Numerical Simulations of E–region Irregularities and Ionosphere–Magnetosphere Coupling

Pekka Janhunen

Preface

The work leading to this thesis was carried out at the Department of Geophysics of the Finnish Meteorological Institute. It was funded by the Academy of Finland as research project 2031010 from summer 1990 to summer 1994. The original name of the project was "STARE radar as a tool of space research", and the aim was to study only questions directly related to the STARE coherent scatter radar. During the project the emphasis shifted toward a more general approach: the dissertation now contains results that are possibly relevant to all coherent scatter measurement. In the course of work it also became evident that the coupling between ionosphere and magnetosphere is important when coherent scatter radar measurements are being assigned to their geophysical context. Therefore the dissertation also studies ionosphere–magnetosphere coupling using numerical simulation techniques.

I wish to thank Prof. Risto Pellinen, the head of the Department of Geophysics, and Dr. Heikki Nevanlinna, the head of the Geomagnetism Division, for providing excellent working conditions at FMI and for continuous support. Especially I want to thank Risto for originally proposing me this topic, which turned out to be a really fascinating one.

My sincere appreciation goes to my supervisor, Dr. Asko Huuskonen. His wide experience and quick understanding of new topics was essential for the successful completion of this work. He also read the manuscript exceedingly carefully and suggested significant improvements. I also want to thank Dr. Johan Silén, who supervised me before Asko. His richness of ideas is constantly helping me, and it was he who recommended using particle simulations right from the beginning, which in retrospect indeed is the most fruitful approach.

I am grateful to Dr. Risto Pirjola and Prof. Christian Sucksdorff, the former head of the Department of Geophysics, for continuous support, patience and interest during the early years of my scientific life. In the same spirit, I also wish to acknowledge the late research professor Ilkka Liede for encouragement and support.

My special thanks go to Dr. Hannu Koskinen, who has constantly given me pieces of advice and ideas on how to proceed, as well as been an inexhaustible knowledge base of space physics literature and references. He also checked the final manuscript very carefully, making a number of useful comments. I also want to thank Dr. Tuomo Nygren from the University of Oulu for his examination of the manuscript.

During spring 1991 I was privileged to visit the Cornell University at Ithaca, New York, and to meet Prof. Don Farley there. Many thanks for his hospitality and for his constructive comments on my ideas and views on E–region irregularities, which at that time were not at a very mature stage.

I am grateful to Dr. Christos Haldoupis, Dr. John Sahr, Dr. Kristian Schlegel and Dr. Mihail Uspensky for many invaluable discussions on E–region plasma waves and coherent scatter radars.

It goes without saying that several colleagues at FMI/GEO have been helpful in various respects. Here I just want to acknowledge numerous discussions with Kirsti Kauristie, Dr. Harri Laakso, Dr. Anssi Mälkki, Dr. Risto Pirjola and Dr. Ari Viljanen.

As this thesis is based on numerical simulations, it could not have been completed without using the national supercomputer at Centre of Scientific Computing (CSC) in Espoo. I am obliged to Jussi Rahola, Juha Ruokolainen, Sami Saarinen and other people at CSC for providing a reliable and user–friendly supercomputer environment. Many thanks also to Lasse Häkkinen, Marko Moilanen, Jouni Rynö and Dr. Johan Silén for keeping the computers at FMI/GEO in a good condition.

Finally I want to thank my parents for their continuos support.

Helsinki, April 1994

Pekka Janhunen

Contents

1 Introduction

The space plasma in the vicinity of the Earth, at distances less than a few ten Earth radii, is largely affected by the Earth's internal magnetic field. As a first approximation, we may regard the geometry of the terrestrial plasma environment (the magnetosphere) as a magnetic dipole distorted by the solar wind. We may use a hydrodynamic analogy when describing this: the solar wind pushes the dipole from the dayside and causes a long tail to be stretched on the nightside. Large–scale phenomena in the magnetosphere obey the so–called magnetohydrodynamic (MHD) equations, which are similar to the neutral fluid Navier–Stokes equations supplemented by Ampère's force.

On the nightside of the magnetosphere, the magnetic field is stretched. Such a field geometry necessarily has nonzero curl, which implies the existence of an electric current due to Ampere's law. The current takes the form of a cross–tail current sheet, and it is a permanent feature of the magnetotail.

Even if the solar wind density, velocity, and magnetic field were constant in time, the magnetosphere might not reach a stationary state. Moreover, the solar wind is not stationary. The flow velocity and the solar wind magnetic field (or the interplanetary magnetic field, IMF) exhibit time variations. Thus the magnetosphere is more like a candle in a weak breeze than a stationary obstacle in a constant flow. Variations in the solar wind velocity can reshape the magnetosphere. The orientation of the IMF in turn determines whether the solar wind plasma can penetrate the magnetosphere. If the cross–tail current sheet is very thin — which is the equivalent of saying that the magnetic field is very stretched — the solar wind forcing may break the tail locally or altogether. The northern and southern hemisphere field lines may reconnect and thus change the topology of the magnetotail. At the same time, the cross–tail current is disrupted. After a global reconnection, a tailward traveling bubble of plasma and its associated magnetic field (a plasmoid) is believed to form. The plasmoid magnetic field is disconnected from the dipole field of the Earth.

While the exact geometry of these "magnetoquakes" (which are related to magnetic substorms) is still not known, it is clear that such events are not without consequences inside the magnetosphere. Consider a local or total disruption of the cross–tail current. At any time, the current must be continuous. A possible direction for the disrupted current to flow is along the field lines (the so–called Birkeland or field–aligned current, FAC) toward the polar regions of the Earth. Near the Earth the plasma β parameter is small, and the magnetic field pressure greatly exceeds the plasma kinetic pressure. In the force balance equation $\mathbf{j}\times\mathbf{B} = \nabla p$, the characteristic magnitude of the left–hand–side is then much larger than the characteristic magnitude of the right–hand–side. Thus the only way substantial current can flow near the Earth is along field lines, whence $\mathbf{j}\times\mathbf{B} \approx 0$.

The field–aligned current can thus proceed toward polar regions, until it hits the atmosphere. In the atmosphere the electrons and ions collide with neutral particles, resisting electric currents: below about 80 km, the neutral collisions are frequent enough to effectively hinder any substantial currents from flowing. Above about 150 km, the currents are mainly FACs. Between these altitudes there is a transition zone where the field–perpendicular conductivities (the magnetic field is nearly vertical) reach their maximum values. It is in the 90–150 km altitude range, the so–called ionospheric E–region, where the FACs can close horizontally. The E–region is then part of a global circuit connecting the ionosphere and the magnetosphere. Both quiet–time and storm–time FACs flow in the circuit; we are here mainly interested in the storm–time currents, which have more dynamical structure and have sufficiently small scale sizes that they can be studied using local earthbound observations.

The nigthside E–region is the region connected to the magnetotail. Since the ionization produced by solar UV radiation is absent, the conductivity of the nightside polar ionosphere is mainly determined by the flux of energetic particles precipitating from the magnetosphere. The precipitating flux is related to the local field–aligned current density, and it can vary steeply in the horizontal direction.

We have just described the ionosphere–magnetosphere system as a system of currents. It is also possible to use the electric field as the primary quantity to be followed. Because the conductivity along field lines is very good, the magnetospheric electric field tends to map to the E–region, and vice versa. The mapping is expected to be nearly complete for features with horizontal scale sizes larger than about 100 km in the ionosphere. Smaller scale structures in the magnetospheric electric field are more or less filtered out before they reach the ionosphere.

It is thus clear that measuring the ionospheric electric field gives information of the whole ionosphere–magnetosphere system. Coherent scatter radars are useful and widely used tools for monitoring ionospheric mesoscale electric fields. These radars probe electron density fluctuations at a fixed wavelength, measuring their intensity and phase velocity distribution as a function of position and time. As the plasma waves are driven unstable by the electric field, information about the electric field can be obtained by measuring the properties of the waves. Unfortunately, the relationship between the plasma wave properties and the electric field is not simple, and using the linearized plasma wave theory in place of the relationship may give misleading results.

The purpose of this dissertation is to give theoretical and simulation background for ionospheric probing with coherent scatter radars, and also relate the mesoscale ionospheric measurements to ionosphere–magnetosphere coupling, again developing simulations. The thesis consists of an Introduction and four original papers:

(1) Janhunen, P., Three-dimensional stabilization mechanism for the auroral Farley–Buneman instability, *J. Atmos. Terr. Phys., 54,* 1633–1643, 1992.

(2) Janhunen, P., Perpendicular particle simulation of the E–region Farley–Buneman instability, *J. Geophys. Res.,* in press, 1994.

(3) Janhunen, P., Implications of flow angle stabilization to coherent E–region spectra, *J. Geophys. Res.,* in press, 1994.

(4) Janhunen, P. and Huuskonen, A., A numerical ionosphere–magnetosphere coupling model with variable conductivities, *J. Geophys. Res., 90,* 9519–9530, 1993.

In Paper 1, the role of height dependence of collision frequencies on auroral E–region irregularities is studied. The paper also contains necessary basic research related to 3D numerical modeling, including a 3D stationary analytic solution to the two–fluid equations in auroral geometry.

Paper 2 is about flow angle dependence of E–region irregularity spatial spectrum. Particle simulation technique is used to show that the flow angle distribution is generally nonsymmetric. This is a completely new finding, which thus far has no direct observational proof or disproof.

In Paper 3, the results of Paper 2 are cast in the form of simple fitted formulae, which allows coherent scatter spectra to be synthesized. The flow angle asymmetry is interpreted as "flow angle stabilization" of the Farley–Buneman waves. It is shown that flow angle stabilization can explain some of the previously unexplained properties of coherent scatter radar observations, in particular those related to the difference of equatorial and auroral spectra. It also provides a new interpretation of the long–standing puzzle of phase velocity saturation to the ion acoustic speed.

Paper 4 is about ionosphere–magnetosphere mesoscale coupling. It shows how precipitation–dependent ionospheric conductivities can be incorporated in mesoscale numerical coupling models. The technique is applied to auroral Ω bands, and it is shown that from the ionospheric point of view, the Kelvin–Helmholtz instability is one possible explanation for these features, although the question of the location of the relevant magnetospheric region is left open.

The Introduction is organized in two main chapters, the first one dealing with ionospheric irregularities and the second one treating ionosphere–magnetosphere coupling. The first chapter provides background for Papers 1, 2 and 3 by reviewing the various theoretical approaches to E–region irregularities. As a new material, a hybrid dispersion relation is derived. The second chapter has a more limited scope, approaching the ionosphere–magnetosphere coupling from the numerical simulation viewpoint. It also discusses novel developments that have been put forward after the writing of Paper 4.

Let us finally consider the significance of the this study from a wider scientific perspective. Studies of E–region irregularities and ionosphere–magnetosphere coupling can be motivated e.g. in the following ways:

1. Understanding E–region irregularities is necessary for the use of coherent scatter radars in monitoring the ionospheric electric fields.

2. Ionosphere–magnetosphere coupling should be understood when electric field patterns produced by coherent scatter radars are being used and interpreted.

3. Ionosphere–magnetosphere coupling is important as such. It is one of the most dynamic and versatile coupling processes occurring in space physics, and it gives a unified view to many seemingly disjoint space plasma phenomena. Any serious theory of magnetic substorms, for instance, probably must involve ionosphere–magnetosphere coupling.

4. E–region irregularities are theoretically interesting. This is evident also from the large scientific literature concerning these irregularities. The popularity can not be a coincidence. The very nonlinear nature of the Farley–Buneman waves gives an intuitive promise that mastering them will also benefit studies of other turbulent phenomena, which are often less nonlinear.

5. The E–region can also be thought of as a natural laboratory for the Farley–Buneman waves, the observation of which is not easy but possible. This may prove valuable if and when similar processes are met elsewhere in the Universe.

2 E–REGION IRREGULARITIES

2.1 INTRODUCTION

Electrostatic plasma waves propagating in the E–region are usually called irregularities, because of their turbulent nature and because they make the electrojet non–laminar. The wavelength of these plasma waves varies from the electron Debye length (a few centimeters) up to the thickness of the current carrying region (several kilometers). We are mainly interested in wavelengths around one meter.

The electron–neutral collision frequency ν_e is much smaller than the electron cyclotron frequency (Larmor frequency) $\Omega_e = eB/m_e$ everywhere in the E–region. The electron Larmor frequency Ω_e is about 9×10^6 s^{-1} in the auroral region and about 5×10^6 s^{-1} in the equatorial electrojet region. Equivalently, the electron mean free path with respect to neutral collisions is much longer than the electron Larmor radius, which is of the order of one centimeter. Thus the electrons are magnetized in the E–region. Their mean drift velocity $\mathbf{V}_{de}$ is almost equal to the $\mathbf{E}\times\mathbf{B}$ velocity $\mathbf{E}\times\mathbf{B}/B^2$.

Because the ions (mainly NO^+ and O_2^+) are about 50000 times heavier than the electrons, their Larmor frequency $\Omega_i = eB/m_i$ is correspondingly smaller than the electron Larmor frequency. In the auroral region where $B \approx 50000$ nT, $\Omega_i \approx 160$ s^{-1}. In the equatorial region, $\Omega_i \approx 90$ s^{-1}. Below about 120 km, $\Omega_i < \nu_i$, where ν_i is the ion–neutral collision frequency. Thus, in the lower E–region, which is the most interesting for us in this work, the ions are collision dominated, i.e. many ion–neutral collisions take place during one ion gyroperiod. Usually it is a good approximation to assume that the ions are unmagnetized, i.e. neglect the gyromotion altogether.

The lower E–region can then be characterized by the inequalities $\Omega_i << \nu_{i,e} << \Omega_e$. Throughout, unless otherwise stated, we perform our calculations in a coordinate frame where the neutrals are at rest.

Farley (1963) and *Buneman* (1963) showed that the plasma is unstable against a specific type of modified two–stream instability, later called the Farley–Buneman instability, if the ambient electric field $\mathbf{E}$ is strong enough. There are many factors affecting the growth rate, but roughly speaking the larger the ion–electron velocity difference $|\mathbf{V}_{de} - \mathbf{V}_{di}|$ is, the stronger the instability. At very low altitudes (less than 80 km, say) the instability is prohibited because both ions and electrons move together with the neutrals, and consequently there is no velocity difference between the charged species. However, regardless of velocity difference, the Farley–Buneman instability is unlikely everywhere below about 95 km, because of collisional damping. At high altitudes (more than 120 km), both ions and electrons $\mathbf{E}\times\mathbf{B}$ drift, and the instability is again quenched because of the lack of velocity difference. Therefore the lower E–region, about 95 to 115 km altitude range, is the primary place where Farley–Buneman unstable waves may be found in both equatorial and auroral electrojets.

