The high-latitude ionosphere and its effects on radio propagation

R. D. Hunsucker
Geophysical Institute, University of Alaska, Fairbanks

J. K. Hargreaves
University of Lancaster

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2003

Printed in the United Kingdom at the University Press, Cambridge

Typeface Times NR MT 10.25/13.5pt System QuarkXPress™ [SE]

A catalogue record for this book is available from the British Library

ISBN 0 521 33083 1 hardback

Contents

Chapter 1

Basic principles of the ionosphere

1.1 Introduction

1.1.1 The ionosphere and radio-wave propagation

The *ionosphere* is the ionized component of the atmosphere, comprising free electrons and positive ions, generally in equal numbers, in a medium that is electrically neutral. Though the charged particles are only a minority amongst the neutral ones, they nevertheless exert a great influence on the electrical properties of the medium, and it is their presence that brings about the possibility of radio communication over large distances by making use of one or more ionospheric reflections.

The early history of the ionosphere is very much bound up with the development of communications. The first suggestions that there are electrified layers within the upper atmosphere go back to the nineteenth century, but the modern developments really started with Marconi's well-known experiments in trans-Atlantic communication (from Cornwall to Newfoundland) in 1901. These led to the suggestions by Kennelly and by Heaviside (made independently) that, because of the Earth's curvature, the waves could not have traveled directly across the Atlantic but must have been reflected from an ionized layer. The name *ionosphere* came into use about 1932, having been coined by Watson-Watt several years previously. Subsequent research has revealed a great deal of information about the ionosphere: its vertical structure, its temporal and spatial variations, and the physical processes by which it is formed and which influence its behavior.

Looked at most simply, the ionosphere acts as a mirror situated between 100 and 400 km above the Earth's surface, as in Figure 1.1, which allows reflected

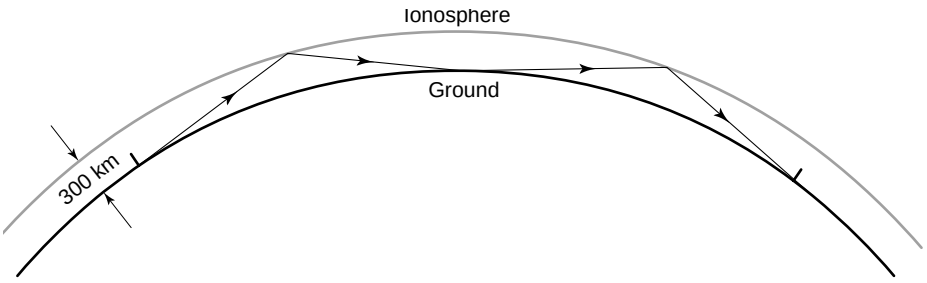

Figure 1.1. Long distance propagation by multiple hops between the ionosphere and the ground.

signals to reach points around the bulge of the Earth. The details of how reflection occurs depend on the radio frequency of the signal, but the most usual mechanism, which applies in the high-frequency (HF) band (3–30 MHz), is actually a gradual bending of the ray towards the horizontal as the refractive index of the ionospheric medium decreases with altitude. Under good conditions, signals can be propagated in this way for several thousand kilometers by means of repeated reflections between ionosphere and ground. Reflection from a higher level (the F region) obviously gives a greater range per "hop" than does one from a lower level (the E region), but which mode is possible depends on the structure of the ionosphere at the time. Higher radio frequencies tend to be reflected from greater heights, but if the frequency is too high there may be insufficient bending and the signal then penetrates the layer and is lost to space. This is the first complication of radio propagation.

The second complication is that the lower layers of the ionosphere tend to absorb the signal. This effect is greater for signals of lower frequency and greater obliquity. Hence, practical radio communication generally requires a compromise. The ionosphere is constantly changing, and the art of propagation prediction is to determine the best radio frequency for a given path for the current state of the ionosphere. Plainly, an understanding of ionospheric mechanisms is basic to efficient radio communication.

Further details about radio propagation are given in Chapter 3, and our central topic of how propagation at high latitudes is affected by the vagaries of the high-latitude ionosphere is discussed later in the book.

1.1.2 Why the ionosphere is so different at high latitude

The terrestrial ionosphere may be divided broadly into three regions that have rather different properties according to their geomagnetic latitude. The mid-latitude region has been explored the most completely and is the best understood. There, the ionization is produced almost entirely by energetic ultra-violet and X-ray emissions from the Sun, and is removed again by chemical recombination processes that may involve the neutral atmosphere as well as the ionized species. The

movement of ions, and the balance between production and loss, are affected by winds in the neutral air. The processes typical of the mid-latitude ionosphere also operate at high and low latitudes, but in those regions additional processes are also important.

The low-latitude zone, spanning 20° or 30° either side of the magnetic equator, is strongly influenced by electromagnetic forces that arise because the geomagnetic field runs horizontally over the magnetic equator. The primary consequence is that the electrical conductivity is abnormally large over the equator. A strong electric current (an "electrojet") flows in the E region, and the F region is subject to electrodynamic lifting and a "fountain effect" that distorts the general form of the ionosphere throughout the low-latitude zone.

At high latitudes we find the opposite situation. Here the geomagnetic field runs nearly vertical, and this simple fact of nature leads to the existence of an ionosphere that is considerably more complex than that in either the middle or the low-latitude zones. This happens because the magnetic field-lines connect the high latitudes to the outer part of the magnetosphere which is driven by the solar wind, whereas the ionosphere at middle latitude is connected to the inner magnetosphere, which essentially rotates with the Earth and so is less sensitive to external influence. We can immediately identify four general consequences.

(a). The high-latitude ionosphere is dynamic. It circulates in a pattern mainly controlled by the solar wind but which is also variable.

