CHAPTER II

THEORIES OF IRREGULARITIES IN THE EQUATORIAL IONOSPHERE

Over the years, irregularities in the equatorial ionosphere have provided a wide variety of perplexing, yet interesting phenomena such as sporadic-\(E\), spread-\(F\), fading of HF radio waves, VHF/UHF scintillation, HF and VHF scatter etc. Therefore, the investigations of these irregularities generated in the equatorial ionosphere have been a scientific endeavour for several decades. A number of theories have been developed to explain the nature of ionospheric E and F region irregularities. In this chapter a brief description of these is outlined.

2.1 E-region Irregularities

Irregularities in the ionosphere have been observed as early as 1932 by Eckersley. An important class of ionospheric irregularity is the so-called 'sporadic E layer' or Es normally found between 100 and 120 km. In its weaker forms, sporadic E consists of clouds of ionization and in its most intense form consists of a thin sheet of ionization some tens or hundreds of meters in thickness. At high latitude such ionization is often associated with the appearance of the aurora. This suggests that polar sporadic-\(E\) might be caused due to particle precipitation. At mid-latitudes one of the prime causes of sporadic-\(E\) is believed to be horizontal wind movements across the earth's magnetic field, coupled with the existence of large vertical gradients in velocity, "wind shears". One type of sporadic-\(E\) (Es-q) observed at
equatorial and low latitudes is believed to be due to plasma instabilities associated with the large electric current flowing in the ionosphere along the magnetic equator - the so called equatorial 'electrojet'. Whitehead (1970) has given a summary of types of Es in all the three latitude regions, viz. high, middle and low.

In the equatorial region, generally Es appears as a patchy and particularly transparent layer, while sometimes it appears in sheets which completely blankets the overlying F-region and is known as 'blanketing' type Es or Es-b. Occurrence of Es is predominantly daytime phenomenon (Rengarajan 1954, Smith 1957, 1962) and is strongly correlated with the equatorial electrojet current (Matsushita 1951). Matsushita (1957) showed that occurrences of Es-q are negatively correlated with the geomagnetic disturbances. Occasionally, even on quiet days, disappearances of Es-q for short periods are associated with abnormally low values of horizontal magnetic field intensity (Cohen et al. 1962, Rastogi et al. 1971, Krishna Murthy and Sen Gupta 1972). This disappearance of Es-q is attributed to the reversal of electrojet currents during that period and is termed as 'counter-electrojet' (Gouin and Mayaud 1967). Rastogi (1972) pointed out that the temporary disappearance of Es-q during daytime is caused by the imposition of additional westward electric field, causing the inhibition of irregularities. The highly transparent Es-q is now mostly attributed to a rather inhomogeneous plasma structure believed to be produced due to
plasma instabilities, e.g. two stream and cross-field instabilities which will be discussed in this section.

2.1.1 Early Theory of E-region Irregularities

Clemmow and Johnson (1959) developed a theory of the motion of weak irregularities in an otherwise homogenous partially ionized gas under the influence of electrostatic and magnetostatic fields. Later their work was extended by Whitehead (1963) to explain how the irregularities may be formed by a low frequency sound wave passing through the medium and producing variations in air density. These variations in air density lead first to corresponding variations in the rate of production of electrons, but of far more importance as it turns out, also a divergence of ionization due to the neutral air movement. This later effect may lead to fractional increases in electron density up to the order of $10^4$/cc of the corresponding changes in air density. The sound waves of importance are those with a velocity component at right angles to the magnetic field. Sound wave moving along the magnetic field produces only fractional electron density irregularities equal to the fractional changes in air density and therefore the irregularities produced in general will be greatly elongated along the magnetic field.

The process leading to the build-up of irregularities is not, however, one of instability, and is limited by the amplitude of the original sound waves and recombination of ionization rather than the onset of a non-linear mechanism. Thus the electron
density may be quite weak. Thus basic problem with this theory is the difficulty of creating, in neutral gas, strong sound waves having the required short wavelengths. The damping of such waves is strong because of the long mean free path of the neutral molecules. Whitehead found the minimum theoretical irregularity wavelength to be about 9 meters. In fact, wavelengths of 1 meter have been observed. Thus the theory of generation of irregularities seems to be inadequate to generate E-region irregularities.

