1

The ionosphere and magnetosphere

1.1. The earth's atmosphere

This book is mainly concerned with the effect on radio wave propagation of the ionised regions of the earth's atmosphere. Near the ground the air is almost unionised and its electrical conductivity is negligibly small, because the ionising radiations have all been absorbed at greater heights. When any part of it is in equilibrium, its state is controlled by the earth's gravitational field so that it is a horizontally stratified system. Although it is never in complete equilibrium, gravity has a powerful controlling effect up to about 1000 km from the ground.

The molecules of the neutral atmosphere have an electric polarisability which means that the refractive index for radio waves is very slightly greater than unity, about 1.000 26, near the ground. The water vapour also affects the refractive index. Thus the neutral air can very slightly refract radio waves. This can lead to important effects in radio propagation. For example in stable meteorological conditions a duct can form near the surface of the sea, acting as a wave guide in which high frequency radio waves can propagate to great distances (see Booker and Walkinshaw, 1946; Brekhovskikh, 1960; Budden, 1961b; Wait, 1962). Spatial irregularities of the refractive index of the air can cause scintillation of radio signals, and also scattering which can be used to achieve radio propagation beyond the horizon (Booker and Gordon, 1950). These effects are beyond the scope of this book. It is here assumed that the non-conducting atmosphere below the ionosphere is the same as free space, with refractive index unity.

Incoming ultra-violet radiation and X-rays from the sun ionise the atmosphere. At great heights of the order 1000 km or more the medium is almost fully ionised, but so tenuous that the ion and electron concentrations are small. At lower levels there is more gas to be ionised so that the radiation is more strongly absorbed and the ion concentrations are greater. At still lower levels the radiation has been used up so that the degree of ionisation is again small. The region where the ion and electron

concentrations are greatest is the ionosphere. The height where maximum ionisation occurs depends on the absorption coefficient of the air for the ionising constituent of the sun's radiation, and on the mechanism by which ions and electrons are removed. Different ionising constituents give maxima at different heights. Thus there are several different ionospheric layers of ionisation. The principal ones are the E-layer at about 100 km and the F-layer at about 300 km. Their formation and structure is discussed in more detail in §§ 1.5–1.8. In the F-layer the ion and electron concentrations are only about 10^{-5} of the concentration of neutral particles, and in the E-layer the figure is about 10^{-11}. Thus gravity, acting mainly on the neutral particles, exerts a dominating control on the hydrostatic configuration of the medium. This means that the ionised layers tend to be horizontally stratified.

Above about 1000 km the atmosphere is a fully ionised ion plasma and therefore a relatively good electrical conductor. In these conditions the earth's magnetic field has a strong effect on the structure of the medium. For a perfectly conducting fluid in a magnetic field of induction $\boldsymbol{B}$ there is an effective stress like a pressure $\frac{1}{2}\boldsymbol{B}^2/\mu_0$ in directions perpendicular to $\boldsymbol{B}$ and a tension of the same magnitude in the direction of $\boldsymbol{B}$; see, for example, Ferraro and Plumpton (1966, § 1.7). The earth's magnetic field $\boldsymbol{B}$ is approximately that of a magnetic dipole near the earth's centre. Thus, at a fixed magnetic latitude, its magnitude B is roughly proportional to the inverse cube of the distance from the centre. If the plasma were perfectly conducting, the magnetic force per unit volume would be the spatial gradient of $\frac{1}{2}B^2/\mu_0$. The gravitational force per unit volume is ρg where ρ is the density and g is the gravitational acceleration. It can be shown that these two forces are about equal at a height of 300–400 km. Here the conductivity, though large, is not large enough to allow the magnetic forces to have their full effect. Gravity still exerts a strong control up to heights of the order of 1000 km. Above this the atmosphere is not horizontally stratified but its structure is controlled by the magnetic field. This region is the magnetosphere and it extends to about $14\,R_e$ on the sunlit side of the earth. The unit R_e is one earth radius, about 6400 km. Distances in R_e are usually measured from the centre of the earth.

When the magnetic field within a conducting medium tends to change, the changing flux density induces electric fields which cause currents to flow, and these currents themselves give a magnetic field. Lenz's law shows that this new field must have a direction which opposes the change that originated it. Thus the effect of the conductivity is to oppose any change of the magnetic field. The rate at which a magnetic field can change depends on the conductivity and on the size of the conductor. (See problem 1.1.) Within a perfect conductor the magnetic induction $\boldsymbol{B}$ cannot change at all. This effect can be observed with metals which become superconducting when cooled to very low temperatures, and it led to the phrase 'freezing in of the magnetic field'. The magnetosphere is like a blob of conducting fluid surrounding the earth, with the earth's magnetic field frozen into it.

The shape of the magnetosphere is greatly influenced by the gas that streams out from the sun, known as the solar wind. This gas is itself a fully ionised plasma and therefore a good conductor. Thus the earth's magnetic field cannot quickly penetrate into it, because it is 'frozen out'. But the solar wind distorts the magnetosphere so that it is drawn out into a long tail, the 'magneto-tail' on the side remote from the sun. This extends out to a distance greater than the radius of the moon's orbit.

Further details of the structure of the ionosphere and magnetosphere are given in §§ 1.5–1.8 and 1.9 respectively.

1.2. Plane and spherical radio waves

Radio waves travelling from a transmitter to a receiver near the earth's surface may take one of several possible paths. A wave may travel over the earth's surface, and it is then known as the ground wave. The earth has an imperfectly conducting curved surface. The problem for a vertical electric dipole transmitter near a plane surface was first propounded and studied by Sommerfeld (1909), and it has intrigued many mathematical writers ever since. One of the best treatments of it is that of Baños (1966). This topic is within the field of surface waves and guided waves and is beyond the scope of this book.

Another wave may travel up to the ionosphere, be reflected there, and return to the receiver. It is with a single reflection of this kind that the present book is largely concerned. The wave comes from a source of small dimensions so that the wave front is approximately spherical, but by the time it reaches the ionosphere the radius of curvature is so large that the wave can often be treated as plane. The approximation involved is examined in §§ 11.4, 11.5 and the error is shown to be negligible except in a few special cases. Similarly the ionospheric layers are curved because of the earth's curvature, but in most problems this curvature can be neglected.

Radio receivers are often used in rockets and satellites within the ionosphere or magnetosphere. It is thus important to be able to study the spatial distribution of the electromagnetic field within these media. At the higher frequencies this can be done by ray tracing (chs. 10, 14). At lower frequencies it is necessary to use the computed solutions of the governing differential equations (chs. 18, 19).

