The High-Latitude Ionosphere and its Effects on Radio Propagation

R. D. Hunsucker, J. K. Hargreaves

Book DOI: http://dx.doi.org/10.1017/CBO9780511535758

Online ISBN: 9780511535758

Hardback ISBN: 9780521330831

Paperback ISBN: 9780521041362

Chapter

Chapter 1 - Basic principles of the ionosphere pp. 1-60

Chapter DOI: http://dx.doi.org/10.1017/CBO9780511535758.003

Cambridge University Press

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

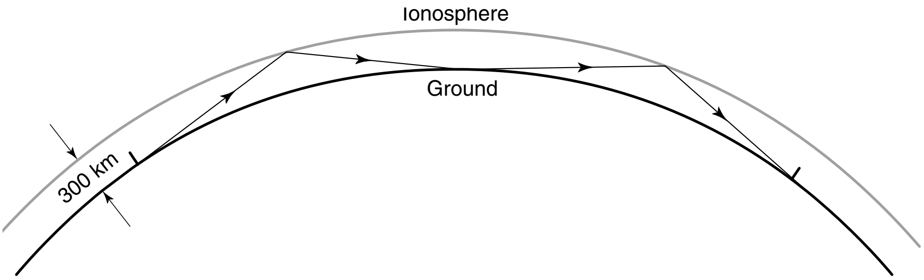

Figure 1.1. Long distance propagation by multiple hops between the ionosphere and the ground.

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

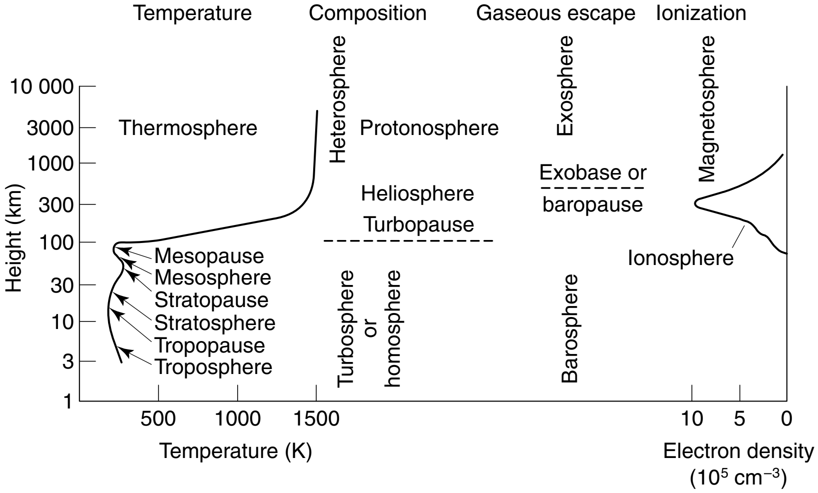

Figure 1.2. Nomenclature of the upper atmosphere based on temperature, composition, mixing, and ionization. (J. K. Hargreaves, *The Solar–Terrestrial Environment*. Cambridge University Press, 1992.)

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:

$$O_2 + h\nu \rightarrow O + O, \qquad (1.12)$$

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

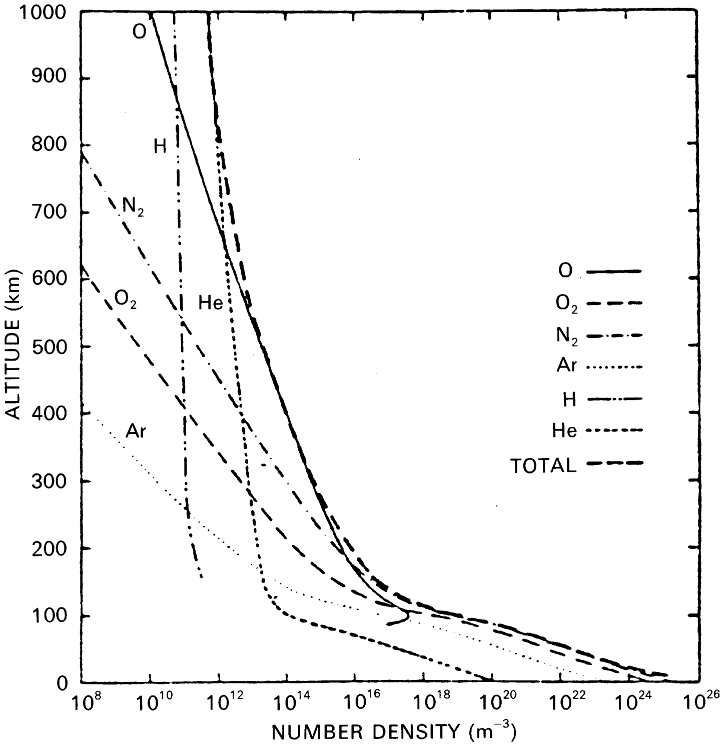

Figure 1.3. Atmospheric composition to 1000 km for a typical temperature profile. (*US Standard Atmosphere*, 1976.)

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,

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

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

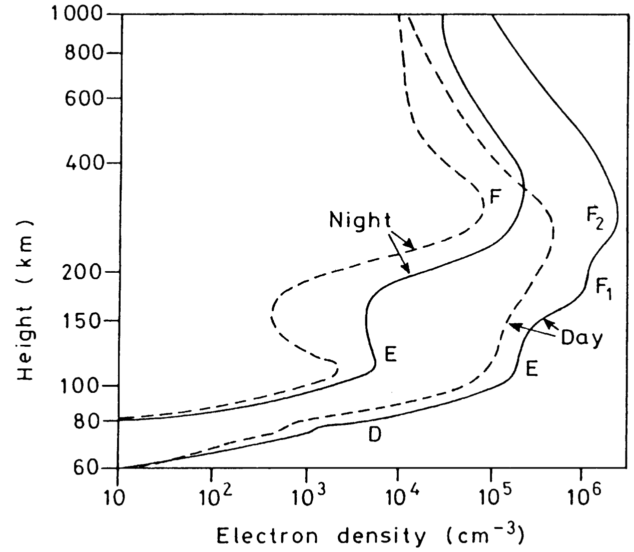

Figure 1.4. Typical vertical profiles of electron density in the mid-latitude ionosphere: ——, sunspot maximum; and – – –, sunspot minimum. (After W. Swider, Wallchart *Aerospace Environment*, US Air Force Geophysics Laboratory.)

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} ($10^5 cm^{-3}$);

- 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,

$$\mathrm{X} + h\nu \leftrightharpoons \mathrm{X}^+ + \mathrm{e}. \tag{1.16}$$

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

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

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

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

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:

$$q=\eta\sigma nI. \tag{1.19}$$

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 σnI. 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

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

Here, z is the reduced height for the neutral gas, $z=(h-h_{\mathrm{m0}})/H$, 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

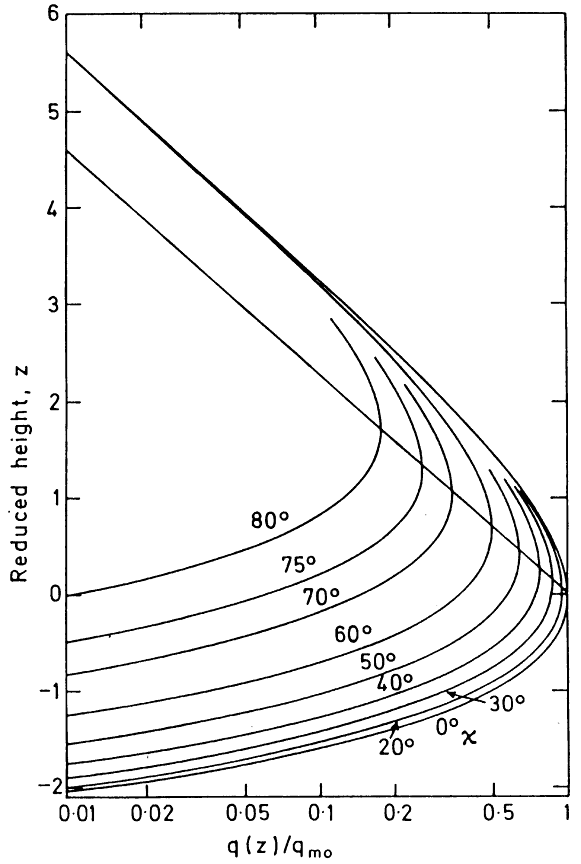

Figure 1.5. The Chapman production function. (After T. E. VanZandt and R. W. Knecht, in *Space Physics* (eds. LeGalley and Rosen). Wiley, 1964.)

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

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

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,

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

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

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

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

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

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

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

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

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

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

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

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 $\ln(q_{\mathrm{m}})$ 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

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

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:

$$\mathrm{X}^+ + \mathrm{A}_2 \rightarrow \mathrm{AX}^+ + \mathrm{A} \tag{1.33}$$

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

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 $\beta[\mathrm{X}^+]$ and that of (1.34) is $\alpha[\mathrm{AX}^+]N_{\mathrm{e}}$. 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 $[\mathrm{AX}^+] = N_{\mathrm{e}}$ for neutrality. At a high altitude β is small, and (1.33) is slow and controls the overall rate. Then $[\mathrm{X}^+] = N_{\mathrm{e}}$ 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

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

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

$$\beta(h_{\mathrm{t}}) = \alpha N_{\mathrm{e}}. \tag{1.36}$$

In the lower ionosphere there are also significant numbers of negative ions. Electrical neutrality then requires $N_{\mathrm{e}} + N_- = N_+$, 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

$$q = \alpha_{\mathrm{e}} N_{\mathrm{e}} N_+ + \alpha_{\mathrm{i}} N_- N_+, \tag{1.37}$$

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 diffusion coefficient whose dimensions are (length)^2/time. From kinetic theory (equating the driving force due to the pressure gradient to the drag force due to collisions as a minority gas diffuses through a stationary majority gas) the diffusion coefficient may be derived in its simplest form as $D=kT/(\mathrm{m}\nu)$. Here k is Boltzmann's constant, T the temperature, m the particle mass and ν the collision frequency.

In the present case the minority gas is the plasma composed of ions and electrons, and the majority gas is the neutral air. However, for drift in the vertical direction the force of gravity also acts on each particle, adding to (or subtracting from) the drag force, and in this case we obtain

$$w=-(D/N)(\mathrm{d}N/\mathrm{d}h+N/H_\mathrm{N}) \tag{1.42}$$

for the upward speed instead of (1.41). Substitution into the continuity equation then gives

$$\frac{\mathrm{d}N}{\mathrm{d}t}=\frac{\partial}{\partial h}\left[D\left(\frac{\mathrm{d}N}{\mathrm{d}h}+\frac{N}{H_\mathrm{N}}\right)\right]. \tag{1.43}$$

This is the basic equation that has to be satisfied by the time and height variations of those regions (specifically the upper F region and the protonosphere) where ion production and recombination are both sufficiently small.

In this equation the scale height H_N merely represents the value of $kT/(mg)$, and does not necessarily describe the actual height distribution. This is given by the *distribution height*, defined as

$$\delta=\left(-\frac{1}{N}\frac{\mathrm{d}N}{\mathrm{d}h}\right)^{-1}. \tag{1.44}$$

Using Equations (1.43) and (1.44) we can easily see that δ is equal to the scale height at equilibrium.

A complication is introduced by the fact that a plasma is composed of two minority species, ions and electrons, which have opposite charges and very different masses. Initially the ions, being heavier, tend to settle away from the electrons, but the resulting separation of electric charge produces an electric field, **E**, and a restoring force $e\mathbf{E}$ on each charged particle. This electrostatic force also affects the drift of the plasma. This problem is handled by writing separate equations for each species and including the electrostatic force on each. We assume

(1). that the electron mass is small compared with the ion mass;

(2). that ion and electron number densities are equal; and

(3). that both species drift at the same speed;

and then it can be shown that Equations (1.42) and (1.43) are still valid for a plasma if one replaces D and H by

$$D_\mathrm{p}=k(T_\mathrm{e}+T_\mathrm{i})/(m_\mathrm{i}\nu_\mathrm{i}) \tag{1.45}$$

and

$$H_p = k(T_e + T_i)/(m_i g), \tag{1.46}$$

respectively known as the *ambipolar* or *plasma diffusion coefficient* and the *plasma scale height*.

In that part of the ionosphere where plasma diffusion is important, the electron temperature usually exceeds the ion temperature. However, taking $T_e = T_i$ by way of illustration, we see that the plasma diffusion coefficient and scale height are then just double those of the neutral gas at the same temperature. Effectively, the light electrons have the effect of halving the ion mass since the two species cannot separate very far. At equilibrium $\mathrm{d}N/\mathrm{d}h = -N/H\mathrm{p}$ and the plasma is exponentially distributed as

$$N/N_0 = \exp(-h/H_p) \tag{1.47}$$

with scale height H_p. Note that this distribution has the same form as the upper part of a Chapman layer but with (about) twice the scale height.

If the plasma is not in equilibrium the distribution changes with time at a rate depending on the value of the diffusion coefficient, which, since it depends on the relevant collision frequency, increases with altitude. If H is the scale height of the neutral gas, then the height variation of the diffusion coefficient can be written as

$$D = D_0 \exp(h - h_0)/H \tag{1.48}$$

where D_0 is the value of D at a height h_0. Thus, diffusion becomes ever more important at greater heights as the photochemistry becomes less important.

Another consequence of the height variation of D is that it leads to a second solution of Equation (1.43) for the case $\mathrm{d}N/\mathrm{d}t = 0$. Substituting $D = D_0 \exp(h - h_0)/H$ and $N = N_0 \exp - (h - h_0)/\delta$ into (1.43) and rearranging, gives

$$\frac{\mathrm{d}N}{\mathrm{d}t} = DN\left(\frac{1}{\delta} - \frac{1}{H_p}\right)\left(\frac{1}{\delta} - \frac{1}{H}\right). \tag{1.49}$$

If $\mathrm{d}N/\mathrm{d}t = 0$ this has two solutions. The first, $\delta = H_p$, is diffusive equilibrium as has already been pointed out, and in this case the vertical drift speed (Equation (1.41)) is $w = -\mathrm{D}(-1/\delta + 1/H_p) = 0$.

The second solution is $\delta = \mathrm{H}$ (H being the scale height of the neutral gas, governing the diffusion coefficient). Here, $\mathrm{d}N/\mathrm{d}t = 0$ as before, but the drift speed is

$$w = D\left(-\frac{1}{\delta} + \frac{1}{H_p}\right) = D\left(-\frac{1}{H} + \frac{1}{H_p}\right), \tag{1.50}$$

which is not zero since $H_p > H$. The upward flow of plasma

$$Nw = ND(1/H - 1/H_p), \tag{1.51}$$

and in fact this is independent of height when $\delta = \mathrm{H}$ because the height variations of D and of N cancel out. Thus, this second solution represents an unchanging distribution of electron density and a constant outflow of plasma.