2.2 Linear theory

In a mathematical treatment of plasma instabilities, and the Farley–Buneman instability in particular, it is usually assumed that the medium is initially homogeneous. There may, however, be a linear trend in the electron density in some spatial direction. We shall call such a medium laminar. If a very small sinusoidal wave perturbation is introduced to the laminar background, it is possible to compute the phase velocity and growth rate of the plane wave. The wave amplitude time dependence is $\exp(-i\omega t)$, where ω is the complex angular frequency, $\omega = \omega_r + i\gamma$, where $\omega_r = \mathrm{Re}\,\omega$, $v_{ph} = \omega_r/|\mathbf{k}|$ is the phase speed, and $\gamma = \mathrm{Im}\,\omega$ is the growth rate. The complex variable ω and the real wavenumber vector $\mathbf{k}$ are related by an equation, which is called the (linear) disperion relation. In the fluid approximation, the dispersion relation is usually polynomial, in the kinetic model it is a more complicated transcendential equation. Usually the dispersion equation will have multiple roots; it is then the ω with largest imaginary part (growth rate) which is of most interest.

This linearized approach is frequently used since it presents only modest computational difficulty and because it provides a classification of wave modes (e.g., *Sudan et al.,* 1973; *Fejer and Kelley,* 1980; *Farley,* 1985). However, as soon as the growth rate $\gamma(\mathbf{k}) > 0$ for any $\mathbf{k}$, the associated mode grows exponentially until it is saturated by nonlinear processes. After this, the assumptions of the dispersion relation formalism break down since the perturbations are no longer "small". This pitfall concerns the system as a whole, not just the individual growing mode. If, on the other hand, the external conditions are such that no mode has positive growth rate, the system turns quiescent and, from our point of view, uninteresting. The conclusion is that the linearized theory can only be used to test whether the system is quiescent or unstable, but, strictly speaking, not to derive its properties if it happens to be unstable.

Still, the linear theory phase velocity can sometimes be used as an estimate to the mean value of the temporal power spectrum. As far as coherent scatter radars are concerned, improving the phase velocity estimate is the main task of numerical simulations and the more advanced theoretical techniques.

2.2.1 Two–fluid approach

The fluid model of a plasma is formally based on the first few moments of the Boltzmann equation (e.g., *Krall and Trivelpiece,* 1973; *Ichimaru,* 1973; *Schunk,* 1977). In the two–fluid approach the electrons and the ions are treated as separate fluids. The neutral atmosphere is usually not explicitly modelled, since we are generally interested in time scales much shorter than it takes to move or heat the neutral gas significantly by plasma collisions.

Applied to the E–region the continuity and momentum equations for charged fluids are in the rest frame of neutrals (e.g., *Sudan et al.,* 1973; *Fejer et al.,* 1984b)

$$\frac{\partial n_s}{\partial t} = -\nabla \cdot (n_s \mathbf{v}_s) - L_s + Q_s \tag{2.1}$$

$$\frac{\partial \mathbf{v}_s}{\partial t} + \mathbf{v}_s \cdot \nabla \mathbf{v}_s = \frac{q_s}{m_s}(\mathbf{E} + \mathbf{v}_s \times \mathbf{B}) - \frac{k_B T_s}{m_s}\frac{\nabla n_s}{n_s} - \nu_s \mathbf{v}_s \tag{2.2}$$

where s labels the particle species, n_s is the number density, $\mathbf{v}_s$ is the flow velocity, L_s is the loss rate, Q_s is the production rate, q_s is the charge of sth species, m_s is the mass of sth species, $\mathbf{E}$ is the electric field, $\mathbf{B}$ is the (external) magnetic field, k_B is Boltzmann's constant, T_s is the temperature of the sth fluid and ν_s is the sth particle–neutral collision frequency. In our case, $s = i, e$.

In addition, we need to specify a relation linking the electrostatic electric field $\mathbf{E} = -\nabla\phi$ to the other quantities. In principle, Poisson's equation

$$-\nabla^2 \phi = \frac{1}{\epsilon_0}\sum_s q_s n_s \tag{2.3}$$

can always be used. A typical value for the electron Debye length is, however, less than one centimeter. As we are interested in wavelengths of the order of one meter, the plasma is quasineutral in a good approximation, and instead of Poisson's equation we can just use the relation $n_e = n_i = n$. It is also a good approximation to neglect the time derivative terms in the electron momentum equation, because we are dealing with frequencies much smaller than the electron cyclotron frequency. Neglecting the recombination and production terms is usually justified as well (*Fejer et al.,* 1984b). Carrying out all these approximations, we are left with the following system of equations:

$$\frac{\partial n}{\partial t} = -\nabla \cdot (n \mathbf{v}_i) \tag{2.4}$$

$$\frac{\partial \mathbf{v}_i}{\partial t} + \mathbf{v}_i \cdot \nabla \mathbf{v}_i = \frac{q_i}{m_i}(\mathbf{E} + \mathbf{v}_i \times \mathbf{B}) - \frac{k_B T_i}{m_i}\frac{\nabla n}{n} - \nu_i \mathbf{v}_i \tag{2.5}$$

$$\frac{1}{1+\psi_e^2}\left[\frac{\psi_e}{B}\nabla \cdot (n \nabla \phi') - \frac{1}{B^2}\mathbf{B} \cdot \nabla n \times \nabla \phi'\right] = \nabla \cdot (n \mathbf{v}_i). \tag{2.6}$$

We have introduced the notation $\psi_s = \nu_s/\Omega_s$ $(s = i, e)$ and

$$\phi' = \phi - \frac{1}{e}k_B T_e \log n \tag{2.7}$$

where $\mathbf{E} = -\nabla\phi$ and ϕ' is an effective potential. We have also assumed that no quantities depend on the field–aligned coordinate.

We use a coordinate system with the x axis parallel to the electric field $\mathbf{E}$, y parallel to $\mathbf{E}\times\mathbf{B}$, and z antiparallel to the geomagnetic field $\mathbf{B}$ (Fig. 2.1). The

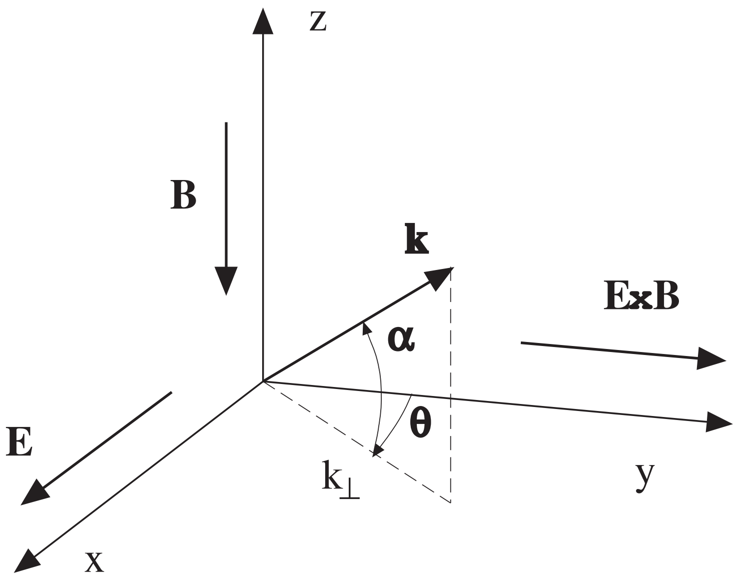

Figure 2.1. This coordinate system is employed throughout, unless otherwise stated.

angle between $\mathbf{k}$ vector and the y axis is called the flow angle θ. The angle between $\mathbf{k}$ and the $\mathbf{B}$-perpendicular plane is called the aspect angle α. The Cartesian components k_x, k_y, k_z of $\mathbf{k}$ are obtained from

$$\begin{aligned} k_x &= k_\perp \sin\theta \\ k_y &= k_\perp \cos\theta \\ k_z &= k_{||} = k \sin\alpha \\ k_\perp &= k\cos\alpha. \end{aligned} \tag{2.8}$$

The linearization of the fluid equations yields the dispersion relation

$$\omega - k_y V_d = \left(\frac{\psi}{\nu_i} - \frac{i k_y}{\mathbf{k}^2 L \Omega_i}\right)\left[\omega(i\omega - \nu_i) - i k^2 C_s^2\right], \tag{2.9}$$

where

$$\psi = \frac{\nu_e \nu_i}{\Omega_e \Omega_i}\left(1 + \frac{\Omega_e^2}{\nu_e^2}\sin^2\alpha\right) \tag{2.10}$$

(*Sudan et al.*, 1973; *Fejer et al.*, 1984b). Here, L is the zeroth order electron density gradient scale length in the x direction, i.e. $n(x) = n_0(1 + x/L)$. The ion acoustic velocity C_s is given by $C_s = \sqrt{[k_B(T_e + T_i)/m_i]}$. In order to obtain a simple tractable solution of Eq. (2.9) for $\omega = \omega_r + i\gamma$ ($\gamma = \mathrm{Im}\ \omega, \omega_r = \mathrm{Re}\ \omega$), we need to assume that L is large in some sense. Let us first obtain an expression for ω_r for marginally stable ($\gamma = 0$) waves. Taking the real part of Eq. (2.9) while assuming a real ω yields a quadratic equation for ω. If the discriminant is expanded to first order in $V_d/L\Omega_i$, we obtain for the real angular frequency

$$\omega_r = \frac{k_y}{1+\psi}\left(V_d - \frac{C_s^2}{L\Omega_i}\right). \tag{2.11}$$

A typical value for both V_d and C_s is 450 m/s. Typical Ω_i is 150 s^{-1}. Therefore, if L is larger than 500 m, we can accurately use the expression

$$\omega_r = \frac{k_y V_d}{1+\psi} = \frac{k V_d \cos\theta}{1+\psi} \tag{2.12}$$

for the real angular frequency.

Assuming $|\gamma| \ll |\omega_r|$, we also readily obtain a solution for the growth rate γ:

$$\gamma = \frac{1}{1+\psi}\left[\frac{\psi}{\nu_i}\left(\omega_r^2 - \mathbf{k}^2 C_s^2\right) + \frac{\omega_r \nu_i k_y}{\mathbf{k}^2 L\Omega_i}\right] \tag{2.13}$$

(*Fejer et al.,* 1984b). The symbol ω_r in Eq. (2.13) is to be substituted from Eq. (2.12).

It should be noted that Eq. (2.13) produces the maximum growth rate at nonzero aspect angle α, as was noted by *Wang and Tsunoda* (1975) and also by *Fejer et al.* (1984b). In Fig. (2.2) we plot the growth rate γ as a function of aspect angle. Both fluid and a more accurate hybrid dispersion relation (see below) are shown, and it is seen that the fluid result is qualitatively correct at 2 meter wavelength and 800 m/s drift. The growth rate is a symmetric function of the aspect angle.

The assumption $|\gamma| \ll |\omega_r|$ is often justified in practice. In particular, it is by definition ideally justified if we are looking for the boundary of the unstable region, $\{\mathbf{k}|\gamma(\mathbf{k}) = 0\}$. The curve $\gamma = 0$ in (k_y, k_x) plane can be solved analytically. The magnitude k of the vector $\mathbf{k}$ as a function of the flow angle θ is obtained as

$$k^2 = \frac{\left(\dfrac{1+\psi}{\psi_e}\right)\left(\dfrac{\nu_i}{V_d L}\right)}{\left[\dfrac{(1+\psi)C_s}{V_d\cos\theta}\right]^2 - 1}. \tag{2.14}$$

Eq. (2.14) is equivalent to

$$\cos^2\theta = \frac{[(1+\psi)C_s/V_d]^2}{1 + \dfrac{1+\psi}{\psi_e}\dfrac{\nu_i}{V_d k^2 L}}, \tag{2.15}$$

from which it is easily seen that in the limit $L \to \infty$ the unstable region becomes a sector (i.e., θ becomes independent of k).

The smallest drift velocity V_d for which γ vanishes when $L = \infty$, or the corresponding electric field $E = V_d B$, is called the Farley–Buneman threshold. In

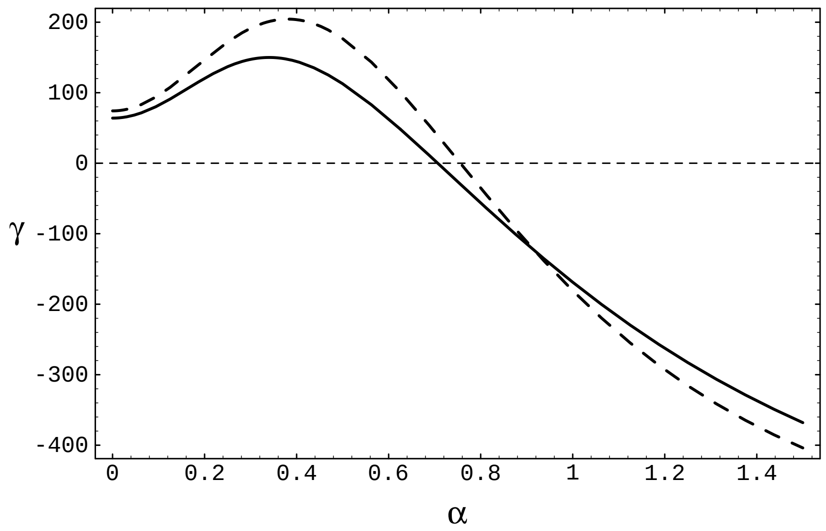

Figure 2.2. Linear growth rate γ (s^{-1}) as a function of aspect angle α (degree) for zero flow angle, 800 m/s electron drift, and 2 m wavelength. Dashed curve is the fluid result, Eq. (2.13), solid curve is a hybrid dispersion relation, Eq. (2.36). The hybrid dispersion relation is more accurate than the fluid model and it is introduced in a later subsection below.

fluid theory the threshold velocity is given by $V_d^{thr} = (1+\psi)C_s$. If $V_d < (1+\psi)C_s$, all waves are damped, unless there is a density gradient with $\mathbf{E} \cdot \nabla n > 0$. In the last case the unstable region is a "droplet" concentrated on large–scale (several tens of meters) waves (Fig. 2.3, topright panel).

When V_d exceeds the threshold, the droplet quickly "opens up" to an instability sector. Within the sector, the growth rate increases without limit as k increases. This is not, of course, physically correct, but is an artifact of the fluid theory. Fig. (2.3) shows the instability region for three different V_d and L values. In the first column L is negative and thus the density gradient is a stabilizing one ("unfavourable"). The instability region is completely absent except in the bottom plot. The middle column displays the $L = \infty$ case, in which case the instability regions are sectors. The effect of a destabilizing L ("favourable") is seen in the rightmost column, where the region transforms from a droplet shape to a distorted sector as V_d grows.

Although the linear theory growth rate should not be taken literally whenever it is positive, its magnitude nevertheless gives some indication of how the nonlinear system will tend to behave. If the growth rate is large in some region of $\mathbf{k}$ space, the corresponding modes (Fourier components in $\mathbf{k}$ space) will probably be present in the nonlinearly saturated spectrum, too. Fig. (2.4) shows how the droplet opens up to a sector and how the growth rate soon becomes large inside the sector.

