Chapter 5. Optimization of Models

The parameters of model auroral current systems were described in the previous chapter. To model the currents at any given time, these parameters must be adjusted to provide an optimal match to the available observations. Once that optimization is done, it is claimed that the data have been inverted within the framework of the proposed forward model. As will be seen in later chapters, the simple forward models proposed do, in fact, allow a good representation of the auroral current systems. Their parameters are also demonstrably amenable to optimization based on ground based data, at least in the case of east-west systems. After a brief discussion of aspects of the forward models which are relevant to the optimization, optimization techniques themselves are discussed and demonstrated below.

Inversion of geophysical data will sometimes not be able to give unique results. If a particular model is put forward to interpret certain data, there may exist other models able to fit the data about equally well. The classic relevant example of this is the long debate (1930's to 1960's) about the nature of the current systems producing geomagnetic disturbances, which has already been alluded to. As pointed out by Akasofu [1991], Chapman's assertion, based on the knowledge at the time, was that only equivalent current systems (in the ionosphere) could be discussed. Despite the economy of Birkeland's assertions at about the turn of the century that three-dimensional systems would explain not only the magnetic variations but the precipitation of particles needed to create the aurora, Chapman's view of ionospheric currents being the physical currents became accepted. This was not solely due to Chapman's dominance in the field and refusal to openly debate. Birkeland's assertions, and the supportive modelling of particle trajectories done by Störmer, were considered by Chapman, and even illustrated in the influential book *Geomagnetism* [Chapman and Bartels, 1940]. The contemporary, and incorrect, context was the view, prevalent at the time (and still held by semi-informed laymen today), that particle streams ejected by the Sun were directly involved in production of aurora. Particularly over the long distances this implied, it appeared to Chapman that a stream of only one charge could not hold its integrity, and that precipitating particles, although present, were in neutral streams [Chapman and Bartels, 1940, Chapter 24]. Such neutral streams would of course have little or no magnetic effect. Despite vigorous championing of the Birkeland hypothesis by Alfvén, it was only in the 1960's that the existence and importance of extra-ionospheric currents was widely realized [Dessler, 1984]. As already described, wide acceptance of the three-dimensional nature of the current systems occurred only after space measurements of magnetic fields which could be plausibly due only to field-aligned (or nearly so) currents. The basic model used here is derived from knowledge since ascertained about the physical current systems. The same techniques could, in fact, be used to give a representation of the data based on a non-physical model (such as Chapman equivalent currents). Thus, the applicability of the model here, and the uniqueness of the solution, rest on knowledge which is not incorporated into the technique. In this sense the numerical techniques are hopefully being used in keeping with Hamming's [1986] maxim that "The Purpose of Computing is Insight, Not Numbers". Nevertheless, in practice the results of the computation are numbers, and these are in general the numerical parameters which are varied in the proposed model until that model represents the data optimally.

a. Parameters

The choice of parameters for a model may be viewed as part of the 'direct problem': for a good model the parameters will allow the reproduction of the data [Sabatier, 1993]. In fact the basic problem faced by the analyst is the 'inverse problem' in which data (usually contaminated by noise) are available and from which the parameters are to be deduced. Generally we will be convinced that the number of parameters being used is far smaller than the actual number required for a complete description of the physical system. For example, in the lower magnetosphere, the current density in SI (MKS) units is a sum over the charge-velocity product of at least thousands of particles. The resulting magnetic field seen at some point is physically due to the complex motion of those thousands of particles, yet may be well represented in practice from knowledge of just an integrated quantity, the current density. The very large-dimensional space of physical states and defined parameters contains such integrated parameters as a subspace. The projection of the state vector of the overall system onto the subspace defined by the parameters is the best approximation possible, within that subspace, to the overall state. The approach to approximation as viewed from a vector space point of view is discussed briefly by Anton [1987, ch. 7]. The fitting of data by models is one step removed from the fitting of functions by other functions since in fact the actual state is not generally projected onto the parameter subspace, but rather the state maps into a data space and the model provides a method of mapping the parameters into that space also. Within that space there exists an optimal choice of parameters bringing the two projections as close as possible. The choice of a model with its associated parameters and the techniques for obtaining the best values of those parameters are related but distinct. Due reflection must be given to the choice of a model. Not only the physical basis of the model may be relevant but also its efficiency. Does the data set provide sufficient information to distinguish the number of parameters chosen?

The usefulness of parameters and their choice depends on their intended application. A spherical harmonic expansion such as the IGRF representation of the Earth's main field [Langel, 1987] has important practical applications, yet most of the parameters do not have an independent physical meaning. Representation of the active fields at high latitudes by similar expansions [Walker *et al.*, 1995] can also not claim to provide parameters having a direct correspondence to physical parameters, although the spatial distribution of approximations to some physical fields (for example equivalent ionospheric currents) may be represented in terms of the parameters derived to represent the magnetic field. The approach of representing fields in terms of expansions does not yield an immediate *mechanistic model* [Bates and Watts, 1988, p. 67] for the phenomena being studied. Such a model, in its most powerful form, has both fewer parameters in general than does a non-mechanistic model, and more meaning attached to those parameters. A useful illustration of the difference may be given by considering the Earth's nearly dipolar field. In practice at least 50 parameters are used in the IGRF, of which the first is the dipolar term. If the field actually were that of a dipole, the statement of this fact in terms of the IGRF as presently used would have all parameters but one equal to zero. This may seem a trivial example, but it may be taken one step further. Consider a shift of the dipole away from the origin of the

expansion (in fact the Earth's field is largely that of such an 'eccentric dipole'). In this case the field would require all terms of the finite number in the IGRF expansion for its description and even then that description would be only an approximation. Yet in fact the difference between this field and that of a centred dipole can be represented by at most only three more parameters representing the origin translation. Obviously the expansion is wasteful of parameters, those parameters having (individually) little relation to the physical situation, and provides little insight into that situation although possibly describing it well. In terms once more of vector spaces, there may well exist a vector space in which relatively few parameters suffice to describe a physical situation. In the case of particles and currents previously described, that efficient space may be the one comprised of the variables usually associated with MHD; in the case of the eccentric dipole, it is clearly that of the dipole parameter and translations. In this study the aim is characterization of substorm current variations in a manner which leads to physical insight. For that reason a parametrization is sought which adequately describes the data (reflecting success of the optimization process) and also has parameters with physical meaning. The aim here, then, is to invert data into a small and efficient subspace in which the parameters have physical meaning and shed light on causality. This point is reflected in the recent statement about the substorm current wedge (SCW) current system, from Tsyganenko [1997]: "due to the very dynamical nature of the disturbances and a relatively small number of available simultaneous data, one can expect to fit meaningfully only a few parameters of the SCW, reflecting its most important characteristics. This also requires the model to be relatively simple."

A further aspect of having a small number of parameters, each having a direct physical interpretation, is the practical aspect. If the physical effects of changing the parameters are understood by the analyst, the analysis itself can be more effectively done. To quote Bates and Watts [1988, p. 72], "one of the best things one can do to ensure a successful nonlinear analysis is to obtain good starting values for the parameters - values from which convergence is quickly obtained." Parameters which lend themselves to initial analysis and choice of starting values by graphical methods, by understanding the response of the goodness of fit to their derivatives, or by insight into behaviour with reduced dimensionality (i.e. holding one or more parameters fixed), are preferred.

An attempt at simple characterization of substorm current systems was made by Cramoysan and Orr [1993] within the framework of the McPherron current wedge. The parameters chosen are only four, the meridians of filamentary field-aligned currents, the latitude of the ionospheric path joining them, and the magnitude of the current. The geometrical parameters are considered to be fixed while the current varies. A slightly more refined, but still limited, approach to this problem allows all of the parameters to vary, yet only considers one wedge to be active. An independent but similar approach was used by Connors [1993] and considerably earlier by Horning *et al.* [1974]. The Cramoysan and Orr approach arguably oversimplifies the problem by having not only one simple current system in the form of the substorm current wedge, but further by having only one parameter (the current) changing in association with it. That the other parameters of the wedge generally also vary is well known (see e.g. Wiens and Rostoker [1975]). The use of more parameters in the other cited studies allows a good fit to the data, but in both cases only stations distant

from the auroral zone current systems were used, so that information was lost and the parameters solved for correspond to 'lower-order' approximations.

In general, and in particular if data from the auroral zone are included, more parameters must be used, as there is more to auroral zone currents during a substorm than simply the substorm current wedge. In practice the global current system during active times is regarded as having two other loops active which may be reasonably represented by a similar overall configuration to that of the wedge. These loops are those of the eastward and westward electrojets in the evening and morning sectors, respectively. They are similar to the substorm current wedge in that evidence exists of regions of net downward field-aligned current feeding them, ionospheric current flow, and then net upward field-aligned current where the current leaves the ionosphere. The global current system is also known to have continuous activity in current systems whose ionospheric paths are aligned nearly north-south and whose physical existence is shown through satellite passes through the associated field-aligned currents [Iijima and Potemra, 1978]. These systems, discussed in section 1.d, generally show considerably less ground effect than do the electrojets or substorm current wedge, for reasons illustrated in Chapter 4. Nevertheless, parameters may be added to the model to represent them, and those parameters optimized. Knowing beforehand that the perturbations associated with this near-solenoidal system may be small, one might expect the problem of finding their parameters to be somewhat ill-posed if based only on surface measurements. For that reason, the examination of satellite data which may better fix those parameters is desirable.

b. Forward Model Choice

The basic possible configurations of current systems may be classified by the flow direction of electric current in the ionosphere. It is now known that auroral zone currents are for the most part due to currents entering the ionosphere, flowing through it over some distance, and then bleeding out into space once more through field-aligned current flow. The specification of entry and exit points is tantamount to specifying the ionospheric current flow direction, provided that this current flow is basically rectilinear. The ionospheric electric current flow direction, considered as a vector in spherical coordinates based on the tilted dipole poles, may be decomposed into a north-south (meridional or poloidal) and an east-west (toroidal) component. This natural geometric division of types of current flow also corresponds roughly to expected flow directions of Pedersen and Hall currents in the observed auroral zone electric potential distribution (see section 2.c.2).

It is observed that east-west currents in the auroral zone may extend coherently across several time zones (tens of degrees), while the auroral zone itself has a north-south extent which is usually less than this, rarely more than ten degrees. Thus a normal parametrization of these currents is through a system of large east-west extent relative to its width. Since in fact the currents flow within an auroral oval having varying latitude at each meridian, in practice any requirement that these current systems have strictly east-west flow is relaxed. For a system spanning several time zones yet restricted to the auroral oval, flow will necessarily be roughly east-west, but that is all. There may be some meridional component

to the roughly toroidal 'Hall' electrojet and this meridional component will be well defined, primarily by the maxima in the magnetic horizontal component underlying the electrojet and the maxima in magnitude of the vertical component near its boundaries.

North-south current flow in the ionosphere may be regarded as having closure currents at or near the borders of the auroral zone. A parametrization involving *strictly* north-south flow is not overly restrictive and in addition forces this type of current system to be different from the east-west type. The requirement for north-south flow in this sense imposes a rough orthogonality condition - there really are two types of current systems present. While in principle the less restricted east-west systems allow choices of parameters which would allow them to become nearly north-south oriented, the data usually furnish no 'motivation' for them to do so. The north-south systems are not allowed to have any toroidal current component, but have flow restricted to meridians. Since these poloidal systems, if long, contribute little perturbation even for relatively large currents, their parameters could be expected to be unresponsive to the observations. If allowed to have an unrestricted toroidal component, such current systems could rearrange themselves to have primarily toroidal current flow since such flow is most effective at producing ground perturbations and thus can be changed to efficiently get a good match to the data. This would change the originally north-south aligned currents into primarily east-west currents. In practice the problem described does occur if initially poloidal systems are not constrained to remain poloidal by not allowing any toroidal flow. By contrast, the east-west systems are constrained by ground observations of the X_m and Z components, to which they primarily respond, to cover large extents east-west. North-south systems are not well constrained from the ground but primarily respond to the Y_m component. As was the case in the discovery of the region 1 and 2 field-aligned currents which was through satellite observations, the primary constraint on the north-south system is through observations above the ionosphere. Such observations are not directly included in the present version of the Automated Forward Modelling routine. As seen near the end of Chapter 4, polar cap observations may provide some basis for solving for the poloidal currents. However, it does not harm to stress once more that inclusion of the north-south currents, if constrained only by ground observations, must be very carefully monitored.

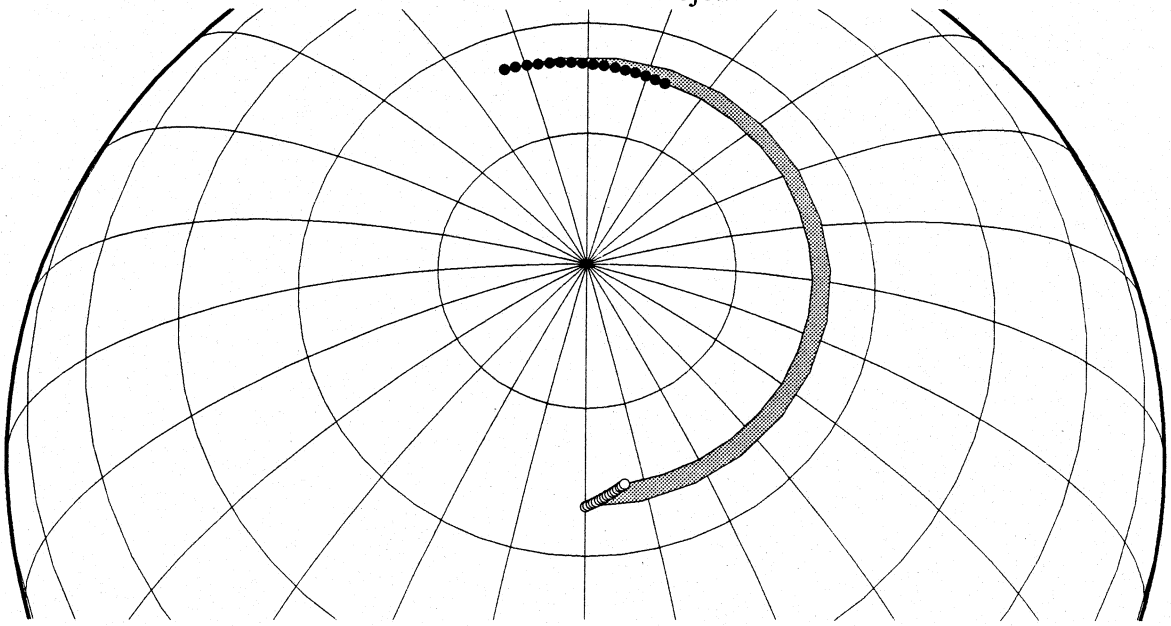
In section 4.d, representative magnetic fields arising from the two basic types of current systems were discussed. Those results parallel those of Kisabeth [1972], but are for systems exceeding the lengths which he studied. The original Kisabeth method breaks down numerically for systems of the length shown, as the spread-out point arrangement of current sources in Kisabeth's implementation of the Biot-Savart integral, using Gaussian integration, becomes dominant. A modified method was used in this study. This method consists of a simple rearrangement of the current elements used in Gaussian integration (section 4.d.1) so that there is a dense placement of elements near any given observation station. This eliminates the problems of Kisabeth's method as applied to long systems at the cost of doubling computation time. Here we briefly discuss the parametrization of the toroidal and poloidal systems. The parameters and the relationships between them indicated here are those used in the modelling.