The effect of a neutral-air wind

Since the flow of ionospheric plasma is constrained by the geomagnetic field, the exact effect varies with latitude. One consequence is that, at middle latitudes, the height distribution of ionization is affected by the neutral-air wind which flows in the thermosphere. Suppose that the wind speed in the magnetic meridian is U and the magnetic dip angle is I. Then the component of the neutral wind along the direction of the magnetic field is $U_{\|} = U\cos I$, and the plasma tends to move in the same way. This motion, along the magnetic field, has a vertical component

$$W = U_{\|}\sin I = \tfrac{1}{2}.\ U\sin(2I). \qquad (1.52)$$

Thus, a horizontal wind in the thermosphere tends to move the ionosphere up or down depending on its direction of flow. The effect is greatest where the magnetic dip angle is 45°. The consequences both for the height and for the magnitude of the peak of the F region can be significant (Section 1.4.5).

1.4 The main ionospheric layers

1.4.1 Introduction

The physical principles which govern the intensity and form of an ionospheric layer were outlined in Section 1.3. To work out what the actual ionosphere should be like on Earth or any other planet, we would have to consider the terms in Equation (1.19) ($q = \eta\sigma nI$) in detail to get the ion production rate, specify the ion chemistry to obtain values for the loss coefficients in Equations (1.29) and (1.31) ($q = \alpha N_e^2$ and $q = \beta N_e$), and, at the higher levels, consider the diffusion coefficient (Equation (1.46)) and take movements into account. We should then require to know about the neutral atmosphere: its composition and physical parameters such as density and temperature. Then we should need full information on the solar spectrum and any fluxes of energetic particles able to ionize the constituents of the atmosphere.

Knowing which gases could be ionized by the incident radiation, we could then determine the ionization rate of each species and sum over all wavelengths and all gases to get the total production rate in a given volume (q). If the loss processes indicated rapid attainment of equilibrium, the electron density (N_e) would be given by Equation (1.29) or (1.31). Otherwise a more complex computation would be required. (Mathematical modeling of the high-latitude ionosphere is discussed in Section 9.2.2.) There is no need to go into all these details here, but a few important points will be made.

Table 1.1 lists the ionization potentials of various atmospheric gases. To be ionized a species must absorb a quantum of radiation whose energy exceeds the ionization potential. Since the energy of a quantum of wavelength λ is $E=hc/\lambda$, there is a maximum wavelength of radiation that is able to ionize any particular gas. These values are included in Table 1.1. For easy reference the wavelengths are given both in ångström units and in nanometers.

These values of λ_{max} immediately identify the relevant parts of the solar spectrum as the X-ray (0.1–17 nm, 1–170 Å) and EUV, (17–175 nm, 170–1750 Å), emissions which come from the solar chromosphere and corona.

The value of the absorption cross-section, σ, generally increases with increasing wavelength up to λ_{max} and then falls rapidly to zero. There is no ionization at all by any radiation with wavelength exceeding λ_{max}, regardless of its intensity.

The ionization efficiency, η, is such that, for atomic species, all the absorbed energy goes into ion production at the rate of one ion–electron pair for every 34 eV of energy. The energy is inversely proportional to the wavelength, and a convenient formula in terms of wavelength is

$$\eta=360/\lambda\ (\text{Å}). \tag{1.53}$$

The Chapman theory (Section 1.3.2) shows that the production rate is a maximum at the level where the optical depth, $\sigma n H \sec\chi$, is unity. If the absorption at a given wavelength is due to several species, then the condition for maximum production is

$$\sum_i \lambda_i n_i H_i \sec\chi = 1.$$

Table 1.1. *Ionization potentials*

		Maximum wavelength λ_{max}	
Species	Ionization potential I (eV)	(Å)	(nm)
NO	9.25	1340	134.0
O_2	12.08	1027	102.7
H_2O	12.60	985	98.5
O_3	12.80	970	97.0
H	13.59	912	91.2
O	13.61	911	91.1
CO_2	13.79	899	89.9
N	14.54	853	85.3
H_2	15.41	804	80.4
N_2	15.58	796	79.6
A	15.75	787	78.7
Ne	21.56	575	57.5
He	24.58	504	50.4

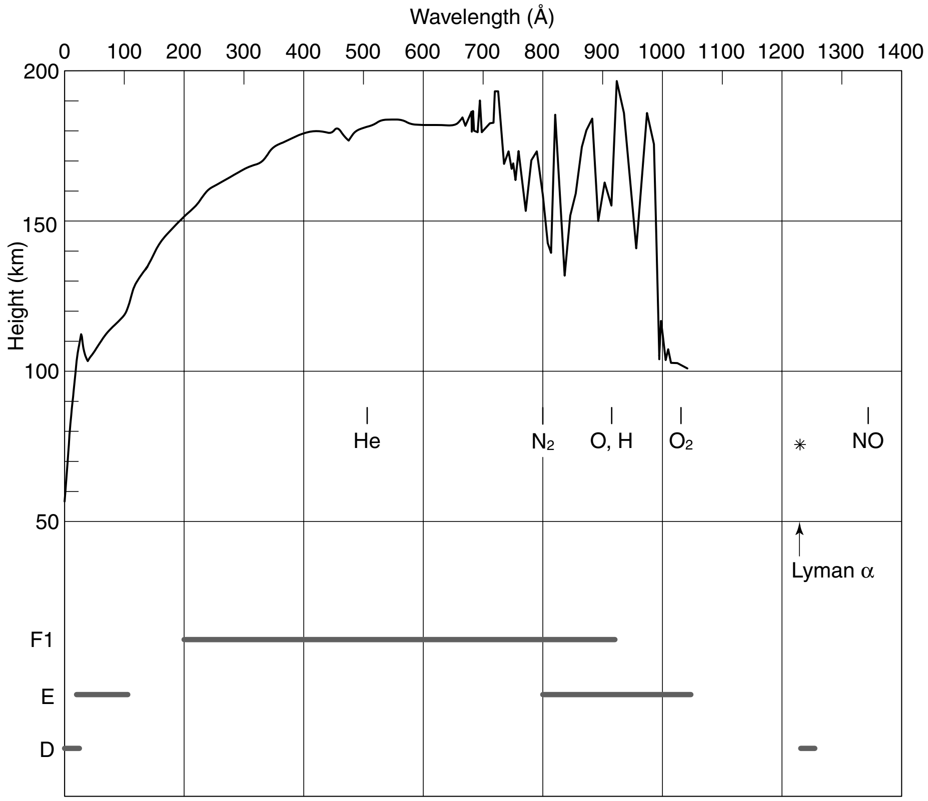

Figure 1.6. The height at which the optical depth reaches unity for radiation vertically incident on the atmosphere. Ionization limits for common gases are marked. (J. D. Mathews, private communication.) The ranges responsible for the major ionospheric layers are indicated below.

The height of unit optical depth in a model terrestrial atmosphere is given as a function of wavelength in Figure 1.6 and this, not the intensity of the ionizing radiation, is what determines the height of the ionospheric layers. This is an important point. It means, simply, that strongly absorbed radiation produces ionization high up, and that low-level ionization must be due to radiation that is more weakly absorbed in the atmosphere.

The simple theory of Section 1.3.2 deals with the shape and intensity of an ionosphere produced by monochromatic radiation acting on a single gas. On a real planet the effect of all gases at a given wavelength has to be considered and then, since the ionosphere is in effect a number of overlapping Chapman layers, the production rate due to all relevant wavelengths has to be summed at each height. The wavelength ranges giving the D, E, and F regions are summarized in Figure 1.6.

1.4.2 The E and F1 regions

Aeronomy

The E region which peaks at 105–110 km, and the F1 region at 160–180 km, are both fairly well understood. The F1 region is attributed to that part of the solar spectrum between about 200 and 900 Å, which is strongly absorbed in atomic oxygen, whose ionization limit is at 911 Å. The optical depth reaches unity from about 140 to 170 km. The band includes an intense solar emission line at 304 Å. The primary reaction products are O_2^+, N_2^+, O^+, He^+, and N^+, but subsequent reactions leave NO^+ and O_2^+ as the most abundant positive ions.

The E region is formed by the less strongly absorbed, and therefore more penetrating, parts of the spectrum. EUV radiation between 800 and 1027 Å (the ionization limit of O_2) is absorbed by molecular oxygen to form O_2^+. The band includes several important emission lines. At the short-wavelength end X-rays of 10–100 Å (1–10 nm) ionize all the atmospheric constituents. The main primary ions are N_2^+, O_2^+, and O^+, but the most numerous are again observed to be NO^+ and O_2^+. The intensity of solar X-rays varies over the solar cycle and they probably make little contribution to the E region at solar minimum.

Direct *radiative recombination* of the type

$$e+X^+\rightarrow X+h\nu \tag{1.54}$$

is slow relative to other reactions and is not significant in the normal E and F regions. *Dissociative recombination*, as

$$e+XY^+\rightarrow X+Y, \tag{1.55}$$

is 10^5 times faster (with a reaction coefficient of 10^{-13} m^3 s^{-1}) and, both in the E region and in the F region, the electron and ion loss proceeds via molecular ions. The main recombination reactions of the E region are therefore

$$\begin{aligned} &e+O_2^+\rightarrow O+O,\\ &e+N_2^+\rightarrow N+N,\\ &e+NO^+\rightarrow N+O. \end{aligned} \tag{1.56}$$

In the F region the principal primary ion is O^+, which is first converted to a molecular ion by a *charge-exchange* reaction

$$O^+ + O_2\rightarrow O_2^+ + O$$

or

$$O^+ + N_2\rightarrow NO^+ + N. \tag{1.57}$$

The molecular ion then reacts with an electron as in Equation (1.56), to give as the net result

$$e + O^+ + O_2 \rightarrow O + O + O$$

or

$$e + O^+ + N_2 \rightarrow O + N + N. \tag{1.58}$$

In the F1 region the overall reaction is controlled by the rate of the dissociative recombination.

Observations show that both the E and the F1 layers behave like, or almost like, α-Chapman layers (Equation (1.30)). On average the *critical frequency*, f_OE or f_0F1 (Section 3.4.2), varies with the solar zenith angle, χ, as $(\cos\chi)^{1/4}$, which means that the peak electron density, N_m, varies as $(\cos\chi)^{1/2}$. The exponent is subject to some variation and ranges between about 0.1 and 0.4 for the E region.

Given that the E region is an α-Chapman layer, the Chapman theory can be applied to determine the recombination coefficient (α) from observations, and this may be done using Equation (1.29):

(1). taking an observed electron density and an observed or computed production rate;

(2). by observing the rate of decay of the layer after sunset and assuming that $q = 0$; or

(3). by measuring the asymmetry of the diurnal variation about local noon, an effect sometimes called the *sluggishness* of the ionosphere, the time delay being given by

$$\tau = 1/(2\alpha N). \tag{1.59}$$

Such methods give values of α in the range 10^{-13}–10^{-14} m^3 s^{-1} (10^{-7}–10^{-8} cm^3 s^{-1}).

The night E layer

The E layer does not quite vanish at night, but a weakly ionized layer remains with electron density about 5×10^9 m^{-3} (against 10^{11} m^{-3} by day). One possible cause is meteoric ionization, though other weak sources might also contribute. Figure 1.7 shows speciman electron-density profiles of the E region for day and night, measured by incoherent-scatter radar.

Sporadic-E

The most remarkable anomaly of the E region is *sporadic-E*, often abbreviated to E_s. On ionograms sporadic-E is seen as an echo at constant height that extends to a higher frequency than is usual for the E layer; for example to above 5 MHz. Rocket measurements, and more recently incoherent-scatter radar, show that, at mid-latitude, these layers are very thin, perhaps less than a kilometer across. Examples are shown in Figure 1.8.

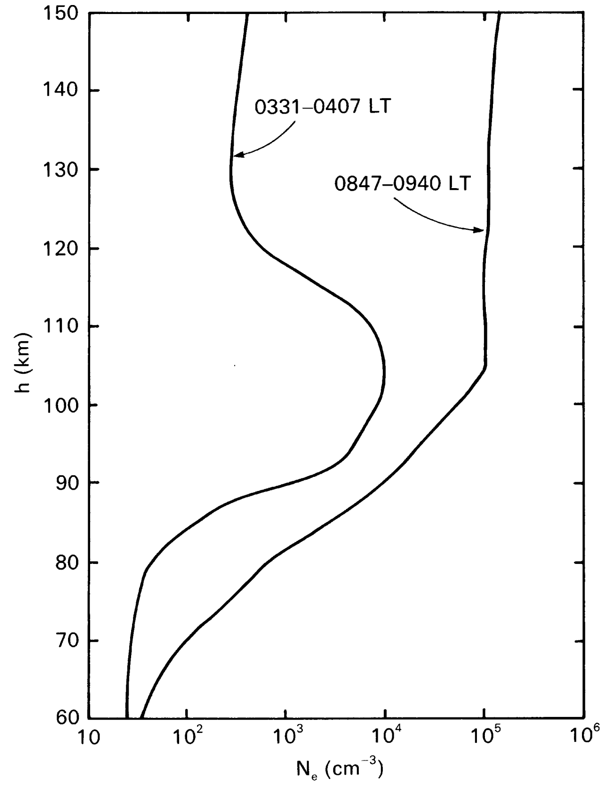

Figure 1.7. Speciman electron-density profiles of the E region for night and day, measured by the incoherent-scatter radar at Arecibo, Puerto Rico (18° N, 67° W), in January 1981. (J. D. Mathews, private communication.)

Figure 1.9 indicates the probability of occurrence of sporadic-E against time of day and season in three latitude zones:

- the equatorial zone, within 20° of the magnetic equator;
- the high-latitude zone, poleward of about 60° geomagnetic;
- and the temperate zone in between.

The high-latitude zone may be sub-divided into the auroral zone (approximately 60°–70° magnetic) and the polar cap (poleward of the auroral zone). A full classification of sporadic-E, particularly regarding its identification on ionograms, is given by Piggott and Rawer (1972). In general, sporadic-E exhibits little direct relationship with the incidence of solar ionizing radiation.

Sporadic-E tends to be particularly severe at low latitude. It occurs frequently during the daytime hours, often with sufficient intensity to reflect radio waves up to 10 MHz. A major cause is the occurrence of instabilities in the equatorial electrojet (Section 1.5.5).