1.3. Waves in ion plasmas

The ionosphere and magnetosphere are composed of ionised gases known as ion plasmas. The negative ions are almost entirely electrons. The plasma must, on the average, be electrically neutral. If any space charge is present it can be shown (§ 3.8) that it must oscillate with the plasma frequency. This is the phenomenon of plasma oscillations. These are damped out with a time constant of a few milliseconds or less. But a radio wave in the plasma modifies the ion and electron concentrations, so that there is, in general, a space charge oscillating with the frequency of the wave.

An ion is about 2000 to 60 000 times more massive than an electron. Thus at the frequencies used for radio communication, that is, greater than a few kilohertz, the range of movement of an ion caused by the electric field of a radio wave is smaller than that of an electron, by about the same factor. This means that the ions can for most purposes be ignored. The effective particles in the plasma are the electrons. The positive ions simply provide a background to keep the plasma electrically neutral on the average. In the lowest parts of the ionosphere there can also be massive negative ions formed by the attachment of electrons to air molecules, and these play some part in the very complicated chemistry of the ionisation process (Rishbeth and Garriott, 1969). But electrons are easily detached from these ions so that their concentration cannot be large. Radio propagation effects attributable to massive negative ions have never been observed as far as the author knows.

The earth's magnetic field has a dominating effect on radio wave propagation in the atmospheric plasmas, and they are therefore sometimes called magnetoplasmas. Because of this they are doubly refracting for radio waves, so that in many problems it is necessary to study a differential equation of the fourth order. The two refractive indices depend strongly on the direction of the wave normal (chs. 4, 5) and the medium is said to be anisotropic. For some purposes, however, it is convenient to neglect the earth's magnetic field, so that the differential equation is only of the second order. There is then only one refractive index, and it is independent of the direction of the wave normal. The medium is said to be isotropic. This is useful for establishing some physical principles.

The neglect of the massive ions is permissible only for frequencies which greatly exceed the ion gyro-frequencies. The gyro- or cyclotron frequency for an ion with charge e and mass m_i in a magnetic field B is $eB/2\pi m_i$. For protons in the upper ionosphere it is about 400 Hz. In the lower ionosphere there are few protons and the gyro-frequencies of the heavier ions are smaller still. Thus the ions need only be considered for frequencies low in the audio range. This is well below the range used for radio communication, but the naturally occurring radio signals such as whistlers and chorus extend down to this range. They are important in studies of the ionosphere and magnetosphere. Some account of the effect of massive ions at these frequencies in the upper ionosphere and magnetosphere is therefore included (§§ 3.11, 13.8, 13.9, 17.5).

The ionosphere is quite hot, 800 K in the E-region and 1000–2000 K in the F-region. Thus an electron has a random thermal velocity of the order of $10^5\ \mathrm{m\,s^{-1}}$, whereas a typical radio wave imparts to it an additional velocity of order only $10^3\ \mathrm{m\,s^{-1}}$. In spite of this it is found, rather surprisingly, that the thermal motions can be neglected. This is called the 'cold plasma' treatment. It can be shown to be justified by using a full kinetic treatment of the problem with the Boltzmann–Vlasov equations. This is done in the numerous books on plasma physics (see, for example:

Stix, 1962; Clemmow and Dougherty, 1969; Shkarofsky, Johnston and Bachynski, 1966).

When the temperature of the plasma is allowed for it is found that, besides the electromagnetic (radio) waves, other types of wave, known collectively as plasma waves, can propagate in the plasma. The main ones are the electron plasma wave also known as the Langmuir wave, and the ion-acoustic wave. These waves are, in general, much more heavily attenuated than radio waves and do not travel over great distances, so they cannot be used for communication. A transition from radio waves to plasma waves sometimes occurs (§ 19.5), but in most radio problems the plasma waves can be ignored. They are however of dominating importance in the theory of the technique of incoherent scatter (§ 1.7).

For most purposes in this book, therefore, the temperature of the electrons can be neglected. But it does have to be considered in the study of electron collisions, § 3.12, and in wave interaction, §§ 13.11–13.13.

Lorentz (1909) studied the theory of the effect of electron collisions on an electromagnetic wave in a plasma and concluded that they have the same effect as a retarding force proportional to the velocity imparted to an electron by the field of the wave. In most of this book the simple Lorentz form of the retarding force is used. It has been realised, however, that the average collision frequency ν of an electron may depend on its total velocity v, including the random thermal velocity which is often very large. If ν were independent of velocity, the Lorentz result would still be true but as a result of laboratory measurements it is now thought that ν is proportional to v^2. This has led to a more complicated treatment of collisions (Sen and Wyller, 1960) described in § 3.12. Collisions can sometimes be neglected, for example in the magnetosphere where ν is very small, or at high frequencies for which 'ray theory' methods are used. Collision damping is important in the lower part of the ionosphere (below about 300 km) and at low frequencies where, usually, the wavelength is so great that ray theory methods are inapplicable and a 'full wave' treatment must be used. Chs. 15–19 are devoted to this. It is often convenient to ignore collisions when describing and classifying the properties of the waves.

The refractive indices n of the two waves in a cold magnetoplasma are strongly frequency dependent and we say that the medium is highly dispersive for these waves. If collisions are ignored the squares of the refractive indices are real. If n^2 is negative the wave is said to be evanescent (§ 2.14) and the frequencies where n^2 is zero are called 'cut-off' frequencies. One n^2 can be infinite at certain frequencies and this is called 'resonance'.

1.4. Relation to other kinds of wave propagation

The theory of radio wave propagation in the terrestrial plasmas is closely related to other branches of physics that deal with wave propagation in media whose

properties vary from place to place. For example in wave mechanics a study is made of the propagation of electron waves in a potential field. The variation of potential is analogous to the variation of the square of the refractive index for radio waves. But the potential is real in nearly all problems of wave mechanics, so that there is nothing analogous to the damping forces which, for radio waves, lead to a complex value of the squared refractive index.

For electron waves in a crystal the wave velocity in general depends on the direction of the wave normal so that the 'medium' is anisotropic. There is an important analogy between the direction and magnitude of the particle velocity, and the ray direction and group velocity of a radio wave in a magnetoplasma. The waves of wave mechanics are highly dispersive and can show cut-off and resonance. Some of the material of chs. 5, 7, 8 and 15 is of great importance when applied to wave mechanics.

In oceanography a study is made of the propagation of sound waves in the ocean and its bed. The system is often horizontally stratified because the temperature, salinity and other properties depend on depth. For sound waves the ocean bed can behave like an inverted ionosphere, and many of the results of chs. 7–19 can be applied to it. But it can nearly always be assumed that sound waves are not dispersive. A plane sound wave does not show cut-off or resonance.

