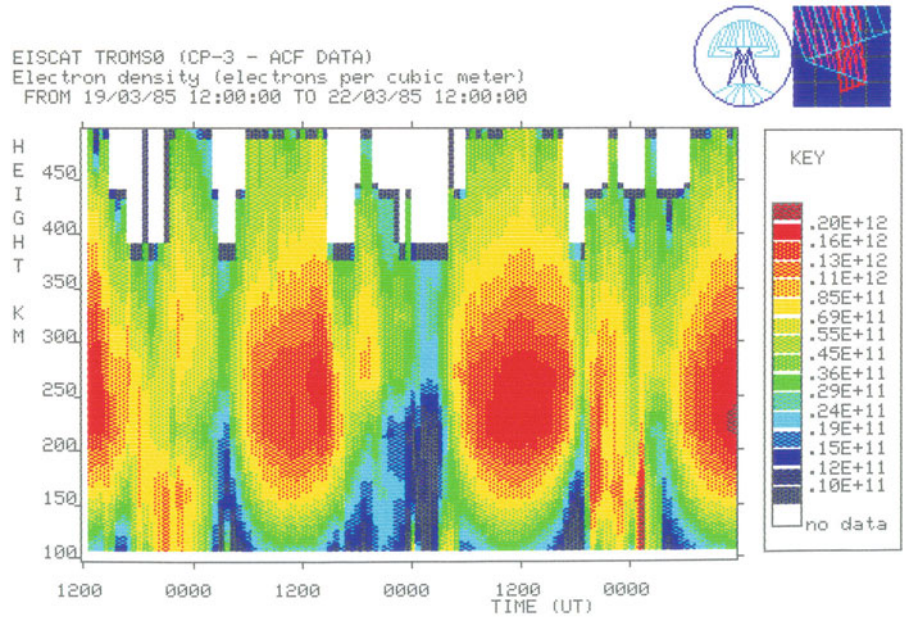


Figure 4.6. Electron density observed between 100 and 500 km altitude by the European Incoherent Scatter Radar (EISCAT) as a function of time for three consecutive days in March 1985. The scale is given to the right in units per cubic meter. (Courtesy C. Hall, 1989.)

Radar (EISCAT). Figures 4.6, 4.7, and 4.8 show some typical observations of electron densities and temperatures, and ion temperatures, respectively, observed between 100 km and 500 km above ground for three consecutive days in March 1985. Electron densities (Figure 4.6) are strongly enhanced during daytime between 150 and 400 km and reduced especially below 250 km at midnight on March 20 (the middle day). These variations are typical of ionization due to solar radiation. In the late evening on March 19 (the first day) and during the midnight hours between March 21 and 22 there are strong enhancements in electron densities reaching down to 100 km and lower. These enhancements are due to ionization caused by precipitating particles or auroral particles.

The electron temperatures shown in Figure 4.7 in the same format as the electron densities in Figure 4.6 display a similar behavior with strong enhancements at daytime and low temperatures or cooling at night-time. Strong temperature enhancements are also seen above 150 km during the events of auroral precipitation discussed above. These temperature enhancements are due to interactions between the beam of auroral particles and background plasma in the ionosphere. The auroral particle beam is found to enhance electron temperature in the F-region and electron density in the E-region.

Finally, the ion temperature between 100 and 500 km is displayed in Figure 4.8 for the same days as the two previous figures. Only small variations are found

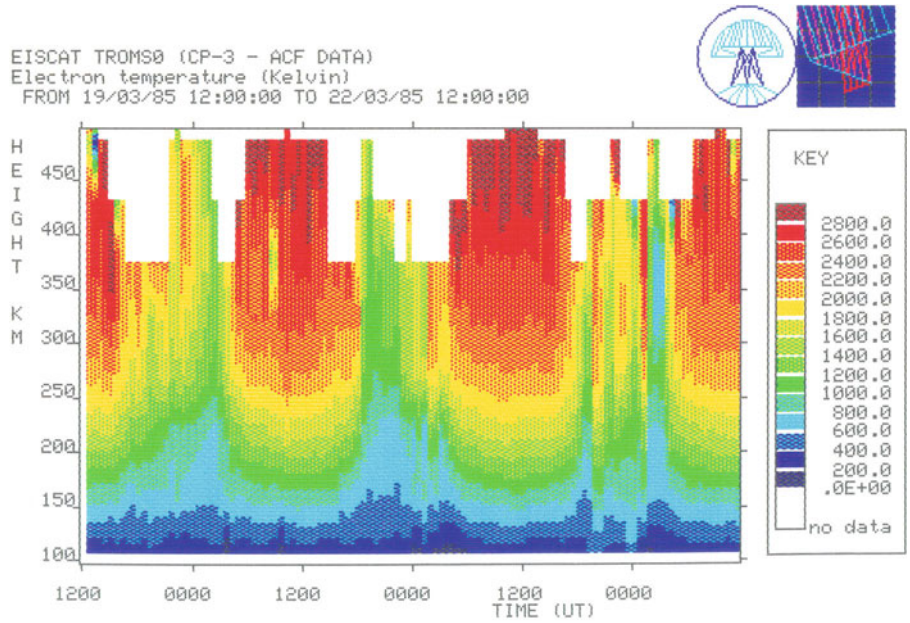


Figure 4.7. Same format as for Figure 4.6 but showing the electron temperature for the same days. The scale is given to the right in kelvins. (Courtesy C. Hall, 1989.)

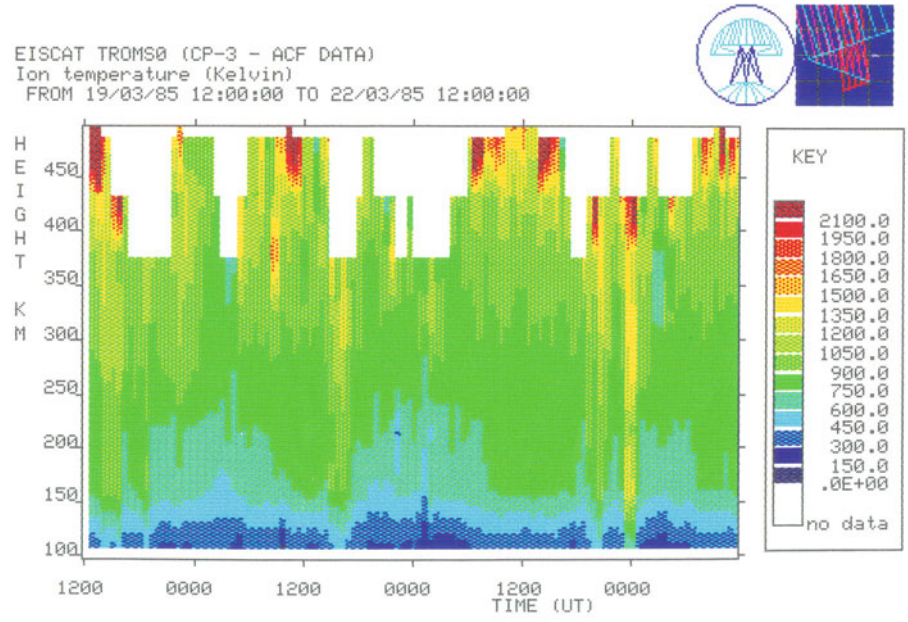


Figure 4.8. Same format as for Figure 4.6 but showing the ion temperature for the same days. The scale is given to the right in kelvins. (Courtesy C. Hall, 1989.)

in ion temperature in this height region. This is due to the fact that ions are much more strongly coupled to the neutral atmosphere by collisions and, therefore, the temperature of ions will be more dominated by the neutral temperature which does not change so strongly at auroral latitudes at equinox. The event observed close to midnight on March 21, however, is related to the event of precipitating auroral particles. This event is associated with a large enhancement in ion velocities, which creates strong ion heating due to collisions with neutrals. It is characteristic for such events that the ion temperature is enhanced through the whole ionosphere between 100 and 500 km since ion velocities are increased in the whole height interval.

Before we go into more detail about the different aspects of ionospheric composition, temperature, and dynamics, we will discuss how ionization in the upper atmosphere is produced by solar radiation.

4.2 THE IONIZATION PROFILE OF THE UPPER ATMOSPHERE

We will assume that the target atmosphere solar radiation penetrates is a horizontally stratified medium, and that this medium obeys the equation of state for an ideal gas. For an atmospheric gas in hydrostatic equilibrium the particle density (m^{-3}) will decrease by altitude as given by (2.11):

$$n = n_0 \exp(-z/H) \tag{4.1}$$

where H is the scale height, and n_0 the density at some reference height $z_0 = 0$.

Incoming solar radiation at wavelength λ will have an intensity $I(\lambda, z)$ at altitude z. The intensity is measured in eV/m^2 s. There will be a cross-section $\sigma(\lambda)$ (m^2) for ionizing of neutral particles in the atmosphere by radiation at wavelength λ. Solar radiation at wavelength λ will therefore ionize a number of neutral particles per cubic meter and second.

Now, if this radiation with intensity $I(\lambda)$ has passed a distance s through the atmosphere, the intensity will be reduced by dI if it passes an additonal infinitesimal distance ds (Figure 4.9) through a slab of the atmosphere. This reduction has to be proportional to the intensity of the radiation, the cross-section for ionization, and the number of targets that can be ionized. The reduction of I per unit distance is as follows:

$$\frac{dI}{ds} = -n \cdot \sigma \cdot I \tag{4.2}$$

Let us assume that for every unit energy absorbed of radiation, there will be formed a number C electrons. Then the production of electrons per cubic meter and second can be expressed as:

$$q = C \cdot \sigma \cdot n \cdot I = -C \cdot \frac{dI}{ds} \tag{4.3}$$

C is called the ionization efficiency. For atomic species the ionization efficiency C

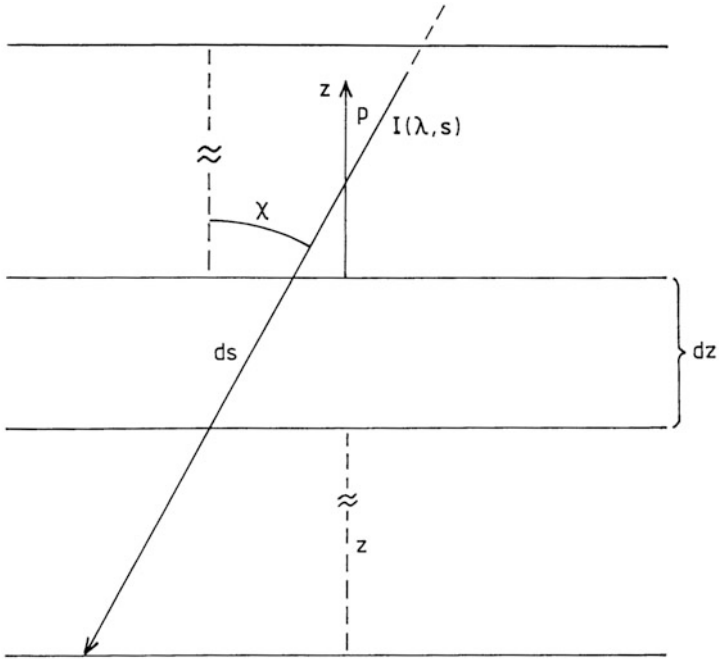


Figure 4.9. Illustration of the geometry related to solar irradiation (I) when penetrating a slab of thickness dz in the Earth's atmosphere at a zenith angle χ. The distance to the source and to the ground are s and z, respectively.

is unity so all the energy of the radiation goes into producing ion–electron pairs; for molecules, however, $C < 1$.

We now realize that since n increases and I decreases as we go down in the atmosphere, the product $n \cdot I$ must reach a maximum somewhere (Figure 4.10), and this maximum is found where:

$$\frac{dq}{ds} = -C \cdot \sigma \cdot \left(I \cdot \frac{dn}{ds} + n \cdot \frac{dI}{ds} \right) = 0$$

since C and σ are constants. By introducing the index m for the maximum we get:

$$\frac{1}{n_m} \left(\frac{dn}{ds} \right)_m + \frac{1}{I_m} \left(\frac{dI}{ds} \right)_m = 0 \tag{4.4}$$

If radiation falls in towards the atmosphere by an angle χ with respect to the zenith (Figure 4.9), we see that

$$ds = -\frac{dz}{\cos \chi} \tag{4.5}$$

and therefore

$$\frac{1}{n} \frac{dn}{ds} = -\frac{1}{n} \frac{dn}{dz} \cos \chi \tag{4.6}$$

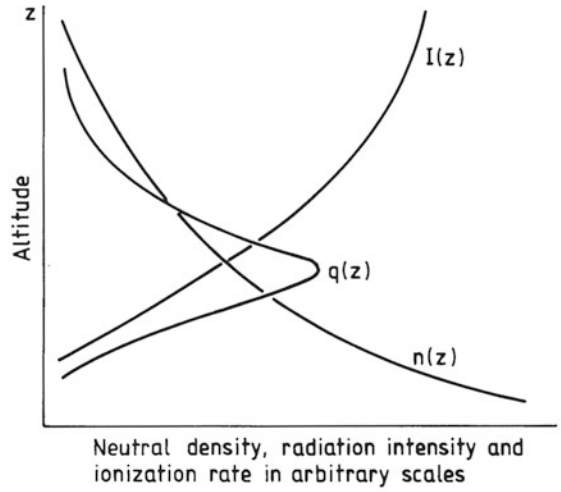


Figure 4.10. Solar irradiation of intensity I decreases due to absorption downward in the Earth's atmosphere where the density increases toward the Earth. A maximum ionization rate q therefore occurs at some altitude.

And by inserting n from (4.1) into (4.6) we get:

$$\frac{1}{n}\frac{dn}{ds} = \frac{\cos\chi}{H} \tag{4.7}$$

at any distance s. In particular, at the production maximum we then have

$$\frac{1}{n_m}\left(\frac{dn}{ds}\right)_m = \frac{\cos\chi}{H} \tag{4.8}$$

From (4.2) we find at this maximum

$$\frac{1}{I_m}\left(\frac{dI}{ds}\right)_m = -\sigma \cdot n_m \tag{4.9}$$

Inserting (4.8) and (4.9) into (4.4) we get

$$\frac{\cos\chi}{H} - \sigma \cdot n_m = 0$$

and then finally

$$\sigma \cdot H \cdot n_m \cdot \sec\chi = 1 \tag{4.10}$$

where $\sec\chi = 1/\cos\chi$. We know from earlier (equation (2.12)) that for an atmosphere with a constant scale height

$$n_0 \cdot H = \mathcal{N}$$

where $\mathcal{N}$ is the total number per unit area between the reference height and infinity. Especially at maximum ionization, therefore

$$H \cdot n_m = \mathcal{N}_m$$

where $\mathcal{N}_m$ is the number of neutral particles above a unit area at height z_m in the atmosphere. Therefore, we get from (4.10)

$$\sigma \cdot \mathcal{N}_m \cdot \sec \chi = 1 \tag{4.11}$$

Inserting (4.5) into (4.2) we find:

$$\frac{1}{I}\frac{dI}{ds} = -\frac{1}{I}\frac{dI}{dz}\cos \chi = -\sigma \cdot n = -\sigma \cdot n_0 \exp\left(-\frac{z}{H}\right)$$

and

$$\frac{dI}{I} = +\sigma \cdot n_0 \exp\left(-\frac{z}{H}\right) \sec \chi \, dz$$

For $z = \infty$, $I = I_\infty$; that is, at the source of radiation

$$\int_{I_\infty}^{I} \frac{dI}{I} = \sigma \cdot n_0 \cdot \sec \chi \int_{\infty}^{z} \exp\left(-\frac{z}{H}\right) dz$$

and

$$\ln \frac{I}{I_\infty} = -\sigma \cdot n \cdot H \cdot \sec \chi \tag{4.12}$$

At the height of maximum ionization, therefore:

$$\ln \frac{I_m}{I_\infty} = -\sigma \cdot n_m \cdot H \cdot \sec \chi = -1 \tag{4.13}$$

and

$$I_m = I_\infty / e \tag{4.14}$$

The intensity of the radiation has therefore decreased to $1/e$ at the height of the ion production maximum. In general, however,

$$I = I_\infty \exp(-\sigma \cdot n \cdot H \cdot \sec \chi) = I_\infty \exp(-\tau)$$

where $\tau = \sigma \cdot n \cdot H \cdot \sec \chi$ is the optical depth, and for the height at maximum ionization this optical depth is:

$$\tau_m = \sigma \cdot n_m \cdot H \cdot \sec \chi = \sigma \cdot n_0 \exp(-z_m/H) \cdot H \cdot \sec \chi = 1$$

and

$$\exp(z_m/H) = \sigma \cdot n_0 \cdot H \cdot \sec \chi$$

Now, for an overhead Sun ($\chi = 0^\circ$)

$$\exp(z_{m,0}/H) = \sigma \cdot n_0 \cdot H$$

Therefore

$$\exp(z_m/H) = \exp(z_{m,0}/H) \sec \chi$$

and

$$\frac{z_m}{H} = \frac{z_{m,0}}{H} + \ln \sec \chi \tag{4.15}$$

The height of the maximum ion production increases as the zenith angle increases, and the lowest height ($z_{m,0}$) it can reach is for overhead Sun. Combining (4.12),

(4.13), and (4.10) we find:

$$\ln \frac{I}{I_m} = -(n - n_m)\sigma \cdot H \cdot \sec \chi = \left(1 - \frac{n}{n_m}\right)$$

and finally

$$\frac{I}{I_m} = \exp\left[1 - \frac{n}{n_m}\right] \tag{4.16}$$

This shows the relationship between the intensity of the radiation at a given neutral density relative to the intensity and neutral density at the maximum of ionization.

Now, neither the ionization cross-section nor the scale height is constant by altitude, therefore a more correct definition of optical depth would be:

$$\tau_\lambda(z) = -\int_\infty^z \sigma_\lambda(z')\, dz' \sec \chi$$

Since each wavelength in the solar spectrum will be absorbed differently at different heights in the atmosphere, each wavelength will have its own optical depth. $\sec \chi$ will always be larger than 1 for a solar zenith angle between 0° and 90°. Unit optical depth will therefore be reached at the lowest altitude for an overhead Sun. This is illustrated in Figure 4.11, where the height in the atmosphere is displayed where wavelengths below 3,000 Å (300 nm) have passed one optical depth

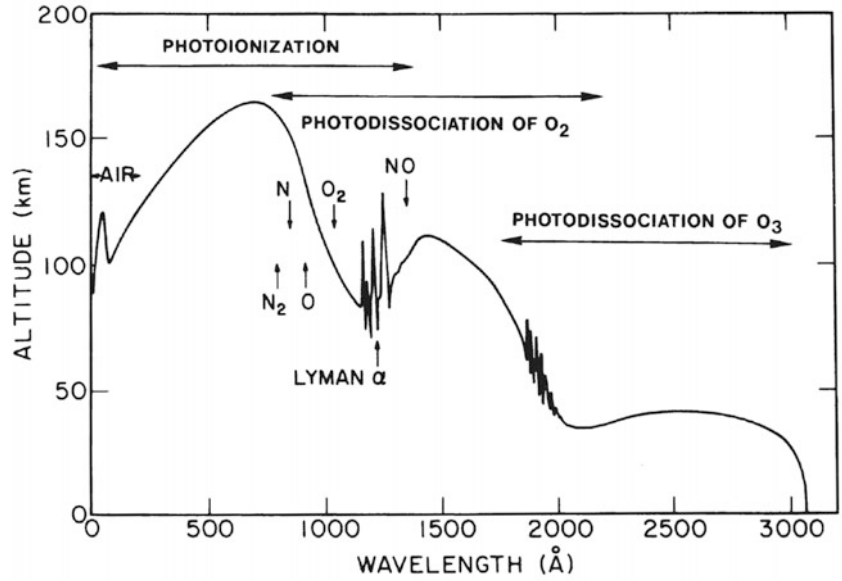


Figure 4.11. The altitude of unit optical depth in the wavelength region below 3,000 Å (300 nm). This corresponds to the level where maximum energy is dissipated at each wavelength. The arrows indicate at what wavelength region the most typical ionization and dissociation processes take place. (From Giraud and Petit, 1978.)

Table 4.1. Ionization and dissociation threshold energies and wavelengths

Species	Ionization		Dissociation	
	V_p (eV)	λ (nm)	V_p (eV)	λ (nm)
N_2	15.58	79.6	9.76	127.0
O_2	12.08	102.6	5.12	242.2
O	13.61	91.1		
N	14.54	85.3		
NO	9.25	134.0	6.51	190.5
H	13.59	91.2		
He	24.58	50.4		

from the source. As already shown, this is equivalent to the height at which the intensity of each wavelength has been reduced to $1/e$ times the intensity at the impact of the uppermost atmosphere.

We notice in particular that there are a few spectral lines that are absorbed very strongly with respect to some neighbor lines, and some lines, especially Ly_α at 1,215 Å (121.5 nm), are reaching deeper down in the atmosphere than lines close by in the spectrum. The Ly_α line, by the way, penetrates all the way down to the D-region where it ionizes NO.

Different constituents in the atmosphere have different ionization threshold energies. In Table 4.1 we have listed some of the most abundant species and given their ionization and dissociation potentials in electronvolts. These potentials are also converted to wavelengths if ionization or dissociation is due to radiation quanta according to

$$V_p = h\nu = h\frac{c}{\lambda}$$

where V_p is the characteristic threshold energy. Only quanta with energies larger than V_p or wavelengths shorter than hc/V_p can ionize or dissociate the given species. From Table 4.1 we therefore notice that radiation in the X-ray region (1–17.0 nm) and in the extreme ultraviolet regime (17.0–175.0 nm) will be especially significant to ionization process in the atmosphere.

In general, the energy of the ionizing radiation $h\nu$ will be much larger than V_p, and the excess energy will then be left by the ejected photoelectron partly as kinetic energy E_e and partly as energy of the excited ion E_λ^* so that

$$h\nu = V_p + E_\lambda^* + E_e$$

Table 4.2. Average energy for ion-pair production.

Target species	*Average energy for ion-pair production $\bar{\varepsilon}$ in eV*
N_2	35
O_2	32
O	27
H_2	36
He	45
Air	34

If E_e is higher than V_p, the photoelectron can yield secondary or higher order ionization during the course of its energy degradation due to further collisions.

According to laboratory experiments it is found that the mean energy $\bar{\varepsilon}$ lost per impact ionization is almost constant as long as $E_e \gg V_p$. The total number of free electrons produced per photoionization is then given by

$$N = 1 + \frac{h\nu - \bar{V}_p}{\bar{\varepsilon}}$$

where $\bar{V}_p$ is the mean ionization energy of the target atoms and molecules. It is customary to set $\bar{V}_p = 15\,\text{eV}$ and $\bar{\varepsilon} = 34\,\text{eV}$ in air. $\bar{\varepsilon}$ is larger than $\bar{V}_p$ since a photoelectron can lose its kinetic energy without ionization in the collisional process. Table 4.2 gives values of the average energy $\bar{\varepsilon}$ in electronvolts for ion-pair production of different species.

Typical ionization cross-sections in the EUV regime are of the order of 10^{-17}–$10^{-18}\,\text{cm}^2$ (10^{-21}–$10^{-22}\,\text{m}^2$). The variation of the cross-section for ionization of O_2 for wavelengths below 2,000 Å (200 nm) is given in Figure 4.12.

For an overhead Sun we therefore typically find that unit optical depth for EUV radiation in the Earth's atmosphere corresponds to:

$$\frac{1}{\sigma} = n \cdot H = 10^{22}\,\text{m}^{-2}$$

From Table 4.3 we notice that $n \cdot H \approx 10^{22}\,\text{m}^{-2}$ at about 110 km, which explains why so much of solar EUV emission is absorbed above 100 km.

Photodissociation, however, has a smaller cross-section of the order of $10^{-24}\,\text{m}^2$. From Table 4.3 we notice that $n \cdot H \approx 10^{24}\,\text{m}^{-2}$ will be close to 80 km in the mesosphere. The collisional cross-section for O_2 is reduced to $10^{-27}\,\text{m}^2$ in the Herzberg continuum (202.6–242.4 nm). This is the reason these wavelengths can penetrate so deeply into the stratosphere where $n \cdot H = 10^{27}\,\text{m}^{-2}$ and where O_2 can dissociate to form O atoms which again can take part in the creation of ozone by the three-body reaction.

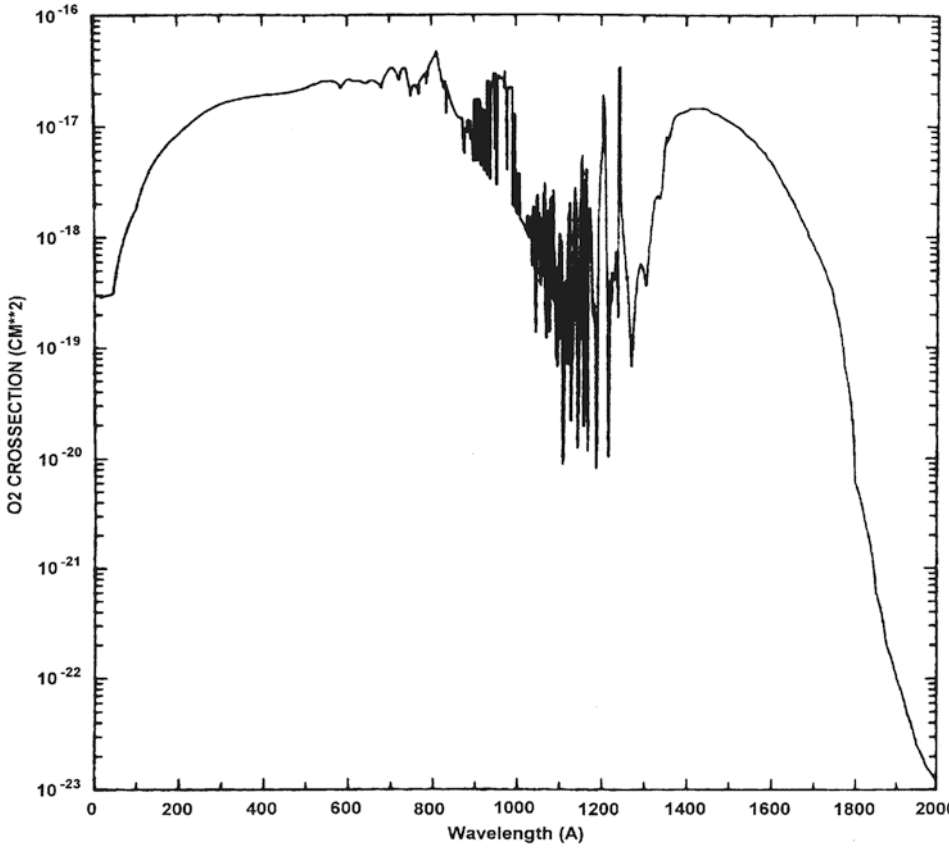


Figure 4.12. Absorption cross-sections for O_2. (From Roble and Emery, 1983.)

4.3 IONIZATION PROFILES

The ion production rate at maximum is according to (4.3), (4.10), and (4.14) given by

$$q_m = C \cdot \sigma \cdot n_m \cdot I_m = C \cdot \sigma \cdot n_m \cdot \frac{I_\infty}{e} = \frac{C \cdot I_\infty}{e \cdot H} \cos \chi \tag{4.17}$$

For an overhead Sun we find ($\chi = 0$):

$$q_{m,0} = \frac{C \cdot I_\infty}{e \cdot H}$$

and therefore

$$q_m = q_{m,0} \cos \chi \tag{4.18}$$

Production at maximum can never be larger than for an overhead Sun, and it decreases as $\cos \chi$ for increasing zenith angle χ.

Table 4.3. Some typical values of parameters below 120 km in the atmosphere

Altitude H (km)	*Temperature* t (K)	*Scale height* H (m)	*Concentration* n (m^{-3})	*Column density* nH (m^{-2})
0	288	8.40(03)*	2.55(25)*	2.14(29)*
5	256	7.50(03)	1.53(25)	1.15(29)
10	223	6.50(03)	8.61(24)	5.60(28)
15	217	6.40(03)	4.04(24)	2.59(28)
20	217	6.40(03)	1.85(24)	1.18(28)
25	222	6.50(03)	8.33(23)	5.41(27)
30	227	6.60(03)	3.83(23)	2.53(27)
35	237	6.90(03)	1.74(23)	1.20(27)
40	250	7.30(03)	6.67(22)	4.87(26)
45	264	7.70(03)	4.12(22)	3.17(26)
50	271	7.90(03)	2.14(22)	1.69(26)
55	261	7.60(03)	1.19(22)	9.04(25)
60	247	7.20(03)	6.45(21)	4.64(25)
65	233	6.80(03)	3.42(21)	2.33(25)
70	220	6.40(03)	1.71(21)	1.09(25)
75	208	6.10(03)	8.36(20)	5.10(24)
80	198	5.80(03)	4.03(20)	2.34(24)
85	189	5.50(03)	1.72(20)	9.46(23)
90	187	5.50(03)	6.98(19)	3.84(23)
95	188	5.50(03)	2.93(19)	1.61(23)
100	195	5.70(03)	1.19(19)	6.78(22)
105	209	6.10(03)	5.20(18)	3.17(22)
110	240	7.00(03)	2.14(18)	1.50(22)
115	300	8.80(03)	9.66(17)	8.50(21)
120	360	1.05(04)	5.03(17)	5.28(21)

*Read 8.40(03), for example, as 8.40×10^3.

For ion production rates we can make the following ratio by using (4.3) and (4.17) and introducing (4.16):

$$\frac{q}{q_m} = \frac{C \cdot n \cdot \sigma \cdot I}{C \cdot n_m \cdot \sigma \cdot I_m} = \frac{n}{n_m} \exp\left[1 - \frac{n}{n_m}\right]$$

Based on (4.1) we can also form the ratio

$$\frac{n}{n_m} = \exp\left(-\frac{z - z_m}{H}\right) \tag{4.19}$$

The production q at height z can then be expressed as

$$q = q_m \exp\left(-\frac{z - z_m}{H}\right) \exp\left[1 - \exp\left(-\frac{z - z_m}{H}\right)\right]$$

or

$$q = q_m \exp\left[1 - \frac{z - z_m}{H} - \exp\left(-\frac{z - z_m}{H}\right)\right] \tag{4.20}$$

Substituting the normalized height reduced to the height of maximum ionization

$$y = \frac{z - z_m}{H}$$

we get:

$$q = q_m \exp[1 - y - \exp(-y)]$$

We notice that independently of the zenith angle χ the ionization profile will maintain its form.

Inserting z_m/H from (4.15) into (4.20) we get:

$$\begin{aligned} q &= q_m \exp\left[1 - \frac{z - z_{m,0}}{H} + \ln \sec \chi - \exp\left(-\frac{z - z_{m,0}}{H} + \ln \sec \chi\right)\right] \\ &= q_m \sec \chi \exp\left[1 - \frac{z - z_{m,0}}{H} - \sec \chi \exp\left(-\frac{z - z_{m,0}}{H}\right)\right] \end{aligned}$$

By then applying (4.18) we have:

$$q = q_{m,0} \exp\left[1 - \frac{z - z_{m,0}}{H} - \sec \chi \exp\left(-\frac{z - z_{m,0}}{H}\right)\right]$$

By introducing the normalized height $x = (z - z_{m,0})/H$ reduced to the height of maximum ionization for an overhead Sun, we find

$$q = q_{m,0} \exp(1 - x - \sec \chi \exp(-x)) \tag{4.21}$$

For very large x ($z \gg z_{m,0}$) the profile takes the form

$$q \approx q_{m,0} \exp(-z/H)$$

Well above the maximum in the ionization profile the profile itself decays by altitude as the density of the target atmosphere. This holds for $x > 2$ (i.e., at a distance more than two scale heights above $z_{m,0}$). This relation arises because at those heights the intensity of the radiation is only weakly reduced, and the rate of production is essentially proportional to the density of the target gas.

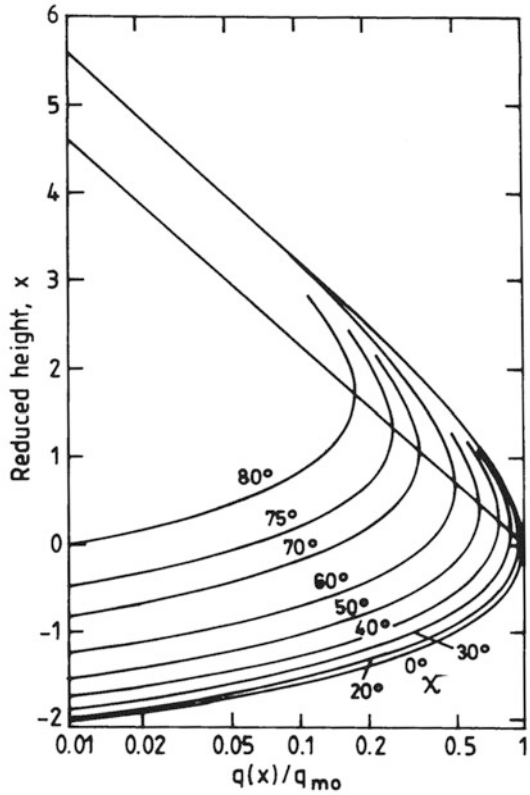


Figure 4.13. Chapman production profiles for different solar zenith angles. (After Van Zandt and Knecht, 1964.)

For very large negative x ($z \ll z_{m,0}$) the production rate takes the form

$$q = q_{m,0} \exp\left[-\sec \chi \exp\left(-\frac{z - z_{m,0}}{H}\right)\right]$$

and the profile decreases very rapidly with height below the altitude of peak production.

Figure 4.13 illustrates the ionization profiles for different zenith angles between 0° and 80°. The altitude parameter used is the reduced height x. The profiles all have the same form, the peak production rate decreases, and the height of the maximum increases when the zenith angle increases in agreement with the mathematical treatment shown above.

For a real target atmosphere and a real solar spectrum it is a rather time-consuming task to derive the final ion production profile. The upper atmosphere is composed of different gases with different scale heights, and the solar radiation spectrum consists of a myriad of lines and bands with different intensities. The approach is therefore to assume a neutral atmosphere model that has the

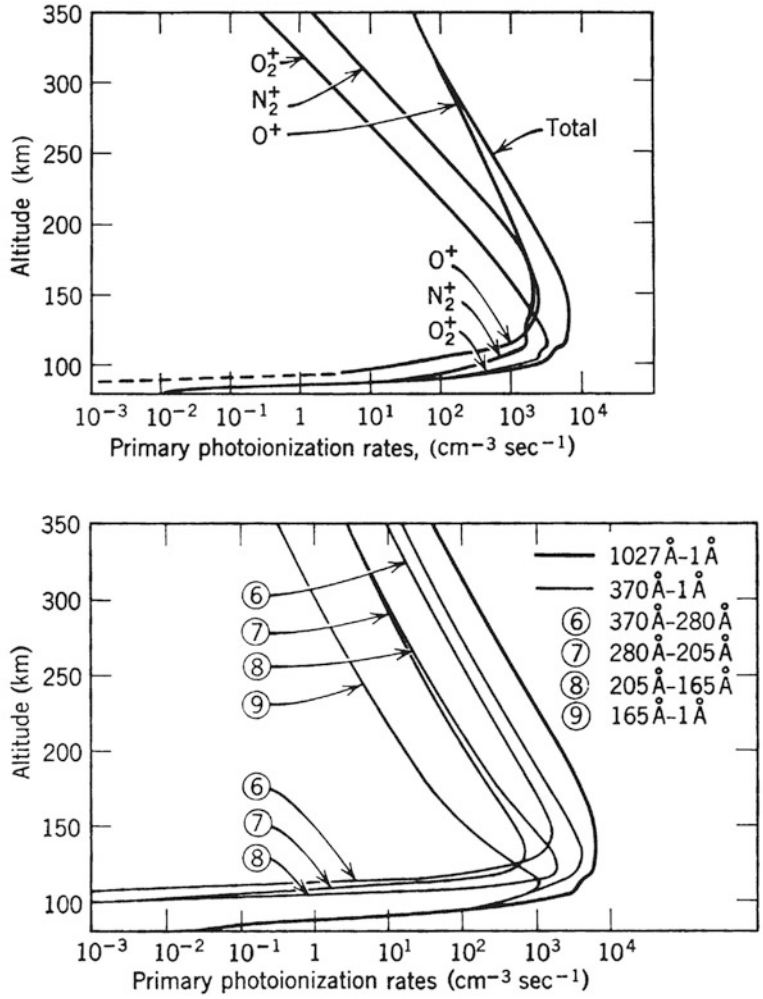


Figure 4.14. Calculated ionization rates in the E- and F-regions. (From Hinteregger *et al.*, 1965.)

height distribution of some of the major gases in question, and then for a finite number of wavelengths derive the individual production profiles. Examples of such a procedure is shown in Figure 4.14. We notice that the shorter the wavelength, the deeper down into the atmosphere the corresponding production maximum, and wavelengths in the X-ray regime (1–17.0 nm) can ionize below 100 km. Individual ion density profiles, being a result of the ionization of the composed radiation spectrum, are shown in the upper panel in Figure 4.14 together with the total ionization profile. The production of O^+ ions is found to be dominant above 200 km and of O_2^+ ions below 130 km according to these model calculations.

4.4 THE RECOMBINATION PROCESS

The ionization thus being formed will immediately be involved in chemical reactions to re-establish the equilibrium condition before the ionization took place. Some of the ionization will also be carried away by transport mechanisms such as diffusion, neutral winds, and electric fields.

The continuity equation for an ion species formed can therefore be written as

$$\frac{\partial n_i}{\partial t} = q_i - l_i - \nabla \cdot (\mathbf{v}_i n_i) \tag{4.22}$$

where n_i is the density of the ion species, q_i is the production, l_i is the loss by chemical and photochemical processes, and the last term is the divergence due to transport phenomena through a volume of interest (the convection term). In the following section we will discuss the loss term l_i for different regions of the ionosphere and study the consequences of this. Later on, we will study the transport phenomena pertinent to the ionized species.

Let us start by neglecting the transport term and assume photochemical equilibrium

$$q_i = l_i$$

Let us further assume that electrons (e) recombine directly with positive ions (X^+) to form a neutral species together with an accompanying radiation of a photon, $h\nu$

$$X^+ + e \rightarrow X + h\nu$$

the so-called radiative recombination process. Another possibility for reducing the number of free electrons is by dissociative recombination of a molecule XY^+ as follows:

$$XY^+ + e \rightarrow X + Y$$

In fact, the latter is more likely to take place in the upper atmosphere because it results in two particles whereas the first reaction results in only one and a photon. The needs for conservation of energy and momentum are not so easily satisfied in the radiative recombination process, and therefore dissociative recombination is more likely. Calculated magnitudes of the dissociative recombination rate are of the order of $10^{-13}\ \mathrm{m^3\ s^{-1}}$, while the rate for radiative recombination is much smaller and of the order of $10^{-18}\ \mathrm{m^3\ s^{-1}}$.

Whether the radiative recombination process or the dissociative recombination process is dominant, we will in both cases for simplicity assume charge neutrality so that the number density of positive ions $[X^+]$ or $[XY^+]$ are equal to the number density n_e of free electrons.

$$[X^+] = n_e \qquad \text{or} \qquad [XY^+] = n_e$$

The loss rate will be proportional to the product of the electron and ion densities. The proportionality factor α is called the "recombination coefficient", thus

$$l_i = \alpha n_e^2$$

Since production is given by (4.21), we find for photochemical equilibrium

$$l_i = q_i - q_{m,0} \exp[1 - x - \sec \chi \exp(-x)] = \alpha n_e^2$$

Electron density at a given height z is therefore given by

$$n_e(z) = \sqrt{\frac{q_{m,0}}{\alpha}} \exp[\tfrac{1}{2}(1 - x - \sec \chi \exp(-x))]$$

where x is the normalized and reduced height parameter as defined above. When neglecting height variations in α, we find that the electron density has a maximum when

$$e^{-x} = \cos \chi$$

showing that the electron density at maximum is given by

$$n_m = \sqrt{\frac{q_{m,0}}{\alpha}} (\cos \chi)^{1/2} = n_{m,0} (\cos \chi)^{1/2}$$

The maximum electron density therefore varies as $(\cos \chi)^{1/2}$ with the zenith angle. This is often called a Chapman α-profile and is representative of the E-region.

If, on the other hand, electrons are lost by attachment to a molecule

$$M + e \rightarrow M^-$$

then the loss rate is proportional to n_e and can be expressed as:

$$l_i = \beta n_e$$

where β is proportional to $[M]$, the density of the neutral molecule M. In equilibrium then

$$q_i = l_i = \beta n_e$$

and the electron density profile is given by applying (4.21) again:

$$n_e = \frac{q_{m,0}}{\beta} \exp(1 - x - \sec \chi \exp(-x))$$

If we neglect height variations in β, the maximum electron density is found when $\exp(-x) = \cos \chi$ and therefore

$$n_m = \frac{q_{m,0}}{\beta} \cos \chi = n_{m,0} \cos \chi$$

The maximum density varies with the zenith angle as $\cos \chi$. This is often called a Chapman β-profile.

It was experimentally demonstrated in the 1920s that the diurnal variation of maximum electron densities in the E- and F-regions followed the solar zenith angle in a $(\cos \chi)^{1/2}$ and $\cos \chi$ law, respectively. When it later was possible to show that these variations could be explained theoretically as a result of ionization of the upper atmosphere by solar radiation, it represented a strong encouragement for scientists developing the new field of ionospheric research. It has also had a profound impact on our understanding of this part of our near environment in space.

The negative ions formed in the attachment process are usually detached by other reactions so that electron loss by attachment can be neglected in the ionospheric E-region and above. In the lower ionosphere, however, it will be of importance.

In the real ionosphere atomic and molecular ions coexist with atoms and molecules and different reactions can take place at the same time. We will now study the importance of these processes at different regions and, in particular, in the transition between E- and F-regions.

From Figure 4.3 we notice that below 150 km molecular ions such as O_2^+ and NO^+ are the dominant species. Above 150 km, however, the O^+ ion is becoming increasingly abundant and will dominate above, say, 200 km.

It is therefore evident that in the E-region dissociative recombination with molecular ions will be the most important loss mechanism for electrons, while in the topside ionosphere loss mechanisms related to ionic oxygen atoms must dominate. In between these regions in the so-called "F_1-region", one or the other loss mechanisms may dominate, depending on the relative values of reaction rate coefficients.

4.5 THE O^+ DOMINANT IONOSPHERE

In the F-region (say, above 150 km), the dominant ions formed by solar irradiance are the O^+ ions. In the following we therefore will assume that these ions are the only ion species resulting from solar radiation and that the other ions are produced as a result of chemical reactions.

The atomic oxygen ions when produced can be lost by several reactions. One possibility is radiative recombination with an electron as follows:

$$O^+ + e \xrightarrow{\alpha_r} O + h\nu$$

Here we denote α_r (Table 4.4) as the radiative recombination coefficient for this type of process. This process is, however, as already mentioned, rather slow for the energy and momentum balance that has to be maintained. A more rapid loss process for the O^+ ions is through a chain of reactions. First, O^+ recombines with the abundant N_2 molecule and forms a NO^+ ion with a rate coefficient k_1 (Table 4.4) as follows:

$$O^+ + N_2 \xrightarrow{k_1} NO^+ + N$$

The NO^+ ion recombines with an electron and dissociates:

$$NO^+ + e \xrightarrow{\alpha_1} N + O$$

where we denote α_1 (Table 4.4) as the relevant dissociative recombination coefficient.

Table 4.4. Typical reaction rates relevant for ionospheric processes as indicated

Processes	*Reaction rates* (m^3/s)
Dissociative recombination	
$NO^+ + e \rightarrow N + O$	$\alpha_1 = 2.1 \times 10^{-13}(T_e/300)^{-0.85}$
$O_2^+ + e \rightarrow O + O$	$\alpha_2 = 1.9 \times 10^{-13}(T_e/300)^{-0.5}$
$N_2^+ + e \rightarrow N + N$	$\alpha_3 = 1.8 \times 10^{-13}(T_e/300)^{-0.39}$
Rearrangement	
$O^+ + N_2 \rightarrow NO^+ + N$	$k_1 = 2 \times 10^{-18}$
$O^+ + O_2 \rightarrow O_2^+ + O$	$k_2 = 2 \times 10^{-17}(T_r/300)^{-0.4}$
$O_2^+ + NO \rightarrow NO^+ + O_2$	$k_3 = 4.4 \times 10^{-16}$
$O_2^+ + N_2 \rightarrow NO^+ + NO$	$k_4 = 5 \times 10^{-22}$
$N_2^+ + O \rightarrow NO^+ + NO$	$k_5 = 1.4 \times 10^{-16}(T_r/300)^{-0.44}$
$N_2^+ + O_2 \rightarrow N_2 + O_2^+$	$k_6 = 5 \times 10^{-17}(T_r/300)^{-0.8}$
Radiative recombination	
$O^+ + e \rightarrow O + h\nu$	$\alpha_r = 7.8 \times 10^{-14}(T_e/300)^{-0.5}$
Electron attachment	
$O_2 + M + e \rightarrow O_2^- + M$	$\beta_1 = 10^{-43}(300/T_e)\, e^{-600/T_e}[N_2]$
	$\beta_2 = 10^{-41}[O_2]$
$O + e \rightarrow O^- + h\nu$	$\beta_r = 10^{-21}$
Electron detachment	
$O_2^- + h\nu \rightarrow O_2 + e$	$\rho = 0.33\ sec^{-1}$
$O_2^- + N_2 \rightarrow O_2 + N_2 + e$	$\gamma_1 = 2 \times 10^{-18}(T_r/300)^{1.5} \exp(5 \times 10^{-3}/T_r)$
$O_2^- + O_2 \rightarrow O_2 + O_2 + e$	$\gamma_2 = 3 \times 10^{-16}(T_r/300)^{0.5} \exp(6 \times 10^{-3}/T_r)$
$O_2^- + O \rightarrow O_3 + e$	$\gamma_3 = 5 \times 10^{-16}$
$O^- + O \rightarrow O_2 + e$	$\gamma_4 = 3 \times 10^{-16}$
Ion–ion recombination	
$O_2^+ + O_2^- \rightarrow O_2 + O_2$	$\alpha_{i_1} = 2 \times 10^{-13}$
$O_2^+ + O_2^- + M \rightarrow O_2 + O_2 + M$	$\alpha_{i_2} = 3 \times 10^{-31}(T/300)^{-2.5}[M]$

Another loss process for O^+ ions is through a similar chain of reactions with O_2^+ ions. First, the O^+ ion charge exchanges with an O_2 molecule

$$O^+ + O_2 \xrightarrow{k_2} O_2^+ + O$$

with a rate coefficient for the process given by k_2 (Table 4.4). The O_2^+ ion formed recombines with an electron and dissociates as follows:

$$O_2^+ + e \xrightarrow{\alpha_2} O + O$$

The dissociative recombination rate for this process is denoted α_2 (Table 4.4).

In a quasi-chemical photoequilibrium condition the production of atomic oxygen ions must be equal to the loss of these ions

$$q(O^+) = l(O^+) = (k_1 \cdot [N_2] + k_2 \cdot [O_2]) \cdot [O^+] \tag{4.23}$$

when we neglect radiative recombination. For molecular oxygen ions we have a similar equilibrium condition when neglecting the production of O_2^+ due to solar radiation

$$q(O_2^+) = k_2[O_2] \cdot [O^+]$$
$$l(O_2^+) = \alpha_2[O_2^+] \cdot n_e$$

and since $q(O_2^+) = l(O_2^+)$

$$[O_2^+] = \frac{k_2}{\alpha_2}\frac{[O_2]}{n_e}[O^+]$$

and for the NO^+ ions when no solar production is assumed

$$q(NO^+) = k_1[O^+] \cdot [N_2]$$
$$l(NO^+) = \alpha_1[NO^+] \cdot n_e$$

and since $q(NO^+) = l(NO^+)$

$$[NO^+] = \frac{k_1}{\alpha_1}\frac{[N_2]}{n_e}[O^+]$$

Since the number of positive and negative charges must be equal we have:

$$n_e = [O^+] + [NO^+] + [O_2^+]$$

and inserting for $[NO^+]$ and $[O_2^+]$ just derived we find:

$$n_e = \left(1 + \frac{k_1}{\alpha_1}\frac{[N_2]}{n_e} + \frac{k_2}{\alpha_2}\frac{[O_2]}{n_e}\right)[O^+]$$

Solving for $[O^+]$ and inserting into (4.23) gives:

$$q(O^+) = \left(\frac{k_1[N_2] + k_2[O_2]}{1 + \frac{k_1}{\alpha_1}\frac{[N_2]}{n_e} + \frac{k_2}{\alpha_2}\frac{[O_2]}{n_e}}\right) n_e = \beta' \cdot n_e \tag{4.24}$$

where β' is the loss rate given by

$$\beta' = \frac{k_1[N_2] + k_2[O_2]}{1 + \frac{k_1}{\alpha_1}\frac{[N_2]}{n_e} + \frac{k_2}{\alpha_2}\frac{[O_2]}{n_e}}$$

The values for k_1 and k_2 are of the order of 2×10^{-18} m^3 s (see Table 4.4), while α_1 and α_2 are of the order of 10^{-13} m^3 s, and we see that for altitudes above 250 km or so where $[N_2] < 10^{15}$ m^{-3} and $[O_2] \approx 10^{14}$ m^3, $\alpha_1 \cdot n_e$ and $\alpha_2 \cdot n_e$ will be

much larger than $k_1[N_2]$ and $k_2[O_2]$, respectively, for an electron density of 10^{11} m^{-3} or less. It is therefore a good approximation to set

$$\beta = k_1[N_2] + k_2[O_2]$$

and since N_2 and O_2 have similar scale heights, we have approximately:

$$\beta = \beta_0 \exp\left[-\frac{z - z_0}{H}\right]$$

where β_0 is the effective loss rate at reference height z_0, and H is the scale height.

If we move lower down in the ionosphere to the lower F-region and the E-region, the reactions involving charge rearrangement between O^+ and O_2 and N_2 will become more abundant because of the increase in molecular neutral density and, therefore, the complete expression for β' must be maintained.

Solving for n_e in (4.24) we get:

$$n_e = \frac{q(O^+)}{\beta'} = q(O^+) \cdot \frac{1 + \left(\frac{k_1}{\alpha_1}[N_2] + \frac{k_2}{\alpha_2}[O_2]\right)\frac{1}{n_e}}{\beta}$$

We can then form

$$n_e = q(O^+)\left[\frac{1}{\beta} + \frac{1}{\alpha_{\text{eff}} n_e}\right]$$

where

$$\frac{1}{\alpha_{\text{eff}}} = \left[\frac{k_1[N_2]}{\alpha_1} + \frac{k_2[O_2]}{\alpha_2}\right]\frac{1}{\beta}$$

Since β decreases by height and $\alpha_{\text{eff}} \cdot n_e$ will increase by height under the F-region maximum, there can be a transition height z_t at which $n_e = n_{et}$ and where

$$\beta_t = \alpha_{\text{eff}} n_{et}$$

Above this height the electron density will vary as q/β. Below this height, however, the electron density varies as $\sqrt{q/\alpha_{\text{eff}}}$.

If the latter situation applied at all heights, the electron density profile would take the form shown in Figure 4.15(b), where a single maximum would occur at height z_m of the production maximum (Figure 4.15(a)). If the first situation applied, then β would decrease by height as shown in Figure 4.15(c), and a "knee" would occur in the electron density profile at height z_m of the production maximum (Figure 4.15(d)). The decay of β by altitude would enhance the electron density above the peak due to a slower loss rate.

Now, if the transition point z_t is above the production maximum, there will still be a maximum in the electron density profile at the peak of production (Figure 4.15(e)), since the loss rate here is still dominated by dissociative recombination. If, however, the transition height z_t is situated below the production maximum at z_m, there would be a "ledge" in the electron density profile at z_t and a "knee" at z_m, since β-loss will dominate over α-loss at z_m.

We have seen from (4.21) that for distances larger than two scale heights above the ionization maximum, the ion production profile decays by altitude with

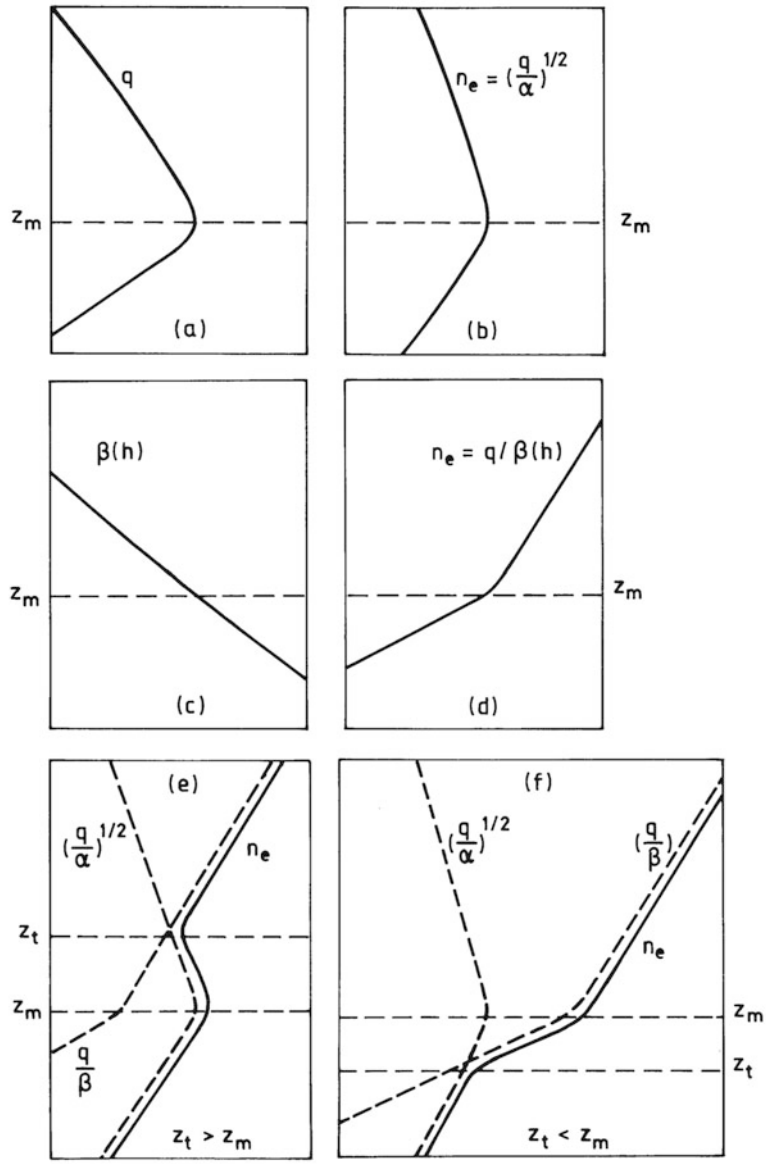


Figure 4.15. Schematic diagram showing the formation of the F_1-layer. (a) The ionization profile q has a maximum at z_m. (b) For a constant recombination constant α by height the electron density profile will have a maximum at z_m. (c) The electron loss rate β decays exponentially with height. (d) For an electron loss process dominated by β the electron density profile n_e would have a knee at z_m. (e) The electron loss due to recombination (α) and attachment (β) are equal at an altitude z_t. If the production maximum at z_m is below z_t, the electron density profile will have a maximum at z_m and a minimum at z_t. (f) If the production maximum at z_m is above z_t, the electron density will have a ledge between z_t and z_m. (After Ratcliffe, 1972.)

the scale height of the background atmosphere (i.e., with the scale height of the oxygen atoms, H_O), since these are the major ionization target above the F-region peak. We then have

$$q(z) \approx \exp\left(-\frac{z}{H_O}\right)$$

Since β is proportional to $k_1[N_2] + k_2[O_2]$, it decays by altitude as the density of N_2 and O_2. As N_2 is the dominant species of these two (Figure 2.2), we assume:

$$\beta(z) \approx \exp\left(-\frac{z}{H_{N_2}}\right)$$

where H_{N_2} is the scale height of the N_2 species. The electron density profile at the topside ionosphere therefore will vary according to the following expression:

$$n_e(z) = \frac{q(z)}{\beta(z)} \propto \exp\left(-\frac{z}{H_O}\right) \cdot \exp\left(\frac{z}{H_{N_2}}\right) = \exp\left[\left(\frac{H_O}{H_{N_2}} - 1\right)\frac{z}{H_O}\right]$$

Since

$$\frac{H_O}{H_{N_2}} = \frac{7}{4}$$

we find

$$n_e(z) \propto \exp\left(\frac{3z}{4H_O}\right)$$

The electron density profile above the F-region peak will increase by a scale height of $1.33 \cdot H_O$ if the ionosphere is in photochemical equilibrium. If our assumptions were appropriate for all regions, the electron density would therefore increase to infinity above the altitude z_t.

4.6 AMBIPOLAR DIFFUSION

There has to be more to it: the ionosphere is simply not in photochemical equilibrium above the F-region peak.

We have to turn our attention to diffusion to see what that can do to the problem.

Let us assume charge neutrality ($n_e = n_i$), that the neutral air is at rest ($\mathbf{u} = 0$), and that no current is present. The last assumption means that $\mathbf{v}_i = \mathbf{v}_e = \mathbf{v}$, where $\mathbf{v}_i$ and $\mathbf{v}_e$ are the ion and electron velocities, respectively. Since the mass of ions is so much greater than the mass of electrons, a charge separation will occur due to the gravity field which creates an electric field. This field, however, will exert a strong force on electrons compared with gravity and the charge separation will be blocked. In particular, for the vertical motion we will have

$$w_i = w_e = w$$

where w_i and w_e are vertical velocity components for the ions and electrons, respectively. We will at present assume no magnetic field and an isothermal

plasma in thermal balance with the neutrals

$$T_e = T_i = T_n = T$$

where T_i, T_e, and T_n are the ion, electron, and neutral temperatures, respectively. The momentum equations in the vertical direction for the ions and electrons now become:

$$n_i m_i \frac{\partial w_i}{\partial t} = -\frac{\partial p_i}{\partial z} - n_i m_i g + n_i eE - n_i m_i \nu_i w_i = 0$$

$$n_e m_e \frac{\partial w_e}{\partial t} = -\frac{\partial p_e}{\partial z} - n_e m_e g - n_e eE - n_e m_e \nu_e w_e = 0$$

when ν_e and ν_i are the electron and ion collision frequencies with the neutrals, respectively and any collision between electrons and ions is neglected. By adding these two equations and leaving the collision terms on the left-hand side:

$$n_e(m_i\nu_i + m_e\nu_e)w = -\frac{\partial}{\partial z}(p_i + p_e) - n_e(m_i + m_e)g$$

Assuming that the ideal gas law applies for the plasma and that T is constant by height:

$$n_e(m_i\nu_i + m_e\nu_e)w = -2kT \cdot \frac{\partial n_e}{\partial z} - n_e(m_i + m_e)g$$

Taking into account that $m_i \gg m_e$ and solving for the vertical particle flux $n_e w$

$$n_e w = -\frac{2kT}{(m_i\nu_i + m_e\nu_e)}\left(\frac{\partial n_e}{\partial z} + \frac{n_e m_i g}{2kT}\right)$$

We notice in this equation that the gravity term has exactly the same form as in the equation of vertical diffusion of the neutral gas (2.21) except that the mass has been changed from m to $\frac{1}{2}m_i$. The scale height of the plasma is therefore given as:

$$H_p = \frac{2kT}{m_i g} = 2H$$

where H is the scale height of the neutral atmosphere (2.6). This is true if neutrals and ions have the same molecular mass, as they have if O^+ is moving in an atomic oxygen-dominated atmosphere or if H^+ is moving in a hydrogen-dominated one. We have noticed that the effect of the polarization electric field is to render the plasma a gas with a mean molecular mass equal to $\frac{1}{2}m_i$.

The diffusion coefficient of the plasma will be given by:

$$D_p = \frac{2kT}{m_i\nu_i + m_e\nu_e} \approx 2D$$

where $m_i\nu_i \gg m_e\nu_e$, D is the diffusion coefficient of the neutral gas (2.20), and D_p is often called the *ambipolar diffusion coefficient* referring to the polarization field between ions and electrons which forces the plasma to move as a whole.

The equation of diffusion (2.22) will now be

$$\frac{\partial n_e}{\partial t} = -\frac{\partial}{\partial z}(n_e w) = \frac{\partial}{\partial z}\left\{ D_p \left(\frac{\partial n_e}{\partial z} + \frac{n_e}{H_p} \right) \right\}$$

Equilibrium can be achieved if

$$\frac{\partial n_e}{\partial z} = -\frac{n_e}{H_p}$$

or

$$n_e = n_{e_0} \exp\left(-\frac{z - z_0}{H_p} \right)$$

where n_{e_0} is the plasma density at a reference height z_0 (e.g., at the F-region maximum). The equilibrium profile of the electron density will therefore be exponential and decay by altitude with a scale height that is twice as large as the scale height of the neutral atmosphere.

Since the diffusion coefficient for $m_i v_i \gg m_e v_e$ is given by:

$$D_p = \frac{2kT}{m_i v_i}$$

it will increase exponentially by altitude as

$$D_p = D_0 \exp\left(\frac{z}{H} \right)$$

since v_i decreases proportionally to neutral density and diffusion becomes more important at greater heights. The electron density profile will, however, decrease by altitude above the F-region peak due to the effect of diffusion.

We notice, however, that if $Te \neq T_i$ and $T_e = cT_i$, where c is a constant, the plasma scale height would be

$$H_p = (1 + c)H$$

and the plasma diffusion coefficient would become ($m_i v_i \gg m_e v_e$):

$$D_p = (1 + c)D$$

The shape of the profile and the effect of diffusion therefore is strongly dependent on the electron–ion temperature ratio. This can indeed be very large (Figure 4.4).

4.7 MULTICOMPONENT TOPSIDE IONOSPHERE

In the topside ionosphere where ions are expected to be in diffusive equilibrium, different species like O^+, H^+, He^+, N^+, etc. are observed. The electric field that occurs will result in a force that is the same for individual species, and it will combine with the gravity force to produce different effective masses for individual species. For some the upward electric field force may be larger than the downward force, and it may appear that some ions have a negative apparent mass and a negative distribution height. Let m_j and n_j be the mass and density of the jth ion

species. The momentum equation when collisions are neglected for these ions in diffusive equilibrium then becomes:

$$-kT_i\frac{\partial n_j}{\partial z} - n_j m_j g + n_j eE = 0$$

when it is assumed that $w_j = w_n$, where w_n is the vertical velocity of neutrals, and T_i is the ion temperature equal for all ion species. The electron momentum equation, when collisions are neglected, becomes:

$$-kT_e\frac{\partial n_e}{\partial z} - n_e eE = 0 \tag{4.25}$$

where we have neglected the gravity force on electrons. T_e is the electron temperature. We now add all equations for all ion species and electrons

$$-kT_i\frac{\partial}{\partial z}\Sigma n_j - kT_e\frac{\partial n_e}{\partial z} - \Sigma n_j m_j g + \left(\Sigma n_j - n_e\right)eE = 0$$

Electrical neutrality requires

$$\Sigma n_j - n_e = 0$$

and by setting the mean mass for the ions $\Sigma n_j m_j / n_e$ equal to m_+ we get:

$$-k(T_i + T_e)\frac{\partial n_e}{\partial z} - n_e m_+ g = 0$$

Letting $T_e = cT_i$, where c is a constant, we finally have:

$$\frac{1}{n_e}\frac{\partial n_e}{\partial z} = \frac{-m_+ g}{k(1+c)T_i}$$

which gives the distribution height for electrons. Solving for $(1/n_e)(\partial n_e/\partial z)$ in (4.25) and inserting

$$\frac{eE}{ckT_i} = \frac{m_+ g}{(1+c)kT_i}$$

and

$$eE = \frac{m_+ gc}{1+c}$$

The electric field is therefore a function of the electron–ion temperature ratio c, and the mean mass of ions. For the jth ion species we then have

$$\frac{1}{n_j}\frac{\partial n_j}{\partial z} = -\frac{m_j g}{kT_i} + \frac{m_+ gc}{(1+c)kT_i} = -\frac{1}{H_j}\left(1 - \frac{c}{1+c}\frac{m_+}{m_j}\right)$$

where $H_j = kT_i/m_j g$. This describes the vertical distribution of the jth ion species, and we see that in diffusive equilibrium this species will have a distribution height different from its scale height

$$\delta_j = H_j\left(1 - \frac{c}{1+c}\frac{m_+}{m_j}\right)^{-1}$$

which depends on the mass of the species with respect to the mean ion mass as

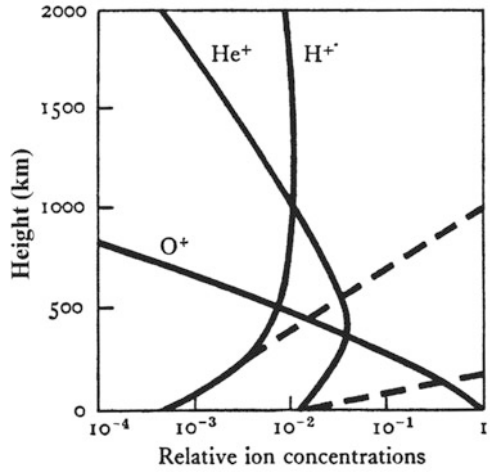


Figure 4.16. Relative ion concentration profiles below 2,000 km altitude for a mixture of H^+, He^+, O^+ ions and electrons in diffusive equilibrium for a temperature equal to 1,200 K. The dashed lines correspond to conditions where photochemical equilibrium prevails. (From Ratcliffe, 1972.)

well as the temperature ratio $c = T_e/T_i$. For $m_j < [c/(1+c)]m_+$, therefore, the distribution height will be negative. In an ionosphere where $T_e > T_i$, the density of the lightest ions will increase with height.

We should notice that δ_j is dependent on the relative mass of the different ion species as well as the temperature ratio c which both change by height. Therefore, the distribution height is not a constant and height variation is not a true exponential. An example of the height distribution of some ion species in the topside ionosphere when c is assumed equal to 1 is given in Figure 4.16.