The paths followed by currents and those defined by the intersections of sheets of field-aligned current with the ionosphere are here, for simplicity, regarded to be 'straight', which here has the following definition. A 'straight' line on the curved surface of the Earth is regarded as a locus of points along which the derivative of the latitude of points along the locus with respect to their longitude is a constant. This is a fairly natural way of parametrizing lines of current flow, but there could be others defined, and Gaussian integration can be redefined to use any parameter. Auroral currents on the largest scale generally flow parallel to the auroral oval, and thus more complex paths can be defined [e.g. Kisabeth 1972]. Here any attempt to represent such a more complex path is made using shorter 'straight' paths, appropriately joined. An example may be found in Chapter 6 where there appears to be a northward 'spur' of electrojet, deviating from what is likely the main electrojet following the auroral oval. These two systems are constrained to intersect and form one joined system in a geometrical sense, although differing currents may flow in each. It was also found necessary, in Chapter 6, to 'bend' the westward electrojet. This was done by having two systems constrained to join, but with the current also constrained to be the same in each.

1. Parameters of East-West (Toroidal) Systems

As indicated, the toroidal systems are generally of great east-west extent compared to their north-south extent. Also mentioned was the fact that it is in practice not necessary for practical reasons, nor in accord with observations, to restrict the current flow in these electrojet systems to be strictly along lines of constant magnetic latitude. In order to specify an east-west system of finite north-south extent, nine parameters are needed. The system has a region of downward current at one extremity, with two points (each consisting of a pair of longitude, latitude parameters) specifying each end of the sheet. Current then flows through the ionosphere to another current sheet, where it flows upward. As each of the two sheets requires four parameters for its general specification, there are eight geometric parameters associated with the system. The current flowing throughout the system is the ninth parameter. Figure 5.1 illustrates this type of system, with a rather distributed current 'feeding' the system and a rather concentrated current 'draining' it. Although the regions of field-aligned current flow are modelled by sheets of current, the effects seen at a distance would be similar to those of distributed regions of current. In most cases there are not enough data to allow these two cases (sheet or distributed region) to be distinguished, so that it would not be justified to add more parameters to specify more detail than in this simplest representation by sheets of field-aligned current. The proof of this statement is to be found in the generally good fits available through these simple systems as evidenced in Chapters 6 to 10.

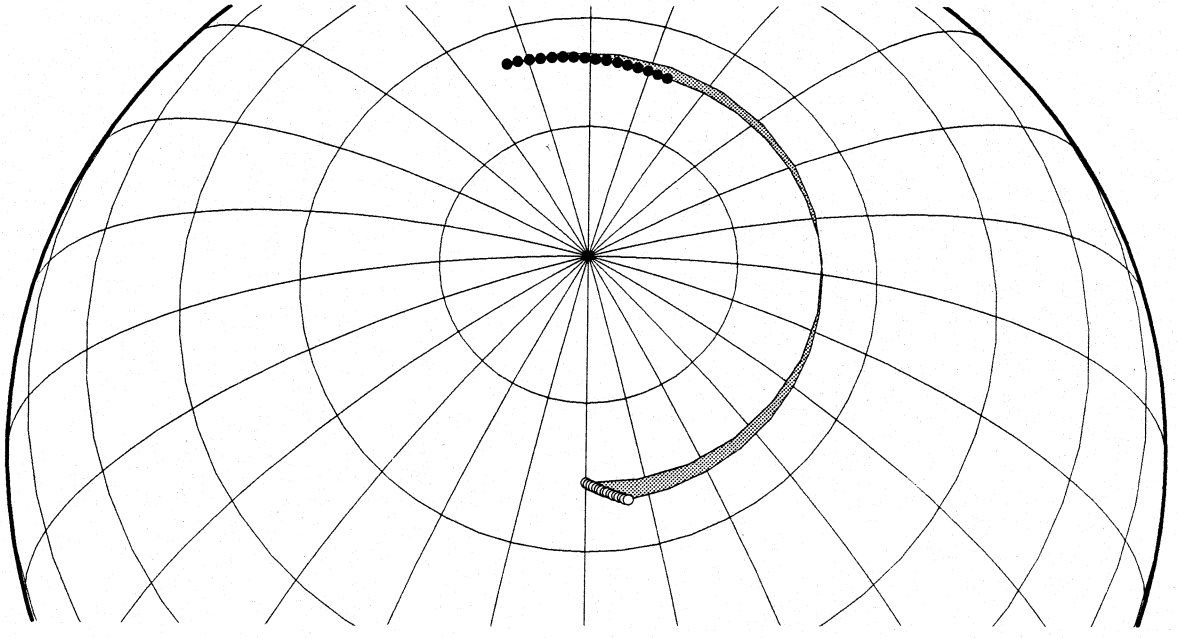
Figure 5.1 Single east-west current system (electrojet) illustrating its parameters. Downward current is indicated by black circles arrayed along a line. Region of current flow is shaded grey in proportion to total current in the system. Upward current is indicated by white circles arrayed along a line. Direction of ionospheric current flow is from black circles to white circles within the electrojet.



The optimization procedure will generally produce meaningful results if long east-west systems are used. The maxima in magnetic perturbations near or bordering the long electrojets which are normally observed in the auroral zone ensure this. In practice, all parameters within such a system may be varied and it could be expected that their optimized values would be meaningful and unique. This is to a large extent true, but certain parameters may be regarded as 'degenerate'. The first, and innocuous, form of such degeneracy may be seen, in that reversal of the northernmost and southernmost points of a current sheet does not in any way change the sheet. It does not matter which is specified first in a list of parameters. Within a current system specified by two sheets, interchanging the northernmost and southernmost points of *both* sheets forming the ends of the system, does not affect the system. A simple example is that of, say, westward current flow bounded by two lines of constant magnetic latitude. It does not matter which is specified first, the northern, or the southern, boundary. A similar and only slightly less innocuous degeneracy is that of direction of current flow. Within an east-west system, a positive sign may be attached to current flow in one direction. With some convention on order of parameters or graphical representation to represent upward or downward field-aligned flow at the ends of the system, reversal of the sign of the current causes the roles of 'upward' and 'downward' to be reversed. In most of the graphical representations in the chapters presenting results, this possible sign reversal has been corrected for, and the regions of up and down current indeed correspond to those directions. In some cases, a convention of positive sign for a direction of current flow has been adopted, and sometimes a negative value is used to indicate flow in the opposite direction. For example, in a case where positive current flow is defined to be toward the west, negative values mean that the current flow was towards the east. Generally little confusion arises from these conventions.

A more troublesome type of degeneracy has a possible physical basis yet results in current systems which 'look' unphysical. As illustrated by Figure 5.2, if the end point parameters of only *one* of the bounding currents are reversed, then the joining of the end points, done according to order in a parameter list, causes crossing of the lines bounding the edges of the region having ionospheric current flow. This results in, rather than a simple quadrilateral, a 'bowtie' shape for the region of current flow. Clearly, not much of this configuration corresponds to realistic ionospheric current flow. The current density at the crossover point is in principle infinite and it is difficult to see that any naturally occurring current system would evolve to look like this. Yet such a system is indeed acceptable under the 'rules' envisaged above, which call for specification of a current system through two lines which are the loci of field-aligned current, and their joining with ionospheric current flow. Clearly these 'rules' need to be modified, and they are in practice, simply by adding the further stipulation that if the boundary lines for the ionospheric current flow cross within the longitude range of the current system, that system is not allowed.

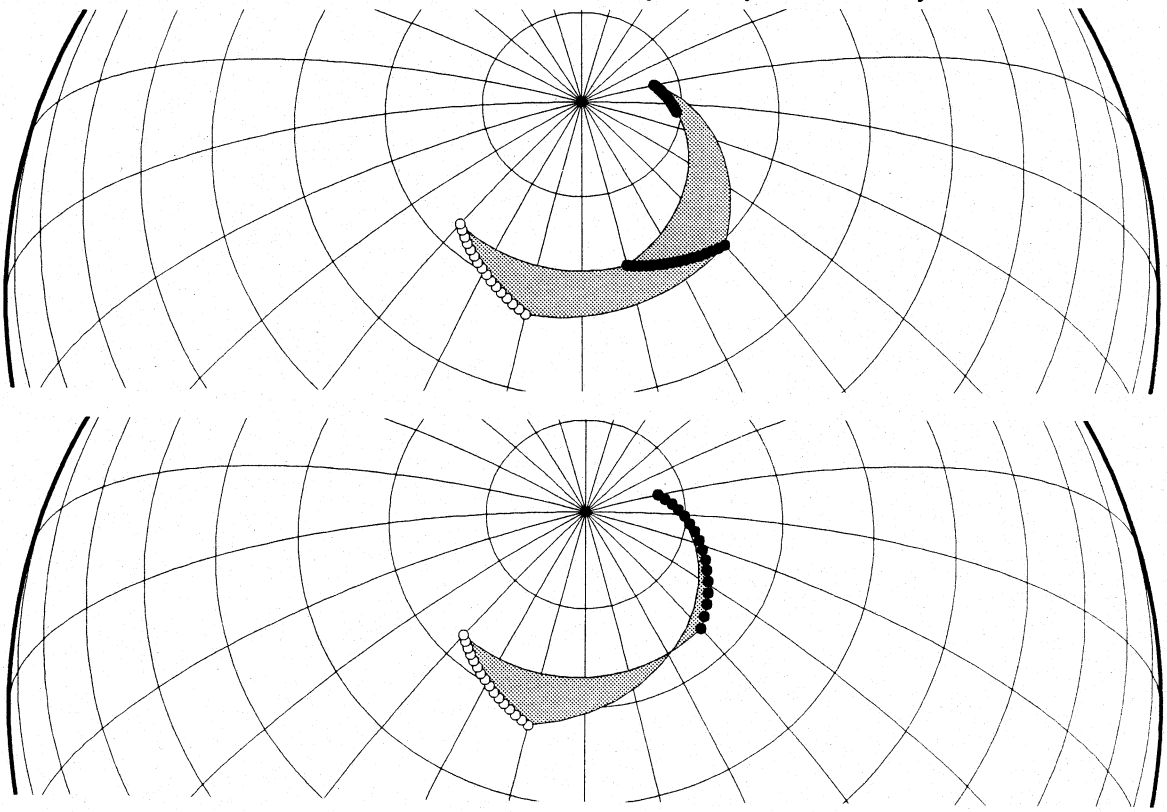
Figure 5.2 Reversal of a single bounding current sheet to result in a 'bowtie'.



While it is easy to say that 'bowtie' systems are not 'allowed' and to exclude them from the realm of acceptable results from the programme, there may be cases in which they do in some sense represent what the currents are doing. In particular, one can imagine a long system, but one not well constrained by stations all along its length. Then it is quite possible that the region of enhanced current density near the crossover point would in fact not be near a station, and thus not cause too much damage to the optimal fit to the data. Further, as opposed to rectilinear flow in the sense of 'constant derivative of latitude versus longitude' along the current flow lines, current flow generally follows the auroral oval, which is a curved path.

If there are stations straddling the auroral oval along two widely separated meridians (see Figure 5.3), then the local current flow across these may be best locally represented by straight line currents across each meridian. Due to the curvature of the auroral oval, these straight line segments may cross each meridian at a different angle. It is possible to imagine bowtie configurations which would best represent the current flow as seen at each meridian since, with arms of unequal length, maximal current flow at different angles may be had in a 'bowtie', as illustrated in Figure 5.3. This is not possible with a simple quadrilateral figure.

Figure 5.3 Optimal fit of a curved current system by a 'bowtie' system.



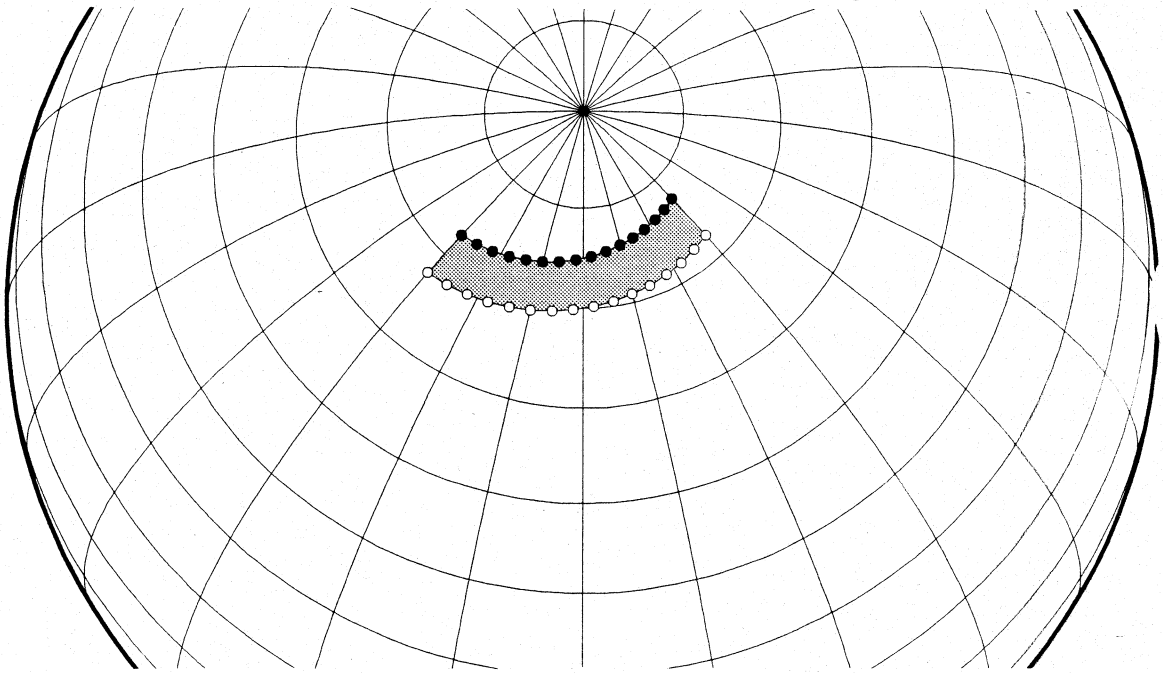
Thus it is possible to see that in some cases a 'bowtie' could give an optimal fit not possible with a 'straight' system. It is even possible that from an initially straight system provided as a starting point, a bowtie system could arise and be optimal. Since there are strong physical grounds for rejecting bowtie systems, this cannot be allowed to happen, and rules are supplied in the programme to prevent it. 'Bowtie' configurations may also arise due to an inadequacy of the forward model in the regions of field-aligned currents. As alluded to above, the simple model used here has simple sheets where physically one might expect field-aligned current flow over relatively large areas. It has been argued above that in many cases data will not be sufficient to discern the details of the regions of field-aligned current flow in any case. This in turn implies that if we are able to make models with good fits (which we are), based on the current sheet approximation, then the models must not be very sensitive to the parameters involved in specifying sheet orientation. Such indeterminacy means that the difference between a 'bowtie' system and a preferred rectilinear system may be small in terms optimality of fit, and it is even possible that the

'bowtie' system will have a lower value of an optimizing parameter (see discussion in next section). Since the routine efficiently finds the optimal fit, it can prefer 'bowties' over rectilinear systems if this is the case. However, since the problem arises precisely because of a lack of sensitivity to the parameters determining whether a system is a 'bowtie' or not, one can repair 'bowties' simply by detecting them and changing these parameters. In practice the 'untying' is done by interchanging longitudes of the ends of one sheet, without interchange of the corresponding latitudes. This untwists the 'bowtie' while retaining the longitudinal range of the field-aligned current sheet, and has proven a very satisfactory approach to resolving the problem.