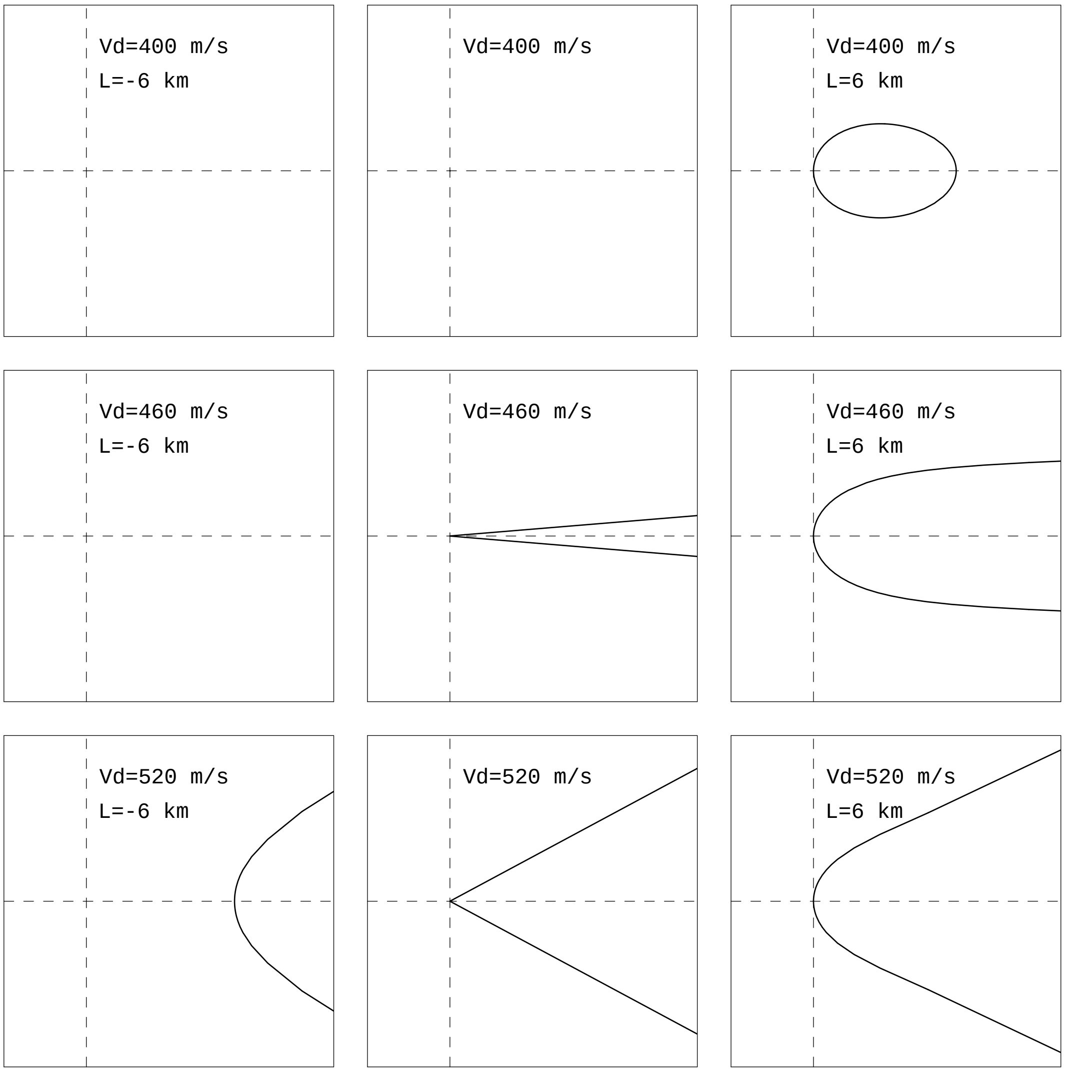

Figure 2.3. Unstable region in $\mathbf{k}$ plane for different combinations of V_d and L. The vertical axis is parallel to the electric field, and the horizontal axis is parallel to the $\mathbf{E}\times\mathbf{B}$ direction. In the middle column the value of L is ∞.

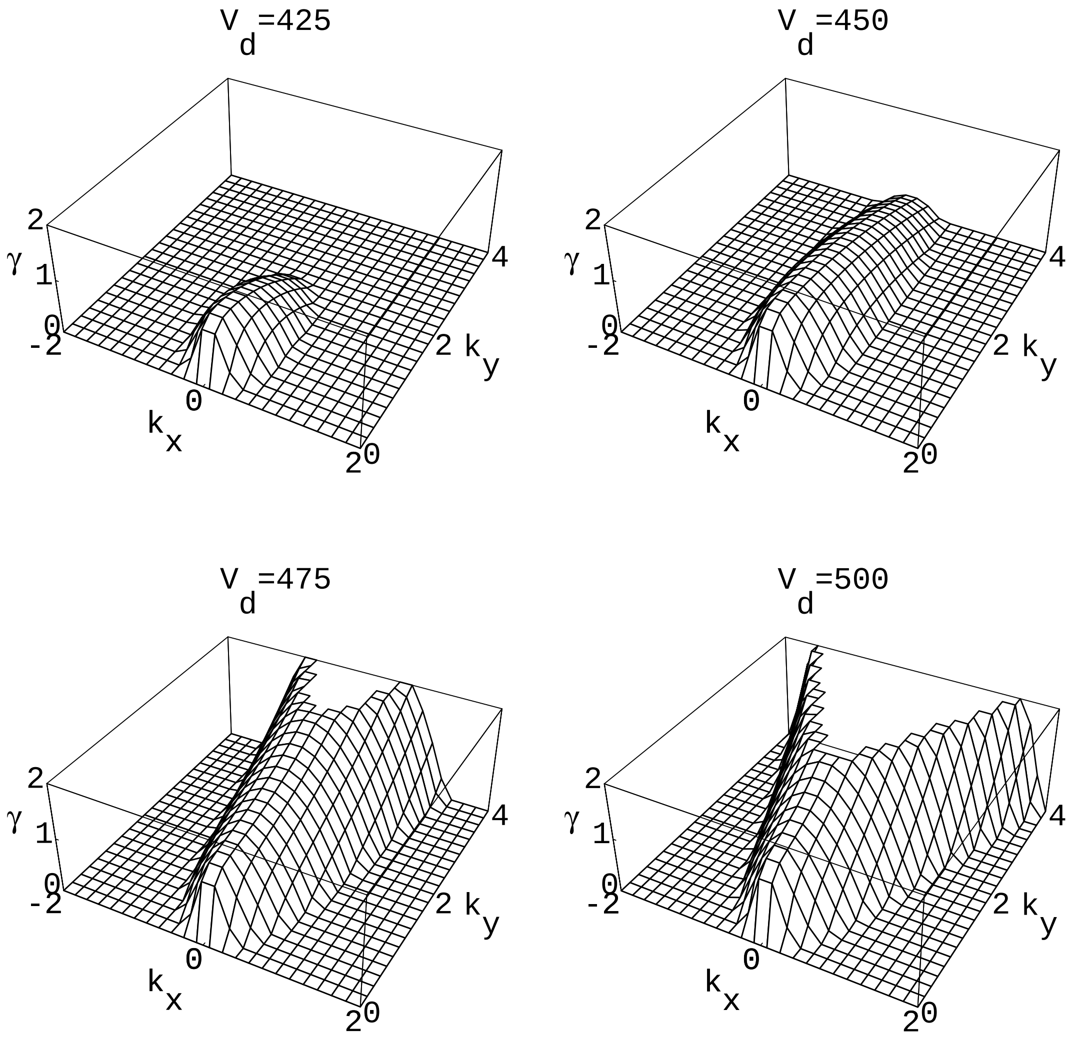

Figure 2.4. Fluid theory growth rate in **k** plane as a series of 3D plots for four drift values ranging from 425 m/s to 500 m/s. Negative growth rates have been replaced with zero. The gradient scale length L is 6 km, and other variables have their typical E–region values.

Considering that the electric field changes only 15 % in Fig. (2.4), the name "threshold" is by no means misleading.

2.2.2 *Kinetic theory approach*

The starting point for kinetic calculations is the Boltzmann equation for both species $s = i, e$:

$$\frac{\partial f_s}{\partial t} + \mathbf{v}_s \cdot \nabla f_s + \frac{q_s}{m_s}(\mathbf{E} + \mathbf{v}\times\mathbf{B}) \cdot \nabla_v f_s = \left(\frac{\partial f_s}{\partial t}\right)_c . \tag{2.16}$$

The linearization of Eq. (2.16) requires the specification of the unperturbed distribution functions for electrons and ions. Furthermore, it requires a simple tractable form for the neutral collision term $(\partial f_s/\partial t)_c$, otherwise the problem seems too difficult to handle even by numerical means.

The original kinetic theory by *Farley* (1963) already contained the most important ingredients. It employed the Bhatnagar–Gross–Krook (BGK) collision term:

$$\left(\frac{\partial f_s}{\partial t}\right)_c = -\nu_s \left[f_{s1}(\mathbf{r}, \mathbf{v}, t) - \frac{N_{s1}(\mathbf{r}, t)}{N_{s0}} f_{s0}(\mathbf{v}, t) \right] \tag{2.17}$$

(*Bhatnagar et al.,* 1954), and considered only fully perpendicular waves without ambient electron density gradient. Here,

$$N_{s0,1}(\mathbf{r}, t) = \int d^3\mathbf{v} f_{s0,1}(\mathbf{r}, \mathbf{v}, t). \tag{2.18}$$

New formulations and extensions have thereafter been constructed. *Schmidt and Gary* (1973) showed that the electrons can practically always be handled as a fluid, but the ions must be treated kinetically unless the wavelength is much larger than the ion mean free path. More specifically, the electrons are fluidlike for wavelengths much larger than the electron Larmor radius and Debye length, which are both generally less than one centimeter. For the ions, on the other hand, the Landau damping is important for wavelengths less than about half a meter.

Parallel propagation effects were studied by *Ossakow et al.* (1975). They found that the fastest growing modes have small but finite $k_{\|}$. Similar result also follows from fluid theory (*Wang and Tsunoda,* 1975; *Fejer et al.,* 1984b). The case of nonequal electron and ion temperatures was elaborated by *Schlegel and StMaurice* (1982) and *Schlegel* (1983).

The collision term (Eq. 2.17) is the weakest point of the kinetic theory. It is a simple relaxation model, which has been corrected to conserve particle density locally. It drives the whole distribution function toward an equilibrium in a common time scale. Therefore, the mean velocity and temperature, for instance, relax in the same time scale. This is qualitatively correct for the ions since the ion and neutral masses are comparable, but not for the electrons, for which the thermalization time scale should be much longer (by factor of m_n/m_e, if only elastic collision are considered) than the momentum transfer time scale. In effect, the BGK collision model forces the fluctuations to be isothermal (*Bhatnagar et al.,* 1954; *Farley,* 1963). In reality, inelastic electron-neutral collisions are much more important than elastic ones, and thus the error with BGK is not as large as m_n/m_e, but it still remains fairly large.

Stubbe (1989) presented an improved collision term, which involved separate time scales for momentum transfer $(1/\nu_s^{(m)})$, isotropization $(1/\nu_s^{(r)})$, and energy transfer $(1/\nu_s^{(e)})$, $s = i, e$. As a relatively good approximation, $\nu_s^{(r)} = \nu_s^{(m)}$, $s = i, e$, and $\nu_i^{(e)} = \nu_i^{(m)}$, that is, ions have only one time scale, but the electron energization time differs from the momentum transfer and isotropization time. Unfortunately, Stubbe's full expressions are complicated to such an extent that, to our knowledge, no one has thus far solved the resulting dispersion relation numerically.

In summary, a satisfactory and tractable kinetic theory for partially ionized gases is still an open problem. In the ionospheric E–region this implies that one is forced to use approximate fluid, kinetic, or hybrid models of the plasma, and one must therefore evaluate the validity of the assumptions separately in each case.

2.2.3 *Hybrid dispersion relation*

In this subsection we derive and present a hybrid dispersion relation, where the electron gas is treated as a fluid in the 5–moment approximation, while the ions are kinetic and unmagnetized. In the 5–moment approximation the continuity and momentum equations are accompanied by an energy equation with isotropic pressure. The resulting dispersion relation is suitable for numerical computations, and various analytical limits are easily derived.

Using the fluid model for the electrons is a good approximation in almost all cases (*Schmidt and Gary,* 1973; *Stubbe,* 1989; *Farley and Providakes,* 1989). Only for large aspect angles there is a real need for full kinetic treatment (*Stubbe,* 1989). Using the 5–moment approximation, we retain the temperature fluctuations, and thus relax the most important limitation of the standard isothermal theory.

The electron fluid equations are (*Schunk,* 1977):

$$\frac{\partial n_e}{\partial t} + \nabla \cdot (n_e \mathbf{v}_e) = 0 \tag{2.19}$$

$$n_e m_e \left(\frac{\partial \mathbf{v}_e}{\partial t} + \mathbf{v}_e \cdot \nabla \mathbf{v}_e \right) + \nabla p_e - n_e q_e (\mathbf{E} + \mathbf{v}_e \times \mathbf{B}) = -n_e m_e \Phi \nu_{en}^{(m)} \mathbf{v}_e \tag{2.20}$$

$$\left(\frac{\partial}{\partial t} + \mathbf{v}_e \cdot \nabla \right) p_e + \frac{5}{3} p_e \nabla \cdot \mathbf{v}_e = \frac{n_e m_e \nu_{en}^{(m)}}{m_e + m_n} \left[2k_B (T_n - T_e) \Psi + \frac{2}{3} m_n \mathbf{v}_e^2 \Phi \right]. \tag{2.21}$$

Here Φ and Ψ are correction factors (*Schunk,* 1977) with basic values of unity. By taking the limit $\Psi \to \infty$, we recover the usual isothermal results. By setting

Ψ larger than unity but finite, we can model electron–neutral inelastic collisions. *Stubbe* (1989) effectively used $\Psi \approx (10000K)/T_{e0}$. Inelastic collisions are ignored if Ψ is set to unity. Notice that Ψ and Φ have nothing to do with the ψ parameter (2.10) or electric potential ϕ.

After linearizing Eqs. (2.19) to (2.21), we eliminate the pressure and velocity perturbations, starting from the higher moment equations and proceeding toward the continuity equation (2.19). As a result, the electron density perturbation n_1^e is expressed as linearly proportional to the assumed electric potential perturbation ϕ_1. We write this relationship in the form

$$\frac{n_1^e}{n_0} = -\frac{q_e}{k_B T_{0e}} F_e \phi_1 \tag{2.22}$$

where F_e is a dimensionless function having the form

$$F_e = i \frac{\Delta_e}{\tilde{\omega}_e + i \dfrac{\Delta_e}{1 - i \frac{\tilde{\omega}_e}{2\Psi \nu_{en}^{(m)}} (\frac{m_n}{m_e} + 1)}} \tag{2.23}$$

where

$$\Delta_e = \left(\mathbf{k} - i\frac{\nabla n_0}{n_0}\right) \cdot \left[D_e^{\perp}(\mathbf{k} + \frac{1}{\tilde{\nu}}\mathbf{k}\times\mathbf{\Omega}_e) + D_e^{\parallel} k_{\parallel} \hat{\mathbf{\Omega}}_e\right] \tag{2.24}$$

$$D_e^{\perp} = \frac{k_B T_{0e} \tilde{\nu}}{m_e(\Omega_e^2 + \delta \cdot \mathbf{k} + \tilde{\nu}^2)} \tag{2.25}$$

$$D_e^{\parallel} = \frac{k_B T_{0e}}{m_e \tilde{\nu}} \left(\frac{\Omega_e^2}{\Omega_e^2 + \delta \cdot \mathbf{k} + \tilde{\nu}^2}\right) \tag{2.26}$$

$$\delta = \tilde{\nu}\mathbf{w} + \mathbf{\Omega}_e \times \mathbf{w} + \frac{1}{\tilde{\nu}}(\mathbf{w} \cdot \mathbf{\Omega}_e)\mathbf{\Omega}_e \tag{2.27}$$

$$\mathbf{w} = \frac{\dfrac{5}{3} i \dfrac{k_B T_{0e}}{m_e}\mathbf{k} + \dfrac{\nabla p_{0e}}{n_0 m_e} - \dfrac{4}{3}\dfrac{m_n}{m_n + m_e}\Phi \nu_{en}^{(m)} \mathbf{v}_{e0}}{\tilde{\omega}_e + 2i \dfrac{m_e}{m_n + m_e} \Psi \nu_{en}^{(m)}} \tag{2.28}$$

$$\tilde{\nu} = \Phi \nu_{en}^{(m)} - i\tilde{\omega}_e \tag{2.29}$$

$$\tilde{\omega}_e = \omega_e - \mathbf{k} \cdot \mathbf{v}_{0e}. \tag{2.30}$$

The notation is similar to *Robinson* (1992) wherever feasible. The quantities $D_e^{\perp}$ and $D_e^{\parallel}$ can be interpreted as electron diffusion coefficients in the perpendicular and parallel directions, respectively (*Robinson,* 1992). No approximations, other than linearity, have been made in the above equations.