2.1.2 Two-stream Plasma Instability Theory

One of the accepted theories of B-region irregularities is two-stream plasma instability. This has been studied quite extensively (Buneman 1959, Bernstein and Kulsrud 1960, Fried and Gould 1961) for laboratory plasmas. Farley (1963) extended the theory so as to explain the generation of irregularities in the E region of the ionosphere particularly in the daytime. If the plasma contains particles with directed energy, for example, due to the electron drift velocity (giving rise to the electron streaming) of the current carrying electrons, another electrostatic instability, namely the two-stream instability can be excited. When the current (or drift velocity) reaches a certain threshold level, the plasma becomes unstable and strong electrostatic type waves with phase fronts aligned with the magnetic field are generated by the excitation of the two-stream instability (Buneman 1959).
The two-stream instability develops due to the interaction of two groups of plasma particles drifting relative to each other. The instability gives rise to growing electrostatic waves (longitudinal), which derive their energy from the kinetic energy associated with the relative drift velocity of the particles of the two groups. These waves are high frequency oscillations with frequency greater or equal to ion gyrofrequency. In the equatorial electrojet, the electron drift relative to ion is larger than the ion acoustic speed. This is one of the important conditions for the plasma to be unstable and is easily satisfied at least during daytime. Thus longitudinal ion-acoustic electrostatic plasma waves can be excited in the electrojet by the two stream instability driven due to cross-field streaming electrons relative to ions. These smaller wavelength waves growing in amplitude, propagate in the direction exactly normal to geomagnetic field and manifest themselves as the irregularities in the ionosphere.

Physically one can understand the mechanism of this instability to a quite similar way as cross-field instability (which would be described later). The E-region of equatorial ionosphere is characterised by magnetized electrons (electron collision frequency is very much less than electron gyrofrequency) and collisional ions (ion collision frequency is very much greater than ion gyrofrequency). The crossed ambient vertical polarization electric field and the horizontal geomagnetic field, then lead to a cross-field streaming of electrons with respect to ions, as the ions are tied up with
the neutrals because of collisions. Now let us assume that the horizontal density interface is perturbed by a small amplitude plane wave oscillations which would give rise to the secondary polarization electric field, alternately in direction at the interval of half wavelengths because of the phase difference between electron and ion density profiles which arises due to the ions lagging behind the electrons as the ions are much heavier than the electrons. These secondary electric fields are crossed with geomagnetic field, give the forces to make crests and troughs of the perturbation to move downwards and upwards respectively. If initially the magnitude of relative electron ion streaming drift velocity is greater than ion acoustic speed, the given perturbations become unstable and in turn, they will continue to grow in amplitudes. Later a number of workers (Skadron and Weinstock 1969, Register 1971, Sato 1972) used a nonlinear approach for two-stream instability to explain some of the features like spectrum of irregularities observed in the equatorial region by in-situ measurements. Linear theory can predict when the instability will occur but can not tell its destiny. On the other hand, observations measure the steady state because the geophysical parameters change very slowly compared to the time scale of instability ($\sim 1$ sec). Therefore, in order to explain the observed features, non linear theory is necessary. Sato (1972) using non-linear approach calculated theoretically $(\Delta N/N)^2 \sim 4 \times 10^{-3}$ which is in agreement with the
observed results (Bowles et al. 1963). All these workers agree that the two stream instability is generated when the electron drift velocity reaches the threshold which is the ion-acoustic velocity.