The principal cause of sporadic-E at middle latitude is a variation of wind speed with height, a *wind shear*, which, in the presence of the geomagnetic field,

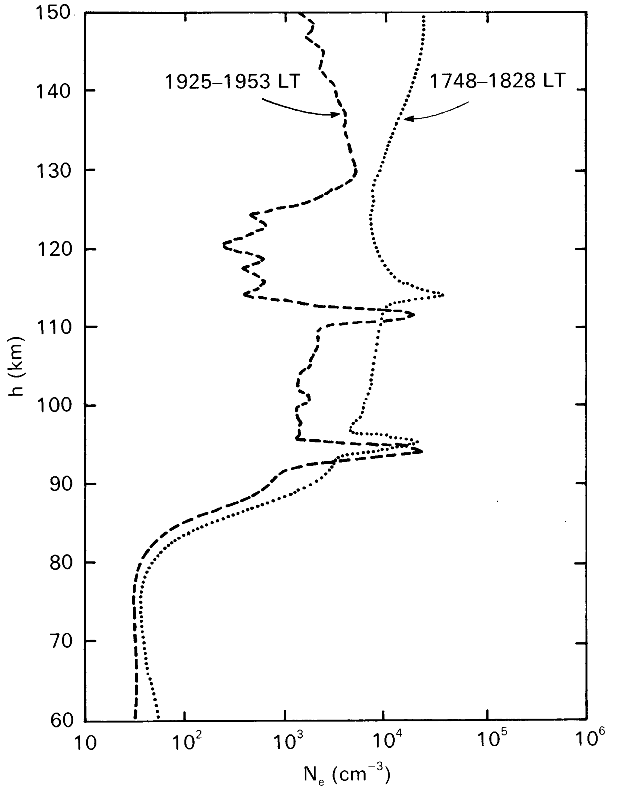

Figure 1.8. Some sporadic-E layers observed at Arecibo by incoherent-scatter radar, January 1981. (J. D. Mathews, private communication.)

acts to compress the ionization by a mechanism similar to that which allows the neutral-air wind in the thermosphere to raise or lower the F region (Section 1.3.4). The time scale of the process needs ions of relatively long life, and it is thought that these are metallic ions of meteoric origin such as Fe^+, Mg^+, Ca^+, and Si^+. Being atomic, these cannot recombine dissociatively and therefore their recombination coefficients are typical of the radiative process (10^{-18} m^3 s^{-1}), which gives them relatively long lifetimes. Temperate sporadic-E occurs at heights of 95–135 km, and the most probable height is 110 km. It occurs most frequently in summer daytime, with maxima in mid-morning and near sunset. The seasonal variation is complex. Its character changes abruptly at about 60° magnetic latitude, the boundary of auroral E_s.

The sporadic-E which occurs at high latitude is attributed to ionization by incoming energetic particles in the energy range 1–10 keV. It is mainly a night-time phenomenon, correlating to magnetic activity (Section 2.5.3), but not to sunspot activity as such. Clouds of auroral E_s drift at speeds between 200 and 3000 m s^{-1}, westward in the evening and eastward in the early morning, much like the aurora. The layer may be either "thick" or "thin". Within the polar caps sporadic-E has

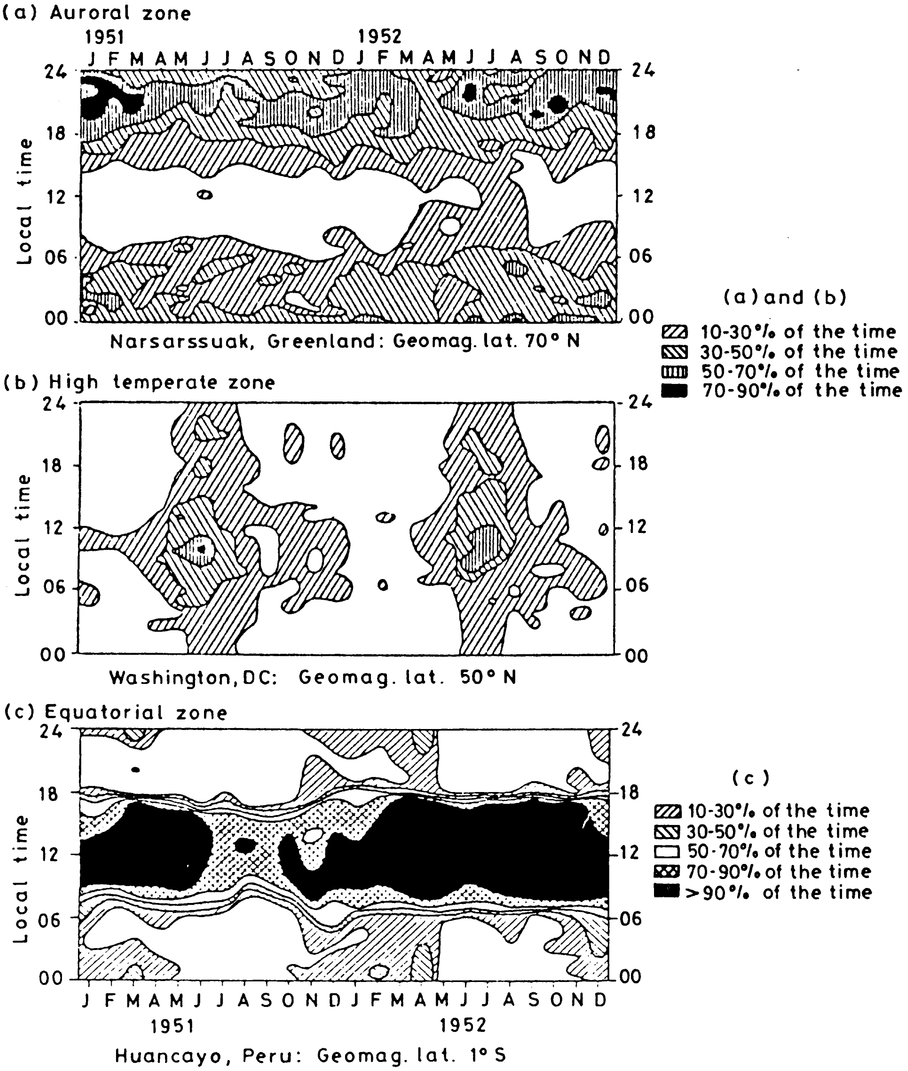

Figure 1.9. Diurnal and seasonal occurrence patterns for three kinds of sporadic-E. (a) The auroral kind maximizes at night but exhibits no seasonal variation. (b) The temperate kind peaks near noon in summer. (c) The equatorial kind occurs mainly by day but has no seasonal preference. (After E. K. Smith, *NBS Circular 582*, US National Bureau of Standards, 1957.)

a different character. It is weaker, and exhibits a negative correlation to magnetic activity. It takes the form of bands or ribbons extending across the polar cap in a roughly sunward direction. The properties and causes of sporadic-E have been reviewed in detail by Whitehead (1970). The high-latitude E region is discussed further in Section 6.5.

Sporadic-E is significant in radio propagation because it may reflect signals that would otherwise penetrate to the F region, though in some cases (for example the equatorial type) it is partly transparent. The irregularities within a sporadic-E layer can scatter radio waves if their dimensions are comparable to half a radio wavelength, and at times they may cause scintillation of trans-ionospheric signals, though F-layer irregularities are the more usual cause of this phenomenon.

The F1 ledge

The strange thing about the F1 region is that it does not always appear! In fact, real-height profiles show that it seldom exists as a distinct peak and for this reason it is more correctly called the *F1 ledge*. The ledge is more pronounced in summer and at sunspot minimum, and it is never seen in winter at sunspot maximum. The explanation is to be found by comparing h_t, the height at which transition between α-type and β-type recombination occurs, as discussed in Section 1.3.3, and h_m, the height of maximum electron-production rate. The F1 ledge appears only if $h_t > h_m$, and, since h_t depends on the electron density (Equation (1.36)), the ledge vanishes when the electron density is greatest.

1.4.3 The D region

Aeronomy

The D region of the ionosphere does not include a maximumum but is that part below about 95 km which is not accounted for by the processes of the E region. It is also the most complex part of the ionosphere from the chemical point of view. This is due, first, to the relatively high pressure, which causes minor as well as major species to be important in the photochemical reactions, and, second, because several different sources contribute to ion production.

The Lyman-α line of the solar spectrum at 1215 Å penetrates below 95 km and ionizes the minor species nitric oxide (NO), whose ionization limit is at 1340 Å. This is the main source at middle latitudes, though not necessarily at all heights. There is a smaller contribution from the EUV spectrum between 1027 and 1118 Å, which ionizes another minor constituent, molecular oxygen in an excited state. At the higher levels ionization of O_2 and N_2 by EUV, as in the E region, makes a contribution. Hard X-rays of 2–8 Å ionize all constituents, the most effect being therefore from the major species O_2 and N_2. Since the intensity of the solar X-ray emissions varies considerably from time to time, this source is sometimes a major one but at other times only minor. The lowest levels are dominated by cosmic-ray ionization, which continues by night as well as by day and affects the whole atmosphere down to the ground. The production rate due to cosmic rays increases downward in proportion to the total air density, and, since the production from other sources is falling off, it is inevitable that the cosmic rays must come to dominate at some level. At high latitudes particles from the Sun or of auroral origin ionize the D region and at times they form the main source. We shall be particularly concerned with those sources and their effects later in the book.

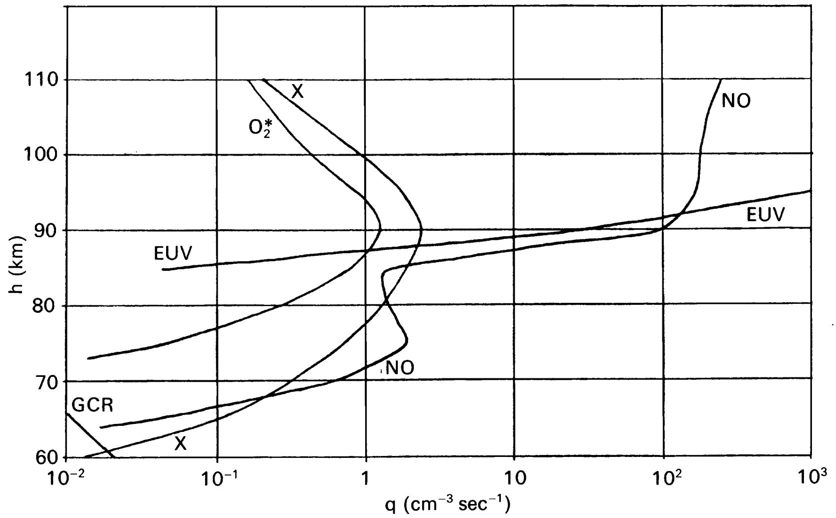

Figure 1.10. Calculated production rates at $\chi = 42°$ due to extreme ultra-violet (EUV), Lyman-α and nitric oxide (NO), X-rays (X), excited oxygen (O_2^*), and galactic cosmic rays (GCR). (J. D. Mathews, private communication.)

Clearly, the relative contributions of these different sources vary with latitude, time of day, and level of solar activity. By way of example, theoretical profiles of the production rate (for solar zenith angle 42° and a 10-cm solar flux of 165 units) are given in Figure 1.10. Note that all the sources mentioned above are significant and that their relative importance depends on the altitude. At greater solar zenith angles the contributions from Lyman-α and X-rays are reduced, and the cosmic rays become relatively more important below 70 km. The X-ray flux varies strongly with solar activity (by a factor of a hundred to a thousand) and is probably not significant in the D region at sunspot minimum.

These production-rate profiles are consistent with measurements of D-region electron densities (Figure 1.11). Friedrich and Torkar (1992) analyzed 164 electron-density profiles of the D region measured by rocket-based wave-propagation techniques (as in Section 4.3.4), to derive an empirical model covering a range of solar zenith angles. Figure 1.12 shows a set of profiles corresponding to a sunspot number of 60.

Following ionization, the primary ions in the D region are NO^+, O_2^+, and N_2^+, but the latter are rapidly converted to O_2^+ by the charge-exchange reaction

$$N_2^+ + O_2 \rightarrow O_2{}^+ + N_2, \tag{1.60}$$

leaving NO^+ and O_2^+ as the major ions. However, below 80 or 85 km, apparently the level of the mesopause, are detected heavier ions that are hydrated species such as $H^+.H_2O$, $H_3O^+.H_2O$, and hydrates of NO^+. These hydrates occur when the concentration of water vapor exceeds about 10^{15} m^{-3}. The level at which hydration first occurs is a natural boundary within the D region.

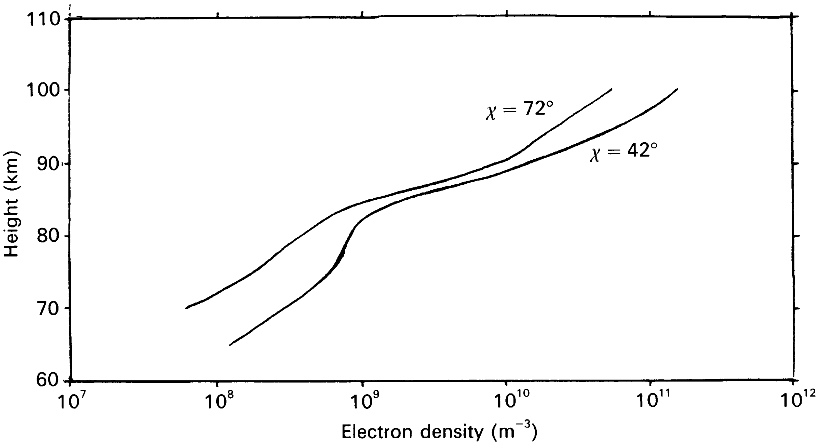

Figure 1.11. Electron-density profiles observed at Arecibo for two solar zenith angles. (J. D. Mathews, private communication.)

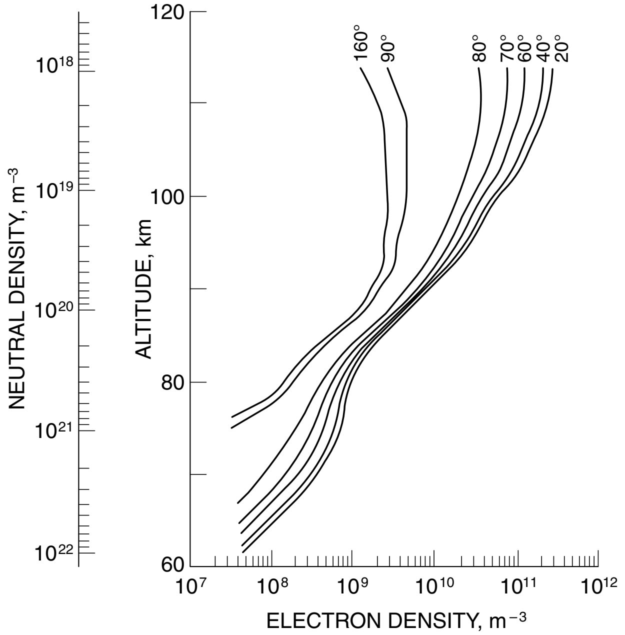

Figure 1.12. Electron-density profiles in the D region derived from rocket measurements for a range of solar zenith angles. The number density of the neutral air is also shown. (M. Friedrich and K. M. Torkar, *Radio Sci.* **27**, 945, 1992. Copyright by the American Geophysical Union.)

Where simple ions dominate, the loss process is dissociative recombination as in the E region, with a recombination coefficient of about 5×10^{-13} m^3 s^{-1}, the reaction of NO$^+$ being somewhat faster than that of O_2^+. In total the situation is much more complex, as illustrated in Figure 1.13. This scheme includes O_2^+, NO$^+$, O_4^+, hydrates and others, and has to be solved by means of a computer program. The hydrated ions, being larger molecules, have greater recombination rates than do the simple ions, of the order of 10^{-12}–10^{-11} m^3 s^{-1}, depending on their size. Thus the equilibrium electron density is relatively smaller in regions where hydrates dominate.