The usual isothermal results in the absence of density gradients can be recovered by taking the limit $\Psi \to \infty$, ignoring electron inertia, setting Φ to unity, and assuming $\nabla n_0 = 0 = \nabla p_{0e}$. Then $\tilde{\nu} = \nu_{en}^{(m)}$, $\mathbf{w} \to 0$, $\delta \to 0$, and the result reduces to that of *Robinson* (1992). (Notice that Robinson's equations (3) and (5) contain an erroneous factor of imaginary unit i.)

For the ions, we write the density–potential relationship in the form

$$\frac{n_1^i}{n_0} = -\frac{q_i}{k_B T_{0i}} F_i \phi_1, \tag{2.31}$$

in analogy with electrons (cf. Eq. (2.22)). The expression for F_i in the unmagnetized BGK approximation becomes (e.g., *Ossakow et al.*, 1975):

$$F_i = \frac{1 + zZ(z)}{1 + \frac{i\nu_{in}}{k}\sqrt{a}Z(z)}, \tag{2.32}$$

where

$$z = \frac{(\omega - \mathbf{k}\cdot\mathbf{v}_{0i} + i\nu_{in})\sqrt{a}}{k}, \tag{2.33}$$

$$a = \frac{m_i}{2k_B T_{0i}}. \tag{2.34}$$

and $Z(z)$ is the plasma dispersion function for any complex z,

$$Z(z) = \frac{1}{\sqrt{\pi}} \int_{-\infty+i\gamma}^{\infty+i\gamma} du \frac{e^{-u^2}}{u - z} \qquad (\gamma < \mathrm{Im}\ z) \tag{2.35}$$

(*Fried and Conte,* 1961).

The dispersion relation is then obtained from Gauss' law as

$$F_e + \frac{T_{0e}}{T_{0i}} Z_i^2 F_i + (k\lambda_{De})^2 = 0 \tag{2.36}$$

where

$$\lambda_{De} = \sqrt{\frac{\epsilon_0 k_B T_{0e}}{n_0 e^2}} \tag{2.37}$$

is the electron Debye length, and Z_i is the ion charge number, usually $Z_i = 1$. It is usually the case that our wavelength is much larger than the Debye length, we may then drop the Debye length term from the dispersion equation.

In summary: our hybrid dispersion relation given by (2.36), (2.23) and (2.32) is valid when $\nu_{in} \gg \Omega_i$, and $\lambda \gg \max(\lambda_{De}, a_e)$ where a_e is the electron Larmor radius. These conditions are usually met for radar wavelengths in the lower and middle E–region. In the higher E–region the ion magnetization should be included, but this would yield to more complicated expressions which are also much less convenient to handle numerically.

2.3 Phenomenological characteristics

2.3.1 Radar observations

There are excellent reviews of the observed properties of the E–region waves, both in equatorial (*Farley,* 1985; *Fejer and Kelley,* 1980) and auroral (*Haldoupis,* 1989) electrojets. Therefore we just briefly mention those properties which are in our view the most important.

In the equatorial electrojet, the effect of the electron density gradient obeys the general picture obtainable from linear fluid theory. A gradient with $\mathbf{E}\cdot\nabla n > 0$ produces broad spectra (called type 2) even if the electric field is below the threshold. On the other hand, a gradient with $\mathbf{E}\cdot\nabla n < 0$ can shut off the instability. The intensity of type 2 spectra at 50 MHz is roughly proportional to V_d^2, this is in agreement with fluid simulations (*McDonald et al.,* 1974; *Ferch and Sudan,* 1977; *Keskinen et al.,* 1979).

The existence of a threshold is also clearly seen in the data. Whenever the drift is large enough, narrow and intense spectra (called type 1) suddenly appear. This corresponds to the "opening up" of the instability droplet (Fig. 2.3). Type 1 spectra can usually be seen also outside the linearly unstable sector, even for flow angle 90°, as soon as the drift exceeds the threshold.

The most peculiar experimental finding in the equatorial spectra is that the phase velocity of type 1 waves appears to be limited by the ion acoustic velocity C_s (e.g., *Farley and Balsley,* 1973). This is in contrast with linear theory which states that the phase velocity should be proportional to the electric field also above the threshold. This phenomenon is called C_s saturation in this work.

In the auroral region, the spectra are more variable, and the C_s saturation is more obscure, although it seems that peaks close to C_s appear more frequently than they should on purely statistical grounds (e.g., *Haldoupis et al.,* 1993). Despite large amounts of auroral coherent scatter data, surprisingly little is known about the geophysical contexts of the spectra.

Fig. (2.5) is a sketch of the spectral types 1, 2, 3 and 4. Type 1 waves are generally believed to be the 'normal' Farley–Buneman waves in both equatorial and auroral regions. Ion cyclotron waves were first considered as a possible explanation for type 3 spectra (*Fejer et al.,* 1984a), but it has later been shown by interferometric measurements that this explanation is incorrect (*Sahr,* 1990; *Sahr et al.,* 1991). Thus the type 3 problem is open again (e.g., *Haldoupis et al.,* 1992). The origin of type 4 spectra is also unknown; elevated electron temperature is one possible mechanism (*Fejer et al.,* 1986).

Electron temperature measurements of the E–region using incoherent scatter radars led to the discovery of anomalous heating (*Schlegel and St.-Maurice,* 1981). Theories trying to explain the observed high temperatures were thereafter built (e.g., *St.-Maurice et al.,* 1981; *StMaurice and Laher,* 1985), based on the idea that the Farley–Buneman waves could nonlinearly heat the electrons. It has, however,

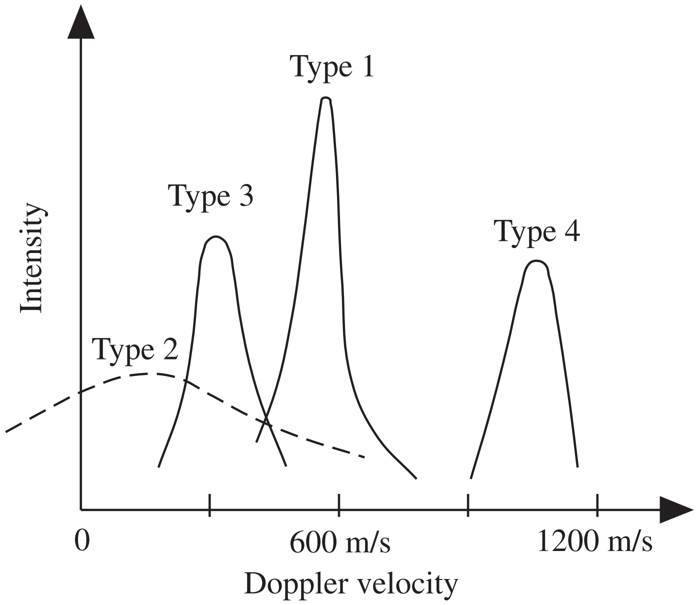

Figure 2.5. Schematic representation of the radar spectral types 1, 2, 3 and 4. Type 1 is the most common, and its phase velocity is close to C_s. Type 4 phase velocity is about twice as large as for type 1. Type 3 phase velocity is close to 250 m/s, or half the type 1 Doppler shift. Type 2 spectra have small average Doppler velocity and they are wide. (Adapted from Fig. 7 of *Haldoupis et al.* (1993)).

also been suggested that the classical frictional heating might be enough and no anomalous heating mechanism is required to explain the observations (*D'Angelo and Merlino,* 1987).

2.3.2 Rocket observations

Sounding rockets provide for *in situ* measurements of the E–region ionosphere. Although rocket campaigns are expensive and the data produced are spatially and temporarily limited, they are essential in confirming and calibrating radar measurements and theories.

The early studies (e.g., *Kelley and Mozer,* 1973; *Pfaff et al.,* 1984) confirmed that the Farley–Buneman waves exist and that their basic properties such as altitude distribution, electron density turbulence level and phase velocity are roughly in agreement with theoretical estimates. The most frequently quoted turbulence levels ($\Delta n/n$) have been about 2–3 %, and less than 10 % in almost all cases.

Some of the later measurements have been able to dwell into the more difficult questions of phase velocities and angular distributions, which are the key to testing the various theories. A work by *Pfaff et al.* (1987) indicated that the phase velocity does not saturate to C_s but rather is proportional to the electron drift as the linear theory suggests. We shall return to this interesting finding later in this Introduction. In some studies (*Schlegel,* 1992), a measurement of the flow angle distribution of the intensity spectrum has been attempted.

2.3.3 Laboratory investigations

Farley–Buneman type instabilities have been studied in laboratory conditions using cylindrical plasma columns with radial driving electrostatic fields (e.g., *D'Angelo et al.,* 1974). The plasma $\mathbf{E}\times\mathbf{B}$ rotates in the cylinder. The most severe limitation of these studies is their non–cartesian geometry.

The early experiments (*D'Angelo et al.,* 1974) confirmed the basic linear theory characteristics: the existence of a threshold and the overall $\mathbf{E}\times\mathbf{B}$ directed motion of plasma structures. *Kustom et al.* (1985) found a steeply falling $k^{-3.5}$ spatial power spectrum and concluded that there is far too little power at short wavelengths (10–30 cm in the ionosphere) to provide any anomalous wave heating. *John and Saxena* (1975) found the spectral exponent to be in the range 2..3.5, and the phase velocity to be between C_s and the linear theory prediction.

Despite the limited accuracy and scope of laboratory investigations, they nevertheless provide an independent approach to the Farley–Buneman problem. However, nowadays and in the future numerical simulations are likely to supersede laboratory experiments, since the former do not suffer from the cylindrical geometry restriction.

2.4 SIMULATIONS

2.4.1 Fluid simulations

A numerical solution of the two–fluid model, Eqs. (2.4) through (2.6), using finite–difference techniques does not seem too difficult at first sight. The pioneering works (*McDonald et al.,* 1974; *Ferch and Sudan,* 1977; *Keskinen et al.,* 1979) indeed solved the full set of equations and obtained results that supported the nonlinear cascade process that *Sudan et al.* (1973) proposed to explain type 2 waves. The simulations gave rise to theoretical developments (*Sudan and Keskinen,* 1977), which completed the current understanding of gradient drift turbulence and type 2 spectra.

An example of a two–fluid model simulation and the associated cascade process is shown in Fig. (2.6) and Fig. (2.7). The system was initialized with a sinusoidal perturbation of 3 % amplitude, the electric field was 30 % of the Farley–Buneman threshold, and a density gradient was arranged such that the sinusoidal wave was weakly unstable according to Eq. (2.13). Fig. (2.6) displays the electron density at time t_1, which corresponds to five wave travel times through the simulation box, which is equipped with periodic boundary conditions. The electric

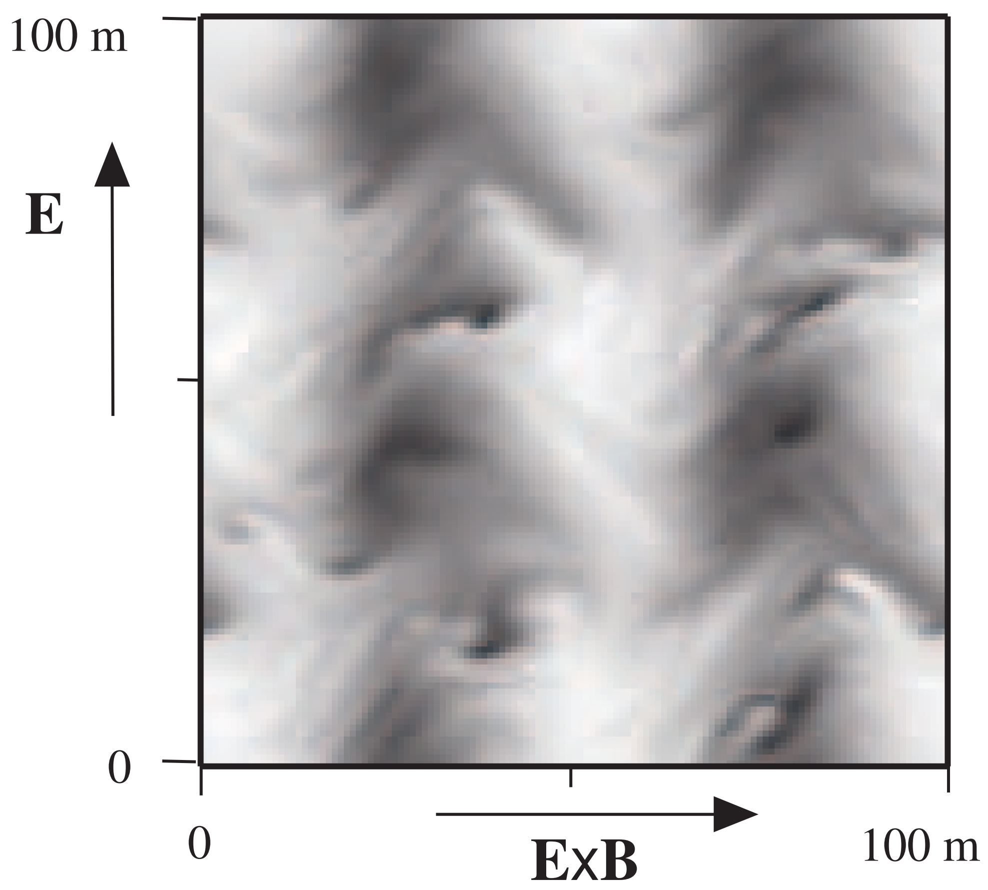

Figure 2.6. 2D fluid simulation electron density at time t_1 (see text).

field is upward and the waves travel from left to right. The simulation box size is 100 m $\times$ 100 m. At time t_1, the initially sinusoidal perturbation has started to erode, and fingerlike structures characteristic to gradient drift instability have started to develop. Fig. (2.7) is at time t_2, which is twice time t_1, that is, it corresponds to ten wave travel times across the box, or about 10 seconds physical time. (It is not reasonable to define the physical quantities more accurately since the scaling from code units to physical units depends on C_s and the ion Larmor radius, for instance.) At t_2, the turbulence is already almost fully developed and nearly isotropic. Almost all traces of the initial sinusoidal wave have disappeared.