In a solid three kinds of elastic wave can propagate, two transverse and one longitudinal, so that the medium might be considered to be triply refracting. These waves are studied in seismology (Bullen and Bolt, 1985; Jeffreys, 1976), and in crystal acoustics (Musgrave, 1970). The interest has been mainly in propagation through homogeneous solids and in reflection and transmission at boundaries. Seismic waves in continuously varying stratified media have been treated by Kennett (1983). Matrix equations very similar to the set (7.80) that forms the basic equations for much of the theory in this book were used for elastic waves by Gilbert and Backus (1966), and for acoustic waves in the ocean by Abramovici (1968). Seismic waves in anisotropic solids have been discussed by Stoneley (1955, 1963) and Crampin (1970) who were mainly interested in surface (guided) waves. Plane seismic waves, like sound waves, are not dispersive and do not show cut-off or resonance.

Finally mention must be made of sound waves of very long period, hours or days, in the earth's atmosphere. The wavelength is of the order of thousands of kilometres so that the waves extend through a large part of the atmosphere. They are strongly influenced by gravity and are known as acoustic-gravity waves or planetary waves. Most of their energy is well below the ionosphere where the air density is large. They cause movements of the atmosphere, including movements of the ionospheric layers, that can be observed from their effect on reflected radio waves. Thus radio probing of the ionosphere is an important method of studying these waves. One type of wave

of this kind is known as a travelling ionospheric disturbance, abbreviated TID; see Munro (1953), Munro and Heisler (1956a, b), Murata (1974).

1.5. Height dependence of electron concentration: the Chapman layer

Before the reflecting properties of the ionosphere for radio waves can be calculated, it is necessary to know how the electron concentration $N(z)$ depends on height z above the ground. To study this, some assumption must be made about how the ionospheric layers are formed. A most important contribution to this problem was made by Chapman (1931a, b, 1939) who derived an expression for $N(z)$ now known as the Chapman law. The full theory of the formation of ionospheric layers has been refined and extended by Chapman and others, and is beyond the scope of this book (see Chapman, 1931a, b, 1939; Rishbeth and Garriott, 1969; Hargreaves, 1979). Only the simplest version of the Chapman theory is given here.

Assume that the air is constant in composition and at a constant temperature T. Then the air density ρ at height z is

$$\rho = \rho_0 \exp(-z/H) \tag{1.1}$$

where ρ_0 is the density at the ground, $H = KT/Mg$, K is Boltzmann's constant, M is the mean mass of a molecule and g is the gravitational acceleration. The curvature of the earth is neglected and g is assumed constant. H is called the 'scale height' of the atmosphere and is approximately 10 km at the ground. The sun's radiation enters the atmosphere at an angle χ from the zenith. Let the mass absorption coefficient of the air for the radiation be σ, and assume that the rate of production of electrons, q, is proportional to the rate of absorption of radiation per unit volume. Let the flux of energy in the incident radiation be I_0 outside the earth's atmosphere, and I at a height z. The energy flux decreases as the radiation is absorbed on passing down through the atmosphere, so that

$$\mathrm{d}I = I\sigma\rho \sec\chi \,\mathrm{d}z. \tag{1.2}$$

This is combined with (1.1) and integrated to give

$$I = I_0 \exp(-\sigma\rho_0 H \sec\chi \, \mathrm{e}^{-z/H}). \tag{1.3}$$

Now let $z_0 = H \ln(\sigma\rho_0 H)$. Then (1.3) gives

$$I = I_0 \exp[-\sec\chi \exp\{-(z - z_0)/H\}]. \tag{1.4}$$

The rate of absorption of energy per unit volume at height z is $\cos\chi\,(\mathrm{d}I/\mathrm{d}z)$ and since this is proportional to the rate of production of electrons q, we have from (1.4)

$$q = q_0 \exp[1 - (z - z_0)/H - \sec\chi \exp\{-(z - z_0)/H\}] \tag{1.5}$$

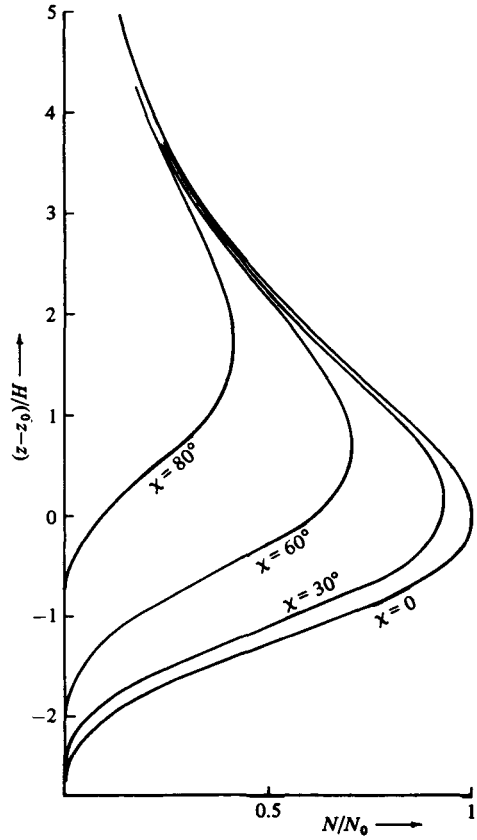
where $q_0 = I_0/\mathrm{e}H$ is a constant (here e is the exponential). The maximum value of q is $q_0 \cos\chi$ and it occurs where $(z - z_0)/H = \ln(\cos\chi)$, so that q_0 is the maximum rate of electron production when $\chi = 0$.

Next it is necessary to consider how electrons are removed. The process is very complicated (Rishbeth and Garriott, 1969) but in the lower part of the ionosphere below about 250 km, the effect is the same as if the electrons simply recombined with positive ions. Assume that the only ions present are electrons and positive ions. Then they both have concentration N, and the rate of removal of electrons is αN^2 per unit volume, where α is a constant called the recombination coefficient. The variation of N with time t is then given by

$$dN/dt = q - \alpha N^2. \tag{1.6}$$

It was at one time believed that in the F-region the electrons were removed by attachment to neutral particles so that $dN/dt = q - \beta N$ and the constant β was called the attachment coefficient. But it is now known that at these greater heights the processes are much more complicated and electron diffusion plays an important part (see Rishbeth and Garriott, 1969). Here we shall study (1.6) which applies fairly

Fig. 1.1. Dependence of electron concentration N on height z according to the simple Chapman theory for a flat earth, for various values of the sun's zenith angle χ.



well in the E-region and below. Now α is of the order $10^{-13}\mathrm{m}^3\mathrm{s}^{-1}$ and the time constant $(\alpha N)^{-1}$ is about 10 min. Thus the processes of formation and removal of electrons come into equilibrium in times of this order. As a first approximation, therefore, $\mathrm{d}N/\mathrm{d}t$ can be neglected compared with the other terms in (1.6) and

$$N = (q/\alpha)^{\frac{1}{2}}. \tag{1.7}$$

This, with (1.5), gives

$$N = N_0 \exp\left[\left[\tfrac{1}{2}\{1 - (z - z_0)/H - \sec\chi \exp\left[-(z - z_0)/H\right]\}\right]\right]. \tag{1.8}$$