4.8 DIFFUSION IN THE PRESENCE OF A MAGNETIC FIELD

The high conductivity along a magnetic field line forces plasma to move along the field and to some extent restrain the true vertical motion. The particle flux along the magnetic field line would therefore read as follows:

$$n_e v_{\parallel} = -\frac{2kT}{m_+ v_i}\left(\frac{\partial n_e}{\partial l} + \frac{n_e m_+ g_{\parallel}}{2kT}\right)$$

where l is measured along B (Figure 4.17).

Since $dl = dz/\sin I$ and $g_{\parallel} = g \sin I$, when I is the angle between the magnetic field and the horizontal plane, we find:

$$n_e v_{\parallel} = -\frac{2kT}{m_+ v_i}\left(\frac{\partial n_e}{\partial z}\sin I + \frac{n_e m_+ g}{2kT}\sin I\right) = -D_p \sin I\left(\frac{\partial n_e}{\partial z} + \frac{n_e}{H_p}\right)$$

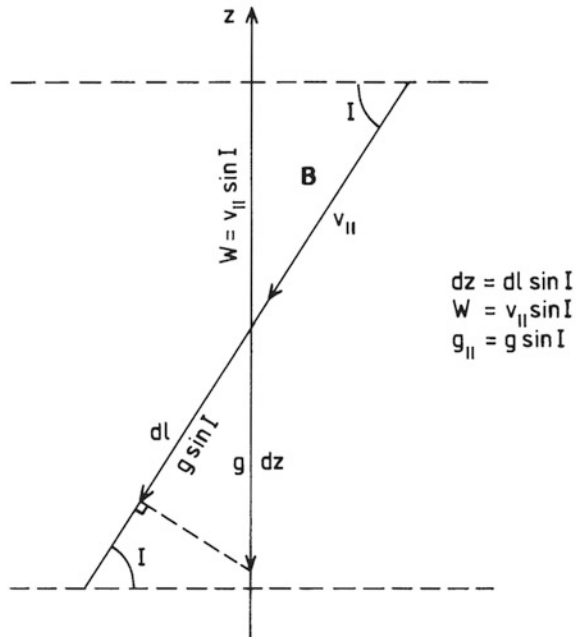


Figure 4.17. The geometry for ionospheric diffusive motion in the presence of a magnetic field making an angle I with the horizontal plane.

Furthermore, we have for the vertical velocity component

$$w = v_{\parallel} \sin I$$

and

$$n_e w = -\sin^2 I \cdot D_p \left(\frac{\partial n_e}{\partial z} + \frac{n_e}{H_p} \right)$$

The vertical diffusion flux increases in magnitude as I increases, which is the situation when moving to higher latitudes. At the magnetic equator, however, this ambipolar diffusion would come to a stop and other forces must be included to complete the picture.

Let us return to the continuity equation (4.22) where we only allow for the vertical motion of plasma and the loss rate is equal to linear loss (βn) in the presence of a magnetic field:

$$\frac{\partial n_e}{\partial t} = q_i - \beta \cdot n_e - \frac{\partial}{\partial z} n_e w = q_i - \beta \cdot n_e + D_p \sin^2 I \left\{ \frac{\partial^2 n_e}{\partial z^2} + \left(\frac{1}{H} + \frac{1}{H_p} \right) \frac{\partial n_e}{\partial z} + \frac{n_e}{H \cdot H_p} \right\}$$

when $D_p = D_0 \exp(z/H)$. Since this is a linear quadratic differential equation, it can be solved analytically by separating the variables. We will only discuss the steady-state solution but, before we do that, we must investigate boundary conditions and asymptotic solutions:

1. At the lowest heights where v_i is very large and $D_p \approx 0$ for a steady state, we then obtain

$$q_i \approx \beta n_e$$

which is the well-known Chapman β-relation.

2. At the upper height where $q_i \approx 0$ and $\beta \approx 0$ we obtain the steady-state solution:

$$\frac{d^2 n_e}{dz^2} + \left(\frac{1}{H} + \frac{1}{H_p}\right)\frac{dn_e}{dz} + \frac{n_e}{H \cdot H_p} = 0$$

This equation has a solution which can be given by:

$$n_e = C_1 \exp\left(-\frac{z}{H}\right) + C_2 \exp\left(-\frac{z}{H_p}\right) \tag{4.26}$$

Since the mass flow at the upper heights must be zero, then

$$\lim_{z\to\infty}(n_e w) = 0$$

and there is no ion outflow. This implies that

$$\lim_{z\to\infty}(n_e w) = \lim_{z\to\infty}\left[-D_p\left(\frac{dn_e}{dz} + \frac{n_e}{H_p}\right)\right] = 0$$

By inserting for n_e from (4.26) and allowing for different ion and electron temperatures we get

$$\lim_{z\to\infty}(n_e w) = \lim_{z\to\infty}\left[(1+c)D_0 C_1\left(\frac{1}{H} - \frac{1}{H_p}\right)\right] = 0$$

and C_1 must be zero in order to fulfill the condition. At the upper boundary, therefore

$$n_e \approx C_2 \exp\left(-\frac{z}{H_p}\right)$$

Since q_i decreases more slowly with altitude than β, the electron density must grow with altitude at lower heights. On the other hand, we have just found that density must decrease with height at upper regions. There must then be a maximum somewhere in between.

4.9 E-LAYER IONIZATION AND RECOMBINATION

As seen from Figure 4.14, O_2^+ and N_2^+ are the main ion species produced by solar radiation between 130 and 90 km. We will therefore study the steady-state balance for these ions ($q_i = l_i$) when no atomic oxygen ion is formed by solar radiation.

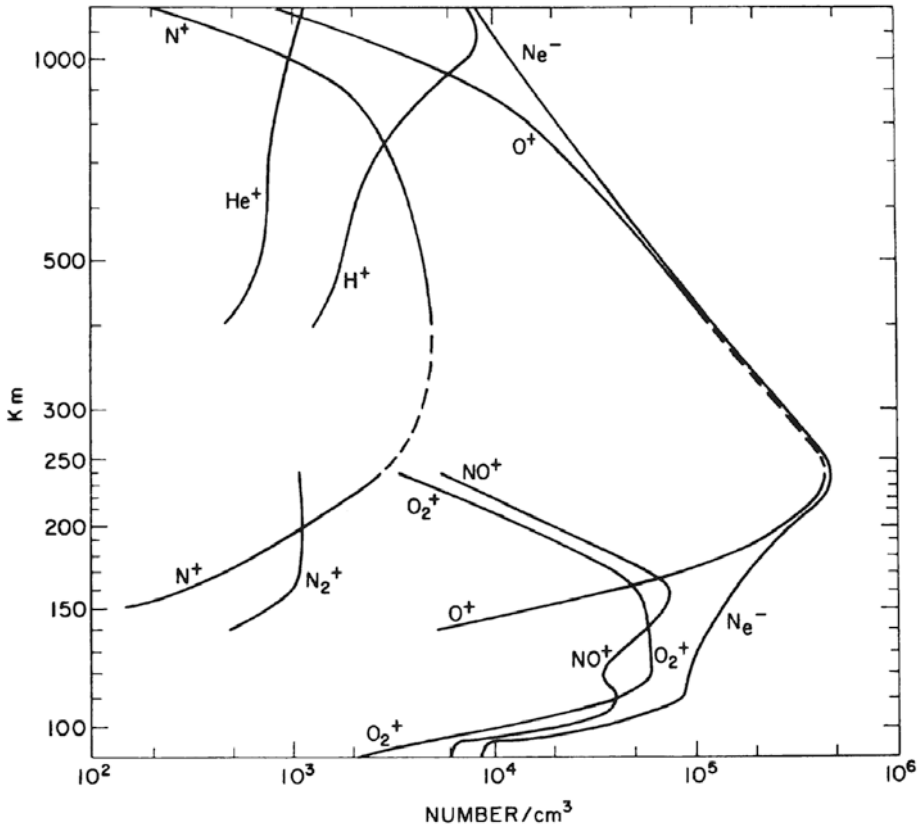


Figure 4.18. Rocket observations of different ion species below 1,000 km for daytime at solar minimum. (From Johnson, 1966.)

The O_2^+ ions produced will recombine with an electron with a dissociative recombination coefficient α_2 (Table 4.4) as before

$$O_2^+ + e \xrightarrow{\alpha_2} O + O$$

or they can rearrange with neutral constituents such as NO and N_2

$$O_2^+ + NO \xrightarrow{k_3} NO^+ + O_2$$
$$O_2^+ + N_2 \xrightarrow{k_4} NO^+ + NO$$

with reaction rates k_3 and k_4, respectively (Table 4.4), forming NO^+ ions which are quite abundant in the E-region (Figure 4.18). Molecular nitrogen ions can also rearrange with O and form NO^+ ions with a reaction rate k_5 (Table 4.4):

$$N_2^+ + O \xrightarrow{k_5} NO^+ + N$$

or they recombine with an electron with a dissociative recombination coefficient α_3 (Table 4.4):

$$N_2^+ + e \xrightarrow{\alpha_3} N + N$$

NO^+ ions are lost through dissociative recombination with an electron:

$$NO^+ + e \xrightarrow{\alpha_1} N + O$$

The dissociative recombination coefficient for this reaction is α_1 (Table 4.4). For O_2^+ ions, production must be equal to loss:

$$q(O_2^+) = l(O_2^+)$$

and it follows:

$$\begin{aligned} q(O_2^+) &= \alpha_2 \cdot [O_2^+] \cdot n_e + k_3[NO] \cdot [O_2^+] + k_4 \cdot [N_2] \cdot [O_2^+] \\ &= \alpha_2 \cdot [O_2^+] \cdot n_e + (k_3[NO] + k_4[N_2]) \cdot [O_2^+] \end{aligned}$$

For N_2^+ ions we have similarly

$$q(N_2^+) = l(N_2^+)$$

and

$$q(N_2^+) = \alpha_3[N_2^+] \cdot n_e + k_5[O] \cdot [N_2^+]$$

and finally for NO^+ ions:

$$q(NO^+) = l(NO^+)$$

where

$$\begin{aligned} q(NO^+) &= k_3[NO] \cdot [O_2^+] + k_4 \cdot [N_2] \cdot [O_2^+] + k_5 \cdot [O] \cdot [N_2^+] \\ &= (k_3[NO] + k_4[N_2]) \cdot [O_2^+] + k_5[O] \cdot [N_2^+] \end{aligned} \tag{4.27}$$

$$l(NO^+) = \alpha_1(NO^+) \cdot n_e \tag{4.28}$$

Since (4.27) and (4.28) are equal

$$\alpha_1[NO^+] \cdot n_e = (k_3[NO] + k_4[N_2]) \cdot [O_2^+] + k_5[O] \cdot [N_2^+] \tag{4.29}$$

The total production of ions due to solar radiation must be equal to the production of electrons:

$$\begin{aligned} q(n_e) &= q(O_2^+) + q(N_2^+) \\ &= \alpha_2[O_2^+] \cdot n_e + \alpha_3[N_2^+] \cdot n_e + (k_3 \cdot [NO] + k_4 \cdot [N_2]) \cdot [O_2^+] + k_5[O] \cdot [N_2^+] \end{aligned}$$

and implementing (4.29):

$$\begin{aligned} q(n_e) &= \alpha_1[NO^+] \cdot n_e + \alpha_2[O_2^+] \cdot n_e + \alpha_3[N_2^+] \cdot n_e \\ &= (\alpha_1[NO^+] + \alpha_2[O_2^+] + \alpha_3[N_2^+]) \cdot n_e \end{aligned} \tag{4.30}$$

At any time there must be charge neutrality, and therefore

$$[NO^+] + [O_2^+] + [N_2^+] = n_e$$

or

$$\frac{[NO^+]}{n_e} + \frac{[O_2^+]}{n_e} + \frac{[N_2^+]}{n_e} = 1$$

Introducing r_1, r_2, and r_3 as the relative abundance ratios of the respective ions

$$r_1 = [NO^+]/n_e \tag{4.31a}$$

$$r_2 = [O_2^+]/n_e \tag{4.31b}$$

$$r_3 = [N_2^+]/n_e \tag{4.31c}$$

we find by inserting the ratios:

$$r_1 + r_2 + r_3 = 1$$

Finally, we derive from (4.30) by introducing the ratios from (4.31):

$$q(n_e) = \alpha_1 r_1 n_e^2 + \alpha_2 r_2 n_e^2 + \alpha_3 r_3 n_e^2 = (\alpha_1 r_1 + \alpha_2 r_2 + \alpha_3 r_3) n_e^2$$

By now introducing the effective recombination coefficient

$$\alpha_{\text{eff}} = \alpha_1 r_1 + \alpha_2 r_2 + \alpha_3 r_3 \tag{4.32}$$

the electron production rate is given by

$$q(n_e) = \alpha_{\text{eff}} n_e^2$$

and the E-region electron density profile behaves like an α-profile.

As observed by many rocket probes passing through the E-region, N_2^+ ions (Figure 4.18) decrease very rapidly below 150 km and can for all practical purposes be neglected ($r_3 = 0$) in the E-region. It is also common practice to infer a mean ion mass $\bar{M} = 30.5$ a.m.u. at these altitudes. When neglecting N_2^+ we therefore have

$$r_1 + r_2 = 1$$

and the mean ion mass is given by

$$r_1 \cdot M_{NO^+} + r_2 \cdot M_{O_2^+} = \bar{M}$$

or

$$r_1 \cdot 30 + r_2 \cdot 32 = 30.5$$

where $M_{NO^+} = 30$ a.m.u. and $M_{O_2^+} = 32$ a.m.u. Solving for r_1 and r_2 gives:

$$r_1 = \tfrac{3}{4} \qquad \text{and} \qquad r_2 = \tfrac{1}{4}$$

The effective recombination coefficient can then be expressed as follows:

$$\alpha_{\text{eff}} = 0.75 \cdot \alpha_1 + 0.25 \cdot \alpha_2$$

where α_1 and α_2 are given in Table 4.4.

4.10 TIME CONSTANT OF THE RECOMBINATION PROCESS

Let us now go back to the electron density continuity equation (4.22) but still neglect any transport processes so that the equation can be simplified to:

$$\frac{dn_e}{dt} = q_e - l_e$$

where q_e and l_e are the production and loss terms of electrons, respectively. For a β-profile we know that the loss rate is linear and given by:

$$l_e = \beta n_e$$

and therefore

$$\frac{dn_e}{dt} = q_e - \beta n_e$$

Let us now assume that the production $q_e = q_0$ is suddenly shut off at a time $t = 0$ ($q_e = 0$ for $t > 0$). Then, the electron density will decay with time as:

$$n_e = n_{e_0} \exp(-\beta t) \tag{4.33}$$

where $n_{e_0} = q_0/\beta$ is density at $t = 0$ (Figure 4.19(a)), and $\tau = 1/\beta$ is the time constant determining how quickly the electron density is reduced to $1/e$. On the other hand, if production q_e is turned on at time $t = 0$ and is constant over a length of time ($q_e = q_0$), then the electron density will grow according to

$$n_e = n_\infty(1 - \exp(-\beta t)) \tag{4.34}$$

where

$$n_\infty = q_0/\beta$$

after a time $T \gg 1/\beta$ (Figure 4.19(b)). For a time-varying $q(t)$, however, the electron density at time t will be given by:

$$n_e(t) = \exp(-\beta t) \int_0^t q(t') \exp(\beta t')\, dt'$$

In the E-region we have just shown that the loss rate is proportional to n_e^2 and, therefore, the continuity equation again for n_e, neglecting transport, is:

$$\frac{dn_e}{dt} = q_e - \alpha_{\text{eff}} n_e^2 \tag{4.35}$$

If we follow the same examples as for the β-profile and assume that $q_e = q_0$ is turned off at $t = 0$, then the electron density will decay as:

$$n_e = \frac{n_{e_0}}{n_{e_0} \cdot \alpha_{\text{eff}} \cdot t + 1}$$

when $n_{e_0} = \sqrt{q_0/\alpha_{\text{eff}}}$ is the electron density at $t = 0$. To introduce a time constant one can estimate the time elapsed until $n_e = \frac{1}{2} \cdot n_{e_0}$ and find:

$$t_{1/2} = 1/n_{e,0} \cdot \alpha_{\text{eff}} \tag{4.36}$$

which depends on the electron density at $t = 0$.

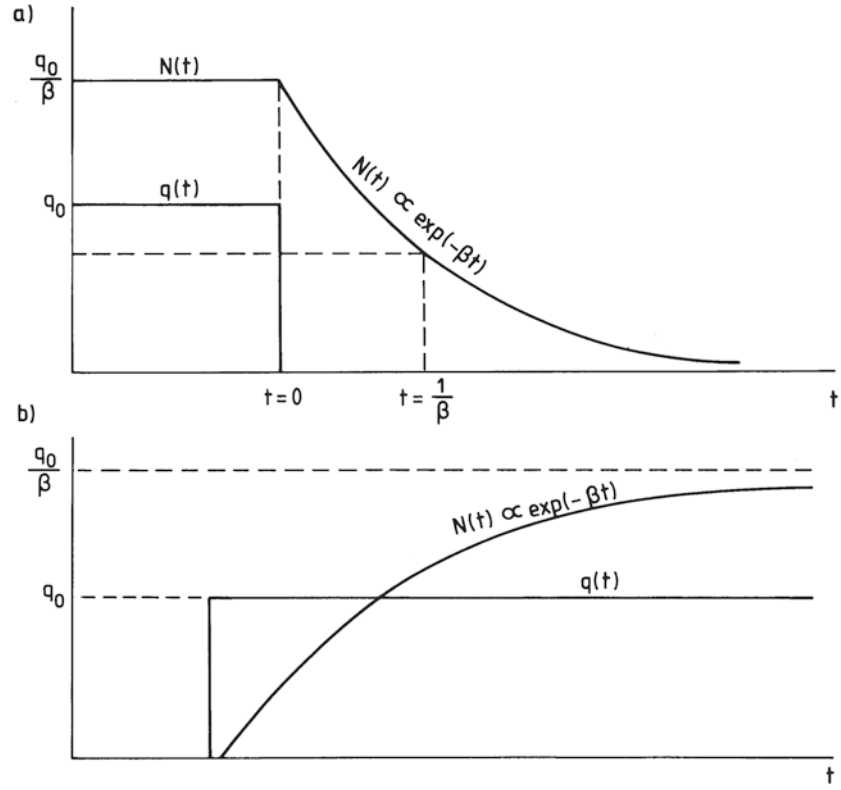


Figure 4.19. Time behavior of the electron density at a given height (a) if production is turned off at $t = 0$, and (b) if it is turned on at $t = 0$.

For an ionization source that is turned on at $t = 0$ and stays constant ($q_e = q_0$) for a length of time, however, we find:

$$n_e = \sqrt{\frac{q_0}{\alpha_{\text{eff}}}} \tanh(\sqrt{q_0 \cdot \alpha_{\text{eff}}} \cdot t) = n_\infty \tanh(n_\infty \cdot \alpha_{\text{eff}} \cdot t) \tag{4.37}$$

where q_0 is the constant production rate, and $n_\infty = \sqrt{q_0/\alpha_{\text{eff}}}$ is the electron density at infinity. We notice for

$$t = 1/(2 \cdot n_{e,0} \cdot \alpha_{\text{eff}}) \tag{4.38}$$

that

$$n_e = 0.46 n_\infty \approx 0.5 n_\infty$$

The conventional use of the time constant concept is therefore not so obvious in the E-region.

For small fluctuations in n_e we may have:

$$n_e = n_0 + n_e'$$

where n_0 is the background electron density, and n'_e is a small variation in the background. If $n'_e \ll n_0$, we get:

$$\frac{dn_e}{dt} = \frac{dn'_e}{dt} = q_e - \alpha_{\text{eff}}(n_0 + n'_e)^2 = q_e - \alpha_{\text{eff}} n_0^2 - 2\alpha_{\text{eff}} n_0 \cdot n'_e$$

when n'^2_e is neglected compared with n_0^2.

$$\frac{dn'_e}{dt} = (q_e - \alpha n_0^2) - 2\alpha n_0 \cdot n'_e$$

and we are back to a linear equation in n'_e where the time constant will be:

$$\tau = \frac{1}{2\alpha_{\text{eff}} n_0} \tag{4.39}$$

This is a similar result for τ as obtained in (4.38). This value of τ is often mentioned in the literature as the recombination time constant of the E-region. When dealing with these parameters—(4.36), (4.38), and (4.39)—we have to discriminate between the different options to avoid confusion.

Figure 4.20 shows a series of E-region electron density profiles observed by the incoherent scatter radar at Chatanika, Alaska, on August 4, 1972, between 20:52.50 and 20:57.50 UT. At 20:54 UT on this day a sudden commencement occurred that released a particle precipitation producing enhanced electron densities. From the figures we notice that at 20:57.50 UT the profile is back to the same shape as before the sudden commencement. It is actually believed that the sudden commencement stopped producing electrons after 20:56.50 UT (see also Figure 6.4).

Background electron production q_0 at an altitude z can be written as:

$$q_0 = \alpha n_0^2$$

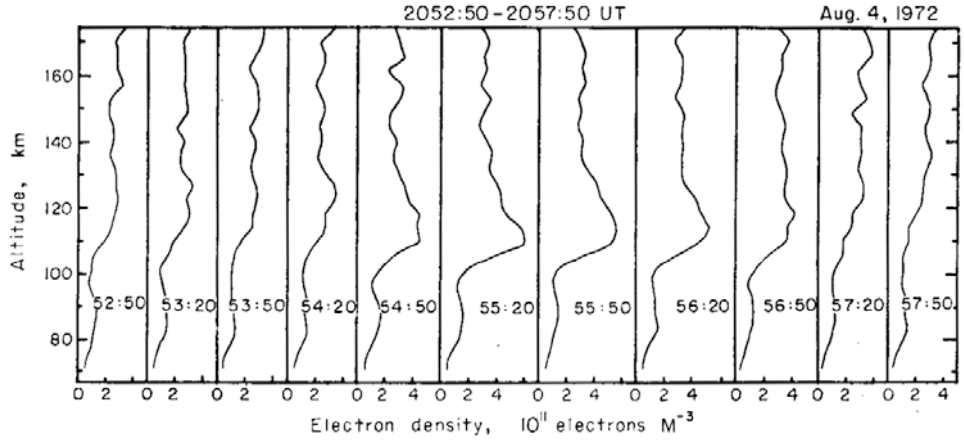


Figure 4.20. Electron density profiles observed every 30 s by the incoherent scatter radar at Chatanika, Alaska on August 4, 1972, between 20:52.50 UT and 20:57.50 UT. A sudden commencement occurred at 20:54 UT.

where n_0 is the corresponding electron density at altitude z. During the decaying phase of the electron densities produced by the sudden commencement between 20:56.50 UT and 20:57.20 UT the total electron density at altitude z is given by

$$n_e = n_0 + n'_e$$

where n'_e is the electron density produced due to the sudden commencement.

The continuity equation (4.35) can now be written as:

$$\frac{dn_e}{dt} = \frac{dn'_e}{dt} = q_0 - \alpha(n_0 + n'_e)^2$$

and since $q_0 = \alpha n_0^2$

$$\frac{dn'_e}{n'_e(2n_0 + n'_e)} = -\alpha\, dt$$

and

$$\ln \frac{\frac{1}{2} + n_0/n_2}{\frac{1}{2} + n_0/n_1} = +2\alpha n_0 \Delta t \tag{4.40}$$

where $n'_e = n_1$ at $t = t_1$ and $n'_e = n_2$ at $t = t_2 = t_1 + \Delta t$ when t_1 and t_2 represent two given times in the decaying phase after the production due to the precipitation is zero.

The dissociative recombination coefficient can then be derived from (4.40):

$$\alpha = \frac{1}{2n_0 \Delta t} \ln \frac{\frac{1}{2} + n_0/n_2}{\frac{1}{2} + n_0/n_1}$$

At an altitude of 110 km, α was found to be $(1.1 \pm 0.3) \times 10^{-13}$ m^3 s^{-1} from this work, which can be compared with the α-values given in Table 4.4.

4.11 D-REGION IONIZATION AND RECOMBINATION

The lower part of the ionosphere below 90 km and down to about 60 km is usually referred to as the D-region. The pressure is about 10^6 times as large as at typical F-region heights and collisions are dominant. The photochemistry is very complex and is not fully understood. Four important ionization sources in the D-region are identified (the contribution from them to ion production is illustrated in Figure 4.21):

1. The Lyman-α line at 1,215 Å (121.5 nm) penetrates the D-region and ionizes nitric oxide (NO), a minor constituent in the neutral atmosphere.
2. The EUV radiation spectrum between 1,027 and 1,118 Å (102.7–111.8 nm) ionizes excited oxygen molecules O_2 ($^1\Delta g$).
3. Hard X-rays between 2 and 8 Å (0.2–0.8 nm) ionizes all constituents, thereby acting mainly on N_2 and O_2.
4. Cosmic rays similarly ionize all constituents.

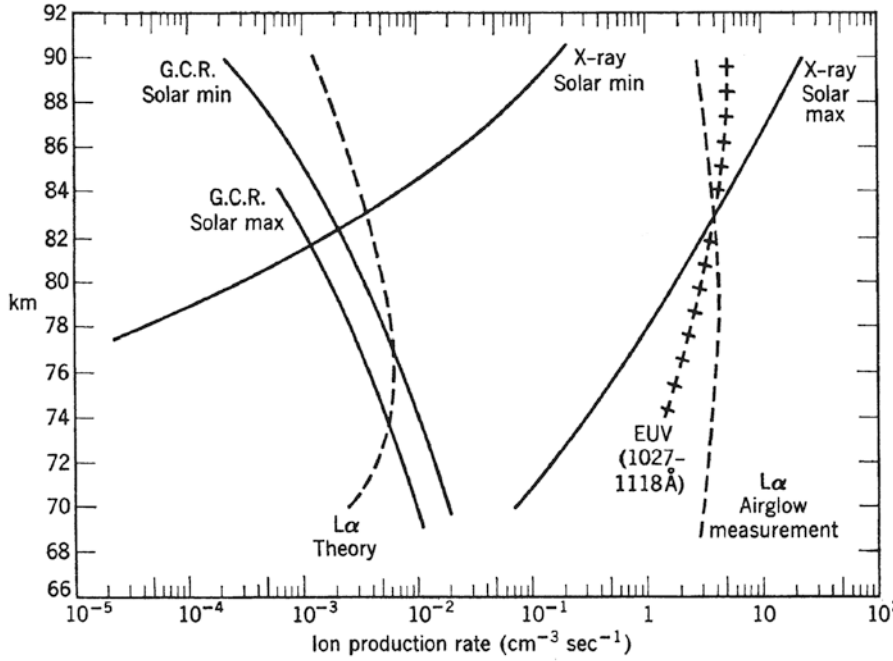


Figure 4.21. Illustrative example of ionization produced by X-rays, EUV radiation, galactic cosmic rays (GCRs), and hydrogen L_α radiation. (From Whitten and Poppoff, 1971.)

Relativistic electrons (100 keV–1 MeV) may penetrate the D-region deeply on occasions and the same is true for protons in the energy range 1–100 MeV, and these particles will also ionize the atmosphere.

From the point of view of radio communication the D-region is of great importance because most of the radiowave absorption in the HF and MF regime occurs in this region.

Due to the high pressure in the D-region, conventional techniques for studying the ionosphere are not applicable. For instance, ionosonde sounding is not possible because the collision frequency is so high that it generally exceeds the critical frequency, and absorption is heavy at all reflected frequencies. For this reason, rocket techniques have been particularly important for D-region studies. The problem with rockets, however, is that the time spent by them in the D-region on both upleg and downleg is extremely short (10–20 s). Modern radar techniques, however, have recently opened up new fields of observational methods which hold out a lot of promise for future D-region studies.

The positive ions formed directly by solar radiation in the D-region are O_2^+, N_2^+, and NO^+. The N_2^+ ion rapidly exchanges charge with O_2 to form an O_2^+ ion by the reaction rate k_6 (Table 4.4) as follows:

$$N_2^+ + O_2 \xrightarrow{k_6} N_2 + O_2^+$$

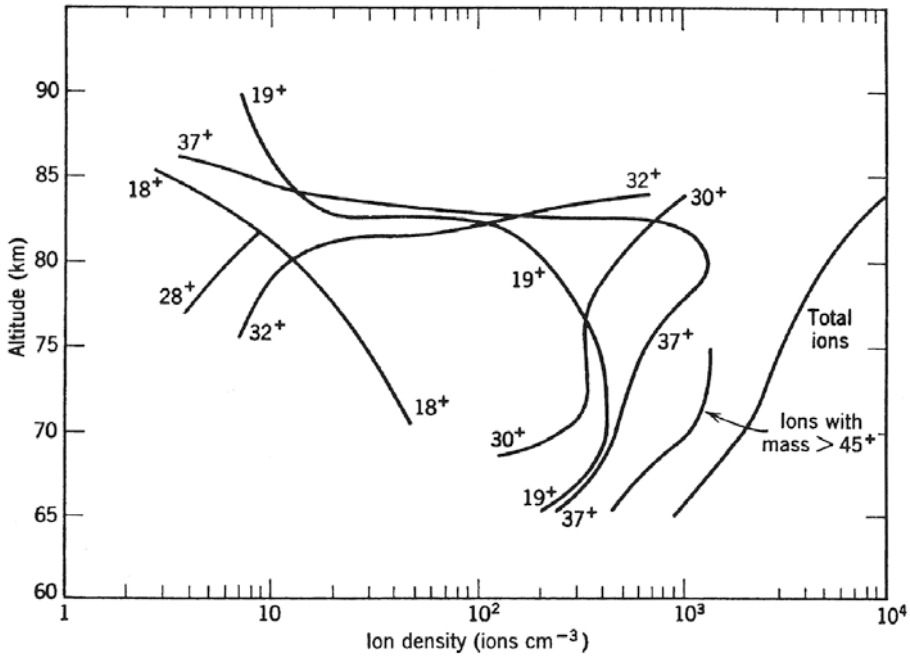


Figure 4.22. Concentration of positive ions detected by mass spectrometers in the D-region. The ions are identified by their molecular mass number. (From Narcisi and Bailey, 1965.)

The O_2^+ ion formed will quickly charge transfer to NO^+ according to rate k_3 or k_4 (see Table 4.4). Because the concentration of NO is very small compared with the density of O_2 (one in about 10^6) at 90 km, a large number of O_2^+ is expected to exist in the D-region during the daytime.

From mass spectrometric analyses of ion species in the D-region (Figure 4.22) it is seen that N_2^+ (28^+) is much less abundant than NO^+ (30^+) and O_2^+ (32^+) above 80 km. Ions with molecular mass 19 and 37 are thought to be hydrated ions, $H^+(H_2O)$ and $H_3O^+ \cdot (H_2O)$; they are very abundant below 80 km but decrease rapidly above 82 km or so. They are generally found where the water vapor concentration exceeds $10^{15}\,m^{-3}$. Heavy ions with molecular masses above 45 are observed below 75 km. These are believed to be metallic ions, probably debris from meteor showers. Above about 85 km it should, however, be safe to assume that NO^+ and O_2^+ are the dominant positive ions.

An important feature of the D-region is the formation of negative ions such as O_2^-. These are produced by a three-body reaction with reaction rates β_1 and β_2, respectively (Table 4.4) as follows:

$$O_2 + M + e \xrightarrow{\beta_1,\beta_2} O_2^- + M$$

where M is any molecule capable of increasing the chance of reaction by removing excess kinetic energy from the reactants.

The electron affinity of O_2 is small, only 0.45 eV, and the electron might therefore be rapidly detached again by quanta of low energy such as infrared or visible light

$$O_2^- + h\nu \xrightarrow{\rho} O_2 + e$$

The rate coefficient for this process is assigned by ρ (Table 4.4). Other processes that can destroy O_2^- ions can be detachment by collisions with neutrals

$$O_2^- + M \xrightarrow{\gamma_1, \gamma_2} O_2 + M + e$$

$$O_2^- + O \xrightarrow{k_3} O_3 + e$$

The reaction rates are defined by γ_1, γ_2, and γ_3 (Table 4.4), where γ_1 applies if $M = [N_2]$, γ_2 applies if $M = [O_2]$, and γ_3 applies for the last reaction, respectively. Other possible loss processes for negative ions in the D-region are ion–ion recombination which can be expressed as

$$O_2^+ + O_2^- \xrightarrow{\alpha_{i_1}} O_2 + O_2$$

$$O_2^+ + O_2^- + M \xrightarrow{\alpha_{i_2}} O_2 + O_2 + M$$

where the reaction rates are α_{i_1} and α_{i_2}, respectively (Table 4.4). We should by now have a relatively complete picture of the major chemical processes taking place in the D-region. On the basis of several studies of the charged constituents of the D-region it is found (Figure 4.23) that the number of negative ions increases

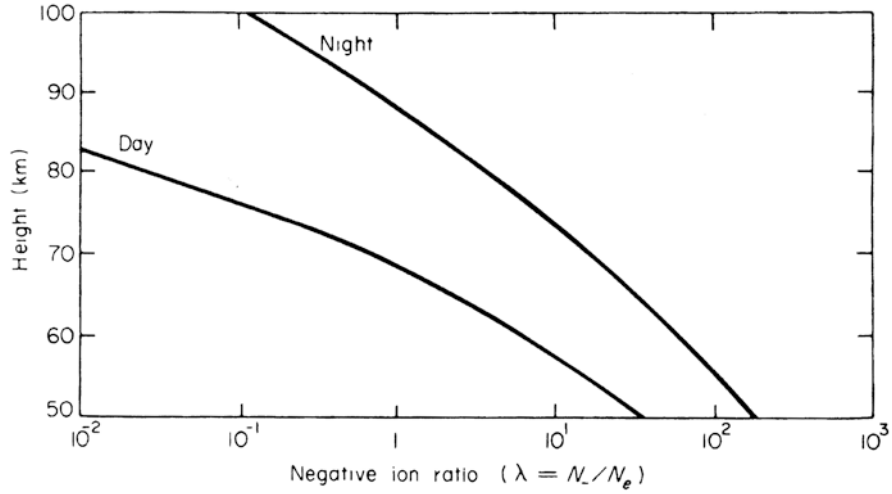


Figure 4.23. Negative ion–electron concentration ratio λ below 90 km under daytime and night-time conditions. The model is based on molecular oxygen ions only. (From Rishbeth and Garriott, 1969.)

sharply with decreasing height, and there are more negative ions at night-time than at daytime.

To give a full treatment of all ion species would be very complicated and time consuming, and probably not so illustrative. Let us therefore identify the electron, positive ion, and negative ion densities in the D-region by n_e, n^+, and n^-, respectively, without differentiating between the different ion species. Let us furthermore adapt an effective recombination loss rate between positive ions and electrons by α_D, between positive and negative ions by α_i, an effective electron attachment rate by β, an effective electron detachment rate by γ, and a radiative detachment rate by ρ. We will in the following neglect radiative recombination as a possible loss rate in the D-region. Any transport term will also be ignored.

The continuity equation for electrons will now be given by:

$$\frac{dn_e}{dt} = q - \alpha_D n^+ \cdot n_e - \beta n_e \cdot n_n + \gamma \cdot n^- \cdot n_n + \rho \cdot n^-$$

where n_n is the number density of neutrals. For positive ions we have a similar continuity equation:

$$\frac{dn^+}{dt} = q - \alpha_D n^+ \cdot n_e - \alpha_i n^- \cdot n^+ \tag{4.41}$$

where q is the same production rate as for electrons due to solar radiation. Negative ions will obey the following continuity equation:

$$\frac{dn^-}{dt} = -\alpha_i n^- \cdot n^+ + \beta n_e \cdot n_n - \gamma n^- \cdot n_n - \rho n^-$$

There has to be charge neutrality in the plasma, therefore

$$n^+ = n_e + n^-$$

By introducing the ratio between negative ions and electrons λ

$$\lambda = \frac{n^-}{n_e}$$

we have

$$n^+ = (1 + \lambda) \cdot n_e \tag{4.42}$$

By inserting (4.42) into the continuity equation for positive ions (4.41) we get:

$$\begin{aligned}(1 + \lambda)\frac{dn_e}{dt} + n_e \frac{d\lambda}{dt} &= q - \alpha_D n^+ \cdot n_e - \alpha_i n^- \cdot n^+ \\ &= q - (a_D + \alpha_i \cdot \lambda) \cdot (1 + \lambda) n_e^2 \end{aligned} \tag{4.43}$$

and

$$\frac{dn_e}{dt} = \frac{q}{1 + \lambda} - (\alpha_D + \lambda \alpha_i) n_e^2 - \frac{n_e}{1 + \lambda}\frac{d\lambda}{dt}$$

We will assume that λ is constant. Then, the continuity equation for the electron density will be given by:

$$\frac{dn_e}{dt} = \frac{q}{1+\lambda} - (\alpha_D + \lambda\alpha_i)n_e^2 \tag{4.44}$$

Since λ is a constant, this equation for n_e has the same form as (4.35), the continuity equation, for the electron density in the E-region. n_e in the D-region behaves in many respects as an α-profile.

The effective recombination coefficient can now be identified as

$$\alpha'_{\text{eff}} = \alpha_D + \lambda\alpha_i$$

and the production rate

$$q' = \frac{q}{1+\lambda}$$

is reduced by the factor $1/(1+\lambda)$ with respect to q_e in (4.35). This factor can indeed be very small in the lower part of the D-region where $\lambda \approx (100\text{–}1{,}000)$. When the electron density ($dn_e/dt = 0$) is in a steady state, we also see from (4.44) that

$$q \approx (1+\lambda)(\alpha_D + \lambda\alpha_i)n_e^2 = \alpha''_{\text{eff}} n_e^2$$

where

$$\alpha''_{\text{eff}} = (1+\lambda)\alpha'_{\text{eff}}$$

α''_{eff} and α'_{eff} are often interchanged in the literature, but they are only equal when $\lambda \approx 0$. It is important to be aware of this in order to avoid confusion as α''_{eff} and α'_{eff} can differ by a factor of 1,000.

The preceding paragraph should have made it abundantly clear that the D-region is very complicated as many different chemical reactions are involved along with a variety of neutral and ionized species.

There has been increased interest in this height region of the polar atmosphere in the last 20 years due to the discovery by radar of so-called PMSEs (polar mesosphere summer echoes). These are echoes that are easily observable by incoherent scatter radars like EISCAT (Figure 4.24) as strong coherent echoes in the height region between 80 and 90 km (i.e., close to the temperature minimum at the mesosphere). These echoes occur in the same height region and at the same time as noctilucent clouds (as shown in Figure 4.25). These phenomena are believed to be due to a combination of charged dust and ice particles. Observations of noctilucent clouds had not been recorded by mankind until the 1880s when the volcanic eruption of Krakatoa in Indonesia revealed their existence. Since then they have been frequently observed in the northern evening sky from latitudes of about 60° N. It is often assumed that these echoes and clouds may be related to release in the atmosphere of methane that reacts with oxygen and forms water vapor diffusing up to the mesosphere where the water vapor freezes to become ice due to the extreme cold there.

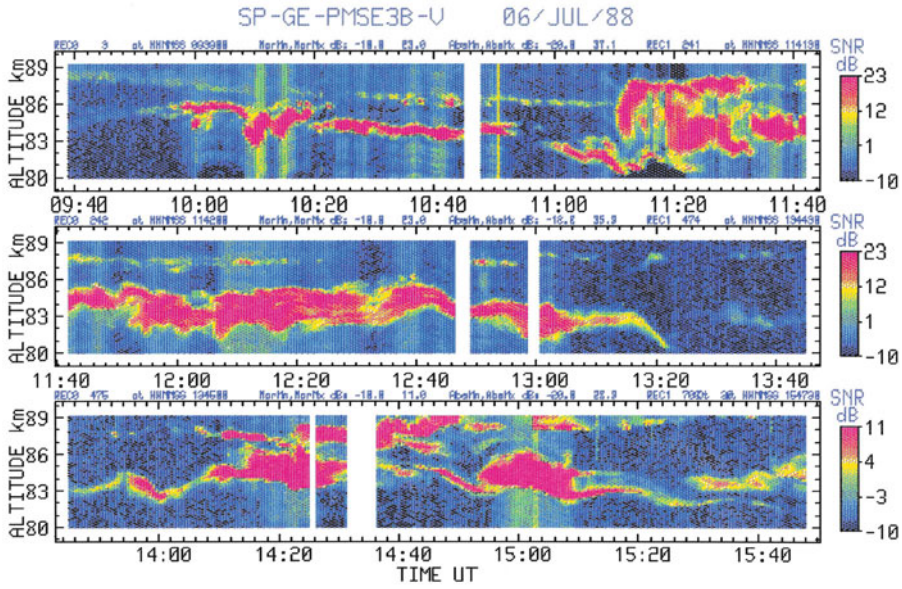


Figure 4.24. Observations of PMSEs (polar mesospheric summer echoes) by EISCAT in Tromsø in July 1988.

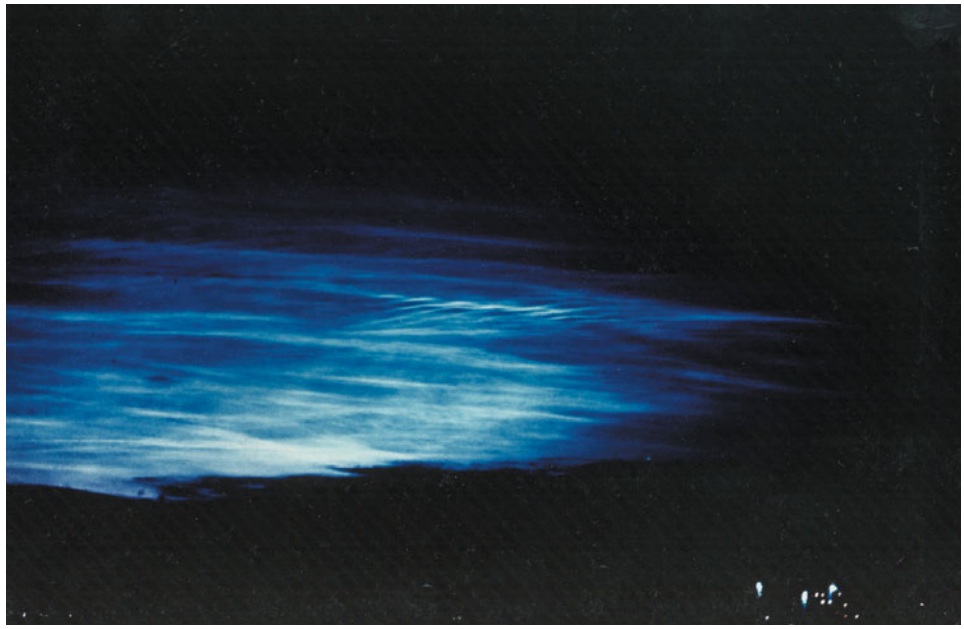


Figure 4.25. Typical noctilucent clouds observed over Finland.

4.12 THE PLASMASPHERE

In the lower ionosphere the motion of the plasma is strongly dominated by interacting with the neutral atmosphere through collisions. Moving upward in the ionosphere, however, the influence from electric fields of external origin and diffusion starts to prevail.

In the lower ionosphere molecular ions such as N_2^+, NO^+, and O_2^+ dominate. In the upper ionosphere, however, lighter ions such as O^+, H^+, and He^+ are more abundant. The lower ionosphere is formed, as we know, by ion production due to solar irradiation below 200 nm and loss due to different recombination and chemical processes. At higher altitudes, however, the transport processes become increasingly dominant.

During the daytime, plasma is produced or lost by different processes in the E- and F-region creating horizontally stratified slowly varying plasma in the ionosphere. Above the F-region peak we expect plasma to be in a state dominated by diffusive equilibrium and, as a result, electron gas because of its light mass diffuses faster than ions creating the ambipolar electric field which enhances ion diffusion, and thereby results in an equilibrium diffusion state of the plasma.

Since this equilibrium is generally established between electrons and O^+ ions, the lighter ions such as H^+ and He^+ may experience outward acceleration due to this ambipolar electric field (as we have seen).

Since hydrogen atoms and molecules can escape the gravitational field of the Earth, the Earth is surrounded by a hydrogen gas called the "geocorona" which interacts with oxygen atoms via charge exchange processes such as:

$$H + O^+ \rightarrow H^+ + O$$

$$H^+ + O \rightarrow H + O^+$$

This leads to a transition between oxygen-dominated and hydrogen-dominated plasma somewhere between 500 km and a few thousand kilometers above ground. Since hydrogen ions actually can also escape the Earth's gravity field, the simple steady state diffusive equilibrium situation cannot be maintained, and net upward flow of plasma is possible from the topside ionosphere during the day. The closed magnetic field lines at lower latitudes will act as an enclosure for the upstreaming plasma and help establish a reservoir of plasma produced by photoionization during the daytime extending a great distance out from the Earth. This is called the plasmasphere and is illustrated by the hatched area in Figure 4.26. During magnetic disturbances plasma will be lost from the plasmasphere, and refilled again in quiet times by the diffusion of protons. At night this plasma can flow back down the magnetic field lines to replenish the ionosphere with plasma lost there by recombination. The result is a complex interaction between the ionosphere and a region of hydrogen plasma trapped by the dipole magnetic field, the plasmasphere.

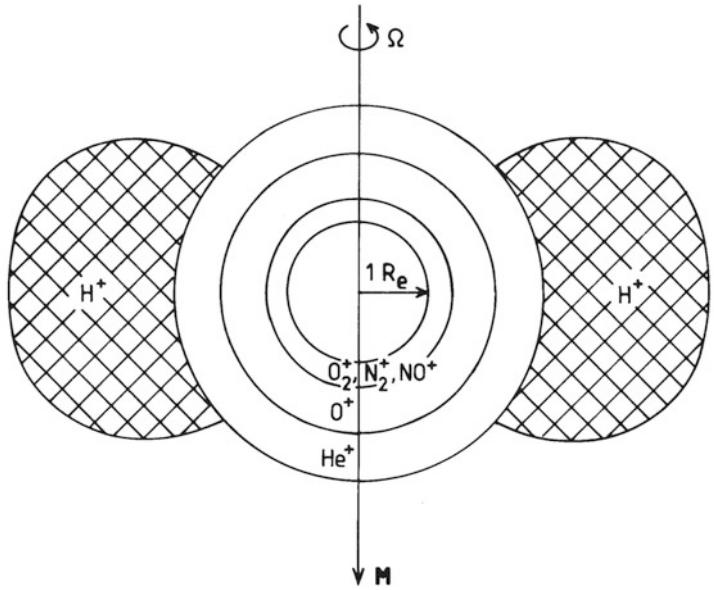


Figure 4.26. A schematic illustration of the predominant ion species in different regions of the ionosphere and plasmasphere. The dimensions are not to scale, the plasmasphere is supposed to reach to about 5 R_e in the equatorial plane, however.

4.13 FERRARO'S THEOREM AND A UNIPOLAR INDUCTOR

Plasma on closed field lines corotates with the Earth. That this indeed is so can be understood from the proof of the so-called Ferraro's theorem:

> *In an equilibrium situation the angular velocity of the highly conducting plasma enclosed by the magnetic dipole field of a body rotating around the magnetic axis will be constant along the magnetic field lines, and the plasma corotates with the body.*

To prove Ferraro's theorem for isorotation we assume that the magnetic field of the Earth can be described by a dipole magnetic field coaxial to the Earth's rotation axis.

We apply Maxwell's equations:

$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E}$$

$$\nabla \cdot \mathbf{B} = 0$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j}$$

where we have neglected the displacement current $\varepsilon_0\mu_0(\partial\mathbf{E}/\partial t)$. The generalized Ohm's law describing the current is given by:

$$\mathbf{j} = \sigma \cdot (\mathbf{E} + \mathbf{v} \times \mathbf{B})$$

where it is assumed that σ is uniform. For an equilibrium situation $(\partial/\partial t = 0)$

$$0 = -\nabla \times \mathbf{E} = -\nabla \times \left(\frac{\mathbf{j}}{\sigma} - \mathbf{v} \times \mathbf{B}\right) = -\frac{1}{\sigma}\nabla \times \mathbf{j} + \nabla \times (\mathbf{v} \times \mathbf{B})$$

For very high conductivity $\sigma \to \infty$ we find:

$$\nabla \times (\mathbf{v} \times \mathbf{B}) = 0$$

If we now choose a cylindrical coordinate system $(\hat{\mathbf{r}}, \hat{\boldsymbol{\varphi}}, \hat{\mathbf{z}})$ so that $\hat{\mathbf{z}}$ is parallel to the rotation axis, then the velocity at a distance r from the rotation axis

$$\mathbf{v} = r\Omega\hat{\boldsymbol{\varphi}}$$

where Ω is the angular velocity. The magnetic field can be described by

$$\mathbf{B} = B_r\hat{\mathbf{r}} + B_\varphi\hat{\boldsymbol{\varphi}} + B_z\hat{\mathbf{z}}$$

where B_r, B_φ and B_z are components along $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\varphi}}$, and $\hat{\mathbf{z}}$ respectively.

In this coodinate system we now have

$$\mathbf{v} \times \mathbf{B} = \Omega r(B_2\hat{\mathbf{r}} - B_r\hat{\mathbf{z}})$$

Let us now apply the curl operator in cylindrical coordinates:

$$\begin{aligned}\nabla \times (\mathbf{v} \times \mathbf{B}) &= \nabla \times (\Omega r B_z\hat{\mathbf{r}} - \Omega r B_r\hat{\mathbf{z}}) \\ &= \frac{1}{r}\left\{-\frac{\partial}{\partial\varphi}(\Omega r B_r)\hat{\mathbf{r}} + \left[\frac{\partial}{\partial r}(\Omega r B_r) + \frac{\partial}{\partial z}(\Omega r B_z)\right]\hat{\boldsymbol{\varphi}} - \frac{\partial}{\partial\varphi}(\Omega r B_z)\hat{\mathbf{z}}\right\}\end{aligned}$$

We will now take advantage of the axial symmetry of the dipole magnetic field, so that

$$\frac{\partial B_r}{\partial\varphi} = \frac{\partial B_\varphi}{\partial\varphi} = \frac{\partial B_z}{\partial\varphi} = 0$$

and likewise $\partial\Omega/\partial\varphi = 0$ for symmetric reasons, and since $\partial r/\partial\varphi = 0$ we have:

$$\begin{aligned}\nabla \times (\mathbf{v} \times \mathbf{B}) &= \left[\frac{1}{r}\frac{\partial}{\partial r}(\Omega r B_r) + \frac{1}{r}\frac{\partial}{\partial z}(\Omega r B_z)\right]\hat{\boldsymbol{\varphi}} \\ &= \left[\left(\Omega\frac{1}{r}\frac{\partial}{\partial r}(r B_r) + \Omega\frac{\partial B_z}{\partial z}\right) + \left(B_r\frac{\partial\Omega}{\partial r} + B_z\frac{\partial\Omega}{\partial z}\right)\right]\hat{\boldsymbol{\varphi}} = 0\end{aligned}$$

Since $\partial\Omega/\partial\varphi = \partial B_\varphi/\partial\varphi = 0$ we can form

$$\begin{aligned}\nabla \times (\mathbf{v} \times \mathbf{B}) &= \left[\Omega\left(\frac{1}{r}\frac{\partial}{\partial r}(r B_r) + \frac{1}{r}\frac{\partial B_\varphi}{\partial\varphi} + \frac{\partial B_z}{\partial z}\right) + \left(B_r\frac{\partial\Omega}{\partial r} + B_\varphi\frac{\partial\Omega}{\partial\varphi} + B_z\frac{\partial\Omega}{\partial z}\right)\right]\hat{\boldsymbol{\varphi}} \\ &= (\Omega\nabla \cdot \mathbf{B} + \mathbf{B} \cdot \nabla\Omega)\hat{\boldsymbol{\varphi}} = \mathbf{B} \cdot \nabla\Omega\hat{\boldsymbol{\varphi}} = 0\end{aligned}$$

since $\nabla \cdot \mathbf{B} = 0$. Therefore, $\mathbf{B} \cdot \nabla\Omega = 0$ and the gradient in Ω along $\mathbf{B}$ is zero. The angular velocity is therefore constant along the magnetic field line. Ferraro's theorem is proven.

The Earth is a rotating magnet surrounded by an atmosphere where the upper part consists of a partly ionized plasma. The motion of this plasma will be strongly influenced by the magnetic field in such a way that plasma moves more easily along the field lines than perpendicular to them. The field lines therefore can act as conductors, triggering interest in investigating whether the rotation of the magnetic field can actually lead to currents along the field lines between the hemispheres.

Let us now derive the magnitude of the electric field associated with rotation in a vacuum of a magnetized sphere of radius R and conductivity σ. In the closed current loop drawn in Figure 4.27, which consists of a conducting sphere where $\mathbf{E}' = \mathbf{E} + \mathbf{v} \times \mathbf{B} = 0$ and a conducting wire, there will be a resulting electromotive force due to the rotating B-field and $\mathbf{E} \neq 0$. This must then give rise to a potential difference between A and B in the conducting wire and on the sphere (as observed outside the sphere).

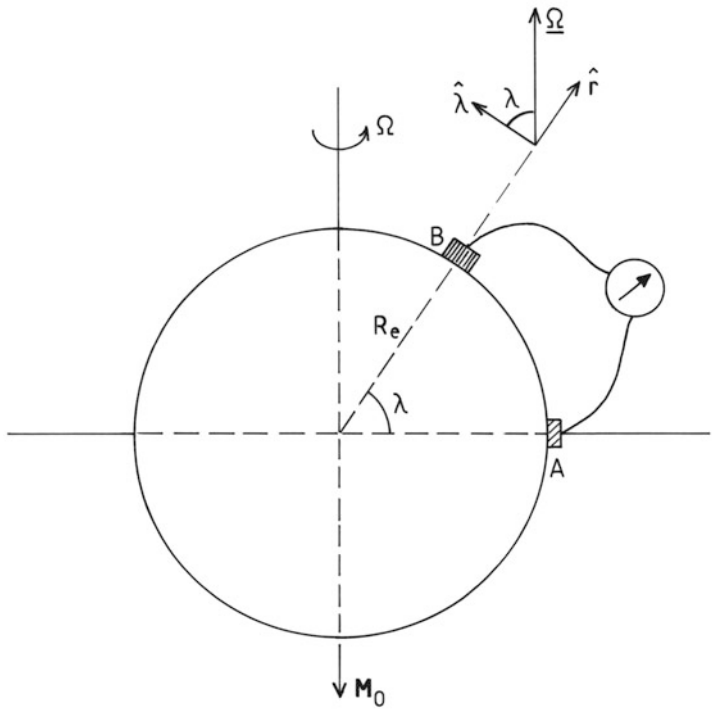


Figure 4.27. The principle for a unipolar inductor including a rotating magnetized conducting sphere with radius R_e and a conductor connected to the surface of the sphere by sliding contacts. A potential difference between points A and B is established.

We will assume that the magnetic field is dipolar and aligned along the rotation axis, then at the surface of the sphere (3.6)

$$B_r = -2H_0 \sin \lambda_B$$

$$B_\lambda = H_0 \cos \lambda_B$$

where H_0 is the magnetic field strength at the equator of the sphere, and λ_B is the latitude of point B.

The potential difference between the two contacts at A and B, where A is at the equator, will be

$$V = -\int_A^B \mathbf{E}\, d\mathbf{s} = -\int_0^{\lambda_B} (\mathbf{E}\cdot\hat{\boldsymbol{\lambda}}) R_e\, d\lambda = -R_e \int_0^{\lambda_B} E_\lambda\, d\lambda$$

From the stationary system (the conductor) we have

$$\mathbf{E} = -(\boldsymbol{\Omega} \times \mathbf{R}_e) \times \mathbf{B}$$

where Ω is the angular velocity of the sphere. In polar coordinates:

$$\boldsymbol{\Omega} = \Omega \sin \lambda\, \hat{\mathbf{r}} + \Omega \cos \lambda\, \hat{\boldsymbol{\lambda}}$$

and

$$\mathbf{E} = \Omega R_e B_r \cos \lambda\, \hat{\boldsymbol{\lambda}} - \Omega R_e B_\lambda \cos \lambda\, \hat{\mathbf{r}}$$

Therefore

$$V = -R_e \int_0^{\lambda_B} \Omega R_e B_r \cos \lambda\, d\lambda$$

$$= +H_0 R_e^2 \Omega \int_0^{\lambda_B} 2 \sin \lambda \cos \lambda\, d\lambda = +H_0 R_e^2 \Omega \sin^2 \lambda_B$$

With appropriate values for the Earth we find the potential difference between the north pole and the equator to be

$$V \sim 9 \times 10^4 \text{ V} = 90 \text{ kV}$$

when $R_e = 6.38 \times 10^6$ m, $\Omega = 7.27 \times 10^{-5}$ rad s^{-1}, and $H_0 = 3.0 \times 10^{-5}$ tesla. We also notice that the total electric field is in the meridional plane and given by

$$E = \Omega R_e H_0 \cos \lambda (1 + 3 \sin^2 \lambda)^{1/2} = 14 \cos \lambda (1 + 3 \sin^2 \lambda)^{1/2} \text{ mV/m}$$

4.14 MAGNETOSPHERIC CONVECTION CLOSE TO THE EARTH

Due to corotation of plasma on closed field lines, the motion of plasma in the inner magnetosphere will partly be dominated by this corotation and partly by magnetospheric convection (Figure 4.28). While corotation represents a constant rotation speed at a given distance, magnetospheric convection is not constant, but rather strongly dependent on interactions between the solar wind and the magnetosphere. The point on the evening side where the corotation velocity $\mathbf{v}_R$ is exactly

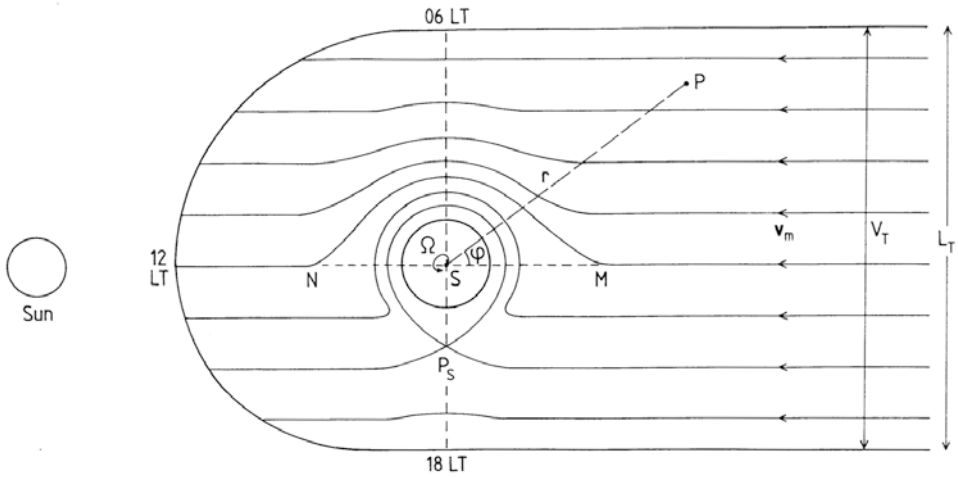


Figure 4.28. A diagram illustrating the analogy between the convection pattern in the magnetospheric equatorial plane including corotation of the plasmasphere and the streamlines in laminar flow around a cylinder. Plasma flow has a uniform flow velocity $\mathbf{v}_m$ directed towards the Sun far down the tail. The Earth rotates with an angular velocity $\boldsymbol{\Omega}$. The stagnation point P_S and a general point P in the flow is indicated. The width of the tail is L_T.

opposite to the convection velocity $\mathbf{v}_m$ will therefore move inwards and outwards from the Earth in concordance with variation in the potential V_T across the tail. The situation is very similar to laminar fluid flow around a rotating cylinder. In the stagnation region in the evening sector the convection velocity will be very low, and plasma can be brought to a halt. Therefore, the plasmasphere often has a bulge in this sector. Plasma may actually be forced to move in closed loops in the evening side depending on the history of the cross-tail potential and its instantaneous magnitude.

By analogy to the streamline function in hydrodynamical flow, we can describe the convection flow lines in a more mathematical sense.

From Maxwell's equation we know that the electric field can be derived from a potential. Let this be ϕ, then:

$$\mathbf{E} = -\nabla\phi = -\frac{\partial\phi}{\partial r}\hat{\mathbf{r}} - \frac{1}{r}\frac{\partial\phi}{\partial\varphi}\hat{\boldsymbol{\varphi}} - \frac{\partial\phi}{\partial z}\hat{\mathbf{z}}$$

The magnetic field in the equatorial plane is directed perpendicular to the plane along the z-direction and therefore

$$\mathbf{B} = B\hat{\mathbf{z}}$$

The velocity due to the $\mathbf{E} \times \mathbf{B}$ drift in the equatorial plane is now

$$\mathbf{v} = \mathbf{E} \times \mathbf{B}/B^2 = \frac{1}{B}\mathbf{E} \times \hat{\mathbf{z}} = -\frac{1}{B}\left[-\frac{\partial\phi}{\partial r}\hat{\boldsymbol{\varphi}} + \frac{1}{r}\frac{\partial\phi}{\partial\varphi}\hat{\mathbf{r}}\right]$$

and the velocity is totally in the equatorial plane.

From the hydrodynamical analogy of streamlines in laminar flow around a rotating cylinder, we have for the velocity field potential or the stream function:

$$\phi_v = -\frac{C_1}{r} + C_2 r \sin\varphi$$

where C_1 and C_2 are constants determined by rotation of a cylinder and flow velocity, and φ and r are as indicated in Figure 4.28. We can introduce this because we are treating flow only in the equatorial plane, such that the cross-section in Figure 4.28 is equivalent to the cross-section of flow around an infinitely long cylinder. By then introducing this stream function into the expression for $\mathbf{v}$, we get:

$$\begin{aligned}\mathbf{v} &= -\frac{1}{B}\left[-\frac{\partial}{\partial r}\left(-\frac{C_1}{r} + C_2 r\sin\varphi\right)\hat{\boldsymbol{\varphi}} + \frac{1}{r}\frac{\partial}{\partial\varphi}\left(-\frac{C_1}{r} + C_2 r\sin\varphi\right)\hat{\mathbf{r}}\right] \\ &= \left(\frac{C_1}{r^2} + C_2\sin\varphi\right)\frac{1}{B}\hat{\boldsymbol{\varphi}} - \frac{C_2}{B}\cos\varphi\,\hat{\mathbf{r}}\end{aligned}$$

We notice that for very large distances ($r \to \infty$)

$$\mathbf{v} = \frac{C_2}{B}(\sin\varphi\,\hat{\boldsymbol{\varphi}} - \cos\varphi\,\hat{\mathbf{r}})$$

which implies that $\mathbf{v}$ is parallel to the MSN line in Figure 4.28 and therefore

$$C_2 = E \qquad \text{and} \qquad \mathbf{v}(r=\infty) = \mathbf{v}_m$$

where $\mathbf{v}_m$ is the uniform convection velocity far down the tail away from Earth. For $r = R_e$ we have

$$\mathbf{v} = \left(\frac{C_1}{R_e^2} + E\sin\varphi\right)\frac{1}{B}\hat{\boldsymbol{\varphi}} - \frac{E}{B}\cos\varphi\,\hat{\mathbf{r}} = \frac{C_1}{BR_e^2}\hat{\boldsymbol{\varphi}} + \mathbf{v}_m$$

Since the velocity due to rotation of the cylinder at the surface of the cylinder is given by

$$\mathbf{v}_R = \Omega R_e\hat{\boldsymbol{\varphi}}$$

we must have

$$C_1 = B\Omega R_e^3$$

and finally:

$$\mathbf{v} = \left(\frac{\Omega R_e^3}{r^2} + \frac{E}{B}\sin\varphi\right)\hat{\boldsymbol{\varphi}} - \frac{E}{B}\cos\varphi\,\tilde{\mathbf{r}}$$

We notice that for $\varphi = \pm\pi/2$ the radial velocity vanishes, which means that $\mathbf{v}$ is parallel to $\mathbf{v}_m$ at the dawn meridian, while at the dusk meridian ($\varphi = -\pi/2$) the azimuthal component depends on the relation between $\Omega R_e^3/r$ and E/B. It can actually be zero when

$$E = \frac{B\Omega R_e^3}{r^2}$$

If this happens at the surface of the cylinder

$$E = B\Omega R_e = E_R$$

which is the corotating electric field at the Earth's equator. At a distance $r = L \cdot R_e$ we notice:

$$E = \frac{B\Omega R_e}{L^2} = \frac{1}{L^2} E_R$$

If the stagnation point occurs at $L = 2$, then the E-field is 1/4 of the corotating field at the equator. Therefore, a large convection electric field, which occurs under disturbed conditions in the magnetosphere, will push the stagnation point P_S towards the Earth and erode the plasmasphere.

We also notice that the potential ϕ at large distances along the dawn–dusk meridian will be positive on the dawn side and negative on the dusk side. The magnitude of the potential depends on the width L_T of the tail. If the stagnation point P_S in Figure 4.28 occurs at $4\,R_e$ and the width of the tail is $50\,R_e$, the dawn–dusk potential gets close to 47 kV. Moving the stagnation point to $3\,R_e$ would almost double the potential ($\sim$83 kV) for the same dimension of L_T.

We realize that plasma between the plasmasphere and the high-altitude F-region is distributed under the influence of a large variety of forces and processes. Due to the large variation in magnetospheric convection, the distance to separation between corotating plasma and plasma convected sunward in the magnetosphere beyond the Earth can change greatly, and thereby change the volume of the plasmasphere. The corotating electric field can be compared with the ionospheric field of magnetospheric origin (Figure 4.29) by assuming the magnetic field lines are equipotentials and applying the mapping factor derived in Chapter 3. We notice that the magnetospheric field corresponding to 1 mV/m at the equator will, due to the mapping factor, be comparable with the corotating field near an L value close to 3. The weaker electric field, however, corresponding to 0.4 mV/m becomes equal to the corotating field at an L value close to 6.

We therefore realize that the corotation field can dominate out to 6–7 R_e, while the magnetospheric field dominates outside this region.

Figure 4.30 illustrates how the position of the outer limit of the plasmasphere depends on magnetospheric disturbance as reflected by the Kp-index.