2. Parameters of North-South (Poloidal) Systems

Much like an east-west system, a north-south system nominally requires eight geometric parameters to specify its four corners. However, due to the ill-posed nature of the problem of determining the parameters of north-south systems from the ground, it is necessary to restrict the current flow in these systems to be strictly along meridians (i.e. purely poloidal). This effectively removes two parameters from play, meaning that only seven parameters are needed. These are the longitude of the most westerly points, the latitudes of up and down current at that longitude, the corresponding three parameters at the most easterly point, and the current. A north-south system is illustrated in Figure 5.4

Figure 5.4 North-south current system as specified by seven parameters.



In practice, the north-south system is not shown in illustrations in this thesis, even where incorporated into the modelling. The practical reason for this is that the east-west systems have by far the largest ground effect and thus must be illustrated, whereas the north-south system has little ground effect and would simply clutter the map. Another aspect is that not only is the north-south system generally acknowledged to be in essentially the same place as the east-west electrojets [Senior and Robinson, 1982], but these parameters for this system would be poorly defined by ground modelling in the first place. Since this is the case, in most cases there has been a relationship established between the east-west and north-south systems (usually that their borders are essentially identical), and the parameters of the north-south system have not been independently derived in modelling, apart from (usually) the total current in the system.

3. Constraints on Parameters

Various parameters may not be entirely independent or may have known properties which are not part of the model. In this case it may be useful to constrain the parameters with respect to one another or an external value. The following discussion explains how this is in practice realized in this project. In other methods which have been discussed(in Chapter 4) the issue of constraint may also arise. For AMIE, some aspects are considered in Appendix I, Section III of Knipp [1989], and use of the statistical expected value matrix $\mathbf{C}_s$ as a constraint was discussed in Section 4.c above, so constraint in other methods is not discussed further here.

A constraint may be implemented in Automated Forward Modelling in at least two ways, both of which have in fact been used. Since in many cases a constraint can be expressed as a non-linear equation or operation, it can be accommodated in a nonlinear fitting routine in a natural way by simply incorporating the rule of constraint into the calculation routine of the forward model. In this way, the forward model may contain rules (the constraints) about the interrelation of its parameters. The second, less rigid, method of enforcing a constraint is to impose a penalty on an optimization variable (see next section) if the constraint condition is not met. This is in principle similar to the approach taken in AMIE, although here it was generally applied in a multiplicative fashion rather than the basically additive fashion employed in AMIE.

An operational example may be offered as follows. Consider a case in which the latitudinal width of an electrojet has been well determined along one meridian where there is a magnetometer chain. Global modelling is to be done, including an adjacent region (a gap) where there are insufficient stations to well constrain the width of the electrojet. Based on past knowledge, or better yet based on some possibly available source of ancillary information such as a satellite image, it may be reasonable to constrain the electrojet latitudinal width to be constant. Incorporating the constraint rule into the forward modelling could be done simply by requiring the parameters specifying the latitudes of the ends of the upward and downward current sheets to be separated by the same amount. In practice, rather than solving for the latitudes of the four corners of the region bounding ionospheric current flow, one would search for the longitudes and for one latitude (say the

most southerly) at each end, the other latitude at each being calculated through addition of the latitudinal width. In this case one effectively removes two parameters from the system, since the sheets of field aligned current at the ends of the region of ionospheric current flow have one less free parameter each. The alternate approach is to incorporate the constraint rule into the optimization variable which is systematically reduced in optimizing. In this case, one would increase that variable based on how far the difference of latitudes in a proposed forward model deviated from the desired value. Since the optimization involves reduction of the variable, increasing it as a proposed solution deviates from some favored solution will result in a tendency for the solution to remain near the favored solution unless strongly pulled away by the data. This latter approach is less rigid since the data could dictate violation of the constraint if sufficiently 'constraining'. For example, if one later got another set of meridian chain data to fill the gap referred to above, the first method mentioned would not respond to the new information this would give about electrojet width, whereas the second method would. The changes in the final result would vary to a degree determined by the relative weighting of the constraint condition and data.

In practice the difficulty in implementing constraints is largely one of programming. In this project a simple parser was developed to operate on an auxiliary file which told which parameters were to be varied and which to be held constrained. In the case of a parameter to be constrained, it was possible to specify that its value was to have some relationship to that of some other parameter, or to be held fixed. Simple operations involving another parameter and a constant were implemented: for example it was possible to constrain a northern corner of the current flow quadrilateral to be some latitudinal separation from the corresponding southern corner. An example will now be given to illustrate use of constraints in modelling. Table 5.1 shows parameters used in an actual model run to specify initial conditions for two currents. As mentioned above, the quadrilateral region within which current flows is specified by eight geometric parameters and by the total current flowing within the region. The geometric parameters may be split into those associated with the sheet of upward 'drain' current from the ionosphere and those of the downward 'feed' current into the ionosphere. These sheets are in turn represented by their extreme points, each specified by a longitude (negative for west longitudes) and a latitude. The set of four such pairs of points, plus the current, completely specifies a three dimensional current system as used in modelling here. Within Table 5.1, the parameters for the upward current sheet in the second system are identical to those for the downward current sheet in the first system. This is simply a way of tying the two systems together: since the current in the second, more westerly system is greater than that in the first system, the net effect is that current flows into the ionosphere at this junction.

Table 5.1 Current Parameters for Two Current Modelling Run

Longitude Up 1	Latitude. Up 1	Longitude. Up 2	Latitude. Up 2	Longitude Down 1	Latitude. Down 1	Longitude Down 2	Latitude. Down 2	Current (MA)
-104.528	62.3278	-10.2782	72.2485	-26.4025	63.4717	41.7283	77.6546	0.196215
-131.174	73.8745	-190.445	76.2815	-104.528	62.3278	-10.2782	72.2485	0.164395

Table 5.2 Rules for Variation and Constraint of Parameters for Two Currents

Longitude Up 1	Latitude. Up 1	Longitude. Up 2	Latitude. Up 2	Longitude Down 1	Latitude. Down 1	Longitude Down 2	Latitude. Down 2	Current (MA)
1	1	1	1	1	1	1	1	1
1	1	1	1	-1.1	-1.2	-1.3	-1.4	1

Table 5.2 shows the parameters which form rules for the variation of the corresponding physical parameters in Table 5.1. In this case these specify that throughout the modelling run, the parameters corresponding to a '1' are allowed to freely vary. This applies to all parameters of the first system, and to those of the upward current sheet and current magnitude of the second system. The parameters of the downward current for the second system have negative numbers associated with them. A negative number indicates that a constraint is to be applied. In this case, the rule is simply that a parameter will be made identical to a parameter within another current system. The number of that system (1, 2, etc.) must be given as the integral part of the rule parameter, while the number of the parameter within that system to use is given by the decimal. For example, the rule parameter for 'Latitude Down 1' within the second system is -1.2. This indicates that the latitude of the first point of the downward current in system 2 is to be numerically equal to the value of the second parameter from system 1. That may be seen to be the latitude of upward current from system 1. In this way the two systems are required to be joined in the manner described in the preceding paragraph, at all times during convergence of the model. If the rule parameters are not specified, they default to 1, allowing free variation of all physical parameters. In that case, an initial system as specified in Table 5.1, with joined currents, could be broken into two unlinked current loops if that would optimize the fit to the data. The number of effective parameters, if all are given free variation, is larger than in the constrained case, and with more freedom of variation, it is quite possible for a better fit to be obtained in such circumstances. The choice to enforce a constraint is motivated by the desire to have a physically realistic system, possibly at the expense of the goodness of fit. Of course, judgment would have to be used. If a much better fit is obtained without constraint, the possibility that the constraint is inappropriate must be considered.

The parsing language allows other constraints to be imposed at run time. Should it be desired to hold a physical parameter fixed at its input value, a '0' may be used as the corresponding rule parameter. For example, if the '1' for the current in the first system in Table 5.2 were changed to a '0', the input value of the current (0.196215 MA) would be used throughout the modelling run, and the optimization would be done based on only the geometric parameters. It was mentioned above that constraints could be introduced to fix the relationship among various parameters, beyond simply making them identical. The example mentioned was that it would be possible to specify a width for the electrojet. To make this clearer, consider the Latitude Up 2 and Latitude Down 2 parameters in Table 5.1. If the corresponding rule parameters in Table 5.2 were changed from '1' (for free variation) to -1.2+5, and -1.5+5, respectively, then the Latitude Up 2 and Latitude Down 2

physical parameters would be constrained to be 5 degrees larger than Latitude Up 1 and Latitude Down 1. In practice, this type of more complex restraint was rarely used, while nonvariation or linking of parameters were used extensively. In practice, use of the simple parsing language proved very effective in constructing joined systems or studying variations based on a limited set of parameters.

A very important application of constraint was in tying the north-south systems, with little ground signal, to east-west systems having sufficient signal to allow solution. In this case the geometric parameters of the systems were related to each other so that those of the east-west system were those of the north-south, and the current in the north-south system freely solved for. It should be noted that this does not mean that the position of north-south system current flow was removed from the system and simply not solved for. Rather, the combined fields of both systems would be matched to determine the geometric parameters, which are constrained to be the same for both systems. Normally those parameters would be more heavily constrained by ground observations due to the east-west part of the total current flow, as these are larger.

c. Optimization

Before considering in detail the optimization by nonlinear least squares which is actually used in Automated Forward Modelling, a brief discussion of optimization theory will be presented. In this way it will be seen that there are specific reasons for the choice of that method and for the Levenberg-Marquardt implementation used.

Rice [1983] discusses three ways in which approximation problems arise. One classical example is in the approximation of mathematical functions. Although such approximations are at the basis of the numeric computation required in digital computing, and thus fundamental in that sense, this is not the approach relevant here. Similarly, smoothing and analysis of data, his third class, may be used, but in his sense there is little 'specialized information' available in this case. A more appropriate approach is found in his class of problems involving 'representation and compactification' of data. Here 'specialized information' is available which allows a fitting model to be proposed. Parameters within the model may be varied to obtain an optimal model. This is the approach which is taken here and referred to as 'forward modelling'. But as opposed to merely the numerical analyst's interest in 'compactification', we ascribe a physical meaning to a model which fits well. By proposing a forward model in which the parameters correspond to physically measurable quantities, we claim that the observed data resulted from a physical system having characteristics analogous to those described by the best-fit model parameters.

Optimization attempts to deliver those best-fit parameters given the proposed model and the data. 'Closeness' of the model to the data is measured by the *norm* [Rice, 1983], of which many may be defined, and of which one must be chosen. To illustrate, consider the approximation of a set of N points of data, y_i, where i varies from 1 to N, by the function $F(\mathbf{a},x_i)$, or simply $F(\mathbf{a})$, where $\mathbf{a}$ is the set of parameters appropriate to the model represented by F and the x_i are ordinates corresponding to the data values. The deviation of

model from observation, with notation $\|\mathbf{y} - \mathrm{F}(\mathbf{a})\|$, is called the *norm*. The maximum deviation or Chebyshev norm is simply $\|\mathbf{y} - \mathrm{F}(\mathbf{a})\|_\infty = \max|y_i - \mathrm{F}(\mathbf{a}, x_i)|$ for $1 \le i \le N$. The sensitivity of this norm to the single largest deviating point in the data set is itself a disadvantage and it is not easy to minimize, largely due to the discrete dependence on whichever point that happens to be. Thus it is not suitable for our type of problem. More responsive to the entire data set is the norm representing average error, the least deviation or $\mathbf{L_1}$ norm, defined as $\|\mathbf{y} - \mathrm{F}(\mathbf{a})\|_1 = \sum_{i=1}^{N} |y_i - \mathrm{F}(\mathbf{a}, x_i)|$. This norm is less responsive to outlier points, as is indicated by its discussion in 'Numerical Recipes' [Press *et al.*, 1992] under the heading of 'robust statistics'. That it is also difficult to use in practice is indicated by the fact that the sole example of its use given there is for fitting straight lines. In practice the easiest norm to use is the least-squares or $\mathbf{L_2}$ norm, defined by

$$\|\mathbf{y} - F(\mathbf{a})\|_2 = \sqrt{\sum_{i=1}^{N} (y_i - F(\mathbf{a}, x_i))^2}.$$

This norm represents the distance between fit and data in an N-dimensional space, and it is clear that lessening the deviation of the fit function from the points reduces the value of the norm. For both the $\mathbf{L_1}$ and $\mathbf{L_2}$ norms, it is possible to introduce weighting functions, which may be chosen so as to make the norms for discrete data sets resemble integrals which would arise through trapezoidal integration in continuous data, or may reflect the relative reliability of each data point. In practice the weights enter as a multiplicative factor (say w_i) in each term of the sum used to form the norm, or as error weight terms inversely proportional to some error value σ_i. This leads naturally to the consideration of the effects of errors in the data in fitting and to a more rigorous reason to prefer a form of weighted $\mathbf{L_2}$ norm. Following Press *et al.* [1992], we consider that the data points y_i are normally distributed about what the 'true' model would give (with this distribution being due to random errors). The probability of obtaining the measured data set of N points is then the product of probability of each point comprising it:

$$P \propto \prod_{i=1}^{N} \left\{ \exp\left[-\frac{1}{2} \left(\frac{y_i - \mathrm{F}(\mathbf{a}, x_i)}{\sigma_i} \right)^2 \right] \Delta y \right\}$$

where Δy is an arbitrary constant reflecting a small interval in y within which the probability density of y_i is evaluated, and σ_i is the standard deviation of the distribution of data points. Intuitively (and we are reminded in 'Numerical Recipes' that statistics is *not* a branch of mathematics) one must identify that set of parameters, considered as variables, which would maximize the probability of the data set (which is given and not to be changed) to be that represented by the parameters.

Minimizing the negative of the logarithm of P will maximize P, and that will require the quantity chi-squared (χ^2),

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - F(\mathbf{a}, x_i)}{\sigma_i} \right)^2$$

to be minimized. This approach to justifying the use of χ^2 minimization for maximum likelihood estimation of parameters is based on the assumed normal distribution of data errors. Even if that is not the case, the inverse errors may be considered as weights, resulting in those points with the smallest errors receiving the highest weighting. It is also possible that the errors are unknown, in which case they may be set to unity and χ^2 minimization is equivalent to seeking the minimal $\mathbf{L_2}$ norm.

In the foregoing there has been no discussion of the functional dependence of F(**a**,x) on either the parameters **a** or the independent variable x. In practice, methods of solution of the optimization problem depend on this functional form. Methods in which the dependence on model parameters is linear are, not surprisingly, relatively amenable to direct solution, while a more general dependence requires a more involved approach. We will consider first how the linear case may be applied to magnetic data. The form of solution presented will be recognized as being the basic form used in the AMIE technique discussed in Section 4.c.