This expression will be called the 'Chapman law'. It is based on a very much oversimplified picture of the actual mechanism of production and removal of electrons. For example it assumes that α is independent of height. But it does give a guide as to how the electron concentration might vary with height. In fig. 1.1 the expression (1.8) is plotted against height z for various values of χ. It is seen that N has a maximum value $N_\mathrm{m} = N_0(\cos\chi)^{\frac{1}{2}} = (I_0 \cos\chi/\mathrm{e}\alpha H)^{\frac{1}{2}}$ at the height $z = z_\mathrm{m} = z_0 + H\ln(\sec\chi) = H\ln(\sigma\rho_0 H\sec\chi)$. It falls off quite steeply below this, and less steeply above it. The maximum concentration N_m depends on H, I_0 and χ but not on the absorption coefficient σ. The height of the maximum z_m depends on H, σ and χ but not on the intensity I_0 of the radiation.

This property suggests that several different ionospheric layers occur because the absorption coefficient σ has different values for the various components of the ionising radiation from the sun. Thus the F-layer is produced largely by ultra-violet radiation in the wavelength range very roughly 17 to 80 nm. The E and D layers are produced largely by X-rays in the wavelength range very roughly 0.1 to 17 nm. (1 nm = 10 ångström units). The solar radiation contains some narrow spectral lines including Lyman-α and Lyman-β which come from atomic hydrogen, and are in the ultra violet range. It happens that Lyman-α is not strongly absorbed at F region heights but makes an important contribution to the ionisation in the D region.

Many radio observations have been used to test the Chapman law (1.8) and have shown good agreement when the ionosphere is not disturbed. For example the penetration frequency (§ 13.2) of a Chapman layer is proportional to $N_\mathrm{m}^{\frac{1}{2}}$, that is to $(\cos\chi)^{\frac{1}{2}}$, and this has been tested (Appleton, 1937, § 2). In the lowest part of the D-region, the dependence on χ of the height where a fixed small N occurs was studied by Budden, Ratcliffe and Wilkes (1939) and found to agree with (1.8).

An alternative form of (1.8) is got by taking

$$\zeta = (z - z_\mathrm{m})/H = (z - z_0)/H - \ln(\sec\chi) \tag{1.9}$$

so that ζH is height measured from the level of maximum N. Then

$$N = N_\mathrm{m} \exp\{\tfrac{1}{2}(1 - \zeta - \mathrm{e}^{-\zeta})\}, \tag{1.10}$$

which shows that the 'shape' of a Chapman layer is independent of the sun's zenith angle χ. The curvature of the curve of $N(z)$ at its maximum is $\frac{1}{2}N_\mathrm{m}/H^2$. This is the

same as the curvature at the apex of a parabolic layer (see § 12.4) whose 'half thickness' is $2H$.

The electrons' height distribution function $N(z)$ appears in the differential equation which has to be solved to find the reflecting properties of the ionosphere for radio waves. This is usually done by numerical computing (chs. 18, 19). But it is also useful to select small ranges of z and use approximate simple expressions for $N(z)$, which permit the differential equation to be reduced to a simple standard form whose solutions have well known properties. For example, it is often permissible to assume, for a small range of z, that $N(z)$ is a linear function. This case is of the greatest importance and is the subject of ch. 8 and §§15.2, 15.4. Near a maximum of $N(z)$ the linear law is not satisfactory but it is useful to approximate $N(z)$ by a parabola (§§ 15.9–15.11). In the lower part of a Chapman layer it is sometimes useful to treat $N(z)$ as an exponential function (§ 15.8).

1.6. Collision frequencies

Radio waves in the ionosphere and magnetosphere are subject to some attenuation because the motions of the electrons and ions are damped through collisions with other particles. In the simplest treatment, originally used by Lorentz (1909), it is assumed that the damping of the electron motions occurs because of a retarding force $-mvV$, (3.11) below, where m is the electron's mass and V is the part of its velocity associated with the ordered motion imposed on it by the electromagnetic field. This is discussed in more detail in §§ 3.4, 3.12. Here ν is an effective collision frequency for electrons. A similar treatment can be used for the collisions of ions. In the formulae for the refractive indices, ν appears in the ratio $Z = \nu/\omega$, where $\omega = 2\pi f$ is the angular frequency. At high frequencies Z is small and can often be neglected.

One way of finding ν is to make measurements of the attenuation of radio waves. The attenuation occurs over a range of height z and care is needed in finding the particular height to which the measured ν applies. There are techniques for dealing with this and careful measurements have given useful values. See, for example, Davies (1969, ch. 6), *Journal of Atmospheric and Terrestrial Physics* (1962, special volume).

Measurements of this type include the use of rockets going through the lower parts of the ionosphere (Kane, 1962), and studies of the effect of electron collisions on the ionospheric reflection coefficients for radio waves of very low frequency (Deeks, 1966a, b, § 4).

There are many other methods of measuring ν. Two radio methods of great importance are measurements of partial reflections, § 11.3, and measurements of wave interaction, §§ 13.11–13.13. Measurements have also been made on plasmas in the laboratory under conditions that are intended to simulate the ionosphere. See, for example, Huxley and Crompton (1974), Phelps and Pack (1959). For a full list of

references see Shkarofsky, Johnston and Bachynski (1966, pp. 177–87, 199–201).

The subject of collisions between particles in a plasma is complicated and is studied in books on plasma physics, for example Clemmow and Dougherty (1969, esp. §9.5), Shkarofsky, Johnston and Bachynski (1966, chs. 4, 5). Any collision frequency is a statistical average with respect to the velocities of the colliding particles and the directions of their paths before and after the collision. In a full theory it is found that in different applications the averages must be taken in different ways, so that many different collision frequencies have to be defined (Suchy and Rawer, 1971; Suchy, 1974b). The one that is needed for studying the refractive indices for radio waves is the one that is used in the theory of transport phenomena, such as thermal conductivity and electrical conductivity. It is called the 'collision frequency for momentum transfer'. For a summary of the main points of the theory see Budden (1965).