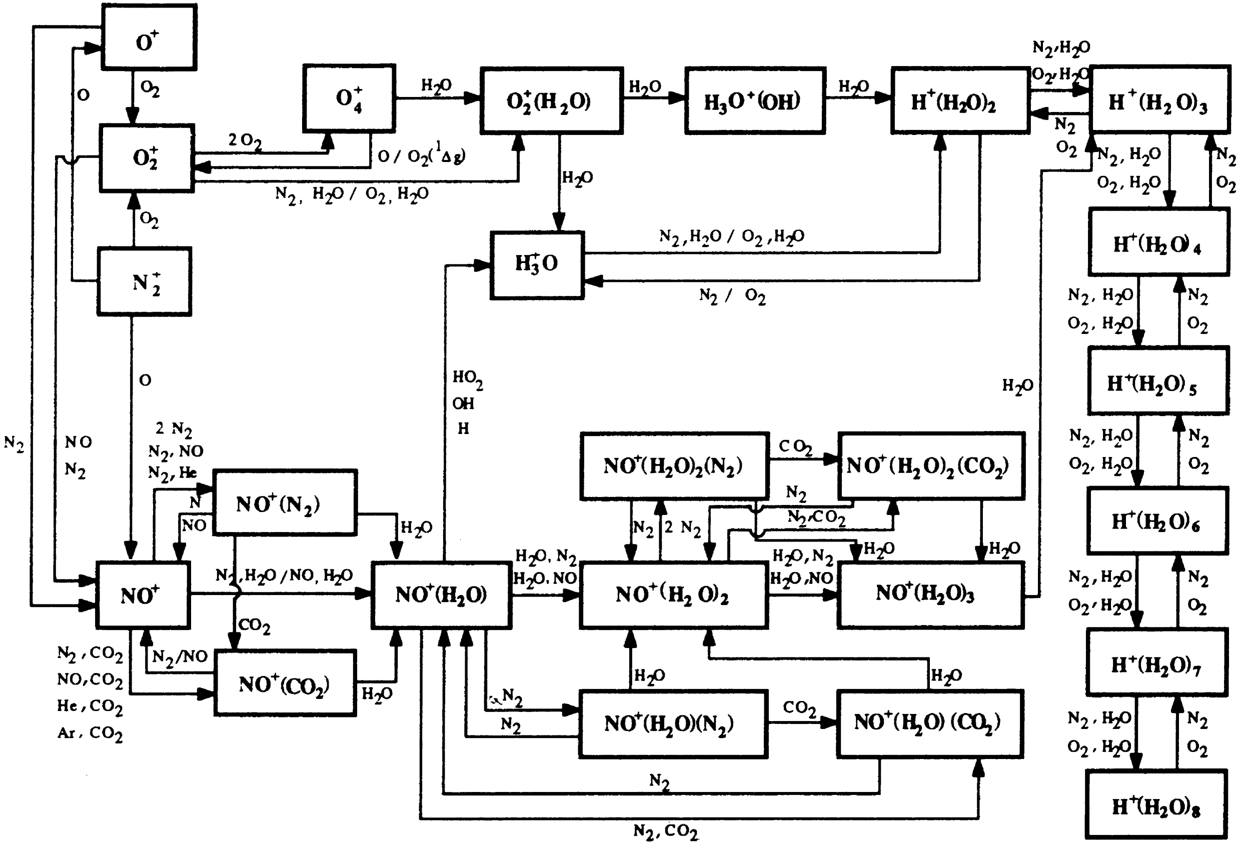

Figure 1.13. A scheme of positive-ion chemistry for the D region. (E. Turunen, private communication.) This model, developed at Sodankylä Geophysical Observatory, Finland, includes 24 positive and 11 negative ions, 35 in all. Later versions include as many as 55 ions.

Below about 70 km by day or 80 km by night much of the negative charge is in the form of negative ions. Their creation begins with the attachment of an electron to an oxygen molecule, forming O_2^-. This is a three-body reaction involving any other molecule, M, whose function is to remove excess kinetic energy from the reactants:

$$e + O_2 + M \rightarrow O_2^- + M. \tag{1.61}$$

This is followed by further reactions forming other and more complex negative ions such as CO_3^-, NO_2^-, and NO_3^- (the most abundant negative ion in the D region) and clusters such as $O_2^-.O_2$, $O_2^-.CO_2$, and $O_2^-.H_2O$. Because the electron affinity of O_2 is small (0.45 eV), the electron may be removed by a photon of visible or near infra-red light:

$$O_2^- + h\nu \rightarrow O_2 + e. \tag{1.62}$$

It may also be detatched through chemical reactions, such as with atomic oxygen (forming ozone), and with excited molecular oxygen. The effect of negative ions on the balance between electron production and loss was included in Equations (1.37)–(1.39). Variations of electron density in the D region can be due to changes in the negative-ion/electron ratio, λ, as well as to changes in production rate.

The complexity and uncertainty of D-region photochemistry is one reason why, when one is relating electron-production rates to electron densities, it is usual to work with an "effective recombination coefficient" (Equation (1.38)), which may be either theoretically or experimentally determined.

Diurnal behavior

Although the mid-latitude D region is complex chemically, observationally its behavior may be deceptively simple. The region is under strong solar control and it vanishes at night. VLF ($f < 30$ kHz) radio waves are, to a first approximation, reflected as at a sharp boundary in the D region because the refractive index changes markedly within one wavelength (Section 3.4.6). For VLF waves incident on the ionosphere at steep incidence, the reflection height, h, appears to vary as

$$h = h_0 + H \ln(\sec \chi), \tag{1.63}$$

where χ is the solar zenith angle. h_0 is about 72 km, and H is about 5 km, which happens to be the scale height of the neutral gas in the mesosphere. This form of height variation is just what is predicted for a level of constant electron density in the underside of a Chapman layer, and it is consistent with the ionization of NO by solar Lyman-α radiation.

At oblique incidence, when the transmitter and the receiver are more than about 300 km apart, the height variation follows a quite different pattern. The

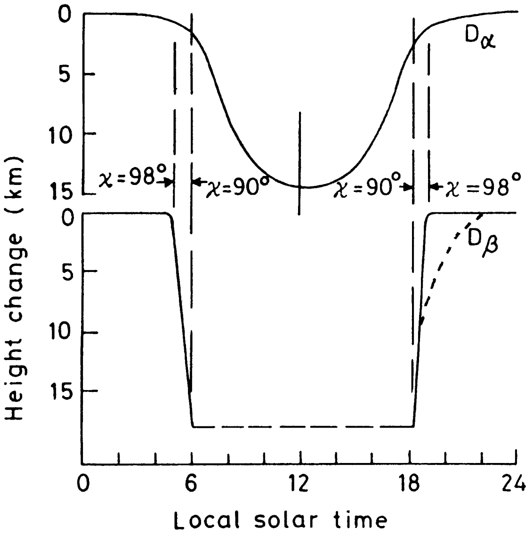

Figure 1.14. Two kinds of diurnal behavior of the D region inferred from VLF radio propagation at vertical and oblique incidence. The regions originally called D_α and D_β are now more usually called D and C. The evening recovery at oblique incidence tends to be more gradual than that in a simple D_β pattern and similar to the dashed curve. (After R. N. Bracewell and W. C. Bain, *J. Atmos. Terr. Phys.*, **2**, 216, Copyright 1952, with permission from Elsevier Science.)

reflection level now falls sharply before ground sunrise, remains almost constant during the day, and then recovers fairly rapidly following ground sunset. The reason has to do with the formation and detachment of negative ions at sunset and sunrise, coupled with electron production by cosmic-ray ionization – a source with no diurnal variation. This lower part of the D region is sometimes called a *C layer*. These patterns of height variation are illustrated in Figure 1.14.

Radio absorption

The D region is the principal seat of radio absorption, and absorption measurements (Section 4.2.4) are one way of monitoring the region. The absorption per unit height depends both on the electron density and on the frequency of collisions between electrons and neutral particles, and the measurement gives the integrated absorption up to the reflection level. Multi-frequency absorption measurements can provide some information about the height distribution.

Generally, the absorption varies with the solar zenith angle as $(\cos\chi)^n$ with n in the range 0.7–1.0. However, the seasonal variation contains an intriguing anomaly, which is that, during the winter months, the absorption exceeds by a factor of two or three the amount that would be expected by extrapolation from summer. Moreover, the absorption is much more variable from day to day in the winter. This phenomenon is the *winter anomaly of ionospheric radio absorption*.

1.4.4 The F2 region and the protonosphere

The peak of the F2 layer

Compared with the good behavior of the lower layers of the ionosphere, the F2 region, on first aquaintance, can be quite puzzling. In the first place it peaks at 200–400 km, whereas Figure 1.6 shows no band of radiation producing a maximum ionization rate at any height above 180 km. The answer is to be found in the height variation of the recombination rate, which forms the F2 region as an upward extension of F1 even though the production rate is now decreasing with height.

Taking O^+ as the major ion, the two-stage recombination process is

$$O^+ + N_2 \rightarrow NO^+ + N \qquad \text{with rate} \propto \beta[O^+]$$

followed by

$$NO^+ + e \rightarrow N + O \qquad \text{with rate} \propto \alpha[NO^+]N_e.$$

As discussed in Section 1.3.3, the second reaction controls the overall rate at low altitude and the first is the rate-determining step at high levels, the transition being where $\alpha N_e = \beta(h_t)$. The transition height, h_t, is generally between 160 and 200 km. The F1 ledge can appear if h_t is above the height of the maximum production rate, h_m: that is, if there is a production maximum within an α-type region.

To explain the F2 region we consider the upper part where the recombination is of β type, and where β depends on the concentration of N_2. On the other hand, the production rate depends on the concentration of O.Thus, at equilibrium,

$$\begin{aligned} N_e = q/\beta &\propto [O]/[N_2] \\ &\propto \exp\left(-\frac{h}{H(O)} + \frac{h}{H(N_2)}\right) \end{aligned}$$

where $H(O)$ and $H(N_2)$ are the scale heights for O and N_2. Since the masses of N_2 and O are in the ratio 1.75: 1, this rearranges to give

$$\begin{aligned} N_e &\propto \exp\left[-\frac{h}{H(O)}\left(1 - \frac{H(O)}{H(N_2)}\right)\right] \\ &= \exp\left(+\frac{0.75h}{H(O)}\right). \end{aligned} \tag{1.64}$$

This is a layer whose electron density increases with height because the loss rate falls off more quickly than does the production rate. It is often called a *Bradbury layer*.

The Bradbury layer explains why the electron density increases with height above the level of maximum ion production, but it does not explain why the F2 layer has a maximum. Here we have to invoke plasma transport. At the higher levels, *in situ* production and loss are less important than diffusion, which has become more important because of the decreasing air density. (That is, the right-hand side of Equation (1.15) is now dominated by the third term.) The F2 layer peaks where chemical recombination and diffusion are equally important. To decide the level at which this will occur, we regard the two loss processes – β-type recombination and transport – as being in competition, and compare their time constants for electron loss on the principle that the more rapid will be in effective control.

The characteristic time for recombination is

$$\tau_\beta = 1/\beta, \tag{1.65}$$

and it may be shown that the corresponding time for diffusion is approximately

$$\tau_D = H_1^2/D, \tag{1.66}$$

where H_1 is a typical scale height for the F2 region. Comparing these two equations places the F2 peak at the level where

$$\beta \sim D/H_1^2. \tag{1.67}$$

The electron density at the peak is given by

$$N_m \sim q_m/\beta_m. \tag{1.68}$$

The protonosphere

At some level in the topside the ionosphere dominated by O^+ gives way to the protonosphere dominated by H^+. It so happens that the ionization potentials for these two ions are almost the same (Table 1.1), and therefore the reaction

$$H + O^+ \rightleftharpoons H^+ + O \tag{1.69}$$

goes rapidly in either direction, and, around the transition level, the equilibrium is given by

$$[H^+][O] = (9/8)[H][O^+]. \tag{1.70}$$

(The factor 9/8 arises for statistical reasons, and there is also a temperature dependence proportional to $(T_n/T_i)^{1/2}$.) Through this reaction ionization can move readily between the ionosphere (as O^+) and the protonosphere (H^+). This is a very important aspect of the behavior of the topside ionosphere.

The transition effectively defines *the base of the protonosphere*. Below that level the H^+ distribution is determined by (1.71), and is related to the distribution of O^+ by

$$\begin{aligned}[H^+] &\propto [H][O^+]/[O]\\ &= \exp[-\mathrm{h}/H(\mathrm{H})]\exp[-h/H(\mathrm{O}^+)]/\exp[-h/H(\mathrm{O})]\\ &= \exp[+7h/H(\mathrm{H})].\end{aligned} \tag{1.71}$$

There is a strong upward gradient in the H^+ concentration below the transition level. Above the transition the concentration of O^+ decreases rapidly, and in this region the protonosphere, when it is in equilibrium, takes an exponential profile with the appropriate scale height (Equation (1.47)). As for the F2 peak, the transition level between ionosphere and protonosphere can be estimated by comparing time constants. If the rate constant of the reaction

$$H^+ + O \rightarrow H + O^+$$

is k, then the lifetime of a proton is $(k[O])^{-1}$. Taking the time constant for diffusion in the protonosphere as H_2^2/D, the boundary occurs where

$$k[O] \sim D/H_2^2. \tag{1.72}$$

This occurs at 700 km or higher, which is always well above the peak of the F2 layer.

1.4.5 Anomalies of the F2 region

The phenomena

The F2 region has the greatest concentration of electrons of any layer, and therefore it is the region of greatest interest in radio propagation. Unfortunately, it is also the region which is the most variable, the most anomalous, and the most difficult to predict. From the point of view of the Chapman theory the F2 region's behavior is anomalous in several ways, and these are sometimes called the *classical anomalies* of the F2 layer. Briefly, they are as follows.