Around the same time when Farley (1963) suggested the two-stream instability mechanism for equatorial E-region irregularities, Bowles et al. (1963) using equatorial observations pointed out that at times when the threshold of two-stream instability is not reached, a vertically pointing radar detects echoes which did not have the spectra similar to one expected theoretically by two-stream instability. The large scale blobs of irregularity were also found to be moving in westward direction. These observations, however, could not be reconciled with the Farley's two-stream instability theory. A detailed study of these irregularities was made by Balsley (1967) who classified the irregularities in two types namely 'type I' and 'type II'. Type I irregularities have a narrow, sharp and stable spectrum with a distinct maximum at a constant Doppler shift that corresponds at least to the ion-acoustic velocity of the medium. Type II have a rather much broader spectrum, with the variable width often greater than the mean Doppler shift. These are observed even when the relative electron ion drift is less than the ion-acoustic velocity in the electrojet. Balsley (1967) suggested that two separate mechanisms must be operating in the electrojet which give rise to these two types of irregularities. This classification is based on backscatter radar and rocket measurements mainly.
Dougherty and Farley (1967) attempted to account for the type II irregularities by invoking non-linear mode coupling of type I irregularities, but this was not acceptable because type II irregularities can be present when the type I are absent. In-situ rocket measurements of Prakash et al. (1971a) showed that the threshold requirement of two stream instability was not fulfilled in some cases when type II echoes were observed. Further, they clearly suggested that these echoes are due to small scale irregularities which are generated through cross-field instability mechanism. Rastogi (1972) reported that the creation of irregularities giving Es-q reflection on ionograms is identified as cross-field (plasma gradient) instability.

Rastogi et al. (1975) studied the sudden appearance and disappearance of sporadic E echoes at Huancayo and compared their results with the estimates of cross-field instability calculations. The comparison reported the view that Es-q type of irregularities in the equatorial electro jet region are generated due to cross-field instabilities.

2.1.3 Cross-field Instability Theory

Cross-field instability is also known in the literature by names like 'gradient-drift instability' because of the fact that the charged particles drifts are involved in the excitation in the presence of density gradient acting as the energy source of the instability and 'drift-resistive instability' because of
en important role played by the neutral particles, giving rise to finite resistance. Martyn (1959a,b) suggested that plasma density enhancement aligned with the magnetic field in the ionosphere may acquire a velocity relative to the background ionization due to polarization effects. The cross-field instability is mainly based on the idea that the relative drift in a direction anti-parallel to a density gradient in the background plasma causes an increase in the relative enhancement.

The cross-field instability was investigated theoretically by Simon (1963) and Hoh (1963) for the case of laboratory plasma. Its possible application to the formation of sporadic-E irregularities in the ionosphere has been studied by Tsuda et al. (1966) and physical mechanism is as follows.

Let us consider low latitude in the northern hemisphere and the background charge density increases vertically upward. Further there exist an applied DC electric field (upward), together with the ambient magnetic field (pointing out of the paper). Let the charge density now fluctuate in horizontal extent in such a way as shown in Fig. 2.1, then the medium will be polarized by drift motions of electrons and ions at right angle to both the electric and the magnetic fields, that is westward in this figure. Since the intensity of induced electric fields depends on the charge density, their direction reverses relatively at intervals of half wavelength of fluctuations. Thus, the charged particles are constrained to move downwards or upwards.
Fig. 2.1 Irregularity formation process, polarization fields (small arrows) and the magnetic field (out of paper) force the ionization along the vertical direction (dashed arrows), causing the density fluctuation shown to enhance if the background density increases with increasing height (after Tsuda et al., 1966).
depending on whether they are immersed in dense or dilute ambient plasma, since northward component of the magnetic field and the induced fields make them drift. Thus charge density fluctuation develops, making clearer contrast against the background density. The irregularities thus formed presents a striped pattern parallel to the magnetic meridian plane.

The irregularities generated by this mechanism were thought to be type II irregularities (Prakash et al. 1971b), having smaller (3-15 meters) and larger (30-300 meters and more) scale sizes and are observed even when the electron drift is below the ion-acoustic speed in the medium. These irregularities rather have a broader spectrum and are practically present always at least at long wavelength (Farley and Balsley 1973).

However, Sudan et al. (1973) reported that the linear approach of this theory cannot account for the 50 MHz radar backscatter observations, which require irregularity wavelength of 3 meters. The marginal unstable fluctuations have wavelengths of several tens of meters or longer; shorter wavelength irregularities are heavily damped by electron thermal diffusion across the magnetic field. Moreover for electron drift velocities less than the ion-acoustic velocity, there appears to be no other linear instability that can directly generate the short wavelength irregularities in the E-region. Thus they worked out the non-linear process for cross-field instabilities, by invoking the mechanism of generation of higher order spatial harmonics.
Both two-stream and cross-field instability mechanisms explain quite satisfactorily many observed characteristics of type I and type II irregularities in the equatorial E-region.