In the Lorentz formula (3.11), $\boldsymbol{V}$ is the ordered part of the electron's velocity. It is superimposed on the random thermal velocity v which is much larger (see problem 3.8). The collision frequency $\nu(v)$ for electrons must depend on v because, if it did not do so, the phenomenon of wave interaction, §§13.11–13.13, would not occur. It was at one time supposed that all electrons have the same mean free path λ_e that is independent of v. Then $\nu(v) = v/\lambda_e$ is proportional to v. As a result of laboratory measurements, however, it is often assumed that $\nu(v)$ is proportional to v^2. This subject is dealt with in § 3.12. The simple Lorentz formula (3.11) uses an effective value of ν, often written ν_{eff}, that is independent of v, and it must be related to the average value ν_{av} of $\nu(v)$. It is usually assumed that, in the ionosphere, the electron velocities have a Maxwellian distribution and this can be used to find ν_{av}. It is thus shown in § 3.12 that if $\nu(v) \propto v^2$ then $\nu_{\text{eff}} = \frac{5}{3}\nu_{\text{av}}$ (3.87), and if $\nu(v) \propto v$ then $\nu_{\text{eff}} = \frac{4}{3}\nu_{\text{av}}$ (3.93).

The propagation properties of radio waves are not very sensitive to the exact value of ν so that it is not necessary to know ν with high precision. Throughout this book the Lorentz formula (3.11) is used and the symbol ν is used for the effective collision frequency, ν_{eff} above. It is useful to have some idea of how this ν depends on height z in the ionosphere. Below about 100 km electron collisions are predominantly with neutral molecules and ν is proportional to their concentration. In an isothermal atmosphere of constant composition it would be expected that ν is proportional to the pressure, thus

$$\nu = \nu_0 \exp(-z/H) \tag{1.11}$$

where H is the scale height defined in § 1.5 and ν_0 is a constant. In practice H takes different values at different heights because of height variations of temperature and composition, and this law can only be expected to hold over ranges of z so small that H may be taken as constant. Throughout the ionosphere the concentration of

positive ions is much less than that of neutral particles but the forces between electrons and ions are of longer range. Consequently at 100 km and above, the contribution to ν from ions must be allowed for; see §§ 3.4, 3.13.

A useful summary of the factors that affect the value of ν was made by Nicolet (1953, 1959). Based on this and other data, a curve was given by Budden (1961a, fig. 1.2) to show how ν depends on height z according to the best estimates then available. A new version of this curve is given in fig. 1.2. It has been revised slightly in the light of data from measurements of partial reflection, § 11.3, and of wave interaction §§ 13.11–13.13, but it differs very little from the 1961 version.

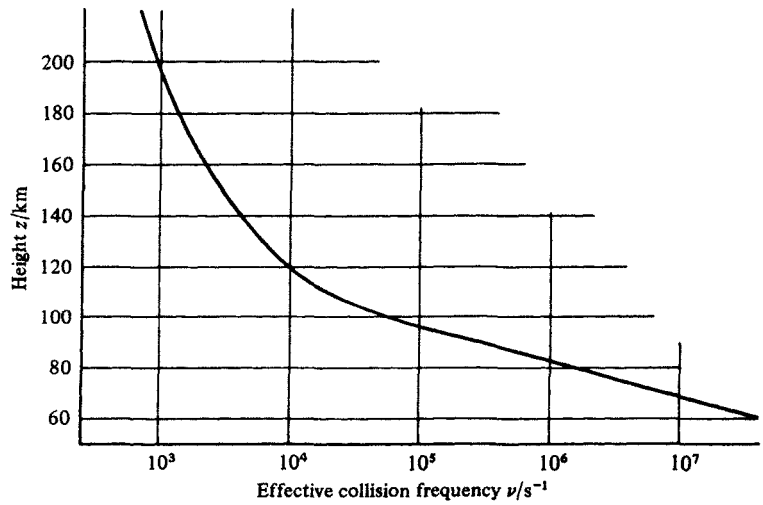
The value of ν does change slightly with time of day and with season, and the height distribution of ν depends on latitude. These changes are, however, very small and for many purposes they can be ignored. In this respect ν is very different from the electron concentration $N(z)$; see § 1.8. Fig. 1.2 applies for temperate latitudes but gives a useful first estimate of ν for any hour and season and latitude.

Because changes of ν affect the propagation of radio waves far less than changes of electron concentration $N(z)$; see §1.8. Fig. 1.2 applies for temperate latitudes but ranges of height z. This is done for many of the topics discussed in this book.

1.7. Observations of the ionosphere

Up to about 1960 our knowledge of the structure of the ionosphere was obtained almost entirely by radio methods, using signals from transmitters near the ground. The principal technique is known as ionosonde. Short pulses of radio waves are sent vertically upwards and return to a receiver on the ground after reflection in the ionosphere. The time of travel t is measured. This is the same as the technique used in

Fig. 1.2. Dependence of the effective electron collision frequency ν upon height z.



radar. It was first applied to the ionosphere by Breit and Tuve (1926). If the pulse travelled entirely in free space it would have to go to a height $h' = \frac{1}{2}ct$ called the equivalent height of reflection. This is strongly frequency dependent. The reflected pulse usually has two components, the ordinary and extraordinary waves with different values of h'.

Ionosonde equipments are commercially available and are used at observing stations all over the world. They produce curves, or records on magnetic tape, of equivalent height versus frequency $h'(f)$, which can be processed by a computer to calculate $N(z)$ (§ 13.6). Some ionosonde equipments now supply the data in digital form (see Bibl and Reinisch, 1978; Downing, 1979), and some include a computer for calculating $N(z)$, (see Reinisch and Huang, 1983). There are more elaborate equipments in which the ionosonde is linked to a computer that controls its frequency and aerial systems. For example an equipment known as the Dynasonde (Wright and Pitteway, 1979) gives information about polarisation, angle of arrival, phase, and other features of the reflected pulses.

Other ground based radio techniques include the following: reflection of waves of very low frequency (VLF), chs. 18, 19; partial reflection, § 11.13; wave interaction or cross modulation, §§ 13.11–13.13; absorption, §§ 1.6, 13.10. The device known as a riometer measures the absorption of the radio signal that comes in from outside the earth, mainly from the galaxy. This method was first suggested by Little and Leinbach (1959). For details see Rawer and Suchy (1967, § 46), Davies (1969, § 6.6.2). There are other methods, not described in this book, including Doppler shift of reflected radio waves (Davies, Watts and Zacharisen, 1962; Jones, T.B., 1964; Jones and Wand, 1965); refraction of radio signals received at the ground from satellites (Titheridge, 1964).

From about 1960 onwards rockets and satellites began to be used and it became possible for the first time to study the ionosphere at heights above the maximum of $N(z)$ in the F-layer, that is the region called the top side of the ionosphere. Two main methods were used. The first was the use of a measuring device or probe carried on the vehicle and designed to measure the electron concentration $N(z)$ and the electron temperature T_e. Such probes are called Langmuir probes. One difficulty is that the plasma being studied is disturbed by the vehicle moving through it (see Boyd, 1968; Willmore, 1970). The second method is the use of the ionosonde technique with a transmitter and receiver in a satellite above the F-region. This is called topside sounding. It is similar to the use of an ionosonde on the ground, but it shows many new features (§ 13.5). A third technique of some interest is the measurement at the ground of the Faraday rotation of a linearly polarised radio signal emitted by a satellite and travelling right down through the ionosphere; see § 13.7.