The success of fluid simulations in explaining type 2 spectra awakened optimism: maybe these simulations can be run for drift velocities approaching and exceeding the Farley–Buneman threshold. *McDonald et al.* (1974) wrote: "We have retained ion inertia ... so that we can later investigate the transition from type 2 instabilities to type 1 as the electron drift velocity is increased in stages until it exceeds the acoustic speed. The results of this investigation will be described in detail elsewhere." *Fejer and Kelley* (1980) were even more optimistic:

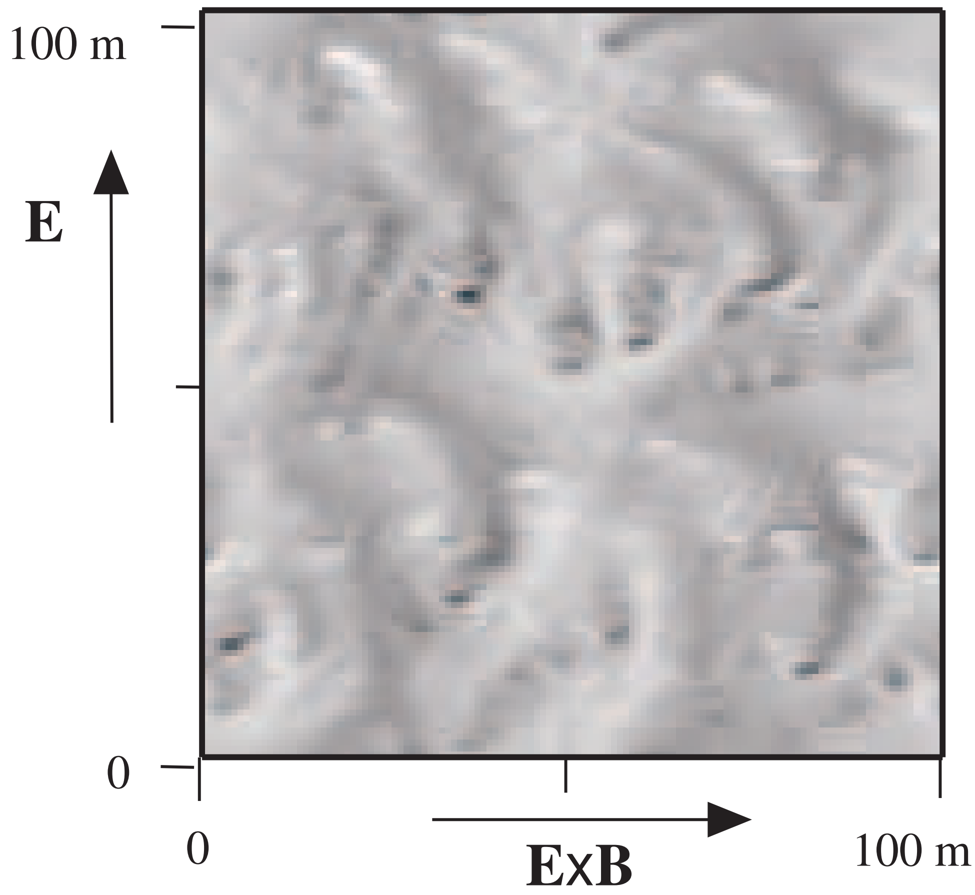

Figure 2.7. 2D fluid simulation electron density at time t_2 (see text).

"The inclusion of the ion inertia, and the use of large drift velocities, may explain eventually the physical processes associated with both type 1 and type 2 irregularities." If the ion inertia is neglected the term proportional to ω_r^2 disappears in Eq. (2.13), and thus the Farley–Buneman instability is suppressed.

Unfortunately, the initial optimism was not justified, and the problem of Farley–Buneman waves is far more difficult than then realized. If one tries to run the fluid codes with electric fields approaching or exceeding the threshold, numerical divergence soon results. The reason for this bad behaviour is that the linear theory growth rate increases without limit as the wavelength is decreased in the unstable sector (Fig. 2.3), and thus the most unstable mode will be the shortest wavelength that the spatial grid can realize. It is hopeless to try to design a stable numerical method for a problem which is so unstable by its very nature.

One might hope to cure this problem by introducing "cutoff" terms in the fluid equations, which damp the smallest scale waves. This has been attempted (*Janhunen,* 1992–1993, unpublished; *Sahr,* 1993, private communication), but it has not been too successful. The cutoffs required are quite severe, and even then the code may occassionally and unpredictably explode. Even if the time development before the onset of numerical instability can be trusted, simulation codes should be more tolerant to be really useful.

The most difficult part in fluid simulations is solving the elliptic equation (Eq. 2.6). The initial investigations (e.g., *McDonald et al.,* 1974) used alternating direction implicit method (ADI, *Press et al.,* 1992), but it is possible that more robust codes would result if the strongly implicit procedure (SIP; *Jesshope,* 1979) was used. For 3D problems, successive overrelaxation (SOR) has been successfully employed (Paper 1). However, even the best elliptic solver can not totally remove the convergence problems intrinsic to the fluid approximation Farley–Buneman waves.

The auroral plasma irregularities were studied in Paper 1 using 3D fluid simulations. The investigation showed that the 3D code is numerically stable for larger electric fields than 2D, but the Farley–Buneman threshold could still not be exceeded. In the auroral region, the collision frequencies and thus ψ (Eq. 2.10) depend on altitude with about 6 km scale height. Therefore, the horizontal and perpendicular phase velocity $v_{ph} \approx V_d/(1+\psi)$ (assuming vertical magnetic field) will also depend on altitude. A typical parallel/perpendicular wavelength ratio is of the order of $\sqrt{m_i/m_e} \approx 240$ (this makes ψ about unity), therefore the parallel component of a wave whose perpendicular wavelength is 10 m will be about 2 km, which is comparable to the scale height 6 km. This will have two effects. 1) Because of altitude dependency of the phase velocity, "wavetower tilting" results, which leads to nonsymmetric nonzero aspect angle propagation, which increases ψ and causes damping. 2) Because there is coupling to lower altitudes where ψ is larger by more frequent collisions, the wave is more damped than it would be in a homogeneous medium. These were studied using a 3D simulation, however, the results were not completely satisfactory because even the 3D code was not unconditionally stable. The wavetower tilting applies only to auroral latitudes, not to the equatorial electrojet.

2.4.2 Particle simulations

To get rid of the limitations and numerical instabilities associated with fluid simulations, particle simulations can be used.

In electrostatic particle simulations, the motion of a large (up to a million, say) number of electrons and ions is solved under the self–consistent electrostatic field (e.g., *Birdsall and Langdon,* 1985).

Traditionally, particle simulation techniques have been used in fully ionized plasmas, while applications to weakly ionized systems have remained rare. Therefore, we shall devote a bit more elaborate discussion on the benefits and difficulties associated with particle simulations in E–region conditions.

In the ionosphere, the neutral collisions are a fundamental part of the system. In Boltzmann equation formalism, they are described by the collision term. In a particle simulation there are more options for including them: we could add "drag" terms to the equations of motion for each particle, or we could use a Monte Carlo type method to randomly collide the particles. We have used the latter type of neutral collisions, which makes it possible and easy to also implement inelastic collisions.

In a fully ionized electrostatic plasma, the sum of plasma kinetic and potential energies is conserved. The particle simulations are initial value problems, where the distribution functions are specified at $t = 0$ and computed at later times. New energy is not usually fed into the system during the run. The results very much depend on the initial conditions, and therefore the interpretation requires great care.

When we add the neutral collisions, the plasma energy is no longer constant. Instead, both ion and electron distribution functions approach thermal equilibrium with the neutrals and themselves, if there is no fresh energy input. The final state is independent of the initial conditions, if the run is sufficiently long. If an external electric field is imposed, energy starts flowing from the electric field through the plasma component to the neutrals. The external field drives unstable plasma waves, which heat the particles. The particles are also heated directly by the external field (frictional heating). The heated plasma particles are cooled by the neutral collisions, until equilibrium is reached. The neutrals also get heated, but in a time scale very much longer than we are interested in, so they can be regarded as an infinite energy sink with fixed temperature. The final state does not depend on the initial conditions if the run is long enough, but it is a function of the imposed electric field only (and, of course, collision frequencies and other such parameters). This is the main difference between fully ionized and weakly ionized plasma simulations, and it is a great benefit for E–region simulators.

The most pressing difficulty with E–region particle simulations is the smallness of the Debye length, as compared to the wavelengths of interest. The Debye length is usually less than one centimeter, while relevant wavelengths are one meter or larger. Therefore, very large number of particles should be used.

The first particle simulations were performed by *Machida and Goertz* (1988). These were 2D simulations with relatively small number of particles in the yz plane ($x \parallel \mathbf{E}$, $y \parallel \mathbf{E}\times\mathbf{B}$, $z \parallel \mathbf{B}$). Because of the geometry, only aspect angle dependencies could be studied. In the early stage of the run, the time development was in qualitative agreement with BGK kinetic theory (the fastest growing modes had nonzero aspect angle), but later the largest amplitude waves had more and more perpendicular $\mathbf{k}$ vectors.

Paper 2 presented fully perpendicular (xy) simulations with much larger number of particles. In this geometry, only flow angle dependencies can be studied. The results showed that the phase velocity was proportional to the electric field, i.e., it followed the linear theory prediction quite well. No C_s saturation was seen. Also, no anomalous wave heating was found, but the heating present was due to the classical frictional heating only. The spatial power spectrum in (k_x, k_y) plane was not symmetric as a function of the flow angle, but the peak was on $\mathbf{k}$ vectors satisfying $\mathbf{k} \cdot \mathbf{E} < 0$. The phase velocity of the most intense wave was close to C_s for above–threshold electric fields, although in any given direction, the phase velocity obeyed the fluid linear theory prediction.

The results of Paper 2 were surprising, and, consequently, several consistency checks (listed in the paper) were performed to ensure that the results were of physical and not of numerical origin. The flow angle asymmetry seemed strange, since both isothermal fluid and kinetic linear theories are manifestly symmetric with respect to the flow angle. The final test was the development of a 3D particle simulation, from which both perpendicular–2D and parallel–2D (*Machida and Goertz,* 1988; *Schlegel,* 1993, private communication) results could be reproduced.

Fig. (2.8) shows an example of the 3D particle simulation results. Plotted is a time series of the electron density at intervals of 1000 ω_{pe}^{-1} in two orthogonal planes. In panel *a* we see the random initial state. After 1000 plasma periods (panel *b*), small wavelength oblique waves in the yz plane have started to develop (lower panel). Meanwhile, in the xy plane (upper panel), an $\mathbf{E}\times\mathbf{B}$ propagating mode fundamental to the simulation box has developed. Thus far the time evolution has more or less obeyed the linear theory, and the fundamental mode in the xy plane has supersonic phase velocity. In panel *c* at $t = 2000\omega_{pe}^{-1}$ we see that the oblique waves in the yz plane have inversely cascaded toward longer wavelengths, and in the xy plane the most intense wave is now oblique. This kind of asymmetry in the flow angle is characteristic to these simulations, and the phase velocity of the oblique mode is no longer supersonic but is close to C_s (transonic). Panel *d* at $t = 3000\omega_{pe}^{-1}$ is very similar to panel *c*, so the system appears to have reached a quasi–steady state.

2.5 Nonlinear models

As pointed out above, strictly speaking the linear theory looses its meaning whenever there is an unstable region of $\mathbf{k}$ space, because then the linearly unstable modes will grow to the nonlinear regime and start interacting with other modes. Physically we expect the system to reach a quasi–steady nonlinearly saturated state, where the time or ensemble averaged nonlinear growth rate of all modes is zero or negative (e.g., *Primdahl and Bahnsen,* 1985). The requirement of a non–positive growth rate can be regarded as a "definition" of quasi–steadiness. The term "nonlinear growth rate" here must be understood phenomenologically:

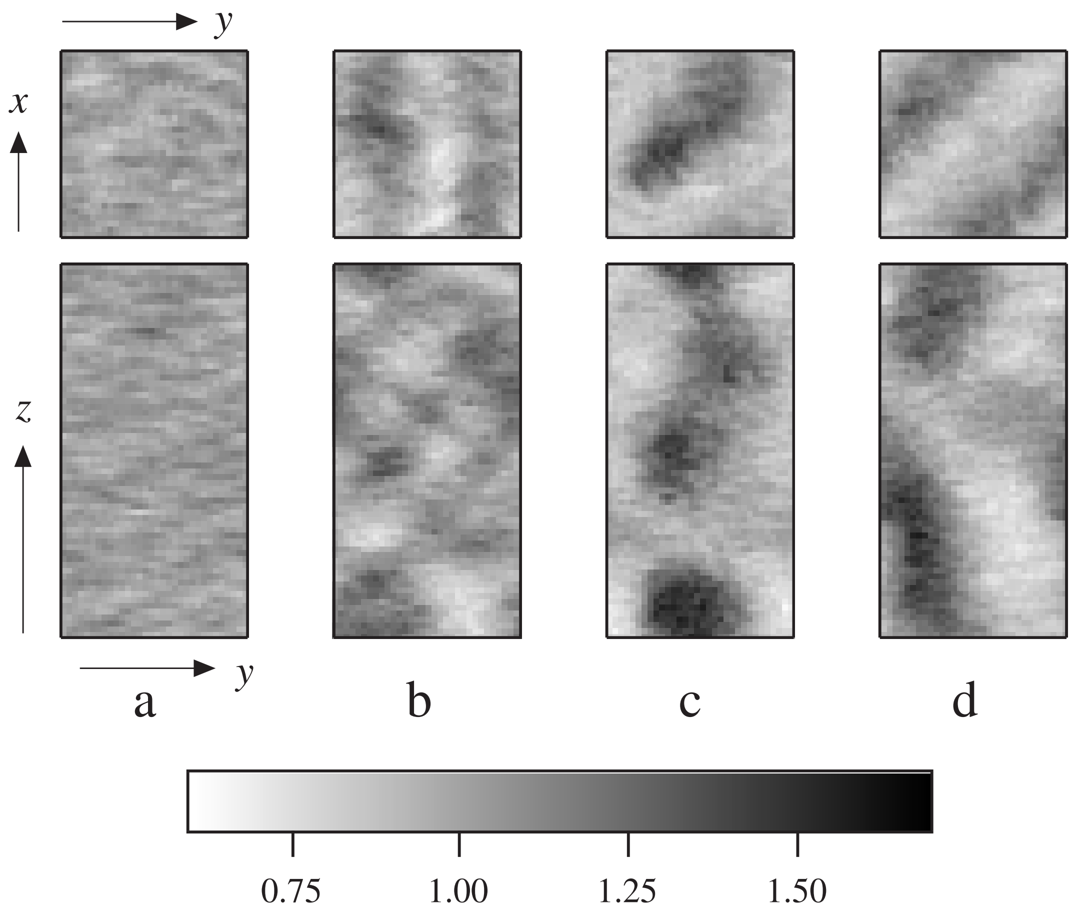

Figure 2.8. 3D particle simulation electron density. Columnar panels a, b, c and d display the electron density in the xy (top) and yz planes (bottom) at times 0, 1000, 2000 and 3000 electron plasma periods, respectively. In the top panels, the x coordinate (the electric field direction) is upwards and the y coordinate (the $\mathbf{E}\times\mathbf{B}$ direction) points from left to right. In the bottom elongated panels, the y coordinate again is directed from left to right and the z coordinate is upwards. The magnetic field is in the negative z direction, or downwards in the bottom panels. Since the $\mathbf{E}\times\mathbf{B}$ direction is to the right, the structures also move from left to right, constantly reappearing because of periodic boundary conditions.

choose a spatial Fourier mode, plot its amplitude as a function of time, and fit a function of the form $\exp(\gamma t)$ to it. Notice that this is *a priori* different from the linear theory growth rate γ, Eq. (2.13). However, to obtain tractable expressions, it is usual practice to equate the nonlinear growth rate with the linear one by using some reinterpretation of the variables appearing in it. This procedure is not unique, however, and different schemes lead to different nonlinear theories. Which one, if any, is correct can be tested only by experiments and numerical simulations.