4.15 THE EQUATORIAL FOUNTAIN EFFECT

At the equator there is an effect that comes into play when discussing the behavior of the upper ionosphere, and that is the so-called fountain effect.

At the equator an eastward electric field is established in the F-region which creates an upward-moving ($\mathbf{E} \times \mathbf{B}$) plasma flow (Figure 4.31). At higher altitudes the recombination of upwelling plasma is slowed down compared with the

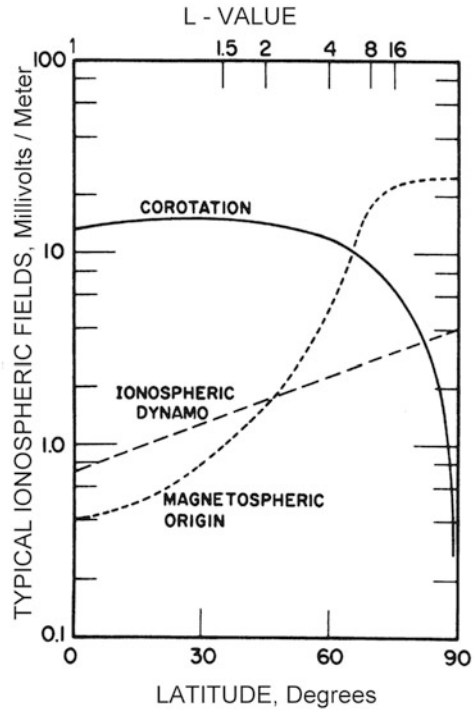


Figure 4.29. Comparison between the corotation electric field and electric fields of magnetospheric origin for different geomagnetic latitudes and L values. The magnetospheric fields are mapped down to the ionosphere assuming equipotential field lines. The ionospheric dynamo electric field is also shown for reference. (From Mozer, 1973.)

situation at lower altitudes just after sunset. Equatorial plasma will rise in the F-region until pressure forces are high enough for the plasma, assisted by gravity, to start to slide down the magnetic field lines toward the tropical ionosphere, where enhanced plasma densities are observed either side of the equator (Figure 4.32).

4.16 EXERCISES

1. Prove the relation given by (4.37).
2. Assume that we have an electron density profile given by:

$$n_e(x) = n_{e_0} \exp(1 - x - \sec \chi \cdot \exp(-x))$$

   (a) Show that the maximum value of $n_e(x)$, $n_m = n_{e_0} \sin \chi$

   (b) Find the total electron content in a unit column of this profile from ground to infinity.

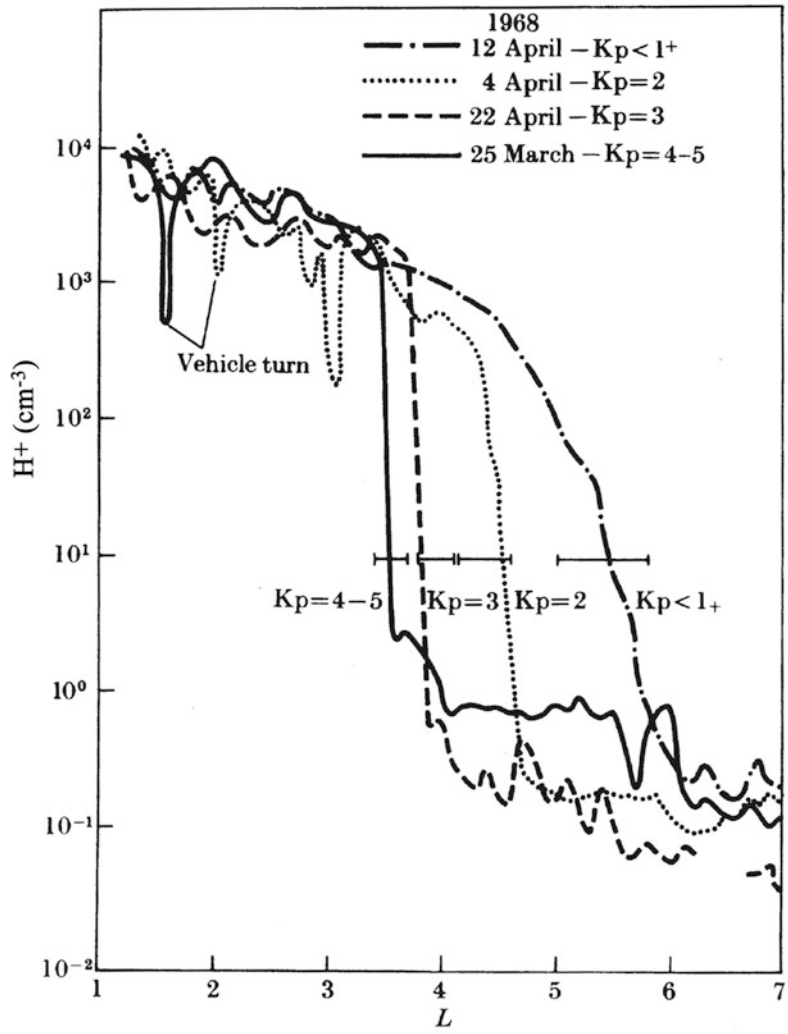


Figure 4.30. Observed H^+ densities for different distances from the Earth surface in the equatorial plane showing a sharp decrease at distances between 3.5 and 6 R_e for different levels of magnetospheric disturbance. The sharp discontinuity marks the position of the plasmapause. (From Chappel *et al.*, 1970.)

Assume instead this profile is given by

$$n_e(x) = n_{e_0} \exp\left[\tfrac{1}{2}(1 - x - \sec\chi \exp(-x))\right]$$

(c) Show that the maximum value of $n_e(x)$, $n_m = n_{e_0}(\cos\chi)^{1/2}$

(d) Find the total electron content of this profile in a unit column from ground to infinity.

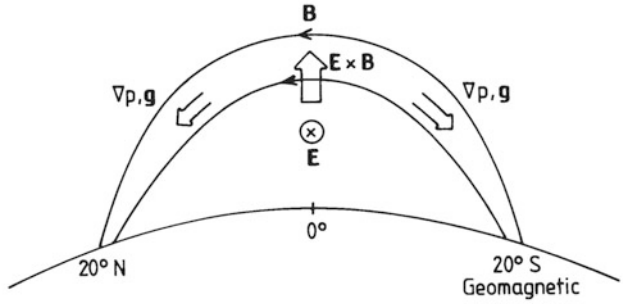


Figure 4.31. A simplified sketch to illustrate the fountain effect at the geomagnetic equator. An eastward field at the geomagnetic equator drives plasma vertically upward. Pressure gradients and gravity drives it down again along the magnetic field lines toward higher latitudes. (After Kelley, 1989.)

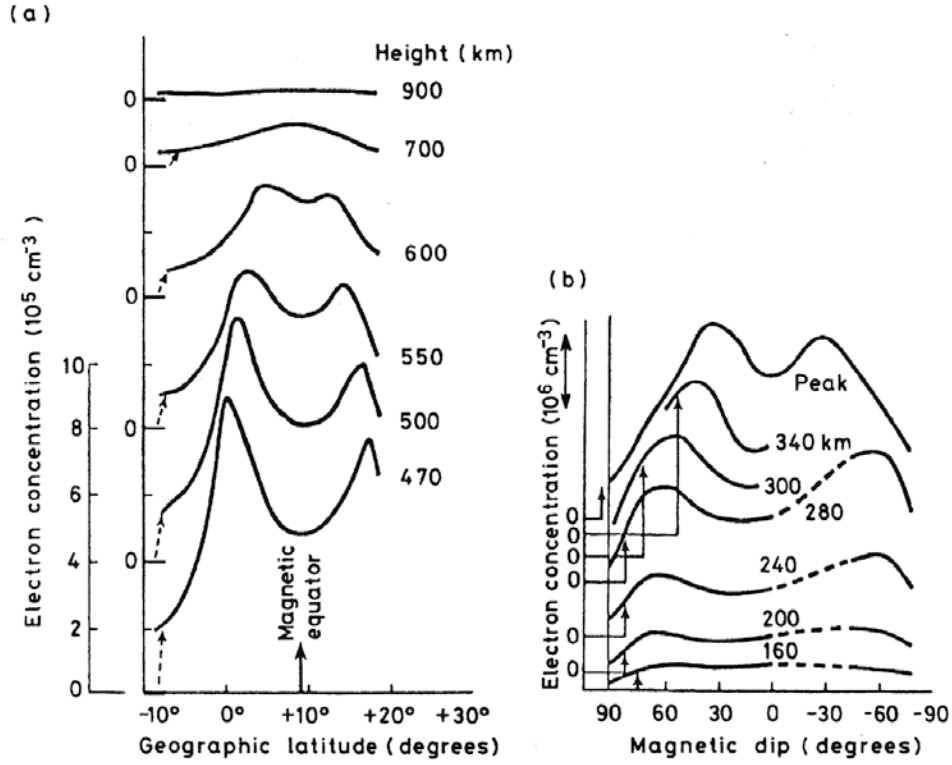


Figure 4.32. Latitudinal variation of electron density across the equator at various altitudes (a) from topside sounding and (b) from bottomside sounding. (From Croom *et al*., 1959; Eccles and King, 1969; and Hargreaves, 1979.)

5

Currents in the ionosphere

5.1 THE STEADY-STATE APPROACH

The momentum equation of the ion species j moving in a plasma in the presence of an electric (**E**) and a magnetic (**B**) field can be expressed as:

$$\rho_j \frac{d\mathbf{v}_j}{dt} = -\nabla p_j + \rho_j \mathbf{g} + \frac{q_j \rho_j}{m_j}(\mathbf{E} + \mathbf{v}_j \times \mathbf{B}) - \sum_{\substack{k \\ j \neq k}} \rho_j \nu_{jk}(\mathbf{v}_j - \mathbf{v}_k) \tag{5.1}$$

where $\mathbf{v}_j$ is the ion velocity vector, ρ_j and p_j are the mass density and pressure of the ion species j, m_j is the ion mass, ν_{jk} is the collision frequency between ion species j and k, q_i is the ionic charge, and **g** is the acceleration of gravity. The summation is taken over all other species except j.

Usually, it is also assumed that the number density of positive ions is set equal to the number density of electrons, thus

$$n_e = n_i = \sum_{\text{ions}} n_j$$

Each ion is then to have only one electric charge.

In order to give a complete description of plasma we would have to treat each individual ion species by (5.1). This would, however, lead to a large set of coupled equations due to the collision terms. In order to understand the basic features of ionospheric dynamics it is therefore more practical to consider a single ion species of mass m_i or, equivalently, assume a pseudo-ion derived by proper weighting of all ions with respect to their individual mass, density, charge, velocity, and temperature.

A. Brekke, *Physics of the Upper Polar Atmosphere*, Springer Praxis Books,
DOI 10.1007/978-3-642-27401-5_5, © Springer-Verlag Berlin Heidelberg 2013

Neglecting the pressure and gravity terms, the momentum equations for ions and electrons can respectively be expressed as:

$$n_i m_i \frac{d\mathbf{v}_i}{dt} = n_i e(\mathbf{E} + \mathbf{v}_i \times \mathbf{B}) - n_i m_i \nu_{in}(\mathbf{v}_i - \mathbf{u}_n) \tag{5.2}$$

$$n_e m_e \frac{d\mathbf{v}_e}{dt} = -n_e e(\mathbf{E} + \mathbf{v}_e \times \mathbf{B}) - n_e m_e \nu_{en}(\mathbf{v}_e - \mathbf{u}_n) \tag{5.3}$$

where neutrality is maintained ($n_e = n_i$) and only single-charged ions are involved ($q_i = e$). Here $\mathbf{E}$ is the electric field one would measure in an Earth-fixed reference frame, which is usually the field measured in ionospheric experiments. We have here neglected ion–ion and ion–electron collision frequencies as these are usually small compared with the ion–neutral and electron–neutral collision frequencies, at least in the lower part of the ionosphere (the E-region).

An approximate formula for the ion–neutral collision frequency found in the literature is:

$$\nu_{in} = 2.6 \times 10^{-15}(n_n + n_i)(M_n')^{-1/2}$$

where M_n' denotes the mean neutral molecular mass. Densities are given in m^{-3}. Sometimes an alternative expression separating the contribution from each neutral species is used, especially in E-region studies:

$$\nu_{in} = k_{\mathrm{N}_2} \cdot [\mathrm{N}_2] + k_{\mathrm{O}_2} \cdot [\mathrm{O}_2] + k_{\mathrm{O}} \cdot [\mathrm{O}]$$

where k_{N_2}, k_{O_2}, and k_{O} are the different collision rates according to the neutral and ionized species involved, and the brackets denote the density of each individual neutral species. For an average ion mass number of 30.7, Table 5.1 gives the characteristic collision rates (Kunitake and Schlegel, 1991).

Table 5.1. The collision rate coefficients in m^3/s.

k_{N_2}	k_{O_2}	k_{O}
4.34×10^{-16}	4.28×10^{-16}	2.44×10^{-16}

An approximate formula for electron collision frequency is:

$$\nu_e \equiv \nu_{en} + \nu_{ei} = 5.4 \times 10^{-10} n_n \cdot T_e^{1/2} + [34 + 4.18 \cdot \ln(T_e^3/n_e)] n_e T_e^{-3/2}$$

The densities of the different species are here given in cm^{-3}, and T_e is given in kelvins.

Typical values for ion–neutral (ν_{in}) and electron–neutral (ν_{en}) collision frequencies are shown as a function of altitude between 90 and 400 km for an auroral latitude station (Tromsø) (Figure 5.1). Also indicated for comparison are the corresponding ion (Ω_i) and electron (Ω_e) gyrofrequencies in the same height region.

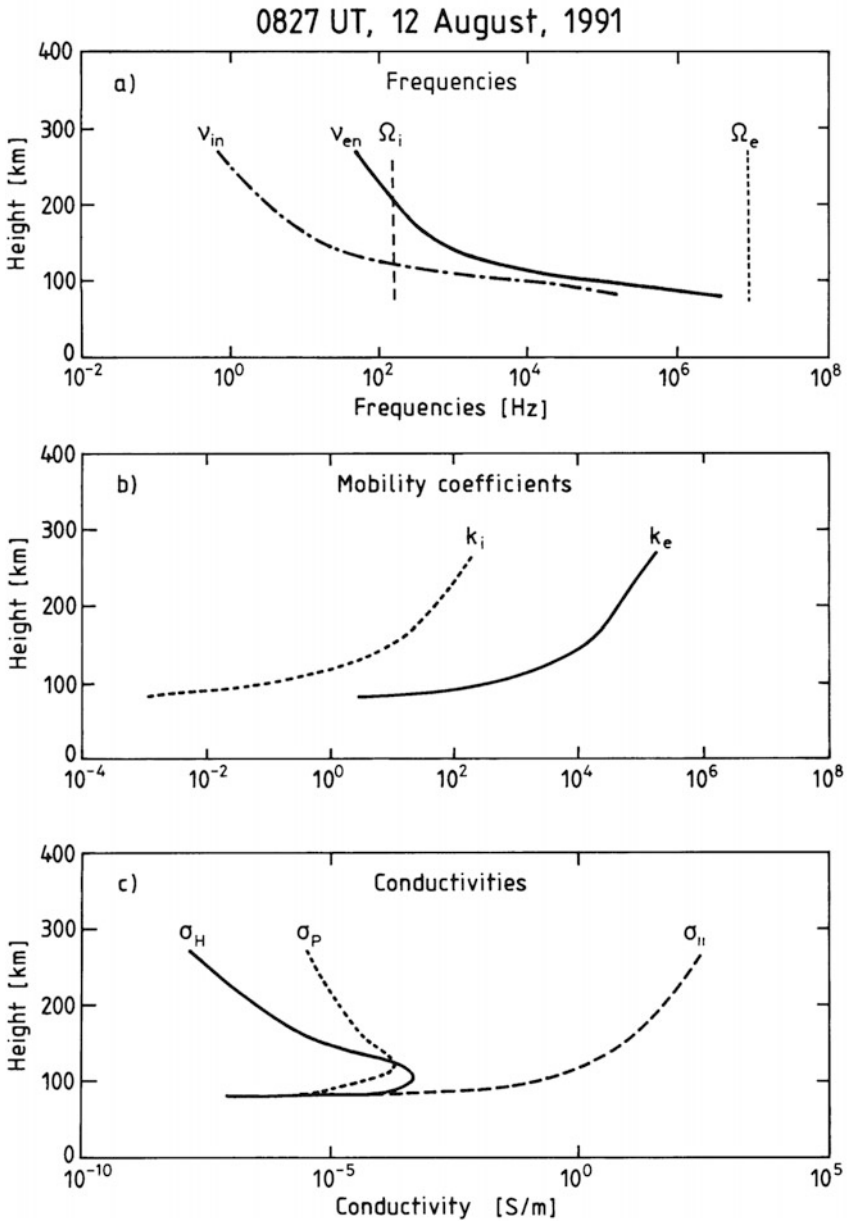


Figure 5.1. (a) Typical altitude profiles of ion neutral (v_{in}) and electron–neutral (v_{en}) collision frequencies for an auroral zone station such as Tromsø (69°39′ N, 18°56′ E). Also indicated for comparison are the ion (Ω_i) and electron (Ω_e) gyrofrequencies. (b) Altitude profiles of the ion ($k_i = \Omega_i/v_{in}$) and electron ($k_e = \Omega_e/v_{en}$) mobility coefficients derived from the profiles given in panel (a). (c) Altitude profiles of the Pedersen (σ_P), Hall (σ_H), and parallel ($\sigma_\parallel$) conductivities as derived by EISCAT at Tromsø at 08:27 UT on August 12, 1991. These profiles are typical of quiet summer days. (Courtesy Blixt, 1995.)

When we compare the terms including velocities in the mobility equation, we find that the acceleration term on the left-hand side is of the order v_j/τ_j or v_j^2/L, where τ_j is the response time and L is a distance characteristic for velocity change. The Lorentz term on the right-hand side is of the order of $v_j\Omega_j$, where Ω_j is the gyrofrequency (q_jB/m_j) for ion species j, and the frictional term is of order v_{jk} where v_{jk} is the collision frequency. As long as $\tau_j \gg \Omega_j^{-1}$ or $\tau_j \gg v_j^{-1}$ the inertia term can be neglected. The collision frequency and gyrofrequency are sufficiently large that in most problems of interest to macroscopic dynamics this is a valid simplification.

The steady-state solution of the ion momentum equation (5.2) is:

$$n_i e(\mathbf{E} + \mathbf{v}_i \times \mathbf{B}) - n_i m_i v_{in}(\mathbf{v}_i - \mathbf{u}_n) = 0 \tag{5.4}$$

and for the electrons (5.3):

$$-n_e e(\mathbf{E} + \mathbf{v}_e \times \mathbf{B}) - n_e m_e v_{en}(\mathbf{v}_e - \mathbf{u}_n) = 0 \tag{5.5}$$

where m_e is the electron mass.

The ion velocity can be derived from (5.4):

$$\mathbf{v}_i = \mathbf{u}_n + \frac{e}{m_i v_{in}}(\mathbf{E} + \mathbf{v}_i \times \mathbf{B}) \tag{5.6}$$

and similarly the electron velocity from (5.5):

$$\mathbf{v}_e = \mathbf{u}_n - \frac{e}{m_e v_{en}}(\mathbf{E} + \mathbf{v}_e \times \mathbf{B}) \tag{5.7}$$

By adding and subtracting a term $\mathbf{u}_n \times \mathbf{B}$ in the bracket on the right-hand side of (5.6) and rearranging we get:

$$\mathbf{v}_i' = \frac{e}{m_i v_{in}}\mathbf{E}' + \frac{e}{m_i v_{in}}\mathbf{v}_i' \times \mathbf{B} \tag{5.8}$$

Here

$$\mathbf{E}' = \mathbf{E} + \mathbf{u}_n \times \mathbf{B} \tag{5.9}$$

and

$$\mathbf{v}_i' = \mathbf{v}_i - \mathbf{u}_n$$

are the electric field and ion velocity measured in a reference frame moving with the neutral wind, respectively.

A similar expression can be derived for the electron velocity from (5.7):

$$\mathbf{v}_e' = -\frac{e}{m_e v_{en}}\mathbf{E}' - \frac{e}{m_e v_{en}}\mathbf{v}_e' \times \mathbf{B} \tag{5.10}$$

where

$$\mathbf{v}_e' = \mathbf{v}_e - \mathbf{u}_n$$

is the electron velocity observed from a reference frame moving with the neutral wind.

Let us now assume that the neutral wind can be neglected, then the ion velocity from (5.6) is given by:

$$\mathbf{v}_i = \frac{k_i}{B}\mathbf{E} + \frac{k_i}{B}\mathbf{v}_i \times \mathbf{B} \tag{5.11}$$

Here $k_i = \Omega_i/\nu_{in}$ is the ion mobility coefficient when $\Omega_i = eB/m_i$. This equation has the same form as (5.8). Therefore, to solve (5.8) we only replace $\mathbf{v}_i$ and $\mathbf{E}$ by $\mathbf{v}_i'$ and $\mathbf{E}'$, respectively, in the following. By now forming $\mathbf{v}_i \times \mathbf{B}$ from (5.11) we get

$$\mathbf{v}_i \times \mathbf{B} = \frac{k_i}{B}\mathbf{E} \times \mathbf{B} + \frac{k_i}{B}(\mathbf{v}_i \cdot \mathbf{B})\mathbf{B} - k_i B\mathbf{v}_i$$

We also find from (5.8) that

$$\mathbf{v}_i \cdot \mathbf{B} = \frac{k_i}{B}\mathbf{E} \cdot \mathbf{B}$$

and therefore

$$\mathbf{v}_i \times \mathbf{B} = \frac{k_i}{B}\mathbf{E} \times \mathbf{B} + \left(\frac{k_i}{B}\right)^2(\mathbf{E} \cdot \mathbf{B})\mathbf{B} - k_i B\mathbf{v}_i$$

By introducing this into (5.11) we find:

$$\mathbf{v}_i = \frac{k_i}{B}\mathbf{E} + \left(\frac{k_i}{B}\right)^2\mathbf{E} \times \mathbf{B} + \left(\frac{k_i}{B}\right)^3(\mathbf{E} \cdot \mathbf{B})\mathbf{B} - k_i^2\mathbf{v}_i$$

and solving for $\mathbf{v}_i$ we derive:

$$\mathbf{v}_i = \frac{1}{1 + k_i^2}\left[\frac{k_i}{B}\mathbf{E} + \left(\frac{k_i}{B}\right)^2\mathbf{E} \times \mathbf{B} + \left(\frac{k_i}{B}\right)^3(\mathbf{E} \cdot \mathbf{B})\mathbf{B}\right] \tag{5.12}$$

This equation represents a decomposition of $\mathbf{v}_i$ into three orthogonal components with respect to $\mathbf{E}$ and $\mathbf{B}$. The first term in the brackets in (5.12) represents the component along $\mathbf{E}$ while the last term represents the component along $\mathbf{B}$, and finally the middle term is along the direction perpendicular to both $\mathbf{E}$ and $\mathbf{B}$.

We also notice that by solving (5.10) while neglecting the neutral wind ($\mathbf{u}_n = 0$) we find the corresponding expression to $\mathbf{v}_e$:

$$\mathbf{v}_e = \frac{1}{1 + k_e^2}\left[-\frac{k_e}{B}\mathbf{E} + \left(\frac{k_e}{B}\right)^2\mathbf{E} \times \mathbf{B} - \left(\frac{k_e}{B}\right)^3(\mathbf{E} \cdot \mathbf{B})\mathbf{B}\right] \tag{5.13}$$

Here k_e $(= \Omega_e/\nu_{en})$ is the electron mobility coefficient when $\Omega_e = eB/m_e$. Figure 5.1(b) presents the altitude variations of k_i and k_e between 90 and 400 km for an auroral zone station (Tromsø).

Now, allowing the neutral wind $\mathbf{u}_n$ to be different from zero, we can, as mentioned above, replace $\mathbf{v}_i$, $\mathbf{v}_e$, and $\mathbf{E}$ by $\mathbf{v}_i'$, $\mathbf{v}_e'$, and $\mathbf{E}'$, respectively, in (5.12) and

(5.13) and obtain

$$\mathbf{v}_i' = \mathbf{v}_i - \mathbf{u}_n = \frac{1}{1+k_i^2}\left[\frac{k_i}{B}\mathbf{E}' + \left(\frac{k_i}{B}\right)^2 \mathbf{E}' \times \mathbf{B} + \left(\frac{k_i}{B}\right)^3 (\mathbf{E}' \cdot \mathbf{B})\mathbf{B}\right] \tag{5.14a}$$

$$\mathbf{v}_e' = \mathbf{v}_e - \mathbf{u}_n = \frac{1}{1+k_e^2}\left[-\frac{k_e}{B}\mathbf{E}' + \left(\frac{k_e}{B}\right)^2 \mathbf{E}' \times \mathbf{B} - \left(\frac{k_e}{B}\right)^3 (\mathbf{E}' \cdot \mathbf{B})\mathbf{B}\right] \tag{5.14b}$$

From Figure 5.1(b) we notice that $k_i \ll 1$ below about 100 km and, therefore, $\mathbf{v}_i \approx \mathbf{u}_n$. The ion velocity is then completely determined by the neutral wind, and the ions are closely coupled to the neutral gas in the bottom part of the E-region. Ions can be used as tracers for neutral gas motion.

5.2 DEPENDENCE OF ION VELOCITY DIRECTION ON ALTITUDE

We will neglect all other forces except the electrostatic field $\mathbf{E}$ for a while and assume that $\mathbf{E}$ is perpendicular to $\mathbf{B}$ ($\mathbf{E} \cdot \mathbf{B} = 0$). Then, (5.12) takes the simplified form:

$$\mathbf{v}_i = \frac{k_i}{1+k_i^2}\frac{\mathbf{E}}{B} + \frac{k_i^2}{1+k_i^2}\frac{\mathbf{E} \times \mathbf{B}}{B^2} \tag{5.15}$$

In a Cartesian coordinate system (x, y, z) where $\mathbf{B} = B\,\hat{\mathbf{z}}$ and $\mathbf{E} = E\,\hat{\mathbf{y}}$, $\mathbf{E} \times \mathbf{B}$ will be in the positive x-direction (Figure 5.2). The ion velocity $\mathbf{v}_i$ will therefore make an angle θ_i with the applied electrostatic field $\mathbf{E}$, and this angle is given by

$$\tan\theta_i = k_i = \frac{\Omega_i}{\nu_{in}} \tag{5.16}$$

Velocity therefore rotates clockwise when seen along $\mathbf{B}$ and ν_{in} increases along $\mathbf{B}$. In the ionosphere at about 125 km altitude $\nu_{in} = \Omega_i$. At this altitude then

$$\theta_i = 45^\circ$$

Lower down in the ionosphere where $\nu_{in} > \Omega_i$ the ion velocity will be increasingly along $\mathbf{E}$ as the collision frequency increases. The ion velocity in the D-region is therefore parallel to the E-field in the absence of other forces. Above 250 km the ion velocity is very close to the $\mathbf{E} \times \mathbf{B}$ direction in the absence of any other forces.

As we can observe ion velocities in the ionosphere by incoherent scatter radar experiments, we can derive at different heights the angle between the electric field and the ion velocity. If this angle does not agree with the angle θ_i derived above, we can infer that there must have been a neutral wind present.

We also notice that the magnitude of $\mathbf{v}_i$ from (5.15) is given by

$$v_i = \left\{\left[\left(\frac{k_i}{1+k_i^2}\right)^2 + \left(\frac{k_i^2}{1+k_i^2}\right)^2\right]\frac{E^2}{B^2}\right\}^{1/2} = k_i \cdot (1+k_i^2)^{-1/2}\frac{E}{B} = \sin\theta_i \frac{E}{B}$$

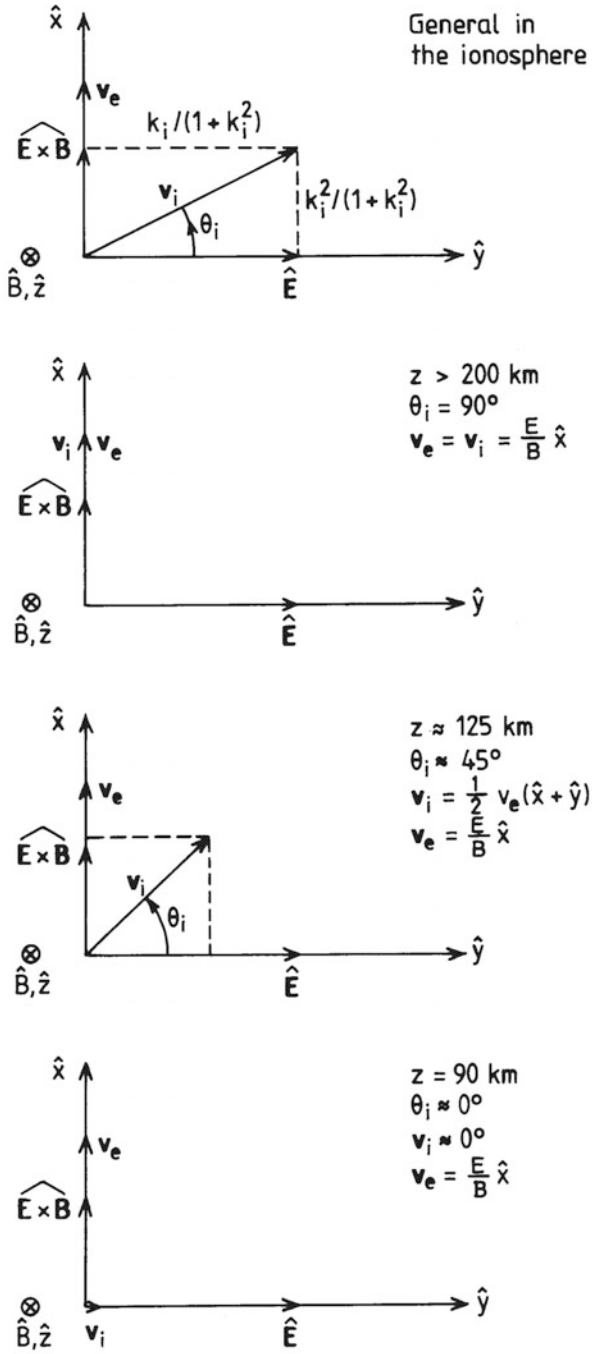


Figure 5.2. Vector diagrams showing the variation of $\mathbf{v}_i$ and $\mathbf{v}_e$ in the applied electric field for three different altitudes in the ionosphere.

At the upper ionosphere where $\theta_i \approx 90^\circ$

$$v_i \approx E/B$$

and in the D-region where $\theta_i \approx 0$

$$v_i \approx 0$$

The variation of $\mathbf{v}_i$ with height in the ionosphere is illustrated in Figure 5.2.

If we do the same analysis for electrons, we get from (5.13):

$$\mathbf{v}_e = -\frac{k_e}{1+k_e^2}\frac{\mathbf{E}}{B} + \frac{k_e^2}{1+k_e^2}\frac{\mathbf{E}\times\mathbf{B}}{B^2} \tag{5.17}$$

and

$$\tan\theta_e = -k_e = -\frac{\Omega_e}{\nu_{en}} \tag{5.18}$$

At 90 km $\theta_e \approx -90^\circ$ or very close to the $\mathbf{E}\times\mathbf{B}$ direction (as illustrated in Figure 5.2). Since ν_{en} decreases above 90 km, θ_e will become increasingly closer to the $\mathbf{E}\times\mathbf{B}$ drift. From (5.17) we find the magnitude of $\mathbf{v}_e$ to be:

$$v_e = k_e(1+k_e^2)^{-1/2}\frac{E}{B} = -\sin\theta_e\frac{E}{B}$$

as $\theta_e = -90^\circ$

$$v_e \approx \frac{E}{B}$$

It is therefore reasonable to assume that electrons are moving in the $\mathbf{E}\times\mathbf{B}$ direction above 90 km when no forces other than the electric field are applied. Above 250 km or so electrons and ions are $\mathbf{E}\times\mathbf{B}$ drifting at the same speed when the E-field is the only applied force.

This rotation of the ion velocity with respect to the electric field and the electron velocity in the ionosphere is, as we have demonstrated above, a result of varying collision frequency with height.

We also notice from (5.8) that the ion velocity $\mathbf{v}_i'$ observed in a frame moving with the neutral wind will rotate with height with respect to $\mathbf{E}'$ similarly as $\mathbf{v}_i$ rotates with respect to $\mathbf{E}$ when the neutral velocity is zero in the Earth's fixed frame of reference.

The neutral wind, however, changes direction quite dramatically with height and, therefore, it is not very practicable to refer to the neutral wind frame when discussing plasma motion in this context.

5.3 CURRENT DENSITY IN THE IONOSPHERE

Current density at a given height in the ionosphere is given by:

$$\mathbf{j} = n_e \cdot e \cdot (\mathbf{v}_i - \mathbf{v}_e)$$

when we again assume singly charged ions and charge neutrality. Note that the

dimension of $\mathbf{j}$ is A/m^2 in standard units. Inserting (5.12) and (5.13) for $\mathbf{v}_i$ and $\mathbf{v}_e$ we find the following expression for current density

$$\mathbf{j} = n_e \cdot e \left\{ \left(\frac{k_e}{1+k_e^2} + \frac{k_i}{1+k_i^2} \right) \frac{\mathbf{E}_\perp}{B} - \left(\frac{k_e^2}{1+k_e^2} - \frac{k_i^2}{1+k_i^2} \right) \frac{\mathbf{E} \times \mathbf{B}}{B^2} + (k_e + k_i) \frac{\mathbf{E}_\parallel}{B} \right\} \quad (5.19)$$

where $\mathbf{E}_\perp$ and $\mathbf{E}_\parallel$ are the electric field components perpendicular to and parallel with $\mathbf{B}$ ($\mathbf{E} = \mathbf{E}_\perp + \mathbf{E}_\parallel$).

Let us now introduce the conductivities:

$$\sigma_P = \frac{n_e \cdot e}{B} \left(\frac{k_e}{1+k_e^2} + \frac{k_i}{1+k_i^2} \right) \quad (5.20a)$$

$$\sigma_H = \frac{n_e \cdot e}{B} \left(\frac{k_e^2}{1+k_e^2} - \frac{k_i^2}{1+k_i^2} \right) \quad (5.20b)$$

$$\sigma_\parallel = \frac{ne \cdot e}{B} (k_e + k_i) \quad (5.20c)$$

where σ_P, σ_H, and $\sigma_\parallel$ are the height-dependent Pedersen, Hall, and parallel conductivities, respectively. Equation (5.19) can then be expressed as:

$$\mathbf{j} = \sigma_P \mathbf{E}_\perp - \sigma_H \frac{\mathbf{E} \times \mathbf{B}}{B} + \sigma_\parallel \mathbf{E}_\parallel \quad (5.21)$$

Figure 5.3 presents some altitude profiles of the electron density as observed at an auroral zone station (Tromsø) for a quiet summer day. Also presented are the Pedersen and Hall conductivity profiles.

Figure 5.4 shows some more examples of Pedersen and Hall conductivity profiles as observed at an auroral zone station (Tromsø). During auroral particle precipitation events enhanced electron densities can increase conductivities at their maxima by a factor of 10 above the typical quiet time value. We notice that while the maximum Pedersen conductivity is approximately determined by height where $\Omega_i = \nu_{in}$, the maximum Hall conductivity is more closely related to the peak in the electron density profile. For this reason the Hall conductivity is more sensitive to high-energy auroral particle precipitation, which enhances ionization below 125 km very strongly.

By inserting (5.14a) and (5.14b) into (5.19) we can also derive the current as:

$$\mathbf{j}' = n_e \cdot e \left\{ \left(\frac{k_e}{1+k_e^2} + \frac{k_i}{1+k_i^2} \right) \frac{\mathbf{E}'_\perp}{B} - \left(\frac{k_e^2}{1+k_e^2} - \frac{k_i^2}{1+k_i^2} \right) \frac{\mathbf{E}' \times \mathbf{B}}{B^2} + (k_e + k_i) \frac{\mathbf{E}'_\parallel}{B} \right\} \quad (5.22)$$

$$\mathbf{j}' = \sigma_P \mathbf{E}'_\perp - \sigma_H \frac{\mathbf{E}' \times \mathbf{B}}{B} + \sigma_\parallel \mathbf{E}'_\parallel \quad (5.23)$$

The forms of (5.22) and (5.23) remain the same as those of (5.19) and (5.21), respectively, except that the electric field $\mathbf{E}'$ ($= \mathbf{E}'_\perp + \mathbf{E}'_\parallel$) is now observed in the reference frame following the neutral gas. Only when $\mathbf{u}_n = 0$ will the current

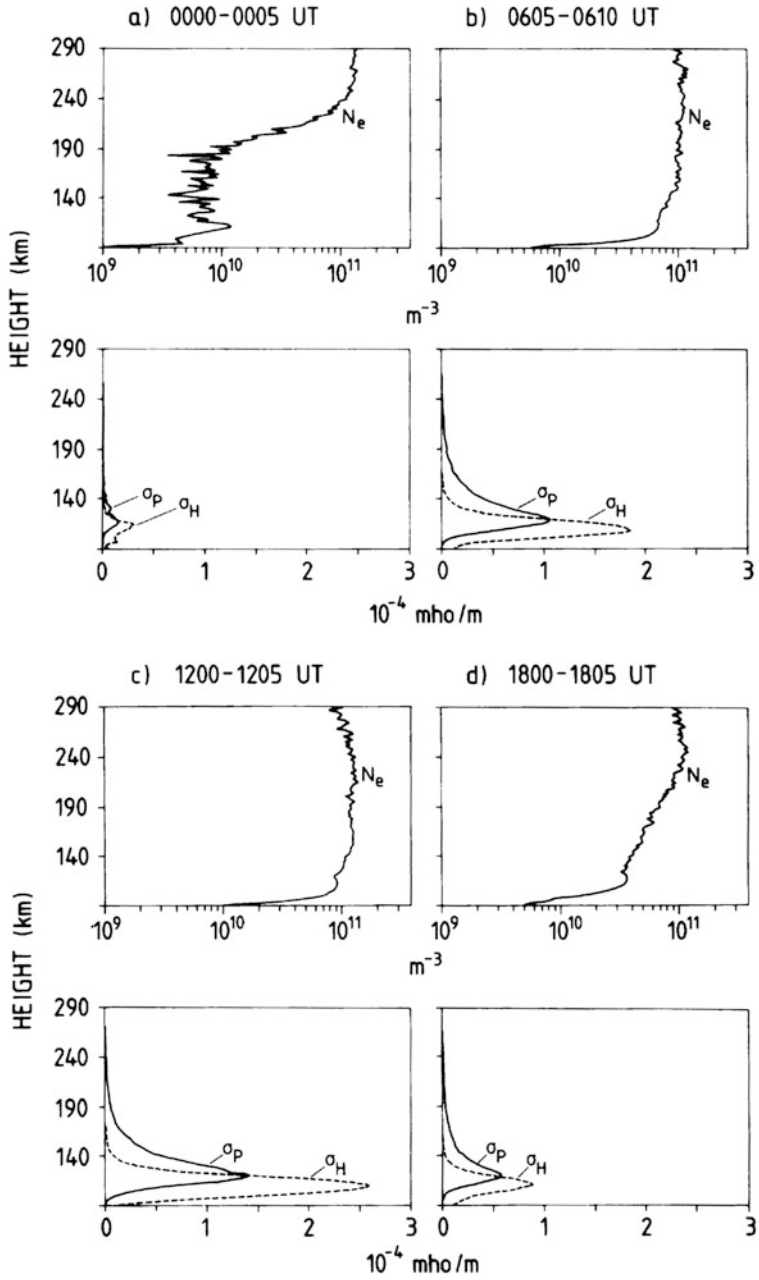


Figure 5.3. (Upper half) Typical quiet summer electron density profiles observed at night, morning, day, and evening in the auroral zone at Tromsø. (Lower half) Ionospheric $\sigma_H(z)$ and $\sigma_P(z)$ conductivity profiles derived from the observed electron density profiles above and model collision frequency. Electron density profiles are observed by the incoherent scatter radar, EISCAT. (From Brekke and Hall, 1988.)

densities given by (5.21) and (5.23) remain the same. For the situation when $\mathbf{u}_n \neq 0$, equation (5.23) is the proper expression for the current.

5.4 HEIGHT-DEPENDENT CURRENTS AND HEATING RATES

The main ionospheric currents contributing to observable effects on the ground flow mainly perpendicular to **B** below 300 km. Especially at high latitudes where the auroral electrojet dominates, it is not expected that perpendicular currents due to gravity and diffusion will play any important role. In the direction parallel to **B**, below 300 km or so, even at high latitudes the contribution to the parallel current due to gravity and diffusion will under normal conditions be rather small.

Similar to the analysis done in Section 5.2 of the variation with altitude of $\mathbf{v}_i$ with respect to the electric field we will now analyze how the current density vector changes direction with height in the ionosphere. Since the current density $\mathbf{j}$ derived for $\mathbf{u}_n = 0$ and $\mathbf{j}'$ derived for $\mathbf{u}_n \neq 0$ have a similar dependence on $\mathbf{E}$ and $\mathbf{E}'$, respectively, we can study the relationship between $\mathbf{j}$ and $\mathbf{E}$ and infer the derived results to the corresponding relationship between $\mathbf{j}'$ and $\mathbf{E}'$.

As we are mostly interested in ionospheric currents that can be related to magnetic fluctuations observed on the ground in auroral latitudes, we will consider only current densities perpendicular to **B**. Furthermore, these currents are predominantly present above 90 km so we can assume, according to Figure 5.1(b), that $k_e \gg 1$. The simplified expression for the height-dependent current density perpendicular to **B** in the auroral ionosphere is then according to (5.19) given by:

$$\mathbf{j}_\perp(z) = \frac{n_e(z) \cdot e}{B} \left\{ \frac{k_i}{1 + k_i^2} \mathbf{E}_\perp - \frac{1}{1 + k_i^2} \frac{\mathbf{E}_\perp \times \mathbf{B}}{B} \right\} \tag{5.24}$$

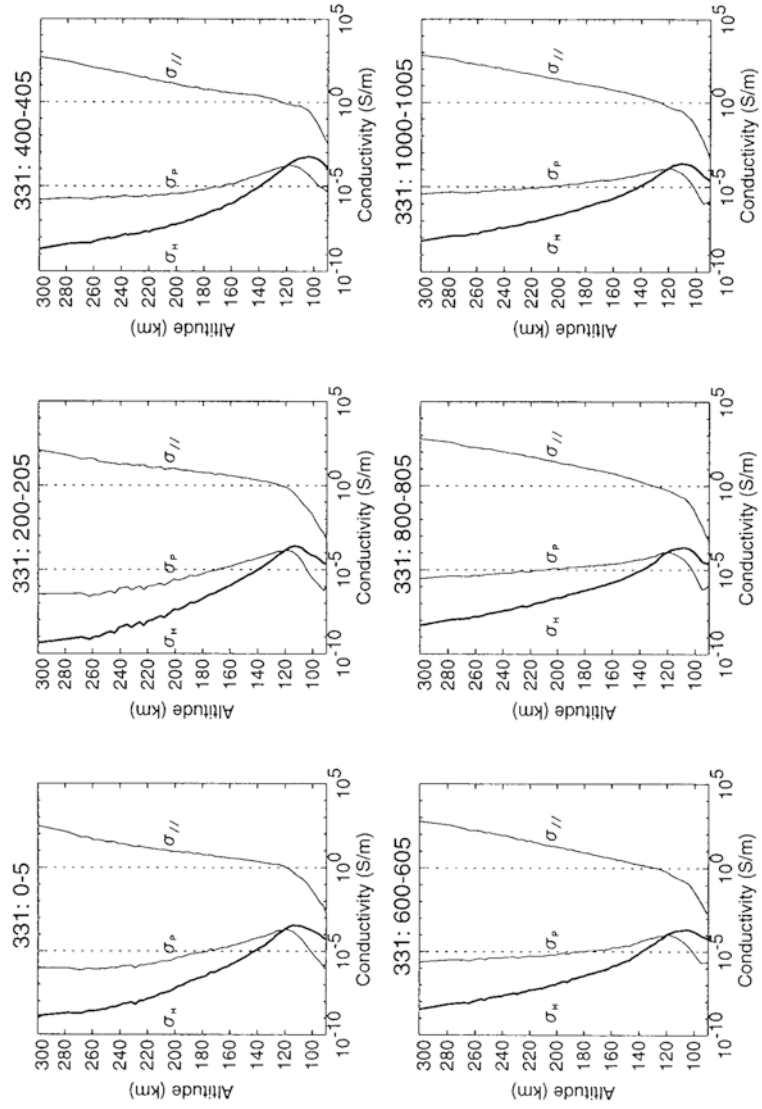
where both $n_e(z)$ and $\mathbf{j}_\perp(z)$ are height dependent. By maintaining the same coordinate system as in Figure 5.2, we find that the angle ϕ this current density makes with the **E**-field is given by:

$$\tan \phi = -\frac{1}{k_i}$$

Therefore, for this approximation $\tan \phi \cdot \tan \theta_i = -1$ where $\tan \theta_i$ is given by (5.16). In the absence of the neutral wind the current density is therefore perpendicular to the ion velocity everywhere in the auroral ionosphere. Furthermore, the current density is parallel to the **E**-field when k_i is very large (i.e., at the top of the ionosphere), and perpendicular to the **E**-field when $k_i \ll 1$ which is at E-region levels and below. The dominant current in the upper ionosphere in the F-region is therefore the Pedersen current, while it is the Hall current that dominates in the E-region.

We also notice that the height-dependent magnitude of current density is given by:

$$j(z)_\perp = \frac{e}{B} \frac{1}{\sqrt{1 + k_i^2}} n_e(z) E_\perp$$


331: 0-5
331: 200-205
331: 400-405
331: 600-605
331: 800-805
331: 1000-1005
Altitude (km)
Conductivity (S/m)
$\sigma_{//}$
σ_P
σ_H

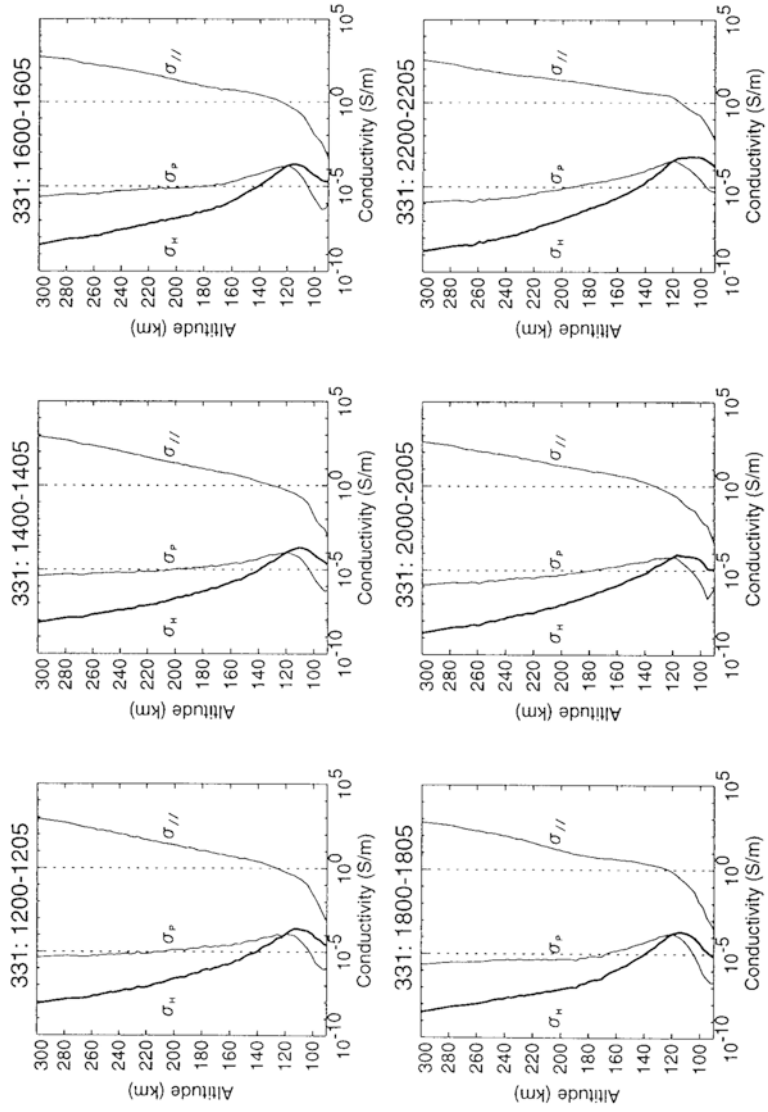


Figure 5.4. Examples of Pedersen, Hall, and parallel conductivity profiles (in siemens/m) as derived at the Tromsø auroral zone station for 12 different time intervals on March 31, 1992. (Courtesy Nozawa, 1995.)

Since $k_i \ll 1$ and the electron density $n_e(z)$ decreases very rapidly below the E-region peak, the current density below 90 km is often negligible.

The height-dependent heat dissipation rate due to current flowing along an electric field perpendicular to **B** (Joule heat dissipation) is given by:

$$q_J(z) = \mathbf{j}(z) \cdot \mathbf{E} = (\sigma_P(z)\mathbf{E}_\perp - \sigma_H(z)\mathbf{E}_\perp \times \mathbf{B}/B) \cdot \mathbf{E}_\perp = \sigma_P(z)E_\perp^2$$

where $\sigma_P(z)$ and $\sigma_H(z)$ are the height-dependent Pedersen and Hall conductivities, respectively. We see that the Hall current does not contribute to heat dissipation. Assuming that the electric field is constant with height and that the neutral wind as well as gravity and diffusion forces can be neglected, the altitude profile of the heat dissipation rate $q_J(z)$ is equivalent to the altitude profile of the Pedersen conductivity scaled by the constant factor $E_\perp^2$.

If we also include the neutral wind but still neglect gravity and diffusion forces, the current density vector, $\mathbf{j}'_\perp(z)$, for $k_e \gg 1$ perpendicular to **B** will according to (5.22) be given by

$$\mathbf{j}'(z) = \frac{n_e(z) \cdot e}{B}\left\{\frac{k_i}{1+k_i^2}\mathbf{E}'_\perp - \frac{1}{1+k_i^2}\mathbf{E}'_\perp \times \mathbf{B}/B\right\}$$

and the angle ϕ' that the current density vector makes with $\mathbf{E}'_\perp$ is given by

$$\tan\phi' = -\frac{1}{k_i} = \tan\phi$$

Since $\mathbf{E}'_\perp$ changes direction with altitude due to the neutral wind, ϕ' is not measured with respect to a fixed direction in space as would be the situation when $\mathbf{u}_n = 0$ and $\mathbf{E}_\perp$ is constant. Furthermore, the height-dependent magnitude of the current density at height z is given by

$$j(z)_\perp = \frac{e}{B}\frac{1}{\sqrt{1+k_i^2}}n_e(z)E'_\perp = \frac{e}{B}\frac{1}{\sqrt{1+k_i^2}}n_e(z)|\mathbf{E}_\perp + \mathbf{u}_{n\perp}(z) \times \mathbf{B}|$$

where $\mathbf{u}_{n\perp}(z)$ is the altitude profile of the neutral wind perpendicular to **B**. When the neutral wind is present, the current density profile with altitude is a product of altitude variations in the electron density profile and the neutral wind. Since the neutral wind in regions where most of the current flows is often strongly variable with height and with time, it is not a straightforward problem to identify the current density profile.

In Figure 5.5 we notice that the current density has a maximum at about 110 km and that it is mainly westward in the morning and eastward in the evening.

The electric field $\mathbf{E}_\perp$ can be measured by several techniques and, if it is measured at a point in the upper part of the ionosphere, it can be assumed constant at all altitudes down to about 90 km. The neutral wind, however, is a very elusive parameter and extremely difficult to measure simultaneously at different altitudes for considerable lengths of time. This makes the neutral wind extra-complicated to implement in models, and some averaging assumptions usually have to be made. Furthermore, since the neutral wind is so variable in the region where

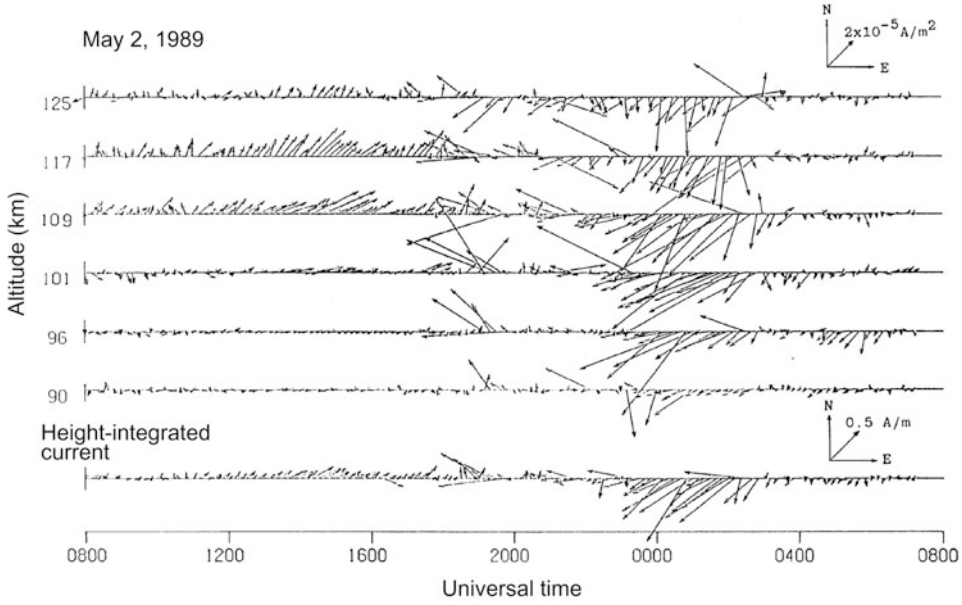


Figure 5.5. Current density vectors in A/m^2 observed at six different heights in the E-region by EISCAT on May 2, 1989. The bottom panel shows the height-integrated current. The scales are indicated by arrows labeled with proper factors. The northward and eastward directions are also indicated. (From Kamide and Brekke, 1993.)

most of the current flows, altitude profiles with reasonable good height resolution are needed.

For the height-dependent heat dissipation rate we now find, when including the effect of the neutral wind:

$$q'_J(z) = \mathbf{j}(z) \cdot \mathbf{E}'_\perp = \sigma_P(z)(\mathbf{E}'_\perp)^2 = \sigma_P(z)(\mathbf{E}_\perp + \mathbf{u}_n(z) \times \mathbf{B})^2$$

We notice that $q'_J(z)$ is always positive. In the presence of a neutral wind the heat dissipation rate has an altitude profile that can be very different from the altitude profile of the Pedersen conductivity (Figure 5.6). If a neutral wind-induced electric field happens to oppose the electrostatic field at one particular altitude such that:

$$\mathbf{E}_\perp + \mathbf{u}_n(z) \times \mathbf{B} = 0$$

then heat dissipation will also vanish at the same altitude. In the opposite sense the Joule heating rate may be strongly enhanced at a certain height if $\mathbf{E}_\perp$ and $\mathbf{u}_n(z)$ mutually enforce each other.

We notice from Figure 5.6 that the heating rate per particle, q_n, has its maximum at a greater height than the heating rate itself, and the height for maximum Pedersen conductivity is generally lower than both. Figure 5.7 shows some more examples of heating rate profiles at an auroral zone station (Tromsø).

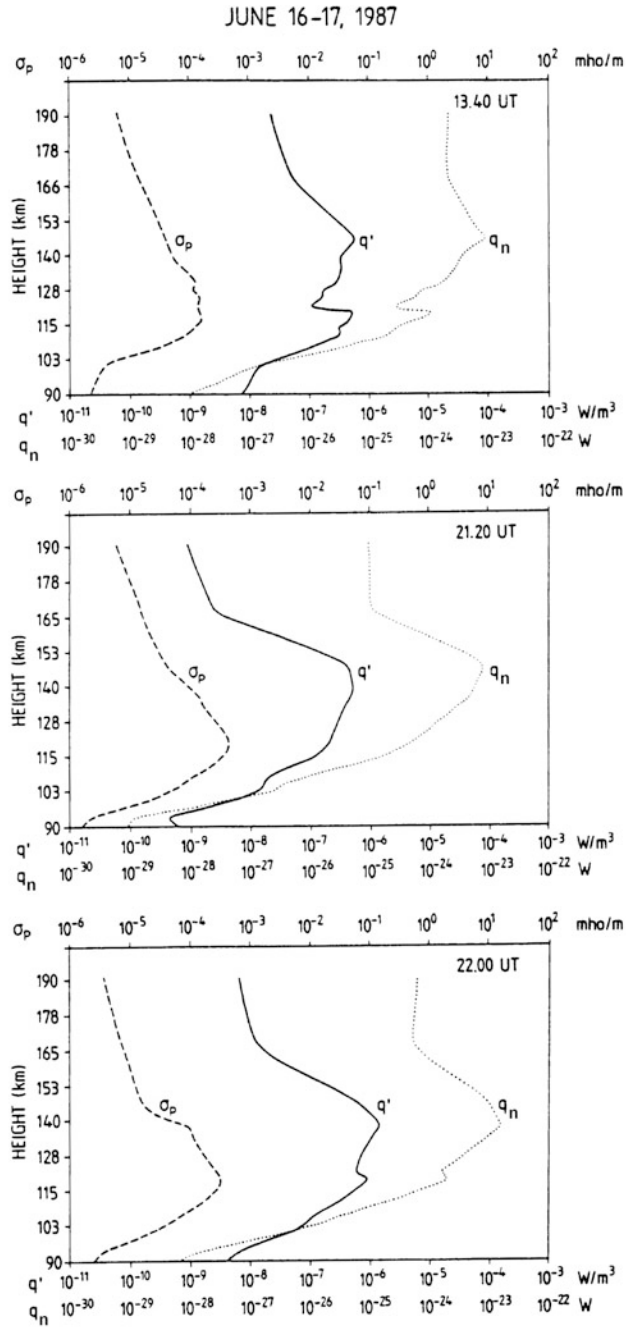


Figure 5.6. Height profiles of the Pedersen conductivity (σ_P) together with the Joule heating rate (q') and the Joule heating rate per particle (q_n) as obtained by EISCAT measurements at three different times on June 16, 1987. (From Brekke *et al.*, 1991.)

By taking the product of (5.23) and $\mathbf{E}'_\perp(z)$ we can form:

$$q'_J(z) = \mathbf{j}'(z) \cdot \mathbf{E}'_\perp(z) = \left(\sigma_P(z) \cdot \mathbf{E}'_\perp(z) - \sigma_H(z)\frac{\mathbf{E}'_\perp(z) \times \mathbf{B}}{B}\right) \cdot \mathbf{E}'_\perp(z)$$

$$= \sigma(z)(\mathbf{E}'_\perp(z))^2 \geq 0 \tag{5.25}$$

$q'_J(z)$ is always larger or equal to zero. Introducing $\mathbf{E}'_\perp = \mathbf{E}_\perp + \mathbf{u}_n(z) \times \mathbf{B}$ into (5.25) gives:

$$q'_J(z) = \mathbf{j}'(z) \cdot (\mathbf{E}_\perp + \mathbf{u}_n(z) \times \mathbf{B}) = \mathbf{j}'(z) \cdot \mathbf{E}_\perp + \mathbf{j}'(z) \cdot (\mathbf{u}_n(z) \times \mathbf{B}) \geq 0$$

Rearranging the last term gives:

$$q'_J(z) = \mathbf{j}'(z) \cdot \mathbf{E}_\perp - (\mathbf{j}'(z) \times \mathbf{B}) \cdot \mathbf{u}_n(z) \geq 0$$

where $\mathbf{j}(z) \times \mathbf{B}$ is a force due to the current. When $\mathbf{j}'(z) \cdot \mathbf{E}_\perp > 0$ then $q'_j(z) > 0$; when $(\mathbf{j}'(z) \times \mathbf{B}) \cdot \mathbf{u}_n(z) < 0$ or $(\mathbf{j}'(z) \times \mathbf{B}) \cdot \mathbf{u}_n(z) > 0$ then $|\mathbf{j}'(z) \cdot \mathbf{E}_\perp| > |(\mathbf{j}'(z) \times \mathbf{B}) \cdot \mathbf{u}_n(z)|$.

If, on the other hand, $\mathbf{j}'(z) \cdot \mathbf{E}_\perp < 0$, $(\mathbf{j}'(z) \times \mathbf{B})'\mathbf{u}_n(z)$ must also be negative and, furthermore, $|(\mathbf{j}'(z) \times \mathbf{B}) \cdot \mathbf{u}_n(z)| > |\mathbf{j}'(z) \cdot \mathbf{E}_\perp|$ in order for $q'_J(z)$ to be positive. When $(\mathbf{j}'(z) \times \mathbf{B}) \cdot \mathbf{u}_n(z) < 0$ the current force is acting against the neutral motion, and the neutrals act as a dynamo.

5.5 HEATING DUE TO COLLISIONS

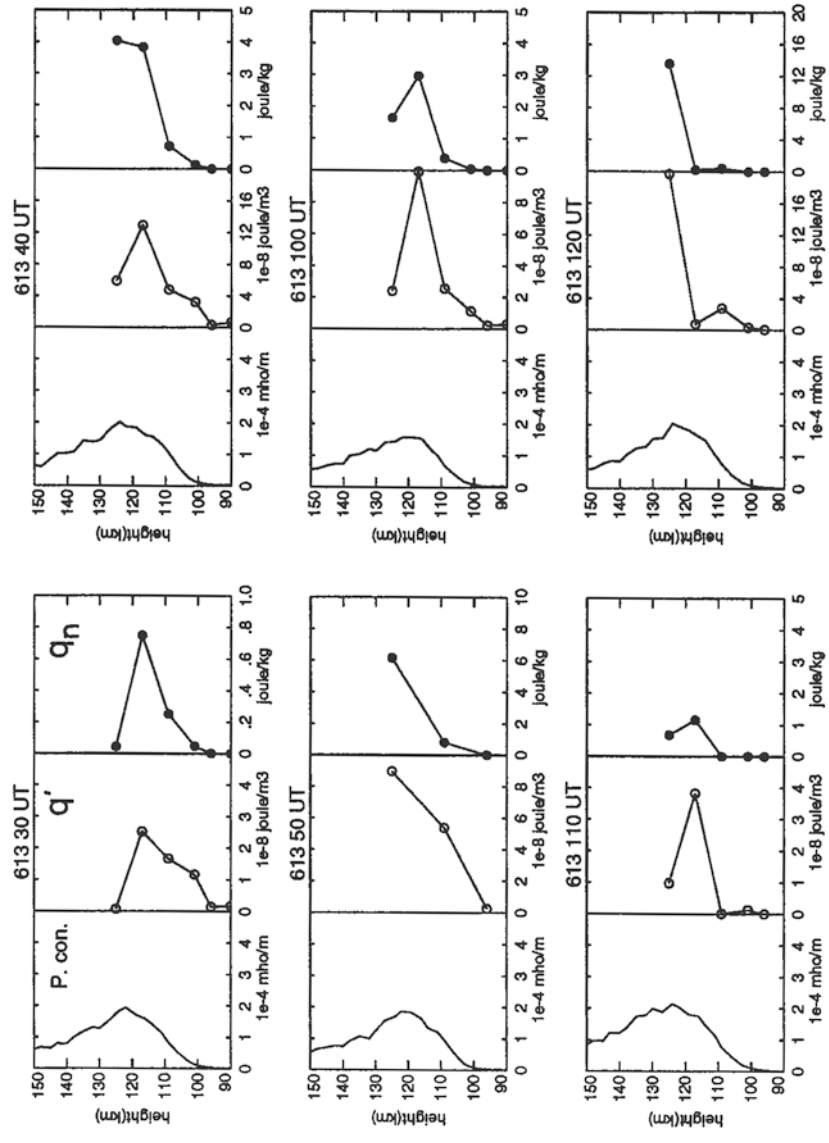
According to (2.31) the rate of energy transfer to ions due to collisions with neutrals, when it is assumed that $T_i = T_n$, is given by:

$$\frac{dQ_i}{dt} = \frac{\rho_i \nu_{in}}{m_n + m_i} m_n(\mathbf{v}_i - \mathbf{u}_n)^2 = \frac{m_n m_i}{m_n + m_i} n_i \nu_{in}(\mathbf{v}'_i)^2 \tag{5.26}$$

where $\mathbf{v}_i$ and $\mathbf{u}_n$ are the ion and neutral velocities observed in the Earth's fixed frame of reference, respectively, $\mathbf{v}'_i$ $(= \mathbf{v} - \mathbf{u}_n)$ is the relative velocity between ions and neutrals or, equivalently, the ion velocity observed in a frame moving with the neutral wind, m_i and m_n are the ion and neutral masses, respectively, and ν_{in} is the ion–neutral collision frequency. The ion mass density $\rho_i = m_i n_i$, where n_i is the ion particle number density equal to the electron density.