Assuming a linear dependence on the model parameters **a** does not imply any particular functional dependence on x, that dependence entering in the M basis functions X(x) put forward as part of the proposed model. Since the following discussion applies equally to discrete and continuous ordinates, the subscript on x is sometimes omitted for convenience. M should be chosen so that there are more data points than parameters being solved for if one expects a unique best solution; in other words N≥M. The quantity ν=N-M is known as the number of degrees of freedom in the problem, and it should be positive (note that ν is not to be confused with the error vector $\boldsymbol{\nu}$ discussed previously in connection with AMIE). The model's form in this linear case is

$$F(\mathbf{a}, x) = \sum_{k=1}^{M} a_k X_k(x)$$

and the χ^2 becomes

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - \sum_{k=1}^{M} a_k X_k(x_i)}{\sigma_i} \right)^2$$

where it is important to note that the basis functions appear evaluated at each data point abscissa. In this equation, it is the parameters a_k which are variables, the set of (x_i,y_i)

observation pairs being preselected, and the basis functions $X_k(x)$ by hypothesis being independent of the parameters a_k. It is thus very clear that minimization takes place in the space of the parameters. The solution of this optimization problem is called a *linear* least squares solution: it is linear because the dependence on each parameter is linear. The dependence of the basis functions on the independent variable(s) may well be nonlinear. That being the case, this approach is widely used in the representation of magnetic data by functional expansions, such as in equivalent current fitting by J. K. Walker *et al.* [1995], or as an important part of the AMIE procedure (see Chapter 4). It is instructive to examine how the optimal parameters are determined in linear least squares fitting, for comparison with the nonlinear method used in this work. Letting **A** be the matrix with elements $A_{ij} = \frac{X_j(x_i)}{\sigma_i}$ (the design matrix), and a weighted observation vector **b**, where $b_i = \frac{y_i}{\sigma_i}$, the expression above becomes $\chi^2 = \sum_{i=1}^{N}(b_i - \sum_{k=1}^{M} a_k A_{ik})^2 = (\mathbf{b} - \mathbf{Aa})^T(\mathbf{b} - \mathbf{Aa})$, leaving the vector **a** of coefficients (parameters) to solve for. This solution may be found by making the derivative of χ^2 with respect to each parameter be zero. Taking the set of such derivatives, one obtains the M normal equations (k=1,...,M) to be solved simultaneously:

$$0 = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left[y_i - \sum_{k=1}^{M} a_j X_j(x_i) \right] X_k(x_i) = (\mathbf{A}^T \cdot \mathbf{A}) \cdot \mathbf{a} - \mathbf{A}^T \cdot \mathbf{b}.$$

To obtain a solution, a matrix must be inverted and there are several standard ways to obtain such an inverse. The matrix form may be rearranged and symbolically solved to give a compact expression of the solution, which is $\mathbf{a}=(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{b}\mathbf{A}^T$. Since all variations in the parameter space are linear in the parameters, the solution is particularly simple.

Linear inversion is suitable in cases for which a sum of weighting factors multiplying basis functions allows good representation of data. The basis functions need not be orthogonal: in a polynomial expansion, for example, they would not be. In many spaces an expansion in orthogonal functions is possible; for near-Earth magnetic fields these expansions could be in spherical harmonics as in the IGRF or in some modified spherical harmonic expansion as has been detailed in Chapter 4 in discussing the KRM and AMIE techniques. As also discussed there, such parameters in general do not have direct physical meaning, which is considered here as a disadvantage. Referring to Section 5.a, we might also note that the linear expansion needed for an eccentric dipole is very inefficient and that a more physical model could have considerably fewer parameters. However, the dependence of the field on parameters specifying the position of the dipole, as may be seen by referring to the expressions for dipole fields in Chapter 1, would be highly nonlinear. Only the parameter specifying the strength of the dipole would affect the observed fields in a linear fashion. The cost of nonlinearity in the model is that a simple linear fitting approach no longer is applicable.

Where the data have linear dependence on the parameters, χ^2 has quadratic dependence on them. Requiring its derivative to be zero restores linearity since the derivative of a quadratic is a linear function. This yielded an algebraically soluble (by matrix inversion) set of equations. One can equally do the same steps for arbitrary dependence of the model on its parameters. The problem arises at the last step since the resulting analogue of the normal equations will generally not be soluble. For this reason, most attempts to solve optimization problems with nonlinear dependence on parameters are done by following gradients of χ^2 in the parameter space. Usually this must be done numerically. Some numerical methods popular about 1970 are discussed in the reprint of the book *Numerical Methods that Work* by Acton [1990] and in the second edition of Hamming's [1973] well-known treatise. The Fletcher-Powell method and its variants appear to have been most popular at that time. The related Levenberg-Marquardt method is now the most widely used, since it 'works very well in practice and has become the standard' [Press *et al.*, 1992, p. 683]. Both are variable metric methods which attempt to find an optimal step size in parameter space, and both are related to Newton's method in the sense of following gradients toward a solution. Here the Levenberg-Marquardt algorithm is described largely through putting together and generalizing material contained in various places in *Numerical Recipes in C* [Press *et al.*, 1992]. The fact that the algorithm is the present-day 'technique of choice' and more references, including several to its widespread commercial implementations, are found in a recent summary [Lampton, 1997]. A concise description of the routine's operation, more coherent than that in *Numerical Recipes* is also given there, along with a test which finds a strategy permitting more computational efficiency.

We assume that χ^2 for a forward model which has parameters can be represented in the space of those parameters as a differentiable function. This is the case for our magnetic field models if using physically meaningful parameters. Then the χ^2 function can be expanded about any point $\mathbf{a_0}$ in M-dimensional (as above) parameter space, to have a value at another point $\mathbf{a}$ (i.e. for a different set of parameters) which is

$$\chi^2(\mathbf{a}) = \chi^2(\mathbf{a}_0) + \sum_{i=1}^{M} \frac{\partial \chi^2}{\partial a_i}\Big|_{a_0} (a_i - a_{0i}) + \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} \frac{\partial^2 \chi^2}{\partial a_i \partial a_j}\Big|_{a_0} (a_i - a_{0i})(a_j - a_{0j}) + \ldots \ .$$

This may be truncated within the domain of convergence and sufficiently close to $\mathbf{a}$, to be written more compactly in a vector form as

$$\chi^2(\boldsymbol{a}) \approx \chi^2(\boldsymbol{a}_0) - \boldsymbol{b}\cdot(\boldsymbol{a} - \boldsymbol{a}_0) + \tfrac{1}{2}(\boldsymbol{a} - \boldsymbol{a}_0)\boldsymbol{A}(\boldsymbol{a} - \boldsymbol{a}_0),$$

where $b_i = -\dfrac{\partial \chi^2}{\partial a_i}\Big|_{\mathbf{a}_0} = -(\nabla \chi^2)_i|_{\mathbf{a}_0}$ is the negative gradient vector of χ^2 in parameter space, and $A_{ij} = \dfrac{\partial^2 \chi^2}{\partial a_i \partial a_j}\Big|_{\mathbf{a}_0}$ is the so-called Hessian matrix, both evaluated at the point of expansion. This expansion is a quadratic form. It is well known that near an extremum (where the gradient is zero) the quadratic form is the lowest non-trivial approximation to a

function. We can see the family resemblance of the Levenberg-Marquardt method to Newton's method by the following considerations. From the expansion, the general form of the gradient of χ^2 can be readily calculated as $\nabla\chi^2(\boldsymbol{a}) = \boldsymbol{A}(\boldsymbol{a}-\boldsymbol{a}_0)-\boldsymbol{b}$. Newton's method could be used to find the extremum in a seemingly straightforward way. If the gradient is set to zero and the point **a** thus regarded as the extremum point, one simply has $\boldsymbol{A}(\boldsymbol{a}-\boldsymbol{a}_0)=\boldsymbol{b}$, with solution $\boldsymbol{a}=\boldsymbol{a}_0+\boldsymbol{A}^{-1}\boldsymbol{b}$. As with Newton's method in one dimension, this solution is to be regarded in an iterative sense, since the Hessian matrix and the gradient vector **b** are both evaluated at the expansion point and may not be representative of the behaviour of the function all the way to its extremum. If the expansion terms retained are sufficient, the solution may be a better approximation to the extremum than was the original point $\mathbf{a}_0$. Despite the fact that convergence of Newton's method is slow, we are likely better off now than with the initial guess. And in fact, since a quadratic form fits well near an extremum, the approximation made above is probably now a better approximation, leading to further improvement on the next step. This might lead one to expect that Newton's method would be an adequate approach to general non-linear χ^2 optimization. However, it may be that the initial guess is not in fact close enough to the solution for a quadratic approximation to be valid. In that case, a linear function, rather than a quadratic, is the lowest-order best approximation to the function near any point of expansion. The gradient from that linear function can be followed downslope and in that manner a minimum might be expected to be approached. The problem with this approach is that one does not know *how far* to follow the gradient. It is this problem which the Levenberg-Marquardt process attempts to solve, and in the process bring in the power of quadratic fitting where appropriate.

Descent along the gradient can be examined through the linear approximation $\chi^2(\boldsymbol{a}) \approx \chi^2(\boldsymbol{a}_0)-\boldsymbol{b}\cdot(\boldsymbol{a}-\boldsymbol{a}_0)$ which is not useful in that form since **a** is unknown and there is no particular way to find it. What we can claim is that *if* we had the right elements for **a** and thus for the step $\mathbf{a}\text{-}\mathbf{a}_0$, we could go down the slope to somewhere near the minimum. We would know if we had done so since χ^2 there would be less than at the starting point. We also know that the step must be locally roughly parallel to the negative gradient, that is, to **b**. This can be written for each component with an undetermined constant which will be called (with some forethought) $\lambda\alpha_{ll}$, so that the equation for the l-th component of the step is $\lambda\alpha_{ll}(\mathbf{a}-\mathbf{a}_0)_l = \mathrm{b}_l$. Recalling that the quadratic approach led to the equation $\boldsymbol{A}(\boldsymbol{a}-\boldsymbol{a}_0)=\boldsymbol{b}$, we note some similarity of form between these two equations. The first is equivalent to a diagonal matrix multiplying the step, whereas the second multiplies by the generally non-diagonal Hessian. The dimensions of the elements of **A** and of $\lambda\alpha_{ll}$ must be the same, so if we regard λ as a dimensionless quantity, then the dimensions of α_{ll} must be those of the elements of $\boldsymbol{A}$. Without any loss of generality, we can claim that the α_{ll} could in fact *be* the diagonal elements of $\boldsymbol{A}$, and that if they are not, then λ could be adjusted so as to at least partially compensate. This rather loose approach allows the *combination* of the two equations into one, with λ to be determined. If δ is the identity matrix, we will regard the system $\boldsymbol{A}'(\boldsymbol{a}-\boldsymbol{a}_0)=\boldsymbol{b}$ as in some way containing the desired solution, with $\boldsymbol{A}'_{ij}=\boldsymbol{A}_{ij}(1+\lambda\delta_{ij})$ the elements of the new combined matrix $\boldsymbol{A}'$. We must

now formulate a method involving λ which allows this new system to solve the optimization problem. First we note that two extremes of value for λ allow extraction of the original two equations. For λ=0 we recover the quadratic approach, while if λ is very large, A' becomes diagonally dominant and the linear case is effectively in force. For intermediate values of λ, the new system will have a solution analogous to the Newton's method case: $\boldsymbol{a} = \boldsymbol{a}_0 + \boldsymbol{A'}^{-1}\boldsymbol{b}$. By choice of λ, one can cause this solution to respond to being near a minimum (quadratic) or on a slope far from a minimum (linear). Control over λ will depend on the acid test of getting nearer the minimum: whether χ^2 at the new **a** is smaller than that at the initially-guessed value $\mathbf{a}_0$. If, for some intermediate value of λ, χ^2 has not become smaller, then presumably one is on the slope and an increased value of λ would make the equation correspond better to the local situation. If (as desired), χ^2 has become smaller, then one is likely nearer to the minimum and a closer resemblance of the modified system to a quadratic form is required. This can be arranged by reducing the value of λ. Ultimately, when the value is reduced to near zero, the pure quadratic form appropriate to being near a minimum is used and convergence accelerates. Based on these observations, the 'recipe' for use of the Levenberg-Marquardt technique given in *Numerical Recipes in C* [Press *et al.*, 1992, p. 684] may now be cited. Their version begins optimistically with a rather low value of λ: in practice an optimal starting value must be determined by the user. The recipe is:

1. For a starting guess at the parameters $\mathbf{a}_0$, compute $\chi^2(\mathbf{a}_0)$.
2. Pick a small value of λ (they suggest 0.001 in hopes of starting near a minimum).
3. Solve $\boldsymbol{A'}(\boldsymbol{a} - \boldsymbol{a}_0) = \boldsymbol{b}$ to get $\boldsymbol{a} = \boldsymbol{a}_0 + \boldsymbol{A'}^{-1}\boldsymbol{b}$, then evaluate $\chi^2(\mathbf{a})$.
4. If $\chi^2(\mathbf{a}) > \chi^2(\mathbf{a}_0)$, one has failed to approach the minimum and must increase λ significantly (they suggest a factor of 10) and repeat step 3.
5. If $\chi^2(\mathbf{a}) < \chi^2(\mathbf{a}_0)$, one is nearer to the minimum, so **a** is a better guess than $\mathbf{a}_0$. One must replace $\mathbf{a}_0$ in step 3 by **a**, decrease λ to reflect being nearer the minimum, and redo 3.

In practice, steps 4 and 5 must also include some criterion to determine when to give up. This may be some tracking of successive values of χ^2 so that when it stops changing much, one stops (which is the method recommended in *Numerical Recipes*), or simply counting the number of iterations. The latter is a more practical method for implementation, but should be coupled with examination of output at each step so that the 'run' of χ^2 is monitored and has a reasonable behaviour (see Section 5.g). Lampton [1997] finds greatest efficiency in a simple problem when λ is added directly to diagonal elements of the Hessian matrix **A**. This is referred to as additive damping while the scheme detailed is called multiplicative damping. The multiplicative approach is suggested when the parameters have widely varying scales, and with the large number of parameters used in this study, that is likely to be the case. In both additive and multiplicative cases, Lampton cites multiplicative factors for rescaling λ of 0.1 and 10 as commonly used, although his study indicates that these should not be reciprocal factors for most efficient additive damping. In this study the upward rescaling of λ has been done with factors from 3 to 10 and reciprocals have been used when it has been necessary to decrease λ.

This is an appropriate place for a brief further discussion of the Hessian matrix $A_{ij} = \frac{\partial^2 \chi^2}{\partial a_i \partial a_j}|_{\mathbf{a}}$. Since the form of χ^2 is known, as are the forward model dependencies on the parameters, there is no problem in taking successive derivatives with respect to the parameters to obtain the functional form of the Hessian elements:

$$A_{ij} = \frac{\partial \chi^2}{\partial a_i \partial a_j} = 2\sum_{k=1}^{N} \frac{1}{\sigma_k^2}\left[\frac{\partial y}{\partial a_i}\frac{\partial y}{\partial a_j} - \left(y_k - y(\mathbf{a})\right)\frac{\partial^2 y}{\partial a_i \partial a_j}\right].$$

In principle, this form of the Hessian may be used with the algorithm outlined above. In practice, however, there are motivations for neglecting the second derivative terms. In the application presented here, the derivatives are calculated numerically and this is computationally expensive. This is equally the case if the minimization technique is applied with analytical derivatives [Connors, 1993]. We may be able to convince ourselves that second derivative terms should be small (this is certainly the case in linear modelling where they are zero, but in that soluble case we do not need to use a gradient following technique). Press *et al.* [1992] supply some more convincing arguments, including that the term multiplying the second derivatives is the error of the model at each point which should be randomly distributed with mean zero. They note also that second derivative terms can be destabilizing while not affecting the final result, only the route to it. So in practice the second derivative terms are rarely used in implementations of the Levenberg-Marquardt algorithm. Without them, the algorithm is referred to as a 'half-Newton' method: with second derivatives included it would be 'full-Newton'.