(a). The diurnal variation may be asymmetrical about noon. There may be a rapid change at sunrise but little or no change in the evening until well after sunset or even until just before the next sunrise (Figure 1.15). The daily peak may occur either before or after local noon in the summer,

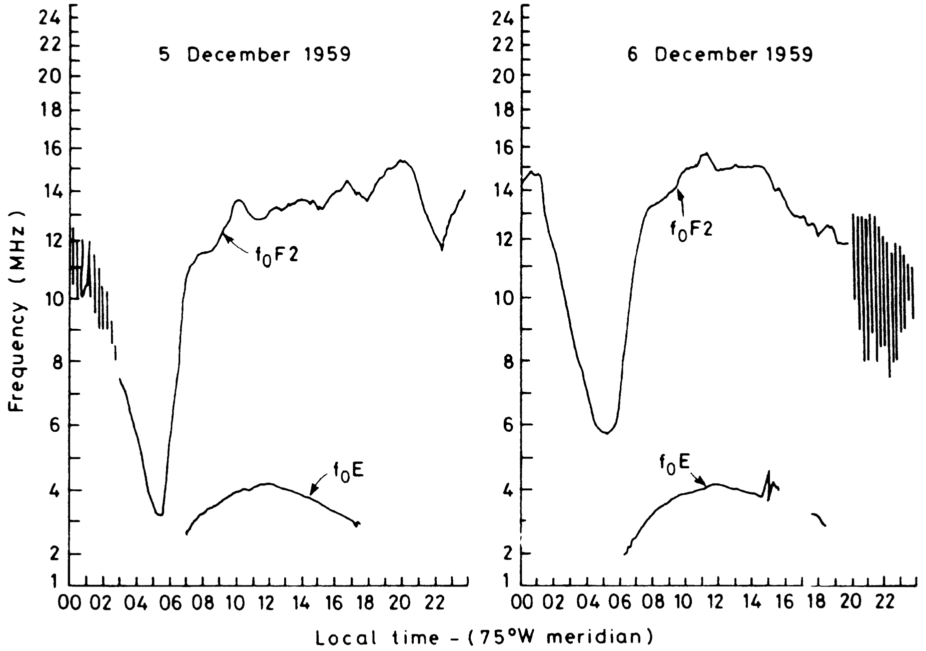

Figure 1.15. The diurnal behavior of f_0F2 on successive days in December 1959 at a low-latitude station, Talara, Peru. Note, by contrast, the regularity of the E layer. (T. E. VanZandt and R. W. Knecht, in *Space Physics* (eds. Le Galley and Rosen), Wiley, 1964.)

though it is likely to be near noon in the winter (Figure 1.16). On some days a secondary minimum appears near noon between the morning and evening maxima (Figure 1.16(a)).

(b). The daily pattern of variation often does not repeat from day to day. (If it did, the next day could at least be predicted from the previous one.) Figure 1.15 illustrates this point.

(c). There are several anomalous features in the seasonal variation. The main one is that noon values of the F-layer critical frequency (see Equation (3.67)) are usually greater in winter than they are in summer, whereas the Chapman theory leads us to expect the opposite. This is the *seasonal anomaly*, which is clear in Figure 1.16. The summer electron content (the summation of electron density in a column through the ionosphere) is greater than the winter value at some stations, but at others it is smaller or about the same. The electron content is abnormally large at the equinoxes, giving the *semi-annual anomaly*. Some stations also show this anomaly in the F-region critical frequency (Figure 1.17).

(d). The mid-latitude F2 region does not vanish at night, but remains through to the next sunrise at a substantial level.

Although not all anomalies have yet been fully explained, it now appears that there are four main causes for this seemingly anomalous behaviour:

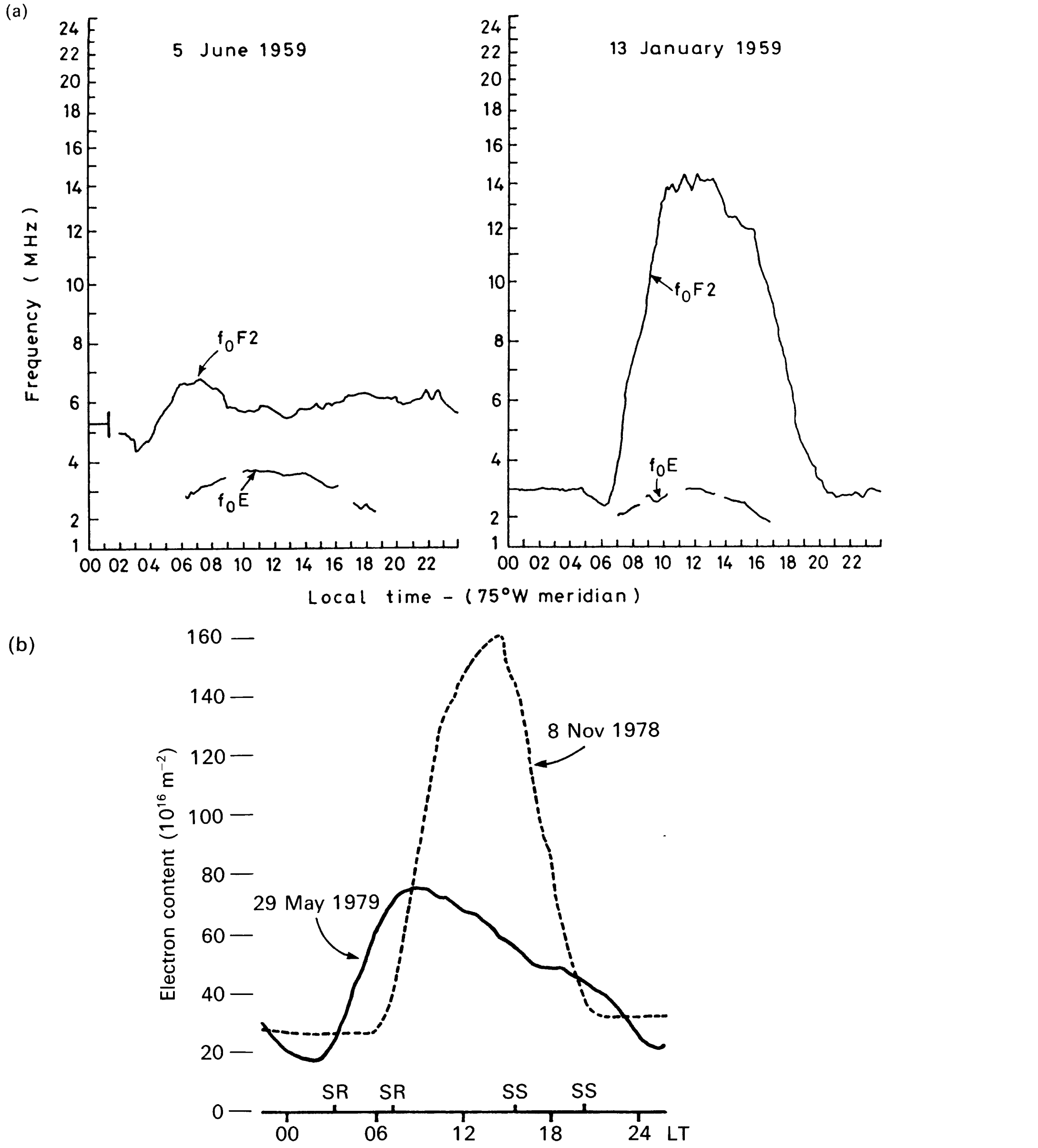

Figure 1.16. (a) The diurnal behavior of f_0F2 in summer and winter at a high-latitude station in the northern hemisphere, Adak, Alaska. The F region is anomalous whereas the E layer behaves as expected according to the Chapman theory. (T. E. VanZandt and R. W. Knecht, in *Space Physics* (eds. Le Galley and Rosen), Wiley, 1964.) (b) Summer and winter electron contents measured at Fairbanks, Alaska. (R. D. Hunsucker and J. K. Hargreaves, private communication.)

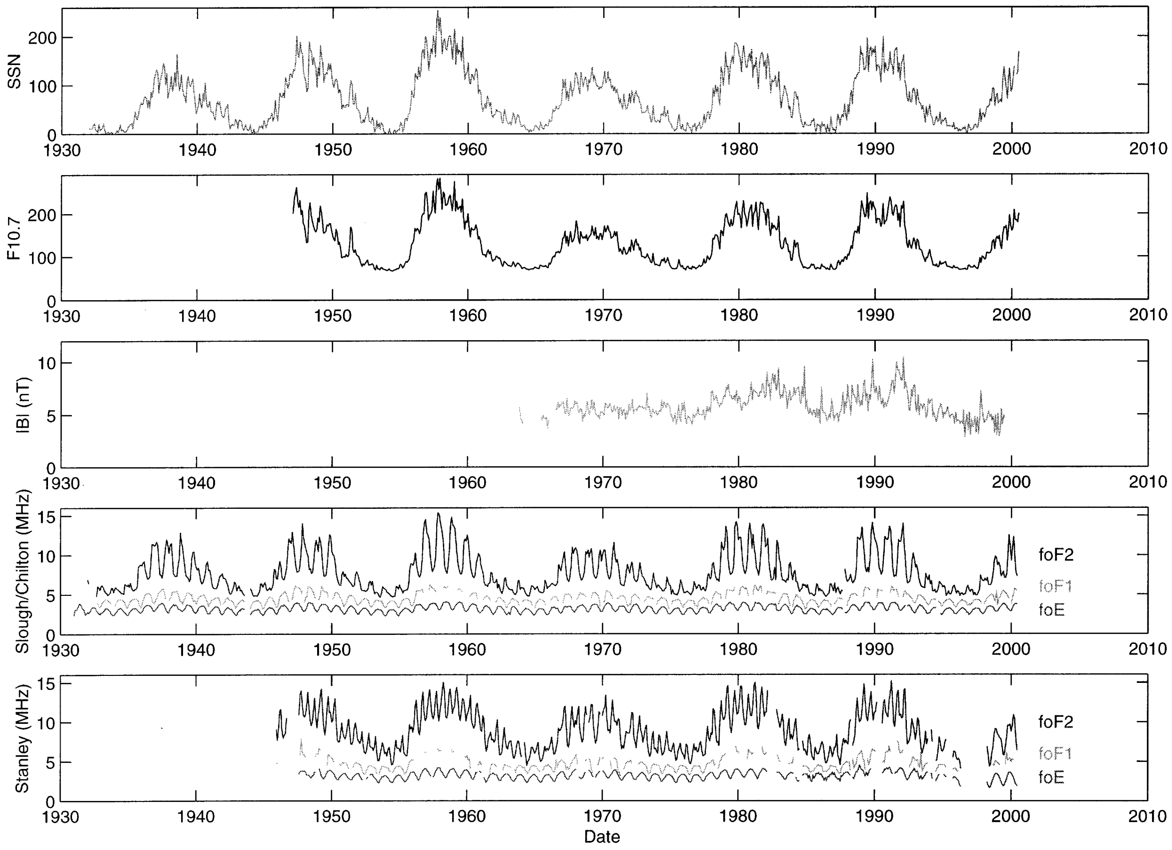

Figure 1.17. Variations of critical frequencies over several sunspot cycles. The three top panels show the sunspot number, the 10.7-cm solar radio flux, and the magnitude of the interplanetary magnetic field. (Diagram provided by M. Wild, Rutherford Appleton Laboratory, Chilton, UK.) Note also the seasonal modulations at Slough and Port Stanley. The E and F1 regions peak in the summer whereas F2 peaks in the winter. The semi-annual anomaly is prominent at Port Stanley.

(a). reaction rates are sensitive to temperature;

(b). the chemical composition varies;

(c). there are winds in the neutral air that lift or depress the layer by the mechanism indicated in Section 1.3.4; and

(d). the ionosphere is influenced by the protonosphere and by conditions in the conjugate hemisphere.

Reaction rates

Reaction rates are generally temperature sensitive. The rate for the reaction

$$O^+ + N_2 \rightarrow NO^+ + N,$$

the first step in an important two-stage loss process (Equations (1.57) and (1.56)), varies strongly with the temperature of neutral N_2 and increases by a factor of 16 between 1000 and 4000 K. This property obviously contributes both to the persistence of the night F region and to the seasonal anomaly.

Composition

Since the electron-production rate depends on the concentration of atomic oxygen, O, whereas the loss rate is controlled by the molecular species N_2 and O_2, increases in the ratios $[O]/[O_2]$ and $[O]/[N_2]$ will increase the equilibrium electron density. Satellite measurements have shown that such variations do occur. The ratio $[O]/[N_2]$ at 250–300 km is measured as about 6 in winter and about 2 in summer, a seasonal change amounting to a factor of three. The change of composition is attributed to the pattern of global circulation in the thermosphere. This is plainly a factor in the seasonal anomaly.

Winds

Mathematical modeling has demonstrated how the meridional component of the thermospheric neutral wind, acting to depress the ionosphere when the wind is flowing equatorward and elevating it when it is flowing poleward (Section 1.3.4), exerts a major influence both on electron densities and on electron content. At 300 km the neutral wind flows poleward by day and equatorward by night at speeds ranging between tens and hundreds of m s^{-1}. Thus its effect is usually to depress the ionosphere and thereby increase the rate of loss by day, but to lift the region and reduce its rate of decay at night. It is estimated (taking $H=60$ km for the neutral scale height, $D=2\times10^6$ m^2 s^{-1} for the diffusion coefficient, and $W=30$ m s^{-1} as a typical vertical drift due to the poleward wind), that by day the peak of the layer is lowered by about 50 km.

The variability of the F region from one day to the next (e.g. Figure 1.15) is one of its most remarkable and puzzling features. This might not be surprising in the polar regions because of the sporadic nature of solar and auroral activity, but these are not dominant influences at middle latitudes. Presumably the origin must be a source in the terrestrial atmosphere or in the solar wind. Variations of the neutral-air wind in the thermosphere are one possible cause.

The plasma temperature and the protonosphere

Variations in the temperature of the plasma affect its vertical distribution. The heating comes from the excess energy of absorbed photons above that needed for ionization. The excess energy is initially in the electrons and it is gradually shared with the positive ions, though transfer to the neutral species is less efficient. Consequently the plasma is hotter than the neutral air, and within the plasma the electrons are hotter than the ions ($T_e > T_i$). The electron temperature can be two or three times the ion temperature by day, though by night the electron and ion temperatures are more nearly equal. These changes in temperature strongly affect the distribution of F2-region plasma. When it is hotter, the plasma has a greater scale height (Equation (1.46)) and so spreads to greater altitudes, where it tends to persist for longer because the loss rate is smaller.

At the greater altitudes the positive ions are protons, and, as discussed in Section 1.4.4, the ionosphere and the protonosphere are strongly coupled through the charge-exchange reaction between protons and atomic oxygen ions (Equation (1.69)). As the F region builds up and is also heated during the hours after sunrise, plasma moves to higher altitudes where protons are created. These then flow up along the field lines to populate the protonosphere. In the evening the proton population flows back to lower levels, where it undergoes charge exchange to give oxygen ions and so helps to maintain the F region at night.

Via the protonosphere the magnetically conjugate ionosphere may also have an effect, since protonospheric plasma, coming mainly from the summer ionosphere, is equally available to replenish the winter ionosphere. Computations show that this is a significant source. Indeed, it is useful to treat the mid-latitude plasmasphere as consisting of winter and summer ionospheres linked by a common protonosphere; the ionospheres act as sources to the protonosphere, which in turn serves as a reservoir to the ionospheres. Overall, the winter ionosphere benefits from the conjugate region in the summer hemisphere. At sunrise, when electron densities are low, the ionosphere may be significantly heated by photoelectrons arriving from the conjugate hemisphere. The effect may show up as an increase of slab thickness (the ratio of electron content to maximum electron density) just before local sunrise.

It appears likely that the various classical anomalies of the F2 region arise from combinations of the factors outlined above, though the details might not be clear in any particular case.