Since our aim here is to obtain the relationship between the rate of mechanical energy input to or frictional interaction with ions and neutrals due to collisions and Joule heating, we search for an alternative way of expressing $dQ_i/dt = \dot{Q}_i$. We therefore apply the simplified equation of motions for ions perpendicular to $\mathbf{B}$ as given by (5.2)

$$m_i \frac{d\mathbf{v}_i}{dt} = e(\mathbf{E} + \mathbf{v}_i \times \mathbf{B}) - m_i \nu_{in}(\mathbf{v}_i - \mathbf{u}_n)$$


613 30 UT
P. con.
q'
q_n
613 40 UT
613 50 UT
613 100 UT
613 110 UT
613 120 UT
height(km)
1e-4 mho/m
1e-8 joule/m3
joule/kg

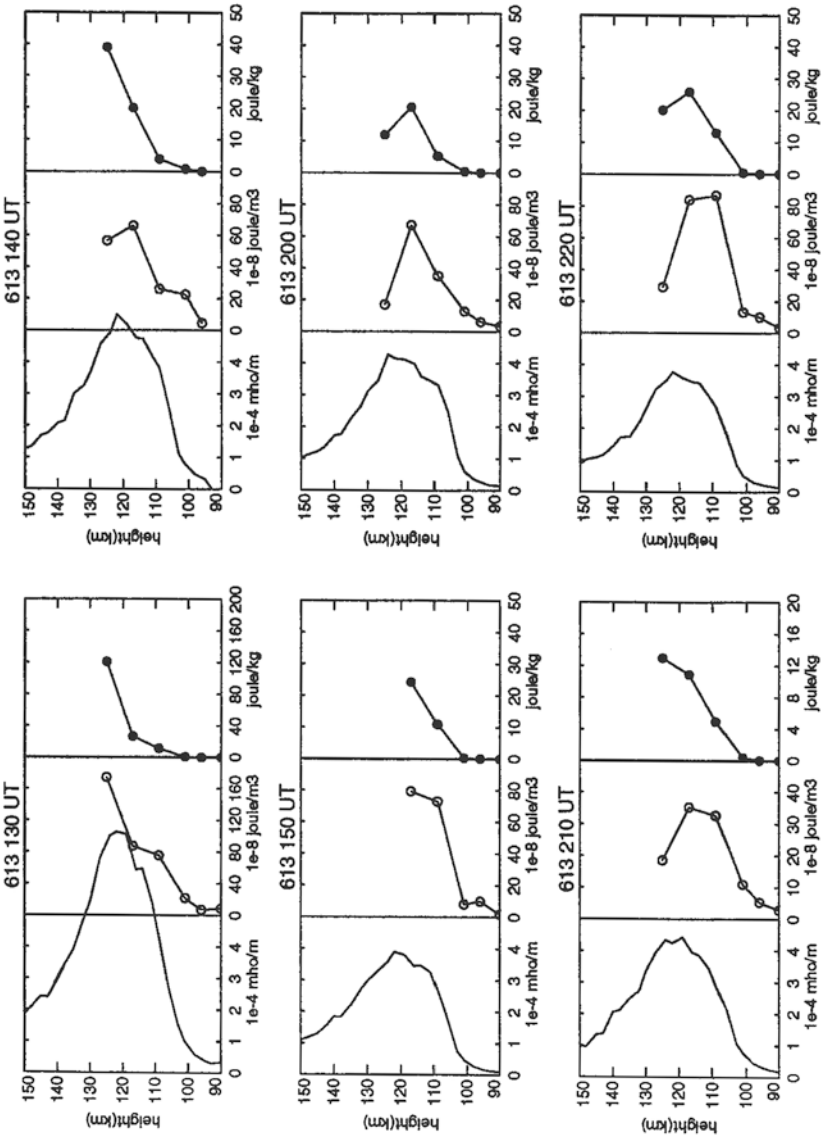


Figure 5.7. Pedersen conductivity P and Joule heating rate q' profiles as derived by EISCAT at Tromsø on June 13, 1990 in the morning hours. q_n represents the Joule heating rate per particle, as explained in the text. (Courtesy Nozawa, 1995.)

Adding and subtracting the term $e(\mathbf{u}_n \times \mathbf{B})$ on the right-hand side gives:

$$m_i \frac{d\mathbf{v}_i}{dt} = e(\mathbf{E}' + \mathbf{v}_i' \times \mathbf{B}) - m_i v_{in} \mathbf{v}_\mathbf{i}'$$

Reorganizing this equation and leaving only terms including $\mathbf{v}_i'$ on the left-hand side, we obtain

$$\Omega_i \mathbf{v}_i' \times \mathbf{B}/B - v_{in}\mathbf{v}_i' = \frac{d\mathbf{v}_i}{dt} - \Omega_i \frac{\mathbf{E}'}{B}$$

where the ion gyrofrequency $\Omega_i = eB/m_i$ is introduced. Squaring this equation and realizing that

$$\mathbf{v}_i' \cdot \mathbf{v}_i' \times \mathbf{B} = 0$$

and

$$(\mathbf{v}_i' \times \mathbf{B}) \cdot (\mathbf{v}_i \times \mathbf{B})/B^2 = (\mathbf{v}_i')^2$$

as $\mathbf{v}_i' \perp \mathbf{B}$, we find

$$(\Omega_i^2 + v_{in}^2)(\mathbf{v}_i')^2 = \left(\frac{d\mathbf{v}_i}{dt} - \Omega_i \frac{\mathbf{E}'}{B}\right)^2$$

The square of the relative velocity $\mathbf{v}_i'$ then becomes:

$$(\mathbf{v}_i')^2 = \frac{\Omega_i^2}{\Omega_i^2 + v_{in}^2}\left(\frac{1}{\Omega_i}\frac{d\mathbf{v}_i}{dt} - \frac{\mathbf{E}'}{B}\right)^2$$

By introducing this into (5.26) the rate of mechanical heat exchange can now be rewritten as:

$$\dot{Q}_i = \frac{m_i m_n}{m_n + m_i} n_i v_{in} \frac{\Omega_i^2}{\Omega_i^2 + v_{in}^2}\left(\frac{1}{\Omega_i}\frac{d\mathbf{v}_i}{dt} - \frac{\mathbf{E}'}{B}\right)^2 \tag{5.27}$$

In the E-region where dominant portions of the ionospheric current are flowing, we can neglect the collisions between electrons and the neutral gas and, therefore, Pedersen conductivity can be approximately expressed as ($n_i = n_e$):

$$\sigma_P = \frac{n_e \cdot e}{B}\frac{k_i}{1 + k_i^2} = \frac{n_e \cdot e}{B}\frac{\Omega_i v_{in}}{\Omega_i^2 + v_{in}^2}$$

In terms of σ_P the rate of mechanical heat exchange now becomes:

$$\dot{Q}_i = \frac{m_n}{m_i + m_n}\sigma_P \cdot B^2\left(\frac{1}{\Omega_i}\frac{d\mathbf{v}_i}{dt} - \frac{\mathbf{E}'}{B}\right)^2 \tag{5.28}$$

Under steady-state conditions ($d\mathbf{v}_i/dt = 0$) we have:

$$\dot{Q}_i^0 = \frac{m_n}{m_n + m_i}\sigma_P(\mathbf{E}')^2 = \frac{m_n}{m_n + m_i}\mathbf{J}' \cdot \mathbf{E}' = \frac{m_n}{m_n + m_i}\dot{Q}_j$$

where $\dot{Q}_j = \mathbf{J}' \cdot \mathbf{E}' = \sigma_P(\mathbf{E}')^2$ is the well-known Joule heating rate due to the ionospheric current density $\mathbf{J}'$. Since the steady-state rate of heat input to neutrals, on the other hand, can be expressed as

$$\dot{Q}_n^0 = \frac{m_i}{m_i + m_n} m_i n_i v_{in}(\mathbf{v}')^2 = \frac{m_i}{m_n}\dot{Q}_i^0$$

we have

$$\dot{Q}_i^0 + \dot{Q}_n^0 = \frac{m_n}{m_n + m_i}\dot{Q}_j + \frac{m_i}{m_n + m_n}\dot{Q}_j = \dot{Q}_j$$

That is, the Joule heating rate is equal to the total rate of mechanical energy input to neutrals and ions due to collisions between the two particle species. It is therefore only under steady-state conditions that the Joule heating rate $\dot{Q}_j$ can be set equal to the rate of mechanical energy input due to collisions. If steady-state conditions are not applicable, the result can be quite different, as we soon will see.

Let us return to the equation of motion again and notice that in the ionosphere where the ion velocity fluctuates more strongly than the neutral velocity:

$$\frac{d\mathbf{v}_i'}{dt} \approx \frac{d\mathbf{v}_i}{dt} \tag{5.29}$$

and that the equation of motion for ions in a reference frame moving with the neutral wind becomes:

$$m_i \frac{d\mathbf{v}_i'}{dt} = e(\mathbf{E}' + \mathbf{v}_i' \times \mathbf{B}) = m_i \nu_{in} \mathbf{v}_i' \tag{5.30}$$

With reference to (5.29) and introducing (5.30) into (5.28), we can express the rate of mechanical energy input as:

$$\begin{aligned}\dot{Q}_i &= \frac{m_n}{m_n + m_i}\sigma_P \cdot B^2 \left(\frac{1}{\Omega_i}\frac{d\mathbf{v}_i'}{dt} - \frac{\mathbf{E}'}{B}\right)^2 \\ &= \frac{m_n}{m_n + m_i}\left\{\sigma_P \cdot (\mathbf{E}')^2 + \sigma_P B^2 \left(\frac{1}{\Omega_i}\frac{d\mathbf{v}_i'}{dt}\right)^2 - \frac{2\sigma_P B}{\Omega_i}\frac{d\mathbf{v}_i'}{dt} \cdot \mathbf{E}\right\}\end{aligned} \tag{5.31}$$

The rate of energy input now consists of three parts: the first is due to the Joule heating rate, the second is related to the inertia motions of ions in the reference frame of neutrals, and the last term depends on the relationship between the electric field $\mathbf{E}'$ and the acceleration $d\mathbf{v}_i'/dt$ of ions in the same reference frame. If the electric field and the acceleration in the last term are in the same direction, then the term represents a maximum negative contribution to $\dot{Q}_i$, while in the opposite sense and when the acceleration is perpendicular to $\mathbf{E}'$ we notice that $\dot{Q}_i$ is larger than the Joule heating rate entering ions. When ions are accelerated up against the electric field, the dynamo action comes into play.

5.6 HEATING OF AN OSCILLATING ELECTRIC FIELD

In order to investigate further the relationship between $d\mathbf{v}_i'/dt$ and $\mathbf{E}'$, we will consider the case in which an oscillating electric field is present in the x–y plane perpendicular to $\mathbf{B}$. Expressing this as a circularly oscillating field by:

$$\mathbf{E}' = E_0(\cos \omega t\, \hat{\mathbf{x}} + \sin \omega t\, \hat{\mathbf{y}})$$

where E_0 is the amplitude, and $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are unit vectors for the x- and y-directions, respectively. Owing to this applied electric field, the ion velocity will also oscillate with the same frequency but out of phase with $\mathbf{E}'$. We therefore infer:

$$\mathbf{v}_i' = v_0(\cos(\omega t + \varphi)\hat{\mathbf{x}} + \sin(\omega t + \varphi)\hat{\mathbf{y}})$$

where φ is the arbitrary phase angle between $\mathbf{E}'$ and $\mathbf{v}_i'$, and v_0 is the amplitude of $\mathbf{v}_i'$.

It is now possible to express the rate of mechanical energy input from (5.31) as:

$$\dot{Q}_i = \frac{m_n}{m_n + m_i}\left\{\sigma_P E_0^2 + \sigma_P\left(\frac{B\omega}{\Omega_i}\right)^2 v_0^2 + \frac{2\sigma_P B\omega}{\Omega_i} v_0 E_0 \sin\varphi\right\} \tag{5.32}$$

In order to find the relationship between v_0 and E_0 and an expression for the phase angle φ we return to the equation of motion of ions (5.30). Introducing the expressions for $\mathbf{E}'$ and $\mathbf{v}_i'$ and separating the x- and y-components we find:

$$x\text{: } v_0(\Omega_i - \omega)\sin(\omega t + \varphi) = \Omega_i \frac{E_0}{B}\cos\omega t - v_0 v_{in}\cos(\omega t + \varphi) \tag{5.33a}$$

$$y\text{: } -v_0(\Omega_i - \omega)\cos(\omega t + \varphi) = \Omega_i \frac{E_0}{B}\sin\omega t - v_0 v_{in}\sin(\omega t + \varphi) \tag{5.33b}$$

when we have introduced $\mathbf{B} = -B\hat{\mathbf{z}}$. Multiplying (5.33a) by $\sin(\omega t + \varphi)$ and (5.33b) by $\cos(\omega t + \varphi)$ and subtracting the resulting equations, we find:

$$v_0 = \frac{\Omega_i}{\Omega_i - \omega}\frac{E_0}{B}\sin\varphi \tag{5.34}$$

Multiplying (5.33a) by $\cos(\omega t + \varphi)$ and (5.33b) by $\sin(\omega t + \varphi)$ and now adding the resulting equations, we find:

$$\Omega_i \frac{E_0}{B}\cos\varphi - v_0 v_{in} = 0$$

By introducing for v_0 from (5.34) we find:

$$\tan\varphi = \frac{\Omega_i - \omega}{v_{in}} \tag{5.35}$$

We notice that in a steady-state situation (namely, $\omega = 0$), the angle φ reduces to φ_0 as given by:

$$\tan\varphi_0 = \frac{\Omega_i}{v_{in}} = \tan\theta_i$$

which is exactly the angle between the ion velocity and the electric field (equation (5.16)). By now introducing the expressions for v_0 and φ from (5.34) and (5.35), respectively, into (5.32), the rate of mechanical energy becomes:

$$\dot{Q}_i = \frac{m_n}{m_n + m_i}\sigma_P E_0^2 \left\{ 1 + \left(\frac{\omega}{\Omega_i - \omega} \right)^2 \sin^2 \varphi + 2 \frac{\omega}{\Omega_i - \omega} \sin^2 \varphi \right\}$$

$$= \frac{m_n}{m_n + m_i}\sigma_P E_0^2 \left(1 + \frac{\Omega_i^2}{\nu_{in}^2} \right) \cos^2 \varphi \tag{5.36}$$

Applying the relation

$$\cos^2 \varphi = \frac{1}{1 + \tan^2 \varphi}$$

and introducing (5.35) we find:

$$\cos^2 \varphi = \frac{\nu_{in}^2}{\nu_{in}^2 + (\Omega_i - \omega)^2}$$

By inserting this last result into (5.36) we derive:

$$Q_i = \frac{m_n}{m_n + m_i}\sigma_P E_0^2 \kappa$$

where

$$\kappa = \frac{\nu_{in}^2 + \Omega_i^2}{\nu_{in}^2 + (\Omega_i - \omega)^2}$$

We notice that for a given ion gyrofrequency, Ω_i, which is reasonably constant in the ionospheric E-region, κ has a maximum when $\omega = \Omega_i$. This maximum is given by:

$$\kappa_{\max} = 1 + \left(\frac{\Omega_i}{\nu_{in}} \right)^2 = 1 + k_i^2$$

$\kappa_{\max}$ is presented as a function of ν_{in}/Ω_i in Figure 5.8(a) showing a strong increase in $\kappa_{\max}$ for ratios of $\nu_{in}/\Omega_i < 1$. In the ionospheric E-region $\nu_{in} \approx \Omega_i$ at about 125 km or at the peak of Pedersen conductivity. At this peak $\kappa_{\max}$ is actually twice as large for an electric field oscillating at the ion gyrofrequency as κ in a steady state ($\omega = 0$). At larger heights in the E-region this factor will be even larger.

In Figure 5.8(b) we have presented κ as a function of ω/Ω_i for different values of ν_{in}. We notice that for all collision frequencies there is an enhancement of κ for oscillating frequencies ω between zero and $2\,\Omega_i$. The enhancement factor is larger the smaller the collision frequency ν_{in}. For $\nu_{in} = 0.1\,\Omega_i$, which corresponds to more than two scale heights above the peak of Pedersen conductivity, the enhancement factor is as high as 100. Even at $\nu_{in} = 0.5\,\Omega_i$, which corresponds to a distance less than a scale height above the peak in Pedersen conductivity, κ is as large as 4 at maximum. Therefore, when an electric field oscillates at a frequency less than $2\,\Omega_i$

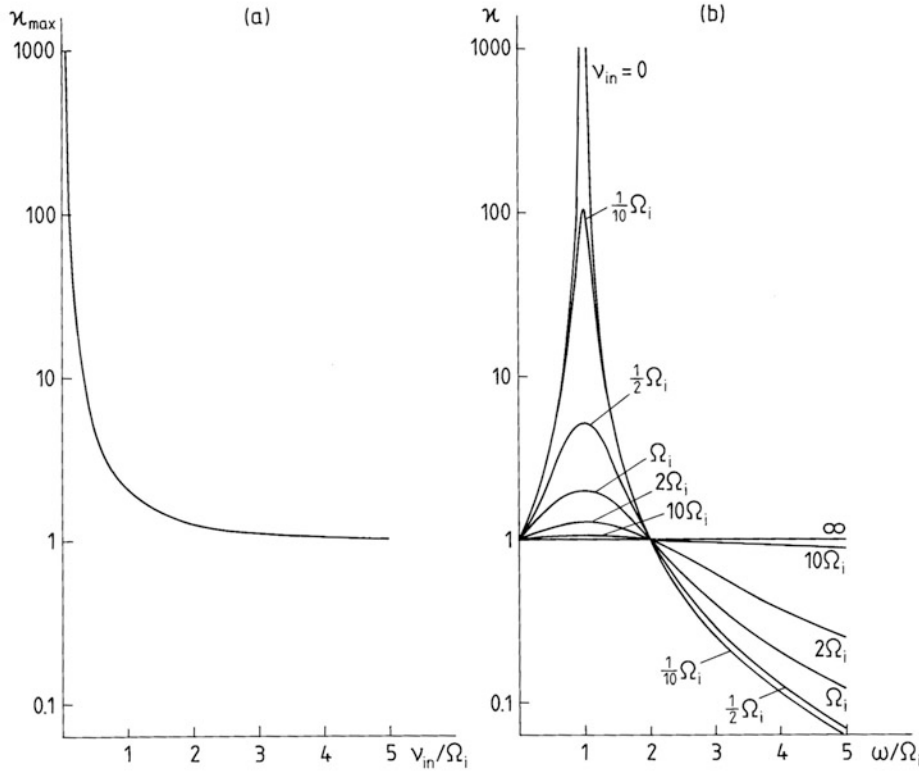


Figure 5.8. The enhancement factor κ of the rate of frictional heating due to the presence of an electric field oscillating at a frequency ω. (a) The maximum enhancement factor κ_{max} that occurs for $\omega = \Omega_i$ shown for different values of ν_{in}/Ω_i. $\nu_{in} = \Omega_i$ occurs at the peak of Pedersen conductivity in the auroral ionosphere. (b) The enhancement factor κ for different values of ν_{in} as a function of the relative oscillating frequency ω/Ω_i. κ is enhanced for $\omega \leq 2\,\Omega_i$ for all values of ν_{in}. (From Brekke and Kamide, 1996.)

the total rate of frictional heating is larger at all heights than the frictional heating rate in a steady state ($\omega = 0$).

From rocket observations of electric field fluctuations in connection with an auroral electrojet it has been demonstrated that considerable power is present in the ionospheric height range between 100 and 120 km in such fluctuating electric fields in the frequency range between 10 and 1,000 Hz, within which typical E-region ion gyrofrequencies are included. Such observations are illustrated in Figure 5.9.

These observations indicate that electric fields fluctuate considerably in the frequency range comparable with the typical E-region ion gyrofrequency.

It is thus reasonable to conclude that these fluctuating fields can represent a significant heat source, one that is comparable with the Joule heating source in the upper atmosphere at high latitudes.

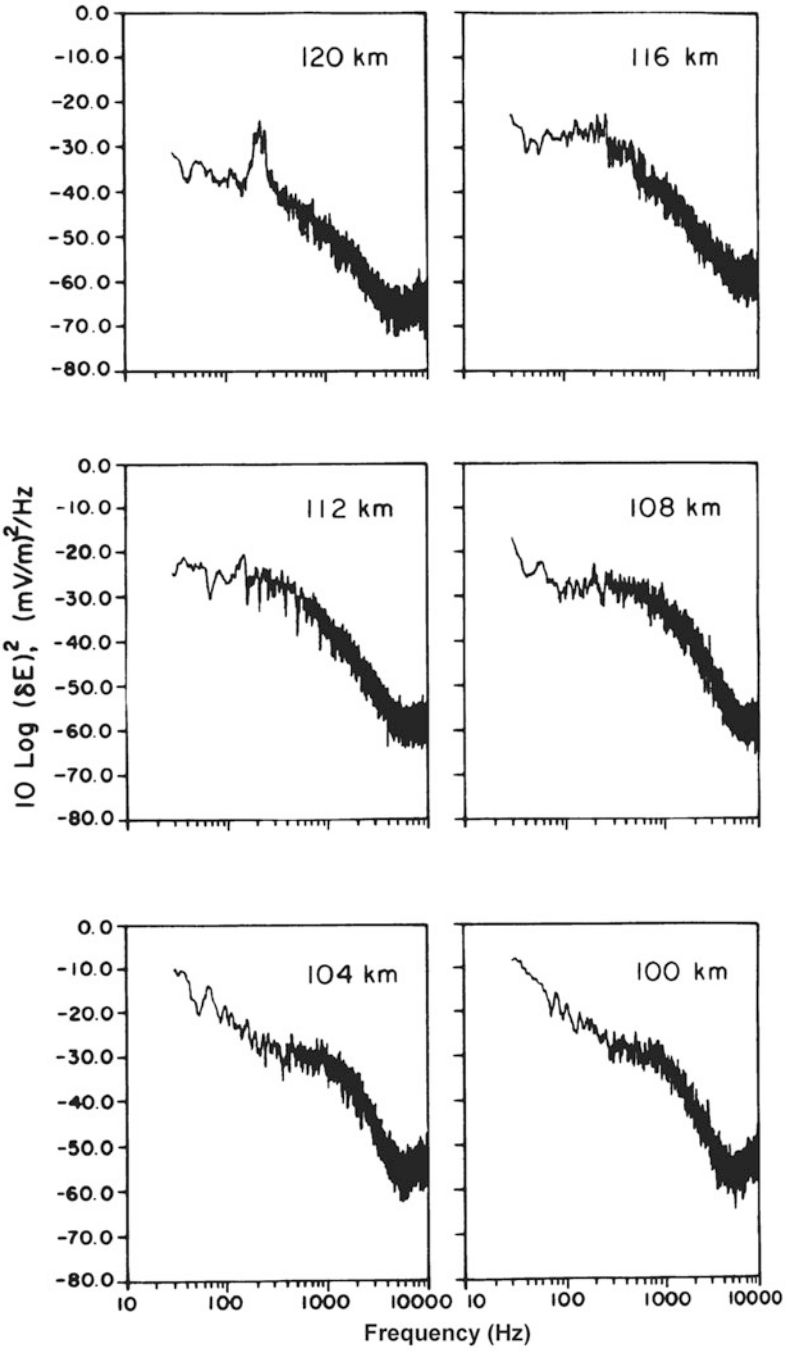


Figure 5.9. Six electric field spectra observed at different E-region heights from a rocket passing through the auroral electrojet. (From Kelley, 1989.)

5.7 CURRENTS DUE TO GRAVITY AND DIFFUSION

Let us return to (5.1) and assume steady-state solutions for ions and electrons. We will also assume that there is no electric field nor neutral wind. The ion velocity is then given by:

$$\mathbf{v}_i = \frac{1}{v_{in}}\mathbf{g} - \frac{kT_i}{m_i v_{in}}\frac{\nabla n_e}{n_e} + \frac{e}{m_i v_{in}}\mathbf{v}_i \times \mathbf{B} \tag{5.37a}$$

and the electron velocity by:

$$\mathbf{v}_e = \frac{1}{v_{en}}\mathbf{g} - \frac{kT_e}{m_e v_{en}}\frac{\nabla n_e}{n_e} - \frac{e}{m_e v_{en}}\mathbf{v}_e \times \mathbf{B} \tag{5.37b}$$

where we have assumed charge neutrality and $n_i = n_e$. By introducing

$$\mathbf{u}_G^i = \frac{1}{v_{in}}\mathbf{g} - \frac{kT_i}{m_i v_{in}}\frac{\nabla ne}{n_e} \tag{5.38a}$$

and

$$\mathbf{u}_G^e = \frac{1}{v_{en}}\mathbf{g} - \frac{kT_e}{m_e v_{en}}\frac{\nabla n_e}{n_e} \tag{5.38b}$$

into (5.37a) and (5.37b), respectively, we have

$$\mathbf{v}_i = \mathbf{u}_G^i + k_i\,\mathbf{v}_i \times \mathbf{B}/B \tag{5.39a}$$

$$\mathbf{v}_e = \mathbf{u}_G^e - k_e\,\mathbf{v}_e \times \mathbf{B}/B \tag{5.39b}$$

We can now solve (5.39a) and (5.39b) for $\mathbf{v}_i$ and $\mathbf{v}_e$, respectively. Then

$$\mathbf{v}_i = (\mathbf{u}_G^i)_\parallel + \frac{1+k_i}{1+k_i^2}(\mathbf{u}_G^i \times \mathbf{B})/B$$

$$\mathbf{v}_e = (\mathbf{u}_G^e)_\parallel + \frac{1-k_e}{1+k_e^2}(\mathbf{u}_G^e \times \mathbf{B})/B$$

where $(u_G^i)_\parallel$ and $(u_G^e)_\parallel$ are parallel to $\mathbf{B}$. The current due to gravity and diffusion can then be derived as:

$$\mathbf{j}_G = n_e \cdot e\left\{\left(\frac{1}{v_{in}} - \frac{1}{v_{en}}\right)\mathbf{g}_\parallel - k\left(\frac{T_i}{m_i v_{in}} - \frac{T_e}{m_e v_{en}}\right)\left(\frac{\nabla n_e}{n_e}\right)_\parallel\right\}$$
$$+ \left[\frac{1+k_i}{1+k_i^2}\left(\frac{1}{v_{in}}\mathbf{g} - k\frac{T_i}{m_i v_{in}}\right) - \frac{1-k_e}{1+k_e^2}\left(\frac{1}{v_{en}}\mathbf{g} - k\frac{T_e}{m_e v_{en}}\right)\right]\frac{\nabla n_e}{n_e} \times \mathbf{B}/B$$

Realizing now that $v_{en} \gg v_{in}$ and that $k_e \gg 1$ in the ionosphere and at the same time letting $\mathbf{g} \parallel \mathbf{B}$ at high latitudes, we get

$$\mathbf{j}_G = n_e \cdot e\left\{\frac{1}{v_{in}}\mathbf{g}_\parallel - k\left(\frac{T_i}{m_i \nu_{in}} - \frac{T_e}{m_e v_{en}}\right)\left(\frac{\nabla n_e}{n_e}\right)_\parallel - \frac{1+k_i}{1+k_i^2}k\frac{T_i}{m_i v_{in}}\frac{\nabla n_e}{n_e} \times \mathbf{B}/B\right\}$$
$$= (\mathbf{j}_G)_\parallel + (\mathbf{j}_G)_\perp \tag{5.40}$$

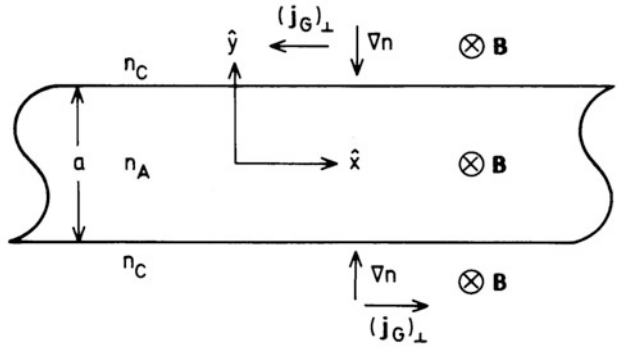


Figure 5.10. A narrow strip of the x–y plane of enhanced density n_A compared with the background density n_C. The strip has a finite width in the y-direction but a uniform and unlimited width in the x-direction. The magnetic field **B** is antiparallel to the z-axis. The gradients in n and current density vectors are indicated.

Along **B** the current will be strongly dependent on $\nabla n_e/n_e$ and the ratio between $T_i/m_i\nu_{in}$ and $T_e/m_e\nu_{en}$. Above the F-region maximum where $\nabla n_e/n_e < 0$ the diffusion current will be downward if $T_e \gg T_i$. Electrons will diffuse upward more rapidly than ions. The current perpendicular to **B** will be fully determined by ions.

Let us consider Figure 5.10 where we have indicated a narrow strip of enhanced density n_A in the x–y plane. The enhancement has a finite width, α, in the y-direction but an unlimited one in the x-direction. The magnetic field **B** is directed in the z-direction. The background density $n_C \ll n_A$. Since ∇n will be pointing towards the strip along the y-axis on both sides of the strip, the current density vector $(\mathbf{j}_G)_\perp$ due to diffusion will be parallel to the strip (as indicated in Figure 5.10), but in opposite directions on the two sides. Such irregularities in the plasma density can therefore create shear currents along irregularities.

5.8 EXERCISES

1. Prove that $\dot{Q}_i$ in (5.36) can be given by:

$$\frac{m_n}{m_n + m_i}\frac{ev_0^2}{\Omega_i}B \cdot n_e \cdot \nu_{in}$$

2. Let the ion–neutral collision frequency be given by

$$\nu_{in} = 2.6 \times 10^{-15} n_n (M_n')^{-1/2}$$

   (a) Derive θ_i in (5.16) at 90 km when $n_n = 6 \times 10^{19}$ m^{-3}, $M_n' = 28.8$, and the ion mass is equal to 30.5 a.m.u. (1 a.m.u. $= 1.660 \times 10^{-27}$ kg).

   (b) Derive θ_i at 180 km where $n_n = 1.3 \times 10^{16}$ m^{-3}, $M_n' = 24$, and the ions are O^+.

(c) Derive θ_i at 250 km where $n_n = 1.3 \times 10^{15}\ \mathrm{m}^{-3}$, $M'_n = 21$, and the ions are O^+.

3. Let the electron neutral collision frequency be given by:

$$\nu_{en} = 5.4 \times 10^{-10} n_n \cdot T_e^{1/2}$$

Derive θ_e in (5.18) at 90 km when $n_n = 6 \times 10^{19}\ \mathrm{m}^{-3}$ and $T_e = 170$ K. ($B = 5.7 \times 10^{-5}$ tesla.)

4. Derive (5.19).

5. Indicate the current vector at the four different panels in Figure 5.2.

6

Magnetic fluctuations in response to height-integrated currents

6.1 HEIGHT-INTEGRATED CURRENTS AND CONDUCTANCE

The ionospheric currents perpendicular to the **B**-field in the ionosphere have to be closed either totally in the ionosphere itself or by vertical currents along magnetic field lines. Since any possible current along the field lines must be of the same order of magnitude as the horizontal current, then:

$$\sigma_P(z) \cdot E_\perp = \sigma_\parallel(z) E_\parallel$$

Above, say, 200 km $\sigma_\parallel(z) > 10^5 \sigma_P(z)$ and, therefore, the parallel **E**-field must be very small compared with the **E**-field perpendicular to **B**,

$$E_\parallel \approx 10^{-5} E_\perp$$

The total electric field is for all practical purposes perpendicular to **B**.

For effects observable on the ground it is usually the height-integrated quantities of currents that count. If these can be derived, they can be compared with the geomagnetic field observations made underneath by ground-based magnetometers.

Since we have shown that the parallel electric field is very small and can be neglected, the height-integrated current density **J** can be derived by height-integrating (5.23) as:

$$\mathbf{J}' = \int_{h_1}^{h_2} \mathbf{j}' \, dh = \int_{h_1}^{h_2} \sigma_P(h) \mathbf{E}'_\perp(h) \, dh - \int_{h_1}^{h_2} \sigma_H(h) \mathbf{E}'_\perp(h) \times \mathbf{B} \frac{dh}{B} \tag{6.1}$$

where h_1 and h_2 are the lower and upper integration limits outside which the current is assumed to be zero. A reasonable choice is $h_1 \approx 90$ km and $h_2 \approx 280$ km. Note that the dimension of **J** is A/m in standard units.

A. Brekke, *Physics of the Upper Polar Atmosphere*, Springer Praxis Books,
DOI 10.1007/978-3-642-27401-5_6, © Springer-Verlag Berlin Heidelberg 2013

As the perpendicular component of the electric field refers to the magnetic field which makes an inclination angle I with the horizontal plane at the point of reference, $\mathbf{E}'_\perp$ and $\mathbf{B}$ should be properly decomposed into a Cartesian coordinate system where z is in the vertical direction.

At high latitudes where $\mathbf{B}$ is almost perpendicular to the Earth's surface or, equivalently, where I is close to 90°, we neglect the inclination of $\mathbf{B}$ for a moment and assume $\mathbf{B} = B\,\hat{\mathbf{z}}$ where $\hat{\mathbf{z}}$ points vertically downward, $\hat{\mathbf{x}}$ points positively northward, and $\hat{\mathbf{y}}$ completes the orthogonal Cartesian system when counted positive eastward. Then:

$$\begin{aligned}\mathbf{J}'_\perp &= \int_{h_1}^{h_2} \sigma_P(h)(E'_x\hat{\mathbf{x}} + E'_y\hat{\mathbf{y}})\,dh - \int_{h_1}^{h_2} \sigma_H(h)(E'_y\hat{\mathbf{x}} - E'_x\hat{\mathbf{y}})\,dh \\ &= \hat{\mathbf{x}}\int_{h_1}^{h_2}(\sigma_P(h)E'_x - \sigma_H(h)E'_y)\,dh + \hat{\mathbf{y}}\int_{h_1}^{h_2}(\sigma_P(h)E'_y + \sigma_H(h)E'_x)\,dh\end{aligned}$$

and

$$J'_x = \int_{h_1}^{h_2}(\sigma_P(h)E'_x - \sigma_H(h)E'_y)\,dh$$

$$J'_y = \int_{h_1}^{h_2}(\sigma_P(h)E'_y + \sigma_H(h)E'_x)\,dh$$

We now neglect the neutral wind and assume that $\mathbf{E}'_\perp = \mathbf{E}_\perp$. Furthermore, since the magnetic field lines act like perfect vertical conductors, $\mathbf{E}_\perp$ can be assumed to be independent of height. Therefore:

$$J_x = \Sigma_P E_x - \Sigma_H E_y \tag{6.2a}$$

$$J_y = \Sigma_P E_y + \Sigma_H E_x \tag{6.2b}$$

where the height-integrated conductivities or conductances are given by:

$$\Sigma_P = \int_{h_1}^{h_2} \sigma_P(h)\,dh \tag{6.3a}$$

$$\Sigma_H = \int_{h_1}^{h_2} \sigma_H(h)\,dh \tag{6.3b}$$

When introducing the expressions for height-integrated conductivities into (6.1) and assuming that $\mathbf{E}'_\perp(h) \geq \mathbf{E}_\perp$ is constant by altitude, we derive for the height-integrated current density:

$$\mathbf{J}_\perp = \Sigma_P\mathbf{E}_\perp - \Sigma_H\frac{\mathbf{E}_\perp \times \mathbf{B}}{B} \tag{6.4}$$

In Figure 6.1 some typical E-field measurements observed by EISCAT close to Tromsø are presented together with the corresponding conductances. We notice that the field is mainly directed in the northward direction before 22:00 UT (23:00 LT) and then turns in the southward direction. This turnover occurs at

earlier times for more strongly disturbed conditions, as indicated by an increasing A_p index. Magnetic midnight in Tromsø is 22:20 LT. The largest conductance enhancements, however, are seen around midnight when the field is mainly in the southward direction.

In general, when the field is increased, it is oriented more in the north–south direction than in the east–west direction and, therefore, the enhanced field is strongly meridional. When mapped into the equatorial plane, this will correspond mainly to radial fields.

6.2 MAGNETIC FIELD FLUCTUATIONS FROM AURORAL CURRENTS

If we now assume that the current density we have measured overhead is representative of the current density in an infinite current sheet, then the magnetic fluctuations we should expect to observe on the ground can be derived.

A simple meridional cross-section in the x–y plane is shown in Figure 6.2. Independent of the position on the ground, the contribution to the magnetic field from this current would be according to Ampère's law

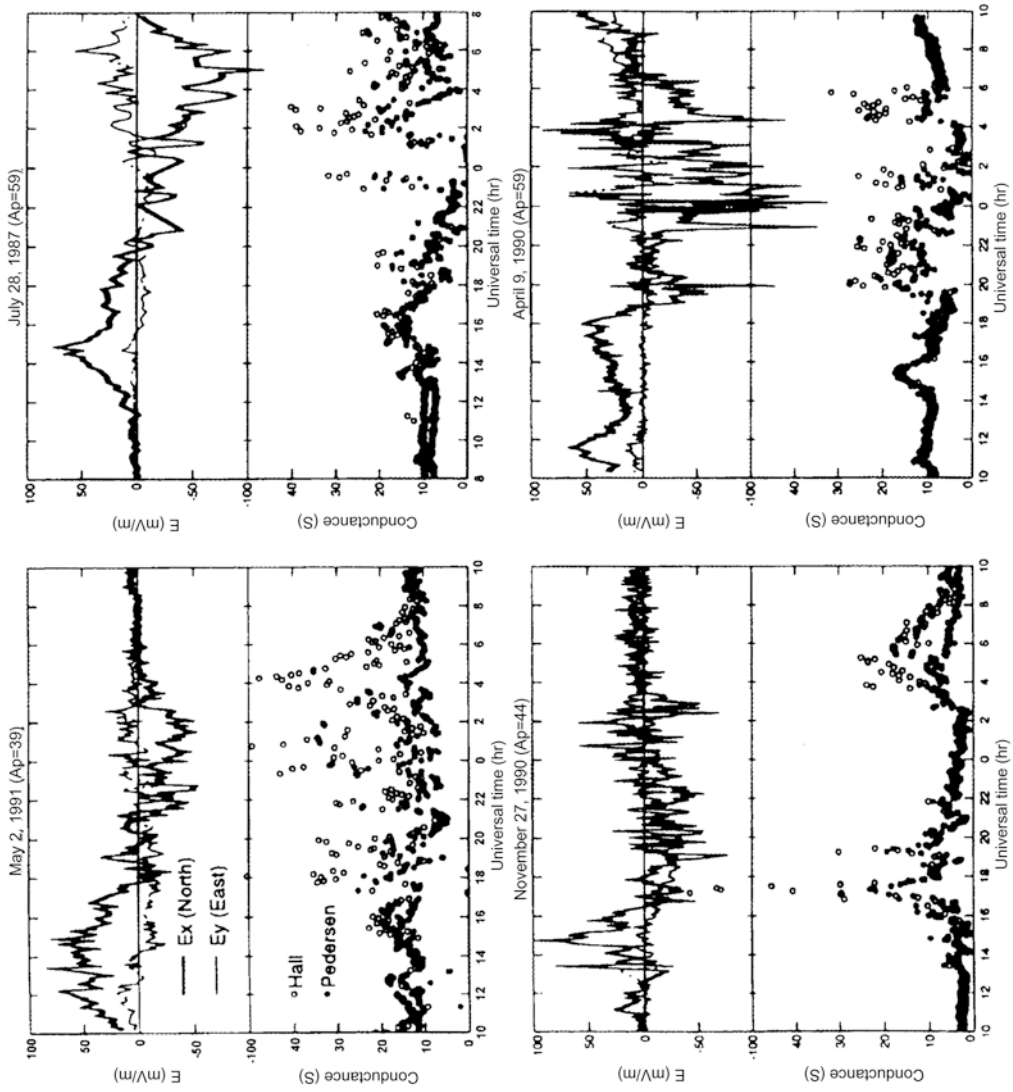
$$\Delta B_x = \frac{\mu_0 J_y}{2}$$

If, however, the current had an arbitrary direction, then:

$$\Delta \mathbf{B} = \Delta B_x \hat{\mathbf{x}} + \Delta B_y \hat{\mathbf{y}} = \frac{\mu_0}{2}(J_y \hat{\mathbf{x}} - J_x \hat{\mathbf{y}})$$

For a conducting Earth the factor 1/2 must be substituted by 3/4. By thus comparing the height-integrated overhead current with the simultaneously observed magnetic field fluctuations one could get a feeling for how good the assumptions were. This is done in the bottom panel of Figure 6.3 where currents derived and integrated overhead Chatanika are compared with the ΔH and ΔD observations at College, Alaska, about 40 km apart. Apart from an arbitrary scaling factor, it is seen that when considering the rather poor time resolution in the current measurements ($\sim$30 min) the overall agreement is good as far as the ΔH- and J_y-components are concerned. The agreement between the ΔD- and J_x-components is, however, less satisfying, probably illustrating that the latitudinal current is much more homogeneous than the meridional currents in the auroral region. The fact that the D-component is also influenced by field-aligned currents which are not included in the calculations of overhead currents in Figure 6.3 may also account for some of this discrepancy.

The two terms in the expression for J_x (6.2a)—the northward component—often cancel each other. The relative uncertainty in subtracting two very similar terms is always greater than if they were added. This may also contribute to some of the discrepancies between J_x and ΔD as observed in Figure 6.3.


July 28, 1987 (Ap=59)
April 9, 1990 (Ap=59)
May 2, 1991 (Ap=39)
November 27, 1990 (Ap=44)
E (mV/m)
Conductance (S)
Ex (North)
Ey (East)
Hall
Pedersen
Universal time (hr)

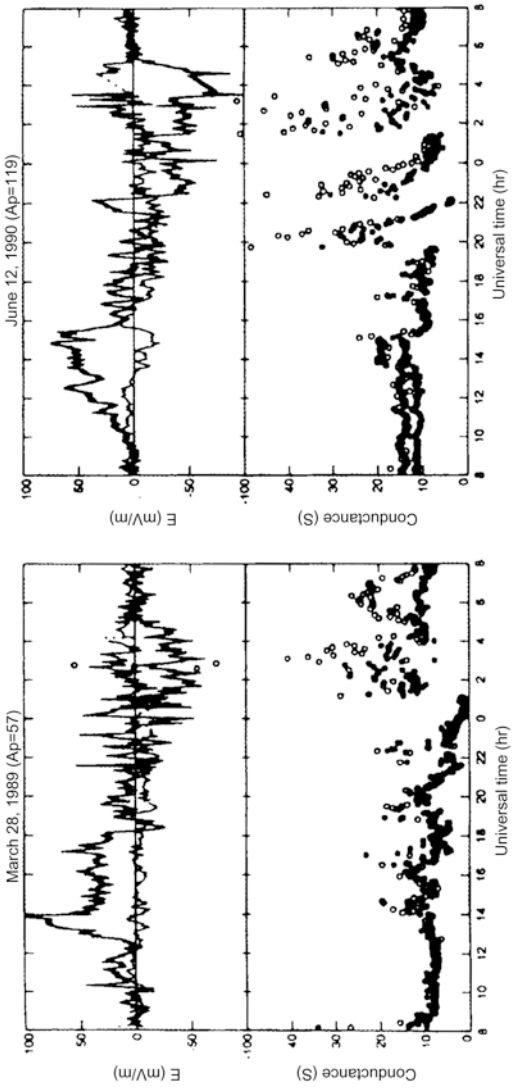


Figure 6.1. Six panels showing time variations in the north and east components of the electric field in the upper half and the Hall and Pedersen conductances in the lower half. The panels are arranged from the upper left corner and downward according to days of increasing A_p index. This index is identified together with the appropriate dates of the EISCAT experiment at the top of each panel. (From Brekke *et al.*, 1995.)

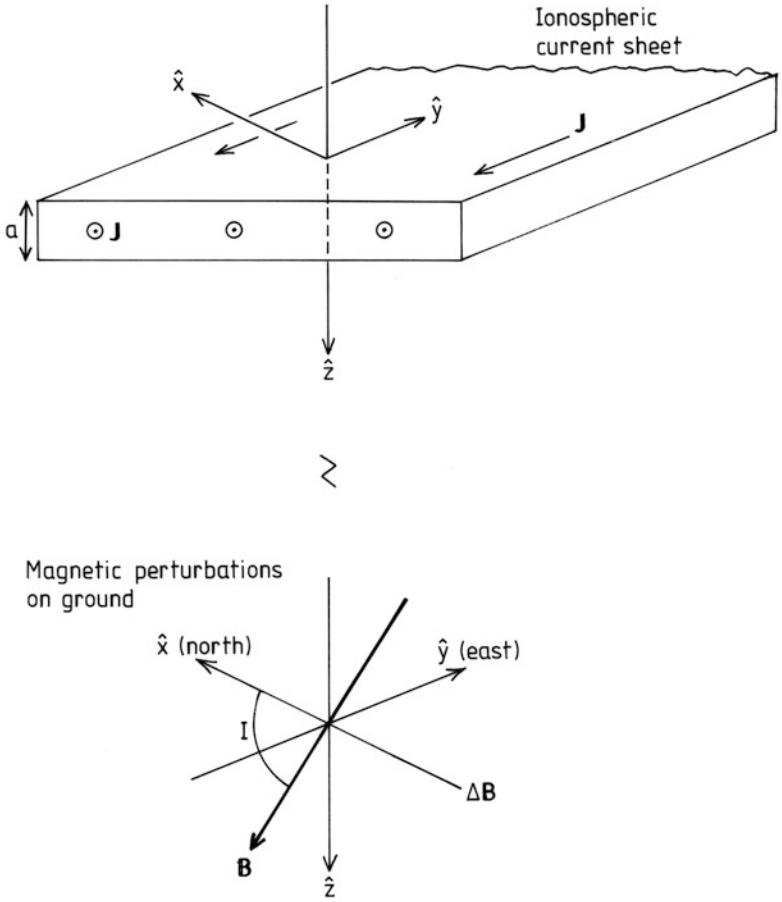


Figure 6.2. (Above) Simplified sketch to illustrate an infinite sheet, carrying a current density **J** (A/m) flowing in the negative y-direction within a thin vertical layer of width a. (Below) An illustration of the magnetic field variation $\Delta\mathbf{B}$ observed on the ground. It is perpendicular to, and turned anticlockwise with respect to, the current direction.

Since current in the north–south direction is much less than current in the east–west direction, we will assume as an illustration that

$$J_x = 0 = \Sigma_P E_x - \Sigma_H E_y = 0$$

where we have neglected the contribution to $\mathbf{E}'$ from the neutral wind. We then get the relation:

$$\frac{E_x}{E_y} = \frac{\Sigma_H}{\Sigma_P}$$

The ratio between the conductances will then be equal to the ratio between the electric field components. Figure 6.3 shows an example where this situation is

nearly fulfilled since

$$\frac{\Sigma_H}{\Sigma_P} \approx 3 \qquad \text{and} \qquad \frac{E_x}{E_y} \approx 3$$

The electric field in the meridional direction can therefore be said to be a polarization field. The east–west current (6.2b) now becomes:

$$J_y = \Sigma_P E_y + \Sigma_H \left(\frac{\Sigma_H}{\Sigma_P} E_y \right) = \left(\Sigma_P + \frac{\Sigma_H^2}{\Sigma_P} \right) E_y$$

and it is proportional to the east–west electric field component. The proportionality coefficient is an enhanced Pedersen conductance:

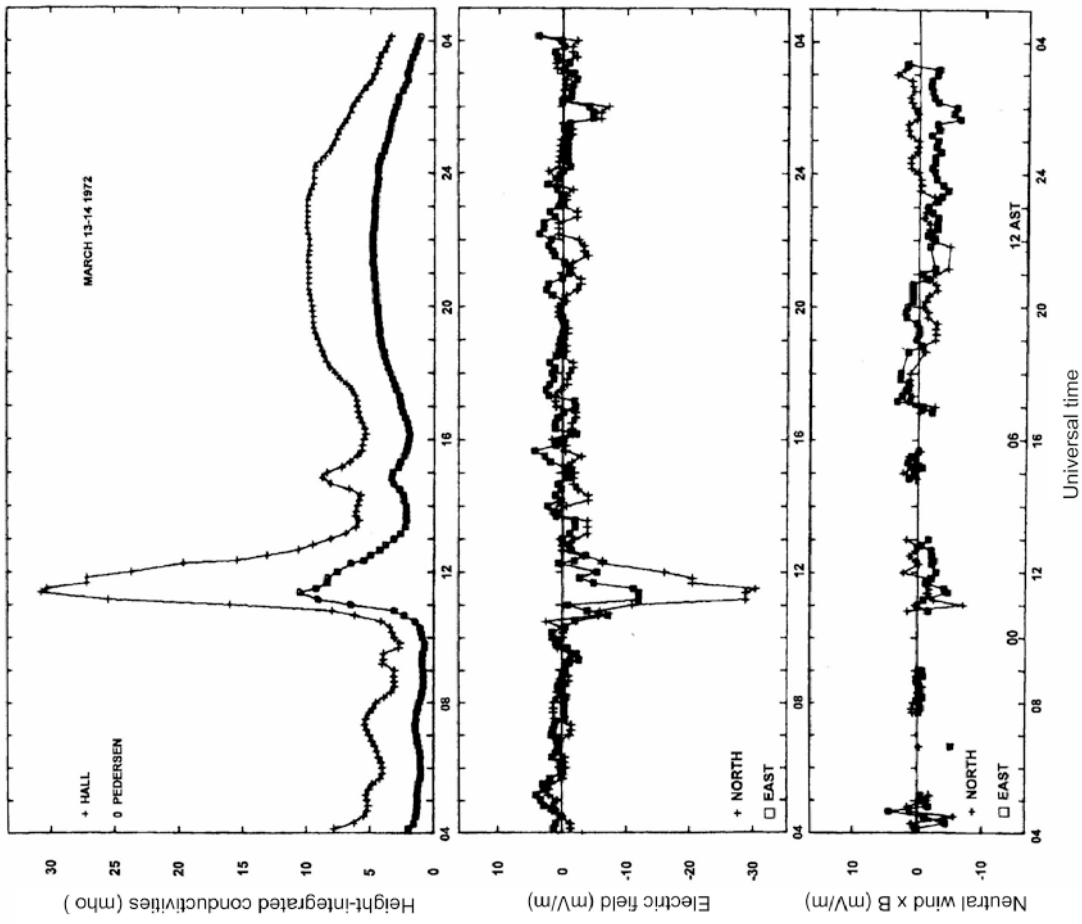
$$\Sigma_A = \Sigma_P + \frac{\Sigma_H^2}{\Sigma_P} = \Sigma_P \left(1 + \left(\frac{\Sigma_H}{\Sigma_P} \right)^2 \right) \tag{6.5}$$

which can be called the "auroral electrojet conductance" or the "auroral Cowling conductance".

The good agreement observed between the height-integrated current, J_y, derived from radar observations at Chatanika and the magnetic field fluctuations, ΔH, as observed by a nearby station at College, Alaska (as shown in Figure 6.3), underlines the significance of the electrojet in the east–west direction. The J_x-component is very small, in agreement with the polarization condition given above. In the lower part of the upper panel of Figure 6.3 estimates of time variations of the $\mathbf{u} \times \mathbf{B}$ term contribution to $\mathbf{E}'$ averaged over the height region of the current are presented. We notice from Figure 6.3 that the $\mathbf{u} \times \mathbf{B}$ term is insignificant during this event of the electrojet.

In early August 1972 there was a period of strong solar disturbances when large magnetic fluctuations indicating the presence of ionospheric currents were observed worldwide. During this period the incoherent scatter radar at Chatanika, Alaska, was operating and observations of electron densities and ion drifts were stored on data tapes for almost 10 consecutive days.

On August 4 at 20:55 UT a sudden enhancement was observed in the magnetic field and at the same time the electron density was found to increase strongly, enhancing Σ_H by a factor of 2 from 30 mhos to 60 mhos (as demonstrated in the lower panel of Figure 6.4). At the same time Σ_P was only enhanced from 12 to 18 mhos. The ion drift velocities used to derive the electric field by assuming $\mathbf{E} \times \mathbf{B}$ drifts went through a strong oscillation between +15 and −20 mV/m. This event was then followed by an oscillating electric field (−20 mV/m) that lasted for about 20 minutes (as demonstrated in the upper panel of Figure 6.4). During this time the Σ_H and Σ_P were very little affected. From these measurements height-integrated currents were derived and compared with the corresponding magnetic fluctuations at College, Alaska (as presented in Figure 6.5). In particular, there is a good correspondence between fluctuations in the H-component and the eastward ionospheric current demonstrating a close relationship between pulsations in the Earth's magnetic field and the oscillating ionospheric electric field.


MARCH 13-14 1972
+ HALL
0 PEDERSEN
Height-integrated conductivities (mho)
+ NORTH
□ EAST
Electric field (mV/m)
+ NORTH
□ EAST
12 AST
Neutral wind x B (mV/m)
Universal time

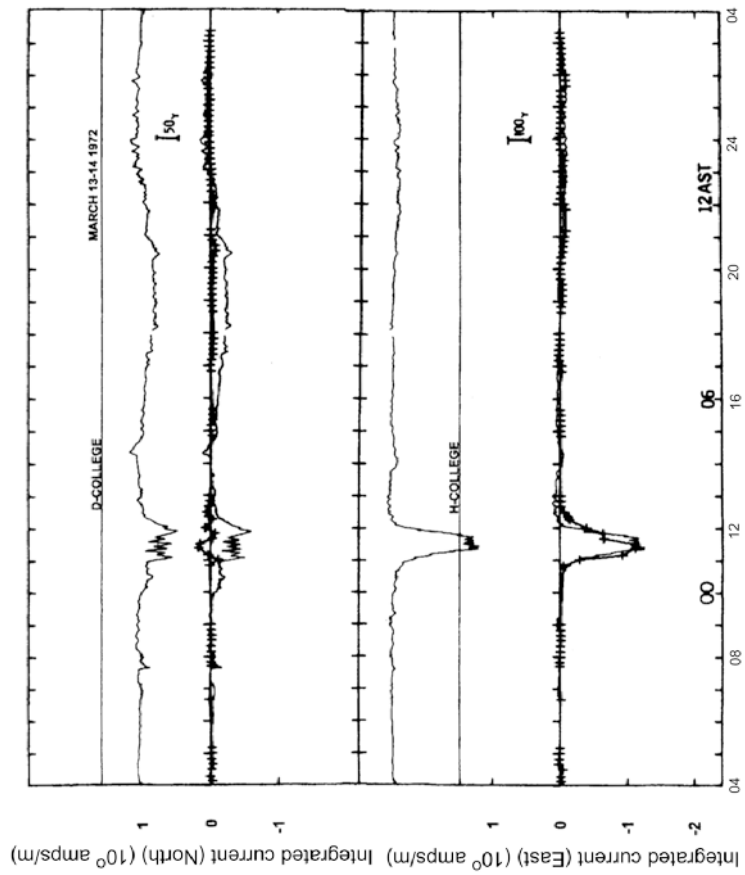


Figure 6.3. Daily variation in the Pedersen and Hall conductances, northward and eastward components of the ionospheric electric field, and the northward and eastward component of the neutral dynamo electric field observed in the auroral zone at Chatanika for the quiet equinox day of March 13–14, 1972. (Lower half) Ionospheric currents deduced from the data in the panels above compared with the magnetic field fluctuations observed by a nearby magnetometer. The northward current is compared with the *D*-component while the eastward component is overlaid on the *H*-component of the magnetic field, respectively. (From Brekke *et al.*, 1974.)

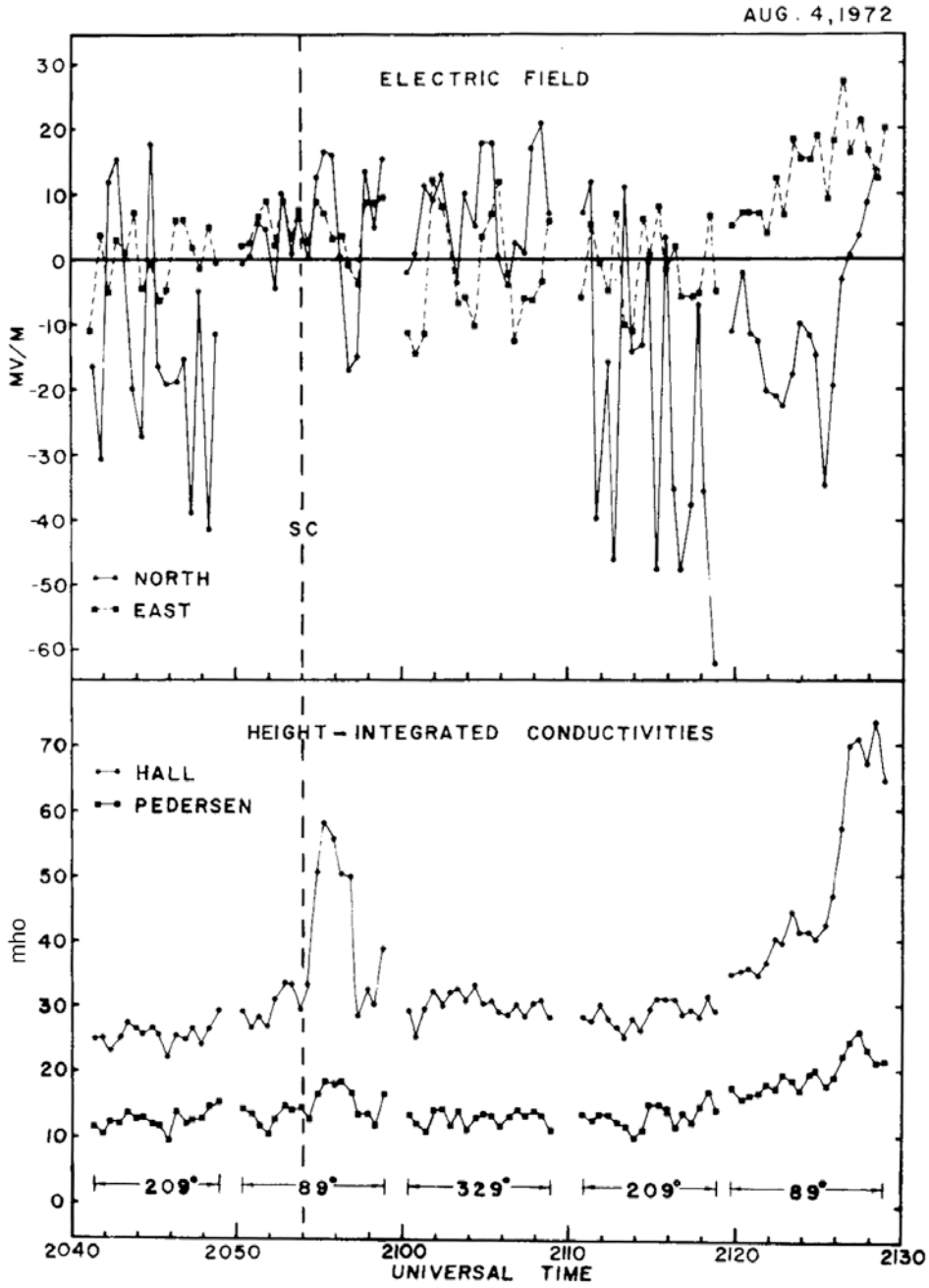


Figure 6.4. Observations by the Chatanika incoherent scatter radar, Alaska, of the electric field components E_x and E_y and the height-integrated conductances Σ_H and Σ_P for 2 hours on August 4, 1972, during a magnetic sudden commencement (SC) and followed by magnetic pulsation. (From Doupnik *et al.*, 1977.)

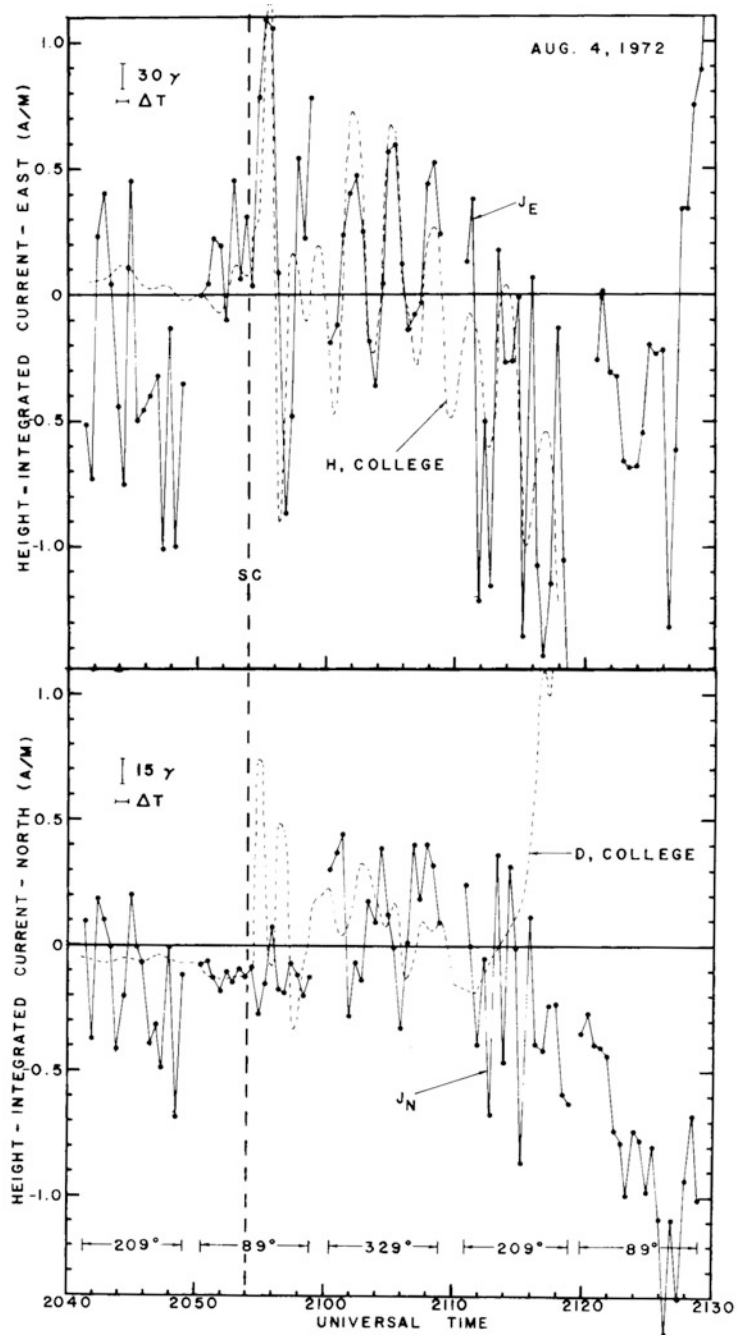


Figure 6.5. Derivations of the height-integrated currents J_x and J_y compared with the magnetic fluctuations in ΔH and ΔD observed at College, Alaska, near Chatanika for the same period as shown in Figure 6.4. (From Doupnik *et al.*, 1977.)

6.3 EQUIVALENT CURRENT SYSTEMS

For more than a century magnetic field observations from the ground at high latitudes have been used to deduce the so-called equivalent current system. Birkeland was the first to do this to study what he called the "elementary storm" in relation to auroral disturbances which today are called "auroral substorms". Figure 6.6 shows two examples of Birkeland's illustrations in which the equivalent current vectors are drawn as arrows at right angles and turned anticlockwise with

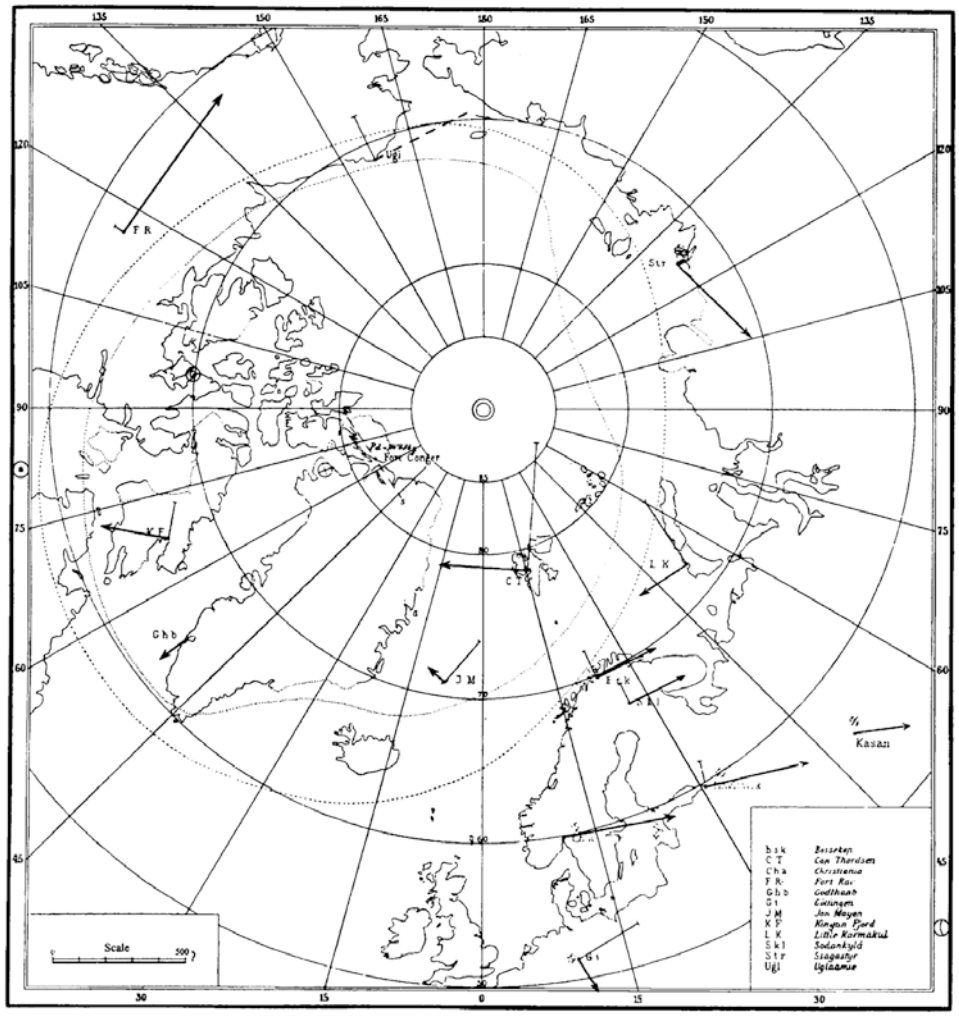


Figure 6.6. One of Birkeland's original drawings showing estimates of the equivalent overhead current vectors as derived from the magnetic fluctuation vectors observed on the ground at different stations inside and outside the auroral zone. (From Birkeland, 1913.)

respect to the magnetic field fluctuations observed from different stations in the northern hemisphere.

Harang expanded on Birkeland's method and used high-latitude magnetic records to draw isointensity curves of the magnetic field fluctuations around the polar cap region. He found that there is a demarcation line between areas of positive and negative ΔH fluctuations on the nightside (as shown in Figure 6.7). This demarcation line has since been called the "Harang discontinuity". The *Harang discontinuity* changes position according to the general activity and is the result of particle precipitation and divergences in the ionospheric electric field (as noticed in relation to Figure 6.1).