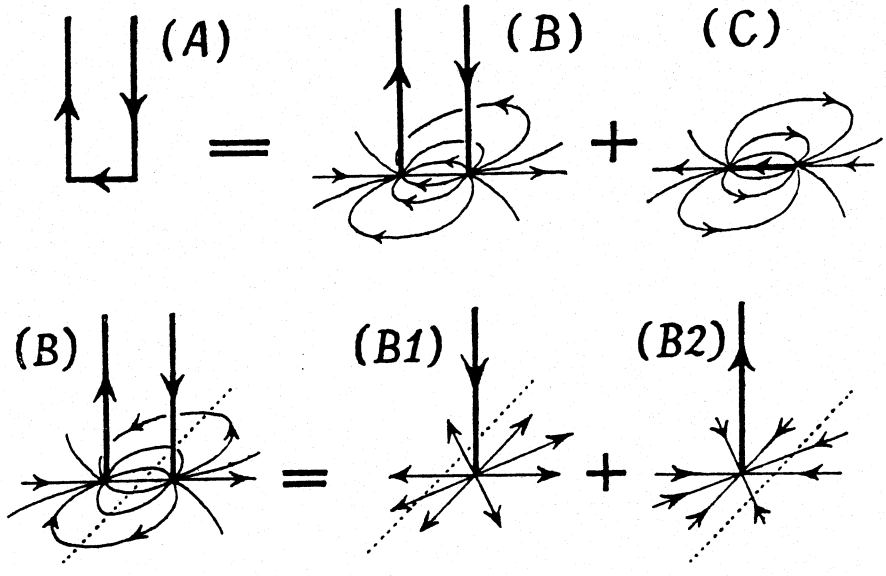
We now proceed to discuss uniqueness of solutions in minimization with ground magnetic data. Other practical aspects of using the Levenberg-Marquardt algorithm to solve problems involving magnetic effects due to current systems are presented in Section 5.g below.

d. Uniqueness

The historic conflict in Space Physics about the nature of the basic current systems causing ground disturbances has been alluded to. Useful insight into why non-uniqueness arises in this case has been given by Fukushima [1969, 1976], although the often-quoted 'Fukushima theorem' is misleading for reasons given below. In Fukushima's illustrative example, field lines are considered to be vertical (a good approximation in the auroral ionosphere), and the ionospheric conductivity is considered uniform and isotropic (a much less correct assumption). The Birkeland current system, a loop consisting of downward field-aligned current, an ionospheric path, and an upward field-aligned current, may have its ground magnetic effects duplicated by a system entirely in the ionosphere with the same ionospheric path and surrounding closure currents, as illustrated in Figure 5.5. Such a system would be similar to the Chapman-Vestine equivalent system previously alluded to. That these produce identical ground effects may be seen by considering the difference currents between the two cases. These amount to the field aligned currents and closure

currents. A set of closure currents may be made by considering each field-aligned current to feed into the ionosphere and flow radially outward in that plane from the point of contact. Each such system produces no ground effect. They are each symmetric about the vertical axis, and this symmetry is at the heart of the 'Fukushima theorem'. The integral form of Ampère's law in this symmetric case states that $2\pi rB = \mu_0 I$, where B (in the azimuthal direction) at radius r is thus proportional to the total current (here I) inside that radius. Below the ionosphere, no current flows and thus the ground-level B from this current system is identically zero. Recalling that this current system is the difference between the Birkeland and the Chapman-Vestine systems, one sees that there are no magnetic signatures of that difference and hence no way to use sub-ionospheric measurements to distinguish the two cases. Although the example is merely illustrative and based partly on unrealistic assumptions, clearly the uniqueness of any model current system put forward must be justified by other considerations than only the ground observations. This simplified theoretical approach is entirely consistent with modelling results for meridional systems as shown in Chapter 4. In the Fukushima example it is *impossible* to determine the current flowing in the field-aligned currents from ground observations. In practice, with poloidal model current systems, the problem is simply very ill-posed and results must be interpreted with great caution and preferably in conjunction with auxiliary data.

Figure 5.5 Fukushima [1969] equivalent currents for a field-aligned current pair. System (B) is the difference current between Birkeland system (A) and Chapman system (C). (B) is equivalent to (B1) plus (B2), neither of which produce ground magnetic fields.



More recently the use of simplified models (such as those discussed in section 4.a.3) by modellers of satellite magnetic observations has led to the belief that the Region 1/2 current systems, characterized by north-south paired sheets of field-aligned current joined by

ionospheric currents, are basically solenoidal in nature and do not produce significant ground effect. The simplified models produce their largest field in the region between the field-aligned sheets and may be used for a reasonable interpretation of field-aligned current based on observations there. Detailed modelling [Kisabeth, 1979], as discussed in Chapter 4, suggests that there may be ground perturbations observable from the Region 1/2 currents, particularly in the polar cap. Simplified models such as those of Fukushima or those used in analysis of spacecraft data would not suggest this to be the case. The more refined models might allow determination of the Region 1/2 current strengths from the ground, but only poor localization of these currents. This is investigated below, in several chapters, where the addition of Region 1/2 currents appears to improve model fit.

In addition to the formal non-uniqueness demonstrated by Fukushima for magnetic problems, there may be non-uniqueness associated with the fitting problem. The investigator searches for a 'best' fit indicating a close resemblance of the proposed model to reality when the appropriate parameters are chosen. It may be relatively straightforward to minimize in parameter space by finding a point where variations are zero but that does not guarantee that the fit is 'best' as there may exist other such points. Such points correspond to local minima in the parameter space. With the effective changes in length scale in the Levenberg-Marquardt technique, such local minima might be expected not to be a great problem. It has been found through experience that a poor choice of initial scale parameter can in fact lead to being caught in a local minimum. Given the large number of parameters involved in magnetic modelling, particularly on a global scale, it is hard to be absolutely certain that local minima are not hindering solution. Such uncertainty must be accepted once minimization with more than a few parameters is attempted. As stated in Numerical Recipes [Press *et al.*, 1992], "this kind of problem is generally quite difficult to solve". Quality of solution, including assessment of the likelihood of having attained a global minimum, must be quantitatively assessed from the behaviour of χ^2 throughout the solution, and from the reasonableness of the final parameters. It is similarly difficult to quantify error bounds when a large number of parameters vary.

e. Error Bounds

The determination of errors in a fitting procedure is difficult for several reasons. If the fit has fully converged to a minimum there is no linear error estimate in at least some parameters since the derivative of the error with respect to at least some parameters is zero at a minimum. If there is also difficulty in telling whether a global extremum has been reached, then there is no real way of knowing whether parameters which do not have zero derivative indicate a slope toward the true solution or simply indicate error. If there are a large number of parameters, as there are in geomagnetic modelling, these limitations prevent much quantitative assessment of error.

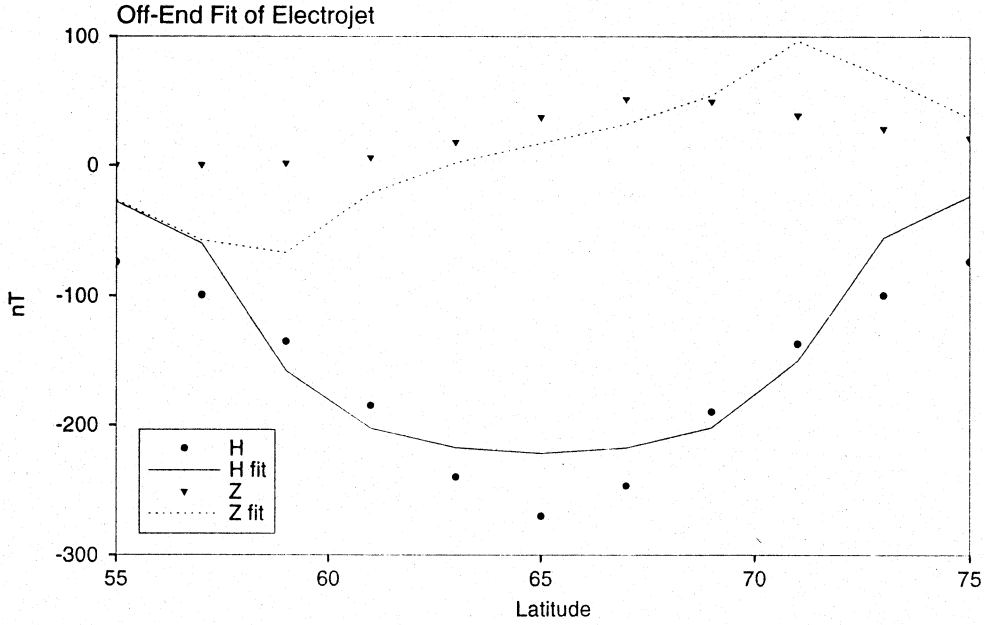
Another aspect of error analysis is the suitability of the forward model itself. If the forward model does not contain elements capable of representing the data, then the optimal model attainable through adjusting its parameters cannot be as good as that of a model which does contain those elements. An example would be an electrojet,

aligned east-west. With such an alignment there is no Y component in the central region. However, it is normal to find a north-south ionospheric current system associated with an electrojet, and such a system produces an eastward perturbation (although usually a small one compared to the X component due to the electrojet). The east-west aligned electrojet forward model would not be an adequate forward model to represent the Y component. Unsuitability of the forward model can also lead to ambiguity. If an electrojet forward model is used with data only from a latitude chain, both the presence of a north-south ionospheric current system and electrojet tilt could cause a Y component to arise. It is difficult to discern and separate the effects of these two sources of Y perturbation with a latitude chain. Thus a latitude chain would lead to ambiguous results. The situation could be made clearer with a two-dimensional distribution of stations (possibly two relatively nearby chains), in which case the electrojet tilt would manifest itself as a change with longitude of the latitude of maximal X perturbation (for example). In the absence of tilt, it would be inferred that north-south currents had produced a Y component. With the known properties of auroral currents, both sources of Y component (electrojet tilt and north-south current flow) would likely be present. With data from a two-dimensional region, the causative currents could be solved for, whereas latitude chain data alone would not permit this.

Yet another source of ambiguity of an electrojet model in a latitude profile is the production of negative X perturbations beyond the ends of a longitudinally limited westward electrojet (and positive X beyond the ends of an eastward electrojet). Such perturbations lead to a spurious equivalent current, giving a false impression that there is overhead ionospheric current. The presence of such perturbations was noted in the modelling of Kisabeth [1972] and they are largely the effects of field-aligned current. In Figure 5.6 such a modelling ambiguity is shown. The perturbation 10° east of the end of a 1 MA 180° electrojet lying between 63° and 67° latitude has been modelled by that from a centred electrojet (also of 180° longitudinal extent) whose latitudinal borders and current intensity were varied, in the manner described later in this chapter, to obtain an optimal fit. This represents an attempt to fit edge effects of an electrojet by a an overhead east-west current. In other terms, an equivalent current is being modelled through overhead current, where there is in fact none. It may be seen in the figure that the general trend of the X perturbation has been very well represented, but by an electrojet whose latitudinal borders (relatively well indicated by Z component extrema) were at 58.3° and 71.5°. The inferred current was 0.39 MA rather than 1 MA. Consideration of only the X component could mislead one into thinking that there is a rather wide (in north-south extent) electrojet overhead but the Z component behaviour has clearly not been well modelled. Rather than having two extrema as is the case for an east-west electrojet, the Z component has one maximum to the north and an ill-defined minimum or simple approach to zero to the south, a result consistent with that of Kisabeth [1972] for systems of smaller east-west extent. Not shown is the large Y component perturbation of the off-end current system, which resembles the forward model Z component. The forward model Y component would be zero in the case an electrojet centred on the latitude profile meridian. To distinguish a situation not well modelled by an electrojet, Z component perturbations must be considered. If

these are small and Y component perturbations are large, the possibility of being off the end of an electrojet must be considered, especially if the X profile is wide. In global modelling as done in later chapters, the longitudes of current ends are allowed to vary (unlike in this example) and this problem should not arise. One should be aware of it in attempting to model perturbations in a meridian, however.

Figure 5.6 Perturbations off the end of a long electrojet, modelled by overhead current. The X component (circles) and Z component (downward triangles) 10° east of the end of a long electrojet have been modelled by a centred electrojet. The X fit is shown by a solid line and the Z fit by a dotted line.



f. Selection Criteria and Data Availability for Events

An initial criterion for event selection was availability of suitable magnetic data from widely distributed stations. In particular, since the method uses mid-latitude stations to constrain positions of field-aligned currents, the classic signatures of the substorm current wedge (bays of duration in the tens of minutes) were sought on midlatitude records. In the final selection, best magnetic data availability and that of other types of data for comparison were found to be from CDAW events. Analysis of the CDAW 9 events by other researchers is not complete and in that sense the analyses presented through automated forward modelling are new for those events. The CDAW 9A event has been most thoroughly modelled by other techniques and those modelling results are discussed for comparison. The CDAW 9B event has been recently modelled by AMIE [B. Emery, private communication, 1995] and global physical parameters are discussed as a complementary aspect of that modelling approach.

More recently, the GEM study of events from early November 1993 has supplied a suitable data set. The generally increasing number of magnetic observatories and the recent ease of interchange of data mean that near real-time inversion of magnetic data could be undertaken with Automated Forward Modelling on a global scale. On a regional scale it should be possible to actually implement real-time modelling. While political and economic considerations which cause gaps to arise in the station distribution (for example the collapse of the Soviet Union) can have adverse affects on any type of modelling effort, there is a continuing trend toward better communication, as exemplified by the rise of the World Wide Web (WWW). This suggests that near-real-time inversion could be done from data sets which come from all parts of the globe. The limit on such a monitoring program is the availability of suitable inversion routines, and this thesis presents one which could be suitable for the task. However, this is demonstrated here using only older data sets.

Data was initially available through specialized CDAW accounts at the NASA Goddard Space Flight Center, accessed by the SPAN dedicated network. Later, CDAW access was through the Internet. The CDAW accounts were eventually superseded by a CD-ROM, although still available for some data sets which were not included on it. Other data were obtained through direct contact with researchers who are listed in the Acknowledgments, and in particular through Barbara Emery at NOAA in Boulder whose gathered data sets were kindly shared. The individual data sets included those of the CANOPUS network. In the most recent months, geomagnetic and related data sets have been most readily accessed through the World Wide Web, which includes search tools which sometimes prove helpful in finding data sources.

g. Optimized Fitting of Magnetic Data

In the most general sense all selected events have in common activity in all of the auroral zone current systems, some of it intense. Perturbations are in each case visible at available mid-latitude stations. Depending on the rough position of the substorm current wedge as evidenced largely by D-component bays at mid-latitude and H-component bays in the auroral zone, stations were selected to give good coverage near the wedge. Distant stations are retained to act as constraints on the current system but need not be as numerous as in the area of rapid spatial change near the wedge. In practice about 30 stations were available for CDAW-based runs; for some events supplementary data were used and the number of stations could be increased to 80 or more. Regional studies can also be done based on 10 or less stations, and have proved particularly valuable if these stations form a latitude chain. The method is completely flexible in terms of number of stations, but judgment must be used to not allow the number of free parameters to become too great in relation to the amount of data available.

The automated forward modelling routine may be used on any subset of magnetic observations for which a reasonable forward model may be proposed. This can in principle include near-Earth satellite data or indeed multipoint observations with the aim of ascertaining the global configuration of the magnetosphere, but severe limitations in the amount of data available prevent such use. In practice, for several of

the events treated below only a global ground-based data set is available with sufficient stations that the forward model may be used to explore details of the current distribution. In others a more general picture must suffice due to insufficient data. In this study, automated forward modelling of satellite perturbation data has not been undertaken and the few passes available are analysed roughly by inspection. In cases where enough data are available, the method may be applied globally and regionally and with models which are fully three-dimensional or essentially two or even one-dimensional. In particular, latitude profiles have formed the basis of many earlier uses of forward modelling [Kisabeth, 1972], and may be readily treated by the new automated technique. A polar-orbiting satellite pass has many similarities to a latitude profile but responds to an associated, but different, current system from that seen on the ground. Both are essentially one-dimensional situations and care is needed in comparing to the three-dimensional modelling. In both satellite and ground-based cases latitudes of current system boundaries are determined and latitude forms the important single independent dimension. In principle tilt of the current system could be determined and this would involve the orthogonal dimension. A forward model could even be proposed which incorporated field-aligned current and was essentially three-dimensional. Obviously, such an approach would have to be used with care not to overinterpret the situation. The effects of field-aligned current can resemble those of tilt of a current system and the ambiguity would have to be resolved by other means. One could also be led to interpret changes in latitudes of current system borders as indicated by one-dimensional modelling in a direct fashion where this is not appropriate. It is impossible to distinguish, in such a case, between actual latitudinal motion of the current systems, and possible east-west motion of a tilted current system. Once again, other means must be called in to resolve the ambiguity.