1.4.6 The effects of the sunspot cycle

The varying activity of the Sun over a period of about 11 years, measured in terms of the number of sunspots visible on the disk, the rate at which flares occur, or the intensity of the 10-cm radio flux, also affects the ionosphere because of variations in the intensity of the ionizing radiations in the X-ray and EUV bands. The temperature of the upper atmosphere also varies with solar activity, approximately by a factor of two between sunspot minimum and maximum. Consequently, the gas density at a given height varies by a large factor.

The maxima of the E, F1, and F2 layers all depend on the number of sunspots, *R*. This influence can be seen in Figure 1.17. (The critical frequencies plotted there, f_OE, f_OF1, and f_OF2, are proportional to the square root of the maximum electron density, and are defined as the highest radio frequencies reflected from the layer at vertical incidence – see Section 3.4.2.) We have seen that the E and the F1 layers both behave as α-Chapman layers. In such a layer (Equation (1.30)) the critical frequency varies with the solar zenith angle χ as $(\cos\chi)^{1/4}$. Taking the number of sunspots into account as well gives two empirical relations:

$$f_OE = 3.3[(1+0.008R)\cos\,\chi]^{1/4}\ \text{MHz} \tag{1.73}$$

$$f_OF1 = 4.25[1+0.015R)\cos\,\chi]^{1/4}\ \text{MHz.} \tag{1.74}$$

Note that the F1 layer is nearly twice as sensitive as the E layer to variations in the sunspot number.

From the status of the E and F1 as α-Chapman layers it follows that the ratios $(f_OE)^4/\cos\chi$ and $(f_OF1)^4/\cos\chi$ are proportional to the ionization rates (q) in the E and F1 layers, respectively. These ratios are called *character figures*. Taking $R=10$ for a typical solar minimum and $R=150$ for a maximum, we see from Equation (1.73) that the E-region production rate varies by a factor of two over a typical sunspot cycle.

The F2 layer does not behave like a Chapman layer but it nevertheless varies with the sunspot number. The dependence may be seen by plotting the noon values of f_OF2, and if these are smoothed over 12 months to remove the seasonal anomalies, a dependence such as

$$f_OF2 \propto (1+0.02R)^{1/2}\ \text{MHz} \tag{1.75}$$

can be recognized.

One measure of the strength of the D region is the radio absorption measured, for example, by the pulsed sounding technique (Section 4.2.4). Other parameters being constant, it is observed that the absorption increases by about 1% for each unit of sunspot number:

$$A(\text{dB}) \propto (1+0.01R). \tag{1.76}$$

At mid-latitude the absorption is expected to vary over a sunspot cycle by about a factor of two.

1.4.7 The F-region ionospheric storm

From time to time the ionosphere suffers major perturbations called *storms*. They last from a few hours to a few days and tend to occur during times of geophysical disturbance resulting from increases in solar activity communicated via the solar wind. There are, on the face of it, connections with magnetic storms (Section 2.5), though some different mechanisms must be involved. Three phases may be identified.

(a). In the *initial* or *positive phase*, which lasts for a few hours, the electron density and the electron content are greater than normal.

(b). Then follows the *main* or *negative phase* when these quantities are reduced below normal values.

(c). Finally, the ionosphere gradually returns to normal over a period of one to several days in the *recovery phase*.

The magnitude of the effect varies with latitude, being greatest at middle and high latitude, where the maximum electron density may be depressed by 30% in a strong storm. At latitudes below about 30° the effect is not likely to exceed a few percent. The beginning can be sudden or gradual, the term *sudden commencement* being used (as for magnetic storms) to describe the former. At middle latitudes ionosondes show the apparent height of the maximium, h'(F2), to be increased, though real-height analysis attributes this mainly to greater group retardation (Section 3.4.2) below the peak rather than to a genuine lifting of the region. The slab thickness (the ratio of the electron content to N_{max}) does increase, however, confirming that the F region broadens during the negative phase. Figure 1.18 compares electron content, electron density, and slab thickness in a typical mid-latitude storm.

The progress of the storm since its time of commencement is the *storm-time variation*, but the time of day is also a significant parameter. Statistical studies, as well as case histories of major storms, show that the magnitude and even the sign of the effect depend on the time of day. The negative phase tends to be weaker in the afternoon and evening, stronger in the night and morning. The positive phase is often missing altogether at stations that were in the night sector at commencement. It has been suggested (Hargreaves and Bagenal, 1977) that the positive phase co-rotates with the Earth on the first day of the storm and does not reappear on the second day.

Seasonal and hemispheric effects are also marked. The negative phase is relatively stronger, and the positive phase relatively weaker, in the summer hemisphere. This holds both for the northern and for the southern hemisphere, though the interhemispheric difference is such that N_{m}(F2) is actually increased during the main phase of storms occurring in the southern hemisphere during winter. The interhemispheric difference arises from the larger separation between the geographic and the geomagnetic poles in the south.

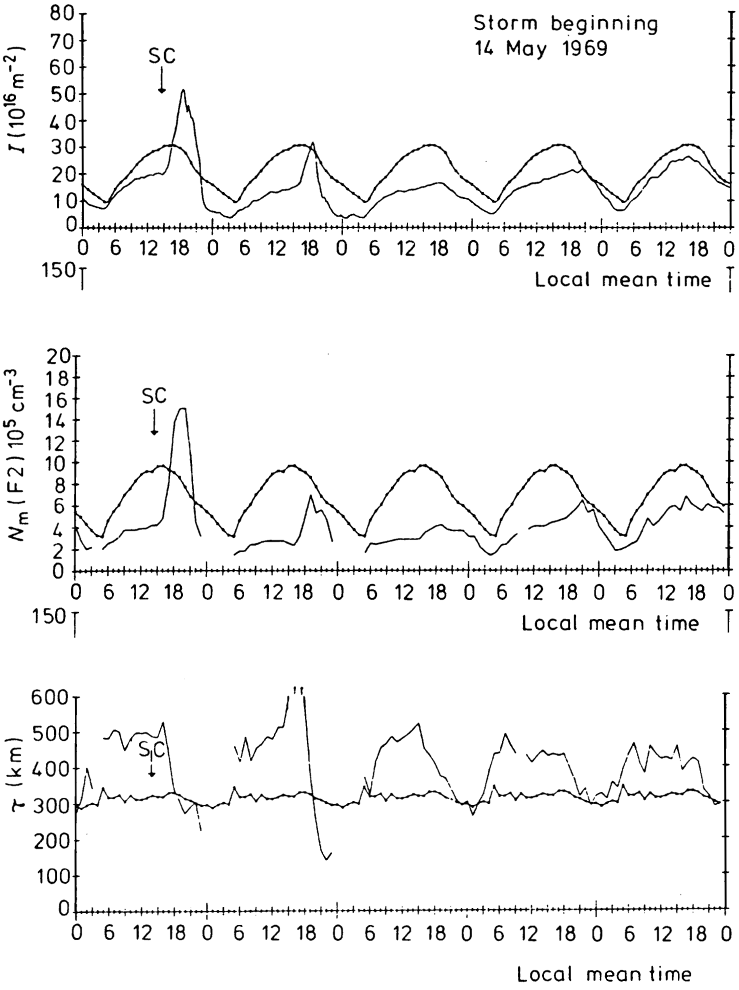

Figure 1.18. The electron content, electron density, and slab thickness at a mid-latitude station during an F-region storm. SC marks the time of sudden commencement. The 7-day mean is shown to indicate normal behavior. (M. Mendillo and J. A. Klobuchar, *Report AFGRL-TR-74-0065*, US Air Force, 1974.)

The most likely cause of the main phase is abnormal heating at high latitude, which also alters the pattern of circulation of the thermospheric wind. The heating reduces the ratio $[\mathrm{O}]/[\mathrm{N}_2]$ at given height in the F region, and the molecularly enriched air is then convected down to the middle latitudes by the changed air circulation. As was pointed out in Section 1.4.5, the effect of a greater proportion of molecular species in the F region is to reduce the equilibrium electron density. This mechanism has been verified by computer modeling (Rishbeth, 1991), though some problems remain to be solved. There appears to be no generally agreed cause of the initial phase, though various mechanisms have been suggested.

1.5 The electrical conductivity of the ionosphere

1.5.1 Introduction

The presence of free electrons and ions allows the ionospheric layers to carry electric currents. The conductivities of the ionosphere lie in the range 10^{-5}–10^{2} Ω^{-1} m^{-1}, a broad middle range between insulators (such as the tropospere, $\sigma \sim 10^{-14}$ Ω^{-1} m^{-1}) and good conductors like metals ($\sigma = 6 \times 10^{7}$ Ω^{-1} m^{-1} for copper), being akin to that of the ground (10^{-7}–1 Ω^{-1} m^{-1}) or a semiconductor (10^{-1}–10^{2} Ω^{-1} m^{-1}).

Radio propagation is generally considered in terms of the electron density of an ionospheric layer rather than its conductivity, and we shall not need to deal with conductivities very much, at least for propagation in the MF, HF, or VHF bands. However, the electric currents of the ionosphere and magnetosphere are a major factor in the behavior of the ionosphere and in the way it is affected by geophysical disturbances. These are particularly important at the high latitudes. The solar–geophysical environment, of which the ionosphere is a part, cannot be understood without including the several current systems that may exist within it. Hence, we give in this section the basis of ionospheric conductivity.

1.5.2 Conductivity in the absence of a magnetic field

If no magnetic field is present, the formula for the conductivity of an ionized gas is a simple one:

$$\sigma_0 = Ne^2/(m\nu), \tag{1.77}$$

where N is the number density of particles each with charge e and mass m, and ν is the collision frequency for collisions of a charged particle with neutral species (which are assumed to be in the majority). The formula is easily proved, remembering that the mobility of a charged particle (its velocity in a unit electric field) is $e/(m\nu)$, and the total charge per unit volume is Ne.

If more than one species of charge is present, for example electrons and positive ions, the total conductivity is the sum of the conductivities for each species separately.

1.5.3 The effect of a magnetic field

Unfortunately the Earth's magnetic field permeates the ionosphere, and this complicates the conductivity enormously. A charged particle moving through a magnetic field experiences a force (the Lorentz force) that acts at right angles both to

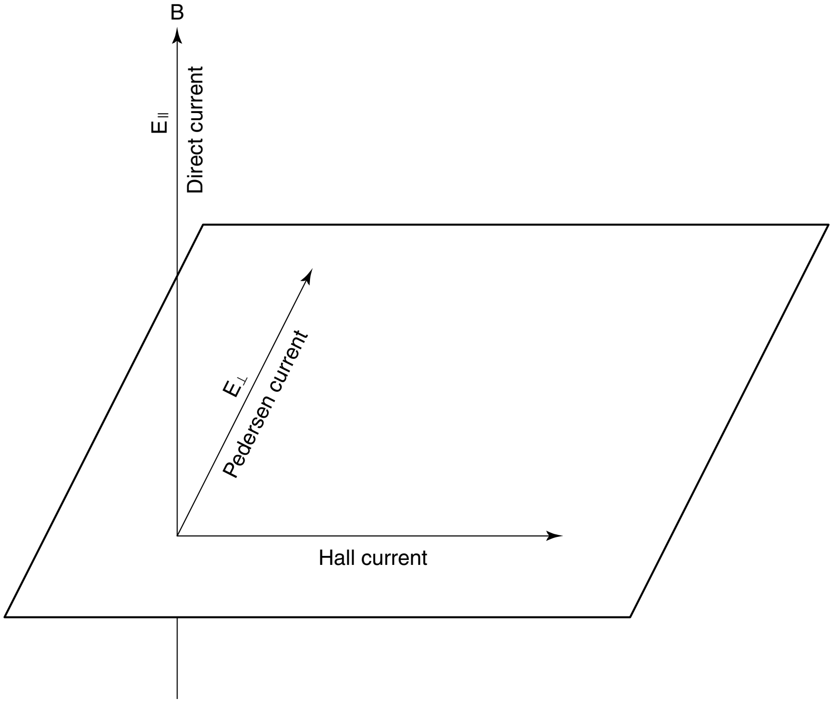

Figure 1.19. Currents due to the electric-field components parallel (E$_{\parallel}$) and perpendicular (E$_{\perp}$) to the magnetic field (**B**). The currents shown are those due to positive charges. The direct and Pedersen currents due to negative charges are the same as those shown, but the Hall current is opposite. The Hall current in the ionosphere is mainly due to electrons.

the direction of the magnetic field and to the direction of motion of the particle. If the particle is moving directly along the magnetic field, the Lorentz force is zero; the magnetic field has no effect and Equation (1.77) applies.

However, if the motion has a component at right angles to the magnetic field, the corresponding conductivity has two parts:

$$\sigma_1 = \left(\frac{N_e}{m_e \nu_e} \frac{\nu_e^2}{(\nu_e^2 + \omega_e^2)} + \frac{N_i}{m_i \nu_i} \frac{\nu_i^2}{(\nu_i^2 + \omega_i^2)} \right) e^2 \tag{1.78}$$

$$\sigma_2 = \left(\frac{N_e}{m_e \nu_e} \frac{\nu_e \omega_e}{(\nu_e^2 + \omega_e^2)} - \frac{N_i}{m_i \nu_i} \frac{\nu_i \omega_i}{(\nu_i^2 + \omega_i^2)} \right) e^2. \tag{1.79}$$

The subscript e here refers to electrons and i refers to positive ions. ω is the relevent gyrofrequency (eB/m, where B is the magnetic flux density). σ_1 is the *Pedersen conductivity*, which gives the current in the same direction as the applied electric field, whereas σ_2 is the *Hall conductivity* giving the current at right angles to it – it being understood that the electric field and the currents are all in the plane normal to the magnetic field. Figure 1.19 may clarify the geometry.

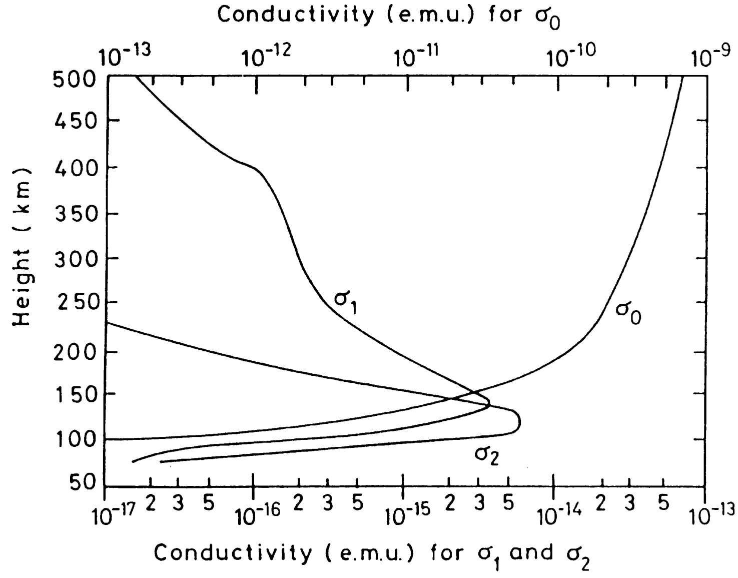

Figure 1.20. Conductivity profiles calculated for middle latitude at noon. (S.-I. Akasofu and S. Chapman (after K. Maeda and H. Matsumoto), *Solar–Terrestrial Physics*, Oxford University Press, 1972. By permission of Oxford University Press.). Multiply the conductivity values by 10^{11} to convert them to the SI unit Ω^{-1} m^{-1}.