Another classical illustration that has emerged thanks to this method is due to the work of Silsbee and Vestine (1942) and is shown in Figure 6.8. A large number of magnetic observatories in the northern hemisphere formed the basis for this study of the global current system related to high-latitude magnetic bays (or "substorms" as they are usually called). The horizontal magnetic field fluctuation vector ($\Delta\mathbf{B}$) observed at each station is depicted, and contour lines of constant current densities are drawn perpendicular to these vectors. Two large current cells appear, one rotating clockwise in the dayside half and the other anticlockwise in the nightside half, when observed from above the North Pole. These cells merge together in an intensified westward current, the *auroral electrojet*, above 60° of latitude on the nightside. This electrojet has a maximum close to 02:00 LT. The current across the polar cap is fairly uniform and directed from evening towards morning close to the 20:00 to 08:00 local meridians.

In later years a large number of magnetic observatories have been installed in the Arctic to investigate high-latitude current systems in more detail, especially those related to auroral substorms. Meridional chains of a multitude of magnetometers have proven extremely useful in this context. Figure 6.9 demonstrates examples of equivalent current patterns derived from the Greenland chain of magnetometers. The figure is divided up into nine separate panels depending on the direction of the interplanetary magnetic field (IMF), which has a clear and observable effect on details in the current pattern system. We recognize for negative B_z a strong similarity with the pattern shown in Figure 6.8. The nightside current cell generally penetrates further toward noon than in Figure 6.8, but the uniform cross-polar cap current from about 20:00 to 08:00 LT is consistent.

When B_z turns more positive, these current cells have a tendency to deform into a more irregular shape and even break up into several cells. This latter effect appears to be more conspicuous when B_y is positive.

Based on such data, it is possible to derive the electric potential pattern, ionospheric currents, and field-aligned currents by implementing a model of ionospheric conductivity.

By height-integrating the ionospheric current density (5.23) and excluding the neutral wind, we have

$$\mathbf{J} = \Sigma_P \mathbf{E}_\perp - \Sigma_H \mathbf{E}_\perp \times \hat{\mathbf{B}} + \Sigma_\parallel (\mathbf{E} \cdot \hat{\mathbf{B}}) \cdot \hat{\mathbf{B}} = \mathbf{J}_\perp + \mathbf{J}_\parallel \tag{6.6}$$

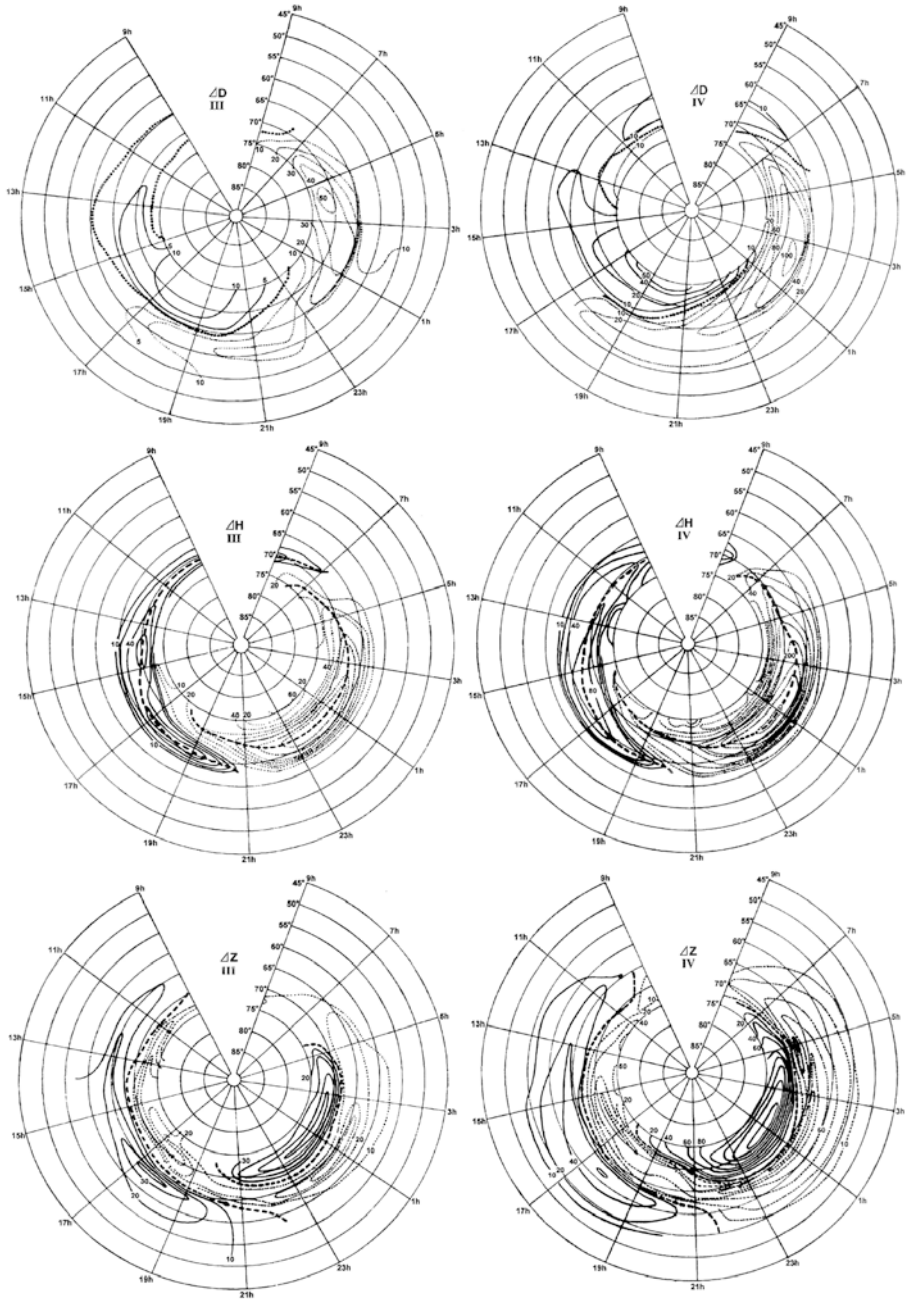


Figure 6.7. Examples of Harang's original drawings showing the contours of constant fluctuations of the magnetic field components (D, H, and Z) for two categories of disturbance, as drawn on the basis of observations at several stations inside and within the auroral zone. The time indicated is the local time at the 120° geomagnetic meridian. (From Harang, 1946.)

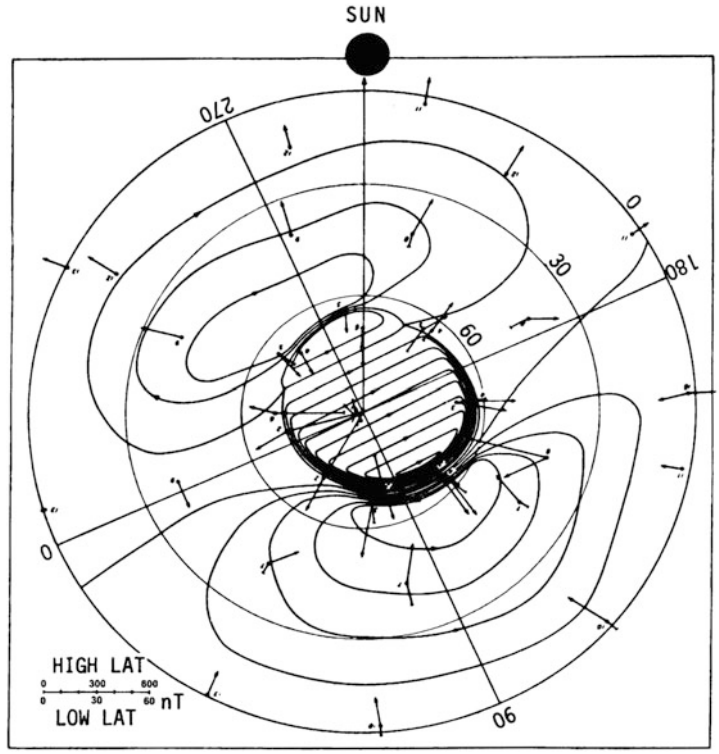


Figure 6.8. The distribution of magnetic disturbance vectors and equivalent ionospheric current lines, which corresponds to the average for many magnetic substorms. (From Silsbee and Vestine, 1942.)

where $\mathbf{J}_\perp$ and $\mathbf{J}_\parallel$ are the height-integrated current density vectors perpendicular to and along the magnetic field, respectively, and $\hat{\mathbf{B}} = \mathbf{B}/B$. Since the divergence of this current must be zero we obtain:

$$\nabla \cdot \mathbf{J} = \nabla_\perp \cdot \mathbf{J}_\perp + \nabla_\parallel \cdot \mathbf{J}_\parallel = \nabla_\perp \cdot \mathbf{J}_\perp + \frac{\partial}{\partial z} J_z = 0 \tag{6.7}$$

where we have split the gradient into a perpendicular and a parallel component. The positive z-axis is counted along the B-field, downward in the northern hemisphere.

Since the current density j_z is given by

$$j_z = \frac{\partial J_z}{\partial z}$$

we have

$$j_z = -\nabla_\perp \cdot \mathbf{J}_\perp \tag{6.8}$$

and the vertical current along the **B**-field can be found from the gradient in the

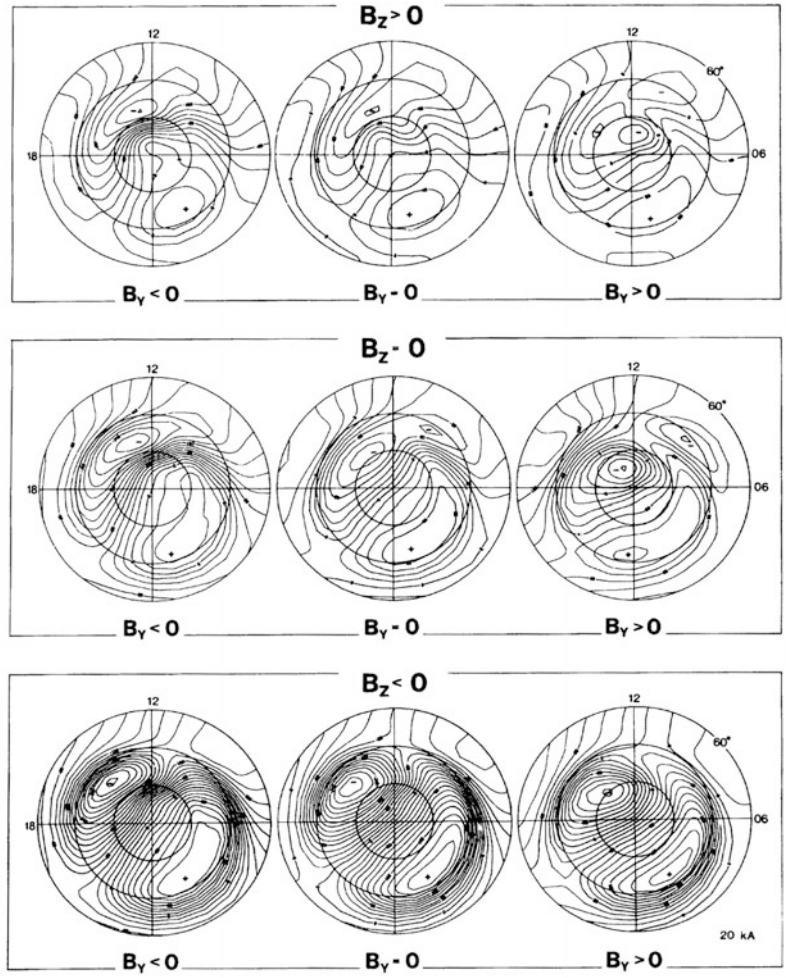


Figure 6.9. Contours showing horizontal equivalent currents for different orientations of B_z and B_y of the IMF. A twin vortex system appears to be present above 70° latitude independent of the IMF. The intensity increases and the orientation of the vortices twists about 3 hours from the noon–midnight meridian as B_z grows large negative. An additional vortex cell has a tendency to occur on the morning side when B_y is positive. (From Friis-Christensen *et al.*, 1985.)

perpendicular current. For $\nabla_\perp \cdot \mathbf{J}_\perp$ we have:

$$\nabla_\perp \cdot \mathbf{J}_\perp = \frac{\partial \Sigma_P}{\partial x} E_x + \Sigma_P \frac{\partial E_x}{\partial x} - \frac{\partial \Sigma_H}{\partial x} E_y - \Sigma_H \frac{\partial E_y}{\partial x} + \frac{\partial \Sigma_P}{\partial y} E_y + \Sigma_P \frac{\partial E_y}{\partial y} + \frac{\partial \Sigma_H}{\partial y} E_x + \Sigma_H \frac{\partial E_x}{\partial y}$$

Since $\nabla \times \mathbf{E} = 0$ then

$$\frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x} = 0 \tag{6.9}$$

and the Hall current will not contribute to the divergence of **J** unless Σ_H is anisotropic. Inserting for $\nabla_\perp \mathbf{J}_\perp$ into (6.8) we find

$$j_z = -\left(\frac{\partial \Sigma_P}{\partial x} + \frac{\partial \Sigma_H}{\partial y}\right) E_x - \left(\frac{\partial \Sigma_P}{\partial y} - \frac{\partial \Sigma_H}{\partial x}\right) E_y - \Sigma_P \left(\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y}\right) \tag{6.10}$$

Even with a constant E-field in the plane perpendicular to the magnetic field we will have a divergence in **J** or a vertical current as long as the conductances are anisotropic. In the neighborhood of auroral arcs we can expect the conductances to vary quite dramatically in the plane perpendicular to **B**.

Figure 6.10 shows a simplified model of the Harang discontinuity, which serves as a boundary towards which the electric field converges, poleward ($\mathbf{E}_1$) towards the boundary on the equatorial side and equatorward ($\mathbf{E}_2$) on the poleward side, respectively. The x-dimension (l) is assumed so small compared with

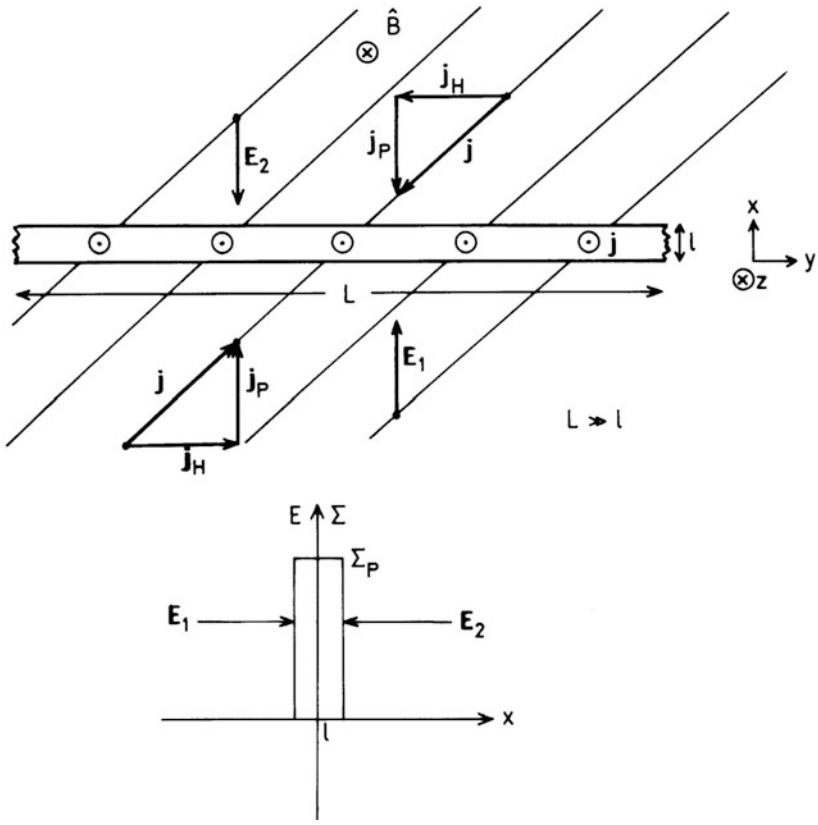


Figure 6.10. Schematic illustration of a simplified model of the Harang discontinuity to show the direction of current flow vectors and the E-field converging toward the boundary represented by upward field-aligned currents. Enhancements to Σ_H and Σ_P are introduced across the boundary.

the y-dimension (L) of the demarcation region that $\partial/\partial y$ can be neglected with respect to $\partial/\partial x$. These electric fields will drive Pedersen currents toward the discontinuity on both sides and Hall currents along the boundary in opposite directions on the two sides.

When neglecting any east–west component (i.e. y-component) of the electric field, the parallel current can be expressed by (6.10)

$$j_{\parallel} = j_z = -\frac{\partial \Sigma_P}{\partial x} E_x - \Sigma_P \frac{\partial E_x}{\partial x}$$

The Pedersen conductivity, which is assumed to have a maximum at the discontinuity as indicated in Figure 6.10, will contribute to an upward current on both sides of the discontinuity, and these currents will be directly related to the field-aligned current due to the precipitating electrons that mark the boundary of enhanced Pedersen conductance.

The currents perpendicular to **B** on both sides of the Harang discontinuity are also shown with an eastward current on the equatorial side and a westward current on the polar side, as also indicated by observations.

If, however, the current perpendicular to **B** around the Harang discontinuity is divergence free, it simply means that the currents there are part of the current system that closes in the ionosphere and there is nothing special about the Harang discontinuity. It just marks the area where two current cells diverge. The crucial point is to establish whether there are gradients in conductivity across the Harang discontinuity.

6.4 EQUIVALENT CURRENTS AT DIFFERENT LATITUDES

Let us introduce a Cartesian coordinate system $(\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3)$ where

$$\mathbf{E}_{\perp} = E_1 \hat{\mathbf{e}}_1$$
$$-\mathbf{E}_{\perp} \times \hat{\mathbf{B}} = E_2 \hat{\mathbf{e}}_2$$
$$\mathbf{E}_{\parallel} = E_3 \hat{\mathbf{e}}_3$$

We then have from (6.6)

$$\begin{pmatrix} J_1 \\ J_2 \\ J_3 \end{pmatrix} = \begin{pmatrix} \Sigma_P & 0 & 0 \\ 0 & +\Sigma_H & 0 \\ 0 & 0 & \Sigma_{\parallel} \end{pmatrix} \begin{pmatrix} E_1 \\ E_2 \\ E_3 \end{pmatrix} \tag{6.11}$$

The conductance tensor has a simple form as long as the electric field is constant with altitude over the height region of integration. In this context the electric field should contain the static electric field **E** and the neutral wind component $\mathbf{u} \times \mathbf{B}$. Since this field changes rather irregularly due to the neutral wind, a coordinate system with reference to the direction of the total electric field is not really practicable to use from an observational standpoint at the Earth's surface. Let us therefore assume an Earth-fixed Cartesian coordinate system (x, y, z) where x

points magnetic northward, y magnetic eastward, and z downward towards the Earth's center. In this coordinate system the magnetic field is given by

$$\mathbf{B} = B(\cos I\,\hat{\mathbf{x}} + \sin I\,\hat{\mathbf{z}})$$

where I is the inclination angle. The unit vector along $\mathbf{B}$ is:

$$\hat{\mathbf{B}} = \cos I\,\hat{\mathbf{x}} + \sin I\,\hat{\mathbf{z}}$$

For this coordinate system it is assumed that the magnetic dipole axis is antiparallel to the Earth's rotation axis and that the magnetic field is symmetric around this axis. The height-integrated current is given by (6.6)

$$\mathbf{J} = \Sigma_P \mathbf{E}_\perp - \Sigma_H \mathbf{E}_\perp \times \hat{\mathbf{B}} + \Sigma_\parallel (\mathbf{E}\cdot\hat{\mathbf{B}}) \cdot \hat{\mathbf{B}}$$

where the electric field is assumed constant with height and may include a constant $\mathbf{u} \times \mathbf{B}$ term. In our coordinate system we then have:

$$\mathbf{E} = E_x\hat{\mathbf{x}} + E_y\hat{\mathbf{y}} + E_z\hat{\mathbf{z}}$$

$$\mathbf{E}_\parallel = (\mathbf{E}\cdot\hat{\mathbf{B}})\hat{\mathbf{B}} = (E_x \cos^2 I + E_z \sin I \cos I)\hat{\mathbf{x}} + (E_x \cos I \sin I + E_z \sin^2 I)\hat{\mathbf{z}}$$

$$\mathbf{E} \times \hat{\mathbf{B}} = E_y \sin I\,\hat{\mathbf{x}} + (E_z \cos I - E_x \sin I)\hat{\mathbf{y}} - E_y \cos I\,\hat{\mathbf{z}}$$

$$\mathbf{E}_\perp = \mathbf{E} - \mathbf{E}_\parallel = (E_x \sin^2 I - E_z \sin I \cos I)\hat{\mathbf{x}} + E_y\hat{\mathbf{y}} + (E_z \cos^2 I - E_x \sin I \cos I)\hat{\mathbf{z}}$$

The current density can then be expressed in component form as:

$$J_x = (\Sigma_P \sin^2 I + \Sigma_\parallel \cos^2 I)E_x - \Sigma_H \sin I\,E_y + (\Sigma_\parallel - \Sigma_P)\sin I \cos I\,E_z \quad (6.12a)$$

$$J_y = \Sigma_H \sin I\,E_x + \Sigma_P E_y - \Sigma_H \cos I\,E_z \quad (6.12b)$$

$$J_z = (\Sigma_\parallel - \Sigma_P)\sin I \cos I\,E_x + \Sigma_H \cos I\,E_y + (\Sigma_P \cos^2 I + \Sigma_\parallel \sin^2 I)E_z \quad (6.12c)$$

or in tensor form:

$$\begin{pmatrix} J_x \\ J_y \\ J_z \end{pmatrix} = \begin{pmatrix} \Sigma_P \sin^2 I + \Sigma_\parallel \cos^2 I & -\Sigma_H \sin I & (\Sigma_\parallel - \Sigma_P)\sin I \cos I \\ \Sigma_H \sin I & \Sigma_P & -\Sigma_H \cos I \\ (\Sigma_\parallel - \Sigma_P)\sin I \cos I & \Sigma_H \exp I & \Sigma_P \cos^2 I + \Sigma_\parallel \sin^2 I \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix}$$

At high latitudes where $I \approx 90^\circ$ and $\hat{\mathbf{B}} \approx \hat{\mathbf{z}}$ we have

$$\begin{pmatrix} J_x \\ J_y \\ J_z \end{pmatrix} = \begin{pmatrix} \Sigma_P & -\Sigma_H & 0 \\ \Sigma_H & \Sigma_P & 0 \\ 0 & 0 & \Sigma_\parallel \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix}$$

and the conductance tensor is fairly simple. In particular, we find for $J_x = 0$

$$E_x = \frac{\Sigma_H}{\Sigma_P} E_y$$

and the auroral electrojet is given by

$$J_y = \left(\Sigma_P + \frac{\Sigma_H^2}{\Sigma_P}\right) E_y = \Sigma_A E_y$$

where Σ_A is the auroral electrojet conductance, as already defined in (6.5). On the other hand, if we assume that

$$E_\parallel = \mathbf{E} \cdot \hat{\mathbf{B}} = (E_x \hat{x} + E_y \hat{y} + E_z \hat{z}) \cdot (\cos I \, \hat{x} + \sin I \, \hat{z}) = 0$$

then

$$E_z = -\frac{\cos I}{\sin I} E_x$$

and by inserting in (6.12) we get

$$J_x = \Sigma_P E_x - \Sigma_H \sin I \, E_y \tag{6.13a}$$

$$J_y = \Sigma_P E_y + \frac{\Sigma_H}{\sin I} E_x \tag{6.13b}$$

$$J_z = -\Sigma_P \frac{\cos I}{\sin I} E_x + \Sigma_H \cos I \, E_y \tag{6.13c}$$

At high latitudes where $I \approx 90^\circ$, J_z becomes very small and the vertical current vanishes when $E_\parallel = 0$. When comparing (6.13a) and (6.13b) with (6.2a) and (6.2b), respectively, we notice that the $\mathbf{E}_\parallel = 0$ constraint reduces the Hall current in the x (north) direction and enhances the Hall current in the y (east) direction. The Pedersen current in the horizontal plane is not influenced by this effect. Equations (6.12) are, however, the expressions that should be used for proper comparison with the X- and Y-components measured by magnetometers.

In equatorial regions where the inclination angle is zero ($I = 0$) and $\hat{\mathbf{B}} = \hat{\mathbf{x}}$, we have:

$$\begin{pmatrix} J_x \\ J_y \\ J_z \end{pmatrix} = \begin{pmatrix} \Sigma_\parallel & 0 & 0 \\ 0 & \Sigma_P & -\Sigma_H \\ 0 & \Sigma_H & \Sigma_P \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix}$$

again a simple tensor solution.

Neglecting vertical currents at the equator, $J_z = 0$, then

$$E_z = -\frac{\Sigma_H}{\Sigma_P} E_y$$

and

$$J_y = \left(\Sigma_P + \frac{\Sigma_H^2}{\Sigma_P}\right) E_y = \Sigma_C E_y$$

which is the equatorial electrojet and where

$$\Sigma_C = \Sigma_P + \frac{\Sigma_H^2}{\Sigma_P}$$

is the Cowling conductance. The auroral electrojet (6.5) and the equatorial electrojet are therefore the results of polarization enforced by the restriction that $J_x = 0$ and $J_z = 0$, respectively.

This last restriction ($J_z = 0$) on the vertical current is usually enforced at low and middle latitudes and not only at the equator. This is because the main currents are assumed to be flowing in stratified layers above the ground.

With this more general restriction we find:

$$E_z = -\big((\Sigma_\parallel - \Sigma_P)\sin I \cos I\, E_x + \Sigma_H \cos I\, E_y\big)/\Sigma_0$$

where $\Sigma_0 = \Sigma_P \cos^2 I + \Sigma_\parallel \sin^2 I$,

$$J_x = (\Sigma_P \Sigma_\parallel E_x - \Sigma_H \Sigma_\parallel \sin I\, E_y)/\Sigma_0$$

$$J_y = (\Sigma_H \Sigma_\parallel \sin I\, E_x + \Sigma_H^2 \cos^2 I\, E_y)/\Sigma_0 + \Sigma_P E_y$$

or in tensor form:

$$\begin{pmatrix} J_x \\ J_y \end{pmatrix} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix}$$

where

$$\Sigma_{xx} = \Sigma_P \Sigma_\parallel / \Sigma_0$$

$$\Sigma_{xy} = -\Sigma_{yx} = -\Sigma_H \Sigma_\parallel \sin I / \Sigma_0$$

$$\Sigma_{yy} = \Sigma_P + \Sigma_H^2 \cos^2 I / \Sigma_0$$

Since $\Sigma_\parallel \gg \Sigma_P$ and $\Sigma_\parallel \gg \Sigma_H$, we find

$$\Sigma_{xx} \approx \frac{\Sigma_P}{\sin^2 I}$$

$$\Sigma_{xy} \approx -\Sigma_{yx} \approx -\frac{\Sigma_H}{\sin I}$$

$$\Sigma_{yy} \approx \Sigma_P$$

except at the equator where

$$\Sigma_{xx} = \Sigma_\parallel$$

$$\Sigma_{xy} = -\Sigma_{yx} = 0$$

$$\Sigma_{yy} = \Sigma_P + \frac{\Sigma_H^2}{\Sigma_P}$$

as before. In the polar regions we find ($I = 90^\circ$):

$$\Sigma_{xx} = \Sigma_P$$

$$\Sigma_{xy} = -\Sigma_{yx} = -\Sigma_H$$

$$\Sigma_{yy} = \Sigma_P$$

As demonstrated in Section 4.4, the electron density profile in the ionosphere is strongly dependent on the solar zenith angle χ. Since ionospheric conductivities at a given height are proportional to the electron density at this height according to (5.20), the conductivity profiles are also strongly dependent on the solar zenith angle. This is the reason several authors have tried to derive empirical formulas relating quiet time conductivities to the solar zenith angle. Furthermore, solar irradiation or insulation also varies according to conditions on the Sun. An often-used parameter for monitoring solar radiation is the radio flux at 2.8 GHz ($\lambda = 10.7$ cm) (as discussed in Section 1.6). Empirical models for the Hall and Pedersen conductance, based on electron density measurements made by the EISCAT system, are represented by the following formulas:

$$\Sigma_H = S_a^{0.53}(0.81 \cos \chi + 0.54 \cos^{1/2} \chi)$$

$$\Sigma_P = S_a^{0.49}(0.34 \cos \chi + 0.93 \cos^{1/2} \chi)$$

where S_a is the radio flux index in units of 10^{-22} W m^{-2} Hz^{-1}.

Figure 6.11(a) gives some examples of how Σ_P is related to the solar zenith angle for four days in Tromsø when the solar index S_a varies from 74 to 248.

When such models of Σ_P and Σ_H are derived, the different values of Σ_{xx}, Σ_{xy}, and Σ_{yy} can be obtained at any observatory around the globe. Figure 6.11(b) shows such model calculations of Σ_{xx}, Σ_{xy}, and Σ_{yy} as functions of latitude between the equator and the poles.

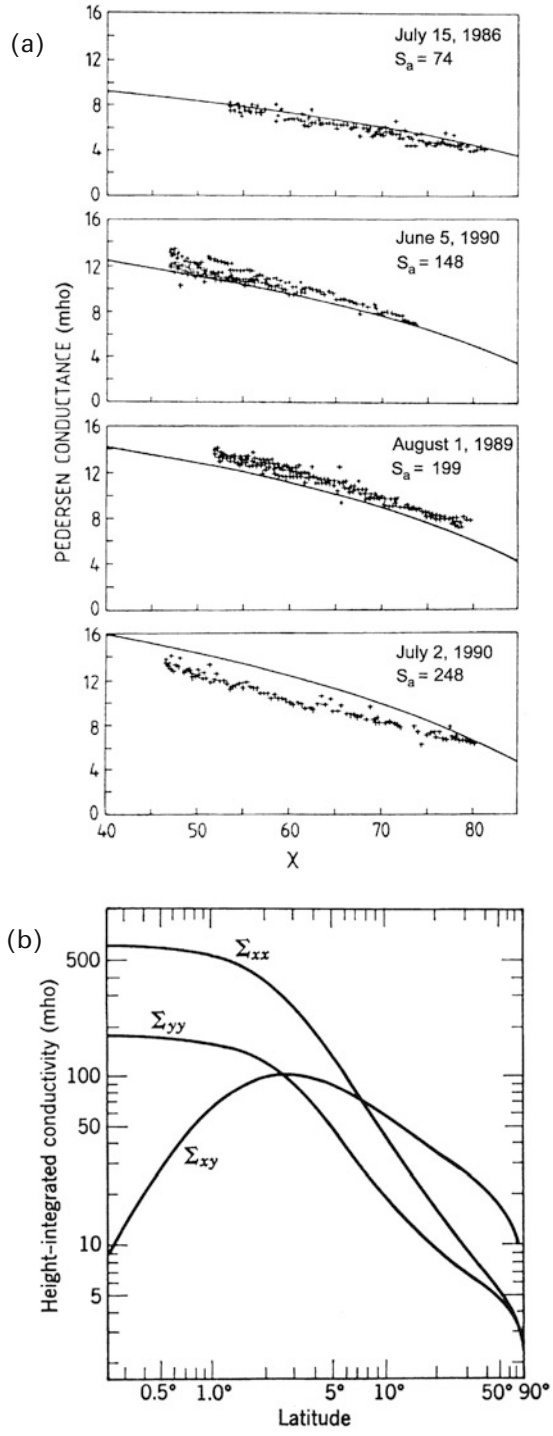
6.5 THE S_q CURRENT SYSTEM

Ground-based magnetic variations observed at stations around the Earth (Figure 6.12) can be used to infer electric fields and neutral winds in the upper atmosphere. The magnetic field fluctuations ΔX and ΔY in the x- and y-directions, respectively, are then assumed to be due to overhead equivalent currents in the x–y plane such that

$$\Delta X \propto J_y \qquad \text{and} \qquad \Delta Y \propto J_x$$

Figure 6.13 illustrates an average global equivalent current system derived from magnetic observations similar to the data shown in Figure 6.12. This current system is called S_q (solar quiet mean) and is presented as seen from the magnetic equatorial plane at the 00:00, 06:00, 12:00, and 18:00 LT meridians. The system at the noon meridian is fairly antisymmetric with respect to the two hemispheres. An

► **Figure 6.11.** (a) Σ_P versus the solar zenith angle for four individual days in Tromsø with different solar flux indexes. Observations are compared with an empirical model derived from EISCAT measurements. (After Moen and Brekke, 1993.) (b) Model calculations of the height-integrated conductivities Σ_{xx}, Σ_{xy}, and Σ_{yy} as functions of magnetic latitude. (From Fejer, 1964.)


(a)
July 15, 1986
S_a = 74
June 5, 1990
S_a = 148
August 1, 1989
S_a = 199
July 2, 1990
S_a = 248
PEDERSEN CONDUCTANCE (mho)
X
(b)
Σ_{xx}
Σ_{yy}
Σ_{xy}
Height-integrated conductivity (mho)
Latitude

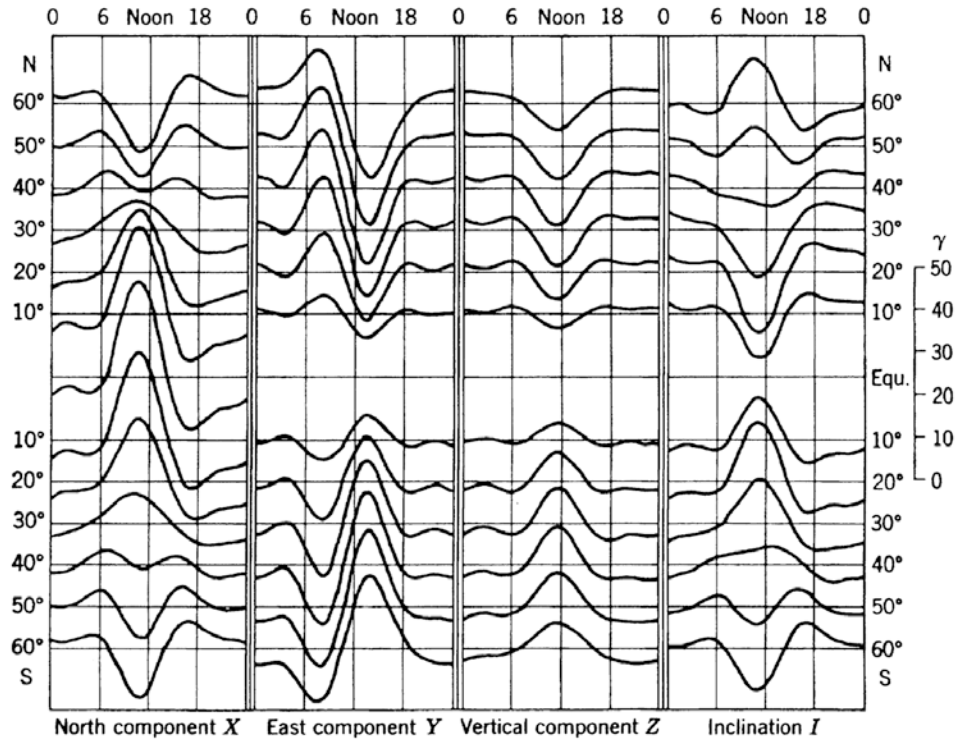


Figure 6.12. Daily solar variation in the X-, Y-, and Z-components and the inclination I at latitudes between 10° and 60° for a sunspot minimum year (1902). (From Chapman and Bartels, 1940.)

anticlockwise current vortex is seen in the northern hemisphere with a maximum of about 182×10^3 A at about 30° latitude, and a clockwise vortex is seen in the southern hemisphere with a minimum of about -155×10^3 A close to the $-30°$ latitude. These two vortices enhance each other at the magnetic equator, creating an eastward current, the equatorial electrojet. This asymmetry is also maintained at midnight except the rotation directions in the vortices are interchanged and the intensities are about one half compared with daytime. Current patterns in the morning and evening represent transitions between the night–day and day–night patterns, respectively.

Since the results presented in Figure 6.13 were obtained on the basis of magnetic field observations at about 35 stations, all below 60° latitude, it is clear that the equivalent current system presented at high latitude and in the polar region does not necessarily represent a true equivalent current system. For this reason special effort has been made to investigate the quiet time high-latitude polar cap equivalent current system (S_q^0). This is particularly important when an understanding of magnetosphere–ionosphere electrodynamic coupling at high latitude is pursued.

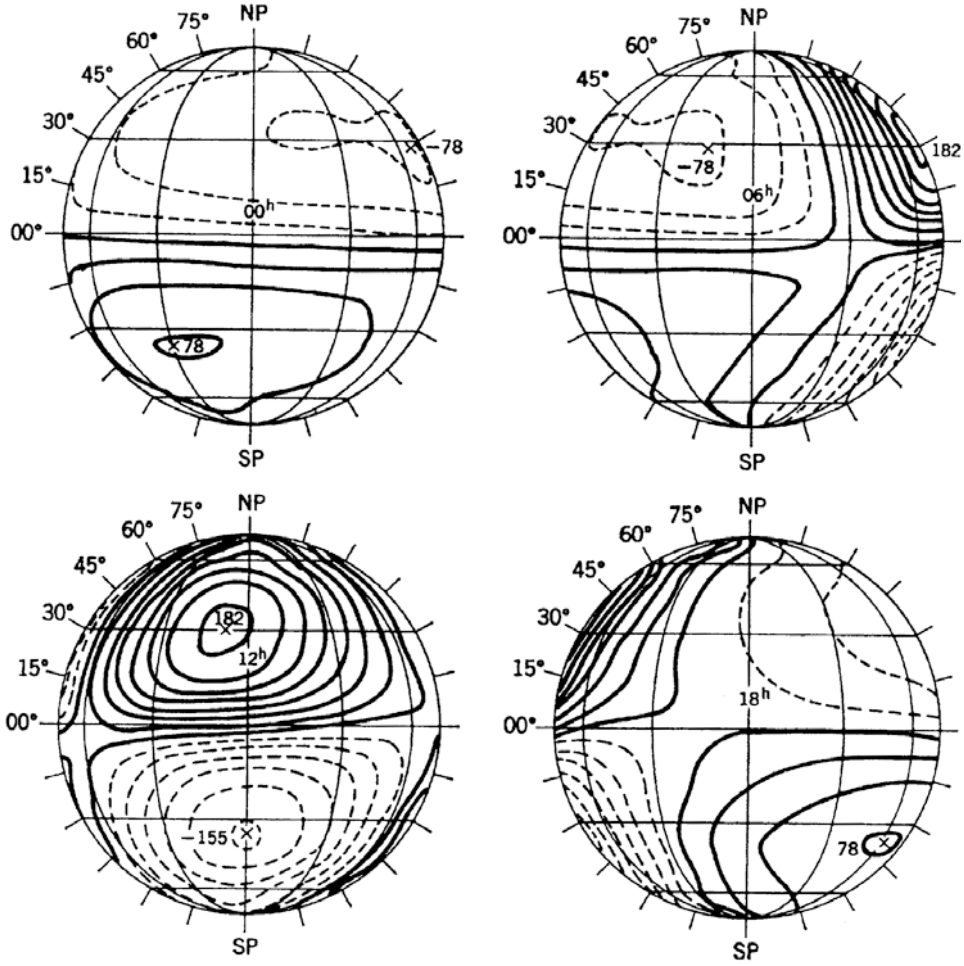


Figure 6.13. Average mean S_q (solar quiet) current systems during the IGY (international geophysical year) viewed from the magnetic equator at the 00:00, 06:00, 12:00, and 18:00 LT meridians. The numbers near the crosses indicate vortex current intensity in 10^3 A. The distance between the current lines corresponds to 2.5×10^3 A. (From Matsushita, 1967.)

The solar quiet polar current (S_q^p) patterns, shown in Figure 6.14 of the northern hemisphere high-latitude region for the summer (a) and winter (b) solstice, are based on data from 13 magnetic observatories, all above 60° latitude. A two-cell pattern is clearly demonstrated in the summer diagram showing anticlockwise rotation on the dayside and clockwise rotation on the nightside when observed from above the North Pole. The cross-polar cap current is reasonably uniform and directed close to the 20:00 and 08:00 LT meridians. Current intensities are, however, more than twice as high in the summer than the winter. This current pattern, which is based on the same data period as the system shown in Figure

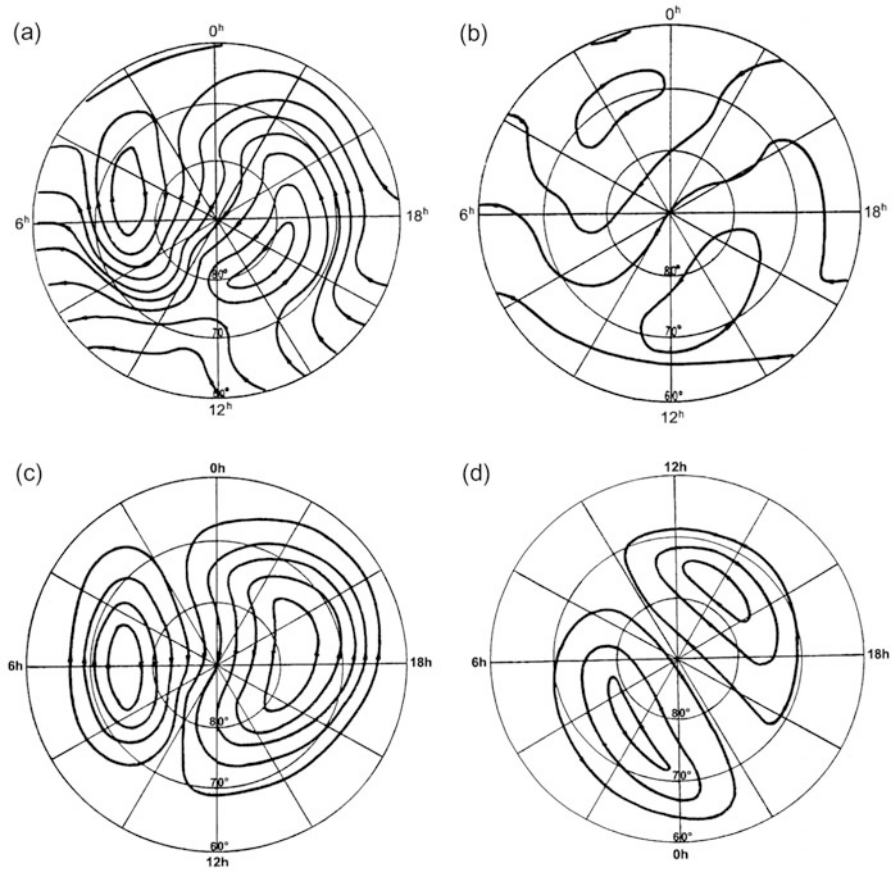


Figure 6.14. (Upper panels) Equivalent current systems of mean daily variations in the north polar region on quiet days: (a) under summer and (b) winter solstice conditions. The separation between isointensity current lines is 2×10^4 A. (Lower panels) Equivalent current systems of additional geomagnetic variation in the northern polar cap, the S_q^p field: (c) under summer and (d) winter solstice conditions. The separation between isointensity current lines is 10^4 A. Note the clock dial in (d) has been turned. (From Nagata and Kokubun, 1962.)

6.13, is believed to consist in extrapolating the S_q current system at middle latitudes, the so-called S_q^0 system, and a polar cap quiet time system called S_p^p. By subtracting S_0^q, as derived from Figure 6.13 from the patterns in Figure 6.14(a) and (b), the special polar cap systems shown in Figure 6.14(c) and (d) can be revealed. These systems are limited to latitudes above 60° and consist of two current cells that depend for magnitude and strength on the season. In the summer season the evening cell which is centered close to the 18:00 local meridian is much larger and stronger than the morning cell which is centered close to the 06:00 local meridian. The evening cell actually extends across the center of the polar cap and into the morning half. Current flow across the polar cap, however, is fairly uniform and almost parallel to the midnight–noon meridian.

In the winter season (Figure 6.14(d)) the two cells are more similar but the axis of symmetry is turned closer to the 22:00 to 10:00 local meridian.

Traditionally, there has been lengthy discussion about how these different current systems (S_q^0 and S_q^p) are maintained and how they interfere with each other. Since no absolute conclusion can be drawn concerning the closure of real currents from ground-based magnetic observations alone, it appears to be a fertile debate to say the least.

It should be noticed, however, that there is a close resemblance between the S_q^p patterns shown in Figure 6.14(c) and (d) and those current cells at high latitudes shown in Figure 6.9, particularly when $B_z = B_y = 0$, which probably represents the average quiet condition rather well (as do the patterns in Figure 6.14).

It is believed that the S_q^p current system, contrary to the S_q^0 system, is not driven by the ionospheric dynamo due to solar heat input but rather by external sources related to processes in the magnetosphere and its interaction with the IMF. Therefore, the S_q^p system is strongly enhanced during geomagnetic disturbances and auroral substorms and does not represent a quiet mean as the index indicates.

By reasonable modeling of Σ_{xx}, Σ_{xy}, and Σyy, the electric fields—whether due to a neutral wind in the ionosphere or a dynamo process in the magnetosphere—can be solved globally. This method of analysis is based on the atmospheric dynamo theory. Since the electrostatic field $\mathbf{E}_\perp$ of magnetospheric origin or due to polarization charges can be deduced from a potential ϕ, we have

$$\mathbf{E}_\perp = -\nabla\phi$$

If it is now assumed that the neutral wind is also derivable from a potential ψ, then

$$\mathbf{u}_n = -\nabla\psi$$

Ohm's law for the height-integrated current can now be written:

$$\mathbf{J} = -\underline{\underline{\Sigma}} \cdot (\nabla\phi + \nabla\psi \times \mathbf{B})$$

where $\underline{\underline{\Sigma}}$ is the conductance tensor. In this sense $\mathbf{F}$ must be derivable from a potential χ such that

$$\nabla\chi = \underline{\underline{\Sigma}} \cdot (\nabla\phi + \nabla\psi \times \mathbf{B})$$

By eliminating ϕ it is possible to solve for ψ when χ is determined by means of the equivalent current system deduced from ground-based magnetic field measurements.

Since the neutral wind is not in fact purely irrotational, it must also include a rotational component in terms of a stream function π. The entire velocity field therefore becomes:

$$\mathbf{u}_n = -\nabla\psi - \hat{\mathbf{r}} \times \nabla\pi$$

where $\hat{\mathbf{r}}$ is a unit vector directed radially outward from the Earth. Complete solutions of this problem are shown in Figure 6.15 in terms of the average ionospheric neutral wind global pattern system. In the upper part of this figure it is assumed

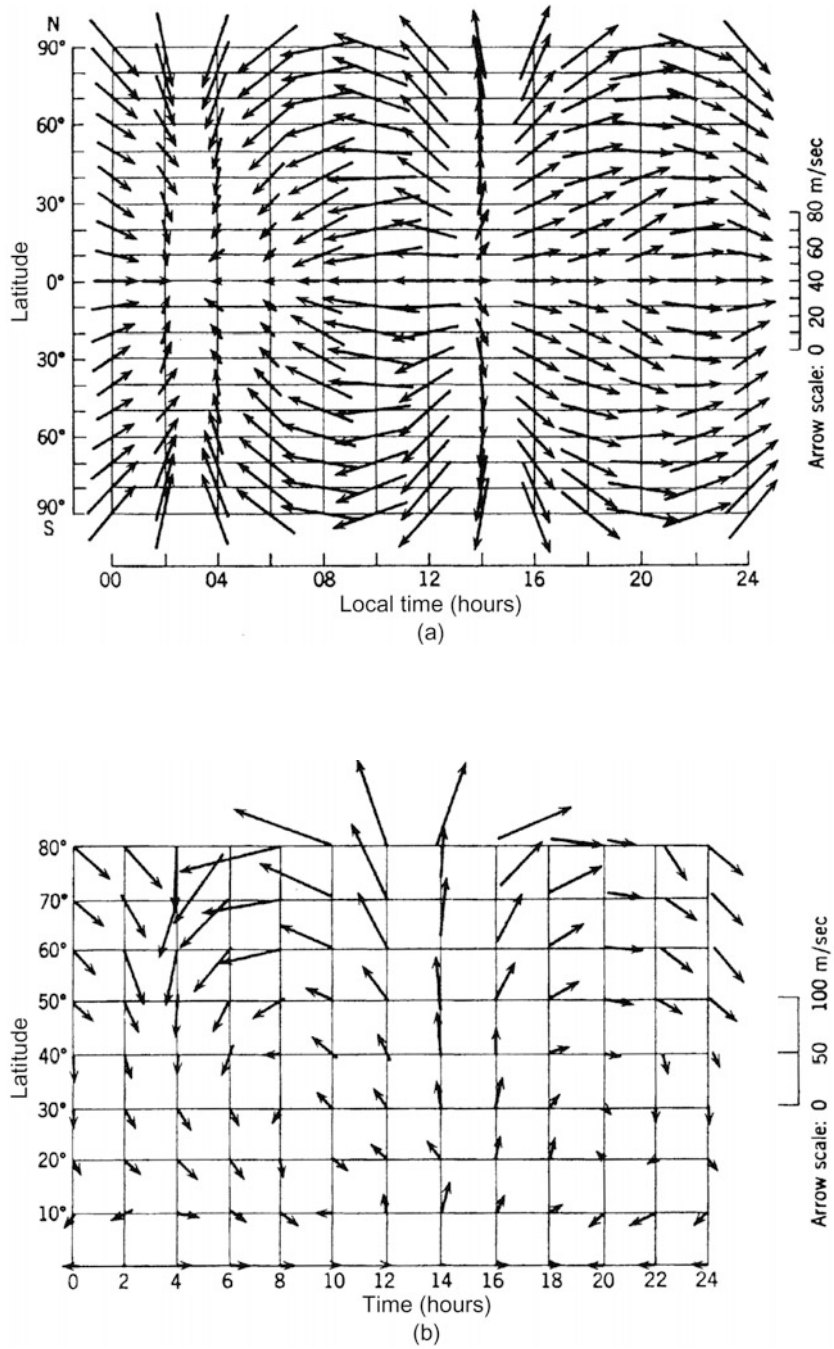


Figure 6.15. Wind systems deduced from S_q variations: (a) with curl $\mathbf{u}_n = 0$; (b) with both rotational and irrotational parts included. (From Maeda and Kato, 1966.)

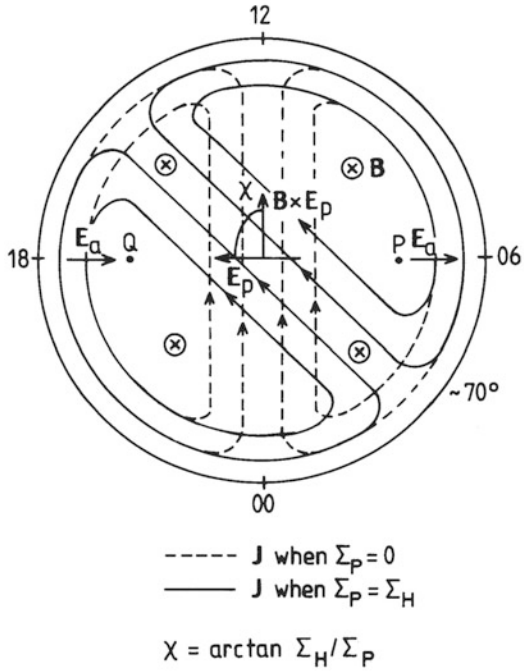


Figure 6.16. Illustration to show the relationship between the cross-polar cap current system and the direction of the cross-polar cap potential gradient when the conductance ratio Σ_P/Σ_H is changed. Dotted lines show the symmetric pattern around the $\mathbf{B} \times \mathbf{E}_P$ vector when $\Sigma_P = 0$, while the full lines show the pattern tilted at an angle $\chi = \arctan \Sigma_P/\Sigma_H$ when $\Sigma_P \neq 0$ ($\Sigma_P = \Sigma_H$ in the drawing).

that $\pi = 0$ ($\text{curl}\, \mathbf{u} = 0$), while in the lower part the rotational term is included. For the curl-free solution the neutral wind mainly blows from the subsolar point at 14:00 LT toward the 03:00 LT meridian at all latitudes. With the curl, however, a two-cell system is established at both hemispheres and velocities increase toward higher latitudes more strongly than for the curl-free situation.

To demonstrate the intimate relationship between the derived ionospheric current pattern from an electric potential distribution and the choice of ionospheric conductivity model, we turn to the simplified picture presented in Figure 6.16.

We assume that there is a constant electric field across the polar cap directed from the high-potential area (P) at dawn to the lower potential area at dusk (Q). Given uniform Hall and Pedersen conductances across the polar cap as Σ_H and Σ_P, the ionospheric current across this region is (6.4) (when $\mathbf{u}_n = 0$):

$$\mathbf{J} = \Sigma_P \mathbf{E}_P + \Sigma_H \mathbf{B} \times \mathbf{E}_P / B$$

where $\mathbf{E}_P$ is the perpendicular electric field across the polar cap, and $\mathbf{B}$ is assumed vertical. Since the vector $\mathbf{B} \times \mathbf{E}_P$ is directed along the midnight–noon meridian, this current will flow from the early morning hours to the early post-noon hours.

Depending on the ratio between the Hall and Pedersen conductances the current will make an angle χ with the $\mathbf{E}_P$ direction given by:

$$\tan\chi = \frac{\Sigma_H}{\Sigma_P}$$

The angle χ will vary with the ratio Σ_H/Σ_P. If $\Sigma_P \ll \Sigma_H$, the current is flowing across the polar cap from midnight to noon (see Figure 6.16).

The combined features of the ionospheric currents derived at high latitudes (as shown in Figure 6.14) probably represent reality rather well; the detailed structure, however, is much less clear. The conductance model used only reflects an envelope of fairly strongly structured conductance patterns, as it is quite likely that the value of $\tan\chi$ will vary drastically throughout the precipitation region. Time variations in Σ_H/Σ_P as derived for the auroral zone station in Tromsø for the same dates as data presented in Figure 6.1 are shown in Figure 6.17. This demonstrates quite dramatic changes in this ratio indicating that the angle χ in Figure 6.16 varies strongly in time.

6.6 MAPPING OF E-FIELDS IN THE IONOSPHERE

We have seen (Section 3.6) how electric fields can map downward from the equatorial plane to the ionosphere or vice versa along the magnetic dipole field lines if they are highly conducting. In the ionosphere, however, conductivity is finite especially perpendicular to magnetic field lines, and the electric potential between two neighboring field lines is not necessarily conserved at ionospheric altitudes. Let us adopt a coordinate system where the z-axis is directed vertically upwards from the ground and let x and y be directed southward and eastward, respectively. Let the magnetic field be directed downward so that

$$\hat{\mathbf{B}} = -\hat{\mathbf{z}}$$

For an electric field given by:

$$\mathbf{E} = E_x\hat{\mathbf{x}} + E_y\hat{\mathbf{y}} + E_z\hat{\mathbf{z}}$$

the ionospheric current density at altitude z will be given by:

$$\begin{aligned}\mathbf{j}(z) &= \sigma_P(E_x\hat{\mathbf{x}} + E_y\hat{\mathbf{y}}) - \sigma_H(E_x\hat{\mathbf{x}} + E_y\hat{\mathbf{y}}) \times \hat{\mathbf{B}} + \sigma_\parallel(\mathbf{E}\cdot\hat{\mathbf{B}})\hat{\mathbf{B}} \\ &= (\sigma_P E_x + \sigma_H E_y)\hat{\mathbf{x}} + (\sigma_P E_y - \sigma_H E_x)\hat{\mathbf{y}} + \sigma_\parallel E_z\hat{\mathbf{z}}\end{aligned}$$

We now assume that conductivities are functions of z only. Since the current must be divergence free we find:

$$\nabla\cdot\mathbf{j} = \sigma_P\frac{\partial E_x}{\partial x} + \sigma_H\frac{\partial E_y}{\partial x} + \sigma_P\frac{\partial E_y}{\partial y} - \sigma_H\frac{\partial E_x}{\partial y} + \frac{\partial}{\partial z}(\sigma_\parallel E_z) = 0$$

Taking advantage of (6.9) we can reduce this to:

$$\nabla \cdot \mathbf{j} = \sigma_P \frac{\partial E_x}{\partial x} + \sigma_P \frac{\partial E_y}{\partial y} (\sigma_\parallel E_z) = 0$$

Since the electric field must be deduced from a potential ϕ, we obtain:

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{1}{\sigma_P} \frac{\partial}{\partial z} \left(\sigma_\parallel \frac{\partial \phi}{\partial z} \right) = 0 \tag{6.14}$$

Because of the anisotropy of conductivity $\sigma_P \neq \sigma_\parallel$ we will transform the z-component by the following substitution (Boström, 1973):

$$dz' = \sqrt{\frac{\sigma_P}{\sigma_\parallel}}\, dz$$

Then we have

$$\frac{\partial}{\partial z} = \sqrt{\frac{\sigma_P}{\sigma_\parallel}} \frac{\partial}{\partial z'}$$

and introducing the geometric mean conductivity

$$\sigma_m = \sqrt{\sigma_P \cdot \sigma_\parallel}$$

we have from (6.14)

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z'^2} + \frac{1}{\sigma_m} \frac{\partial \sigma_m}{\partial z'} \frac{\partial \phi}{\partial z'} = 0 \tag{6.15}$$

This equation will be greatly simplified if we assume that σ_m can be represented at least in limited intervals by an exponential function:

$$\sigma_m = \sigma_0 \exp\left(-\frac{z'}{\alpha} \right)$$

where α is a constant scale height. This does not limit the applicability of the solutions obtained since any reasonable conductivity profile can be closely approximated by a series of simple exponential functions. Equation (6.15) will now read:

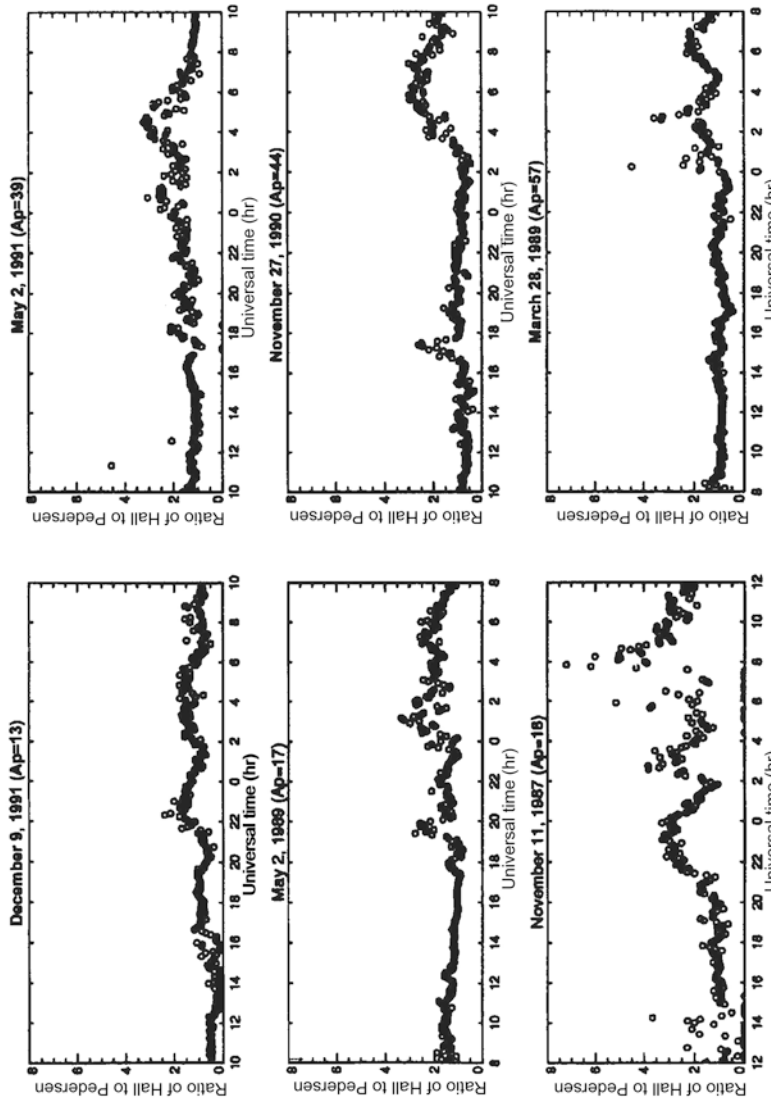
$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z'^2} - \frac{1}{\alpha} \frac{\partial \phi}{\partial z'} = 0$$

which can be solved by the method of separable variables. We will assume that the electrostatic potential is generated at a source level z'_0 and that it has sinusoidal spatial variations with wavelength λ in both the x- and y-directions. The solution is then given by:

$$\phi = \phi_0 \exp[A(z' - z'_0)] \exp\left[\frac{2\pi i}{\lambda} (x + y) \right]$$

where A satisfies the equation

$$A^2 - \frac{1}{\alpha} A - \frac{8\pi^2}{\lambda^2} = 0$$


May 2, 1991 (Ap=39)
November 27, 1990 (Ap=44)
March 28, 1989 (Ap=57)
December 9, 1991 (Ap=13)
May 2, 1989 (Ap=17)
November 11, 1987 (Ap=18)
Ratio of Hall to Pedersen
Universal time (hr)

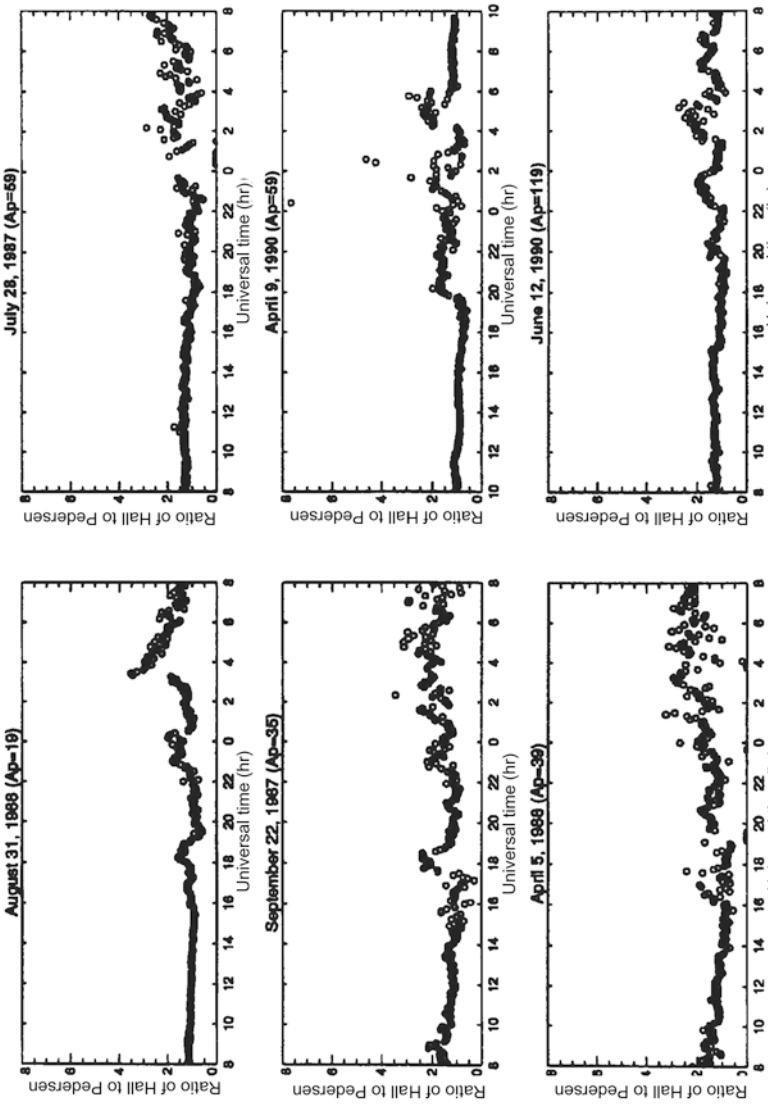


Figure 6.17. Variations in the conductance ratio $R = \Sigma_H/\Sigma_P$ as a function of local time for disturbed days as observed by the EISCAT radar from Tromsø. (Courtesy Nozawa, 1995.)

and

$$A = \frac{1}{2\alpha} \pm \sqrt{\left(\frac{1}{2\alpha}\right)^2 + \frac{8\pi^2}{\lambda^2}}$$

If we let these two roots be denoted by

$$A_1 = \frac{1}{2\alpha} - \sqrt{\left(\frac{1}{2\alpha}\right)^2 + \frac{8\pi^2}{\lambda^2}} \quad (A_1 < 0)$$

$$A_2 = \frac{1}{2\alpha} + \sqrt{\left(\frac{1}{2\alpha}\right)^2 + \frac{8\pi^2}{\lambda^2}} \quad (A_2 > 0)$$

then the complete solution of (6.15) can be expressed as:

$$\phi = \{S_1 \exp[A_1(z' - z'_0)] + S_2 \exp[A_2(z' - z'_0)]\} \times \exp\left[\frac{2\pi i}{\lambda}(x + y)\right]\phi_0$$

where S_1 and S_2 are constants to be determined by the boundary conditions of the problem. To make a close approximation to true physical conditions the ionosphere can be divided below z'_0 into a number of slabs in each of which the conductivity σ_m is assumed to obey an exponential law with a fixed scale height α and to be continuous across the interfaces. Examples of such solutions, which depend on wavelength λ, are shown in Figure 6.18. We notice that A_1 and A_2 determine how fast the field is damped around z'_0 and that this damping depends on the perpendicular scale length λ and the scale height α for the conductivity σ_m. S_1 and S_2 have to be determined from the boundary conditions applied to the problem for each slab (S_{1m} and S_{2m}). For a source region at the top of the ionosphere, $z' - z'_0$ will be negative in the ionosphere and A_2 must be the damping factor while $S_1 = 0$. For a source region in the lower atmosphere, $z' - z'_0$ is positive in the atmosphere and A_1, which is negative, is the damping factor. S_2 is then zero.