2.2 F-region Irregularities

Currently there has been a very keen interest in the equatorial F-region irregularities, which for many years have been detected by the spreading of the echoes as seen in ionograms. This was first reported by Booker and Wells (1938). F-region irregularities have been investigated by various techniques like topside and bottomside ionosonde, VHF backscatter radar, rocket, drift and radio star and satellite scintillation. As a consequence, a wealth of observational data exists regarding various features of these irregularities.

Several comprehensive morphological studies of the occurrence of spread-F were made (Rober 1956, Shimazaki 1962, Singleton 1960, 68). Further, from careful examination of the ionograms the characteristics of spread-F have been studied by a number of workers (Bhargava 1958, Marasigian 1960a,b, Lyon et al. 1961, Rastogi and Kulkarni 1969, Huang 1970, Skinner and Kelleher 1971, Chandra and Rastogi 1972, Sastri and Murthy 1975). At high latitudes due to earth's magnetic field configuration the solar wind particle precipitation causes spread-F (Dyson and Winningham 1974) which is absent at the equatorial latitudes. This suggest that the equatorial and non-equatorial spread-F are two distinct phenomena. However, both involve field-aligned irregularities of electron density in the
F-region but it is likely that the irregularities are produced by different mechanism. Here discussion is mainly restricted to equatorial F-region irregularities.

Several theoretical attempts have been made to understand the generation mechanism of the irregularities in the F-region of the ionosphere. Some of the important theories are given below.

2.2.1 The Drift Instability Theory

Martyn (1959a,c) pointed out that a cylindrical excess or deficiency of ionization would move at velocity different from that of the background ionization in the presence of the orthogonal electric and magnetic fields owing to the surface charges built upon it. The apparent magnitude of the irregularity would grow exponentially on the underside of an upward drifting layer or the topside of a downward drifting one. This suggestion was negated by Dougherty (1959) and by Fejer (1959) on the grounds that any surface-charge effect in the F-region would be very small and that the real control of motion of such an irregularity would reside in the dynamo region; if the fields were uniform, the irregularity will move as though embedded in the F-region. Farley et al. (1970) indicated the inability of this theory to generate irregularities in the F-region, because spread-F occurs simultaneously above, below and at the F peak irrespective of the vertical motion of the layer, whereas drift instability requires
a special relationship of the irregularity with the ambient ionization.

2.2.2 Cross-field Instability Theory

The cross-field instability was first investigated by Simon (1963) for laboratory plasma and later applied to the ionospheric E-region as described in Sec. 2.1.3 where the Pedersen conductivity is the largest. This instability relies on ion-neutral collisions in the $E \times B$ field, and therefore it is much less effective at F-region altitudes. Reid (1968*) suggested that this difficulty can be overcome if the instability is taken to be in the E-region at the latitudes connected by the magnetic field lines to the equatorial F-region. However, the problem with a mechanism involving the coupling between the E and F regions is that it is effective for wavelengths of the order of a kilometer or more (Farley et al. 1970) and therefore irregularities of 3 meter scale-size cannot be explained by this mechanism.

2.2.3 Rayleigh-Taylor Instability (Including Its Extension to Hole or Bubble) Theory

Originally Dungey (1956) proposed the Rayleigh Taylor (R-T) instability as a source of equatorial F-region irregularities. This is also known as gravitational instability. The basic requirement for R-T mode to be operative is the presence of electron density gradient anti-parallel to acceleration due to
To understand the physical mechanism, let us consider a plasma boundary lying in the X-Z plane and density gradient $\Delta N$ in the negative y direction. We assume acceleration due to gravity $\vec{g}$ and a magnetic field $\vec{B}$ as shown in Fig. 2.2. Under this condition, in the equilibrium, electrons and ions will experience opposite drifts as given by