2.5.1 Sudan–Robinson theory

Sudan (1983a) and *Sudan* (1983b) proposed that the electron–neutral collision frequency is effectively enhanced by the presence of turbulent small–scale waves. That is, $\nu_e^{\text{eff}} = \nu_e + \nu_e^*$, where the anomalous collision frequency ν_e^* is interpreted as emerging from electron–wave collisions. In the perpendicular plane, $\psi = \nu_e \nu_i / \Omega_e \Omega_i$ is thus replaced by $\psi^{\text{eff}} = \nu_e^{\text{eff}} \nu_i / \Omega_e \Omega_i$. The fluid theory dispersion relation becomes (assuming $L = \infty$)

$$\omega_r = \frac{k V_d \cos\theta}{1 + \psi^{\text{eff}}} \tag{2.38}$$

$$\gamma = \frac{\psi^{\text{eff}}}{1 + \psi^{\text{eff}}} \frac{1}{\nu_i} \left[\left(\frac{k V_d \cos\theta}{1 + \psi^{\text{eff}}} \right)^2 - \mathbf{k}^2 C_s^2 \right]. \tag{2.39}$$

This corrected γ is then interpreted as the true nonlinear growth rate, although it is formally similar to linear theory. The quasi–steadiness condition $\gamma = 0$ then implies

$$V_d \cos\theta = C_s (1 + \psi + \psi^*). \tag{2.40}$$

If this equation is understood as defining ψ^*, that is

$$\psi^* = \frac{V_d}{C_s} \cos\theta - 1 - \psi, \tag{2.41}$$

then we see that ψ^*, and thus ν_e^*, will explicitly depend on the flow angle. [Notice that in formulas like this, it is to be understood that ψ^* is never negative: in fact we should write $\psi^* = \max((V_d/C_s)\cos\theta - 1 - \psi, 0)$].

Sudan's theory predicts the following "C_s–limited cosine law" for the phase velocity:

$$v_{ph} = \frac{\omega_r}{k} = \min\left[\frac{V_d \cos\theta}{1 + \psi}, C_s \right]. \tag{2.42}$$

Other predictions are harder to make. Many authors have tried to relate ψ^*, or ν_e^*/ν_e, to the turbulence spectrum, but mutual agreement has not been particularly good.

Robinson (1986) generalized Sudan's mechanism to include self–consistent heating (*Robinson,* 1986). The mechanism was also applied to the BGK kinetic linear dispersion relation (*Robinson and Honary,* 1990). The results of Robinson's heating calculations were in good agreement with incoherent scatter radar observations (*Williams et al.,* 1992).

2.5.2 Aspect angle stabilization

To generalize or modify Sudan's theory, we notice that in order to obtain the experimental result $v_{ph} \approx C_s$ we only need the condition $\gamma = 0$, not necessarily anomalous collisions. The full expression for ψ (Eq. 2.10) contains the aspect angle. If α were nonzero, then we would have enhanced ψ also. In analogy with Sudan's theory, this would yield

$$\sin^2 \alpha = \frac{\nu_e}{\nu_i}\frac{m_e}{m_i}\left(\frac{V_d}{C_s}\cos\theta - 1 - \frac{\nu_e \nu_i}{\Omega_e \Omega_i}\right). \tag{2.43}$$

The interpretation of this expression is that the most intense waves have nonzero aspect angle and their phase velocity is equal to C_s, if the electric field above the threshold. The radar will see mainly these most intense modes, not the weaker ones propagating exactly perpendicularly, and will consequently report $v_{ph} \approx C_s$ (e.g., *Ossakow et al.,* 1975).

Both aspect angle stabilization theory and Sudan–Robinson theory are formally similar in that both have enhanced ψ. By having a larger ψ, the growth rate can be turned to zero. However, it must be kept in mind that the growth rate is not a monotonically decreasing function of ψ. Instead, the growth rate will have a maximum at some value of ψ, which is larger than the classical perpendicular value but smaller than the anomalous value. Thus, too small anomalous ψ destabilizes rather than stabilizes the waves. This follows directly from Eq. (2.39), see also *Wang and Tsunoda* (1975) and *Fejer et al.* (1984b).

2.5.3 Flow angle stabilization

As mentioned above, the particle simulations carried out in Paper 2 produced surprising results: 1) the phase velocity was not saturated by C_s, 2) the flow angle of most intense waves was shifted away from drift direction such that $\mathbf{k} \cdot \mathbf{E} < 0$. These results can be understood in terms of a different kind of nonlinear theory, called flow angle stabilization (Paper 3). It is very similar to aspect angle stabilization. In order to satisfy

$$V_d \cos\theta = C_s(1 + \psi) \tag{2.44}$$

we do not assume an enhanced ψ, but simply that the most intense modes have a suitable flow angle θ:

$$\theta = \arccos\left[\frac{C_s}{V_d}(1+\psi)\right]. \tag{2.45}$$

The interpretation is exactly similar to aspect angle stabilization: the radar will see mainly waves with nonzero θ, and this will lead, in a statistical sense, to $v_{ph} \approx C_s$ whenever the mean drift is above the threshold. For this scheme to work, it must be assumed that there is enough large–scale turbulence so that there is sufficient flow angle mixing. In the equatorial electrojet this seems to be the case (*Farley and Balsley,* 1973). In the auroral region there is in general less gradient drift turbulence because the perpendicular gradient introduced by the height dependence of the electron density is small. Thus the flow angle stabilization predicts less dramatic C_s saturation than in the equatorial zone. This is in accordance with observations: in the auroral region the C_s saturation is seldom clear, but is usually realized only in a statistical sense.

Flow angle stabilization is conceptually similar to aspect angle stabilization. In both cases, a modification in the angular dependence of the spatial power spectrum is assumed. However, in flow angle stabilization, contrary to aspect angle stabilization, the growth rate decreases monotonically with the angular variable.

2.6 Current status and future directions

The gradient drift turbulence (type 2 spectra) is well understood. Both simulations (*Keskinen et al.,* 1979) and theory (*Sudan and Keskinen,* 1977) are in good agreement with radar observations (*Fejer and Kelley,* 1980). If required, the simulations could nowadays be performed with good spatial resolution. The spatial power spectrum obeys a power law, and it is almost isotropic with respect to the flow angle. The remaining problem with gradient drift turbulence is how it interacts with the Farley–Buneman waves, but obviously this question can be meaningfully studied only if the Farley–Buneman instability itself is first properly understood.

The Farley–Buneman problem is not yet solved, but a solution may be emerging. The original puzzle with equatorial observations was the C_s saturation (*Farley and Balsley,* 1973). Throughout the 1970's, many theoretical explanations were proposed (e.g.,*Fejer and Kelley,* 1980), but no one was really compatible with all observations. Sudan's theory (*Sudan,* 1983b) explained the phenomenon in a way compatible with observations, but a rigorous mathematical foundation was still missing.

Sudan's theory gained wide popularity, and it was applied to the auroral region by several authors, including *Robinson* (1986). Perhaps this was a misleading step, since in the auroral region a strict C_s saturation was never really observed.

Therefore, it was necessary to modify Sudan's theory to account for phase velocities larger than C_s; also for type 4 waves (*Fejer et al.,* 1986) which have phase velocities about twice C_s. Most authors assumed elevated electron temperatures (high C_s) to explain the high phase velocities.

In situ rocket measurements (*Pfaff et al.,* 1987) indicated that the phase velocity followed linear theory, and no C_s saturation was seen. This was enigmatic, not least because simultaneous radar observations showed C_s saturation. The particle simulations (Paper 2) may provide an answer. According to these simulations, the phase velocity obeys linear theory, but the most intense modes have flow angles such that their phase velocity is close to C_s. When this information is combined with the turbulent picture of the equatorial electrojet (*Farley and Balsley,* 1973), both radar and rocket observations can be at least qualitatively explained.

The flow angle stabilization may also explain the high Doppler shift auroral spectra without need to assume very high electron temperatures. Elevated electron temperatures probably exist, but it is not likely that they can alone explain all high Doppler velocity events (*Haldoupis and Schlegel,* 1990). The flow angle stabilization scheme can produce high Doppler velocities directly, if the flow angle is small and if the scattering volume is not too turbulent. This may be the primary reason why auroral and equatorial spectra are so different: in the auroral region there is no significant permanent density gradient perpendicular to $\mathbf{B}$, as at the equator, and therefore there is less gradient drift turbulence. Consequently, the spectra obey more the linear theory and are less C_s saturated on the average (Paper 3).

When there is auroral particle precipitation, the situation may change drastically since then there may be steep (scale length less than one kilometer) density gradients associated with the auroral arcs. Then there may be a lot of gradient drift turbulence, and C_s saturation is expected. On the other hand, the gradients may be large enough to appreciably raise or lower the Farley–Buneman instability threshold, which may again modify the spectra (*Prikryl et al.,* 1992). This may lead to diverse and irregular spectra during auroral precipitation.

Flow angle stabilization thus appears to provide general answers to many puzzling questions, which are difficult to explain by other means. In particular the flow angle stabilization provides a simple and natural reason for the observed difference of equatorial and auroral spectra.

It remains to be seen to what extent the flow angle stabilization will realize its current promise. Above all we need an experimental verification. To this end, a new combined coherent/incoherent scatter radar campaign would be ideal. The numerical simulations should proceed toward more quantitative estimates of the spatial turbulence spectrum. This could be accomplished by using implicit particle codes or hybrid simulations. The theoretical estimates like those given in Paper 3 should include the density gradients and electron heating.

If this program is successful, it may open new possibilities to address the old question of coherent scatter data 'inversion', i.e. the estimation of the electric field, when the temporal power spectrum is known. If the numerical simulations can produce the spatial power spectrum accurately enough, a statistical inversion theoretical approach for the estimation of **E** becomes feasible.

3 Ionosphere–magnetosphere coupling simulations

3.1 Introduction

The auroral ionosphere is coupled through magnetic field–aligned currents (FAC) to the magnetosphere (Fig. 3.1). Upward and downward FACs are closed mainly by E–region horizontal currents. Because E–region irregularities are current–driven, the regions of strong irregularities and intense horizontal electrojet currents generally coincide. The upward FAC is carried mostly by energetic electrons precipitating from the magnetosphere to the ionosphere. At nighttime, the electron precipitation is usually the most important source of ionization. It is not uncommon for the height–integrated Hall conductivity to get enhanced by an order of magnitude because of precipitation. These conductivity changes, which can be very complicated spatially and temporally, tend to alter the horizontal electric field (in almost any medium, the electric field and the conductivity are more or less anticorrelated). Because the field lines are good conductors, the modified electric field tends to propagate back to the magnetosphere, where the horizontal field was initially generated. Thus we have a feedback loop between ionosphere and magnetosphere, where the auroral electron precipitation and associated conductivity changes play a decisive and controlling role.

The structure of this Chapter is as follows. The component regions of mesoscale ionosphere–magnetosphere coupling, namely the ionosphere, the field lines, and the magnetosphere are covered in turn. The intent is not to pursue the ionosphere–magnetosphere coupling physics in full, but only to the extent that is needed for numerical simulations. Our contributions to the field include Paper 4 and *Janhunen and Huuskonen* (1992). The latter one presents an earlier version of the program. These studies are largely based on the pioneering works of *Lotko and Schultz* (1988) and *Lotko and Shen* (1991). A general review for ionosphere–magnetosphere coupling can be found in e.g. *Lysak* (1990). Recent numerical developments that have partly utilized techniques similar to ours, include *Wei and Lee* (1993) and *Keskinen et al.* (1992).

3.2 Ionosphere

It is customary to assume that the ionospheric electric field is electrostatic without further arguments. However, since in the magnetosphere the inductive (rotational, divergence–free) part of the electric field is non–negligible, we shall consider more carefully under which circumstances the ionospheric electric field can be assumed curl–free.

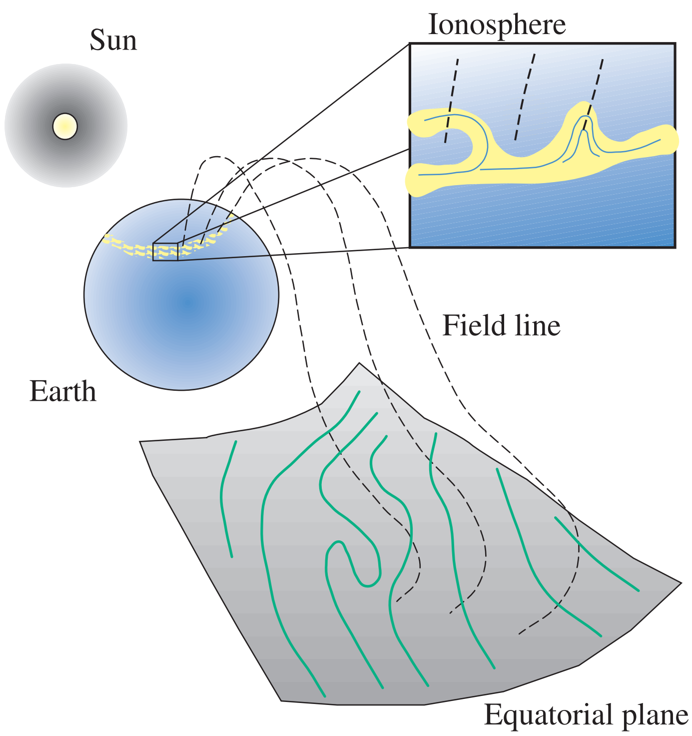

Figure 3.1. Schematic representation of the ionosphere, the field lines, and the magnetospheric equatorial plane. The figure is not to scale.