(b). The region is generally more accessible to energetic particle emissions from the Sun that produce additional ionization. Thus it is affected by sporadic events, which can seriously degrade polar radio propagation. Over a limited range of latitudes the dayside ionosphere is directly accessible to material from the solar wind.

(c). The auroral zones occur within the high-latitude region. Again, their location depends on the linkage with the magnetosphere, in this case into the distorted tail of the magnetosphere. The auroral phenomena include electrojets, which cause magnetic perturbations, and there are "substorms" in which the rate of ionization is greatly increased by the arrival of energetic electrons. The auroral regions are particularly complex for radio propagation.

(d). A "trough" of lesser ionization may be formed between the auroral and the mid-latitude ionospheres. Although the mechanisms leading to the formation of the trough are not completely known, it is clear that one fundamental cause is the difference in circulation pattern between the inner and outer parts of the magnetosphere.

This monograph is concerned mainly with the ionosphere at high latitudes, but before considering the special behavior which occurs in those regions we must review some processes affecting the ionosphere in general and summarize the more normal behavior at middle latitudes. In order to do that, we must first

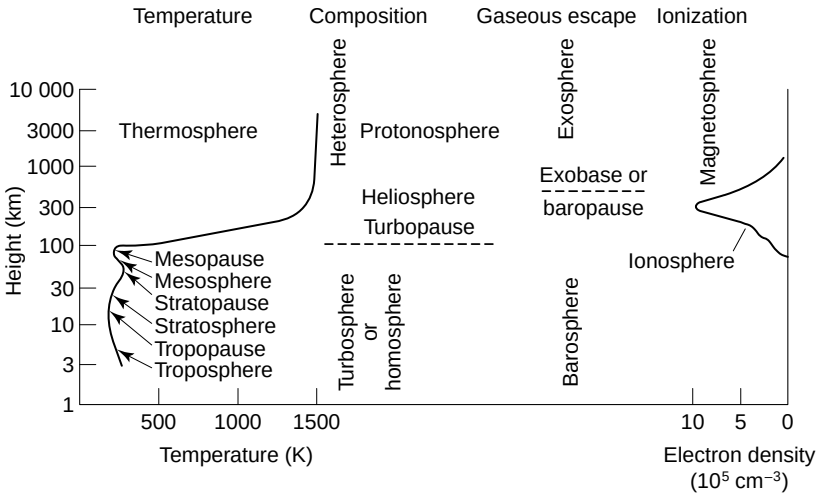

Figure 1.2. Nomenclature of the upper atmosphere based on temperature, composition, mixing, and ionization. (J. K. Hargreaves, *The Solar–Terrestrial Environment*. Cambridge University Press, 1992.)

consider the nature of the neutral upper atmosphere in which the ionosphere is formed.

1.2 The vertical structure of the atmosphere

1.2.1 Nomenclature

A static planetary atmosphere may be described by four properties: pressure (P), density (ρ), temperature (T), and composition. Since these are not independent it is not necessary to specify all of them. The nomenclature of the atmosphere is based principally on the variation of temperature with height, as in Figure 1.2. Here, the different regions are called "spheres" and the boundaries between them are "pauses". The lowest region is the *troposphere*, in which the temperature falls off with increasing height at a rate of 10 K km^{-1} or less. Its upper boundary is the *tropopause* at a height of 10–12 km. The *stratosphere* which lies above it was once thought to be isothermal, but it is actually a region where the temperature increases with height. At about 50 km is a maximum due to the absorption of solar ultra-violet radiation in ozone; this is the *stratopause*. Above that the temperature again decreases in the *mesosphere* (or *middle atmosphere*) and passes through another minimum at the *mesopause* at 80–85 km. At about 180 K, this is the coldest part of the whole atmosphere. Above the mesopause, heating by solar ultra-violet radiation ensures that the temperature gradient remains positive, and this is the *thermosphere*. Eventually the temperature of the thermosphere becomes

almost constant at a value that varies with time but is generally over 1000 K; this is the hottest part of the atmosphere.

Though the classification by temperature is generally the most useful, others based on the state of mixing, the composition or the state of ionization are also useful. The lowest part of the atmosphere is well mixed, with a composition much like that at sea level except for minor components. This is the *turbosphere* or *homosphere*. In the upper region, essentially the thermosphere, mixing is inhibited by the positive temperature gradient, and here, in the *heterosphere*, the various components separate under gravity and as a result the composition varies with altitude. The boundary between the two regions, which occurs at about 100 km, is the *turbopause*. Above the turbopause the gases separate by gaseous diffusion more rapidly than they are mixed by turbulence.

Within the heterosphere there are regions where helium or hydrogen may be the main component. These are the *heliosphere* and the *protonosphere*, respectively. From the higher levels, above about 600 km, individual atoms can escape from the Earth's gravitational attraction; this region is called the *exosphere*. The base of the exosphere is the *exobase* or the *baropause*, and the region below the baropause is the *barosphere*.

The terms *ionosphere* and *magnetosphere* apply, respectively, to the ionized regions of the atmosphere and to the outermost region where the geomagnetic field controls the particle motions. The outer termination of the geomagnetic field (at about ten Earth radii in the sunward direction) is the *magnetopause*.

1.2.2 Hydrostatic equilibrium in the atmosphere

Between them the properties temperature, pressure, density, and composition determine much of the atmosphere's behaviour. They are not independent, being related by the universal gas law which may be written in various forms, but for our purposes the form

$$P = nkT, \tag{1.1}$$

where n is the number of molecules per unit volume, is the most useful. The quantity n is properly called the *concentration* or the *number density*, but "*density*" alone is often used when the sense is clear.