In recent years a new probing technique has been coming into use. This is called incoherent scatter or Thomson scatter. It was suggested by Gordon (1958) and first

successfully used by Bowles (1958, 1961). A radio beam of high frequency (typically 40–400 MHz) and high power is sent into the ionosphere from a transmitter on the ground. The plasma scatters some of the radiation and this is detected by a receiver also on the ground. The observed scattering region can be selected by adjustment of the highly directive transmitting and receiving aerials, or by time gating of the receiver.

The received scattered signal is extremely weak so that sophisticated signal integration methods have to be used in the receiver. But the method works well and gives the electron and ion concentrations and temperatures. For an account of it see Dougherty and Farley (1960), Fejer (1960, 1961), Evans (1969), Beynon (1974), Bauer (1975). The theory of the method involves a detailed study of the physics of warm plasmas and is beyond the scope of this book.

For a survey and assessment of various methods of measuring $N(z)$ and other properties of the ionosphere, see Booker and Smith (1970).

1.8. The structure of the ionosphere

A great deal of knowledge has now accrued about the structure of the ionosphere and how it depends on time of day, on season, on latitude and longitude and on solar activity. This has been incorporated into a computer program known as the International Reference Ionosphere (abbreviation: IRI), which predicts the height distribution of electron concentration $N(z)$ and of other features such as ion concentration and electron, ion and neutral temperatures. It is published jointly by the International Union of Radio Science (URSI) and the Committee on Space Research (COSPAR). For details see Rawer (1981, 1984). It is available in the computer libraries of some of the larger computers. It does not give concentrations of neutral particles or pressure (these can be obtained from the International Reference Atmosphere, CIRA 1972), and it does not give electron collision frequencies (see § 1.6). The version of IRI used here for fig. 1.3 is dated 1978. The program is revised from time to time and the user is advised to check that he has the latest version. A new version was announced by URSI in March 1983.

The ionosphere consists of two main layers known as E and F. For the E-layer, the electron concentration $N(z)$ has a maximum at a height of 100 to 110 km. Above this $N(z)$ is only slightly less for a height range extending right up to the base of the F-layer. Thus the E- and F-layers are not really distinct. For many purposes it is useful to assume that the part of the E-layer below its maximum is a Chapman layer having $N(z)$ given by (1.8), with $H = 10\,\text{km}$, $z_0 = 115\,\text{km}$ and $N_0 = 2.8 \times 10^{11}\,\text{m}^{-3}$. This means that the penetration frequency (§ 13.2) is $4.7 \times (\cos\chi)^{\frac{1}{4}}$ MHz. The actual behaviour of the E-layer is very much more complicated than this. At night the E-layer has a penetration frequency of the order of 0.5 MHz, corresponding to an electron concentration of $3 \times 10^{9}\,\text{m}^{-3}$.

A thin but intensely ionised layer sometimes appears at a height near 115 km. Its occurrence is sporadic and it is therefore known as 'sporadic E', or E_S. It nearly always gives partial reflection and penetration of radio waves over a wide frequency range. There are several different types of E_S. One form is believed to have horizontal variations of electron concentration N so that radio waves can penetrate through where N is small, but are reflected from where N is large. Another form is believed to be a uniform thin layer so that partial penetration and reflection occurs because of the thinness. For the theory of this see § 19.8. An E_S layer can be so strong that its effective penetration frequency is comparable with or greater than the penetration frequency of the F-layer, and then radio observations of the F-layer from the ground are made difficult or impossible.

Several mechanisms may contribute to the formation of E_S. It is believed that wind shears, that is a large vertical gradient of horizontal wind speed, play a major part. Ionisation caused by meteorites may have some effect. In auroral latitudes precipitation of charged particles may contribute. For a full account see Smith, E. K. and Matsushita (1962), Rawer and Suchy (1967, § 40), Rishbeth and Garriott (1969, § 6.3).

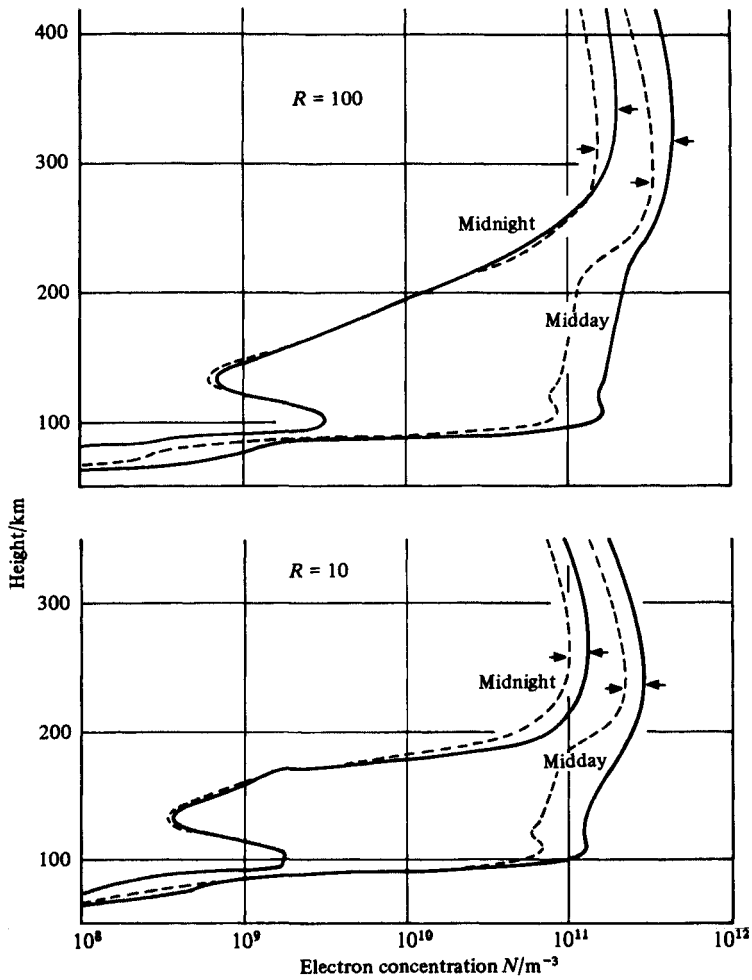
The F-layer is more heavily ionised, with a maximum of $N(z)$ in the range 200 to 400 km. Its diurnal and seasonal variations are more complicated than those of the E-layer. During daylight the function $N(z)$ often shows a bulge below the maximum (see fig. 1.3). This is known as the F1-layer, and it may occasionally attain an actual maximum. The main maximum above it is known as the F2-layer. It has been suggested that the formation of the whole F-layer is caused mainly by a single ionising agency which would form a simple layer like a Chapman layer if the attachment coefficient β (§ 1.5) were constant at all heights. Bradbury (1938) suggested that β decreases as the height z increases, so that the maximum of $N(z)$ in the F2-layer arises, not from a fast rate of production but because of a slow rate of removal of electrons. But this alone would not be adequate to explain the F2 maximum. For a detailed discussion see Rishbeth and Garriott (1969, § 3.62 and Ch. IV).