$$
V_i = \frac{M_i}{e} \frac{\vec{g} \times \vec{B}}{B^2}
$$

and

$$
V_e = -\frac{m_e}{e} \frac{\vec{g} \times \vec{B}}{B^2}
$$

where $V_e$, $V_i$ and $m_e$, $M_i$ are the drift velocities and masses of electron and ion respectively of the plasma. As the ratio of $m_e/M_i$ is very large, the electron drift $V_e$ can be neglected in comparison to ions' drift $V_i$. Now if we assume a small ripple in the interface as shown in Fig. 2.2b, the drift $V_i$ will cause the ripple to grow. The drift of ions causes a charge to build up on the sides of the ripple, and due to this an electric field $\vec{E}$ develops as shown in the figure. As a result of this electric field an additional $\vec{E} \times \vec{B}$ drift is always upward in those regions where the surface has moved upward and downward where it has moved downward. Thus ripple grows as a result of these properly phased $\vec{E} \times \vec{B}$ drifts.

Balesley et al. (1972) proposed that the long wavelength end of the spectrum can be attributed to the collisional limit of the R-T instability. Thus R-T mode is not applicable to the topside
(a) A PLASMA SURFACE SUBJECT TO A GRAVITATIONAL INSTABILITY

(b) PHYSICAL MECHANISM OF THE GRAVITATIONAL INSTABILITY

FIG. 2.2
spread-F phenomenon because the density gradient is parallel to acceleration due to gravity $g$ which is against the basic requirement for the R-T mode to be operative.

Hudson and Kennel (1975) showed that the collisional drift mode can grow on topside as well as on bottomside density gradients in the F-region of the equatorial ionosphere and thus this mode is applicable to the topside spread-F as well. However, they concluded that for the bottomside spread-F, both the R-T and drift modes can contribute to the total spectrum of spread-F irregularities, the drift mode peaking at shorter perpendicular wavelengths. They argued that long wavelength irregularities provide density gradients which in turn linearly excite collisionless drift waves (two-step process).

Chaturvedi and Kaw (1976) showed that the two-step process (Hudson et al. 1973) can not give rise to short wavelengths growing on long wavelength drift waves. This is because of the fact that in a drift wave the first order electron diamagnetic drift, which supports the secondary drift mode is absent. They suggested that a drift wave can grow on the density gradients of a R-T mode or some other mode. As the large scale irregularities increase in amplitude, the gradient associated with them becomes sharper. When this gradient exceeds a critical value, the long wavelength mode becomes unstable to collisional as well as collisionless drift waves by a two-step process. This mechanism is applicable
to small and large scale irregularities in the bottomside of the F-region.

Woodman and LaHoz (1976) based on their radar observations of equatorial F-region irregularities suggested that hierarchy of instabilities, starts from R-T node which takes place at large scale size in the bottomside of the ionosphere and later cascade down to 3-meter wavelength seen by the Jicamarca radar. In order to explain the irregularities in the region of negative gradient at the top of the F-region, they considered a rough model consisting of three fluids the lightest of them at the bottom, the heaviest at the middle and one of medium density at the top. The lower interface would represent the bottom of the F-region, and the second one the region of negative gradient. They envisioned that bubbles of lighter fluid will form in the lower interface (by R-T mechanism) and will be transported up by buoyancy effect all the way to the stable interface. But since the top fluid, although lighter than the one on the second level, is denser than the fluid contained in the bubble, the latter will break the stable interface and will continue to float through the top fluid until a region of equal or smaller density is reached (as shown in Fig. 2.3). The discovery of plume like structures which relate to irregularities at high altitudes on the topside with those on the lower side give support to R-T instability and to the interpretation that this instability propagates nonlinearly to the stable high altitudes in the form of bubbles or intruding fingers of low density plasma.
(a) Typical Equatorial Night Time
Electron density (e_i/c_c)

(b) Formation of Bubble
(After Woodman & LaHoz)

FIG. 2-3
In-situ rocket and satellite observations (Hanson and Sanatani 1973, Kelley et al. 1976, McClure et al. 1977, Morse et al. 1977) reveal sharp bite-outs of plasma density, also interpreted as Rayleigh-Taylor bubbles or depleted density regions. Recently Ott (1978) considering the collision-dominated and inertia-dominated situation in the F-region of the equatorial ionosphere, obtained the cylindrical bubble rise velocity without making any assumption but the relative strengths of ion inertia and ion neutral collisions and found that in the collisional limit the bubble velocity (vertical) depends on its altitude but not on size, while in the inertial limit, this velocity is independent of altitude but depends on size.