Apart from its composition, the most significant feature of the atmosphere is that the pressure and density decrease with increasing altitude. This height variation is described by the *hydrostatic equation*, sometimes called the *barometric equation*, which is easily derived from first principles. The variation of pressure with height is

$$P = P_0 \exp(-h/H), \tag{1.2}$$

where P is the pressure at height h, P_0 is the pressure where $h = 0$, and H is the scale height given by

$$H = kT/(mg), \tag{1.3}$$

in which k is Boltzmann's constant, T is the absolute temperature, m is the mass of a single molecule of the atmospheric gas, and g is the acceleration due to gravity.

If T and m are constant (and any variation of g with height is neglected), H is the vertical distance over which n falls by a factor e (= 2.718), and thus it serves to define the thickness of an atmosphere. H is greater, and the atmosphere thicker, if the gas is hotter or lighter. In the Earth's atmosphere H varies from about 5 km at height 80 km to 70–80 km at 500 km.

Using equation (1.1), the hydrostatic equation may be written in differential form as

$$\mathrm{d}P/P = \mathrm{d}n/n + \mathrm{d}T/T = -\mathrm{d}h/H. \tag{1.4}$$

From this, H can be ascribed a local value, even if it varies with height.

Another useful form is

$$P/P_0 = \exp[-(h - h_0)/H] = \mathrm{e}^{-z}, \tag{1.5}$$

where $P = P_0$ at the height $h = h_0$, and z is the *reduced height* defined by

$$\mathrm{z} = (h - h_0)/H. \tag{1.6}$$

The hydrostatic equation can also be written in terms of the density (ρ) and the number density (n). If T, g, and m are constant over at least one scale height, the equation is essentially the same in terms of P, ρ, and n, since $n/n_0 = \rho/\rho_0 = P/P_0$. The ratio k/m can also be replaced by R/M, where R is the gas constant and M is the relative molecular mass.

Whatever the height distribution of the atmospheric gas, its pressure P_0 at height h_0 is just the weight of gas above h_0 in a column of unit cross-section. Hence

$$P_0 = N_\mathrm{T} mg = n_0 k T_0, \tag{1.7}$$

where N_T is the total number of molecules in the column above h_0, and n_0 and T_0 are the concentration and the temperature at h_0. Therefore we can write

$$N_\mathrm{T} = n_0 k T_0/(mg) = n_0 H_0, \tag{1.8}$$

H_0 being the scale height at h_0. This equation says that, if all the atmosphere above h_0 were compressed to density n_0 (that already applying at h_0), then it would

occupy a column extending just one scale height. Note also that the total mass of the atmosphere above unit area of the Earth's surface is equal to the surface pressure divided by g.

Although we often assume that g, the acceleration due to gravity, is a constant, in fact it varies with altitude as $g(h) \propto 1/(R_E + h)^2$, where R_E is the radius of the Earth. The effect of changing gravity may be taken into account by defining a *geopotential height*

$$h^* = R_E h/(R_E + h). \tag{1.9}$$

A molecule at height h over the spherical Earth has the same potential energy as one at height h^* over a hypothetical flat Earth having gravitational acceleration $g(0)$.

Within the homosphere, where the atmosphere is well mixed, the mean relative molecular mass determines the scale height and the variation of pressure with height. In the heterosphere, the partial pressure of each constituent is determined by the relative molecular mass of that species. Each species takes up its own distribution, and the total pressure of the atmosphere is the sum of the partial pressures in accordance with Dalton's law.

1.2.3 The exosphere

In discussing the atmosphere in terms of the hydrostatic equation we are treating the atmosphere as a compressible fluid whose temperature, pressure, and density are related by the gas law. This is valid only if there are sufficient collisions between the gas molecules for a Maxwellian velocity distribution to be established. As the pressure decreases with increasing height so does the collision frequency, and at about 600 km the distance traveled by a typical molecule between collisions, the *mean free path*, becomes equal to the scale height. At this level and above we have to regard the atmosphere in a different way, not as a fluid but as an assembly of individual molecules or atoms, each following its own trajectory in the Earth's gravitational field. This region is called the *exosphere*.

While the hydrostatic equation is strictly valid only in the barosphere, it has been shown that the same form may still be used if the velocity distribution is Maxwellian. This is true to some degree in the exosphere, and the use of the hydrostatic equation is commonly extended to 1500–2000 km, at least as an approximation. However, this liberty may not be taken if there is significant loss of gas from the atmosphere, since more of the faster molecules will be lost and the velocity distribution of those remaining will be altered thereby. The lighter gases, helium and hydrogen, are affected most.

The rate at which gas molecules escape from the gravitational field in the exosphere depends on their vertical speed. Equating the kinetic and potential energies of an upward-moving particle, its escape velocity (v_e) is given by

$$v_e^2 = 2gr, \tag{1.10}$$

where r is the distance of the particle from the center of the Earth. (At the Earth's surface the escape velocity is 11.2 km s^{-1}, irrespective of the mass of the particle.)

By kinetic theory the root mean square (r.m.s.) thermal speed of gas molecules ($\overline{v^2}$) depends on their mass and temperature, and, for speeds in one direction, i.e. vertical,

$$m\overline{v^2}/2 = 3kT/2. \tag{1.11}$$

Thus, corresponding to an escape velocity (v_e) there can be defined an *escape temperature* (T_e).

T_e is 84 000 K for atomic oxygen, 21 000 K for helium, but only 5200 K for atomic hydrogen. At 1000–2000 K, exospheric temperatures are smaller than these escape temperatures, and loss of gas, if any, will be mainly at the high-speed end of the velocity distribution. In fact, the loss is insignificant for O, slight for He, but significant for H. Detailed computations show that the resulting vertical distribution of H departs significantly from the hydrostatic at distances more than one Earth radius above the surface, but for He the departure is small.