1.5.4 The height variation of conductivity

It is clear from Equations (1.78) and (1.79) that the conductivity due to a single species depends on the ratio ω/ν. Indeed, the ratio between the Hall and Pedersen conductivities for a given electron (or ion) density is just ω/ν, and is therefore strongly height-dependent. Note, also, that, in Equation (1.79) the electron and ion terms are of opposite sign, so the total Hall conductivity depends on the difference between the electron and ion conductivities, not on their sum. Figure 1.20 illustrates the height variations of the direct, Pederson, and Hall conductivities in a typical mid-latitude ionosphere. The Hall conductivity peaks in the E region, the Pedersen conductivity peaks somewhat higher, and the direct conductivity continues to increase with height. The Hall conductivity is very small in the F region because the electron and ion components almost cancel out there.

Figure 1.21 indicates the motions of ions and electrons, and the resulting electric current, at various key altitudes. In the upper panel the driving force is a wind in the neutral air, which induces ion motion through collisions. The effect of an electric field is shown in the lower panel.

1.5.5 Currents

For there to be an electric current there must also be a driving force (either a wind or an electric field) and a path of conductivity providing a complete circuit. Where

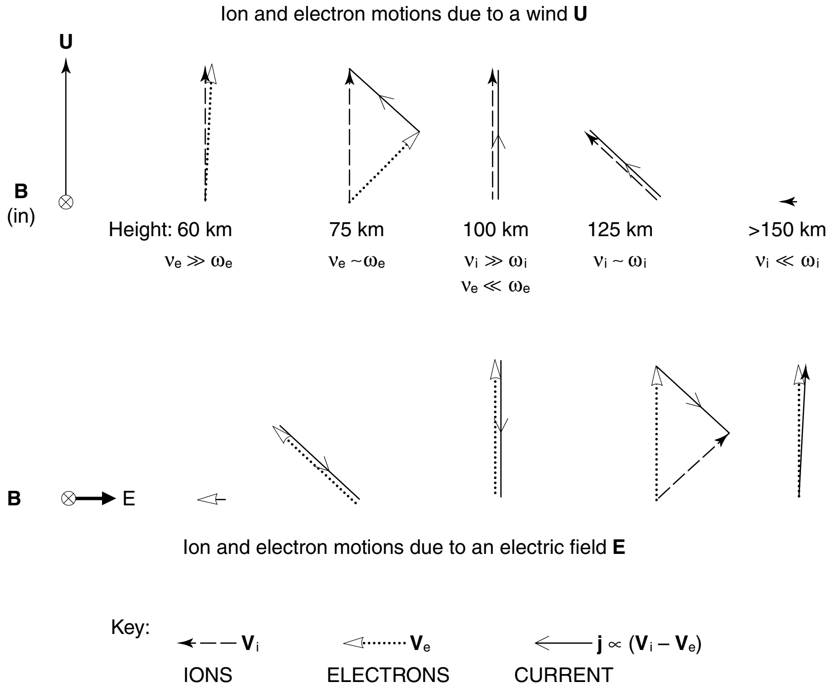

Figure 1.21. Ion and electron motions due to a neutral-air wind (top) and an electric field (bottom) at selected key altitudes. The current is proportional to the vector difference between the ion and electron velocities. (After H. Rishbeth. *J. Inst. Electronic Radio Engineers* **58**, 207, 1988.)

the latter is not present, the flow of current is inhibited or modified by the electric potentials created at the boundaries.

The geomagnetic equator is one interesting case. Here the magnetic field runs horizontally and therefore the current which would otherwise flow normal to the field is inhibited in the vertical direction. Charges are created at the upper and lower boundaries, and the resulting electric field acts to increase the current in the horizontal plane. It can be shown that, in this special situation, the conductivity across the magnetic field and in the horizontal direction is given by

$$\sigma_3 = \sigma_1 + \sigma_2^2/\sigma_1, \tag{1.80}$$

called the *Cowling conductivity*. The value of the Cowling conductivity is comparable to that of the direct conductivity (Equation (1.77)), and therefore the current over the magnetic equator is abnormally large. This is the *equatorial electrojet*.

The large value of the direct conductivity suggests that current should be able to flow readily along the geomagnetic field direction. The existence of *field-aligned currents* was suggested by K. Birkeland in 1908, but the idea lay dormant for many years due to lack of evidence, and magnetic perturbations observed at the ground were interpreted in terms of currents flowing purely horizontally. It was not until field-aligned currents were detected by satellite-borne magnetometers in the early 1970s that the *Birkeland current* came into fashion and current systems became three-dimensional. Birkeland currents are particularly important in the auroral regions.

A fuller treatment of conductivity and the current systems of the solar–geophysical environment is given in several of the standard textbooks.

1.6 Acoustic-gravity waves and traveling ionospheric disturbances

1.6.1 Introduction

The familiar acoustic wave, in which the compression of the gas provides the force restoring a displaced particle towards its original position, is actually the high-frequency limit of a more general class, the *acoustic-gravity wave* (*AGW*). A parcel of air displaced vertically in a stratified atmosphere tends to be restored by buoyancy (due to gravity), and the AGW family results when both gravity and the compressional force are taken into account. We are here concerned mainly with atmospheric waves towards the low-frequency end of the AGW range, whose periods range from a few minutes to an hour or two. They have horizontal wavelengths from several hundred to about a thousand kilometers. Gravity waves in the atmosphere (which should not be confused with cosmological gravity waves, to which they have no connection whatsoever) are transverse waves, the displacement of the gas being normal to the direction of travel of the wave. Their properties, in fact, are complex and in many respects not at all obvious.

Several sources of AGWs are known: the motion of the ground during an earthquake, man-made explosions, weather systems, and ionospheric disturbances at high latitude. Table 1.2 shows a classification based on period and wavelength. Waves of small scale come mainly from the troposphere; the medium-scale waves may be tropospheric or ionospheric in origin; and the large-scale events generally have their source in the high-latitude ionosphere – hence their appearance in this opus! Some AGWs are, no doubt, a consequence of events in the solar–terrestrial system: for example, perturbations in the solar wind can produce magnetospheric

Table 1.2. *A classification scheme for AGWs*

Nomenclature	Horizontal trace velocity (m s^{-1})	Period (min)	Wavelength
Large-scale	~250–1000	>70	>1000 km
Medium-scale	90 to ~250	~15–70	Several hundred kilometers
Small-scale	>300	~2–5	—

effects, which couple to the high-latitude ionosphere as particle–precipitation events and electric-field disturbances, which in turn generate medium- and large-scale AGWs. We shall meet other examples of solar–geophysical chains of events later in the book.

The ionospheric manifestation of AGWs is the *traveling ionospheric disturbance* (*TID*), which is due to ion movement communicated from the motion of the neutral air through collisions. There are, however, some complications, the principle one being that, in the F region, the ion motion is constrained along the geomagnetic field.

The generation of atmospheric waves at high latitude is discussed in Sections 6.5.6 and 6.5.7.

1.6.2 Theory

Wave motions in the upper atmosphere have been known for over 100 years and TIDs have been noted in ionospheric observations since the 1940s, but not until the 1950s did adequate explanations start to emerge, the key theory being developed by C. O. Hines (Hines, 1960). The underlying concepts of wave propagation are given in Sections 3.2.1 and 3.2.3 in the context of electromagnetic waves. We outline here some of the basic theory governing the properties and behavior of AGWs.

In a planar, horizontally stratified, isothermal, single-species, windless, nonrotating atmosphere, the AGW obeys a dispersion relation

$$\omega^4 - \omega^2 s^2(k_x^2 + k_z^2) + (\gamma - 1)g^2 k_x^2 - \omega^2\gamma^2 g^2/(4s^2) = 0. \tag{1.81}$$

where

- ω is the angular frequency of the wave,
- k_x is the horizontal wave number $(=2\pi/\lambda_x)$,
- λ_x being the wavelength in the horizontal,
- k_{z} similarly is the vertical wave number,
- γ is the ratio of specific heats (constant pressure/constant volume),
- s is the speed of sound, and
- g is the acceleration due to gravity.

This equation states the relation between the frequency and the wavelength (or wave number) in the vertical and the horizontal directions for an AGW. k_y does not appear in the equation because there is no asymmetry between the x and y directions.

Two significant frequencies are the *acoustic cut-off frequency*,

$$\omega_{\mathrm{a}} = \gamma g/(2s) \tag{1.82}$$

and the *buoyancy* or *Brunt–Väisala frequency*,

$$\omega_{\mathrm{b}}=(\gamma-1)^{1/2}g/s. \tag{1.83}$$

ω_{a} is the resonance frequency in the acoustic mode of a column of air extending through the whole atmosphere, whereas ω_{b} is the natural frequency of oscillation of a displaced parcel of air when buoyancy is the restoring force.

Substituting these frequencies into Equation (1.81) and rearranging gives

$$k_z^2=\left(1-\frac{\omega_{\mathrm{a}}^2}{\omega^2}\right)\frac{\omega^2}{s^2}-k_x^2\left(1-\frac{\omega_{\mathrm{b}}^2}{\omega^2}\right). \tag{1.84}$$

Putting $\omega^2 \gg \omega_{\mathrm{b}}^2$ gives

$$k_x^2+k_z^2=\left(1-\frac{\omega_{\mathrm{a}}^2}{\omega^2}\right)\frac{\omega^2}{s^2}\ [=(2\pi/\lambda)^2], \tag{1.85}$$

where λ is the wavelength. There is now no distinction between the x and z coordinates, and this is the acoustic regime. If we go a stage further by putting $\omega^2 \gg \omega_{\mathrm{a}}^2$, we get

$$s=\omega/(k_x^2+k_z^2)=\omega\lambda/(2\pi). \tag{1.86}$$

This represents a sound wave. In the acoustic regime the phase speed is independent of direction. Putting, now, $\omega^2 \ll s^2k_x^2$, which removes the effect of compressibility, gives

$$k_z^2=k_x^2\left(\frac{\omega_{\mathrm{b}}^2}{\omega^2}-1\right), \tag{1.87}$$

which represents a pure gravity wave.

Since k_x and k_z must both be positive in a propagating wave, the frequency ω must be either larger than ω_{a} or smaller than ω_{b}. These, the acoustic and the gravity regimes, are illustrated in Figure 1.22, which plots the regimes of AGW in terms of the frequency (ω) and the horizontal wave number (k_x). Between the acoustic and the gravity regimes the waves are evanescent and do not propagate.

The angle of propagation with respect to the horizontal is

$$\theta=\tan^{-1}(k_z/k_x). \tag{1.88}$$

If ω^2 is small compared with ω_{b}^2, the ratio k_z/k_x is large and then the wave propagates almost vertically. This is for the propagation of phase. The energy, on the other hand, travels at the group velocity, given (Equation (3.21)) by

$$u=(\mathrm{d}k/\mathrm{d}\omega)^{-1},$$

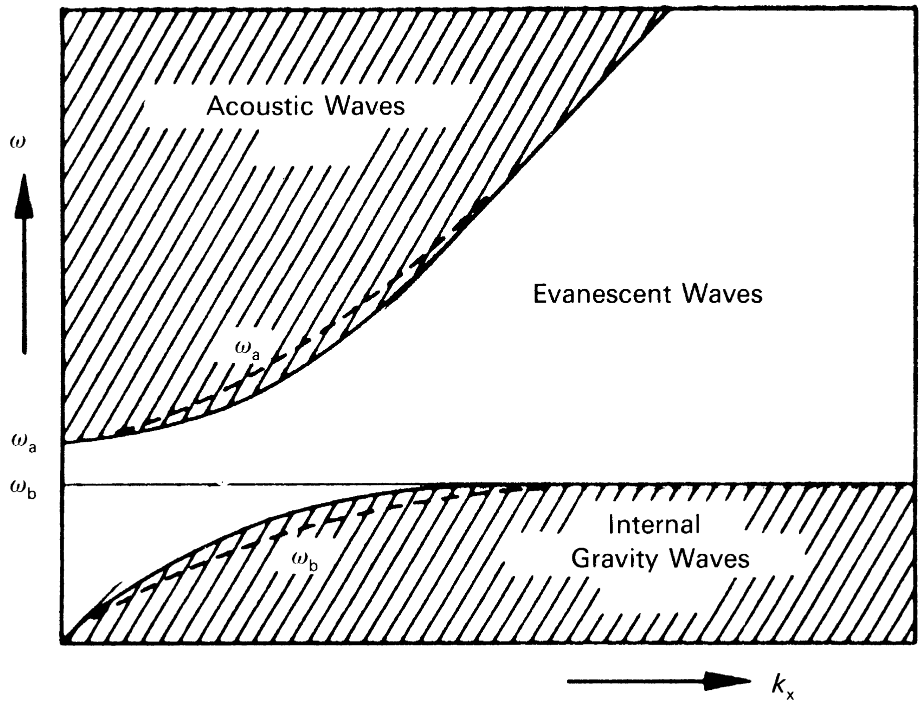

Figure 1.22. The acoustic, evanescent, and gravity regimes of acoustic-gravity waves. The dashed lines show the effects of neglecting gravity and compressibility, repectively. At ionospheric levels, waves with periods longer than 10–15 min are likely to be gravity waves, and any with periods of only a few minutes are probably acoustic. (After J. C. Gille, in *Winds and Waves in Stratosphere, Mesosphere and Ionosphere* (ed. Rawer). North-Holland, 1968. Elsevier Science Publishers.)

and, in a gravity wave, the energy flow is at right angles to the direction of phase propagation. Figure 1.23 illustrates the relations amongst particle displacement, phase propagation, and group propagation in a gravity wave. Note that, if the source is below, the energy flows upward (as it must) but the phase propagation is downward. Furthermore, the amplitude of the air displacement increases with altitude so that the energy flux may be constant (provided that there are no losses).

Figure 1.24 shows how the horizontal component of the group velocity varies with the wave period (normalized by the Brunt frequency as ω_b/ω) at fixed values of the ascent angle of the energy (i.e. the angle between the group velocity and the horizontal). The energy flow approaches horizontal when the wavelength is very large. A distinction between the sections of the curves labeled "buoyancy" and "gravity" needs to be made when one is considering AGW propagation over large distances (Francis, 1975).