The "curl–freeness" of the **E** field can be tested by estimating

$$\frac{|\nabla\times\mathbf{E}|}{E/L} \tag{3.1}$$

where L is a typical horizontal length scale for **E** variations. From Faraday's law,

$$|\nabla\times\mathbf{E}| = \left|\frac{\partial\mathbf{B}_1}{\partial t}\right| \tag{3.2}$$

where $\mathbf{B}_1$ is the variation geomagnetic field. (The main field $\mathbf{B}_0$ is constant in time, and $\mathbf{B} = \mathbf{B}_0 + \mathbf{B}_1$.) Thus we obtain

$$\frac{|\nabla\times\mathbf{E}|}{E/L} = \frac{\left|\frac{\partial\mathbf{B}_1}{\partial t}\right|}{E/L} \sim \frac{1}{T}\frac{B_1}{B_0}\frac{B_0 L}{E} = \frac{B_1}{B_0}\frac{L}{TV} \tag{3.3}$$

where T is a typical time scale and V is the magnitude of the $\mathbf{E}\times\mathbf{B}$ velocity. If we now assume that the spatial and temporal scales L and T are related by

$$\frac{L}{T} \sim V = \frac{E}{B_0} \tag{3.4}$$

then it follows that

$$\frac{|\nabla\times\mathbf{E}|}{E/L} \sim \frac{B_1}{B_0} \ll 1, \tag{3.5}$$

i.e., that $\mathbf{E}$ is approximately electrostatic since the variation field B_1 is much smaller than the main field B_0. Notice that the result is not necessarily valid if L/T much exceeds the $\mathbf{E}\times\mathbf{B}$ velocity. In practice this means that we cannot completely rule out the possibility of inductive ionospheric electric fields during very intense geomagnetic storms.

We shall from now on nevertheless assume that the electric field is electrostatic. Then the current density is given by

$$\mathbf{j} = -\sigma \cdot \nabla\phi \tag{3.6}$$

where σ is the conductivity tensor. The current continuity equation $\nabla \cdot \mathbf{j} = 0$ will then imply an elliptic second order partial differential equation for ϕ:

$$\nabla \cdot (\sigma \cdot \nabla\phi) = 0. \tag{3.7}$$

Often in mesoscale studies (concerning structures larger than individual auroral arcs but smaller than the auroral zone), the horizontal current layer thickness (about 20 km) is much smaller than the relevant horizontal scale sizes. In this case we may use the height–integrated version,

$$-\nabla \cdot (\Sigma \cdot \nabla\phi) = j_{||} \tag{3.8}$$

where $\Sigma = \int dz\sigma$ is the height–integrated conductivity tensor and $j_{||}$ is the field–aligned current entering the ionosphere. The height–integrated conductivity tensor is defined by

$$\Sigma \cdot \mathbf{E} = \Sigma_P \mathbf{E} + \Sigma_H \hat{\mathbf{B}}\times\mathbf{E} \tag{3.9}$$

for any $\mathbf{E}$, if the magnetic field $\mathbf{B}$ can be assumed vertical. Here Σ_P and Σ_H are the height–integrated Pedersen and Hall conductivities, respectively. The electric field $\mathbf{E}$ and the corresponding potential ϕ in the above formulas are expressed in a coordinate frame co–moving with the ionospheric neutral particles. If neutral winds or corotation must be taken into account, a corresponding $\mathbf{v}\times\mathbf{B}$ term must be added to $\mathbf{E}$.

In Paper 4, the height–integrated conductivities Σ_P and Σ_H are computed from a 3D electron density. The electron density is computed using a height–dependent model for the ionization production rate. The model has the ionosphere–magnetosphere potential difference as a parameter, which is taken self–consistently from the simulation. As a crude approximation, these complicated steps give the following general behavior: the Hall conductivity Σ_H depends linearly on the ionosphere–magnetosphere potential difference $\Delta\phi$, while Σ_P is proportional to the square root of $\Delta\phi$. It is a peculiarity of nature, however, that this simple scheme cannot be used in the simulation: doing so will result in convergence problems, among other things.

3.3 Field lines

Satellite studies (e.g., *Weimer et al.,* 1987) have revealed that an approximately linear relationship between the parallel potential drop and the FAC exists:

$$j_{||} = K\Delta\phi = K(\phi_m - \phi_i), \tag{3.10}$$

where $j_{||}$ is the FAC at the ionospheric level, ϕ_i is electric potential at ionospheric level, ϕ_m is the electric potential in the magnetosphere (on the same field line as ϕ_i), and K is the constant of proportionality, the Lyons–Evans–Lundin constant. An often quoted numerical value of K for upward current is 10^{-9} mho m^{-2}. A typical 'large' voltage $|\Delta\phi|$ is 10 kV, the corresponding FAC is then 10 μAm^{-2}. In this Introduction we use the convention that the FAC is positive downward.

Field–aligned currents associated with dynamic mesoscale processes are mainly carried by electrons. When the potential difference $\Delta\phi$ is positive (FAC downward), the current–carrying electrons are of ionospheric origin, and no additional ionization is generated in the ionosphere. Downward FAC is sometimes called return current. When the potential difference is negative (FAC upward), the current–carrying electrons are of magnetospheric origin. In this case the electrons precipitate into the ionosphere, they may have high energy, and they may enhance the ionospheric electron density.

For upward FAC configuration, Eq. (3.10) can be derived relatively simply by considering single particle motion (e.g., *Friedman and Lemaire,* 1980). The reason why K is finite is the 'mirror force' of converging field lines on precipitating electrons. The numerical value of K derived from the theory is in rather good agreement with observations.

In return current regions, the single–particle theory predicts that K is nearly infinite, i.e. very much larger than 10^{-9} mho m^{-2} (e.g., *Lemaire and Scherer,* 1983). This is because the mirror force tries to bounce back only earthward–moving particles, not particles coming from the ionosphere. Observations in return current regions are much more rare than in electron precipitation regions. There are some indications (*Goertz et al.,* 1991) that Eq. (3.10) could also hold for return currents regions. If this is the case, then some nonlinear mechanism must be involved in generating the necessary resistivity. On the other hand, some investigators are willing to believe that the single–particle theory is basically correct also for return current regions (*J. Lemaire,* 1993, private comm.); in this case the K parameter should be very large for return current regions. The issue of current–voltage relationship in return current regions is perhaps the biggest open question in mesoscale field line coupling in general, and it deserves further observational studies.

3.4 Magnetosphere

Paper 4 presented a version of the mesoscale ionosphere–magnetosphere coupling where the magnetosphere was modeled as a two–dimensional incompressible fluid with constant magnetic field. We call it the old model here. Currently we have an enhanced model under development. This is called the new model or the semiglobal model in the following.

When the curl operation is applied to the old model momentum equation (Paper 4), the pressure term, it being a gradient, vanishes, and only a term proportional to the field–aligned current is left from the $\mathbf{j}\times\mathbf{B}$ force density. The resulting equation for the vorticity function is easy to handle numerically. The main drawback of this approach is that the magnetic field is not a dynamical variable but is a fixed background quantity. The formulation is therefore suitable only for the near–earth region. In particular, many of the substorm–related processes such as magnetic reconnection are excluded.

It has proved difficult to improve the old model gradually. Therefore, full three–dimensional magnetohydrodynamic equations (e.g., *Sato,* 1985) in a semiglobal simulation box are used in the new model. 'Semiglobal' in this regard means that a significant part of the magnetotail is explicitly included, but not the magnetopause or the solar wind. For example, the simulation box might be $\{(x,y,z)|x < -6, -5 < y < 5, -4 < z < 4\}$, where (x,y,z) are the GSM coordinates expressed in units of Earth radii (Fig. 3.2).

There are two main problems with the new semiglobal model: the boundary conditions in general, and the coupling to the ionosphere, which is also related to boundary conditions.

In any fluid model, including magnetohydrodynamics, the boundaries should be classified as either inflow or outflow boundaries. At inflow boundaries, the fluid quantities can, and shall, be specified (Dirichlet boundary condition). At outflow boundaries, on the other hand, Dirichlet conditions are generally useless and trying to use them only leads to numerical instabilities. One has to specify the normal derivatives (Neumann conditions), or the second normal derivatives. The reason for this behavior is intuitively clear. As an example, consider ordinary unmagnetized fluid flow in a pipe. One can specify the flow velocity at an upstream boundary, but not at a downstream boundary.

The six boundary planes of the semiglobal simulation box can not, however, be simply classified as inflow and outflow boundaries. The character of a boundary point may change dynamically during the simulation. A reasonable way out of this dilemma is to use 'silky glove' boundary conditions. By this we mean adding suitable friction terms to the MHD equations such that the friction coefficient becomes larger as the numerical boundaries are approached. For example, the

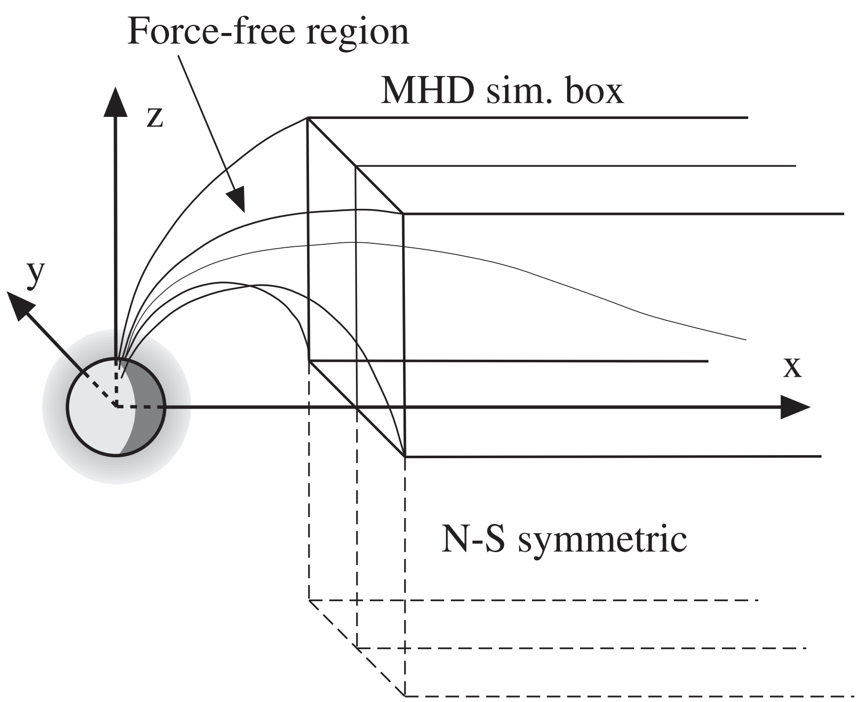

Figure 3.2. Geometry of the new semiglobal coupling model. The model includes the ionosphere, the field lines (force–free region), and the magnetosphere, which is modeled as 3D MHD fluid. The figure is approximately to scale.

momentum equation becomes

$$\frac{\partial \mathbf{v}}{\partial t} = \langle \text{other terms} \rangle - \frac{\mathbf{v} - \mathbf{v}_{BC}}{\tau} \qquad (3.11)$$

where τ is the time scale for frictional relaxation, and $\mathbf{v}_{BC}$ is the boundary condition velocity field. The time scale τ depends on position (x, y, z) in such a way that τ is very large in the central regions of the grid but becomes smaller as the point moves close to the boundaries. In this arrangement, the flow velocity will be gradually forced to follow $\mathbf{v}_{BC}$ near the boundaries. The actual "hard" boundary conditions are then less important, they may be just of Dirichlet type everywhere. Similar forcing terms may be added to other MHD equations, too.

It seems that emplying silky glove boundary conditions is a promising technique to deal with simulation boxes that have no natural boundaries. Numerical instabilities are much easier to control than in traditional "hard" boundary conditions. The programmer is free to choose the spatial distribution of τ for each variable to control the amount of external forcing. The method is also easy to program and as such it takes only little extra computing time and memory. A drawback is that the grid should be made larger than the actual domain of interest. This increases the computational workload by a factor of two or so, which is usually tolerable.

The interaction with the ionosphere is the other problematic part. In the old model, the magnetospheric potential ϕ_m was explicitly obtained from the vorticity ω by solving Poisson's equation. The ionospheric part used ϕ_m as input, and returned the ionospheric potential ϕ_i. The FAC could then be computed from the potential difference, and the FAC was just one of the input terms in the vorticity equation. Thus the magnetospheric part used the FAC as input, and ϕ_m as output.

In the semiglobal model, the ionospheric interaction must be treated differently. The interaction takes place at the earthward boundary of the MHD simulation box (e.g. at $x = -6$ $\mathrm{R_E}$). Imposing FAC into the MHD box is difficult because the MHD equations are written in terms of the magnetic field, not the current density. However, reversing the roles of FAC and ϕ_m as input and output quantities helps. Obtaining the FAC as an output quantity is straightforward. Inputting the potential ϕ_m in the MHD box is also possible: it amounts to specifying the **B**–perpendicular velocity

$$\mathbf{v}_\perp = \frac{-\nabla\phi_m \times \mathbf{B}}{B^2}. \tag{3.12}$$

The velocity can be specified if it is accompanied with a suitable specification of the variation magnetic field $\mathbf{B}_1$. Let us define two auxiliary quantities $\beta_\pm = \mathbf{B}_1 \pm \sqrt{\mu_0\rho}\mathbf{v}$ representing in– and outgoing Alfvén waves. β_- and β_+ propagate parallel and antiparallel, respectively, to the background magnetic field. Using these formulas, we can prepare an ingoing Alfvén wave, whose velocity variation part is the one obtained from ϕ_m.

With a slight modification, the ionospheric routine can also accept the FAC as an input quantity; it then computes ϕ_i from $j_{||}$. ϕ_m can then be computed from ϕ_i and FAC using the current–voltage relationship. Thus the reverse interaction scheme works also in companion with the ionosphere.

3.5 Numerical techniques: elliptic solving

The biggest numerical problem that had to be addressed in Paper 4 was the elliptic solver. The situation is very similar in the new model, thus the following discussion applies to both. Consider an elliptic equation of the form $\nabla \cdot (\Sigma_P \nabla\phi)... = s$. Considering only x dependence, we have for the left–hand side

$$\Sigma_P \frac{\partial^2\phi}{\partial x^2} + \frac{\partial\Sigma_P}{\partial x}\frac{\partial\phi}{\partial x} \quad ... \tag{3.13}$$

Using the usual second order accurate discretizations for the derivatives, this becomes

$$\Sigma_P \frac{\phi_{i+1} + \phi_{i-1} - 2\phi_i}{(\Delta x)^2} + \frac{\partial\Sigma_P}{\partial x}\left(\frac{\phi_{i+1} - \phi_{i-1}}{2\Delta x}\right) \quad ... \tag{3.14}$$

Considering the linear system that this equation represents in matrix form, we easily see that the coefficients of ϕ_i become diagonal matrix elements, while coefficients of $\phi_{i\pm1}$ contribute to off–diagonal elements. Let us form the quotient

$$R \equiv \left| \frac{\text{coeff-of } \phi_{i+1}}{\text{coeff-of } \phi_i} \right| = \frac{1}{\Sigma_P} \frac{\partial \Sigma_P}{\partial x} \frac{\Delta x}{4}. \tag{3.15}$$

For simplicity, we have dropped the term proportional to Σ_P in the coefficient of ϕ_{i+1}. Replacing $\partial\Sigma_P/\partial x$ by its typical value Σ_P/L (L being a typical length scale for Σ_P), we find

$$R \sim \frac{\Delta x}{L}. \tag{3.16}$$

A numerical linear system solution by almost any iterative method generally becomes unstable as the off–diagonal elements grow in magnitude. That is, for any given method and problem, there exists an R_{max}, above which the method is unstable. Often in practice, R_{max} is of the order of unity. In order to prevent numerical problems, the spatial grid spacing should therefore be smaller than the scale of spatial variations of the conductivity. The smallness of R is difficult to guarantee beforehand, since the conductivity is computed self–consistently during the simulation. A typical conductivity gradient length scale is of the order of 100 km, but because the problem is highly nonlinear, much steeper gradients may occasionally arise.