1.2.4 The temperature profile of the neutral atmosphere

The atmosphere's temperature profile results from the balance amongst sources of heat, loss processes, and transport mechanisms. The total picture is complicated, but the main points are as follows.

Sources

The troposphere is heated by convection from the hot ground, but in the upper atmosphere there are four sources of heat:

(a). Absorption of solar ultra-violet and X-ray radiation, causing photodissociation, ionization, and consequent reactions that liberate heat;

(b). Energetic charged particles entering the upper atmosphere from the magnetosphere;

(c). Joule heating by ionospheric electric currents; and

(d). Dissipation of tidal motions and gravity waves by turbulence and molecular viscosity.

Generally speaking, the first source (a) is the most important, though (b) and (c) are also important at high latitude. Most solar radiation of wavelength less than 180 nm is absorbed by N_2, O_2 and O. Photons that dissociate or ionize molecules or atoms generally have more energy than that needed for the reaction, and the excess appears as kinetic energy of the reaction products. A newly created photoelectron, for example, may have between 1 and 100 eV of kinetic energy, which

subsequently becomes distributed throughout the medium by interactions between the particles (optical, electronic, vibrational, or rotational excitation, or elastic collisions, depending on the energy.) Elastic collisions redistribute energy less than 2 eV, and, since this process operates mainly between electrons, these remain hotter than the ions. Some energy is reradiated, but on average about half goes into local heating. It can generally be assumed that in the ionosphere the rate of heating in a given region is proportional to the ionization rate.

The temperature profile (Figure 1.2) can be explained as follows. The maximum at the stratopause is due to the absorption of 200–300-nm (2000–3000-Å) radiation by ozone (O_3) over the height range 20–50 km. Some 18 W m^{-2} is absorbed in the ozone layer. Molecular oxygen (O_2), which is relatively abundant up to 95 km, absorbs radiation between 102.7 and 175 nm, much of this energy being used to dissociate O_2 to atomic oxygen (O). This contribution amounts to some 30 mW m^{-2}. Radiation of wavelengths shorter than 102.7 nm, which is the ionization limit for O_2 (See Table 1.1 of Section 1.4.1), is absorbed to ionize the major atmospheric gases O_2, O, and N_2 over the approximate height range 95–250 km, and this is what heats the thermosphere. Though the amount absorbed is only about 3 mW m^{-2} at solar minimum (more at solar maximum), a small amount of heat may raise the temperature considerably at great height because the air density is small. Indeed, at the greater altitudes the heating rate and the specific heat are both proportional to the gas concentration, and then the rate of increase in temperature is actually independent of height.

At high latitude, heating associated with the aurora – items (b) and (c) – is important during storms. Joule heating by electric currents is greatest at 115–130 km. Auroral electrons heat the atmosphere mainly between 100 and 130 km.

Losses

The principal mechanism of heat loss from the upper atmosphere is radiation, particularly in the infra-red. Emission by oxygen at 63 μm is important, as are spectral bands of the radical OH and the visible airglow from oxygen and nitrogen. The mesosphere is cooled by radiation from CO_2 at 15 μm and from ozone at 9.6 μm, though during the long days of the polar summer the net effect can be heating instead of cooling.

Transport

The thermal balance and temperature profile of the upper atmosphere are also affected by processes of heat transport. At various levels conduction, convection, and radiation all come into play.

Radiation is the most efficient process at the lowest levels, and the atmosphere is in radiative equilibrium between 30 and 90 km. *Eddy diffusion*, or convection, also operates below the turbopause (at about 100 km), and allows heat to be carried down into the mesosphere from the thermosphere. This flow represents a major loss of heat from the thermosphere but is a minor source for the mesosphere.

In the thermosphere (above 150 km) thermal conduction is efficient because of the low pressure and the presence of free electrons. The large thermal conductivity ensures that the thermosphere is isothermal above 300 or 400 km, though the thermospheric temperature varies greatly from time to time. *Chemical transport* of heat occurs when an ionized or dissociated species is created in one place and recombines in another. The mesosphere is heated in part by the recombination of atomic oxygen created at a higher level. There can also be horizontal heat transport by large-scale winds, which can affect the horizonal distribution of temperature in the thermosphere.

The balance amongst these various processes produces an atmosphere with two hot regions, one at the stratopause and one in the thermosphere. The thermospheric temperature, in particular, undergoes strong variations daily and with the sunspot cycle, both due to the changing intensity of solar radiation.

1.2.5 Composition

The upper atmosphere is composed of various major and minor species. The former are the familiar oxygen and nitrogen in molecular or atomic forms, or helium and hydrogen at the greater heights. The minor constituents are other molecules that may be present as no more than mere traces, but in some cases they can exert an influence far beyond their numbers.

Major species

The constant mixing within the turbosphere results in an almost constant proportion of major species up to 100 km, essentially the mixture as at ground-level called "air", although complete uniformity cannot be maintained if there are sources and sinks for particular species. Molecular oxygen is dissociated to atomic oxygen by ultra-violet radiation between 102.7 and 175.9 nm:

$$O_2 + h\nu \rightarrow O + O, \tag{1.12}$$

where $h\nu$ is a quantum of radiation. An increasing amount of O appears above 90 km. The atomic and molecular forms are present in equal concentrations at about 125 km, and above that the atomic form increasingly dominates. Nitrogen is not directly dissociated to the atomic form in the atmosphere, though it does appear as a product of other reactions.