Having seen that considerable ambiguity can accompany one-dimensional modelling using latitude profiles, it may be useful to consider its advantages. Actual current flow through a meridian is readily established by such modelling if the latitudinal coverage is sufficient. In terms of automated modelling one can start with a poor fit to the data (as an initial condition, possibly derived by some very rough procedure which is readily automated) and arrive at a good fit. This is largely due to the reduced dimension of the modelling space. The three-dimensional approach has a very complex topology in a fit parameter space having typically about 30 dimensions. Initial conditions too far from the "actual solution" may not arrive at a globally optimum situation. Having reliable local conditions is a very effective way of constructing useful initial conditions which are derived from the data. In this way one can use initial one dimensional modelling along a meridian to form initial conditions for global modelling.

The use of weighted χ^2 (chi-square) as a fitting criterion implies that weights must be chosen. The absolute value of the weights can easily be seen to be of little importance as long as machine precision limits are not compromised by their choice. The relative values of weights determines to what extent goodness of fit at various stations creates a depression in parameter space. The programme attempts to find minima and will 'steer into' such depressions. Since auroral zone stations have larger perturbations (i.e. signal) but

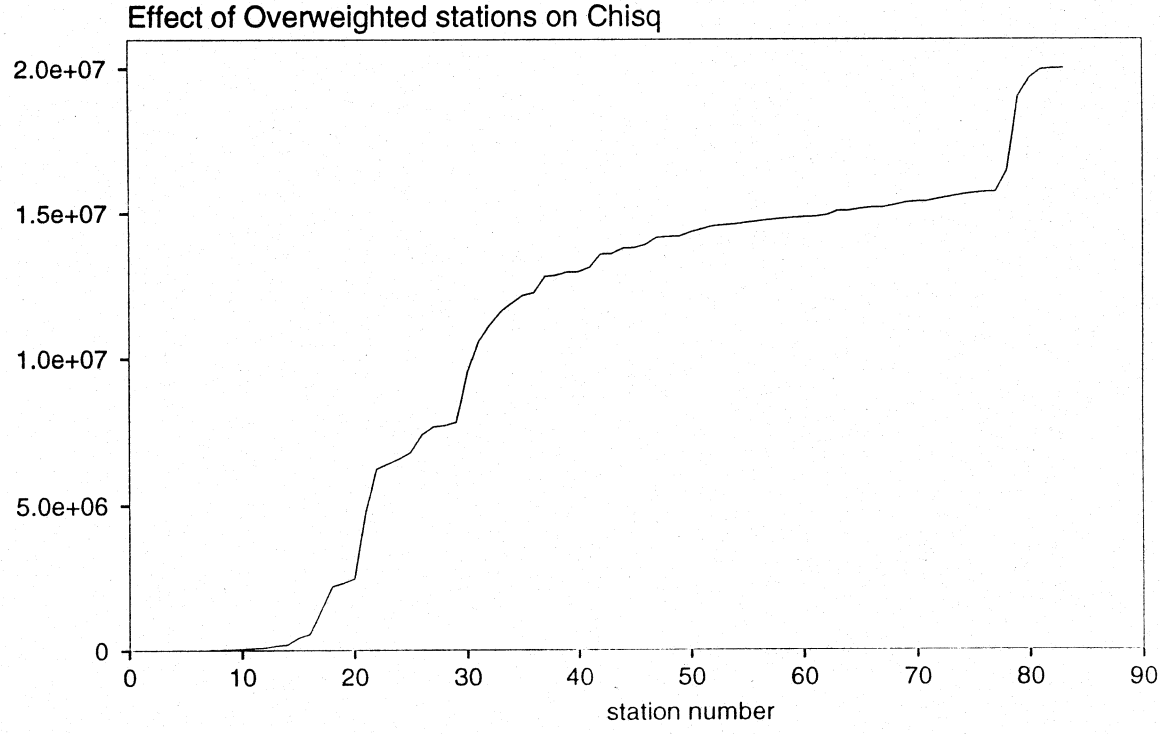
the events treated below only a global ground-based data set is available with sufficient stations that the forward model may be used to explore details of the current distribution. In others a more general picture must suffice due to insufficient data. In this study, automated forward modelling of satellite perturbation data has not been undertaken and the few passes available are analysed roughly by inspection. In cases where enough data are available, the method may be applied globally and regionally and with models which are fully three-dimensional or essentially two or even one-dimensional. In particular, latitude profiles have formed the basis of many earlier uses of forward modelling [Kisabeth, 1972], and may be readily treated by the new automated technique. A polar-orbiting satellite pass has many similarities to a latitude profile but responds to an associated, but different, current system from that seen on the ground. Both are essentially one-dimensional situations and care is needed in comparing to the three-dimensional modelling. In both satellite and ground-based cases latitudes of current system boundaries are determined and latitude forms the important single independent dimension. In principle tilt of the current system could be determined and this would involve the orthogonal dimension. A forward model could even be proposed which incorporated field-aligned current and was essentially three-dimensional. Obviously, such an approach would have to be used with care not to overinterpret the situation. The effects of field-aligned current can resemble those of tilt of a current system and the ambiguity would have to be resolved by other means. One could also be led to interpret changes in latitudes of current system borders as indicated by one-dimensional modelling in a direct fashion where this is not appropriate. It is impossible to distinguish, in such a case, between actual latitudinal motion of the current systems, and possible east-west motion of a tilted current system. Once again, other means must be called in to resolve the ambiguity.

Having seen that considerable ambiguity can accompany one-dimensional modelling using latitude profiles, it may be useful to consider its advantages. Actual current flow through a meridian is readily established by such modelling if the latitudinal coverage is sufficient. In terms of automated modelling one can start with a poor fit to the data (as an initial condition, possibly derived by some very rough procedure which is readily automated) and arrive at a good fit. This is largely due to the reduced dimension of the modelling space. The three-dimensional approach has a very complex topology in a fit parameter space having typically about 30 dimensions. Initial conditions too far from the "actual solution" may not arrive at a globally optimum situation. Having reliable local conditions is a very effective way of constructing useful initial conditions which are derived from the data. In this way one can use initial one dimensional modelling along a meridian to form initial conditions for global modelling.

The use of weighted χ^2 as a fitting criterion implies that weights must be chosen. The absolute value of the weights can easily be seen to be of little importance as long as machine precision limits are not compromised by their choice. The relative values of weights determines to what extent goodness of fit at various stations creates a depression in parameter space. The programme attempts to find minima and will 'steer into' such depressions. Since auroral zone stations have larger perturbations (i.e. signal) but also more

'noise' in the sense that deviations from an ideal current system (as presented by the forward model) are nearby and therefore 'seen', they are weighted less than are mid-latitude stations which have smaller perturbations but also do not respond significantly to small irregular auroral zone variations. In practice the mid-latitude signal is about one-tenth of that typical in the auroral zone. To obtain about the same relative errors in the weighted χ^2, they are weighted about three times as much as are auroral zone stations. Sub-auroral zone stations received a weight about two times that of auroral zone stations. Depending on their strength and desire to attempt to model them, polar cap signals were weighted either the same or less than auroral zone stations. Figure 5.7 shows a practical way of determining that station weighting has been done properly by illustrating a case where it was not well done. The behaviour of χ^2 with station number (out of 83) is shown from an actual modelling run. It may be seen that it shows step-like behaviour with rather less variation in between the steps. This suggests that the stations at the steps are contributing too much to χ^2 and that those elsewhere may not be contributing enough. Such a diagram can be used to guide the choice of weights. It is not entirely clear that the weights would need to be changed in even such a case, however. The χ^2 will depend on deviations of the model fields from the observed fields, and these latter can vary enormously from station to station. In an early stage of the fitting process, a pattern such as that shown might even be expected and desired. Toward the end of the process, the χ^2 pattern should be smoother. Also, these considerations will affect the efficiency of the fitting computations, but in principle a fit can be reached despite poor weighting.

Figure 5.7 Behaviour of χ^2 as a function of station number.



To demonstrate operation of the fitting algorithm, a test with a simple current system is first described. The fitting parameter λ will be shown explicitly as it is changed by the algorithm according to the steps described in Section 5.c. The simple system consists of downward current along a meridian, ionospheric flow at constant latitude in a belt of finite latitudinal width, and upward current flow along a second meridian, as illustrated in Figure 5.1. Due to the simple nature of the system, parametrization in terms of longitudes, central latitudes, and latitudinal width of the current systems was used. In most other work presented here, the parameters (as presented in Table 5.1) are instead the geometric corners of the quadrilateral region delimiting current flow in the ionosphere. Table 5.3 shows the parameters of the test system and those of the initial guess. This current system is rather short in longitudinal extent, only 20°, and is of 5° north-south extent in the ionosphere, centred at 70°. This is very similar to current systems studied by Kisabeth [1972] and illustrations are given there of the perturbations from such a system. The guess deviates by several degrees in all parameters except one of the longitudes of upward current. The guess has about one sixth the total current of the target solution.

Table 5.3 Parameters of test (target) system and initial guess

	Long. Down 1	Long. Down 2	Latitude Down	Width Down	Long. Up 1	Long. Up 2	Latitude Up	Width Up	Current (MA)
Target	40	40	70	5	20	20	70	5	0.100
Guess	42	50	65.2	8	25.2	20	66	11	0.016

Figure 5.8 Chi-squared on linear scale throughout test convergence

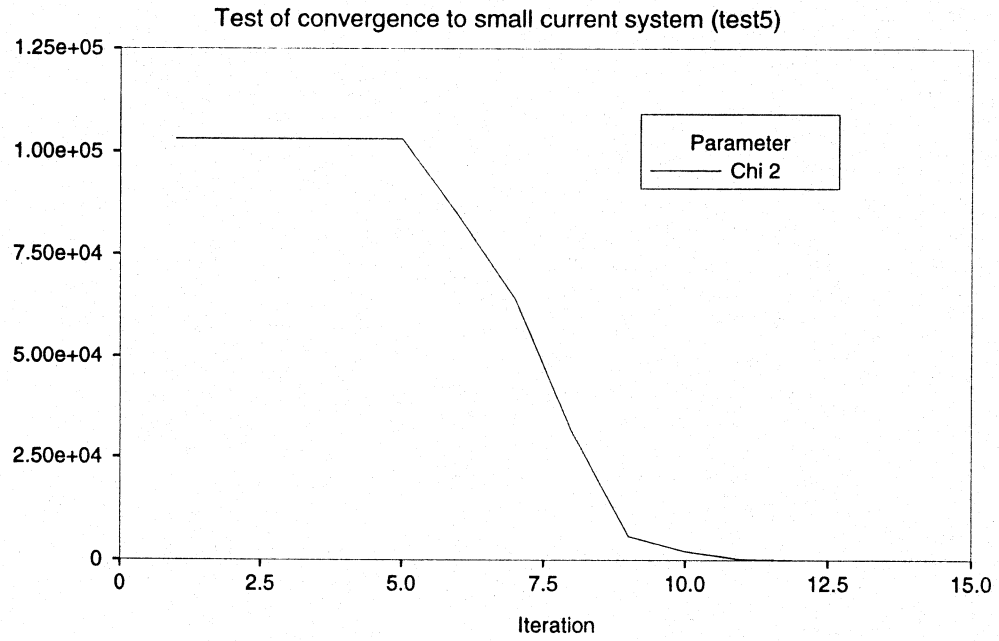
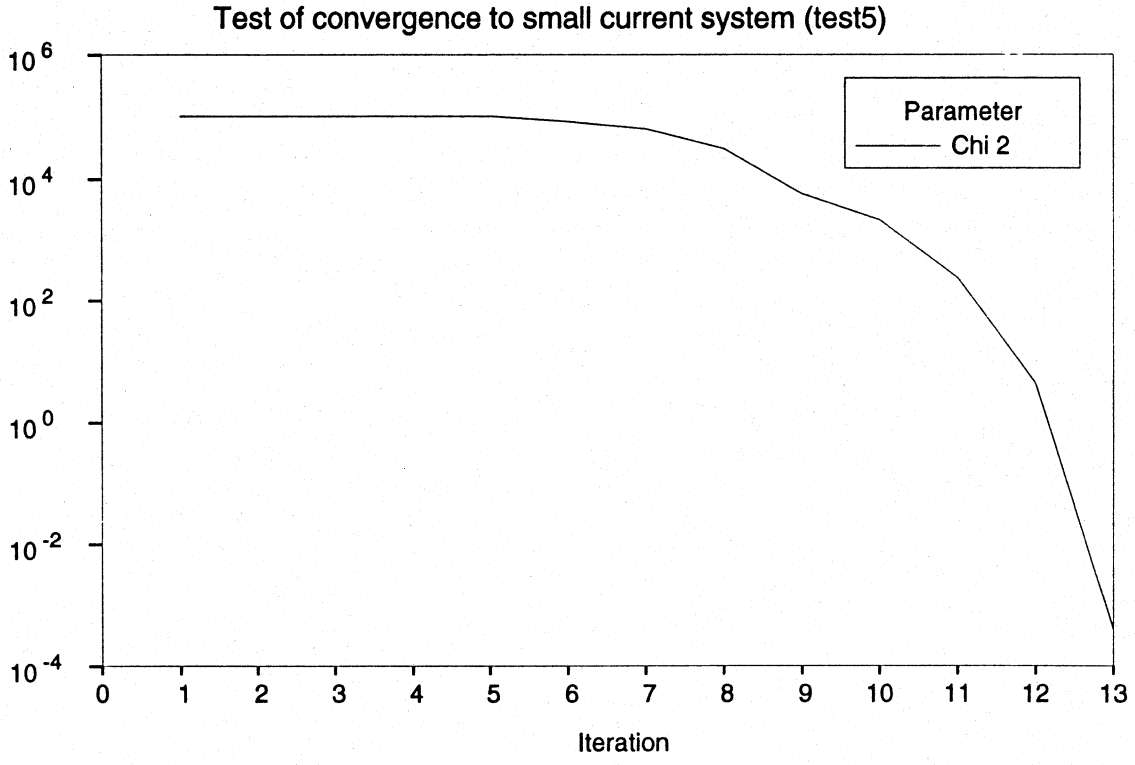


Figure 5.8 shows the values of χ^2 throughout this short run. Stations were selected on a grid covering the current system with one degree spacing in latitude, two degree spacing in longitude, and equal but arbitrary weighting was used at each station. χ^2 is thus an arbitrary number and its behaviour is more important than its actual value. It is seen to have values which initially remain as high as the starting value and after 5 iterations begin to decrease, going down by a factor of 10 after 4 more steps and then becoming too small to easily see on this figure.

Figure 5.9 better illustrates the behaviour of χ^2 late in the convergence of this test case. It may be noted that the convergence, once it actually starts after step five, is better than exponential (which would be a straight line in this logarithmic graph). Also, χ^2 goes to extremely low values, reflecting the fact that there is no noise in this test run and that the forward model's parameters are exactly those of the 'data' and thus can be optimized almost arbitrarily well.