In chs. 10 and 12 some account is given of methods of finding the function $N(z)$ in the F-layer from radio observations using the ionosonde technique. This work shows that, at the maximum, the curvature of the $N(z)$ curve is about the same as that at the apex of a parabola of half thickness 100 km, or that of a Chapman layer with scale height about 50 km. The F-layer is therefore much thicker than the E-layer, as well as more heavily ionised. In temperate latitudes the penetration frequency of the F-layer ranges from about 2 MHz at night to 8 MHz in a summer day. These correspond to electron concentrations of 5×10^{10} and $8 \times 10^{11} \mathrm{m}^{-3}$ respectively.

In the daytime there is another layer below the E-layer, known as the D-layer with a maximum of $N(z)$ near $z = 80$ km. This overlaps the lower part of the E-layer so

that although $N(z)$ is enhanced, it is still a monotonically increasing function as shown in the example of fig. 1.3. The first experimental study of the D-layer came from radio observations at very low frequencies (VLF, of the order of 16 kHz). Here the wavelength is so great that interpretation of the observations is less direct than at high frequencies. One of the most important features of the full wave theory given in this book is that it has helped to disentangle the numerous radio observations at VLF. The procedure most often used is to adopt a trial function $N(z)$ and work out its reflecting properties. If these do not agree with observations, some other $N(z)$ is

Fig. 1.3. Examples of the dependence of electron concentration on height in the ionosphere as computed from the International Reference Ionosphere. The continuous curves are for June and the broken curves for January. R is the sunspot number. $R = 100$ is typical of a year near sunspot maximum and $R = 10$ of a year near sunspot minimum.



tried until a satisfactory result is obtained. The process is laborious because of the variability of the radio observations with time of day and season, and with ionospheric disturbances, but it has been successfully used by Deeks (1966a, b). The problem of finding $N(z)$ from the observations, directly without a trial and error process, is known as 'inversion', and techniques for doing it have been given by Backus and Gilbert (1967, 1968, 1970). Several workers have used this kind of technique for radio observations. For details see § 19.7.

1.9. The magnetosphere

For many purposes the earth's magnetic field can be regarded as that of a magnetic dipole at the centre of the earth. For greater accuracy the dipole is displaced from the centre and quadrupole and higher multipole terms must be included but these refinements are not needed in this book. If the earth were in a vacuum the strength of its magnetic field would decrease as the inverse cube of the distance from the centre. This is approximately true in the ionosphere but does not hold beyond about three or four earth's radii from the centre because outside the earth there is a fully ionised plasma. This is a stream of gas coming out from the sun, called the 'solar wind'. Near the earth's orbit its outward speed is about $200\,\mathrm{km\,s^{-1}}$ and it consists mainly of electrons and protons each with a concentration of about $10^7\,\mathrm{m^{-3}}$. This plasma has a strong effect on the outer parts of the earth's magnetic field.

When a magnetic field is present within an electric conductor, it cannot be changed quickly because any change induces e.m.f.'s that cause currents to flow, whose magnetic fields oppose the attempted change. This effect is well known for metals made superconducting by cooling them to the temperature of liquid helium. Any change of magnetic field in a superconductor is impossible and the field is said to be 'frozen in'. In a conducting body that is not superconducting, the magnetic field can change but only within a time period of order $\tau \approx \mu_0 \sigma l^2$ that depends on the electric conductivity σ and the linear dimension l of the body. Here μ_0 is the permeability of a vacuum; § 2.1. (See problem 1.1.) For the same reasons, a magnetic field cannot quickly penetrate into a conducting body. For the solar wind σ is of order $0.03\,\mathrm{S\,m^{-1}}$. If l is of order two or three times the earth's radius $\mathrm{R_e}$, this gives τ of order several months.

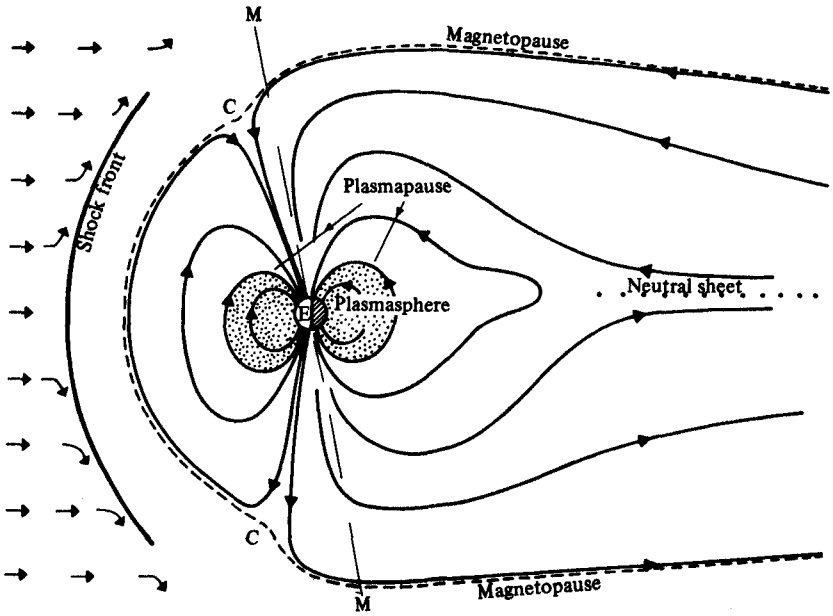
The earth's magnetic field therefore cannot penetrate into this advancing plasma stream. On the sunward side of the earth there is a surface at about 10 to 12 $\mathrm{R_e}$ from the centre where the earth's field ends abruptly. It is called the 'magnetopause' and surrounds the earth as sketched in fig. 1.4. Beyond it is a region of turbulence called the 'magnetosheath' and outside this on the sunward side is a shock front on which the solar wind impinges. The region inside the magnetopause is the magnetosphere, and the plasma here moves with the earth and is regarded as part of the earth's upper atmosphere. The solar wind plasma flows round the outside of the magnetosphere.