Above the turbopause mixing is less important than diffusion, and then each component takes an individual scale height depending on its relative atomic or molecular mass ($H = kT/(mg)$). Because the scale heights of the common gases vary over a wide range – H = 1, He = 4, O = 16, $N_2 = 28$, $O_2 = 32$ – the relative composition of the thermosphere is a marked function of height, the lighter gases becoming progressively more abundant as illustrated in Figure 1.3. Atomic oxygen dominates at a height of several hundred kilometers. Above that is the

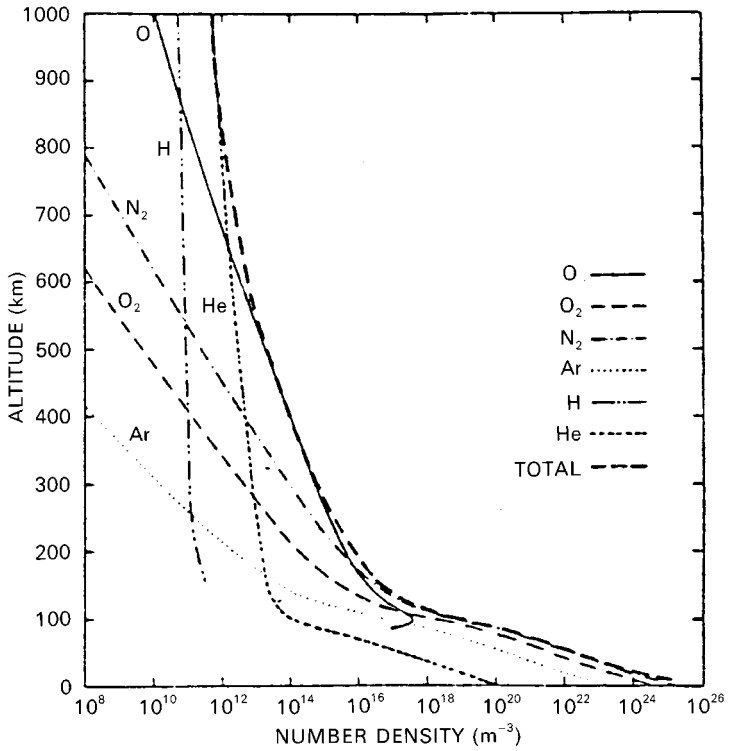

Figure 1.3. Atmospheric composition to 1000 km for a typical temperature profile. (*US Standard Atmosphere*, 1976.)

heliosphere, where helium is the most abundant, and eventually hydrogen becomes the major species in the protonosphere. Because the scale height also depends on the temperature, so do the details of the composition. The protonosphere starts much higher in a hot thermosphere, and the heliosphere may be absent from a cool one.

Two of the important species of the upper atmosphere, helium and hydrogen, are no more than minor species in the troposphere. Helium comes from radioactive decay in the Earth's crust. It diffuses up through the atmosphere, eventually escaping into space. The source of atomic hydrogen is the dissociation of water vapor near the turbopause from where it, also, flows constantly up through the atmosphere.

Minor species

Water, carbon dioxide, oxides of nitrogen, ozone, and alkali metals are all minor species of the atmosphere, but not all of them are significant for the ionosphere.

Water does not have the same dominating influence in the upper atmosphere as in the troposphere. It is important nevertheless, first as a source of hydrogen, and second because it causes ions to be hydrated below the mesopause. Carbon dioxide, also, plays a part in the chemistry of the D region.

Nitric oxide (NO), on the other hand, makes an important contribution to the lower ionosphere since it is ionized by the intense Lyman-α line of the solar spectrum and is thereby responsible for much of the ionospheric D region at middle latitudes (Section 1.4.3). The chemical story of NO is complicated because several production and loss mechanisms are at work and the distribution is affected by the dynamics of the mesosphere.

Nitric oxide in the mesosphere comes from two sources. One source is in the stratosphere and involves the oxidation of nitrous oxide (N_2O) by excited atomic oxygen. The second one peaks in the thermosphere, at 150–160 km, and involves a reaction with neutral or ionized atomic nitrogen, for example

$$N^* + O_2 \rightarrow NO + O, \tag{1.13}$$

where the * indicates an excited state. The resulting NO diffuses down to the mesosphere by molecular and then by eddy diffusion. Loss by photodissociation and recombination, aided by the effect of the low temperature at the mesopause, is sufficient to create a minimum at 85–90 km. The diffusion is weaker in the summer, and that is when the minimum is most marked. The depth of the minimum also varies with latitude.

The production of these atomic-nitrogen species is closely linked to ionization processes, and it is estimated that 1.3 NO molecules are produced on average for each ion produced. The concentration of nitric oxide therefore varies with time of day, latitude, and season. It is 3–4 times greater at high latitude than it is at middle latitude, and more variable. The production rate increases dramatically during particle precipitation events, and this is plainly an important mechanism in the high-latitude ionosphere.

The ozonosphere peaks between heights of 15 and 35 km, well below the ionosphere. The small amounts of ozone that occur in the mesosphere are involved in certain reacions in the D region, but we shall not be particularly concerned with them in this monograph. It is, however, of some general interest that there is a reaction between ozone and nitric oxide that tends to remove ozone at mesospheric levels. Thus,

$$\begin{aligned} &O_3 + NO \rightarrow O_2 + NO_2 \\ &O + NO_2 \rightarrow O_2 + NO \\ &\hline \\ &O_3 + O \rightarrow 2O_2. \end{aligned} \tag{1.14}$$

The net result, in the presence of atomic oxygen, is a catalytic conversion of ozone back to molecular oxygen. In this way the ozone concentration is affected by the natural production of nitric oxide discussed above.