From Figure 6.18 we see that large-scale (>100 km) horizontal electric potentials penetrate from the top of the ionosphere below the E-region almost without damping. For small-scale fields (<10 km) damping is severe. It is also noticed that damping is more severe under night-time conditions, when conductivities are low. We can also conclude that electric potential variations with characteristic horizontal scale lengths shorter than 1 km cannot easily propagate to the E-region from the outer magnetosphere, while those with scale size larger than 10 km do so easily.

While the perpendicular electric field component is damped when it penetrates the ionosphere from above, the parallel component is enhanced (Figure 6.19). A parallel component of a 10^{-4} order of magnitude of the perpendicular component at 300 km, after penetrating 90 km into the ionosphere, will be equal to or larger than the perpendicular component, depending on the scale size. This is partly a consequence of isotropic conductivities at these lower heights.

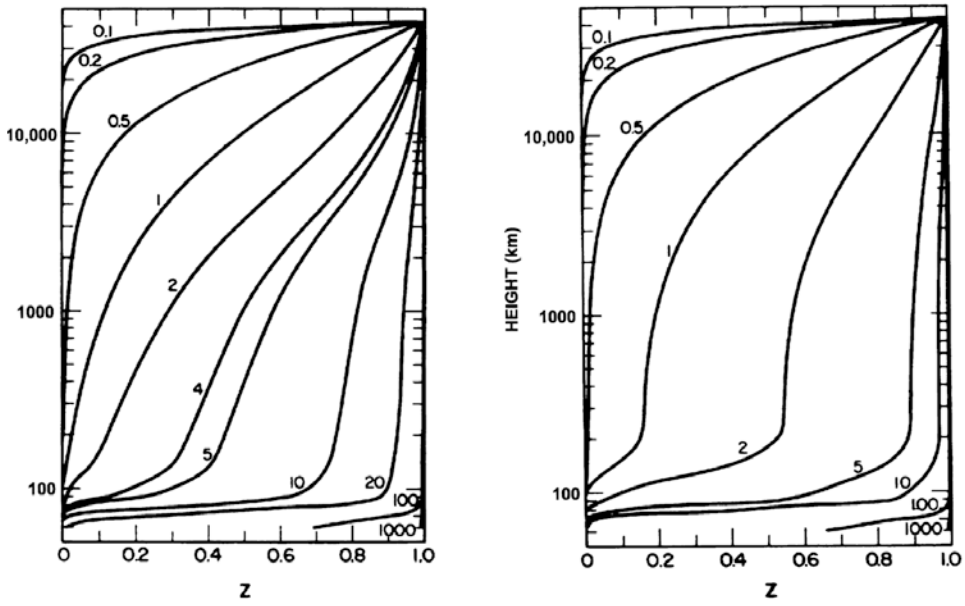


Figure 6.18. (Left panel) The potential transmission factor (Z) as a function of altitude under daytime conditions. Z is the ratio between the potential at a particular altitude and at the source region. The labels on the curves refer to spatial wavelength (λ) in kilometers. (Right panel) The potential transmission factor (Z) as a function of altitude under night-time conditions. (From Reid, 1965.)

For time-varying electric fields the damping relation will be strongly modified, and more strongly so for higher frequency variations.

Very large-scale fields (100 km) will penetrate well below the ionosphere and all the way down to 20–40 km where they can be observed from balloons (as can be seen from Figure 6.18). It is likely that electric fields with extreme horizontal dimensions originating in the magnetosphere can affect the vertical electric field at the surface of the Earth, which is essential for the current circuit associated with weather conditions.

6.7 POLARIZATION FIELDS AROUND AN AURORAL ARC

An auroral arc is associated with an area of enhanced electron precipitation and ionospheric conductance. Associated with the arc there is usually a strong auroral electrojet due to strong electric fields. How the current system is formed and closed around an auroral arc is not well understood since local conductance enhancements, polarization fields, and feedback current loops are all intermingled in a dynamically complicated manner.

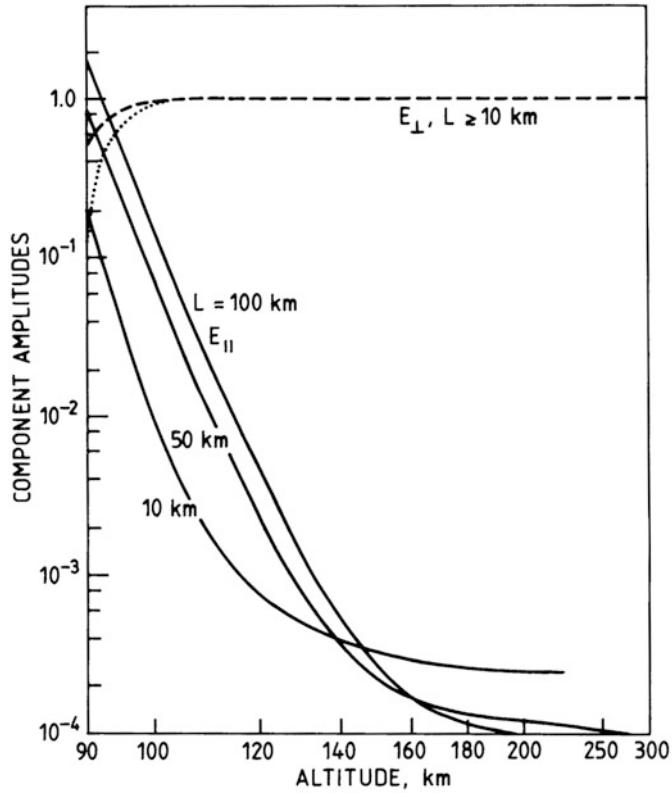


Figure 6.19. The results of self-consistent mapping of perpendicular ($E_\perp$) and parallel ($E_\parallel$) components of electric fields between 90 and 300 km. The dashed curve ($E_\perp$) and its associated solid curves ($E_\parallel$) for different perpendicular scales (L) are obtained for the boundary conditions $E_\perp$ (300 km) = 1.0 and $E_\perp$ (90 km) = 0.5. The dotted curve shows the change in $E_\perp$ if the lower boundary value is altered to 0.01. (After Park and Dejnakarintra, 1974.)

However, we will study a simple model of the configuration of the current within and close to an auroral arc in order to get a better understanding of the physics taking place.

Let the magnetic field be directed vertically downward $\mathbf{B} = B\,\hat{\mathbf{z}}$, the x-axis northward, and the y-axis eastward (Figure 6.20). We will now assume that an electric field with a large (>100 km) horizontal scale length (in the x–y plane) is penetrating the ionosphere from the magnetosphere above. The field $\mathbf{E}_0 = E_0\hat{\mathbf{y}}$ points eastward and will be constant in the whole region, which also includes the auroral arc. The auroral arc is represented by a narrow strip (~10 km) of enhanced electron densities between $x = a$ and $x = -a$. Inside the arc the Hall and Pedersen conductances are given by Σ_H^A and Σ_P^A, respectively. The arc is therefore associated with a thin uniform layer in the vertical direction. Outside the arc the conductances are small but constant and we denote them by Σ_H^C and Σ_P^C, respectively.

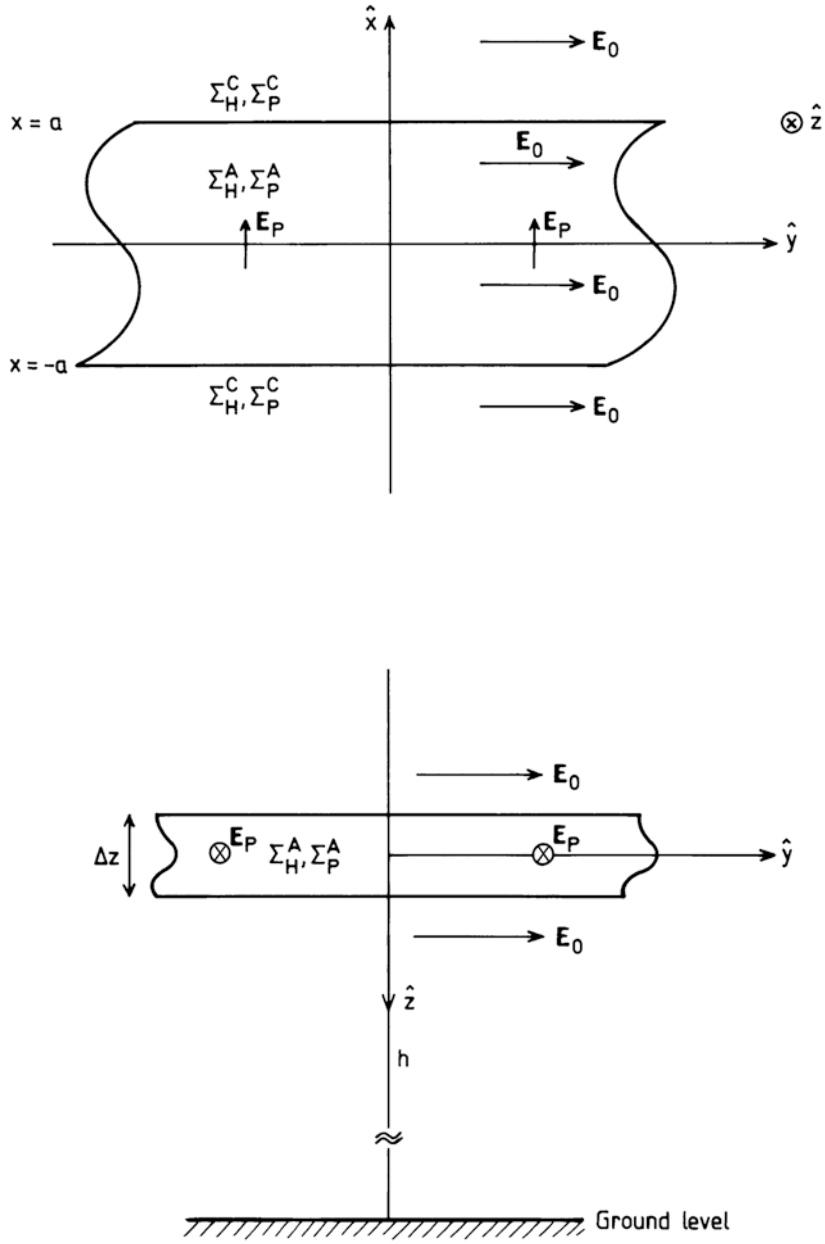


Figure 6.20. Idealized presentation of the polarization effects associated with an auroral arc at height h above the ground. The arc is represented by a slab of enhanced conductivity Σ_P^A and Σ_H^A, within a narrow region in the x-direction $(+a, -a)$, thickness Δz in the z-direction, and an infinite extension in the y-direction. An applied electric field $\mathbf{E}_0$ in the positive y-direction gives rise to a Hall current which polarizes the arc such that a polarization field $\mathbf{E}_p$ is established in the positive x-direction. Altogether this leads to enhanced current in the positive y-direction.

Since the electric field is curl free it must be continuous across the interface between the background and the arc. This electric field

$$\mathbf{E}_0 = E_0\hat{\mathbf{y}}$$

will drive a Pedersen current

$$\mathbf{J}^A_{P_0} = \Sigma^A_P E_0\hat{\mathbf{y}}$$

within the arc and outside the arc:

$$\mathbf{J}^C_{P_0} = \Sigma^C_P E_0\hat{\mathbf{y}}$$

This electric field will also drive a Hall current

$$\mathbf{J}^A_{H_0} = -\Sigma^A_H E_0\hat{\mathbf{x}}$$

within the arc and outside the arc:

$$\mathbf{J}^C_{H_0} = -\Sigma^C_H E_0\hat{\mathbf{x}}$$

The Hall current inside the arc will be larger than the current perpendicular to the arc on the outside of the arc, and charges will accumulate at the interface ($x = \pm a$); positive charges at $x = -a$ and negative charges at $x = +a$. This will again lead to a polarization field perpendicular to the arc in the x–y plane directed parallel to the x-axis. It can be expressed as:

$$\mathbf{E}_p = +E_p\hat{\mathbf{x}}$$

This will drive a Pedersen current within the arc given by

$$\mathbf{J}^A_{P_p} = +\Sigma^A_P E_p\hat{\mathbf{x}}$$

and a Hall current within the arc given by

$$\mathbf{J}^A_{H_p} = +\Sigma^A_H E_p\hat{\mathbf{y}}$$

The current perpendicular to the arc must be continuous as long as charges are not allowed to move along the magnetic field lines. Then

$$\mathbf{J}^A_x\hat{\mathbf{x}} = -\Sigma^A_H E_0\hat{\mathbf{x}} + \Sigma^A_P E_p\hat{\mathbf{x}}$$

$$\mathbf{J}^C_x\hat{\mathbf{x}} = -\Sigma^C_H E_0\hat{\mathbf{x}}$$

If we now assume that $\Sigma^C_H \ll \Sigma^A_H$ such that the current perpendicular to and outside the arc can be neglected, then

$$J^A_x = -\Sigma^A_H E_0 + \Sigma^A_P E_p = 0$$

and

$$E_p = +\frac{\Sigma^A_H}{\Sigma^A_P}E_0$$

The polarization field is enhanced with respect to the applied field E_0 by the

conductivity ratio Σ_H^A/Σ_P^A. The current along and inside the arc is now given by

$$J_y^A\hat{\mathbf{y}} = \Sigma_P^A E_0\hat{\mathbf{y}} + \Sigma_H^A E_p\hat{\mathbf{y}}$$

and finally:

$$J_y^A = \left[\Sigma_P^A + \frac{(\Sigma_H^A)^2}{\Sigma_p^A}\right]E_0 = \Sigma_P^A\left[1 + \left(\frac{\Sigma_H^A}{\Sigma_p^A}\right)^2\right]E_0$$

and the auroral electrojet flows along the applied **E**-field $E_0\hat{\mathbf{y}}$, but is enhanced with respect to the Pedersen current. The enhancement factor within the brackets can be much larger than 1. The conductivity is now the auroral electrojet conductance as derived in (6.5)

$$\Sigma_A^A = \Sigma_P^A + \frac{(\Sigma_H^A)^2}{\Sigma_P}$$

due to the polarization effect of the Hall current. An illustration of this polarization effect is also demonstrated in Figure 6.3.

If the current perpendicular to and outside the arc is not completely neglected and the current along the magnetic field line is still assumed to be vanishing, the polarization field could be expressed as:

$$E_p = -\frac{\Sigma_H^A - \Sigma_H^C}{\Sigma_P^A}E_0$$

and the auroral electrojet will be:

$$J_y^A = \left(\Sigma_P^A + \frac{(\Sigma_H^A)^2}{\Sigma_P^A} - \frac{\Sigma_H^A\Sigma_H^C}{\Sigma_P^A}\right)E_0$$

The electrojet would then be reduced compared with the ideal case when the current perpendicular to and outside the arc is zero.

Finally, a current moving vertically upward on the southern edge of the arc would reduce the electrojet even more. By studying conductivities and electric fields outside and inside the arc, one can deduce the extent of the polarization actually taking place.

6.8 CURRENTS RELATED TO AN AURORAL ARC

Auroral displays associated with strong electron precipitation from the magnetosphere often appear as narrow elongated arcs. It is also observed that plasma motion, which can be very strong adjacent to the arc, changes direction across the arc. In fact, this electron precipitation tends to occur in regions in which flow vorticity has a positive sign. Figure 6.21 illustrates a cross-section of such an auroral arc directed along the x-axis. The length L of the arc is much larger than the width a. The z-axis is parallel to the vertical magnetic field and

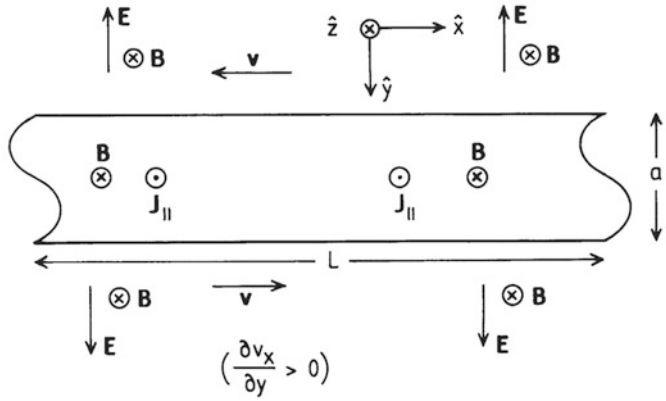


Figure 6.21. Idealized model of a narrow strip in the ionosphere where precipitation takes place associated with wind shear across the strip. Vertical currents associated with precipitating electrons flow antiparallel to the **B**-field and out of the region in the northern hemisphere when flow vorticity is positive. The strip is then associated with high potential as the **E**-field is also directed away from the strip.

points into the paper. The y-axis then completes the right-handed coordinate system.

Flow vorticity is defined as the curl of the velocity field

$$\boldsymbol{\xi} = \nabla \times \mathbf{v}$$

When $E_z = 0$, plasma flow velocity $\mathbf{E} \times \mathbf{B}/B^2$ is independent of z and, assuming that the flow is parallel to the arc everywhere and along the x-axis, vorticity has only a z-component

$$\boldsymbol{\xi} = \frac{\partial}{\partial y}\hat{\mathbf{y}} \times (v_x\hat{\mathbf{x}} + v_y\hat{\mathbf{y}}) = -\frac{\partial v_x}{\partial y}\hat{\mathbf{z}} = \xi_z\hat{\mathbf{z}}$$

as there are no variations in the z- and x-directions of $\mathbf{v}$, and the v_z-component is zero. Again the parallel current density is given by (remember $\mathbf{B} = B\hat{\mathbf{z}}$):

$$j_\parallel = -\nabla_\perp(\underline{\underline{\Sigma}}_\perp \cdot \mathbf{E})$$

When assuming uniform conductances, we obtain

$$j_\parallel = -\Sigma_P \nabla_\perp \cdot \mathbf{E} = -\Sigma_P \frac{\partial E_y}{\partial y}$$

since $\partial/\partial x \equiv 0$. When velocity is due to the $\mathbf{E} \times \mathbf{B}$ drift, we find that

$$\mathbf{E} = -\mathbf{v} \times \mathbf{B} = -v_y B\hat{\mathbf{x}} + v_x B\hat{\mathbf{y}}$$

and

$$\frac{\partial E_y}{\partial y} = +\frac{\partial v_x}{\partial y} B$$

since **B** is independent of y. Finally

$$j_{\parallel} = -\Sigma_P B \frac{\partial v_x}{\partial y} = \Sigma_P B \xi_z$$

The current parallel to the magnetic field (i.e., upward on the northern hemisphere) is therefore associated with positive vorticity. Precipitating electrons, however, are associated with currents out of the ionosphere or antiparallel to **B** in the northern hemisphere. Therefore, this current is related to positive vorticity, as stated above.

We also notice from Figure 6.21 that the electric field will point away from the arc on both sides if there is excess positive charge in the arc. The current then tends to neutralize the charge either by positive charges leaving the ionosphere or electrons precipitating into the ionosphere.

6.9 EXERCISES

1. Derive (6.15).
2. Let $\mathbf{E}_0$ in Figure 6.20 be given by

   $$\mathbf{E}_0 = -E_0 \hat{\mathbf{y}}$$

   and derive the auroral electrojet.
3. An infinitely long line current I is situated at a height h above the ground. The current is directed in the eastward (y) direction (Figure 6.22). At two stations S and T separated a distance $L > h$ along the x-axis and perpendicular to the current, variations in the Earth's magnetic field are observed. We will assume that the background quiet time magnetic field $\mathbf{B}_0$ is the same at both stations. Let the observed components of magnetic field fluctuations at a certain moment be $(\Delta x_S, 0, \Delta z_S)$ and $(\Delta x_T, 0, \Delta z_T)$ at station S and T, respectively.

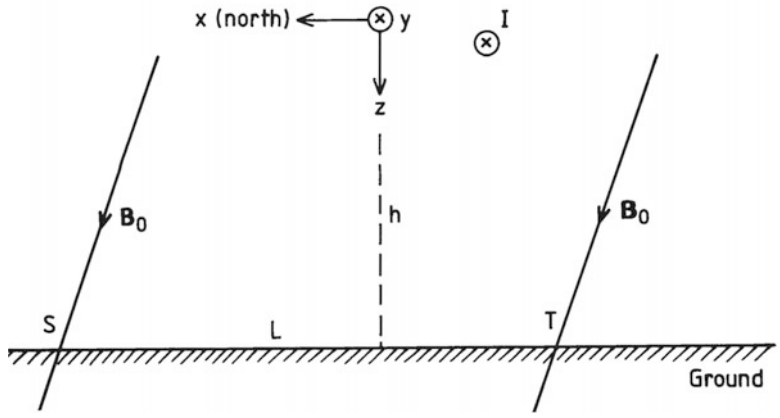


Figure 6.22.

(a) Find an expression for the height h of the current when the distance L is known.

Let $L = 200\,\text{km}$ and the observed components:

$$\Delta x_T = 1.42 \times 10^{-6}\ \text{tesla} \qquad \Delta z_T = 0.59 \times 10^{-6}\ \text{tesla}$$

$$\Delta x_S = 0.65 \times 10^{-6}\ \text{tesla} \qquad \Delta z_S = -0.81 \times 10^{-6}\ \text{tesla}$$

(b) Determine the height and the strength of the current ($\mu_0 = 4\pi \times 10^{-7}\ \text{H/m}$).

The current is now moving in the x-direction with a constant velocity v.

(c) Show that the constant velocity v can be expressed by:

$$v = h\frac{d}{dt}\left(\frac{\Delta z_T}{\Delta x_T}\right)$$

7

The aurora

7.1 AN HISTORICAL INTRODUCTION

No one knows where and when the aurora was first observed. Most certainly it has embellished part of the dark night sky and danced between the stars as long as man has lived; its overall appearance and display, however, might have changed.

Some people think they find traces of auroral allegories in the Bible or in mythological deliberations, as for example in Norse mythology. Others believe they can recognize the auroral motif in ancient wall carvings, as for example in India.

The aurora has been given many different names: the Chinese of old called the lights "flying dragons"; the Scots "merry dancers"; the ancient Greeks "chasmata"; and the Norwegians "northern lights" or "weather lights".

Several cultures in the Arctic have related the aurora to the realm of death. In some areas of Scandinavia, for example, it was believed that the aurora was the spirit of dead maidens. The Eskimos of Greenland believed the lights were the spirits of stillborn children playing ball. Others believed the aurora related to an eternal fight between those who were killed in war or by other brutal means.

In historical times the aurora was observed as far south as Greece and Italy and in the northern part of Japan, and in modern times in the southern states of the U.S.A.

As in so many areas of science, auroral research had a faultering start at the end of the Middle Ages. From pamphlets distributed in Germany as late as the 17th and 18th centuries, the aurora was depicted as fighting hordes (Figure 7.1) and imaginary monsters in the sky, believed to be bad omens of war, pestilence, famine, and fire. It was often seized as an opportunity by both the Church and the various governments of the day to menace the people into paying their tithes or taxes.

A. Brekke, *Physics of the Upper Polar Atmosphere*, Springer Praxis Books,
DOI 10.1007/978-3-642-27401-5_7, © Springer-Verlag Berlin Heidelberg 2013

Figure 7.1. A fantastic illustration of an aurora observed in central Europe on February 10, 1681. The Sun is just dipping below the horizon to the right, and the aurora is represented by burning castles and an advancing troop of cavalry. The river at the front is the Danube. (From Réthly and Berkes, 1963.)

However, individuals such as the celebrated Danish astronomer Tycho Brahe (1546–1601), who meticulously observed the night sky from Uraniborg, also noticed a few auroral displays without reflecting further about their cause. Nevertheless, he did make a few notes (Figure 7.2), which in later times have been used to study the occurrence of aurora in history.

The 17th century with all its wars and depressions was not favorable for auroral recording, and may even have been a period of low auroral activity. Nothing much was reported about the aurora between Brahe's notes and the beginning of the 18th century. As a matter of fact, the phenomenon was considered new in many academic circles when strong auroral displays occurred in March 1716, which could be observed over large areas of Europe. When looking back at the large number of papers devoted to these events in the early 18th century, it is not exceptional to find articles dealing with the question of the age of the aurora. A notable group of scholars in Scandinavia took exception to the contentions from more renowned scientists in central Europe that the aurora was a new heavenly display marking the entrance to a new century. The Norwegian historian Gerhard Schøning (1722–1786), for example, wrote a 120-page treatise in

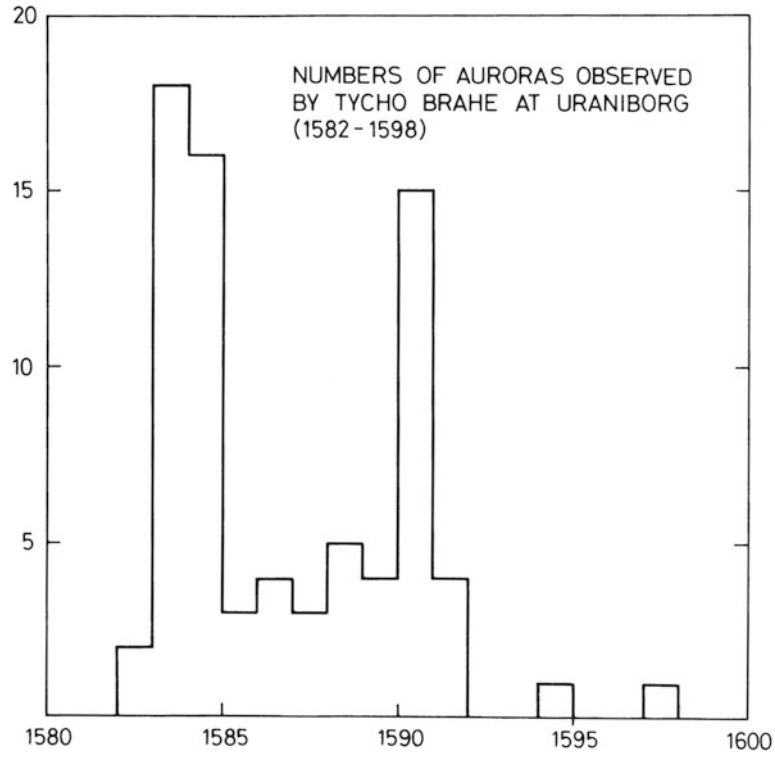


Figure 7.2. Number of auroras observed per year between 1582 and 1597 by Tycho Brahe. (From Brekke and Egeland, 1983.)

two parts in the *Proceedings of the Royal Norwegian Academy* in 1760, carrying the title: "The age of the Northern Lights proven with testimonies from ancient writers." He strongly rejected the absurd claim that the aurora was a new phenomenon since he could prove that the Vikings had known about this celestial "apparition" which they so descriptively called the "northern lights". The facts that Schøning referred to could be found in an ancient book from about AD 1230 entitled *The King's Mirror*. In the book the aurora was referred to as a phenomenon often observed by people sailing to Greenland. Schøning himself lived in the middle of Norway. This statement has remained a puzzle for generations. Since the aurora today is probably more often observed in Norway than in Greenland, it is surprising that a Norwegian author talked about it as if he had not seen it. However, this might be understood in the context of the close connection between the appearance of the aurora and the configuration of the Earth's magnetic field (as mentioned in Section 3.4).

In *The King's Mirror* attempts were also made to explain the cause of the phenomenon. A big fire encircling the Earth, the radiation of stored sunlight in ice and snow, and the reflection of solar rays by snow crystals in the air were all

considered ingredients in the scholarly comprehension of the aurora among the Nordic peoples.

Sir Edmund Halley (1656–1742) was afraid of leaving this Earth without having observed the aurora, the only known light phenomenon in the sky that he had not witnessed. When he finally experienced the gigantic display of March 1716, he immediately sat down and wrote about his observations. He also launched a theory far more advanced than the almost 500-year-old allegations put forward by the author of *The King's Mirror*. Halley took advantage of his detailed knowledge of the Earth's magnetic field (see Figure 3.2) and, as a result of the appearance of the auroral display with its arched form encircling the northern sky, he proposed it had to be linked in some way to the geometry of this field. He turned the matter upside down and proposed that so-called "magnetic effluvia" streamed out of pores in the ground and, on contact with moisture high in the atmosphere, mingled with it thus forming the aurora.

Modern auroral science is rooted in scholarly discussions in parts of the world where the aurora was seldom seen. However, it was in these circles that the phenomenon was to become known by its international name, the "aurora borealis", a misleading appellation which has nothing to do with the beautiful goddess of dawn. Since the appearance of the aurora is dominated by its crimson color when observed in Paris or Rome, the name is quite adequate. For a northerner, however, used to observing the phenomenon with its spectrum of colors and wild wavy display, the name is rather unseemly.

7.2 THE HEIGHT OF THE AURORA

In the early days of science the aurora was considered to be a meteorological phenomenon and was often called a "meteoron". Traditionally, the aurora was often called "weather lights", and all around Norway there were weather signs related to the aurora. The northern lights were prognostic as far as the weather was concerned and, interestingly enough, local weather signs throughout the country with respect to the aurora were quite similar.

As the repressive Middle Ages in Europe with its hostile attitude towards individual and progressive science faded away after the turmoil following the Reformation and the subsequent Thirty Years' War (1618–1648), interest in the auroral phenomenon grew, as substantiated by the increasing numbers of auroral recordings made at the beginning of the 18th century.

The height of the phenomenon became an important issue. Rather conflicting records appeared, where the aurora was said to have been observed below the peaks of mountains and even as far down as the surface of the sea. Others maintained that it was a phenomenon appearing in the upper strata of the atmosphere. How far the atmosphere reached above the ground, however, was still an unsolved problem around the turn of the 18th century, although Halley had given evidence of an exponentially decaying pressure with height—in principle, an infinite atmosphere.

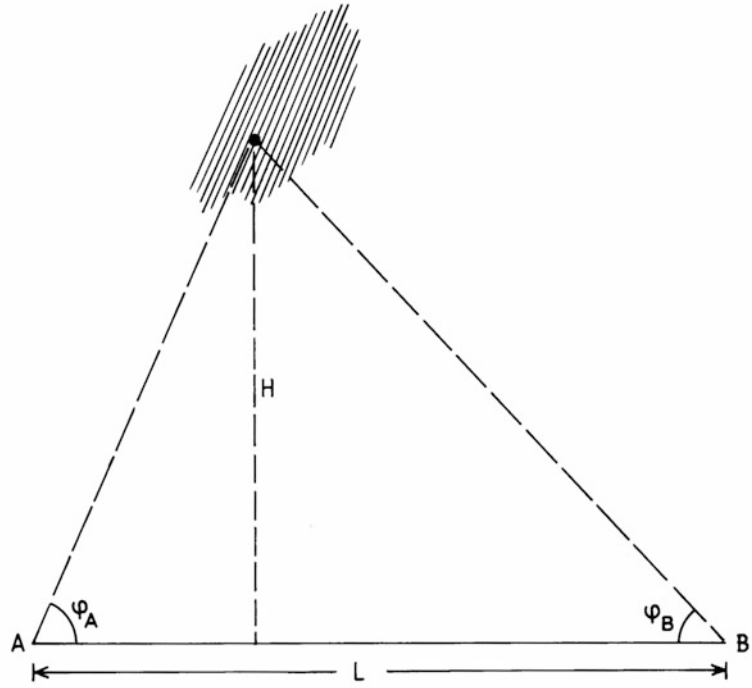


Figure 7.3. Schematic illustration of the principle behind auroral triangulation. Two observers A and B at a known distance L simultaneously observe a point in an aurora at elevation angles φ_A and φ_B, respectively. By simple geometric calculations the height can then be derived.

The famous French astronomer Gassendi had as early as 1621 estimated the height of the aurora to be about 850 km as observed from Paris. A few individual measurements of the auroral altitude were reported from places like Paris, Geneva, Copenhagen, and St. Petersburg between 1730 and 1750, giving values varying from 200 to 1,000 km; most of the events observed, however, were reported to be above 600 km. From Uppsala in Sweden the well-known physicist Anders Celsius (1701–1744) and his colleagues reported on 13 observations of the auroral altitude which, by means of triangulation, were between 380 and 1,300 km with a mean height of 760 km.

Precise triangulation (Figure 7.3) of an aurora in the early part of the 18th century was not an easy task. Because of the real height of the aurora, which is more than 100 km above the Earth's surface, two observers on the ground had to be separated by a substantial distance in order to obtain reliable observations. Furthermore, these observers had to observe exactly the same point in the aurora at exactly the same time, and this was more or less impossible without online communication.

The height of the aurora could therefore not be resolved before the invention of the telephone and the camera, which took place toward the end of the 19th

century. The Norwegian mathematician Carl Størmer (1874–1957) made auroral triangulation a lifelong enterprise. From his more than 20,000 auroral pictures he derived the height of about 10,000 of them, the statistical distribution of which is presented in Figure 7.4.

The distribution maximum is found to be close to 100 km, none is observed below 70 km, and only about 6.5% are found above 150 km.

Although improvements have been made to Størmer's method, thanks to TV cameras and modern automation, his results have proven to be a frame of reference for most discussions concerning the height of auroras. There are indications, however, that thin auroral layers occurring quite frequently below 100 km may have escaped Størmer's careful analysis of photographic plates.

7.3 THE OCCURRENCE FREQUENCY OF THE AURORA

When the solar cycle was discovered in 1844 by Schwabe, it was soon realized that auroral occurrence was strongly related to this cycle. The fact that auroral displays appeared to agitate the magnetic needle had already been reported in 1741 by Hiorter and Celsius, and in 1826 Ørsted was able to explain this relationship by an electric discharge along the auroral arc. A close relationship between variations in the Earth's magnetic field and the solar cycle was established by the middle of the 19th century by Sabine and Wolf. The foundation for auroral research to become an integral part of planetary science was grounded in electromagnetic coupling between the Sun and the Earth.

As early as 1733 the French astronomer de Mairan, in a beautiful treatise entitled "Traité Physique et Historique de l'Aurore Boréale", introduced a hypothesis that the aurora was the result of an interaction between the atmospheres of the Sun and the Earth. Lacking any knowledge of electromagnetism, he leant upon the forces of gravity only. His basic idea was brought to light again toward the end of the 19th century by the theory that particles from the solar atmosphere were a source of the aurora. This was highlighted by the work of Birkeland who proposed that the aurora was caused by cathode rays (electrons) streaming out from the Sun and guided by the magnetic field towards the polar atmosphere of the Earth.

One early graph showing the high correlation between the number of auroras reported per year in Scandinavia between 1761 and 1877 and the annual sunspot number is due to Tromholt (1851–1896) and is shown in Figure 7.5. The correlation is good, especially until 1865; after that, however, the number of auroral observations appears to increase disproportionately. This small discrepancy serves as a warning to people employing statistics of visual auroral sightings in scientific work, because they are the sources of many errors. In this example the apparent increase in auroral occurrence after 1865 is more a result of Tromholt's dedicated work to encourage people to observe the phenomenon (he had about 2,000 correspondents), and of the fact that aurora occurrence had became a part of

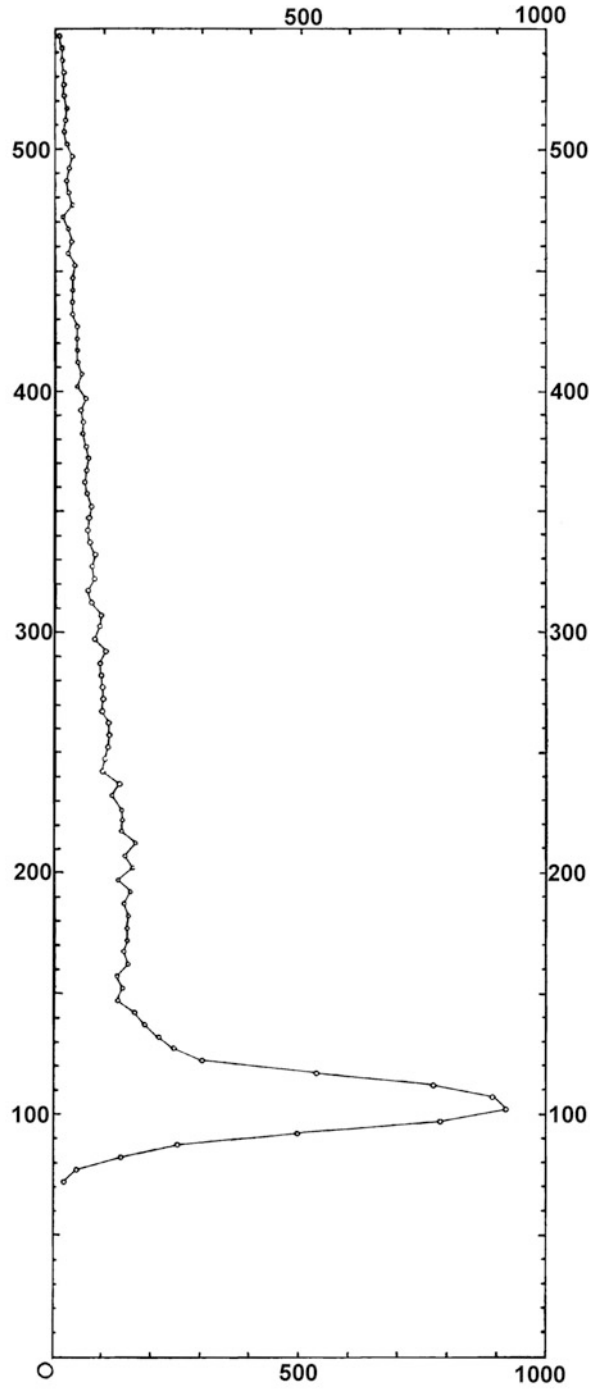


Figure 7.4. The distribution of observed heights of auroras. (From Størmer, 1955.)

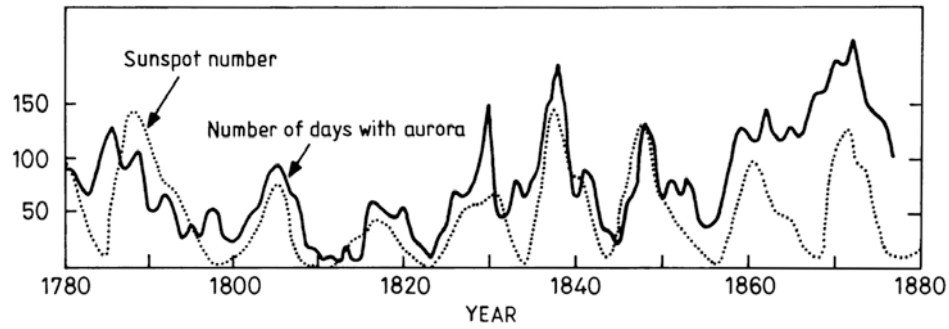


Figure 7.5. The annual numbers of auroras recorded in Scandinavia between 1780 and 1877 according to a survey made by Sophus Tromholt (1851–1896). These numbers are compared with the annual sunspot number for the same period. (After Brekke and Egeland, 1983.)

meteorological observations at many weather stations in Scandinavia, than a genuine increase in auroral occurrence.

A well-known and much used catalog of auroral observations is due to Rubenson, the director of the Swedish Meteorological Institute in the last part of the 19th century. The annual numbers of recordings are shown in Figure 7.6. The increase in auroral observations towards the end of the 19th century is again evident with respect to the earlier part of the time series. Although the Rubenson catalog represents a meticulous work in collecting newspaper articles, notes, diaries, meteorological journals, etc. for almost 200 years, the statistics have little scientific value quantitatively because the catalog does not represent a homogeneous dataset. When studying auroral activity from annals and suspect old notes, one also needs to know something about the activity of individual recorders and this is often the most difficult part of the task. The likelihood of observing an aurora and getting it written down on paper for later use was far higher toward the end of the 19th century than at the beginning of the 18th. The fact that the aurora has a tendency to appear in relation to variations in the solar cycle cannot be denied, but to express this in a quantitative manner by, for instance, a correlation coefficient, is far from well established, let alone an understanding of the physical processes linking the appearance of sunspots with that of auroral displays.

Looked at annually, it appears that the aurora has a tendency to occur more frequently at equinoxes than solstices (Figure 7.7), even when corrected for cloudiness and number of dark hours per night. The reason for this is not quite clear, but is most likely due to seasonal difference in electromagnetic coupling between the solar wind and interplanetary field with the Earth's magnetic field and magnetosphere.

The aurora frequently shows a 27-day periodicity in relation to disturbances on the Sun, which have a tendency to repeat themselves with a period determined by the rotation period of the Sun as seen from the Earth (close to 27 days).

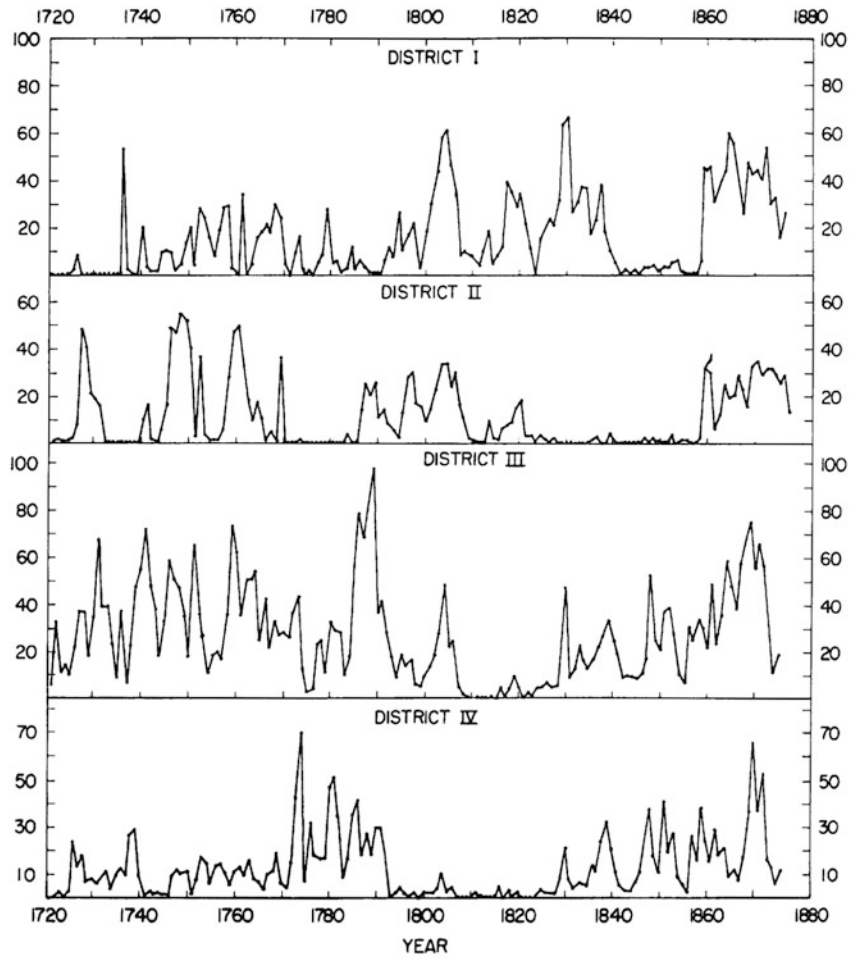


Figure 7.6. Annual numbers of aurora recorded in Sweden for the period 1720–1875 according to a data catalog prepared by Rubenson (1879, 1882).

7.4 GLOBAL DISTRIBUTION OF THE AURORA

The existence of an auroral zone encircling the polar regions was envisaged as early as 1833 by the German scientist Muncke. In 1860 the American physicist Elias Loomis had collected enough data globally to make an attempt to draw the geographical position of the auroral zone. However, in 1881 the Swiss physicist Hermann Fritz published a book called *Das Polarlicht* in which he, on the basis of a large set of global data, was able to draw so-called isochasms (the word "chasms" is borrowed from Aristotle who used this expression to describe the aurora) in the northern hemisphere (i.e., lines where the aurora occurs with equal frequency at midnight).

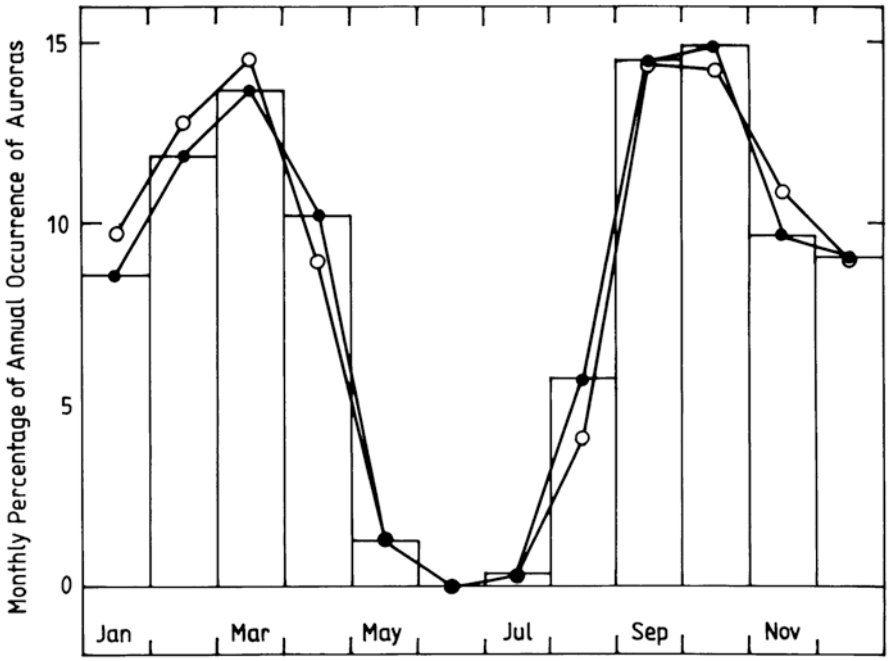


Figure 7.7. Average monthly percentage of auroral observations according to data compiled in Sweden (dots) from 1720 to 1877 (Rubenson, 1882) and data compiled in southern Finland (circles) between 1748 and 1843 (Hällström, 1847). (After Nevanlinna, 1995.)

Figure 7.8 is a reproduction of Fritz's drawing. The numbers labeled on the isochasms indicate the number of nights per year when the aurora occurs. Inside the maximal zone this is, according to Fritz, more than 100 nights. However, in Rome and Madrid it is less than one. On average, it corresponds to about one auroral display over Madrid per solar cycle.

When we track the maximal zone around the polar region, we notice that it approximately follows the geomagnetic latitude close to 67° in Europe, touches the northern part of Scandinavia, crosses Novaya Zemlya, passes north of Siberia, and reaches the northern part of North America. Greenland and Iceland, however, are on the poleward side of the maximal zone where auroral occurrence is less.

The mild climate in Scandinavia and the relative ease of access to this part of the world has, because of the position of the auroral zone, given these countries an advantage in ground-based auroral research, as substantiated by the many Scandinavian pioneers in this field of science.

Since the early 18th century, however, the notion has been put forward that the auroral arc formed part of a luminous ring encircling the global pole. As a matter of fact, this concept was used as a reference for some height estimates of the aurora based on measurements by one observer only (Figure 7.9). Assuming

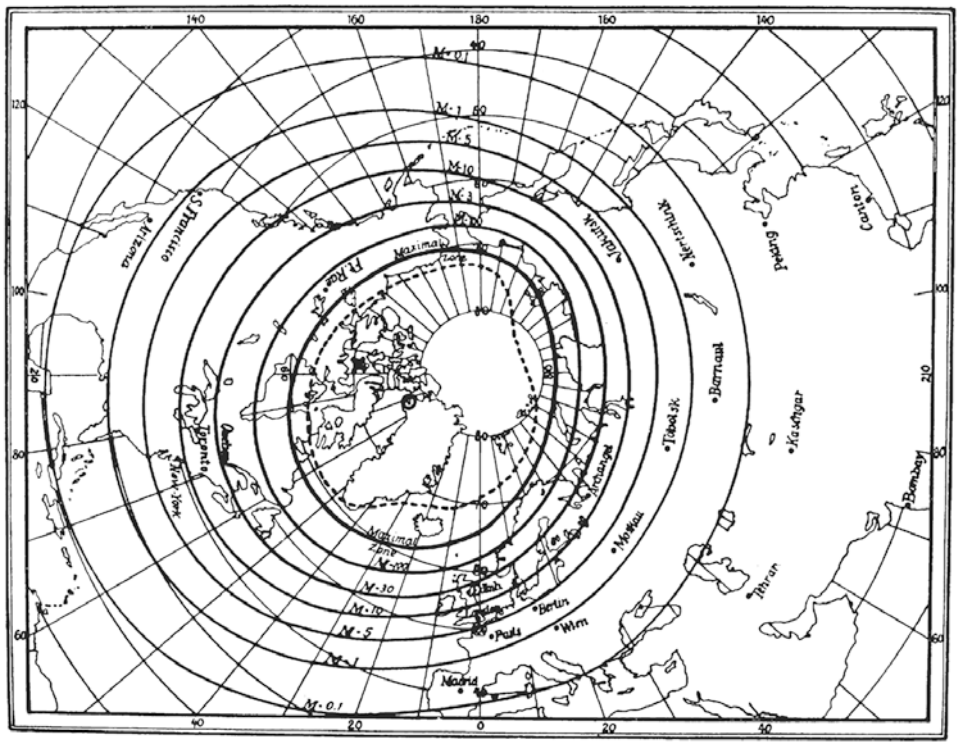


Figure 7.8. Fritz's map of isochasms showing the occurrence frequency of auroral displays at midnight for different places on Earth. The labels marked on the curves show the number of nights per year when the aurora can be observed. The geomagnetic pole is marked by a small circle in the northwestern corner of Greenland. The magnetic pole is shown by a cross in northern Canada. The maximal zone within which the aurora can be observed more than 100 nights per year is located roughly at 67° in the European sector. (From Fritz, 1881.)

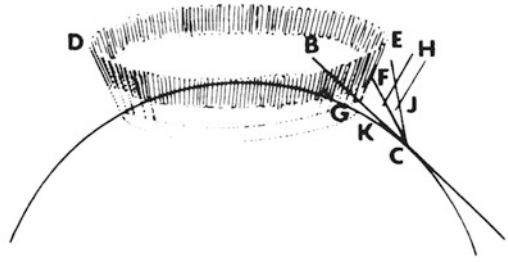


Figure 7.9. Schematic illustration to show the concept of triangulation of auroral heights from one point only, assuming the auroral arc as part of a complete circumpolar ring. (From Hansteen, 1827.)

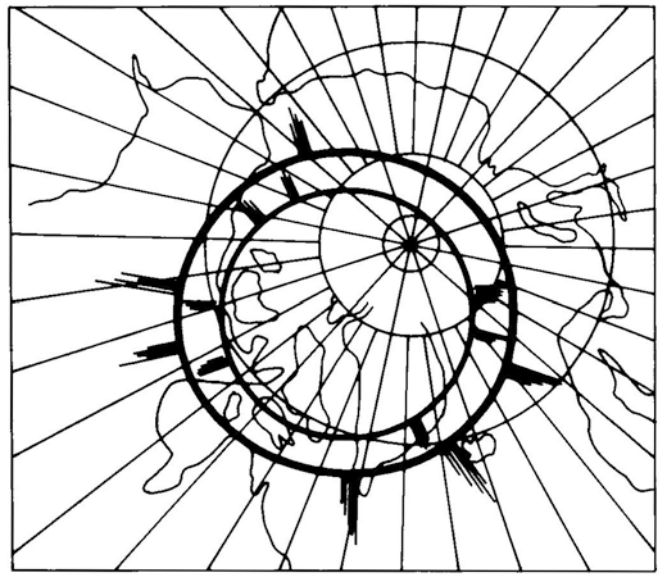


Figure 7.10. Illustration from the book, *The Vega Expedition*, by A. E. Nordenskiöld, showing the auroral ring centered at the geomagnetic pole as defined by Gauss. (After Nordenskiöld, 1880–81.)

the aurora was a complete ring at a fixed height encircling the pole, observations from one point of the elevation and azimuth angles to the apex of the auroral arc together with the azimuth angles to points where the arc went below the horizon were in principle sufficient to derive the height of this ring if the radius of the Earth was known. In 1881, however, Nordenskiöld argued that the center of the ring was not the geographic pole but rather the geomagnetic pole, at that time determined by Gauss to be in the northern corner of Greenland. Figure 7.10 is an illustration representing the concept of the auroral ring as redrawn from Nordenskiöld's book, *The Vega Expedition*.

At the start of the 20th century when Birkeland set up his laboratory experiments to prove his theory that the aurora was caused by cathode rays (electrons) streaming out of the Sun and caught by the Earth's magnetic field, he was able to produce two luminous rings around terrella poles, which supported the old concept of auroral rings (Figure 7.11).

During the International Geophysical Year 1956–1957 this theory was put to the test when a large array of all-sky cameras was installed in the auroral zones of Arctic and Antarctic regions. When analyzing all-sky films from this network of stations, Feldstein was able to show (Figure 7.12) that the aurora at any instant indeed forms an annular belt around the geomagnetic pole, which was to become known as the *auroral oval*.

This oval has a fixed position with respect to the Sun. It is compressed somewhat on the sunward side where it reaches down to approximately 78° geo-

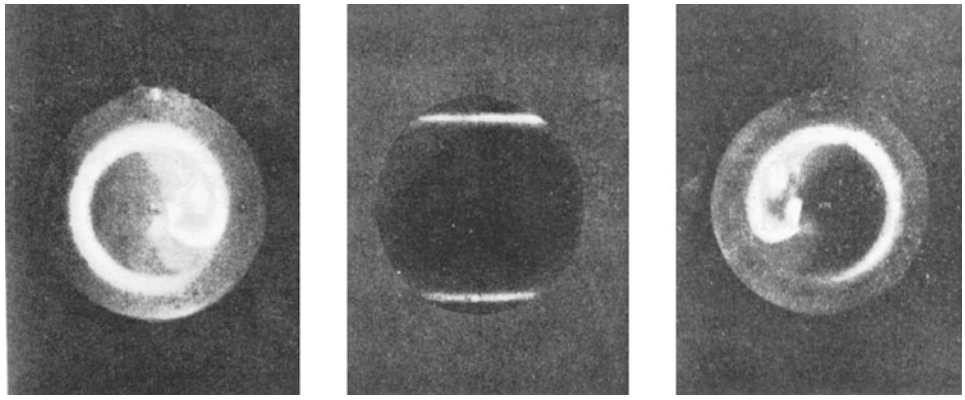


Figure 7.11. Examples from Birkeland's terrella experiments showing the two luminous rings around a magnetized fluorescent sphere bombarded by cathode rays or electrons, thus supporting his own theory of the aurora. (From Birkeland, 1913.)

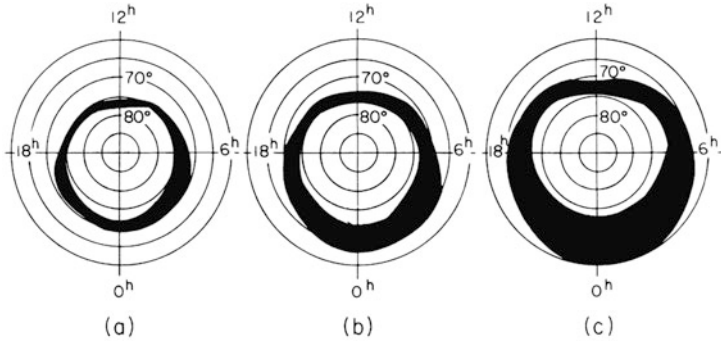


Figure 7.12. The auroral oval for different levels of disturbance: (a) under quiet conditions; (b) under medium disturbance conditions; and (c) under strongly disturbed conditions. (From Starkov and Feldstein, 1967.)

magnetic latitude on average, and is stretched out toward the antisunward side where it reaches down to approximately 67° geomagnetic latitude on average. As a station on Earth at auroral latitudes rotates around the geographic pole in the course of a day, it can be partly inside, partly poleward, and partly equatorward of the oval. A station at very high latitude, however, may be poleward of the oval throughout the whole day.

This oval, however, is not static as it expands and contracts with respect to the situation on the Sun and the potential distribution across the polar cap, the dawn–dusk potential. Figure 7.13 shows variation of the polar cap area inside the auroral oval as a function of this dawn–dusk potential indicated by the equivalent solar wind electric field. When the potential grows, the oval expands to lower latitudes as it also widens in latitudinal width.

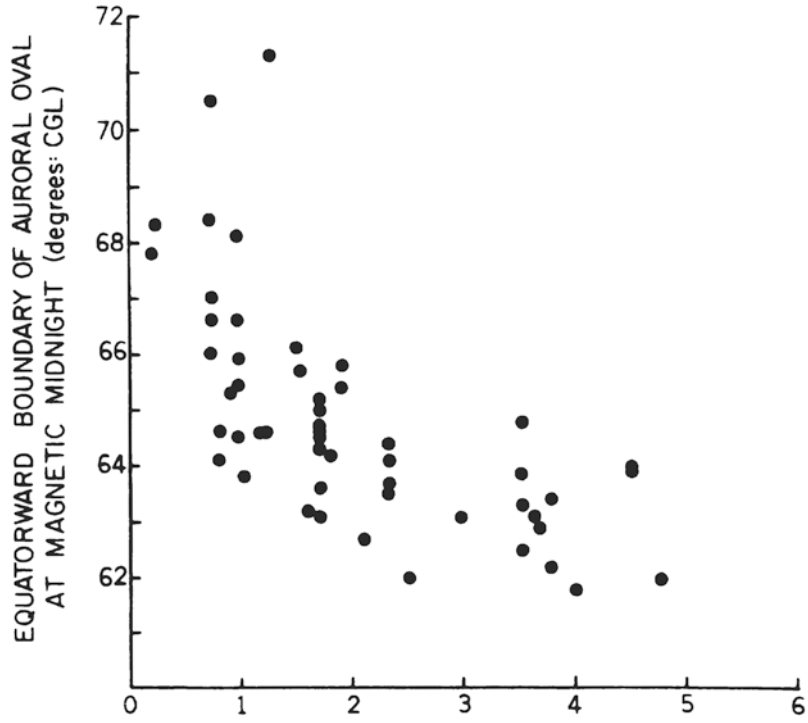


Figure 7.13. Location of the equatorward border of the auroral oval versus the maximum 30 min average $\mathbf{B}_S \cdot \mathbf{V}_S$ value over the six hours preceding the time of each auroral observation. $\mathbf{B}_S$ and $\mathbf{V}_S$ are the solar wind magnetic field and velocity, respectively. (From Kamide, 1988.)

In recent years it has been possible to image this oval in its total extent from space (Figure 7.14 gives a few examples of the oval). It is becoming evident from these images that the oval is always present although it changes shape and may sometimes be bifurcated by auroras crossing the polar cap and sometimes appears as a spoked ring when torches of aurora stretch toward the pole from the main luminous oval. Furthermore, it is important to notice that the aurora is not only a night-time phenomenon, as we may think when watching it by eye from the ground, but it is also always active in the day (as observed in the ultraviolet from space as these figures show).

7.5 THE AURORAL APPEARANCE

As already mentioned, when the aurora is observed at lower latitudes (i.e., when the oval has expanded equatorward), it appears red. Consequently, the phenomenon came to be called the "northern dawn" or aurora borealis named for the Roman goddess of the dawn. At auroral latitudes the phenomenon exposes a mul-

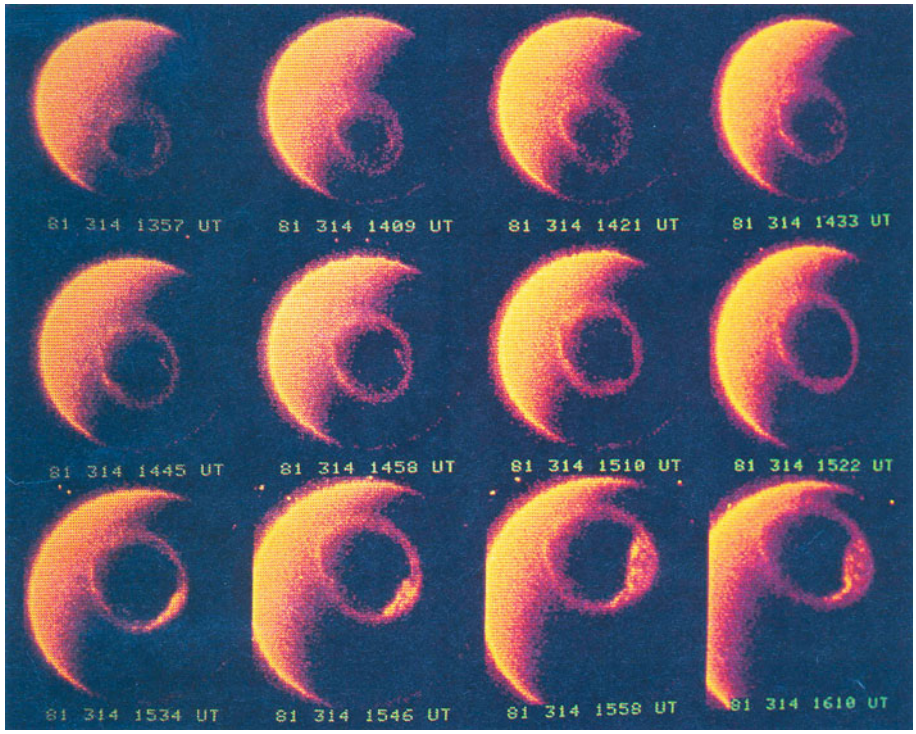


Figure 7.14. A series of images of the auroral oval prior to and during an isolated modest auroral substorm on November 10, 1981. The dayglow and auroral oval are seen primarily in the emissions of atomic oxygen at 130.4 nm. (From Frank and Craven, 1988.)

titude of colors, often mingled together in such a drapery of beauty that makes it hard to believe the aurora seen from Madrid is the same as the one seen from Tromsø.

Auroral spectroscopy was introduced by Ångström. In 1866 he observed the most prominent auroral emission, the yellow–green emission, and determined its wavelength to be 5,577 units, the unit became known as the Ångström at a later time. However, it took nearly 60 years before an explanation for this emission was found. Then, it was revealed that it was due to a transition in atomic oxygen from the metastable excited state ^1S to the lower ^1D state. It is still a puzzle why this emission is so strong in the aurora, because the metastable state makes it very vulnerable to collisions, which are very frequent at auroral heights. The green line emission is centered around 100 km.

Auroral spectra have been obtained in great numbers by many different observers. Because spectra are extremely variable as a result of different auroral conditions, the study of auroral spectra has developed into a scientific field of its own. The different emissions are due to different excited atmospheric atoms, molecules, and ions. Careful studies of the intensities and height variations of these

emissions have considerably increased our knowledge of the composition and temperature in the upper atmosphere above 90 km. As a matter of fact, before the 1960s when rockets and incoherent scatter radars were brought into use in ionospheric science, our knowledge of the composition of the upper polar atmosphere resulted mainly from optical auroral research.

Figure 7.15 gives some examples of typical spectra for medium to bright auroras in the visible part of the spectrum. The features observed in the spectrum are almost all due to lines and bands of neutral or ionized N_2, O, O_2, and N, roughly in the same order of importance. Some of the most outstanding lines and bands are indicated in the spectrum such as the oxygen emissions at 5,577 Å (557.7 nm) and 6,300 Å (630.0 nm) and the N_2^+ 1N band with a maximum at 4,278 Å (427.8 nm). Some of the most characteristic emissions together with their emitting source are listed in Table 7.1.

Therefore, auroral emissions are due to transitions between energy states in atoms and molecules. We will not go into detail here about the many possibilities for transitions in atoms and especially molecules, but will only indicate that the wavelength of light emission as a result of transition between two energy states, E_1 and E_2, can be expressed as

$$\lambda = \frac{hc}{E_2 - E_1}$$

In the reverse direction, light emission with this wavelength can be absorbed by an atom or a molecule to leave the particle in an excited state ($E_2 > E_1$). For the atmospheric species of interest here, we have seen that the ionization potential is of the order of $E_I = 15$ eV. Therefore, different emissions in the aurora are usually associated with wavelengths larger than the one corresponding to this energy:

$$\lambda_I = \frac{hc}{E_I} = 82.4 \text{ nm}$$

Figure 7.16(b) shows the energy levels of different auroral emissions from O, N_2, and N_2^+. The characteristic auroral green line is found to correspond to an upper level with energy of about 4.2 eV, while the oxygen red line at 6,300 Å (630.0 nm) and the N_2^+ 1N emission at 4,278 Å (427.8 nm) correspond to upper states with energies close to 2 eV and 18.75 eV, respectively.

7.6 AURORAL PARTICLES

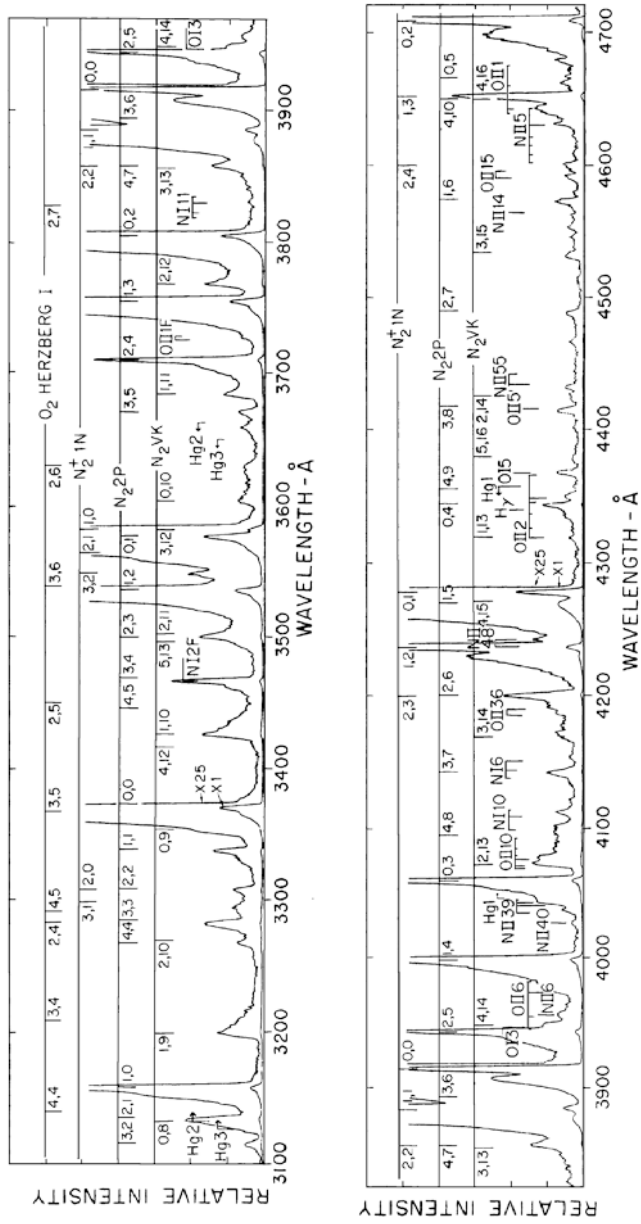
It has long been known that energetic electrons and protons penetrating the polar atmosphere from above are the ultimate cause of auroral emissions. Not until rockets equipped with instruments to catch incoming electrons and to measure their energies could be launched into auroras did the energy spectrum of precipitating electrons become known. Figure 7.17 gives a few examples of energy spectra of auroral electrons. The number of electrons with energies above a few tens of kiloelectronvolts are typically falling rapidly, while there is a tendency for a peak in the number of electrons at a few kiloelectronvolts. The number of electrons

Table 7.1. Characteristic auroral emissions and their emission source (upper state). The lifetime and dominant quenching particle for each upper state is also given.