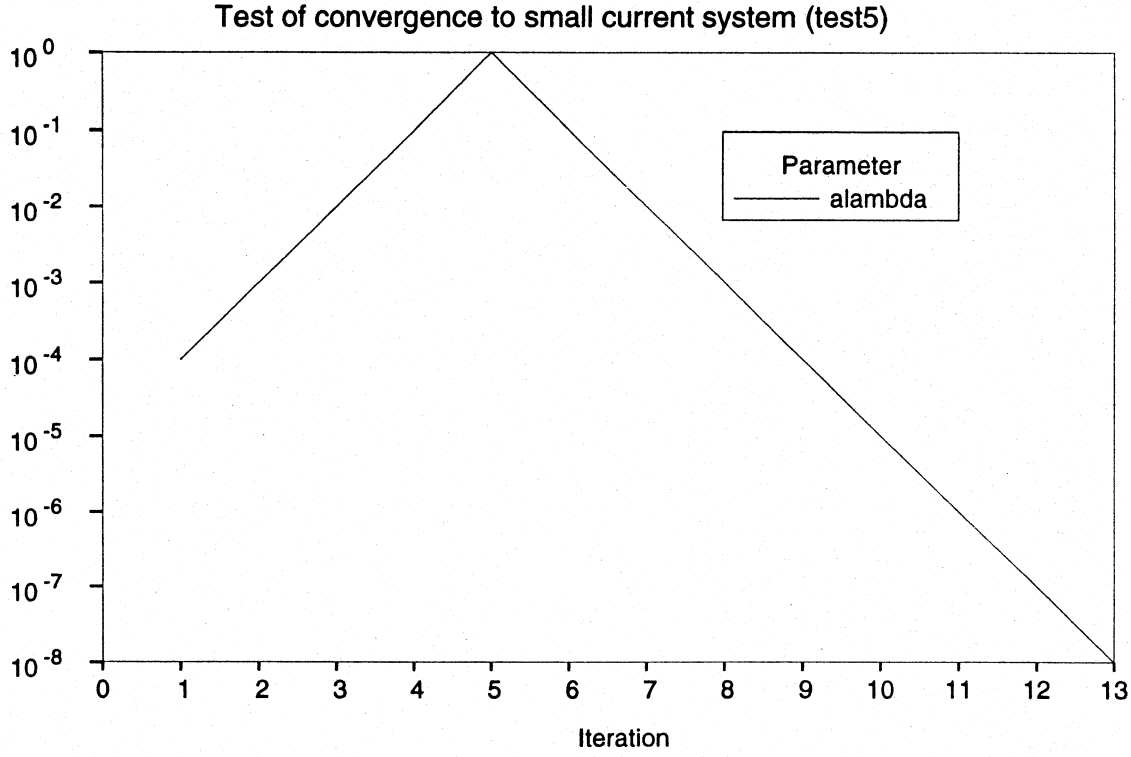
Figure 5.9 Logarithmic graph of χ^2 in test convergence.



To better understand the mechanics of algorithm operation, the behaviour of the scale parameter λ must be examined, and is shown in Figure 5.10. As mentioned in Section 5.b, the authors of *Numerical Recipes* [Press *et al.*, 1992] seem to take a rather optimistic view of how close one is starting to a minimum, and recommend using a rather low value of λ initially so as to exploit the curvature (quadratic) aspects near one. In this case, initial parameter values are in fact rather far from those of the target, as is in part evidenced by

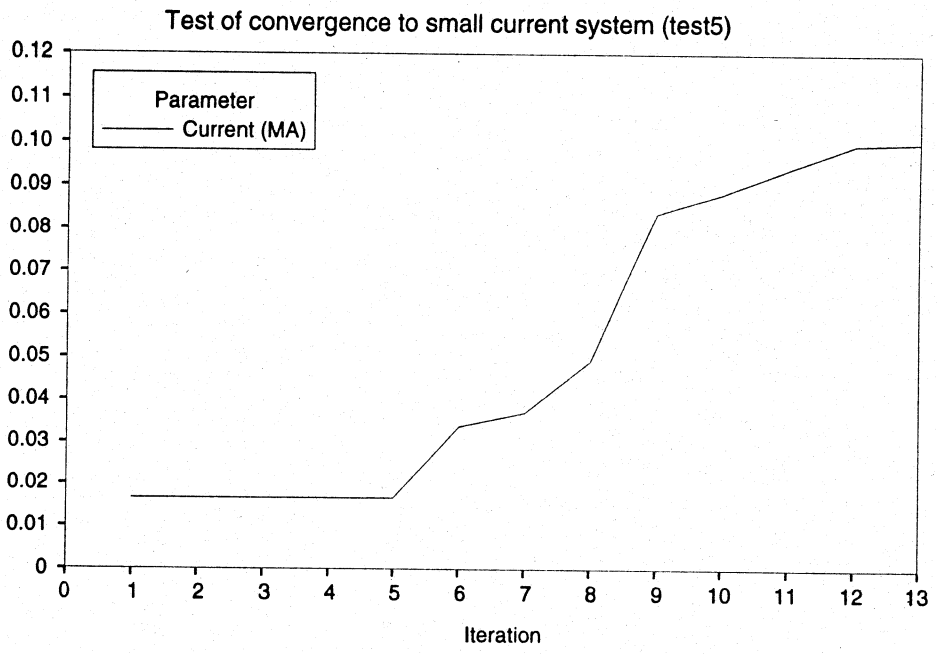
the huge decrease in values of χ^2 through the run. At the initial stages, one is far from a minimum and the quadratic terms, brought into play by an inappropriate choice of the scale parameter, prevent moving toward the minimum, rather than helping. This is not directly seen in the graphs but may be easily inferred by recalling that if the proposed new parameter vector results in a higher χ^2 value, then it is rejected, and λ scaled up by a significant factor. The retention of a constant χ^2 for the first four steps as seen in Figure 5.9 or 5.10 simply means that the proposed new parameters, derived from solution of the **A**' matrix as detailed in Section 5.b, were rejected, and the old ones, with the accompanying value of χ^2, retained as the so-far-best guess. It may be easily deduced by inspection of Figure 5.10 that the 'significant' factor by which λ was scaled up was 10. Once λ has been adjusted to a value of one, at step 5, the χ^2 begins its steady decline, and λ is also adjusted downward, appropriately now as the new guesses are increasingly close to the minimum. Once 'over the hump' of choosing the right scale factor, the solution never slows down. The algorithm can rescale λ upwards once more if progress is not sufficient in some less ideal case, in hopes that such a re-emphasis on linear terms helps out.

Figure 5.10 Value of scale parameter λ during convergence of test case.



The steady decrease of χ^2 might lead one to expect that the physical parameters would all show similar steady improvement toward their final target values. In the case of the current, perhaps because it is a linear parameter in the system, this is indeed the case, as is shown in Figure 5.11.

Figure 5.11 Convergence of total current in system from poor to good value.



This parameter, initially far too small, is essentially doubled at the first effective step, and after that increases steadily, especially from step 8 to step 9 after (as will be seen) the other important system parameters of current latitudinal width and latitude have been brought near their target values. These nonlinear parameters vary in a more complex way during the optimization, as may be seen in Figure 5.12, which shows the latitudinal width of the system. Here also, nothing happens until the first effective step (step 5) and then changes occur in what we might initially consider to be the *wrong* direction, with the current widths, already too large in the guesses, made even larger. In a multidimensional nonlinear system it is sometimes difficult to visualize what is happening, which is a large part of the reason that we need algorithms like the one under discussion to help us navigate, but here we can usefully speculate. The initial latitudes (see Figure 5.13) were initially too low, and in fact the initially guessed system hardly even overlaps the target system. These are brought in the correct (northward) direction at the first step, the widths were increased, and the current was increased. The net result is that on average less magnetic perturbation was felt south of the target system where there was too much from the initial guess, more in the vicinity of the target system where there was initially too little, and the chi-squared decreased although some individual parameters might have become 'worse'. It is left to later steps (better due to being nearer the minimum) to correct this.

Figure 5.12 Uneven convergence of system width parameters toward target values.

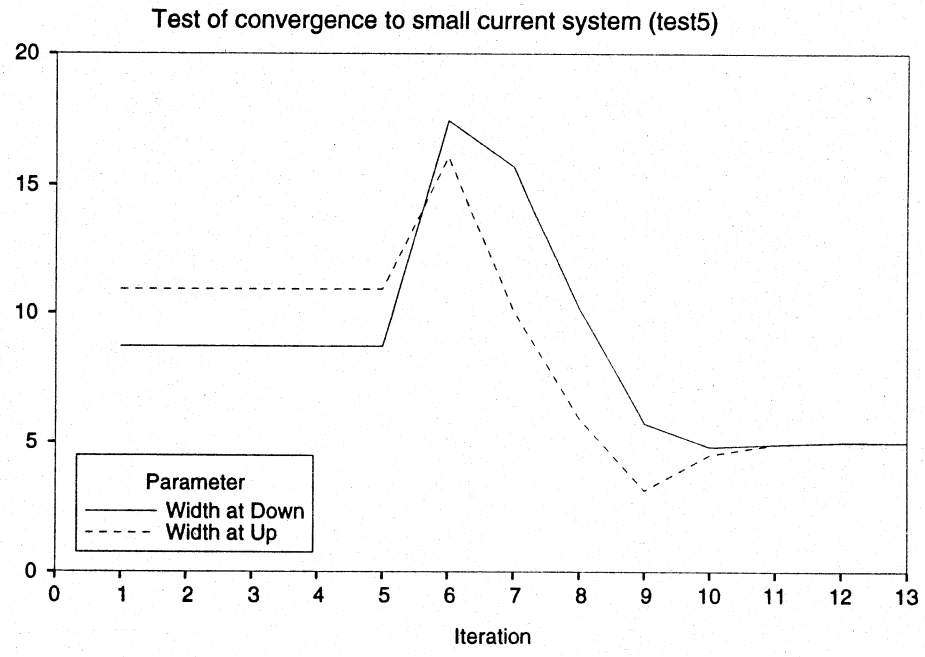
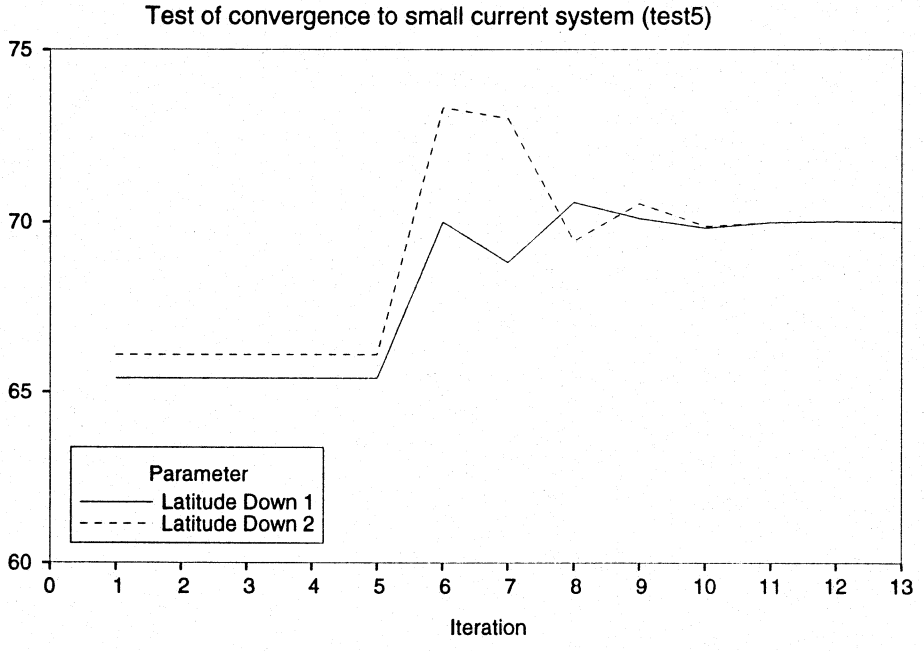
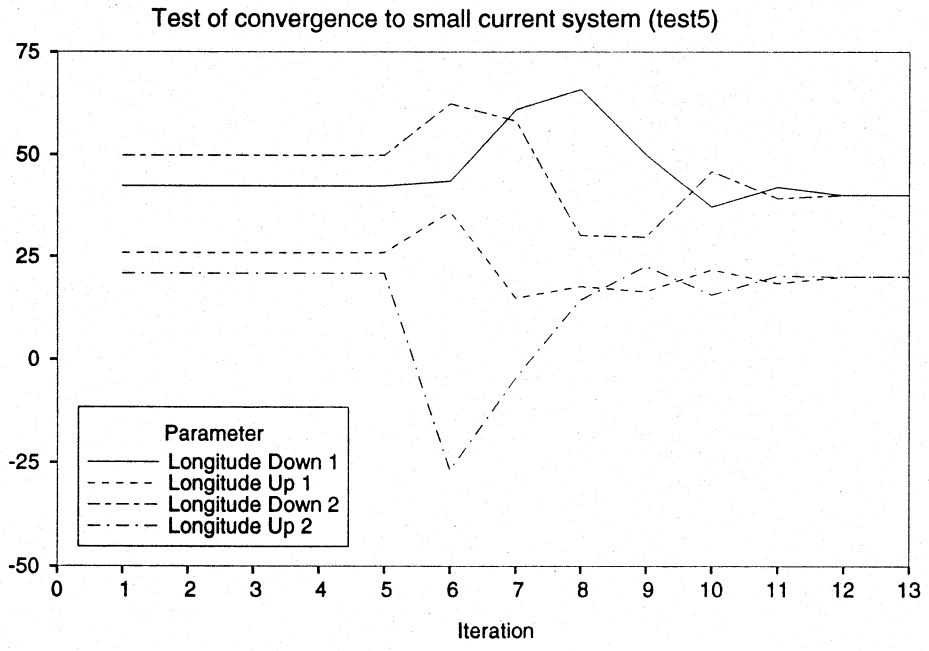


Figure 5.13 Variation of latitudes of centre points of up and down current sheets during convergence.



The latitudes of the midpoints of downward current sheets at either end of the overall system are shown in Figure 5.13. Some overshoot occurs, particularly at the 'up-current' end. Although the target system is symmetric, the initial guess is not, and the nonlinear behaviour seems to have magnified that asymmetry at intermediate steps. By step 8 the latitudes seem to have reached values close to the target values. At the following step, both the widths and the currents make large changes to near their final values. It seems, not unreasonably, that the very important parameter of latitude had to have about the right value before the current and width can settle in on their best values. The various longitudes in Figure 5.14 reinforce the idea that overshoot can occur and that at intermediate steps certain individual parameters can get 'worse' before getting better. The astute reader will have remarked that, in Table 5.3, the initial parameter for 'Up Longitude 2' was in fact set to the known value of the target solution. It may be noted in Figure 5.8 that this 'fix' seems quite irrelevant: this value deviates widely from the initial (i.e. correct) value at steps 6 and 7, returning and staying near the target value at step 8. One's concepts of how to converge on a good solution must be relaxed considerably. This is one of two related problems faced in attempting to solve this sort of optimization manually [Connors *et al.*, 1991]. Changing individual parameters in the solution space can be quite misleading in itself, and the calculation of gradients in the space to guide such a choice (either explicitly or by 'inspection') is very tedious. Letting a proven algorithm do the work, with of course some inspection, is much more productive.

Figure 5.14 Variation of longitudes of extremal points of current sheets during iterations.



The above discussion used one current system and illustrated changes in parameters during a solution, including the changes in the χ^2 and scale parameters which guide the behaviour of the algorithm. Due to this simplicity, once effective convergence started, it continued until convergence to the target resulted. It was stated that the algorithm can adapt to cases in which this does not happen: in effect that a certain amount of navigation around 'corners' in the solution space can be done. An example of this and of a further algorithmic step to double-check the optimality of a solution is now given. From an actual inversion of the data discussed in Chapter 6, Figure 5.15 shows a parameter which is used to track progress of a run and in practice decides when a run is finished. This parameter is a count of convergence failures in the run up to that point. With real data, the forward model will never be perfect, as it was in the simplified example above, and χ^2 will not become arbitrarily small. Rather, the minimum of χ^2 will have some finite value and once at that value, if it corresponds in fact to a global minimum, further reduction is not possible. Thus, further attempts at convergence will fail. For this reason it is useful to keep track of the number of failures to converge, since that may indicate that it is not productive to continue to attempt to optimize. It may also occur that there is never any improvement over the initial guess, and in that case the whole initial guess may need modification. Once the number of failures to converge exceeds five, the routine is once (only) restarted with a much larger value of scale parameter to ensure that one is not trapped in a local minimum, embedded within a larger scale parameter space slope which leads to a global minimum. After this check, and possible further convergence resulting from it, the result is accepted as a potential solution and calculations stop.