Metallic atoms are introduced into the atmosphere in meteors, whose flux over the whole Earth amounts to 44 metric tons per day. In the ionized state, metals

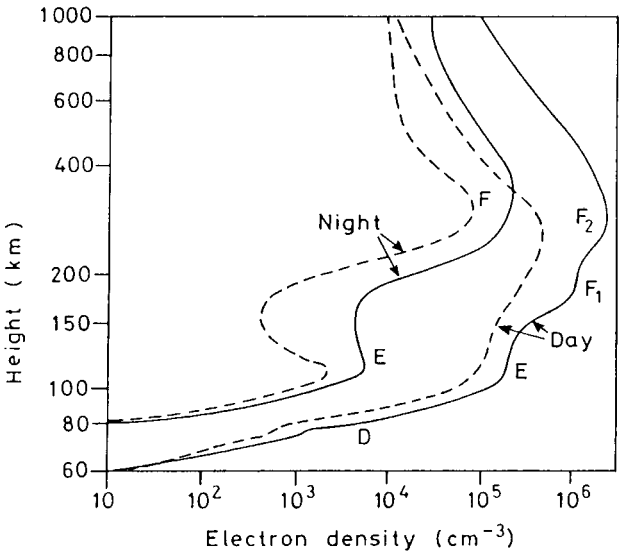

Figure 1.4. Typical vertical profiles of electron density in the mid-latitude ionosphere: ——, sunspot maximum; and – – –, sunspot minimum. (After W. Swider, Wallchart *Aerospace Environment*, US Air Force Geophysics Laboratory.)

such as sodium, calcium, iron, and magnesium are significant to the aeronomy of the lower ionosphere in various ways, but they will not be of great concern to us at high latitudes.

1.3 Physical aeronomy

1.3.1 Introduction

The topic of *physical aeronomy* covers the physical considerations governing the formation and shape of an ionospheric layer. The detailed photochemical processes which are involved in a particular case are generally considered under *chemical aeronomy*; however, we shall include such chemical details as we require in Section 1.4 as part of our description of the actual terrestrial ionosphere.

Typical vertical profiles of the ionosphere are shown in Figure 1.4. The identification of the regions was much influenced by their signatures on ionograms (see Section 4.2.1), which tend to emphasize inflections in the profile, and it is not necessarily the case that the various layers are separated by distinct minima. The main regions are designated D, E, F1, and F2, with the following daytime characteristics:

- D region, 60–90 km: electron density 10^8–10^{10} m^{-3} (10^2–10^4 cm^{-3});
- E region, 105–160 km: electron density of several times 10^{11} m^{-3} (10^5cm^{-3});

- F1 region, 160–180 km: electron density of several times 10^{11} to about 10^{12} m^{-3} (10^5–10^6 cm^{-3});
- F2 region, height of maximum variable around 300 km: electron density up to several times 10^{12} m^{-3} (10^6 cm^{-3}).

All these ionospheric regions are highly variable, and in particular there is generally a large change between day and night. The D and F1 regions vanish at night, and the E region becomes much weaker. The F2 region, however, tends to persist, though at reduced intensity.

The ionosphere is formed by the ionization of atmospheric gases such as N_2, O_2, and O. At middle and low latitude the required energy comes from solar radiation in the extreme ultra-violet (EUV) and X-ray parts of the spectrum. Once they have been formed, the ions and electrons tend to recombine and to react with other gaseous species to produce other ions. Thus there is a dynamic equilibrium in which the net concentration of free electrons (which, following standard practice, we call the *electron density*) depends on the relative speed of the production and loss processes. In general terms the rate of change of electron density is expressed by a *continuity equation*:

$$\partial N/\partial t = q - L - \mathrm{div}(N\mathbf{v}) \tag{1.15}$$

where q is the production rate (per unit volume), L is the rate of loss by recombination, and $\mathrm{div}(N\mathbf{v})$ expresses the loss of electrons by movement, $\mathbf{v}$ being their mean drift velocity.

If we consider a representative ionization and recombination reaction and neglect movements,

$$\mathrm{X} + h\nu \rightleftharpoons \mathrm{X}^+ + \mathrm{e}, \tag{1.16}$$

the "law of mass action" tells us that, at equilibrium,

$$[\mathrm{X}][h\nu] = \text{constant} \times [\mathrm{X}^+][\mathrm{e}], \tag{1.17}$$

where the square brackets signify concentrations. Thus, since $[\mathrm{e}] = [\mathrm{X}^+]$ for electrical neutrality,

$$[\mathrm{e}]^2 = \text{constant} \times [\mathrm{X}][h\nu]/[\mathrm{X}^+] \tag{1.18}$$

During the day the intensity of ionizing radiation varies with the elevation of the Sun, and the electron density responds to the variation of $[h\nu]$. At night the source of radiation is removed and so the electron density decays. From this simple model we can also see that the electron density must vary with altitude. The intensity of ionizing radiation increases with height but the concentration of ionizable gas [X]

decreases. It is reasonable to expect from this that the electron density will pass through a maximum at some altitude.

1.3.2 The Chapman production function

In 1931, S. Chapman developed a formula that predicts the form of a simple ionospheric layer and how it varies during the day. Although it is only partly successful in explaining the observed behavior of the terrestrial ionosphere – and this because of phenomena that it does not include – Chapman's formula is at the root of our modern understanding of the ionosphere and therefore it deserves a brief mention in this section.

At this stage we deal only with the rate of production of ionization (q), and the formula expressing this is the *Chapman production function*. In the simple treatment, which is sufficient for our purposes, it is assumed that

- the atmosphere is composed of a single species, exponentially distributed with constant scale height;
- the atmosphere is plane stratified: there are no variations in the horizontal plane;
- radiation is absorbed in proportion to the concentration of gas particles; and
- the absorption coefficient is constant: this is equivalent to assuming that we have monochromatic radiation.

The rate of production of ion–electron pairs at some level of the atmosphere can be expressed as the product of four terms:

$$q = \eta\sigma n I. \tag{1.19}$$

Here, I is the intensity of ionizing radiation and n is the concentration of atoms or molecules capable of being ionized by that radiation. For an atom or molecule to be ionized it must first absorb radiation, and the amount absorbed is expressed by the *absorption crossection*, σ: if the flux of incident radiation is I (J m^{-2} s^{-1}) then the total energy absorbed per unit volume of the atmosphere per unit time is $\sigma n I$. However, not all this energy will go into the ionization process, and the *ionization efficiency*, η, takes that into account, being the fraction of the absorbed radiation that goes into producing ionization.