In practice it has been found that numerical instabilities do not arise, if the grid spacing is small enough (of the order of 10 km or less) and if a robust elliptic solver is used. The successive overrelaxation method (SOR) (*Press et al.,* 1992) was found to be divergent in most cases. The alternating direction implicit routine (ADI) (*Press et al.,* 1992) works well if the ionospheric grid is isotropic and rectangular: $\Delta x = \Delta y = const$. For a more general grid, however, ADI tends to diverge, and other methods have to be sought. The strongly implicit procedure (SIP) (*Stone,* 1968; *Jesshope,* 1979) is more complicated to program and the number of arithmetic operations per iteration is larger than in SOR and ADI, but we applied it successfully to the case of nonuniform ionospheric grid. SIP is our current method of choice for solving difficult elliptic equations. The only problems we have met have been with very large (more than 100 grid points along one dimension) spatial grids; for these, SIP sometimes fails to converge. SIP can be made very efficient on vector machines by modifying the algorithm such that the grid is traversed diagonally. We know of no implementation of SIP that would be efficient on massively parallel computers.

3.6 Summary and future models

A mesoscale or semiglobal ionosphere–magnetosphere coupling model contains the ionosphere, the field lines (acceleration region, force–free region), and the magnetosphere as separately modeled regions. Information between these regions is carried by electric fields (potentials), field–aligned currents and precipitating electrons.

The semiglobal model is still in development stage, but some of its features can already be discussed. The scale sizes are suitable for studying e.g. complete substorm activations. The computational requirements are larger than for the mesoscale model, mainly due to the three–dimensionality of the MHD code, but still clearly smaller than for any serious global MHD model (e.g., *Kageyama et al.,* 1992). Although the development from a mesoscale model to a semiglobal model almost totally concerns the magnetospheric part, the magnetotail is still the poorest modeled region in the system. The most important features not included are explicit Alfvén waves propagating along field lines, and non–MHD effects in the magnetotail (e.g., chaotization of particle orbits, nonideal Ohm's law, anisotropic pressure).

Let us finally briefly compare the semiglobal coupling model with global MHD models. Global MHD models have the benefit that they contain the solar wind — magnetosphere interaction explicitly. They are also easier to code, since only one set of equations in a single domain is needed, and the boundary conditions are simple. Their most important drawback is the poor representation of the ionosphere. The semiglobal model is more complicated to program, and the solar wind is only implicitly contained in the boundary conditions. There are still some conceptual difficulties associated with boundary conditions between different regions. The main strength of a semiglobal coupling model is the adequate representation of the ionosphere. It is expected that there are many auroral region phenomena in which the ionosphere plays such an important role that the semiglobal coupling model is the best way to proceed, regardless of the problems mentioned.

References

Bhatnagar, P.L., Gross, E.P., and Krook, M., A model for collision processes in gases, 1, Small amplitude processes in charged and neutral one-component systems, *Phys. Rev., 94,* 511-525, 1954.

Birdsall, C.K. and Langdon, A.B., Plasma physics via computer simulation, McGraw–Hill Book Company, Singapore, 1985.

Buneman, O., Excitation of field aligned soundwaves by electron streams, *Phys. Rev. Lett., 10,* 285–287, 1963.

D'Angelo, N., Pécseli H.L., Petersen, P.I., The Farley instability: A laboratory test, *J. Geophys. Res., 79,* 4747–4751, 1974.

D'Angelo, N. and Merlino, R.L., Comment on "Are observed broadband plasma wave amplitudes large enough to explain the enhanced electron temperatures of the high–latitude E region?" by J.-P. St.-Maurice and Russ Laher, *J. Geophys. Res., 92,* 321–322, 1987.

Farley, D.T., A plasma instability resulting in field–aligned irregularities in the ionosphere, *J. Geophys. Res., 68,* 6083–6097, 1963.

Farley, D.T. and Balsley, B.B., Instabilities in the equatorial electrojet, *J. Geophys. Res., 78,* 227–239, 1973.

Farley, D.T., Theory of equatorial electrojet plasma waves: new developments and current status, *J. Atmos. Terr. Phys., 47,* 729–744, 1985.

Farley, D.T. and Providakes, J., The variation with T_e and T_i of the velocity of unstable ionospheric two–stream waves, *J. Geophys. Res., 94,* 15415–15420, 1989.

Fejer, B. and Kelley, M.C., Ionospheric irregularities, *Rev. Geophys. Space Phys.,* 18 401–454, 1980.

Fejer, B.G., Reed, R.W., Farley, D.T., Swartz, W.E., and Kelley, M.C., Ion cyclotron waves as a possible source of resonant auroral radar echoes, *J. Geophys. Res., 89,* 187–194, 1984a.

Fejer, B.G., Providakes, J., and Farley, D.T., Theory of plasma waves in the auroral E region, *J. Geophys. Res., 89,* 7487–7494, 1984b.

Fejer, B.G., Providakes, J., Farley, D.T., and Swartz, W.E., Auroral E region plasma waves and elevated electron temperatures, *J. Geophys. Res., 91,* 13583–13592, 1986.

Ferch, R.L. and Sudan, R.N., Numerical simulations of Type II gradient drift irregularities in the equatorial electrojet, *J. Geophys. Res., 82,* 2283–2288, 1977.

Fried, B.D. and Conte, S.D., The plasma dispersion function, Academic Press Inc., New York, 1961.

Friedman, M., and Lemaire, J., Relationship between auroral electron fluxes and field aligned electric potential difference, *J. Geophys. Res., 85,* 664–670, 1980.

Goertz, C.K., Whelan, T., and Nishikawa, K.-I., A new numerical code for simulating current–driven instabilities on auroral field lines, *J. Geophys. Res., 96,,* 9579–9593, 1991.

Haldoupis, C., A review on radio studies of auroral E–region ionospheric irregularities, *Ann. Geophys., 7,* 329–258, 1989.

Haldoupis, C. and Schlegel, K., Direct comparison of 1-m irregularity phase velocities and ion acoustic speeds in he auroral E–region ionosphere, *J. Geophys. Res., 95,* 18989–19000, 1990.

Haldoupis, C., Schlegel, K., and Nielsen, E., On type 3 auroral VHF coherent radar backscatter, *J. Geophys. Res., 97,* 4109–4120, 1992.

Haldoupis, C., Koehler, J.A., Sofko, G.J., Danskin, D.W., McKibben, M.J., and André, D., Preferential phase velocities for type 4 irregularities in the auroral E region plasma, *J. Geophys. Res., 98,* 6173–6179, 1993.

Ichimaru, S., Basic principles of plasma physics, A statistical approach, W.A. Benjamin, Inc., London, 1973.

Janhunen, P., and Huuskonen, A., On role of precipitation–enhanced conductivities in mesoscale ionosphere–magnetosphere coupling, *Proceedings of First International Conference on Substorms,* Kiruna, Sweden, ESA–SP 335, 33–37, 1992.

Jesshope, C.R., SIPSOL, a suite of subprograms for the solution of the linear equations arising from elliptic particle differential equations, *Comp. Phys. Comm., 17,* 383–391, 1979.

John, P.I. and Saxena, Y.C., Observation of the Farley–Buneman instability in laboratory plasma, *Geophys. Res. Lett., 2,* 251–254, 1975.

Kageyama, A., Watanabe, K., and Sato, T., A global simulation of the magnetosphere with a long tail: No interplanetary magnetic field, *J. Geophys. Res., 97,* 3929–3943, 1992.

Kelley, M.C., and Mozer, F.S., Electric field and plasma density oscillations due to the high–frequency Hall current two–stream instability in the auroral E region, *J. Geophys. Res., 78,* 2214–2221, 1973.

Keskinen, M.J., Sudan, R.N., and Ferch, R.L., Temporal and spatial power spectrum studies of numerical simulations of Type II gradient drift irregularities in the equatorial electrojet, *J. Geophys. Res., 84,* 1419–1430, 1979.

Keskinen, M.J., Chaturvedi, P.K., and Ossakow, S.L., A low–altitude mechanism for mesoscale dynamics, structure, and current filamentation in the discrete aurora, *Geophys. Res. Lett., 19,* 2199–2202, 1992.

Krall, N.A. and Trivelpiece, A.W., Principles of plasma physics, McGraw–Hill Book Company, New York, 1973.

Kustom, B., D'Angelo, N., and Merlino, R.L., A laboratory investigation of the high–frequency Farley–Buneman instability, *J. Geophys. Res., 90,* 1698–1704, 1985.

Lemaire, J., and Scherer, M., Field–aligned current density versus electric potential characteristics for magnetospheric flux tubes, *Ann. Geophys., 1,* 91–96, 1983.

Lotko, W., and Schultz, C.G., Internal shear layers in auroral dynamics, in Moore, T.E., and Waite, J.H. Jr. (eds.), Modeling Magnetospheric Plasma, *Geophys. Monograph Series 44,* AGU, Washington D.C., 1988.

Lotko, W., and Shen, M.–M., On large–scale rotational motions and energetics of auroral shear layers, *J. Geophys. Res., 96,* 9549–9565, 1991.

Lysak, R.L., Electrodynamic coupling of the magnetosphere and ionosphere, *Space Sci. Rev., 52,* 33–87, 1990.

Machida, S. and Goertz, C.K., Computer simulation of the Farley-Buneman instability and anomalous electron heating in the auroral ionosphere, *J. Geophys. Res., 93,* 9993–10000, 1988.

McDonald, B.E., Coffey, T.P., Ossakow, S., and Sudan, R.N., Preliminary report of numerical simulation of type 2 irregularities in the equatorial electrojet, *J. Geophys. Res., 79,* 2551–2554, 1974.

Ossakow, S.L., Papadopoulos, K., Orens, J., and Coffey, T., Parallel propagation effects on the type 1 electrojet instabilities, *J. Geophys. Res., 80,* 141–148, 1975.

Pfaff, R.F., Kelley, M.C., Fejer, B.G., Kudeki, E., Carlson, C.W., Pedersen, A., and Hausler, B., Electric field and plasma density measurements in the auroral electrojet, *J. Geophys. Res., 89,* 236–244, 1984.

Pfaff, R.F., Kelley, M.C., Kudeki, E., Fejer, B.G., and Baker, K.D., Electric field and plasma density measurements in the strongly driven daytime equatorial electrojet 2. Two–stream waves, *J. Geophys. Res., 92,* 13597–13612, 1987.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., Numerical Recipes in C: The art of scientific computing, 2nd edition, Cambridge, 1992.

Prikryl, P., André, D., Koehler, J.A., and Sofko, G.J., Type 4 VHF radio aurora observations during low to moderate geomagnetic activity, *Planet. Space Sci., 40,* 1371–1389, 1992.

Primdahl, F. and Bahnsen, A., Auroral E–region diagnosis by means of nonlinearly stabilized plasma waves, *Ann. Geophys., 3,* 57–62, 1985.

Robinson, T.R., Towards a self–consistent non–linear theory of radar auroral backscatter, *J. Atmos. Terr. Phys., 48,* 417–422, 1986.

Robinson, T.R. and Honary, F., A resonance broadening kinetic theory of the modified two–stream instability: implications for radar auroral backscatter experiments, *J. Geophys. Res., 95,* 1073–1085, 1990.

Robinson, T.R., The effects of the resonance broadening of Farley–Buneman waves on electron dynamics and heating in the auroral E–region, *J. Atmos. Terr. Phys., 54,* 749–757, 1992.

Sahr, J.D., Observation and theory of the radar aurora, Ph. D. thesis, Cornell University, New York, 1990.

Sahr, J.D., Farley, D.T., and Swartz, W.E., The altitude of type 3 auroral irregularities: radar interferometer observations and implications, *J. Geophys. Res., 96,* 17805–17811, 1991.

Sato, T, Principles of magnetohydrodynamic simulation in space plasmas, in *Computer simulation of space plasmas,* edited by H. Matsumoto and T. Sato, Terra, Tokyo, 1985.

Schmidt, M.J. and Gary, S.P., Density gradients and the Farley–Buneman instability, *J. Geophys. Res., 78,* 8261–8265, 1973.

Schlegel, K. and St.-Maurice, J.P., Anomalous heating of the polar E region by unstable plasma waves, 1, Observations, *J. Geophys. Res., 86,* 1447–1452, 1981.

Schlegel, K. and St-Maurice, J.-P., Note on the parallel propagation effects of unstable Farley–Buneman waves at high latitudes, *Planet. Space Sci., 30,* 315–318, 1982.

Schlegel, K., Interpretation of auroral radar experiments using a kinetic theory of the two–stream instability, *Radio Sci., 18,* 108–118, 1983.

Schlegel, K., Measurement of electron density fluctuations during the ROSE rocket flights, *J. Atmos. Terr. Phys., 54,* 715–723, 1992.

Schunk, R.W., Mathematical structure of transport equations for multispecies flows, *Rev. Geophys. Space Phys.,* 15 429–445, 1977.

St.-Maurice, J.-P., Schlegel, K., and Banks, P.M., Anomalous heating of the polar E region by unstable plasma waves 2. Theory, *J. Geophys. Res., 86,* 1453–1462, 1981.

St.-Maurice, J.-P. and Laher, R., Are observed broadband plasma wave amplitudes large enough to explain the enhanced temperatures in the high-latitude E region?, *J. Geophys. Res., 90,* 2843–2850, 1985.

Stone, H.I., Iterative solution of implicit approximations of multidimensional partial differential equations, *Siam J. Numer. Anal., 5,,* 530–558, 1968.

Stubbe, P., Theory of electrostatic waves in an E region plasma 1. General formulation, *J. Geophys. Res., 94,* 5303–5315, 1989.

Sudan, R.N., Akinrimisi, J., Farley, D.T., Generation of small–scale irregularities in he equatorial electrojet, *J. Geophys. Res., 78,* 240–248, 1973.

Sudan, R.N. and Keskinen, M., Theory of strongly turbulent two–dimensional convection of low–pressure plasma, *Phys. Rev. Lett., 38,* 966–970, 1977.

Sudan, R.N., Unified theory of type I and type II irregularities in the equatorial electrojet, *J. Geophys. Res., 88,* 4853–4860, 1983a.

Sudan, R.N., Nonlinear theory of type I irregularities in the equatorial electrojet, *Geophys. Res. Lett., 10,* 983–986, 1983b.

Wang, T.N.C. and Tsunoda, R.T., On a crossed field two–stream plasma instability in the auroral plasma, *J. Geophys. Res., 80,* 2172–2182, 1975.

Wei, C.Q., and Lee, L.C., Coupling of magnetopause–boundary layer to the polar ionosphere, *J. Geophys. Res., 98,* 5707–5725, 1993.

Weimer, D.R., Gurnett, D.A., Goertz, C.K., Menietti, J.D., Burch, J.L., and Sugiura, M., The current–voltage relationship in auroral current sheets, *J. Geophys. Res., 92,* 187–194, 1987.

Williams, P.J.S., Jones, B., and Jones, G.O.L., The measured relationship between electric field strength and electron temperature in the auroral E–region, *J. Atmos. Terr. Phys., 54,* 741–748, 1992.

Paper 1

Paper 2

Paper 3

Paper 4