For an AGW, the refractive index (μ) is defined as the ratio between the speed of sound and the phase velocity of the wave. (Compare with Section 3.2.3.) Then

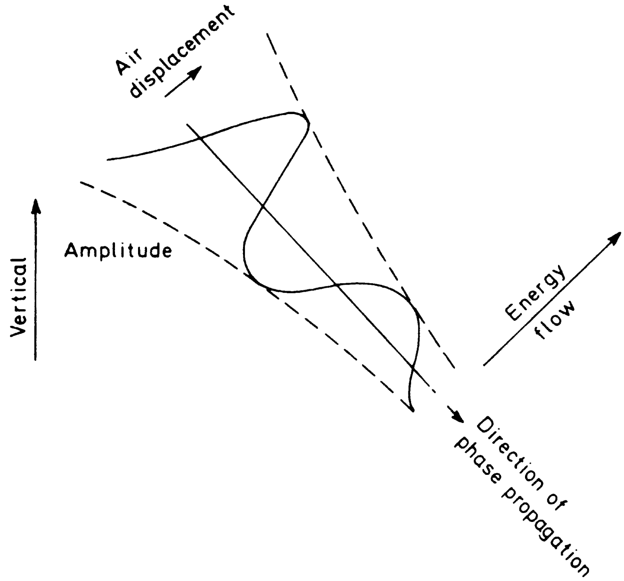

Figure 1.23. A simple gravity wave, showing the essential relations amongst phase propagation, air displacement, and energy flow.

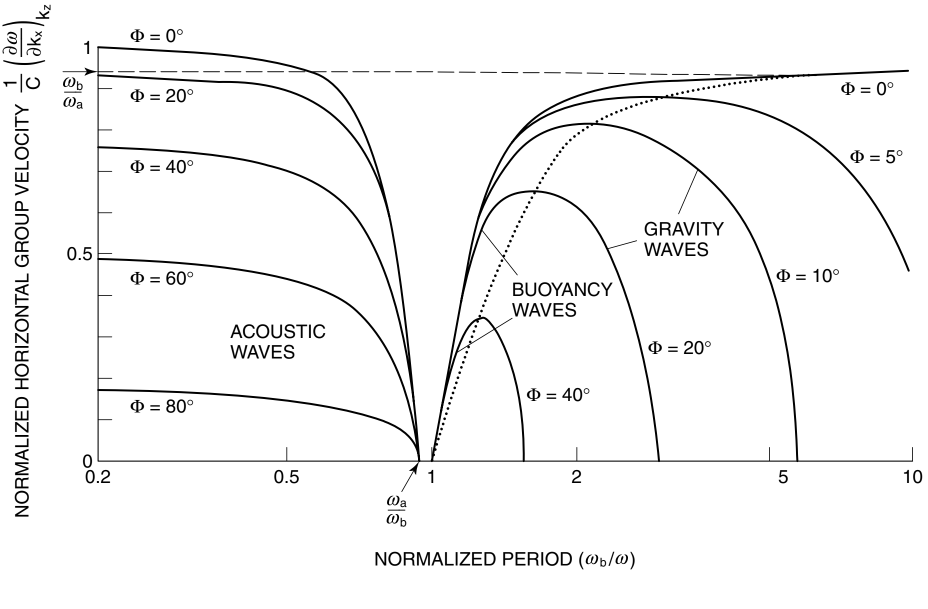

Figure 1.24. Contours of constant ϕ, the ascent angle of the group velocity from the horizontal, against the wave period and the horizontal component of the group velocity. The details of the diagram depend on the values assumed for the acoustic (ω_a) and Brunt (ω_b) frequencies. (Reprinted from S. H. Francis, *J. Atmos. Terr. Phys.* **37**, 1011, Copyright 1975, with permission from Elsevier Science.)

$$\mu^2 = \left(1 - \frac{\omega_a^2}{\omega^2}\right) \bigg/ \left[\left(1 - \frac{\omega_b^2}{\omega^2}\right)\cos^2\theta\right]. \qquad 1.89$$

If $\omega \gg \omega_a, \omega_b$, $\mu \rightarrow 1$, and if $\omega \ll \omega_a, \omega_b$,

$$\mu \rightarrow \frac{\omega_a}{\omega_b}\frac{1}{\cos\theta}.$$

In general the particle motion is elliptical in an AGW, combining the longitudinal displacement of an acoustic wave with the transverse displacement of a gravity wave. There are alternate compressions and rarefactions at successive zero-displacement points in Figure 1.23. At extremely low frequency, the air motion and the group velocity would be horizontal, the phase propagation vertical, and the compression and rarefaction zero.

Complexities neglected by the simple theory, but which affect AGWs in real life, are energy loss through the viscosity of the air, non-linear effects if the amplitude becomes too large at the higher altitudes, reflection and ducting due to the change of atmospheric properties with altitude, the curvature of the Earth's surface, and winds.

1.6.3 Traveling ionospheric disturbances

The mechanism by which AGWs produce ionospheric disturbances (TID) is collisional coupling between neutral and ionized particles. This force acts in the direction of motion of the neutral air, but, in the ionospheric F region, the effect is strongly modified by the geomagnetic field which permits ion motion along the field only. Thus, while there are several radio techniques able to measure properties of a TID, to interpret these data as properties of the AGW causing it may be less than straightforward. This, however, hardly matters if propagation effects are the principal concern.

Figure 1.25 is an elegant example of a TID observation by ionosonde (Section 4.2.1). It shows the period of the wave and its wavelength, the latter derived using the velocity estimated from spaced observations. The downward phase propagation is clearly seen.

1.6.4 The literature

The literature of published research on the topics of AGW and TID is very large. Surveys of the earlier work have been published by Yeh and Liu (1974) and Francis (1975). Studies performed from the mid-1970s up to 1981 have been reviewed by Hunsucker (1982), and those between 1982 and 1995 by Hocke and Schlegel (1996). Work since then is addressed by Kirchengast (1996), Bristow and Greenwald (1997), Balthazor and Moffett (1997, 1999), Huang *et al.* (1998) and Hall *et al.* (1999).

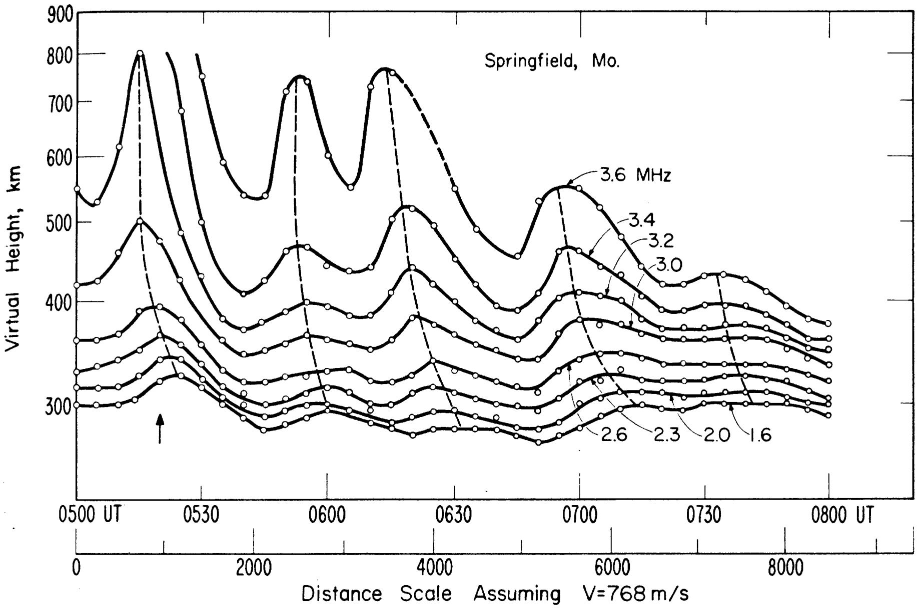

Figure 1.25. A train of gravity waves observed by ionosonde over Missouri, USA, in December 1966, identified from the virtual heights of echoes at frequencies between 1.6 and 3.6 MHz. (T. M. Georges, *Ionospheric Effects of Atmospheric Waves*. Institutes for Environmental Research, report IER 57-ITSA 54, 1967, Boulder, Colorado, USA.)

1.7 References and bibliography

1.2 The Vertical structure of the atmosphere

Hargreaves, J. K. (1992) *The Solar–Terrestrial Environment*. Cambridge University Press, Cambridge.

Richmond, A. D. (1983) Thermospheric dynamics and electrodynamics. *Solar–Terrestrial Physics* (eds. R. L. Carovillano and J. M. Forbes), p. 523. Reidel, Dordrecht.

1.3 Physical aeronomy

VanZandt, T. E. and Knecht, R. W. (1964) The structure and physics of the upper atmosphere. *Space Physics* (eds. D. P. LeGalley and A. Rosen), p. 166. Wiley, New York.

1.4 The main ionospheric layers

Bracewell, R. N. and Bain, W. C. (1952) An explanation of radio propagation at 16 kc/sec in terms of two layers below E layer. *J. Atmos. Terr. Phys.* **2**, 216.

Friedrich, M. and Torkar, K. M. (1992) An empirical model of the nonauroral D region. *Radio Sci.* **27**, 945.

Hargreaves, J. K. and Bagenal, F. (1977) The behavior of the electron content during ionospheric storms: a new method of presentation and comments on the positive phase. *J. Geophys. Res.* **82**, 731.

Piggott, W. R. and Rawer, K. (1972) *URSI Handbook of Ionogram Interpretation and Reduction*, Chapter 4. Report UAG-23A, World Data Center A, NOAA, Boulder, Colorado.

Rishbeth, H. (1991) F-region storms and thermospheric dynamics. *J. Geomag. Geoelectr.* **43** suppl., 513.

VanZandt, T. E. and Knecht, R. W. (1964) The structure and physics of the upper atmosphere. *Space Physics* (eds. D. P. LeGalley and A. Rosen), p. 166. Wiley, New York.

Whitehead, J. D. (1970) Production and prediction of sporadic E. *Rev. Geophys. Space Phys.* **8**, 65.

1.5 The electrical conductivity of the ionosphere

Akasofu, S.-I. and Chapman, S. (after K. Maeda and H. Matsumoto) (1972) *Solar–Terrestrial Physics*, Oxford University Press, Oxford.

Kelley, M. (1989) *The Earth's Ionosphere*. Academic Press, New York.

Rishbeth, H. (1988) Basic physics of the ionosphere – a tutorial review. *J. Inst. Electronic Radio Engineers* **58**, 207.

1.6 Acoustic-gravity waves and traveling ionospheric disturbances

Balthazor, R. L. and Moffett, R. J. (1997) A study of atmospheric gravity waves and travelling ionospheric disturbances at equatorial latitudes. *Ann. Geophysicae* **15**, 1048.

Balthazor, R. L. and Moffett, R. J. (1999) Morphology of large-scale traveling atmospheric disturbances in the polar thermosphere. *J. Geophys. Res.* **104**, 15.

Bristow, W. A. and Greenwald, R. A. (1997) On the spectrum of thermospheric gravity waves observed by the Super Dual Auroral Radar Network. *J. Geophys. Res.* **102**, 11585.

Francis, S. H. (1975) Global propagation of atmospheric gravity waves: a review. *J. Atmos. Terr. Phys.* **37**, 1011.

Gille, J. C. (1968) The general nature of acoustic-gravity waves. Winds and Turbulence in Stratosphere, Mesosphere and Ionosphere (ed. Rawer). Elsevier Science Publishers, Amsterdam.

Hall, G. E., MacDougall, J. W., Cecile, J.-F., Moorcroft, D. R. and St.-Maurice, J. P. (1999) Finding gravity wave positions using the Super Dual Auroral Radar network. *J. Geophys. Res.* **104**, 67.

Hines, C.O. (1960) Internal atmospheric gravity waves at ionospheric heights. *Can. J. Phys.* **38**, 1441.

Hocke, K. and Schlegel, K. (1996) A review of atmospheric gravity waves and travelling ionospheric disturbances: 1982–1995. *Ann. Geophysicae* **14**, 917.

Huang, C.-S., Andre, D. A. and Sofko, G. (1998) High-latitude ionospheric perturbations and gravity waves: 1. Observational results. *J. Geophys. Res.* **103**, 2131.

Hunsucker, R. D. (1982) Atmospheric gravity waves and traveling ionospheric disturbances. *Encyclopedia of Earth System Science*, p. 217. Academic Press, New York.

Kirchengast, G. (1996) Elucidation of the physics of the gravity wave–TID relationship with the aid of theoretical simulations. *J. Geophys. Res.* **101**, 13 353.

Yeh, K-C. and Liu, C-H. (1974) Acoustic-gravity waves in the upper atmosphere. *Rev. Geophys. Space Phys.* **12**, 193.

General reading on the topics of Chapter 1

Books

Akasofu, S.-I. and Chapman, S. (1972) *Solar–Terrestrial Physics*. Oxford University Press, Oxford.

Banks, P. M. and Kockarts, G. (1973) *Aeronomy*. Academic Press, New York.

Bauer, S. J. (1973) *Physics of Planetary Atmospheres*. Springer-Verlag, Berlin.

Brasseur, G. and Solomon, S. (1984) *Aeronomy of the Middle Atmosphere*. Reidel, Dordrecht.

Carovillano, R. L. and Forbes, J. M. (eds.) (1983) *Solar–Terrestrial Physics*. Reidel, Dordrecht.

Dieminger, W., Hartmann, G. K. and Leitinger, R. (eds.) (1996) *The Upper Atmosphere – Data Analysis and Interpretion*. Springer-Verlag, Berlin.

Hess, W. N. and Mead, G. D. (eds.) (1968) *Introduction to Space Science*. Gordon and Breach, New York.

Jursa, A. S. (ed.) (1985) *Handbook of Geophysics and the Space Environment*. Air Force Geophysics Laboratory, US Air Force, National Technical Information Service, Springfield, Virginia.

Kato, S. (1980) *Dynamics of the Upper Atmosphere*. Center for Academic Publication Japan, Tokyo.

Matsushita, S. and Campbell, W. H. (eds.) (1967) *Physics of Geomagnetic Phenomena*. Academic Press, New York.

Ratcliffe, J. A. (ed.) (1960) *Physics of the Upper Atmosphere*. Academic Press, New York.

Rawer, K. (1956) *The Ionosphere*. Frederick Ungar Publishing Co., New York.

Rees, H. M. (1989) *Physics and Chemistry of the Upper Atmosphere*. Cambridge University Press, Cambridge.

Rishbeth, H. and Garriott, O. K. (1969) *Introduction to Ionospheric Physics.* Academic Press, New York.

VanZandt, T. E. and Knecht, R. W. (1964) The structure and physics of the upper atmosphere. In *Space Physics* (eds. D. P. Le Galley and A. Rosen). Wiley, New York.

Whitten, R. C. and Poppoff, I. G. (1965) *Physics of the Lower Ionosphere*. Prentice-Hall, Englewood Cliffs, New Jersey.

Whitten, R. C. and Poppoff, I. G. (1971) *Fundamentals of Aeromony*. Wiley, New York.

Conference reports

McCormac, B. M. (ed.) (1973) *Physics and Chemistry of Upper Atmosphere*. Reidel, Dordrecht.

McCormac, B. M. (ed.) (1975) *Atmospheres of Earth and Planets*. Reidel, Dordrecht.