On the side of the earth remote from the sun the earth's magnetic field and the magnetosphere extend back into a region called the 'magnetotail', that goes out to a distance of 100 R_e or more.

At heights greater than about 1000 km above the earth's surface the plasma in the magnetosphere is a good conductor and its distribution is strongly influenced by the magnetic field, as explained in § 1.1. The effect of gravity here is small. Observations have shown that the magnetospheric plasma is in two main parts. The inner part is continuous with the upper ionosphere and is called the 'plasmasphere'. It extends out to a surface which, at the equator on the sunward side, is at about 4 to 5 R_e and it follows the magnetic lines of force; see fig. 1.4. It is called the 'plasmapause'. Just inside it the electron concentration is very roughly $10^8 m^{-3}$ and just outside it about $10^7 m^{-3}$. Thus there is a fairly sudden decrease when the plasmapause is crossed. This was first discovered in observations of whistlers, § 13.8, Carpenter (1963, 1968, 1983), Carpenter and Park (1973), and has since been confirmed and studied by measurements in space vehicles.

The figures given in this section and in fig. 1.4 for the dimensions of the magnetosphere and for the electron and ion concentrations are very approximate. In

Fig. 1.4. Sketch showing cross section of the magnetosphere by a plane containing the magnetic axis MM and the sun–earth line, at a time when this plane is at right angles to the ecliptic. The continuous lines with arrows are magnetic lines of force. The short arrows show the direction of flow of the solar wind. The letters CC denote polar cusp regions. For explanation of these and of the neutral sheet see references in text.



practice they vary with time and are strongly influenced by radiation from the sun, by fluctuations in the solar wind and by the magnetic fields frozen into the solar wind plasma.

For a most illuminating discussion of the processes that result from the solar wind impinging on the earth's magnetic field see Ratcliffe (1972). For a description of the magnetosphere and further references see Al'pert (1983, especially tables 2.1, 2.2), Ratcliffe (1970, 1972), Kennel, Lanzerotti and Parker (1979).

1.10. Disturbances of the ionosphere and magnetosphere

The earth's atmosphere is subjected to many disturbing influences that affect the ionosphere and magnetosphere. These do not play any essential part in the theoretical topics discussed in this book, but some of them should be mentioned briefly. Most of them originate in events on the sun and are more frequent and intense near times of maximum sunspot number. A good general account is given by Ratcliffe (1970, 1972).

An eruption on the sun's surface, known as a solar flare, can occur very suddenly. It is seen as a brightening of a small area, usually near a sunspot, and it results in intense emission of ultraviolet radiation that causes ionisation in the D-region of the ionosphere. This is known as a 'sudden ionospheric disturbance' (SID). It occurs sometimes within a few seconds and results in strong absorption of radio waves, so that communication circuits that rely on ionospheric reflection are interrupted. This is referred to as a 'radio fadeout', or 'short wave fadeout'. There are accompanying disturbances in the earth's magnetic field, and marked changes in the propagation of radio waves of very low frequency (10 to 100 kHz) which are strongly influenced by the D-region. The ionisation decreases again and returns to its normal value in a time of order one hour.

Energetic charged particles are emitted from the sun, particularly from solar flares. Those that enter the magnetosphere are deviated by the earth's magnetic field. They move on helical paths in the direction of the lines of force but cannot easily move across the field. They can therefore reach low altitudes only at high latitudes, that is in polar regions. Here they produce increased ionisation in the D-region. This results in increased absorption of radio waves, known as 'polar cap absorption' (PCA). The particles that cause this are mainly protons.

A prolonged and complicated disturbance can be produced by changes in the solar wind. There can sometimes be an increase in the concentration and the velocity of the particles in the wind, accompanied by a magnetic field frozen in to the moving plasma. When this strikes the earth it causes changes in the earth's magnetic field, known as a 'magnetic storm', changes in the ionosphere, known as an 'ionosphere storm', and changes in the magnetosphere. The current flows in the ionosphere, magnetosphere and magnetotail are altered and the size of the plasmasphere

becomes much reduced. The modified solar wind can reach the earth suddenly and this gives a 'sudden commencement'. The subsequent disturbance evolves in a complicated way and may last several days. The main effect in the ionosphere is an increase of ionisation. This may be in the D-region and then leads to increased absorption of radio waves. There may also be an increase at higher levels including the production of sporadic E-layers (§§ 1.8, 19.8).

Meteors and meteorites can produce enhanced ionisation, mainly in the D- and E-regions. The ionisation in a meteor trail has a cylindrical structure and its effect on radio waves is an interesting theoretical problem but beyond the scope of this book.

Eclipses of the sun disturb the production process of ionisation in the upper atmosphere. Radio observations during eclipses have given valuable information about the complicated chemical processes involved in the formation and decay of ionised layers. See Beynon and Brown (1956), Rishbeth and Garriott (1969, § 6.2).

The ionosphere can be influenced by large scale motions of the earth's atmosphere. Both solar and lunar tidal motions have been observed. Acoustic gravity waves or planetary waves have already been mentioned; see end of § 1.4. Waves of this type were generated in the atmosphere by a large chemical explosion at Flixborough, England, in 1974 and their effect on the ionosphere was studied by radio methods; see Jones and Spracklen (1974).

Many of the results of radio observations can be explained by assuming that the ionosphere is horizontally stratified, that is that the electron concentration and collision frequency are functions only of the height z. This assumption is adopted throughout this book, except in § 11.13. There are, however, many other observations that show that there must also be horizontal variations of electron concentration. These are often irregular and are subject both to steady movements and to random change. Scattering from ionospheric irregularities can be used for some forms of communication link (Booker, 1959). One of their effects is to cause fading or scintillation of radio signals. These permit measurements to be made of steady motions, usually called winds, in the ionosphere. For reviews see Briggs and Spencer (1954), Ratcliffe (1956), Yeh and Liu (1982).

This subject has become part of the much wider study of waves in random media, with applications in radio astronomy and in ocean acoustics. See Uscinski (1977).

PROBLEMS 1

1.1. Derive a partial differential equation relating the variation with space and time coordinates of a magnetic field in a homogeneous isotropic solid of high electrical conductivity, when the displacement current is negligible compared with the conduction current, and obtain a solution for it in one case of physical interest. In what circumstances is the neglect of the displacement current permissible?

Estimate a rough order of magnitude for the decay time of a magnetic field in (a) a

copper sphere of diameter 1 m and resistivity $1.7 \times 10^{-8}\,\Omega$m, (b) a cloud of interplanetary plasma of diameter 6×10^{4} km containing electrons with concentration $10^{8}\mathrm{m}^{-3}$ if the average number of collisions made by each electron is 10 per second.

[Natural Sciences Tripos. Part II Physics, 1968.]