The Chapman production function is usually written in a normalized form as

$$q = q_{m0}\exp(1 - z - \sec\chi\, e^{-z}). \tag{1.20}$$

Here, z is the reduced height for the neutral gas, $z = (h - h_{m0})/H$, H being the scale height. χ is the solar zenith angle, h_{m0} is the height of the maximum rate of production when the Sun is overhead (i.e. h_m when $\chi = 0$), and q_{m0} is the production

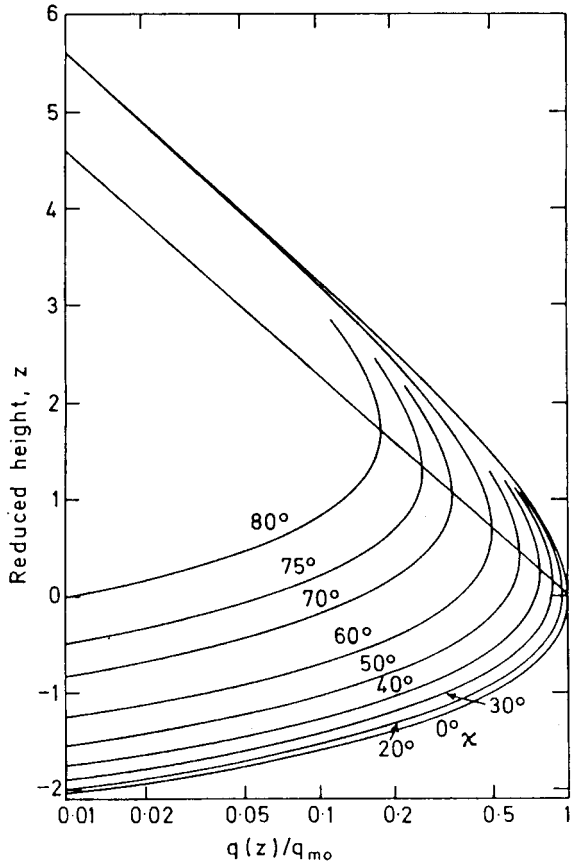

Figure 1.5. The Chapman production function. (After T. E. VanZandt and R. W. Knecht, in *Space Physics* (eds. LeGalley and Rosen). Wiley, 1964.)

rate at this altitude, also when the Sun is overhead. Derivations of equation (1.20) are given in many of the standard textbooks (see the list of further reading). Equation (1.20) can also be written

$$q/q_{m0} = \mathrm{e}\mathrm{e}^{-z}\mathrm{e}^{[-\sec\chi.\exp(-z)]}, \tag{1.21}$$

where the first term is a constant, the second expresses the height variation of the density of ionizable atoms, and the third is proportional to the intensity of the ionizing radiation.

Figure 1.5 illustrates some general properties of the production-rate profile. At a great height, where z is large and positive,

$$q \rightarrow q_{m0}\mathrm{e}\mathrm{e}^{-z}. \tag{1.22}$$

Thus the curves merge above the peak, becoming independent of χ and exhibiting an exponential decrease with height due to the decreasing density of the

neutral atmosphere. In the region well below the peak, when z is large and negative, the shape becomes dominated by the last term of Equation (1.21), producing a rapid cut-off. Thus, as predicted in the previous section, the production rate is limited by a shortage of ionizable gas at the greater altitudes and by a lack of ionizing radiation low down. On a plot of $\ln(q)$ against z all the curves are the same shape, but they are displaced upwards and to the left as the zenith angle, χ, increases.

The intensity of radiation in an absorbing atmosphere may be written as

$$I = I_{\text{inf}} \mathrm{e}^{-\tau} \tag{1.23}$$

where τ is the *optical depth*, which is equal to the absorption coefficient times the number of absorbing atoms down to the level considered:

$$\tau = \sigma N_{\text{T}}; \tag{1.24}$$

and I_{inf} is the intensity at great height. This leads to an important theorem:

The production rate is greatest at the level where the optical depth is unity.

From this general result there follow some particularly useful rules.

(1). The maximum production rate at a given value of χ is given by

$$q_{\text{m}} = \eta I_{\text{inf}}/(eH \sec \chi). \tag{1.25}$$

(2). The reduced height of the maximum depends on the solar zenith angle as

$$z_{\text{m}} = \ln(\sec \chi). \tag{1.26}$$

(3). The rate of production at this maximum is

$$q_{\text{m}} = q_{\text{m0}} \cos \chi. \tag{1.27}$$

These simple results are important in studies of the ionosphere because the maximum of a layer is the part most readily observed. From Equations (1.26) and (1.27) we see that a plot of $\ln(q_{\text{m}})$ against z_{m} is effectively a plot of $\ln(\cos \chi)$ against $\ln(\sec \chi)$, which obviously gives a straight line of slope -1. This line is shown in Figure 1.5.

The Chapman production function is important because it expresses fundamentals of ionospheric formation and of the absorption of radiation in any exponential atmosphere. Although real ionospheres may be more complicated, the Chapman theory provides an invaluable reference point for interpreting observations and a relatively simple starting point for ionospheric theory.

1.3.3 Principles of chemical recombination

Working out the rate of electron production is just the first step in calculating the electron density in an ionized layer, and the next step is to reckon the rates at which electrons are removed from the volume under consideration. This is represented in the continuity equation (1.15) by two further terms, one for the recombination of ions and electrons to reform neutral particles, and the other to account for movement of plasma into or out of the volume. We deal first with the principles of chemical recombination. The question of which individual reactions are most important in different parts of the ionosphere will be addressed in Section 1.4.