Ossakow and Chaturvedi (1978) attempted a more general solution in that the result for elliptical bubbles up to an axial ratio of 10:1 and found a common feature to all models that the higher the altitude and/or the higher the density depletion, the higher is the vertical rise velocity. The concept that bite-outs and spread-F can be one and the same phenomenon has been found to be in agreement with the recent ion mass-spectrometer data from atmosphere explorer-C (Szuszczewicz 1978). The chemistry and transport model which emerged from his work is one which allows a given chemical volume on the bottomside F layer (\([\text{NO}^+]\), \([\text{O}_2^+]\)>\([\text{O}^+]\)) to move upward through a stationary neutral atmosphere and appear at higher altitudes as a bite-out in the local plasma density.

2.2.4 The Spatial Resonance Mechanism

Beer (1974) suggested that the generation of F-region irregularities is a multistage process. The first stage starts
when the phase velocity of an atmospheric gravity wave matches the component of the ionization velocity in the direction of the wave's phase velocity. Atmospheric gravity waves are manifested by a wavelike perturbation in the neutral gas density and temperature. Wavelike perturbations of the ionization then occur through collisions between the charged particles and the neutral particles and also from charges induced by atmospheric waves in the rate at which the temperature-dependent loss processes operate. If the velocity of the resultant irregularity has a component which is the same as the phase velocity of the gravity wave and in the same direction then the peaks of the perturbation will continue to stay in the same position relative to the wave. The perturbation will maintain the same phase relation to the wave and as the process continues the irregularity will continue to increase. This has been called as spatial resonance (Whitehead 1971, Beer 1973a). The spatial resonance cannot continue indefinitely and it will cease when the charged particle density of the irregularity becomes so large that the system satisfies the collision-dominated R-T instability criterion. This is the second stage of generation of irregularities.

Further Beer (1974) calculated the time between the start of spatial resonance and the onset of instability in the absence of any background ionization gradients. His calculations also predicted lowest height limit of instability which he used in explaining the observed relationship between equatorial spread-F
and Fe\textsuperscript{+} ions (Hansen and Sanatomi 1971) as follows. That the presence of Fe\textsuperscript{+} is merely an indicator as to whether the daytime rise of the ionosphere has been sufficient to bring the F-region peak above the lower limit of the instability, in which case Fe\textsuperscript{+} ions will have also been transported upwards. If the upward drift during the day is weak, Fe\textsuperscript{+} ions will not be transported into the F-region; most of the F-region will lie below the lower limit of instability and thus no spread-F. This mechanism is critically dependent on the gravity waves required for producing the irregularities through spatial resonance. Chimonas (1970) considered theoretically the effectiveness of the equatorial electrojet in generating gravity waves and concluded that it is not a very efficient source. Beer (1973b) proposed that the supersonic motion of the earth's terminator may generate gravity waves, but indicated that the westward phase speed of the waves so generated, makes them unlikely contenders for the production of F-region irregularities which are in general observed to drift eastwards. Although certain fraction of the wave energy generated lower in the atmosphere leaks upward, it may not be sufficient to trigger or maintain the irregularities in the F-region for a considerable time. Thus gravity waves necessary for the generation of irregularities in the F-region of the equatorial ionosphere will have to be generated in the altitude region above 100 km. Recently Röttger (1978) has shown that the spatial resonance condition can be fulfilled in the equatorial region during post-sunset hours.
In spite of the fact that a large amount of experimental and theoretical work have been carried out, the study of the F-region irregularities still has a number of unsolved problems. As concluded by Aarons (1977) there is a need of a unified theory which must bring out the following:

(a) An explanation for the latitudinal extent and longitudinal variation of the irregularity region.

(b) An answer to the magnetic control question, i.e. sometime negative correlation with the irregularity occurrence and intensity, while sometime either no correlation or positive correlation.

(c) The causes of the variation in the thickness of irregular region and of the clouds of small scale irregularities.