Emission (Å)	*Upper state*	*Lifetime*	*Dominant quenching particle, Q*
λ5577	$O_I\ ^1S$	0.7 s	O_2
λ2972			O
$\lambda\lambda$6300–64	$O_I\ ^1D$	110 s	N_2
$\lambda\lambda$7319–30	$O_{II}\ ^2P$	5 s	N_2
$\lambda\lambda$3727–9	$O_{II}\ ^2D$	2.9 h	N_2
$\lambda\lambda$10395–404	$N_I\ ^2P$	12 s	N_2
λ3466			O_2
$\lambda\lambda$5199–201	$N_I\ ^2D$	26 h	O_2, O, NO
λ5755	$N_{II}\ ^1S$	0.9 s	O_2, O
λ6584	$N_{II}\ ^1D$	246 s	O_2, O
N_2(LBH)	$a^1\pi_g$	0.14 ms	
N_2(VK)	$A^3\Sigma_u^+$	~2 s	O
O_2(Atm.)	$b^1\Sigma_g^+$	12 s	N_2
O_2(IR Atm.)	$a^1\Delta_g$	60 mn	O_2
N_2^+(1N)	$B^2\Sigma$	70 ns	$N_2 + O_2$
N_2^+(M)	$A^2\pi$	14 μs	N_2
N_2(1P)	$B^3\Sigma$	6 μs	N_2
N_2(2P)	$C^3\pi$	50 ns	O_2
O_2^+(1N)	$b^4\Sigma$	1.2 μs	N_2

increases strongly towards lower energy, below, say, a couple of hundred electron-volts.

Based on such energy spectra of incoming particles it would be possible to derive the auroral emission spectra if all cross-sections for the various transitions between the numerous energy states of atoms and molecules in respect of these energies were known. Since this is extremely difficult to obtain in an exact manner, one has to depend on only a few typical cross-sections, some of which are shown in Figure 7.18 as a function of the energy of precipitating particles.


O2 HERZBERG I
N2+ 1N
N2 2P
N2 VK
Hg2
Hg3
X25
X1
RELATIVE INTENSITY
WAVELENGTH - Å
3100
3200
3300
3400
3500
3600
3700
3800
3900
4000
4100
4200
4300
4400
4500
4600
4700

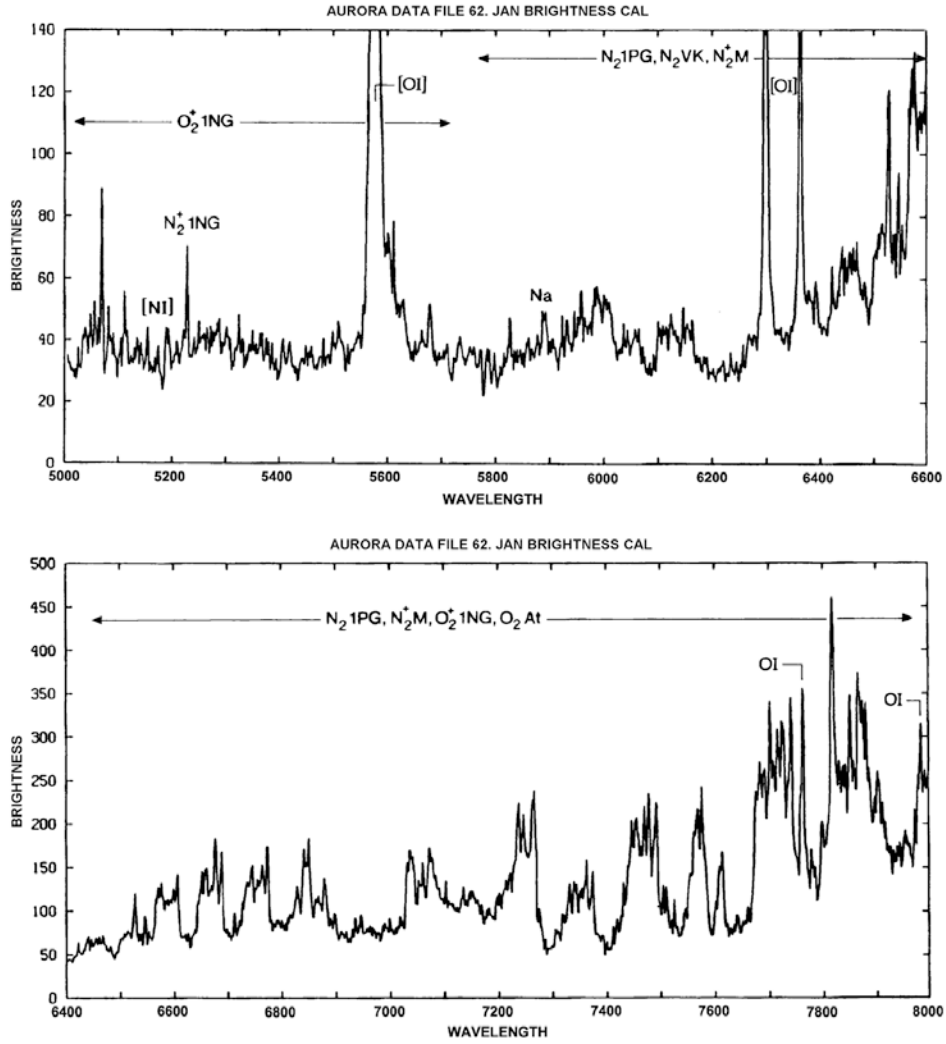
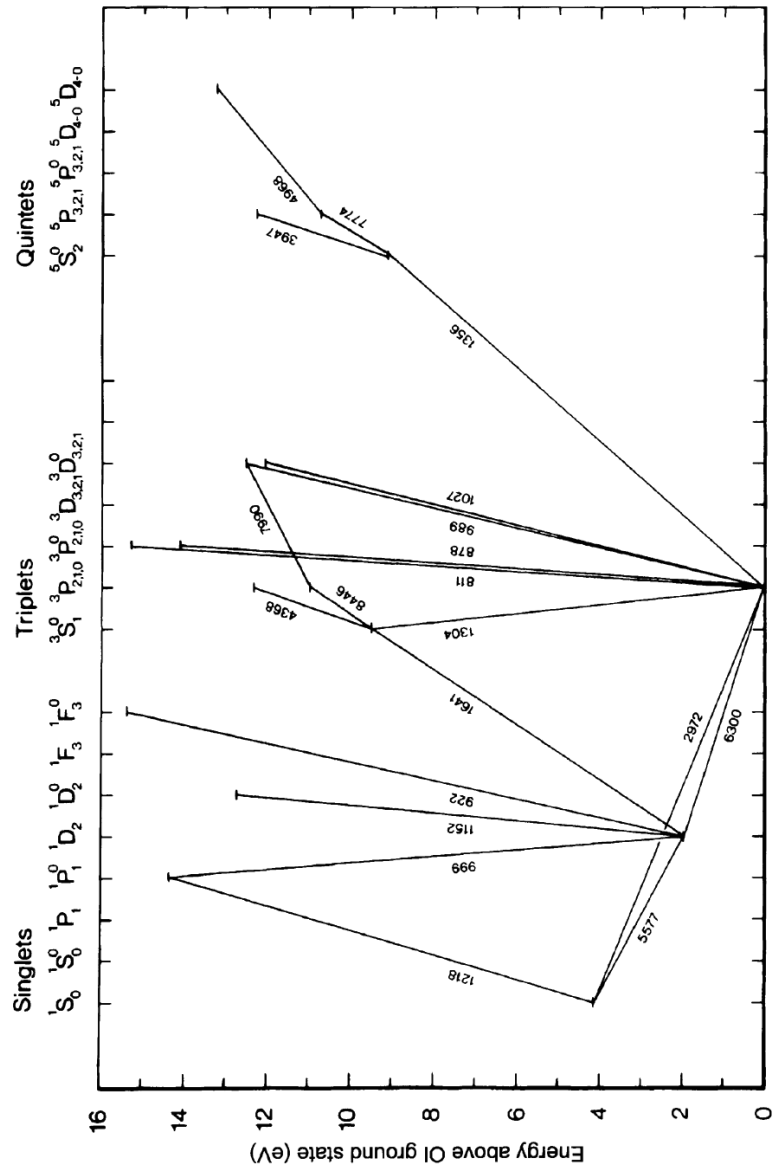


Figure 7.15. (Left page) Auroral spectrum between 3,100 and 4,700 Å (310.0–470.0 nm). The two different curves are at the gains shown (X_1 and X_{25}). The bands and atomic features present or possibly present are indicated. The ordinates are relative intensity uncorrected for atmospheric transmission. (From Vallance Jones, 1974.) (This page) Auroral spectrum in the visible wavelength region 5,000–7,000 Å (500.0–700.0 nm) showing a diversity of spectral emission features, some of which are indicated by their assumed sources. (From Rees, 1989.)

(a)


Singlets
Triplets
Quintets
$^1S_0$ $^1S_0^0$ $^1P_1$ $^1P_1^0$ $^1D_2$ $^1D_2^0$ $^1F_3$ $^1F_3^0$
$^3S_1^0$ $^3P_{2,1,0}$ $^3P_{2,1,0}^0$ $^3D_{3,2,1}$ $^3D_{3,2,1}^0$
$^5S_2^0$ $^5P_{3,2,1}$ $^5P_{3,2,1}^0$ $^5D_{4\text{-}0}$ $^5D_{4\text{-}0}^0$
1218
999
1152
922
1641
4368
8446
1304
7990
811
878
989
1027
3947
4963
7774
1356
5577
2972
6300
Energy above OI ground state (eV)
0
2
4
6
8
10
12
14
16

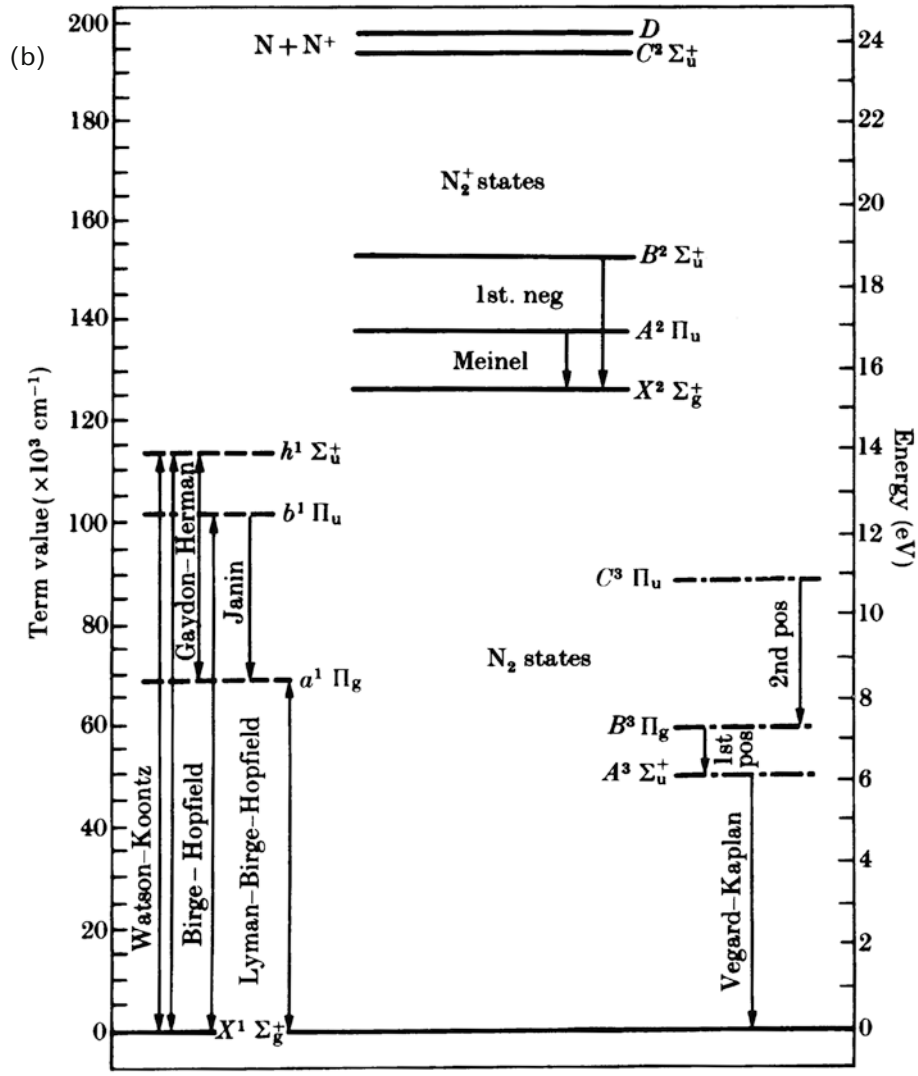


Figure 7.16. (Left page) Energy levels in atomic oxygen related to some of the most characteristic auroral emissions such as 6,300 and 5,577 Å (630.0 and 557.7 nm). (From Rees, 1989.) (This page, b) Energy levels in molecular nitrogen related to some of the auroral emissions such as 4,278 and 3,914 Å (427.8 and 391.4 nm). (From Akasofu and Chapman, 1972.)

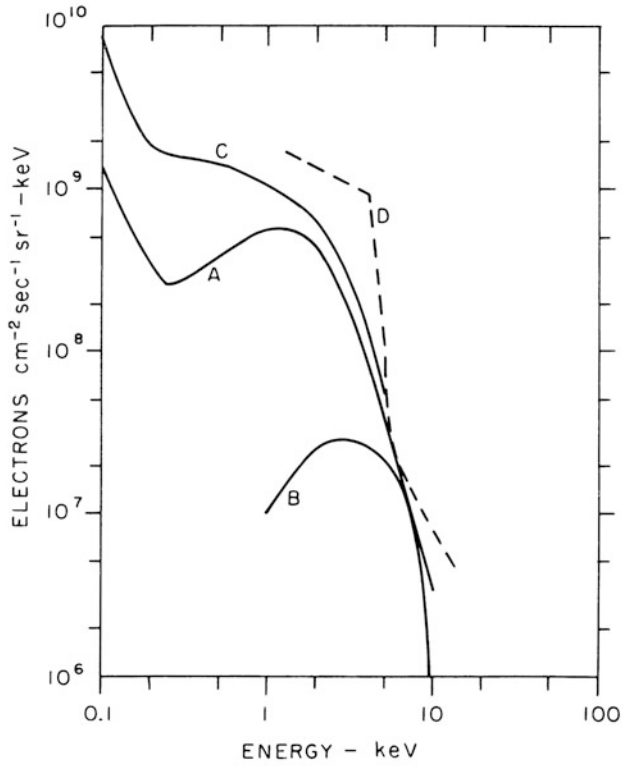
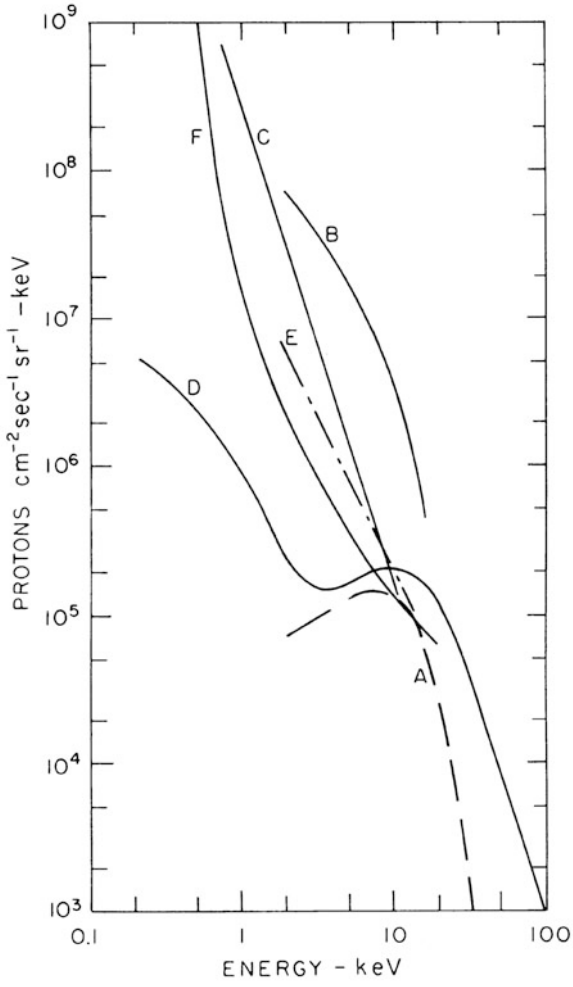


Figure 7.17. Energy spectra of precipitating particles related to different auroral forms. (This page) Electron spectra observed by satellites above the auroral oval: A, early evening auroral oval; B, early morning auroral oval; C, broad weak homogeneous arc; D, auroral oval under very disturbed conditions and strong aurora at midnight sector. (Facing page) Proton spectra observed in rocket flights over aurora: A, early evening proton aurora; B, medium to strong aurora; C, post-breakup aurora; D, recovery phase aurora; E, recovery phase aurora; F, breakup aurora. (From Vallance Jones, 1974.)


PROTONS cm⁻²sec⁻¹sr⁻¹-keV
ENERGY - keV
10⁹
10⁸
10⁷
10⁶
10⁵
10⁴
10³
0.1
1
10
100
F
C
B
E
D
A

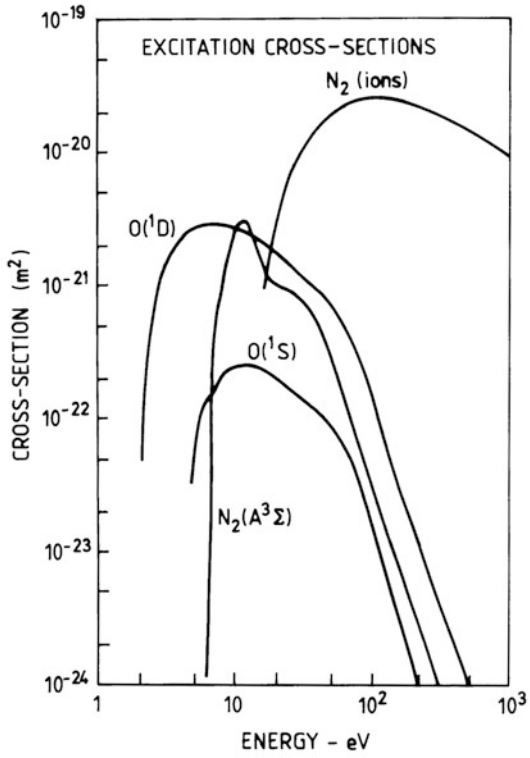


Figure 7.18. Excitation and ionization cross-sections for electron impact of some characteristic states of importance in aurora. (After Vallance Jones, 1974.)

We notice that the cross-sections are of the order of 10^{-24}–10^{-19} m^2 in the energy range between 1 eV and 1 keV. Table 7.2 presents the maximum cross-section as well as the typical energy at this maximum for a few important energy states of atomic and molecular species in the auroral atmosphere.

When comparing the energy at the maximum of the different cross-sections with the energy spectra of penetrating auroral particles, it becomes clear that there is a high-energy tail of particles which must be decelerated before they can become efficiently active in ionization and excitation processes. A 1 keV particle has to go through a whole series of secondary, tertiary, and higher order collisions before it finally has its energy reduced below the average 15 eV to excite an atom. During this process about 300 ion pairs have been formed as a result of one auroral particle with 1 keV energy at the top of the ionosphere. We notice here that the maximum cross-sections for the $O(^1S)$ and $O(^1D)$ states are 0.25×10^{-21} and 0.28×10^{-20} m^2, respectively, and that the maxima occur at roughly the same energy. Let us then assume a monoenergetic electron beam of intensity I_∞ at the top of the atmosphere, which will be degraded as it penetrates downward due to collisions with atomic oxygen. The maximum production of the two states will

Table 7.2. Representative data on cross-sections in the auroral atmosphere.

Excited state	σ_{max} (m^2)	E_{max} (eV)
O(^1S)	0.25×10^{-21}	10
O(^1D)	0.28×10^{-20}	5.6
N2($A^3\Sigma$)	0.28×10^{-20}	10
N2($B^3\pi$)	0.11×10^{-19}	12
O2($a^1\Delta$)	0.85×10^{-21}	6.5
O2($b^1\Sigma$)	0.20×10^{-21}	6
N_2^+ 1N(0,0)	0.17×10^{-20}	100
O_2^+ 1N(1,0)	0.43×10^{-21}	100

occur when (see Section 4.2)

$$\sigma \cdot n_m \cdot H = 1$$

where σ is the cross-section, n_m the density of target atoms at maximum, and H the scale height. When assuming H to be constant for the density of target atoms, then

$$n = n_0 \exp(-z/H)$$

where n_0 is the density of target atoms at $z = 0$, which is a reference height. The height of maximum production is then

$$z_m = H \ln(\sigma \cdot n_0 H)$$

Since $\sigma(O^1D) = 10 \cdot \sigma(O^1S)$, we find that

$$z_m(O^1D) = z_m(O^1S) + H \ln 10$$

The scale height at about 100 km, where the production of the ^1S state has a maximum, is about 7.0 km. The production maximum of the O(^1D) state should then be about 16 km higher according to this simple calculation. In reality, the height difference is greater than this.

7.7 PRECIPITATION PATTERNS OF AURORAL PARTICLES

There are other ways in which auroral particles can be mapped in addition to rocket probes launched through auroras. During the deceleration process X-rays are formed in the auroral ionosphere which can be observed from balloons

at lower heights. These X-ray emissions then give information on precipitation above the balloon height. From particle detectors in satellites the particles can be mapped before they hit the atmosphere below, and from the energy and direction of particle movement the precipitation zone can be deduced.

From the vast datasets now available on auroral particle precipitation it has been found, in addition to the visible auroral oval, that there are other zones of distinct precipitation (Figure 7.19). The central region of the average oval appears to correspond to the poleward boundary of trapped electrons with energies greater than 40 keV. This tends to be an important clue with respect to the origin of the auroral particles, since trapped electrons have to be on closed field lines. The central region of the average oval therefore appears to be the demarcation between open and closed field lines. Equatorward of the auroral oval on the late morning side and partly covering the oval at the early morning sector is a pattern of precipitating electrons with energies higher than 21 keV. On the equatorward side of the oval in the evening sector there is a zone of precipitating protons with energies higher than 4 keV. The auroral oval in itself appears to be closely related to the precipitation zone of electrons carrying energies less than 21 keV.

We have demonstrated that the energy spectra of auroral electrons as measured by rockets have a strongly decaying slope with increasing energy and that they sometimes display an enhancement at a few kiloelectronvolts. One can imagine these spectra as composed of two parts, where one forms the background spectrum, fairly well represented by an exponential or a power law distribution. The other part, however, represents a quasi-monoenergetic beam with a beam width of a few kiloelectronvolts. It is evident that these spectra must be replicas of energy source processes. This is a problem that engenders great interest among auroral scientists searching for the ultimate source of auroral particles.

Figure 7.20 compares an electron spectrum observed in the dayside auroral oval and a similar spectrum observed in the magnetosheet by a satellite at $L = 20$. The scales are in relative units to enhance the similarity between them. This indicates the magnetosheet is a likely candidate as a reservoir for auroral electrons.

7.8 ENERGY DEPOSITION PROFILES OF AURORAL PARTICLES

We have already discussed the stopping cross-section (Section 4.2) for a charged particle penetrating the atmosphere and the distance it has to penetrate before it is thermalized (expressed as the range). The stopping cross-section for protons and electrons in O, O_2, and N_2 is illustrated in Figure 7.21, which is derived partly from theory and partly from laboratory experiments.

Based on such a cross-section, a neutral atmosphere model, and monoenergetic beams of auroral particles penetrating the atmosphere from above, it is in principle possible to derive the energy deposition profile for each individual energy. Figure 7.22 demonstrates such profiles derived for energies between 2 and 20 keV. As expected, the peak of the deposition rate profile occurs at a lower height when energy increases. The peak energy deposited per unit length in the

Table 7.3. Height (H_m) at which energy deposition peaks, and the maximum energy deposition (ε_l) per unit length according to the energy (ε_0) of the precipitating particle.

ε_0 (keV)	H (km)	ε_l (eV m^{-1})
20	97	2.22
10	105	0.97
5	115	0.34
2	136	0.07

atmosphere, however, increases with energy. Table 7.3 gives the maximum energy deposition per unit length and the height of this maximum as a function of the initial energies used in Figure 7.22. As seen, a 10 keV particle has a maximum deposition rate at about 105 km. Since the aurora is found to occur most frequently in this height region, one can conclude—also based on rocket measurements of auroral particles, of course—that the typical energy of these particles is below 10 keV.

In this context the range–energy relation $R(\varepsilon_0)$ is often introduced. $R(\varepsilon_0)$ gives the depth of penetration in a particular medium as a function of the incident particle energy ε_0. The unit of R is typically given in g/cm^2. The important quantity is not the particle path length but rather the matter traversed by the particle along its path, and this is given by:

$$N(l) = \int_0^l n(s)\, ds$$

where l is the point of interest, and $n(s)$ is the particle density of the matter along the path. For vertical incidence the corresponding function will be

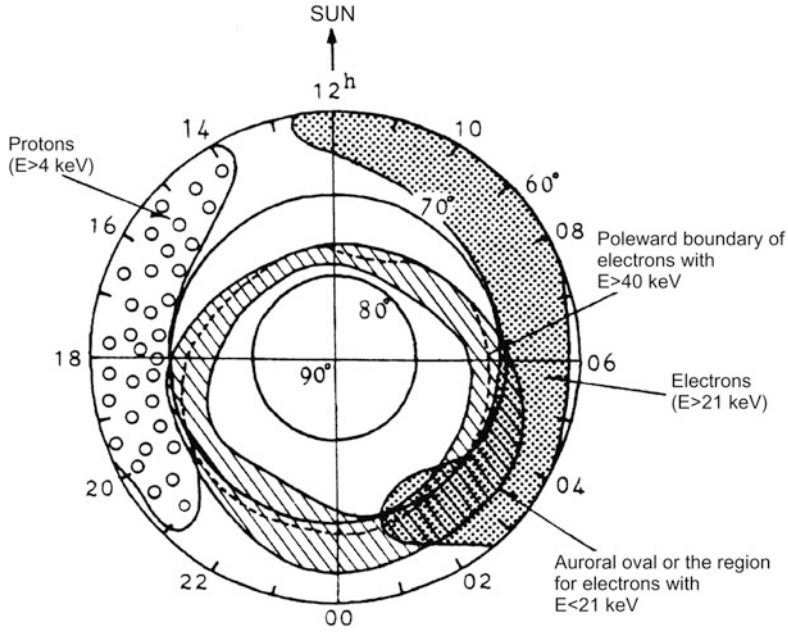
$$N(h) = \int_h^\infty n(z)\, dz$$

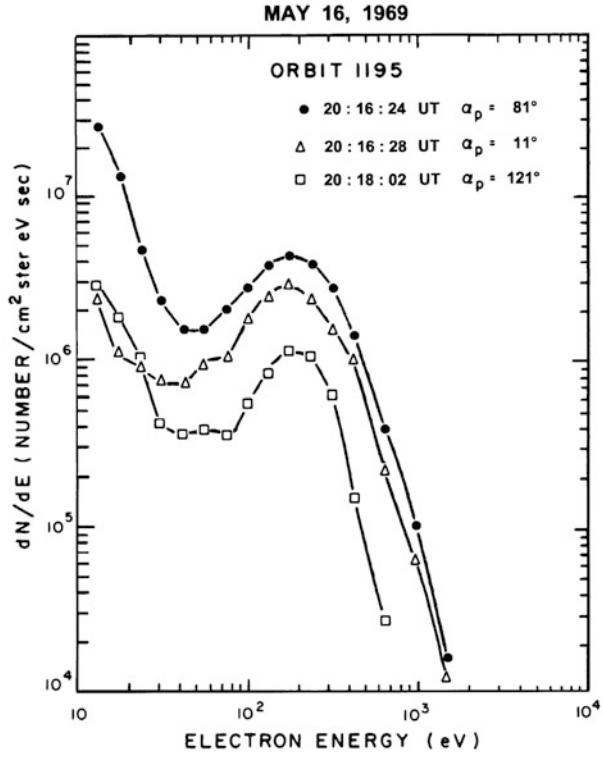
where h is the height of interest. For protons in air, the range–energy relation is given by (Rees, 1989):

$$R(\varepsilon_0) = 5.05 \cdot 10^{-6} \varepsilon_0^{0.75} \text{ g cm}^{-2}$$

for ε_0 incident proton energy between 1 keV and 100 keV. The corresponding range–energy relation for incident electron energies between 200 eV and 50 keV is given by (Rees, 1989):

$$R(\varepsilon_0) = 4.30 \cdot 10^{-7} + 5.36 \cdot 10^{-6} \varepsilon_0^{-1.67} \text{ g cm}^{-2}$$


SUN
12h
14
10
16
08
18
06
20
04
22
02
00
70°
80°
90°
60°
Protons
(E>4 keV)
Poleward boundary of
electrons with
E>40 keV
Electrons
(E>21 keV)
Auroral oval or the region
for electrons with
E<21 keV


MAY 16, 1969
ORBIT 1195
20 : 16 : 24 UT α_p = 81°
20 : 16 : 28 UT α_p = 11°
20 : 18 : 02 UT α_p = 121°
dN/dE (NUMBER/cm² ster eV sec)
10^7
10^6
10^5
10^4
10
10^2
10^3
10^4
ELECTRON ENERGY (eV)

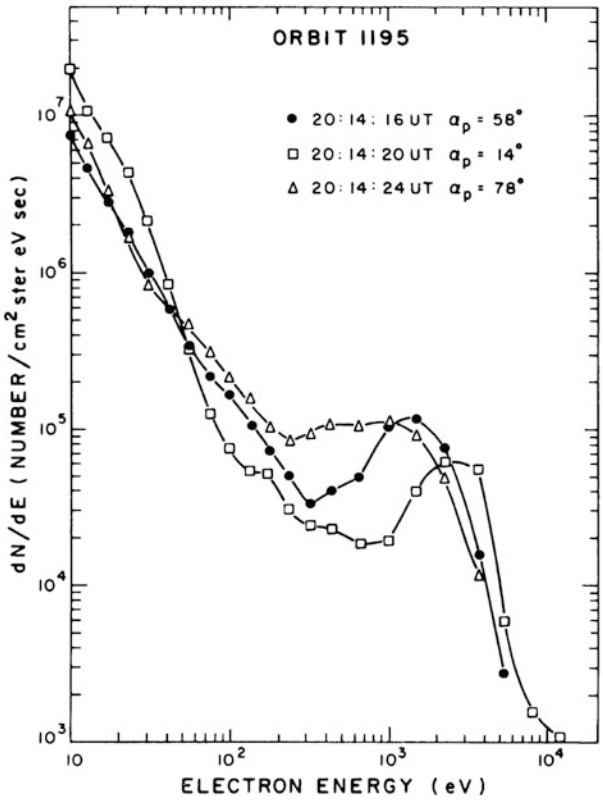


Figure 7.19. (a) Characteristic precipitation zones at high latitude. The typical auroral oval is associated with electrons with energies less than 21 keV. Equatorward of the oval on the evening side is the proton precipitation zone corresponding to protons with energies above 4 keV. On the morning side there is an equatorward zone of highly energetic electrons (>21 keV). The poleward boundary of trapped electrons with energies above 40 keV is illustrated by the dotted line. (From Tohmatsu, 1990). (b) Typical electron spectra observed in the polar cap at three different pitch angles from the ISIS-I satellite. (From Winningham and Heikkila, 1974.) (c) Typical electron spectra observed in the dayside oval at three different pitch angles from the ISIS-I satellite. (From Winningham and Heikkila, 1974.)

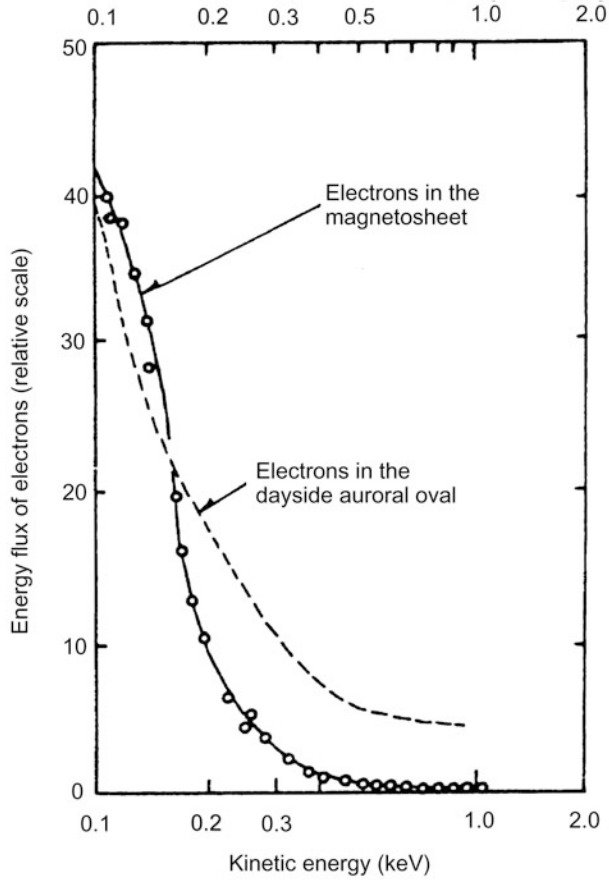


Figure 7.20. Energy spectrum from the dayside auroral oval of auroral particles related to the energy spectrum of electrons in the magnetosheet at 19.8 R_e. (From Sharp and Johnson, 1974.)

When the incident energy ε_0 is known, $N(h)$ is calculated by adjusting the integration limit h until the stopping altitude h_s is found, given by:

$$N(h_s) = R(\varepsilon_0)$$

The corresponding stopping altitudes for typical electron and proton incident energies are shown between 40 and 130 km in Figure 7.23.

By now applying a library of all the individual cross-sections for ionization in the auroral atmosphere one can derive ionization profiles for monoenergetic particle beams and composite particle spectra. Figure 7.24 shows typical ionization profiles derived for monoenergetic electron beams with energies between 2 and 100 keV normalized to the incident electrons per cm^2 and presented in units of ion pairs per cm^3 s. The shape of the profiles is very similar to the profiles for the energy deposition rate indicating that ionization per energy is constant and inde-

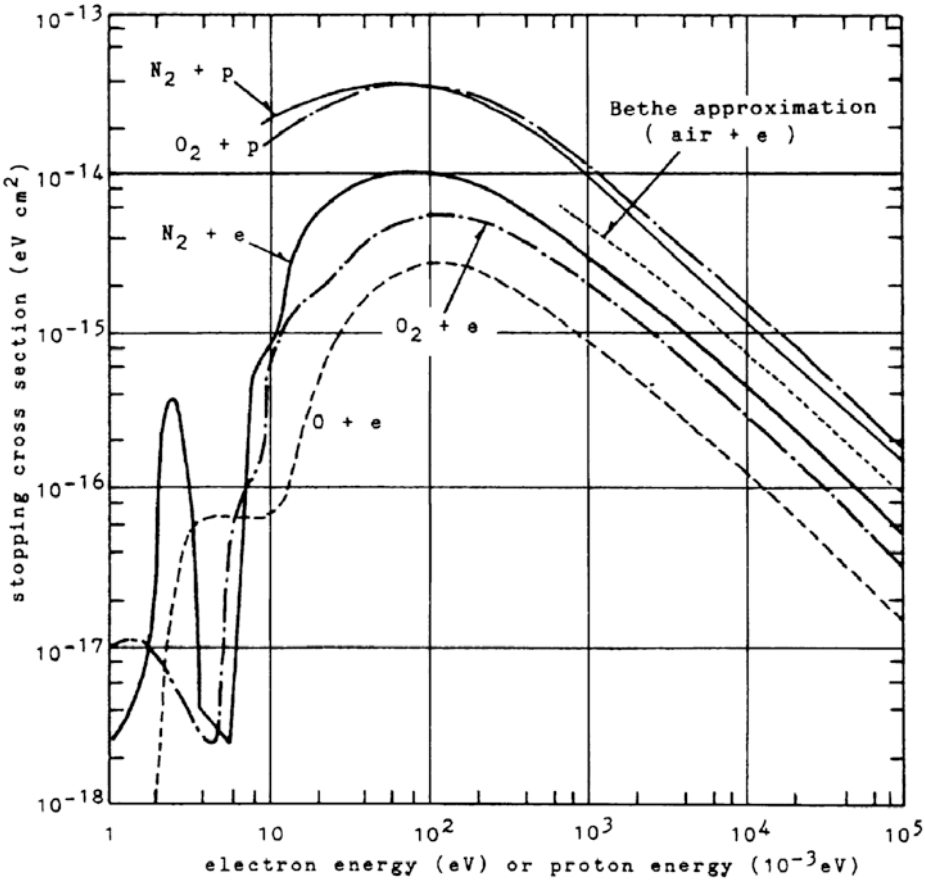


Figure 7.21. Stopping cross-sections of electrons and protons in N_2, O_2, and O. (From Tohmatsu, 1990.)

pendent of height. Of course, electron beams are not monoenergetic, but having developed a library of such ionization profiles for a variety of energies within the energy range of interest, composed energy spectra can be used either from measurements or theoretically and more realistic ionization profiles can be derived. Figure 7.25 shows two ionization profiles derived from such composed spectra represented by the Maxwellian function

$$\phi = \phi_0 \varepsilon \exp^{-\varepsilon/\varepsilon_0}$$

where ε_0 is the characteristic energy, and ϕ_0 is the normalized flux. The characteristic energies used are 1.0 and 5.0 keV, respectively. These profiles are compared with the ionization profiles derived from monoenergetic electron beams with 0.56 and 10 keV, respectively. Composite spectra produce a broader profile with respect

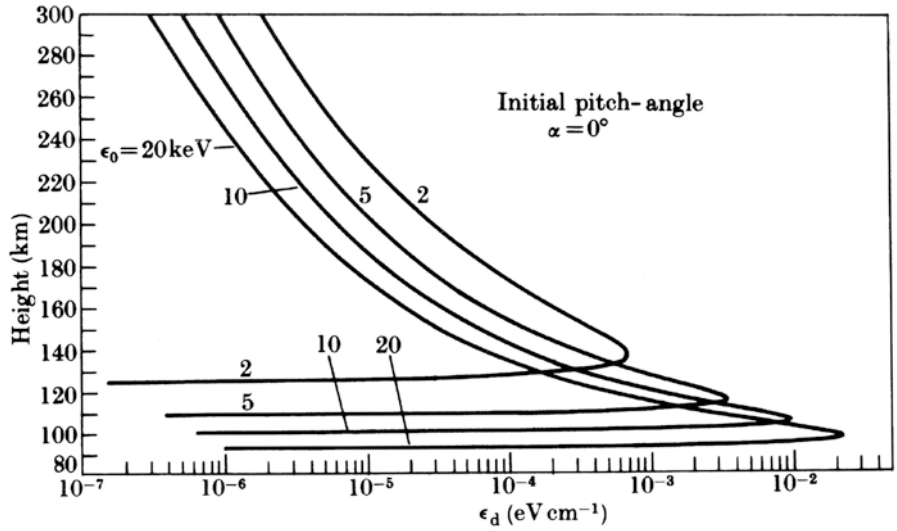


Figure 7.22. The height profile of energy deposition by electrons in the CIRA (1965) mean atmosphere normalized to one incident electron. (From Berger *et al.*, 1970.)

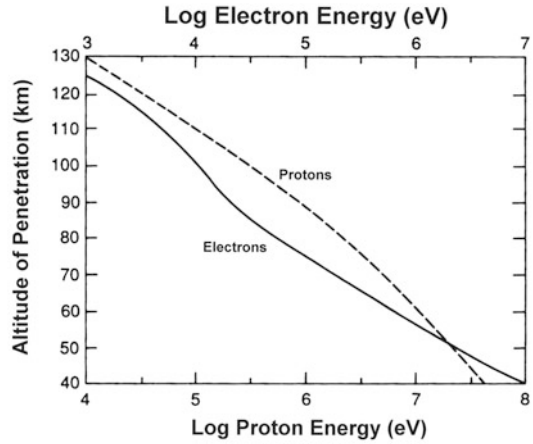


Figure 7.23. Stopping altitude for electrons (1–10^4 keV) and protons (10–10^5 keV) in the case of vertical incidence to the atmosphere. (From Luhmann, 1995.)

to height and, in particular, they reach down to lower altitudes due to the highly energetic tail in the spectrum.

Having then a library of excitation cross-sections for a number of reactions adequate for auroral processes involving primary and secondary electrons, it is possible to estimate production profiles as a function of height. In Figure 7.26 we show some excitation rate profiles in units of $cm^{-3}\ s^{-1}$ as a function of height for

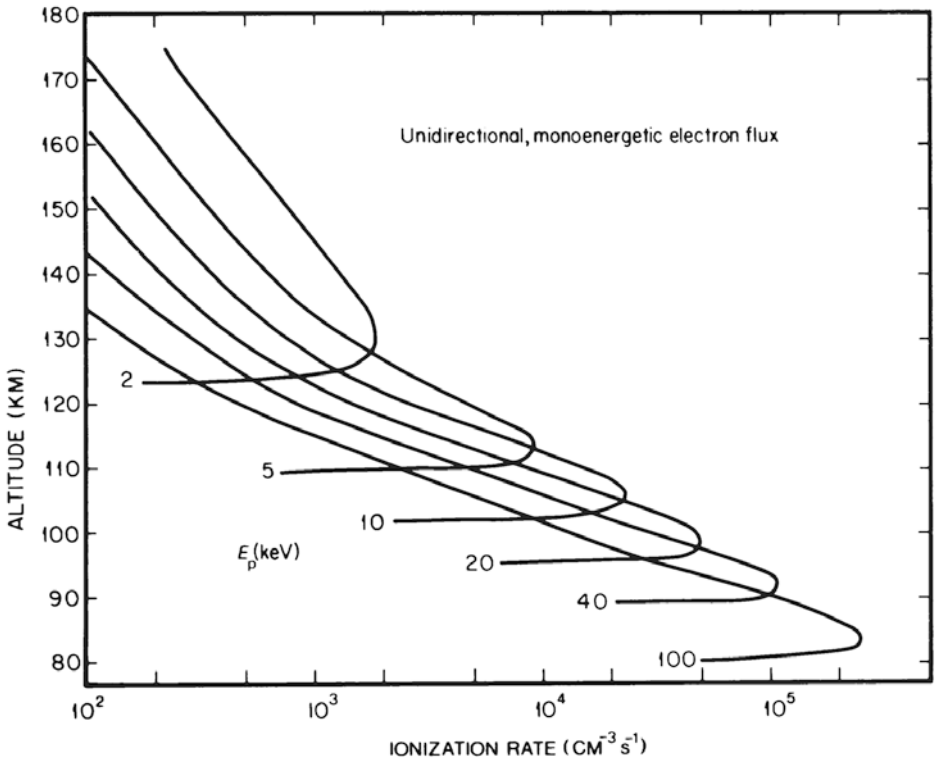


Figure 7.24. Ionization profiles in the Earth's atmosphere of unidirectional and monoenergetic electron beams. The labels on the curve indicate the energy in kiloelectronvolts. The ionization rate is given in ion pairs $cm^{-3}\ s^{-1}$. The primary electron flux is 10^8 erg $cm^{-2}\ s^{-1}$. (From Rees, 1989.)

the ^1D and ^1S states in atomic oxygen as well as the $B^2\Sigma_u^+$ state in molecular nitrogen, corresponding to the upper state of the 630.0, 557.7, and 427.8 nm emissions, respectively. The $N_2^+(B^2\Sigma_u^+)$ and the $O(^1S)$ states peak approximately at the same height, while the ionized nitrogen profile falls off very rapidly with height above the peak, due to the relatively strong decrease in the density of N_2 compared with the density of O above 150 km. The energy spectrum used in these calculations has an *e*-folding energy of 5 keV.

7.9 DERIVING ENERGY SPECTRA FROM ELECTRON DENSITY PROFILES

In the reverse sense it is possible to retrieve the energy spectrum of precipitating auroral particles if the ionization profile can be measured by, for example, a rocket probe or an incoherent scatter radar. If we assume that a steady state is

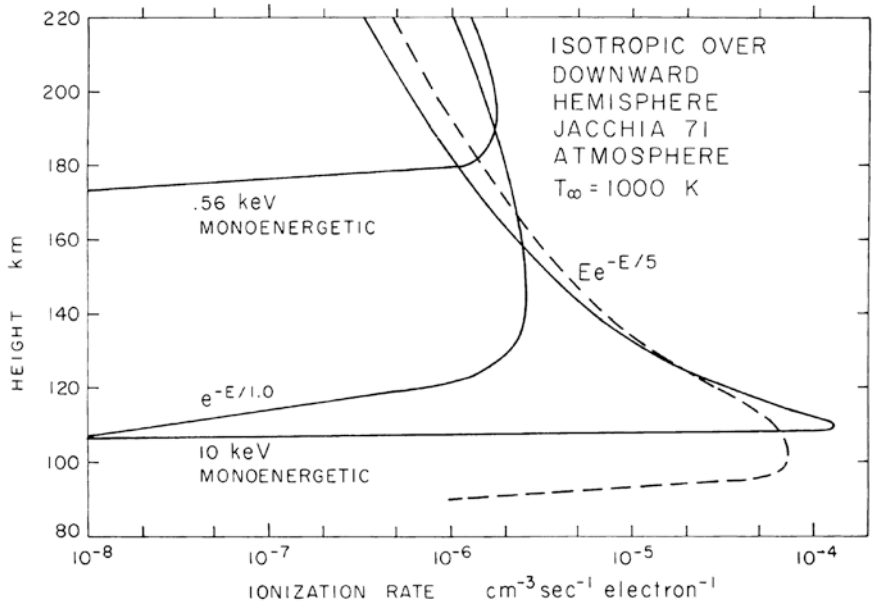


Figure 7.25. Calculated ion production rates for electrons of various initial energy spectra. E is in kiloelectronvolts. The particles are assumed to have an isotropic distribution over the downward hemisphere. (From Vallance Jones, 1974.)

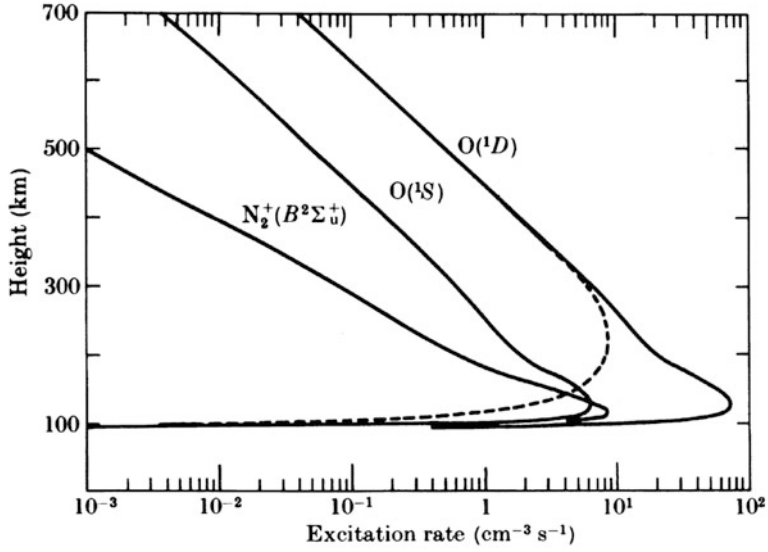


Figure 7.26. Height distribution of the rate of excitation of the $N_2^+(B^2\Sigma_u^+)$, $O(^1S)$, and $O(^1D)$ states for primary auroral electrons of spectral form $J_0 \exp(-\varepsilon/5000)$ in units of cm^2 s ster eV, where ε is in electronvolts. $J_0 = 10^3$ $(cm^2 \text{ ster s eV})^{-1}$. The dashed curve also takes deactivation into account. (From Kamiyama, 1966.)

valid, then the electron production at height z in the E-region is approximately given by (see Section 4.4):

$$q_0(z) = \alpha(z) \cdot (n_e(z))^2$$

where $\alpha(z)$ is the recombination coefficient, and $n_e(z)$ is the electron density as a function of height. An observed electron density profile therefore can, when transport terms are neglected and a steady state is assumed, yield the ion-pair production profile to a first approximation. By now applying a library of ionization profiles $q_i(\varepsilon_i, z)$ derived for a large variety of monoenergetic beams of energy ε_i, a combination of these profiles should yield the observed ion production profile. Then, assuming we can form a linear combination of $q_i(\varepsilon_i, z)$ at each discrete height z_j, we have

$$q(z_j) = \sum_{i=1}^{N} a_i \cdot q_i(\varepsilon_i, z_j)$$

where the energy range is ε_1–ε_N and the number of heights can be chosen arbitrarily. The factor a_i now represents the weight at which the different monoenergetic energy beams contribute to the total profile.

By applying a least squares fit to the deduced ionization profile $q_0(z_j)$, by means of the electron densities observed at the discrete heights z_j, we try to minimize the sum of squared differences:

$$\delta = \sum_{j=1}^{M} [q_0(z_j) - q(z_j)]^2 = \sum_{j=1}^{M} \left[q_0(z_j) - \sum_{i=1}^{N} a_i q_i(\varepsilon_i, z_j) \right]^2$$

where M is the number of heights between the lower height z_1 and the upper height z_M. By minimizing δ with respect to the coefficients $\{a_i\}$ we derive the spectrum. Figure 7.27 shows an electron density profile observed during an auroral event together with the energy spectrum of the electrons derived as just described. Also shown are the synthetic profiles derived from these calculated spectra. The spectra are rather similar to spectra observed in auroral forms directly by rocket probes. Therefore, measured electron density profiles can be used at least as a first approximation to derive the energy spectra of auroral particles from the ground in a quasi-continuous manner. The method can actually be improved by allowing for time variations in n_e. This is certainly more appropriate since there are rather dramatic time variations in n_e during auroral displays.

An auroral electron emits electromagnetic radiation when it is deflected by Coulomb collisions with atmospheric molecules. This radiation is called "bremsstrahlung" and represents a flat continuum in the spectral range ν given by

$$\nu < \frac{\varepsilon}{h}$$

where ε is the energy of the penetrating particle. The X-ray range of this spectrum is especially important since an X-ray photon at a given energy is able to reach lower altitudes than its parent charged particle. Therefore, bremsstrahlung X-rays can produce ionization much lower down in the atmosphere than the stopping

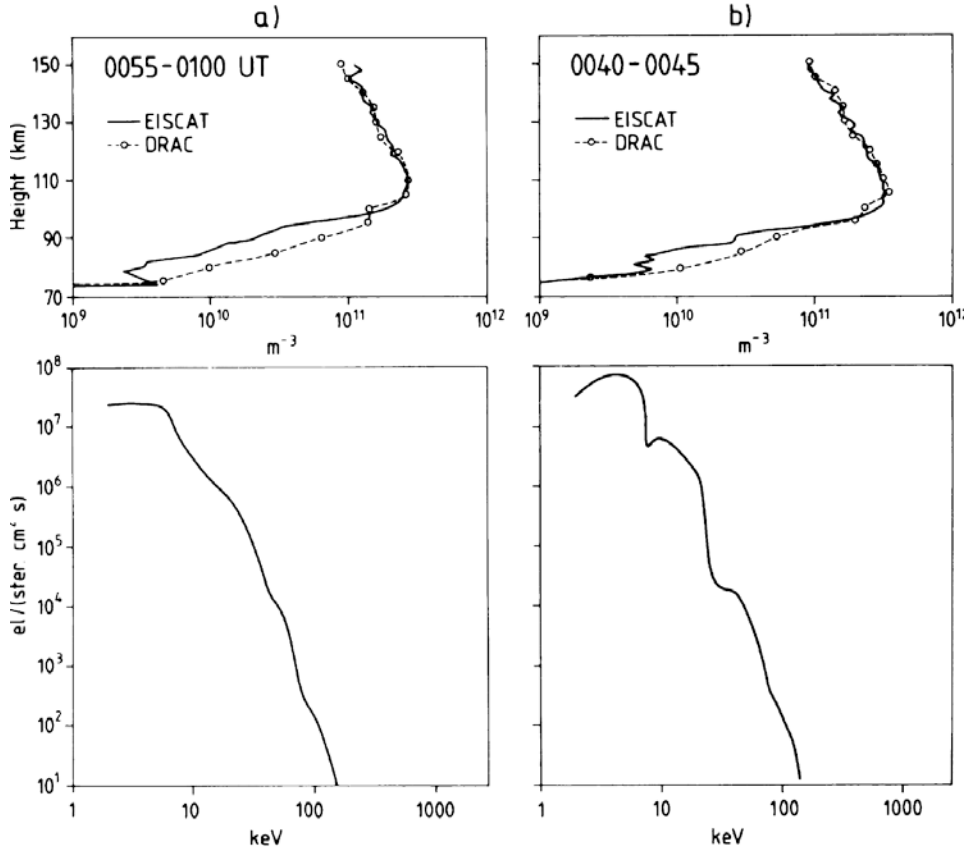


Figure 7.27. (Upper panels) Night-time electron density profiles observed by EISCAT on August 14, 1985 at 00:40 to 00:45 UT and 00:55 to 01:00 UT, respectively. Also included (dashed line) are synthetic profiles as calculated by using the corresponding electron spectra shown in the bottom panels. (Bottom panels) Energy spectra derived by following the procedure described in the text from the corresponding measured electron profiles in the upper panels. (From Brekke *et al.*, 1989.)

altitude of primary electrons. Figure 7.28 represents model calculations of the ionization profiles produced by penetrating electrons and their accompanying X-rays when the primary particles have an *e*-folding energy spectrum with a characteristic energy of 5 keV. While primary and secondary particles are stopped above 90 km where they ionize atmospheric constituents, X-rays may reach below 30 km where they create ionization that can be of importance for the fair weather electric circuit.

This means that auroral X-ray emissions can also be monitored at balloon heights. Balloons therefore represent an alternative platform to study spatial and temporal variations in auroral precipitation. Care has to be taken, however, since X-rays at balloon heights are mainly formed by the highly energetic tail of auroral

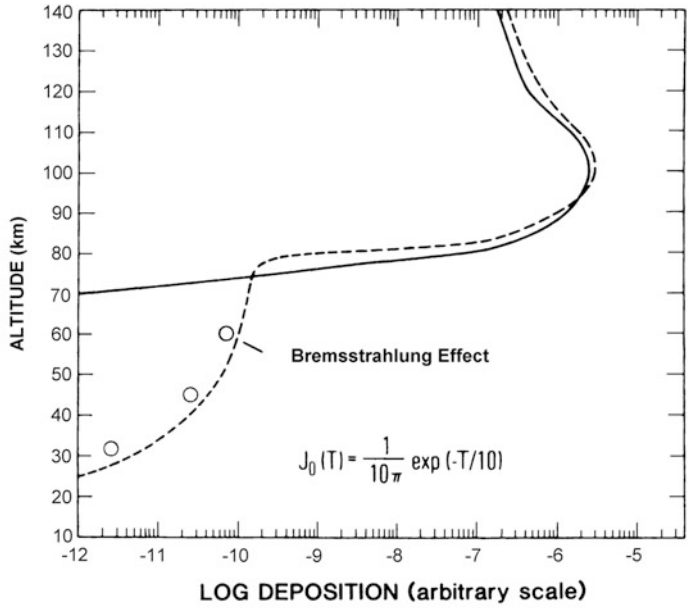


Figure 7.28. An energy deposition profile for an incident electron spectrum of the form indicated, including the contribution from absorbed bremsstrahlung photons. (From Luhmann, 1995.)

particles ($\varepsilon > 25$ keV) while auroras are mainly due to particles with energies less than 10 keV. It is not well known how particles in these different energy ranges are related to each other.

Since auroral X-rays contribute so strongly to ionization of the D-region, auroral events are often found to be related to fairly strong Galactic cosmic radio noise at 20–40 MHz.

7.10 EXCITATION PROCESSES IN THE AURORA

One of the best understood auroral emissions is related to excitation of the N_2^+ ion, more specifically to the $B^2\Sigma_u^+$ state which has a maximum cross-section of excitation close to 100 eV (see Figure 7.18). It is most often monitored by the (0, 0) or (0, 1) emission bands at 427.8 or 391.4 nm, respectively. Because they are spontaneous emissions, radiation occurs at the incidence of primary and secondary particles (within 10^{-7} s). Approximately 1 photon at 391.4 nm appears, according to laboratory experiments, to be produced per 50 ion pairs formed in the atmosphere by auroral particles. Therefore, the number of photons resulting from an incident electron with an initial energy ε_0 can be calculated to be

$$\eta(391.4\,\text{nm}) = 0.02 \cdot \frac{n(N_2)}{n} \cdot \frac{\varepsilon_0}{W}$$

where W is the mean ionization energy equal to about 35 eV, and $n(N_2)$ and n are the number densities of nitrogen molecules and the total number density in the atmosphere at the height of the emission, respectively. Since $n(N_2) \approx 0.80n$, we find that

$$\eta(391.4\text{ nm}) \approx 0.5 \cdot \varepsilon_0$$

where ε_0 is given in kiloelectronvolts. There is therefore a close relationship between the emission rate η of N_2^+ emissions and the energy ε_0 of the precipitating particle. Note that $\eta(391.4\text{ nm})/\eta(427.8\text{ nm}) = 3.3$.

The most predominant emission in the high-latitude aurora is the yellow–green line at 557.7 nm. The line is known to be due to the ^1S–^1D transition in atomic oxygen (see Figure 7.16), where the ^1S state has a lifetime against radiation of approximately 0.75 s. Therefore, the transition is so-called forbidden, leaving the oxygen atom long enough in the excited ^1S state for it to be quenched by collisions with other gas particles in the atmosphere. How the excitation occurs is still not clear, and much work has been contributed to the attempt to solve the problem.

The direct impact of secondary and higher order electrons produced during the stopping process of an incoming highly energetic auroral electron may be the source of this excitation. This process can be illustrated by the following reaction equation:

$$O(^3P) + e \longrightarrow O(^1S) + e \tag{7.1}$$

Another possible source for excitation of the $O(^1S)$ state is dissociative recombination with an O_2^+ ion:

$$O_2^+ + e \xrightarrow{k_{O_2^+}} O(^1S) + O \tag{7.2}$$

or energy transfer from an excited N_2 molecule as follows:

$$N_2(A^2\Sigma) + O \xrightarrow{k_{N_2}} O(^1S) + N_2 \tag{7.3}$$

These last two processes have a delay time with respect to the initial ionization and excitation process brought about by auroral primary and secondary particles. If the reaction rates for the two processes are $k_{O_2^+}$ and k_{N_2}, respectively, then the time constants involved will be

$$\tau_{O_2^+} = \frac{1}{k_{O_2^+} \cdot [O_2]}$$

$$\tau_{N_2} = \frac{1}{k_{N_2} \cdot [N_2]}$$

respectively. If these last two processes are effective at producing the $O(^1S)$ state, there should be a delay in the 557.7 nm emission compared with the 427.8 nm emission which is spontaneously emitted at the incidence of excitation by primary and secondary particles.

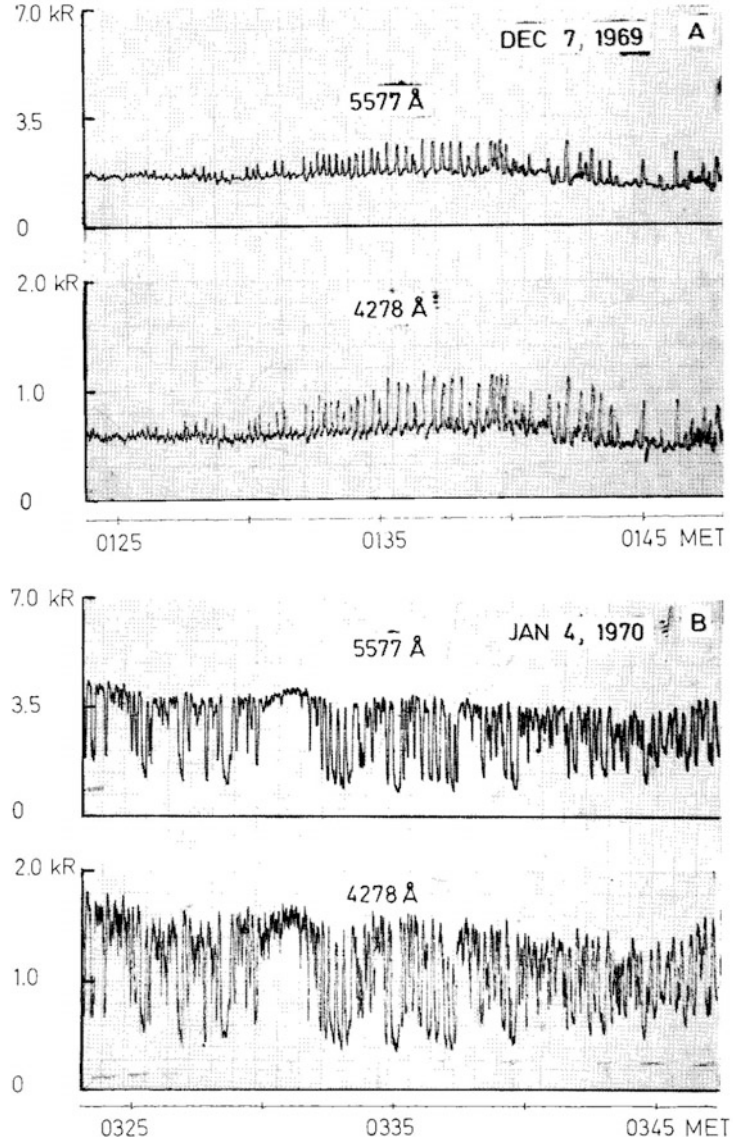


Figure 7.29. Intensity variation for two events of simultaneous pulsating aurora at 557.7 and 427.8 nm observed on December 7, 1969 and January 4, 1970 in Tromsø. The intensity in kiloRayleighs (kR) is indicated to the left and local time MET (Middle European Time) at the bottom of each record. The paper speed was 0.1 mm/s. (From Brekke and Henriksen, 1972.)

When comparing the 557.7 nm emission carefully with the 427.8 nm emission, for example, in so-called pulsating aurora (Figure 7.29) which often quasi-regularly displays sinusoidal variations in intensity, one often finds a phase shift between the two time series. This is shown in Figure 7.30 where the 427.8 nm

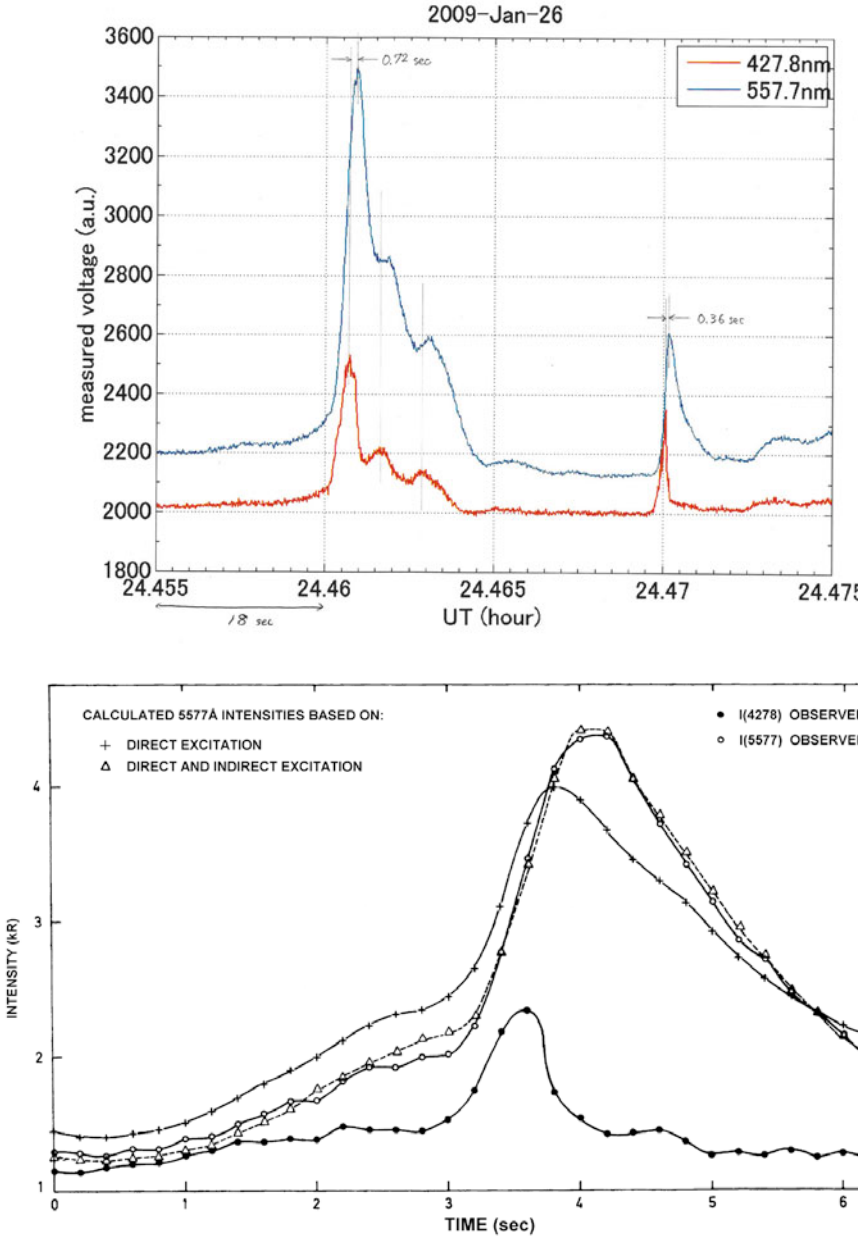


Figure 7.30. (Top) Intensity variations in the 557.7 and 427.8 nm emission in two pulses of a pulsating aurora showing a phase shift between the two emissions of 0.72 and 0.36 s, respectively. (Courtesy Oyama, 2010.). (Bottom) Intensity variations in the 557.7 and 427.8 nm emission in a pulse of a pulsating aurora showing a phase shift between the two emissions. Also indicated are theoretical estimates of the 557.7 nm emission depending on different excitation sources. (From Henriksen, 1974.)

emission leads the other in time. Whether this time delay is only due to the time constant against radiation of the $O(^1S)$ state or is a combination of that with delayed production due to recombination of O_2^+ and energy transfer of N_2 remains uncertain. All agents probably contribute.