First we assume that the electrons recombine directly with positive ions and that no negative ions are present: $X^+ + e \rightarrow X$. Then the rate of electron loss is

$$L = \alpha[X^+]N_e = \alpha N_e^2 \tag{1.28}$$

where N_e is the electron density (equal to the ion density $[X^+]$) and α is the *recombination coefficient*. At equilibrium, therefore,

$$q = \alpha N_e^2. \tag{1.29}$$

The equilibrium electron density is proportional to the square root of the production rate, which may be replaced by the Chapman production function (1.20) to get the variation of electron density with height and solar zenith angle. In particular, it is seen that the electron density at the peak of the layer varies as $\cos^{1/2}\chi$:

$$N_m = N_{m0}\cos^{1/2}\chi. \tag{1.30}$$

A layer with these properties is called an *α-Chapman layer*.

If one is concerned particularly with electron loss, then attachment to neutral particles to form negative ions can itself be regarded as another type of electron-loss process. In fact, as we shall see, this becomes the dominant type at somewhat higher levels of the ionosphere (though by a different process). Without at this stage specifying chemical details, we can see that the attachment type of reaction can be written $M + e \rightarrow M^-$, and the rate of electron loss is $L = \beta N$, where β is the *attachment coefficient*. The loss rate is now linear with N because the neutral species M is assumed to be by far the more numerous, in which case removing a few of them has no significant effect on their total number and [M] is effectively constant.

At equilibrium,

$$q = \beta N_e \tag{1.31}$$

and taking q from the Chapman production function as before shows that the peak electron density now varies as

$$N_m = N_{m0} \cos \chi. \tag{1.32}$$

Such a layer is a β-Chapman layer.

This simple formulation assumes that β does not vary with height, though this restriction does not affect the validity of Equation (1.31) at a given height.

In fact β is expected to vary with height because it depends on the concentration of the neutral molecules (M), and this has important consequences for the form of the terrestrial ionosphere. It is known that electron loss in the F region occurs in a two-stage process:

$$X^+ + A_2 \rightarrow AX^+ + A \tag{1.33}$$

$$AX^+ + e \rightarrow A + X \tag{1.34}$$

in which A_2 is one of the common molecular species such as O_2 and N_2. The first step moves the positive charge from X to AX, and the second one dissociates the molecular ion through recombination with an electron, a *dissociative-recombination* reaction. The rate of Equation (1.33) is $\beta[X^+]$ and that of (1.34) is $\alpha[AX^+]N_e$. At low altitude β is large, (1.33) goes quickly and all X^+ is rapidly converted to AX^+; the overall rate is then governed by the rate of (1.34), giving an α-type process because $[AX^+] = N_e$ for neutrality. At a high altitude β is small, and (1.33) is slow and controls the overall rate. Then $[X^+] = N_e$ and the overall process appears to be of β-type. As height increases, the reaction type therefore alters from α-type to β-type. The reaction scheme represented by Equations (1.33) and (1.34) leads to equilibrium given by

$$\frac{1}{q} = \frac{1}{\beta(h) N_e} + \frac{1}{\alpha N_e^2}, \tag{1.35}$$

where q is the production rate as before. The change from α- to β-type behaviour occurs at height h_t where

$$\beta(h_t) = \alpha N_e. \tag{1.36}$$

In the lower ionosphere there are also significant numbers of negative ions. Electrical neutrality then requires $N_e + N_- = N_+$, where N_e, N_- and N_+ are, respectively, the concentrations of electrons, negative ions, and positive ions. Since the negative and positive ions may also recombine with each other, the overall balance between production and loss is now expressed by

$$q = \alpha_e N_e N_+ + \alpha_i N_- N_+, \tag{1.37}$$

α_e and α_i being recombination coefficients for the reactions of positive ions with electrons and negative ions, respectively. The ratio between negative-ion and electron concentrations is traditionally represented by λ – which has nothing to do with wavelength! In terms of λ, $N_- = \lambda N_e$ and $N_+ = (1+\lambda)N_e$, and thus

$$q = (1+\lambda)(\alpha_e + \lambda\alpha_i)N_e^2, \tag{1.38}$$

which, in cases for which $\lambda\alpha_i \ll \alpha_e$, becomes

$$q = (1+\lambda)\alpha_e N_e^2. \tag{1.39}$$

In the presence of negative ions the equilibrium electron density is still proportional to the square root of the production rate but its magnitude is changed. The term

$$(1+\lambda)(\alpha_e + \lambda\alpha_i)$$

is often called the *effective recombination coefficient*. As we shall see in Section 1.4.3, the chemistry of the D region is complicated because of the presence of many kinds of positive and negative ions.

1.3.4 Vertical transport

Diffusion

The final term of the continuity equation (1.15) represents changes of electron and ion density at a given location due to bulk movement of the plasma. Such movements can have various causes and can occur in the horizontal and the vertical planes in general, but since our present emphasis is on the overall vertical structure of the ionosphere, we shall concentrate here on the vertical movement of ionization, which, indeed, is very important in the F region. We assume now that photochemical production and loss are negligible in comparison with the effect of movements, and then the continuity equation becomes

$$\frac{\mathrm{d}N}{\mathrm{d}t} = -\frac{\partial(wN)}{\partial h}, \tag{1.40}$$

where w is the vertical drift speed and h is the height.

We now suppose that this drift is entirely due to diffusion of the gas, and then we can put

$$w = -\frac{D}{N}\frac{\partial N}{\partial h}, \tag{1.41}$$

D being the *diffusion coefficient*. This equation simply states that the bulk drift of a gas is proportional to its pressure gradient, and it effectively defines the diffu-