Now, let the effective lifetime against radiation for the $O(^1S)$ state be τ. Then, time variation in the 557.7 nm emission ($I_O(t)$) can be expressed by:

$$\frac{dI_O(t)}{dt} + \frac{I_O(t)}{\tau} = K_1 I_N(t) \tag{7.4}$$

where $I_N(t)$ is the intensity of the 427.8 nm emission that is proportional to the production of excited states. Therefore, it is also proportional to the production of the $O(^1S)$ by the proportionality constant K_1, if direct electron impact is the only source (7.1). If the 427.8 nm emission now exhibits a sinusoidal variation, then we denote:

$$I_N(t) = I_{N_0} \exp(i\omega t) \tag{7.5}$$

Due to the phase lag of the 557.7 nm emission we express

$$I_O(t) = I_{O_0} \exp(i(\omega t - \varphi)) \tag{7.6}$$

and find:

$$\exp(i\varphi) = \frac{I_{O_0}}{K \cdot I_{N_0}} \left(\frac{1}{\tau} + i\omega\right)$$

which gives

$$\tan \varphi = \omega\tau \tag{7.7}$$

Given the frequency ω of the pulsations, it is therefore possible in principle to derive the effective lifetime τ of the upper state of the excited atom related to the 557.7 nm emission. If one performs a cross-spectral analysis of the time signals and derive φ for each frequency component, then there is a linear relationship between ω and $\tan \phi$ where the slope in the line is represented by the lifetime τ (Figure 7.31).

We notice from Figure 7.31 that the measured $\tan \varphi$ in (7.7) is not a linear function of ω and, therefore, it is believed that other mechanisms may take place in exciting $O(^1S)$ than the one given by (7.1); delayed processes like (7.2) or (7.3) may also take part. Let us therefore include another production term in (7.5) such that

$$\frac{dI_O(t)}{dt} + \frac{1}{\tau'} I_O(t) = K_1' I_N(t) + K_2 n(t) \tag{7.8}$$

where K_2 is a proportionality constant, the production of $n(t)$ is proportional to and in phase with the ionization and excitation of the N_2^+ ions emitting the 427.8 nm emission, and the number density of $n(t)$ decays with a time constant τ_1:

$$\frac{dn(t)}{dt} + \frac{1}{\tau_1} n(t) = K_3 I_N(t) \tag{7.9}$$

where K_3 is another proportionality constant.

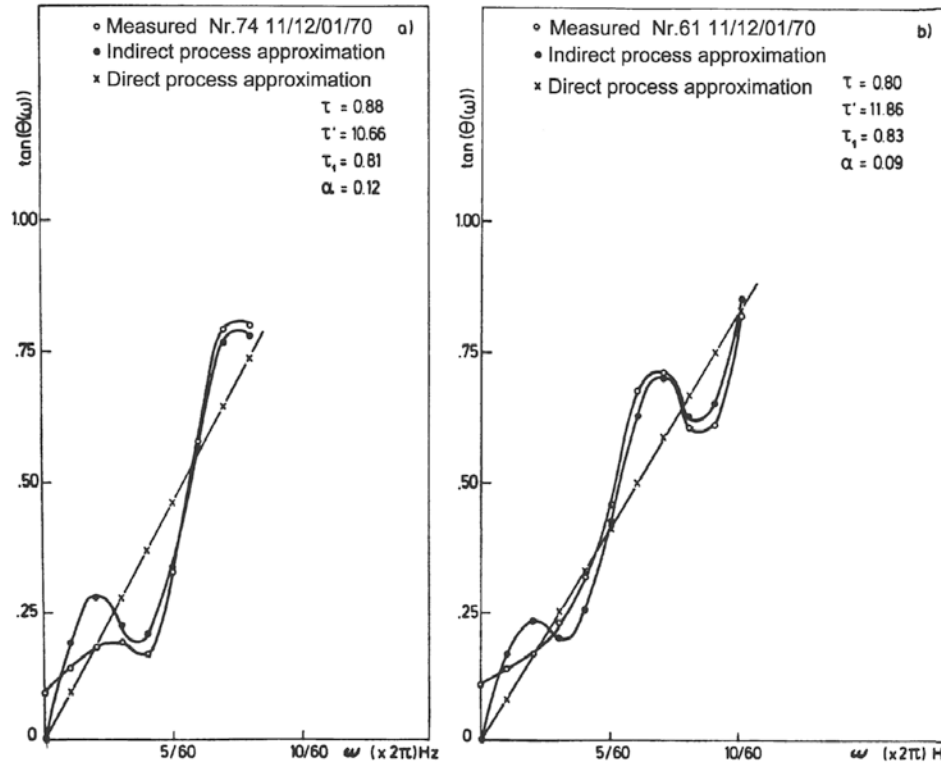


Figure 7.31. The tangent of the phase angle φ between the 557.7 nm and 427.8 nm emission as a function of the frequency ω as measured in pulsating aurora. (Derived from cross-spectral analysis.)

By now assuming the expressions for $I_N(t)$ and $I_O(t)$ given by (7.5) and (7.6), respectively, and that

$$n(t) = n_0 \exp(i(\omega t - \varphi_1))$$

where φ_1 is another phase angle, we can solve for equations (7.8) and (7.9) simultaneously and find:

$$\exp(i\varphi) = \frac{I_{O_0}}{K_1' I_{N_0}} \frac{\left(\frac{1}{\tau_0} + i\omega\right)\left(\frac{1}{\tau_1} + i\omega\right)}{\frac{1}{\tau_1} + \frac{K_2 K_3}{K_1'} + i\omega} \tag{7.10}$$

Figure 7.31 shows two curves of tan φ as a function of the frequency observed in a pulsating aurora where there is no linear relationship (as derived in (7.7)). Therefore, a least squares fit has been applied to (7.10) which appears to correspond better to the observed relationship than (7.7), indicating that there is a delayed excitation process (as introduced in (7.8) and (7.9)). It turns out that

(7.10) has two sets of solutions depending on the ratio between K' in (7.4) and K_1 in (7.8) in the following way:

(A) If $K_1' > K_1$ then $\tau' < \tau_1/(1 + \tau_1 K_2 K_3/K_1')$

(B) If $K_1' < K$ then $\tau' > \tau_1/(1 + \tau_1 K_2 K_3/K_1')$

which means that if $K_1' > K_1$ then the effective lifetimes of $O(^1S)$ atoms is always less than the decay time τ_1 and, in the reverse case, τ' may be greater than τ_1.

Figure 7.32 shows a set of the lifetimes of $O(^1S)$ atoms derived according to (7.4) and (7.8) where (7.9) is included. The inclusion of the delayed excitation process reduces the effective lifetimes estimated for $O(^1S)$ atoms. Figure 7.33 shows the derived values of τ_1 in the situation where $K_1' > K_1$ as a function of the derived values of τ'. It is seen that τ_1 is always larger than τ'. It has been indicated that this delayed excitation process can be process (7.2).

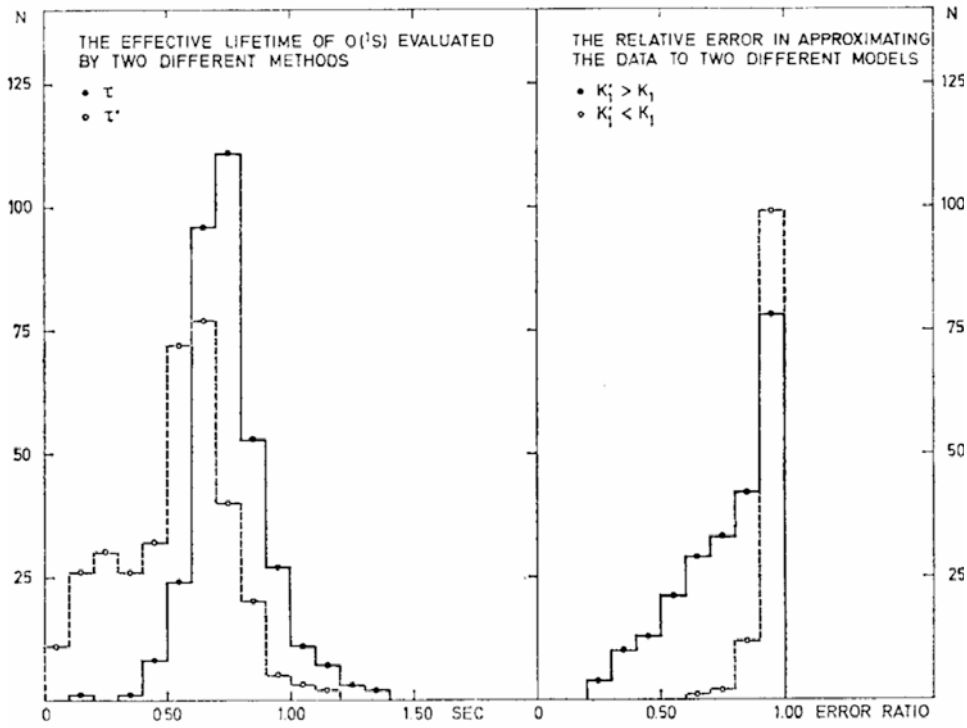


Figure 7.32. (Left) The number of measured effective lifetimes of $O(^1S)$ atoms. The dots and circles indicate results derived from equations (7.4) and (7.8), respectively. (Right) Histogram of the relative error in approximating the data to two different models. The dots and circles indicate results derived for two different cases of K_1'. (From Brekke and Pettersen, 1972.)

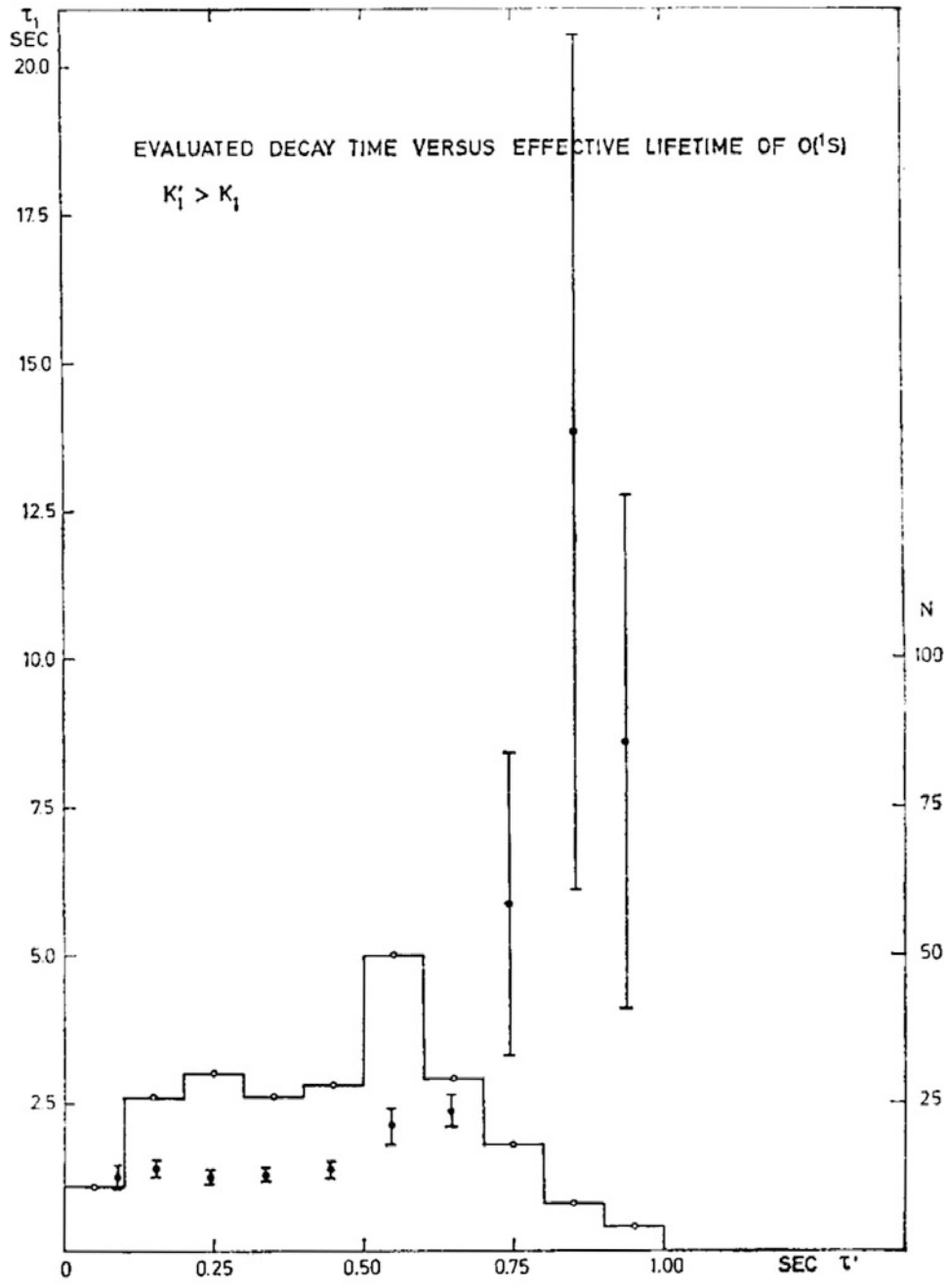


Figure 7.33. Number of evaluated decay times versus effective lifetimes of $O(^1S)$ atoms for case A (i.e., when $K_1' > K_1$). The bars and circles indicate the standard deviation of mean values and the number of evaluated records, respectively.

7.11 THE QUENCHING PROCESS

When a so-called forbidden transition occurs, its upper state is metastable and has a lifetime against radiation that is much longer than the lifetime against spontaneous emission, which is of the order of 10^{-7} s. The metastable states of major interest in the aurora are $O(^1D)$, $O(^1S)$, and $N_2(A^3\Sigma_u^+)$ having time constants corresponding to 110, 0.75, and 2 seconds, respectively.

For an emission due to a transition between an upper energy level E_m and a lower level E_n, the probability of radiation is expressed by the Einstein factor A_{nm}. The lifetime for the energy state E_m against a spontaneous transition is then given by

$$\tau_m = \frac{1}{A_{mn}}$$

Now there may be several lower levels E_n to which a transition can take place, and each individual transition will have its own Einstein coefficient. The final lifetime for the energy state m can then be expressed as:

$$\tau_M = \left(\sum_n A_{mn} \right)^{-1}$$

for all $E_n < E_m$. If, on the other hand, there is a number density N_m of excited states of energy E_m, then the volume emission rate of a photon with wavelength λ corresponding to the transition E_m–E_n will be

$$\eta(\lambda) = N_m \cdot A_{mn}$$

If the upper state with energy E_m is now metastable, it may be deactivated by a collision before it can emit a photon and the volume emission rate will be reduced. Let the metastable atom or molecule in the excited state E_m be denoted by $X(m)$, and let it be deactivated by a collision with another unspecified agent M. Then this collision can be expressed as:

$$X(m) + M \xrightarrow{k^q} X + M^*$$

where X is in a lower state or the ground state compared with $X(m)$, and M^* is in any arbitrary state. If this process has a reaction rate denoted by k^q, then the loss of $X(m)$ species per unit time due to these collisions is

$$L_q(N_m) = k^q \cdot N_m \cdot [M]$$

where the square brackets mean number densities as usual. In equilibrium, the number of excited states produced per unit time $\eta(N_m)$ must equal the number lost per unit time which is the sum of those lost by radiation and those lost by quenching:

$$\eta(N_m) = (A_{mn} + k^q \cdot [M])N_m$$

The emission rate per unit volume now becomes

$$\eta(\lambda) = N_m \cdot A_{mn} = \frac{\eta(N_m)}{A_{mn} + k^q \cdot [M]} \cdot A_{mn} \tag{7.11}$$

where λ is the wavelength corresponding to the energy transition E_m–E_n.

Now, in the $O(^1S)$ state there are two transition probabilities: $A_{3,2}$ which corresponds to $O(^1S)$–$O(^1D)$ at 557.7 nm and $A_{3,1}$ which corresponds to $O(^1S)$–$O(^3P)$ at 297.2 nm. Therefore, the probability of radiation transition between the ^1S and ^1D states is just a fraction of the total transition probability

$$A_3 = A_{3,1} + A_{3,2}$$

from the ^1S state.

Furthermore, the $O(^1S)$ atom may go through different kinds of deactivating collisions such as

$$O(^1S) + O \xrightarrow{k_1^q} O(^3P) + O$$

with a rate coefficient k_1^q and

$$O(^1S) + O_2 \xrightarrow{k_2^q} O(^3P) + O_2$$

with a rate coefficient k_2^q. Then the total loss of excited $O(^1S)$ atoms by quenching is given by

$$L(N_m) = (k_1^q[O] + k_2^q[O_2])$$

and the 557.7 nm volume emission rate is now

$$\eta(557.7\text{ nm}) = \eta(O(^1S)) \cdot \frac{A_{3,2}}{A_{3,1} + A_{3,2} + k_1^q[O] + k_2^q[O_2]}$$

The effective lifetime of the $O(^1S)$ state then becomes

$$\tau_{\text{eff}} = [A_{3,2} + A_{3,1} + k_1^q[O] + k_2^q[O_2]]^{-1} \tag{7.12}$$

Theoretical estimates of the lifetimes of $O(^1S)$ and O_2^+ ions are given in Figure 7.34, together with the observed lifetimes when a delayed process is accounted for.

Concerning the 297.2 nm emission

$$\eta(297.2\text{ nm}) = \eta(557.7\text{ nm})\frac{A_{3,1}}{A_{3,2}}$$

for the 630.0 nm emission. Each emission of a 557.7 nm photon results in a ^1D state that represents the upper state.

The ^1D state is deactivated by collisions between N_2 and O_2 molecules according to the following schemes

$$O(^1D) + N_2 \xrightarrow{k_3^q} O(^3P) + N_2$$

$$O(^1D) + O_2 \xrightarrow{k_4^q} O(^3P) + O_2$$

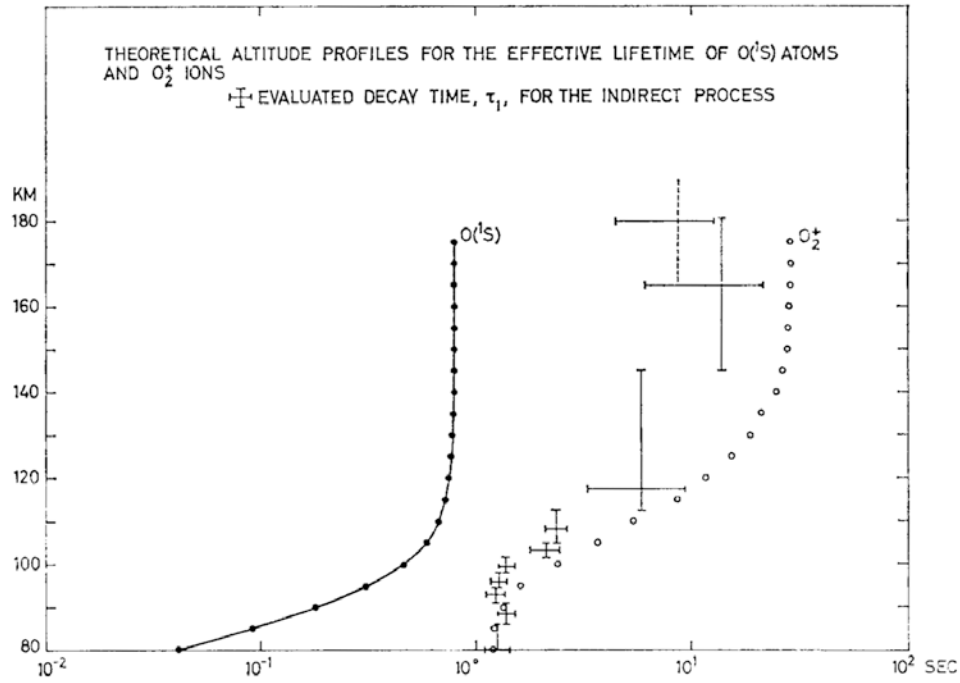


Figure 7.34. Theoretical altitude profiles for the lifetimes of $O(^1S)$ atoms and O_2^+ ions in the atmosphere at typical auroral altitudes. The crosses with horizontal and vertical bars indicate the decay time evaluated by the standard deviation of mean values and the uncertainty in estimating the altitude. (From Brekke and Pettersen, 1972.)

with reaction rate coefficients k_3^q and k_4^q. We now obtain

$$\eta(630\text{ nm}) = (\eta(O(^1D)) + \eta(557.7\text{ nm}))\frac{A_{2,1}}{A_{2,1} + k_3^q \cdot [N_2] + k_4^q \cdot [O_2]}$$

where $\eta(O(^1D))$ is the direct excitation rate of the $O(^1D)$ state. Quenching is most important below 115 km for the $O(^1S)$ state while it is important below 250–300 km for the $O(^1D)$ state. The latter will, however, depend heavily on the thermospheric temperature.

7.12 THE PROTON AURORA

The behavior of protons penetrating the atmosphere is fundamentally different from that of penetrating electrons. First, the probability of protons being deflected when colliding with particles in the atmosphere is almost negligible and, second,

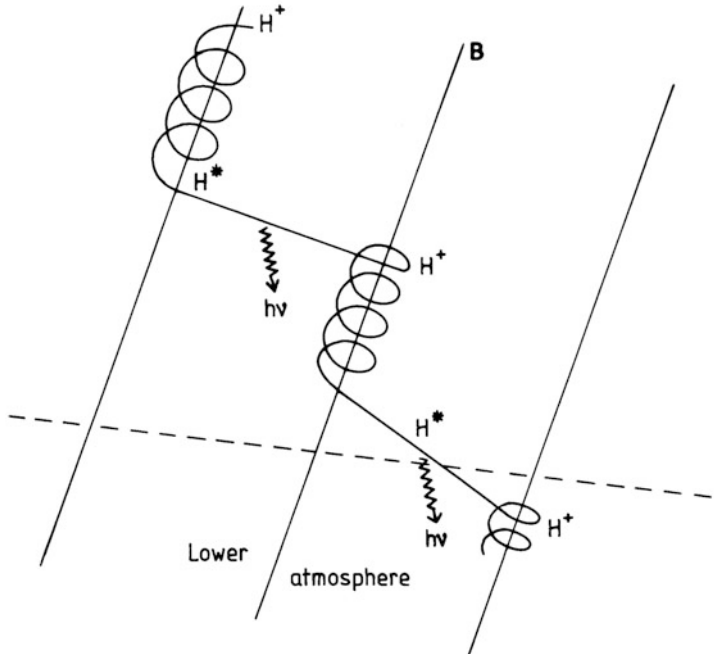


Figure 7.35. Schematic illustration showing charge loss for a proton colliding with a hydrogen atom and charge capture of the latter when colliding with a proton. In the neutral state a hydrogen atom can be excited to emit either H_α (656.3 nm) or H_β (486.1 nm). Due to the velocity of an H^* atom along the field line, Doppler shifts will be observed in emission lines.

the possibility of a fast proton capturing an electron to form a fast neutral hydrogen atom is considerable. In a neutral state the hydrogen atom is free to move with respect to the magnetic field until a new collision occurs and the hydrogen atom is stripped of its electron. This proton–hydrogen interchange leads to horizontal diffusion of protons rendering the proton aurora less structured than the electron aurora. Figure 7.35 illustrates the proton–hydrogen interchange and the diffusive action of this process.

After precipitation into the atmosphere, protons lose their kinetic energy through the following collisions with atmospheric molecules to finally merge into the ambient density of neutral hydrogen:

1: Ionization

$$H^+ + M \longrightarrow H^+ + M^+ + e \qquad (\Delta_{p1})$$

2: Excitation

$$H^+ + M \longrightarrow H^+ + M^* \qquad (\Delta_{p2})$$

3: Charge capture

$$H^+ + M \longrightarrow H^* + M^{*(+)} \qquad (\Delta_{p3})$$

Neutral hydrogen will then be involved in the following collision processes:

4: Charge loss or ionization

$$\mathrm{H} + M \longrightarrow \mathrm{H}^+ + M^* + e \qquad (\Delta_{\mathrm{H1}})$$

5: Double ionization

$$\mathrm{H} + M \longrightarrow \mathrm{H}^+ + M^{*(+)} + 2e \qquad (\Delta_{\mathrm{H2}})$$

6: Excitation

$$\mathrm{H} + M \longrightarrow \mathrm{H}^* + \mathrm{H} \qquad (\Delta_{\mathrm{H3}})$$

where M^* and M^+ represent excited and ionized neutral atmospheric species, and Δ_{p1}, Δ_{p2}, Δ_{p3}, Δ_{H1}, Δ_{H2} and Δ_{H3} are the collision cross-sections indicated for each process, respectively. These cross-sections (Figure 7.36) are typical between 10^{-23} and $10^{-19}\,\mathrm{m}^2$ in the energy range 0.1 keV to 1 MeV. Maximum cross-sections are found, however, around 10 keV.

Let us assume that at a given point in the beam of precipitating particles there are $n(p)$ protons and $n(\mathrm{H})$ hydrogen atoms per unit volume (Figure 7.37).

If Δ_p and Δ_{H} are the electron capture and loss cross-sections, respectively, then variation in the number density of protons in the beam as it traverses at distance dz along the beam is:

$$dn(p) = \Delta_{\mathrm{H}}\, dN \cdot n(\mathrm{H}) - \Delta_p\, dN \cdot n(p)$$

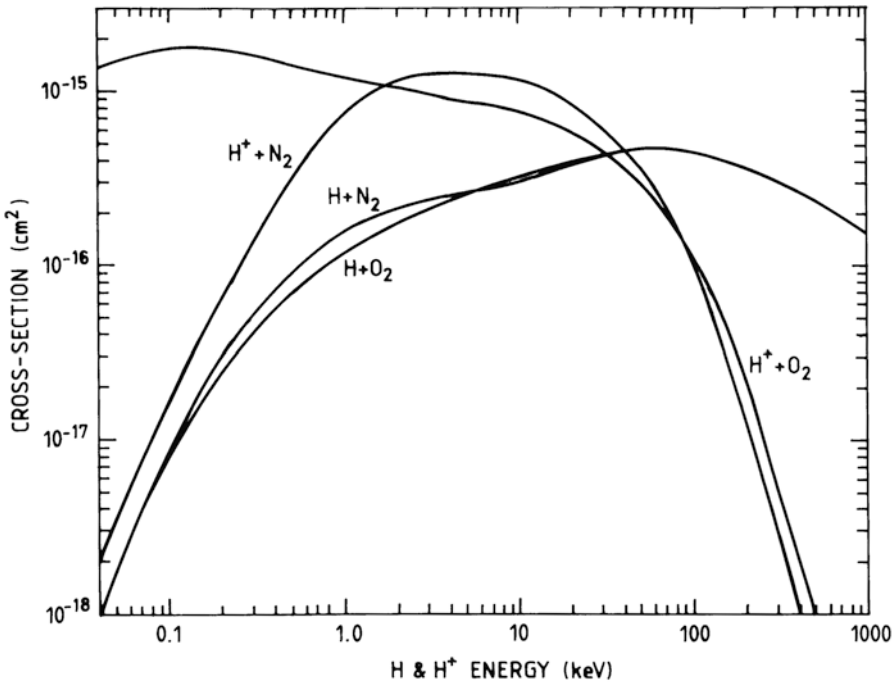


Figure 7.36. Cross-sections for electron capture and electron stripping for H^+ and H in nitrogen and oxygen, respectively. (After Rees, 1989.)

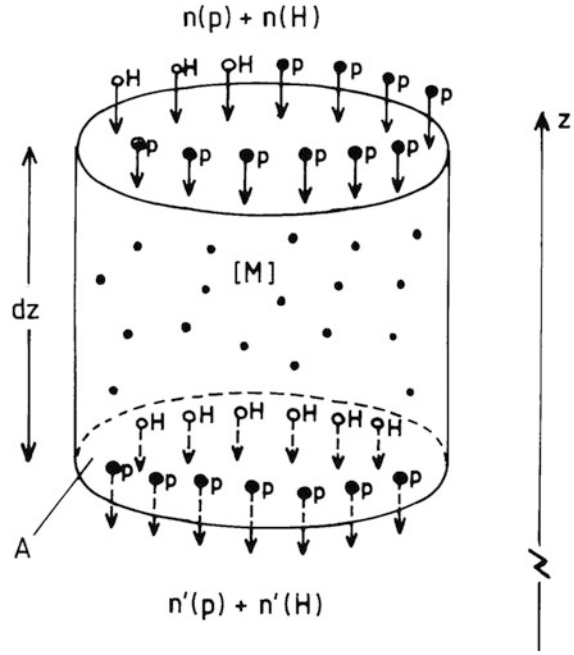


Figure 7.37. A schematic drawing to illustrate a beam of protons and hydrogen atoms passing through a distance dz of the atmosphere. The sum of the densities of the protons and the hydrogen atoms must be the same at the top of the cylinder with area A as at the bottom.

where $dN = [M] \cdot dz$ is the number per unit area of neutral particles traversed by the beam. Let us consider a part of the beam in the form of a cylinder with cross-section A and height dz. After penetrating the height dz the densities of the proton and hydrogen atoms in the beam are $n'(p)$ and $n'(\mathrm{H})$, respectively, and since no particle is lost from the cylinder in Figure 7.37

$$n(p) + n(\mathrm{H}) = n'(p) + n'(\mathrm{H}) = n_0$$

where n_0 is a constant number density. Let the relative number of protons be denoted by α; then

$$\alpha = \frac{n(p)}{n_0} = \frac{n(p)}{n(p) + n(\mathrm{H})}$$

and

$$d\alpha = \frac{d(n(p))}{n_0} = \left(\Delta_{\mathrm{H}} \frac{n(\mathrm{H})}{n_0} - \Delta_p \frac{n(p)}{n_0}\right) dN$$

Since $n(\mathrm{H}) = n_0(1 - \alpha)$, we have:

$$\frac{d\alpha}{dN} = -(\Delta_{\mathrm{H}} + \Delta_p)\alpha + \Delta_{\mathrm{H}}$$

This is a linear differential equation with constant coefficients that can be solved analytically as:

$$\alpha = \frac{\Delta_{\mathrm{H}}}{\Delta_{\mathrm{H}} + \Delta_p} + C \exp[-(\Delta_{\mathrm{H}} + \Delta_p)N]$$

where C is a constant to be determined according to the boundary conditions. At the top of the column where $N = 0$, all particles in the beam are protons ($\alpha = 1$), and therefore

$$\alpha = \frac{\Delta_{\mathrm{H}}}{\Delta_{\mathrm{H}} + \Delta_p}\left(1 + \frac{\Delta_p}{\Delta_{\mathrm{H}}}\exp[-(\Delta_{\mathrm{H}} + \Delta_p)N]\right)$$

A steady-state value of α given by

$$\alpha_s = \frac{\Delta_{\mathrm{H}}}{\Delta_{\mathrm{H}} + \Delta_p}$$

is reached at a rate determined by

$$N_s = \frac{1}{\Delta_{\mathrm{H}} + \Delta_p}$$

From Figure 7.36 we notice that Δ_p and Δ_{H} are typically of the order of $5 \times 10^{-20}\,\mathrm{m}^2$ in the energy range between 10 and 300 keV. The total number N of traversed neutral particles in the column must therefore be larger than $2 \times 10^{19}\,\mathrm{m}^{-2}$ before equilibrium can be reached. Since this occurs at a height close to 300 km in the Earth's atmosphere, a steady state is virtually complete in the beam below this height. Table 7.4 gives the ionization degree and the altitude of 90% of the ionization degree for different initial energies of the protons in kilo-electronvolts.

The incident proton flux will dominate when the beam has a high-energy level, while neutral hydrogen will dominate at lower energy levels.

Table 7.4. Ionization degrees of protons for different energies.

Proton energy ε (keV)	*Ionization degree* $\Delta_p/(\Delta_p + \Delta_{\mathrm{H}})$	*Height of 90% of ionization degree* (km)
3	0.11	310
10	0.26	305
30	0.46	295
100	0.80	268
300	0.99	252

Between 1939 and 1941 in Tromsø, the Norwegian physicist Vegard observed auroras in which emissions from the hydrogen lines H_α (656.3 nm) and H_β (486.1 nm) showed clear frequency shifts toward shorter wavelengths when observed along the magnetic field line (Figure 7.38(a)). Observed perpendicularly to the field line the H_α and H_β did not show Doppler shift but rather Doppler broadening (Figure 7.38(b)). These observations indicate that the emissions are due to hydrogen atoms moving toward the observer at relative speed. The average Doppler shift corresponds to a velocity of 300–400 km/s or a kinetic energy of 0.5–1 keV. This finding, however, does not represent the initial energy of the protons but rather the intermediate energy after the protons have been slowed down to kinetic energies in the neighborhood of the maximum cross-section for hydrogen excitation. The wing of the emission line in Figure 7.38(b) does correspond to a speed of 2,000 km/s, however. Some of the incident protons must therefore have energies of the order of 20 keV or more.

Vegard interpreted his measurements as evidence of protons penetrating the polar atmosphere together with electrons during an aurora and that the emissions observed were due to excited hydrogen atoms moving at high downward velocities due to charge exchange with protons (as explained). This gave important support to the idea that auroral particles originated in the solar atmosphere.

7.13 EXERCISES

1. What is the most dominant emission of the aurora?

2. When an aurora is moving, what color will you observe in the front of the aurora, and what color will the trailing edge have?

3. Assume that the $O(^1S)$ state is also produced by an indirect chemical reaction such as the one indicated by (7.3). Let the time-dependent density of the $N_2(A^2\Sigma)$ state be indicated as $N_2(t)$. We can then write the continuity equation for $I_O(t)$ as follows:

$$\frac{dI_O(t)}{dt} = k_1 I_N(t) + k_2 N_2(t) - \frac{I_O(t)}{\tau_1}$$

We also assume that production of the $N_2(A^2\Sigma)$ state is proportional and in phase with the ionization and excitation of N_2^+ ions. The continuity equation for $N_2(A^2\Sigma)$ can then be written as:

$$\frac{dN_2(t)}{dt} = k_2 I_N(t) - \frac{N_2(t)}{\tau_2}$$

Now, find the phase–angle relationship between $I_O(t)$ and $I_N(t)$.

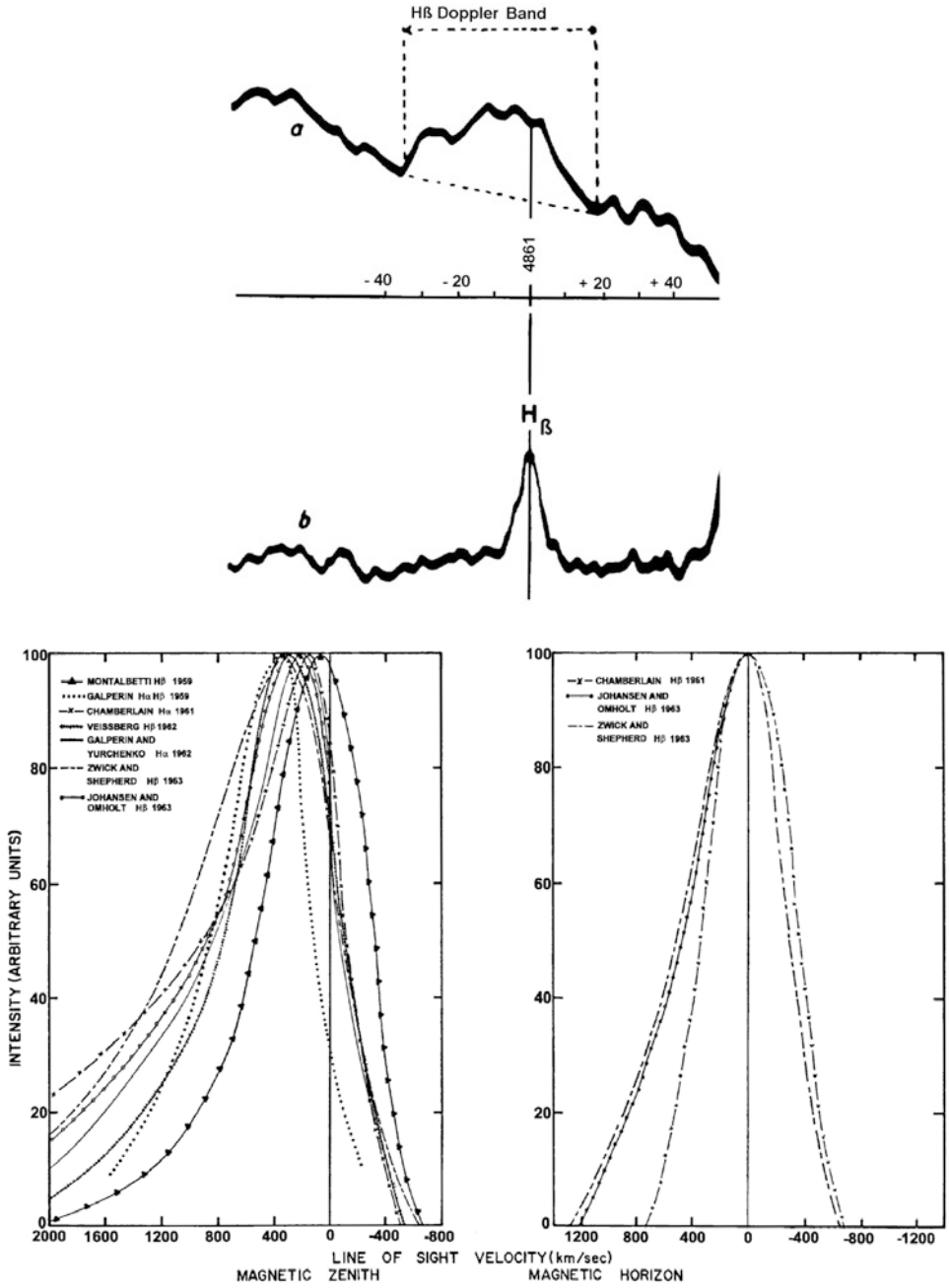


Figure 7.38. (a) Original observations done by Vegard, showing Doppler shift of the H_β line and also Doppler broadening of the line due to the velocity distribution of H^* atoms along the magnetic field line. (From Vegard, 1952.) (b) Observed hydrogen line profiles. (From Eather, 1967.)

References

Akasofu, S.-I. and Chapman, S. (1972) *Solar–Terrestrial Physics*, Clarendon Press, Oxford, U.K.

Alexandrescu, M., Courtillot, V., and Monël, J. L. (1997) High-resolution secular variation of the geomagnetic field in western Europe over the last 4 centuries: Comparision and integration of historical data from Paris and London. *J. Geophys. Res.*, **102**, 20245–20258.

Baron, M., Watt, T., Rino, C., and Petriceks, J. (1971) *DNA Project 617Radar: First Auroral-Zone Results* (DNA 2826F), Stanford Research Institute, Menlo Park, CA.

Barraclough, D. R. (1974) Spherical harmonic analysis of geomagnetic field for eight epochs between 1600 and 1910. *Geophys. J. Roy. Astron. Soc.*, **36**, 497–513.

Berger, M. J., Seltzer, S. M., and Maeda, K. (1970) Energy deposition by auroral electrons in the atmosphere. *J. Atmos. Terr. Phys.*, **32**, 1015–1045.

Birkeland, K. (1913) *The Norwegian Aurora Polaris Expedition 1902–3*, Vols. I and II, H. Aschehoug, Christiania, Norway.

Blixt, M. (1995) Private communication.

Boström, R. (1973) Electrodynamics of the ionosphere. Chapter 12 in: A. Egeland, Ø. Holter, and A. Omholt (Eds.), *Cosmical Geophysics*, Universitetsforlaget, Oslo, pp. 181–192.

Brekke A. and Egeland A. (1983) *The Northern Lights: From Mythology to Space Research*, Springer-Verlag, New York.

Brekke, A. and Hall, C. (1988) Auroral ionospheric quiet summer time conductances. *Ann. Geophysicae*, **6**, 361–376.

Brekke, A. and Henriksen, K. (1972) The intensity ratio I(5577)/I(4278) and the effective lifetime of O(^{1}S) atoms in pulsating aurora. *Planet. Space Sci.*, **20**, 53–56.

Brekke, A. and Kamide, Y. (1996) On the relationship between Joule and frictional heating in the polar ionosphere. *J. Atmos. Terr. Phys.*, **58**, 139–144.

Brekke, A. and Pettersen, H. (1972) A possible method for estimating any indirect process in the production of the O(^{1}S) atoms in aurora. *Planet. Space Sci.*, **20**, 1569–1576.

Brekke, A., Doupnik, J. R., and Banks, P. M. (1974) Observations of neutral winds in the auroral E region during the magnetospheric storm of August 3–9, 1972. *J. Geophys. Res.*, **79**, 3773–3790.

A. Brekke, *Physics of the Upper Polar Atmosphere*, Springer Praxis Books,
DOI 10.1007/978-3-642-27401-5, © Springer-Verlag Berlin Heidelberg 2013

Brekke, A., Hall, C., and Hansen, T. L. (1989) Auroral ionospheric conductances during disturbed conditions. *Ann. Geophysicae*, **7**, 269–280.

Brekke, A., Moen, J., and Hall, C. (1991) Studies of the conductivities in the auroral zone. *J. Geomag. Geoelectr.*, **43**, 441–465.

Brekke, A., Nozawa, S., and Sparr, T. (1994) Studies of the E-region neutral wind in the quiet auroral ionosphere. *J. Geophys. Res.*, **99**, 8801–8825.

Brekke, A., Nozawa, S., and Sato, M. (1995) Samples of auroral E-region parameters derived from EISCAT experiments. *J. Geomag. Geoelectr.*, **47**, 889–909.

Cahill, L. J. and Amazeen, P. G. (1963) The boundary of the geomagnetic field. *J. Geophys. Res.*, **68**, 1835–1843.

Cain, J. C. (1987) The Earth as a magnet. Chapter 4 in: S.- I. Akasofu, and Y. Kamide (Eds.), *The Solar Wind and the Earth*, Terra Sci. Publ. Co., Tokyo, Japan.

Chapman, S. and Bartels, J. (1940) *Geomagnetism*, Vol. 1, Clarendon Press, Oxford, U.K.

Chappel, C. R., Harris, K. K., and Sharp, G. W. (1970) A study of the influence of magnetic activity on the location of the plasmapause and measured by OGO 5. *J. Geophys. Res.*, **75**, 50–56.

CIRA (1965) COSPAR International Reference Atmosphere 1965.

Cox, A., Dalrymple, G. B., and Doell, R. R. (1967) Reversals of the Earth's magnetic field. *Scient. Am.*, **216**, 44–54.

Croom, S. A., Robbins, A. R., and Thomas, J. D. (1959) Two anomalies in the behaviour of the F2 Layer of the ionosphere. *Nature*, **184**, 2003–2004.

Danielson, R. E. (1961) The structure of sunspot umbras, I: Observations. *Ap. J.*, **134**, 275–287.

Dickinson, R. E., Ridley, E. C., and Roble, R. G. (1975) Meridional circulation in the thermosphere: Equinox conditions. *J. Atmos. Sci.*, **32**, 1737–1754.

Dickinson, R. E., Ridley, E. C., and Roble, R. G. (1984) Thermospheric general circulation with coupled dynamics and composition. *J. Atmos. Sci.*, **41**, 205–219.

Doupnik, J. R., Brekke, A., and Banks, P. M. (1977) Incoherent scatter radar observations during three sudden commencements and a Pc5 event on August 4, 1972. *J. Geophys. Res.*, **82**, 499–514.

Dryer, M. (1987) Solar wind and heliosphere, Chapter 2 in: S.-I. Akasofu and Y. Kamide (Eds.), *The Solar Wind and the Earth*, Terra Sci. Publ. Co., Tokyo, Japan, pp. 19–35.

Dungey, J. W. (1961) Interplanetary magnetic field and the auroral zones. *Phys. Rev. Lett.*, **6**(2), 47–48.

Dütsch, H. U. (1978) Vertical ozone distribution on a global scale. *Pure Appl. Geophys.*, **116**, 511–529.

Eather, R. H. (1967) Auroral proton precipitation and hydrogen emission. *Rev. Geophys.*, **5**, 207–285.

Eccles, D. and King, J. W. (1969) A review of topside sounder studies of the equatorial ionosphere. *Proc. Inst. Elect. Electronics Engr.*, **57**, 1012.

Eddy, J. A. (1976) The Maunder Minimum. *Science*, **192**, 1189–1202.

Egeland, A., Henriksen, T., og Kanestrøm, I. (1990) *Drivhuseffekten: Jordens Atmosfære og Magnetfelt*, Universitetet i Oslo [in Norwegian].

Evans, J. V. (1967) Midlatitude F-region densities and temperatures at sunspot minimum. *Planet. Space Sci.*, **15**, 1387–1405.

Fälthammar, C.-G. (1973) The solar wind. Chapter 7 in: A. Egeland, Ø. Holter, and A. Omholt (Eds.), *Cosmical Geophysics*, Universitetsforlaget, Oslo, pp. 121–142.

Fejer, J. A. (1964) Atmospheric tides and associated magnetic effects. *Rev. Geophys.*, **2**, 275.

Frank, L. (1994) Private communication.

Frank, L. A. and Craven, J. D. (1988) Imaging results from dynamic Explorer 1. *Rev. Geophys.*, **26**, 249–283.

Friis-Christensen, E., Kamide, Y., Richmond, A. D., and Matsushita, S. (1985) Interplanetary magnetic field control of high-latitude electric fields and currents determined from Greenland magnetometer data. *J. Geophys. Res.*, **90**, 1325–1338.

Fritz, H. (1881) *Das Polarlicht*, F. A. Brockhaus, Leipzig, Germany [in German].

Giraud, A. and Petit, M. (1978) *Ionospheric Techniques and Phenomena*, D. Reidel, Dordrecht, The Netherlands.

Hall, C. (1989) Private communication.

Hällström, G. G. (1847) De apparitionbus Aurorae Borealis in septentrionalibus: Europae pertibus. *Acta Soc. Scient. Fenn.*, **2**, 363–376 [in Latin].

Handbook of Geophysics and the Space Environment, Air Force Geophysics Laboratory, 1985.

Hansteen, C. (1827) On the polar lights, or Aurora Borealis and Australis. *Philos. Mag. Ann. Philos. New Ser.*, **2**, 333–334.

Harang, L. (1946) The mean field disturbance of the polar Earth magnetic storm. *Geophys. Publ.*, **16**(12), 1–44.

Hargreaves, J. K. (1979) *The Upper Atmosphere and Solar–Terrestrial Relations*, Van Nostrand Reinhold, London.

Heelis, R. A. (1984) The effects of interplanetary magnetic field orientation on dayside high-latitude ionospheric cusp. *J. Geophys. Res.*, **89**, 2873–2880.

Henriksen, K. (1974) Thesis, Auroral Observatory, Tromsø, Norway.

Herman, J. R. and Goldberg, R. A. (1978) *Sun, Weather and Climate*, NASA, Washington, D.C.

Hess, W. N. (1967) *The Radiation Belt and the Magnetosphere*, Ginn Blaisdell, Waltham, MA.

Hinteregger, H. E. (1980) *Proceedings of the Workshop on Solar UV Irradiance Monitors*, NOAA Environmental Research Laboratories (ERL), Boulder, CO.

Hinteregger, H. E., Hall, L. A., and Schmidtke, G. (1965) Solar XUV radiation and neutral particle distribution in July 1963 thermosphere. In: D. G. King-Hele, P. Müller, and G. Righini (Eds.), *Space Research*, Vol. 5, North-Holland, Amsterdam, pp. 1175–1190.

http://wdc.kugi.kyoto-u.ac.jp/igrf.2010

Hultqvist, B. (1988) *IRF Scientific Report 106*, Kiruna, Sweden.

Hundhausen, A. J., Bame, S. J., Ashbridge, J. R., and Sydoriak, S. J. (1970) Solar wind proton properties: Vela 3 observations from July 1965 to June 1975. *J. Geophys. Res.*, **75**, 4643–4657.

Ijiima, T. and Potemra, T. A. (1976) Field aligned currents in the dayside cusp. *J. Geophys. Res.*, **83**, 5971–5979.

Jacchia, L. G. (1965) New static models of the thermosphere and exosphere with empirical temperature profiles. In: D. G. King-Hele, P. Müller, and G. Righini (Eds.), *Space Research*, Vol. V, pp. 1152–1174, North-Holland, Amsterdam.

Jin, R.-S. and Thomas, D. M. (1977) Spectral line similarity in the geomagnetic dipole field variations and length of day fluctuations. *J. Geophys. Res.*, **82**, 828–834.

Johnson, C. Y. (1966) Ionospheric composition and density from 90 to 1200 kilometres at solar minimum. *J. Geophys. Res.*, **71**, 330.

Joint NOAA–USAF Space Weather Operations (1996) *Solar Geophysical Data* (SWO PRF 1087).

Kamide, Y. (1988) *Electrodynamic Processes in the Earth's Ionosphere and Magnetosphere*, Kyoto Sangyo University Press, Kyoto, Japan.

Kamide, Y. and Brekke, A. (1993) The importance of incoherent scatter radars in substorm studies. *Geophys. Res. Lett.*, **20**, 309–312.

Kamiyama, H. (1966) Ionization and excitation by precipitating electrons. *Rep. Ionosph. Space Res. Japan*, **20**, 171–187.

Kelley, M. C. (1989) *The Earth's Ionosphere: Plasma Physics and Electrodynamics*, Academic Press, New York.

Killeen, T. L. and Roble, R. G. (1986) An analysis of the high-latitude thermospheric wind patterns calculated by a thermospheric general circulation model, 2: Neutral parcel transport. *J. Geophys. Res.*, **91**, 11291–11307.

Kochanski, A. (1963) Circulation and temperatures at 70- to 100-kilometer height. *J. Geophys. Res.*, **68**, 213–216.

Kohl, H. and King, J. W. (1967) Atmospheric winds between 100 and 700 km and their effects on the ionosphere. *J. Atmos. Terr. Phys.*, **29**, 1045–1062.

Kunitake, M. and Schlegel, K. (1991) Neutral winds in the lower thermosphere at high latitudes from five years of EISCAT data. *Ann. Geophysicae*, **9**, 143–155.

Lean, J. (1991) Variations in the Sun's radiation output. *Rev. Geophys.*, **29**, 505–536.

London, J. (1985) The observed distribution of atmospheric ozone and its variation. Chapter 1 in: R. C. Whitten and S. S. Prasad (Eds.), *Ozone in the Free Atmosphere*, Van Nostrand Reinhold, London, pp. 11–80.

Luhmann, J. G. (1995) Ionospheres. In: M. G. Kievelson and C. T. Russell (Eds.), *Introduction to Space Physics*, Cambridge University Press, Cambridge, U.K., pp. 183–202.

Maeda, K. and Kato, S. (1966) Electrodynamics of the ionosphere. *Space Sci. Rev.*, **5**, 57.

Marklund, G. (1984) Auroral arc classification scheme based on the observed arc-associated electric field pattern. *Planet. Space Sci.*, **32**, 193–211.

Matsushita, S. (1967) Solar quiet and lunar daily variation field. Chap. III-1 in: S. Matsushita and W. H. Campbell (Eds.), *Physics of Geomagnetic Phenomena*, Academic Press, New York, pp. 302–424.

McElhinny, M. W. and Senanayake, W. E. (1982) Variations in the geomagnetic dipole, 1: The past 50000 years. *J. Geomag. Geoelect.*, **34**, 39–51.

Merrill, R. T. and McElhinny, M. (1983) The Earth's magnetic field. *International Geophysics Series*, **32**, Academic Press, London.

Moen, J. and Brekke, A. (1993) The solar flux influence on quiet time conductivities in the auroral ionosphere. *Geophys. Res. Lett.*, **20**, 971–974.

Mozer, F. S. (1973) Electric fields and plasma convection in the plasmasphere. *Rev. Geophys. Space Phys.*, **11**, 755–765.

Murgatroyd, R. J. (1957) Winds and temperatures between 20 km and 100 km: A review. *Quart. J. Roy. Meteorol. Soc.*, **83**, 417–458.

Nagata, T. and Kokubun, S. (1962) An additional geomagnetic daily variation field (S_q^p field) in the polar region on a geomagnetically quiet day. *Rep. Ionosph. Space Res. Japan*, **16**, 256–274.

Narcisi, R. S. and Bailey, A. D. (1965) Mass spectrometric measurements of positive ions at altitude from 64 to 120 kilometres. *J. Geophys. Res.*, **70**, 3687–3700.

Ness, N. F. (1965) The Earth's magnetic tail. *J. Geophys. Res.*, **70**, 2989–3005.

Ness, N. F. (1969) Preprint X-616-69-334, NASA Goddard Space Flight Center, Greenbelt, MD

Ness, N. F., Scearce, C. S., and Seek, J. B. (1964) Initial results of Imp 1 magnetic field experiment. *J. Geophys. Res.*, **69**, 3531–3569.

Nevanlinna, H. (1995) Auroral observations in Finland: Visual sightings during the 18th and 19th centuries. *J. Geomag. Geoelectr.*, **47**, 953–960.

NOAA–NESS (1979) *Earth–Atmosphere Radiation Budget Analyses Derived from NOAA Satellite Data June 1974–February 1978*, National Oceanic and Atmospheric Administration, Washington, D.C.

Nordenskiöld, A. E. (1880–81) *Vegas Färd Kring Asien och Europa*, F. & G. Beijer, Stockholm [in Swedish].

Nozawa, S. (1994, 1995) Private communication.

Nozawa, S. and Brekke, A. (1995) Studies of the E-region neutral wind in the disturbed auroral ionosphere. *J. Geophys. Res.*, **100**, 14717–14734.

Nozawa, S. and Brekke, A. (1999) Studies of the auroral E-region neutral wind through a solar cycle: Quiet days. *J. Geophys. Res.*, **104**, 45–66.

Nozawa, S., Brekke, A., Maeda, S., Aso, T., Hall, C., Ogawa, Y., Buchert, S. C., Röttger, J., Richmond, A. D., Roble, R., and Fujii, R. (2005) Mean winds, tides and quasi-2 day wave in the polar lower thermosphere based on EISCAT eight day run data in November 2003. *J. Geophys. Res.*, **110**, A12309, doi: 10.1029/2005JA0011128.

Oguti, T. (1994) Private communication.

Oyama, S.-I. (2010) Private communication.

Paetzold, H. K. and Zschörner, H. (1960) Bearings of Sputnik 3 and the variable acceleration of satellites. In: H. Kallmann (Ed.), *Space Research*, North-Holland, Amsterdam, pp. 24–36.

Park, C. G. and Dejnakarintra, M. (1974) Paper presented at *Fifth International Conference on Atmospheric Electricity, Garmisch-Partenkirchen, West Germany, September 2–7, 1974*.

Pettit, E. (1951) In *Astrophysics*, edited by J. A. Hynek, p. 259, McGraw-Hill, New York.

Polar Geophysical Institute, Kola Branch (2011) Private communication.

Ratcliffe, J. A. (1972) *An Introduction to the Ionosphere and the Magnetosphere*, Cambridge University Press, Cambridge, U.K.

Rees, M. H. (1989) *Physics and Chemistry of the Upper Atmosphere* (Cambridge Atmospheric and Space Science Series), Cambridge University Press, Cambridge, U.K.

Reid, G. C. (1965) Ionospheric effects of electrostatic fields generated in the outer magnetosphere. *Radio Sci.*, **69D**, 827–837.

Reiff, P. H. and Burch, J. L. (1985) IMF B_y-dependent plasma flow and Birkeland currents in the dayside magnetosphere, 2: A global model for northward and southward IMF. *J. Geophys. Res.*, **90**, 1595–1609.

Réthly, A. and Berkes, Z. (1963) *Nordlichtbeobachtungen in Ungarn (1523–1960)*, Akadémiai Kiadó, Verlag der Ungarischen Akademie der Wissenschaften, Budapest [in German].

Richmond, A. (1987) The ionosphere. Chapter 7 in: S.-I. Akasofu and Y. Kamide (Eds.), *The Solar Wind and the Earth*, Terra Sci. Publ. Co., Tokyo, Japan, pp. 123–140.

Rikitake, T. (1958) Oscillations of a system of disk dynamos. *Proc. Camb. Phil. Soc. Math. Phys. Sci.*, **54**, 89–105.

Rishbeth, H. and Garriott, O. K. (1969) *Introduction to Ionospheric Physics*, Academic Press, London.

Roberts, W. O. (1975) *Relationship between Solar Activity and Climate Change* (NASA SP-366), edited by W. E. Bandeen and S. Maran, p. 13, NASA Goddard Space Flight Center, Greenbelt, MD.

Roble, R. G. (1987) The Earth's thermosphere. Chapter 12 in: S.-I. Akasofu, and Y. Kamide (Eds.), *The Solar Wind and the Earth*, Terra Sci. Publ. Co., Tokyo, Japan, pp. 243–264.

Roble, R. G. and Emery, B. A. (1983) On the global mean temperature of the thermosphere. *Planet. Space Sci.*, **31**, 597–614.

Rubenson, R. (1879, 1882) *Kungl. Svenska Vetensk. Acad. Handl.*, Part 1, 15(5); Part 2, 18(1).

Schwabe, H. (1844) Sonnen-Beobachtungen im Jahre 1843. *Astron. Nachr.*, **21**, 233 [in German].

Sharp, R. D. and Johnson, R. G. (1974) Low energy auroral particle measurements from polar satellites. In: B. M. McCormack (Ed.), *The Radiating Atmosphere*, D. Reidel, Dordrecht, The Netherlands, p. 239.

Shimazaki, T. (1987) Ozone and the stratosphere. In: S.-I. Akasofu, and Y. Kamide (Eds.), *The Solar Wind and the Earth*, Terra Sci. Publ. Co., Tokyo, Japan.

Silsbee, H. C. and Vestine, E. H. (1942) Geomagnetic bays, their frequency and current systems. *Terr. Mag.*, **47**, 195–208.

Smith, R. W. and Sweeney, P. J. (1980) Winds in the thermosphere of the Northern Polar Cap. *Nature*, **284**, 437–438.

Starkov, G. V. and Feldstein, Ya. I. (1967) Change in the boundaries of the oval auroral zone. *Geomagnetism and Aeronomy*, **7**, 48–54, Fig. 7.12.

Størmer, C. (1955) *The Polar Aurora*, Oxford University Press, London.

Stroud, W. G., Nordberg, W., Bandeen, W. R., Bartman, F. L., and Titus, P. (1959) Rocket-grenade observation of atmospheric heating in the Arctic. *J. Geophys. Res.*, **64**, 1342–1343.

Teske, R. G. (1969) Observations of the solar soft X-ray components: Study of its relation to transient and slowly varying phenomena observed at other wavelengths. *Solar Physics*, **6**, 193.

Tohmatsu, T. (1990) *Compendium of Aeronomy*, Terra Sci. Publ. Co., Tokyo, Japan.

U.S. Standard Atmosphere (1976) NOAA, NASA, USAF, Washington, D.C.

Vallance Jones, A. (1974) *The Aurora*, D. Reidel, Dordrecht, The Netherlands.

Van Zandt, T. E. and Knecht, R. W. (1964) The structure and physics of the upper atmosphere. In: D. P. Le Galley and A. Rosen (Eds.), *Space Physics*, John Wiley & Sons, New York.

Vegard, L. (1952) Doppler displacement of auroral hydrogen lines and its bearing on the theory of aurora and magnetic disturbances. *Geofys. Publ.*, **18**(5), 1–15.

Waldmeier, M. (1955) *Ergebnisse und Problem der Sonnenforschung*, Akademische Verlagsgesellschaft, Leipzig, Germany [in German].

Walterscheid, R. L. (1989) Solar cycle effects on the upper atmosphere: Implications for satellite drag. *J. Spacecraft and Rockets*, **26**, 439–444.

Whitten, R. C. and Poppoff, I. G. (1971) *Fundamentals of Aeronomy*, John Wiley & Sons, New York.

Wild, J. P., Smerd, S. F., and Weiss, A. A. (1963) Solar bursts. *Ann. Rev. Astron. Astrophys.*, **1**, 291.

Willson, R. C. and Hudson, H. S. (1988) The Sun's luminosity over a complete solar cycle. *Nature*, **322**, 810.

Winningham, J. D. and Heikkila, W. J. (1974) Polar cap auroral electron fluxes observed with Isis 1. *J. Geophys. Res.*, **79**, 949–957.

Yukutake, T. (1967) The westward drift of the Earth's magnetic field in historic times. *J. Geomag. Geoelect.*, **19**, 103.

Zirin, H. (1966) *The Solar Atmosphere*, Ginn Blaisdell, Waltham, MA.

Symbols

UNIVERSAL CONSTANTS

c	Speed of light ($= 3 \times 10^8$ m/s)
σ	Stephan–Boltzmann constant ($= 5.67 \times 10^{-8}$ W m^{-2} K^{-4})
m_e	Electron mass ($= 9.1 \times 10^{-31}$ kg)
m_p	Proton mass ($= 1.672 \times 10^{-27}$ kg)
ε_0	Permittivity constant ($= 8.854 \times 10^{-12}$ F/m)
μ_0	Permeability constant ($= 4\pi \times 10^{-7}$ H/m)
m_0	One atomic unit (1 a.m.u. $= 1.660 \times 10^{-27}$ kg)
κ	Boltzmann constant ($= 1.38 \times 10^{-23}$ J/K)
h	Planck's constant ($= 6.63 \times 10^{-34}$ J s)
e	Electronic charge ($= 1.60 \times 10^{-19}$ C)
G	Constant of gravity ($= 6.67 \times 10^{-11}$ N m^2 kg^{-2})

PLANETARY CONSTANTS USED IN THE TEXT

R_e	Mean radius of the Earth ($= 6.37 \times 10^6$ m)
$R_\odot$	Mean solar radius ($= 6.96 \times 10^8$ m)
M_e	Earth's mass ($= 5.98 \times 10^{24}$ kg)
$M_\odot$	Solar mass ($= 1.99 \times 10^{30}$ kg)
E_e	Solar constant at 1 AU ($= 1{,}380$ W/m^2)
1 AU	One astronomical unit ($\approx 1.5 \times 10^{11}$ m)

A. Brekke, *Physics of the Upper Polar Atmosphere*, Springer Praxis Books,
DOI 10.1007/978-3-642-27401-5, © Springer-Verlag Berlin Heidelberg 2013

H_0 Mean magnetic field strength at the equator ($= 3.0 \times 10^{-5}$ tesla)

M_0 Magnetic dipole moment of the Earth ($= 7.74 \times 10^{22}$ A m^2) ($H_0 R_e^3 = 7.74 \times 10^{15}$ Wb m)

$g_\odot$ Acceleration of gravity at the solar surface ($=2.74 \times 10^2$ m/s^2)

H Mean scale height of the Earth's atmosphere ($= 8.47$ km at the surface)

α^* Adiabatic lapse rate ($= -9.8$ K/km at the Earth's surface)

$v_{\rm esc}$ Escape velocity ($= 11.2$ km at the Earth's surface)

$v_{\odot e}$ Escape velocity at the solar surface ($= 6.2 \times 10^5$ m/s)

r_a Aphelion ($= 1.52 \times 10^{11}$ m)

r_p Perihelion ($= 1.46 \times 10^{11}$ m)

$T_\odot$ Mean solar radiation temperature ($= 6{,}000$ K)

T_e Mean temperature of the Earth ($= 288$ K)

γ Adiabatic constant of atmospheric gas ($=1.4$ at the Earth's surface)

M' Molecular mass number ($= 28.8$ of the Earth's atmosphere close to the surface)

δ Angle between the Earth's magnetic and rotation axis ($= 10^\circ$)

c_v Heat capacity at constant volume ($= 712$ J/K kg for air at the Earth's surface)

c_p Heat capacity at constant pressure ($= 996$ J/K kg for air at the Earth's surface)

$\bar{V}_p$ Mean ionization threshold in air ($= 15$ eV)

$\bar{\varepsilon}$ Mean energy for ion-pair production in air ($= 34$ eV)

Ω Rotation frequency of the Earth ($= 7.27 \times 10^{-5}$ s^{-1})

g Acceleration of gravity at the Earth's surface ($= 9.81$ m/s^2)

Index

(f. = and the following page; ff. = and the following pages;
fff. = and following pages to the end of the chapter)

A. Brekke, *Physics of the Upper Polar Atmosphere*, Springer Praxis Books,
DOI 10.1007/978-3-642-27401-5, © Springer-Verlag Berlin Heidelberg 2013