Figure 5.15 Count of convergence failures in a full data run of the optimization. Decimal part of iteration number on ordinate is the number of iterations done. Abscissa indicates number of failures to converge since initializing the scale length.

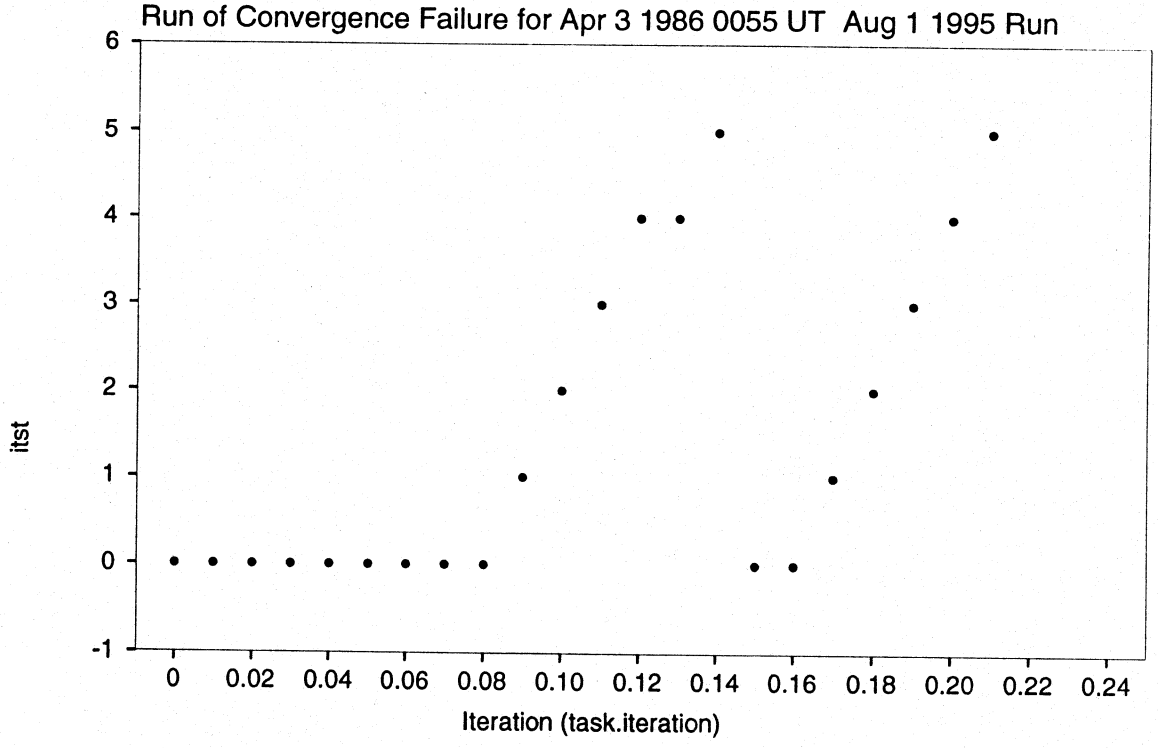
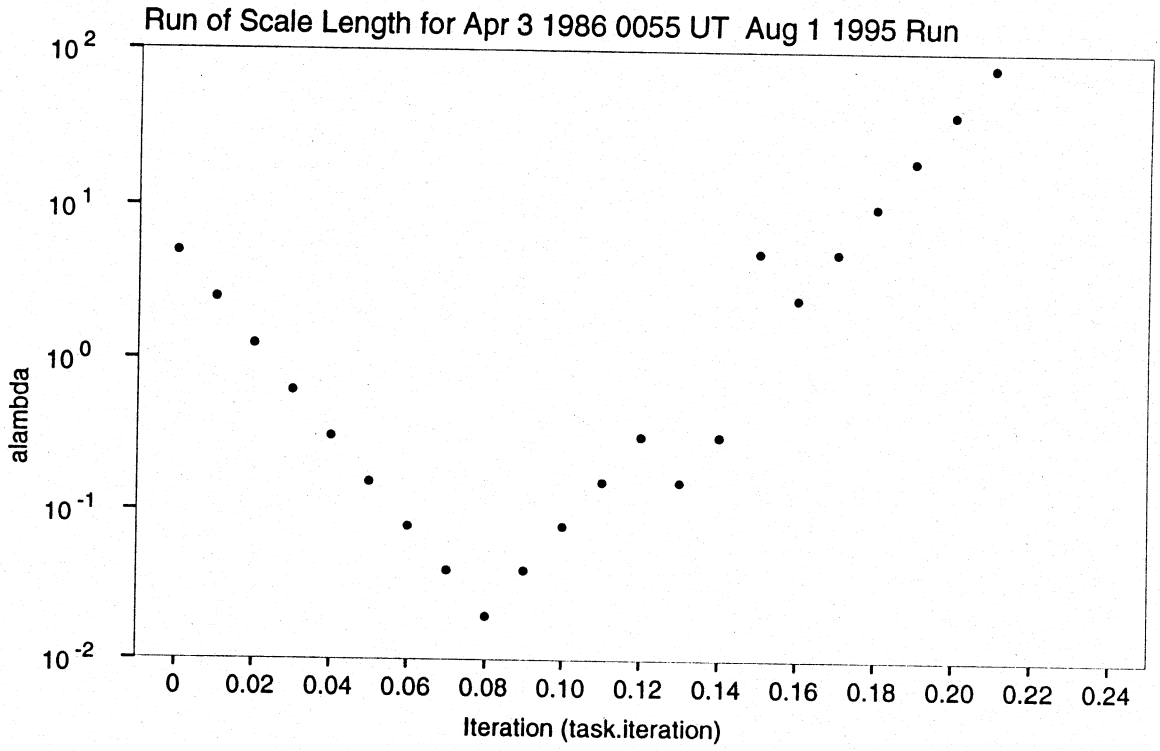


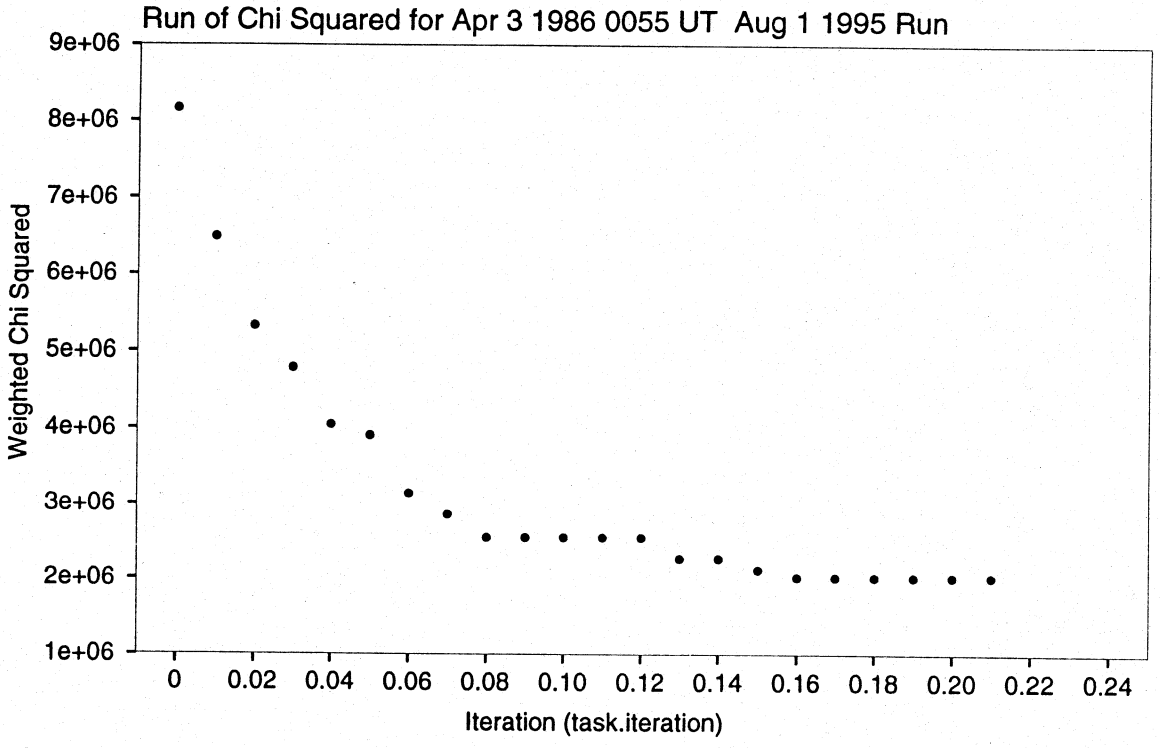
Figure 5.16 Scale length parameter λ in an actual program run. Ordinate labeled as in Figure 5.15. λ decreases each time convergence succeeds, increases on failure.



The convergence failure count is best interpreted in conjunction with the behaviour of scale length shown in Figure 5.16. The initial lack of convergence failure means that useful convergences were being obtained; that is, that the new parameter values were closer to the minimum than the guesses preceding them and thus were accepted. The scale length in such a case is to be decreased and this is seen to be what took place. Experience has shown that in these problems the initial guess is not very good and that a large value of λ (in fact 10) is appropriate. The best value for the 'large factor' by which λ must be scaled proves to be 3. Improvements continued to be made, and no failures recorded, until step 8. At this point a failure to improve χ^2 occurred, and λ was scaled up, to a larger value in case convergence failure occurred due to being on a slope rather than actually near a minimum. This did not improve the situation except at step 12, and after 5 such attempts it was concluded that one is in fact near the optimal solution. In case the scale is still not large enough to detect being in a local minimum on the edge of a large slope in parameter space, a new set of iterations is attempted with the much larger value of λ that was initially used with the first guess. This does produce a small improvement, so the failure count in this second run through remained 0 and λ was decreased as per the recipe. However, no further improvement occurred even when λ, after one failure, was once more increased in value.

After a string of five failures to reduce χ^2 with yet larger values of λ, it may be safely concluded that the solution had been attained since the slope in χ^2 space on many scales was zero. Figure 5.17 shows the initial large decline in χ^2 and its subsequent flattening when near the optimal set of parameters. The set of parameters at the end of such a run may with some confidence be accepted as being a best fit of the forward model in use to the data.

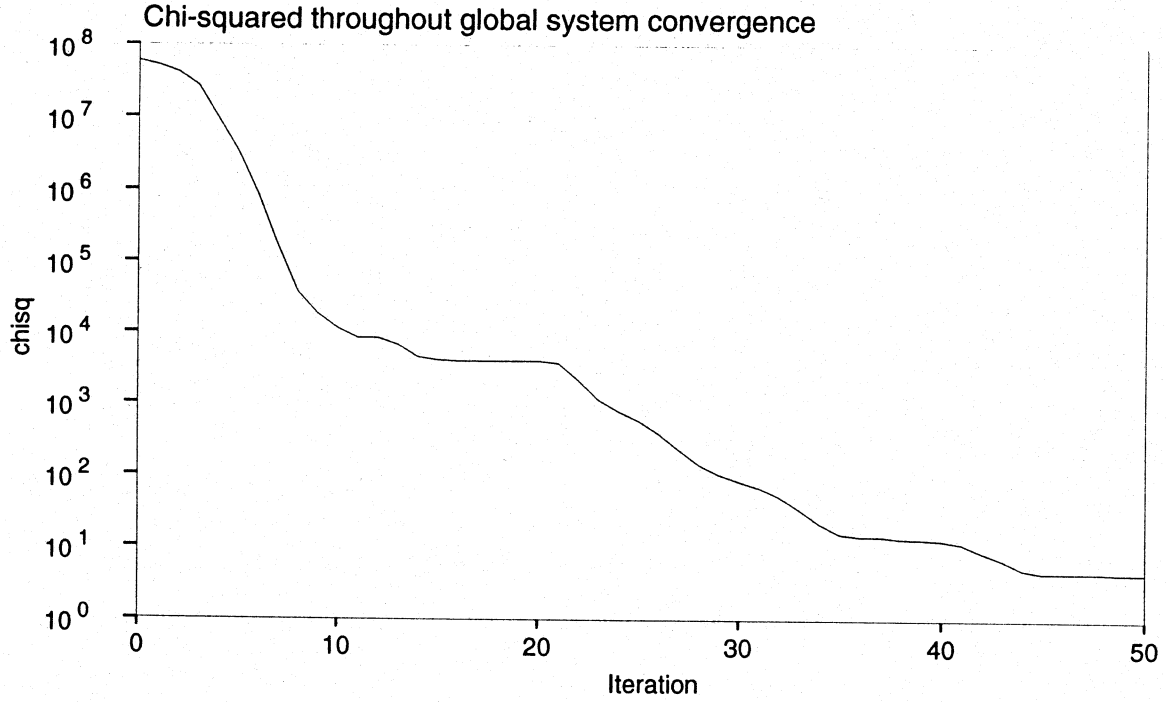
Figure 5.17 Behaviour of χ^2 through a real optimization run. Ordinate labeled as in Figure 5.15.



As has been discussed above, generally nine parameters are needed to characterize each toroidal current system, and generally three systems are used to represent global perturbations. Thus, at least 27 parameters are needed to represent the data. At least one other ('D_{st}-like') parameter must be added to represent low latitude perturbations, likely due to external current sources. There may also be free parameters associated with the poloidal current systems, although in most cases one would use colocation (see above) to reduce the number of parameters associated with these systems. By constraining the geometric parameters to be close to those of the toroidal systems, only the current in poloidal systems remains as a free parameter. This means that 31 parameters should suffice to represent the global current systems. In principle, with three components per station, about 10 stations should suffice (with essentially a one-to-one parameter remap from magnetic data space to current parameter space) to determine these 31 parameters. In practice, for global modelling, more than 30 stations must be used in order to have enough data to determine these parameters with some degree of accuracy in global modelling. Although each station provides 3 components of data, some stations are not well placed and contribute little information. In the most complete global modelling presented in this

work, 80 or more stations are used. The good fits attained imply an economization by about a factor of 8 through use of current parameters as opposed to the raw data. On the other hand, the 31 (or more) dimensional current parameter space must be considered 'large-dimensioned' in several practically important fashions. Visualization of data and model parameters in this space is itself a challenge. Further, it is obviously difficult to be sure that any extremum found in the space is truly a global extremum. Even crude tests of the extrema, such as taking steps away from a purported minimum, are difficult given the large dimensionality of the space. For these reasons, it has been pointed out that one must carefully survey the steps taken to get to a minimum, and the overall behaviour of the programme while it does its optimization. Further, one must understand the operation of the Levenberg-Marquardt algorithm, as described above, to interpret the behaviour. It has been shown that one should not attach too much importance to the changes in individual parameters as the optimization proceeds, provided that both χ^2 decreases in a systematic fashion, and that the final solution produces perturbations which are comparable to the input data. As a final example/test in this section, a 'poor' initial guess is used with an idealized yet realistic forward model. This demonstrates that an initial guess can be far from the target and that the routine can still obtain convergence, in other words, it is a demonstration in principle that the method under discussion works well on the type of data available.

Figure 5.18 χ^2 throughout convergence of a realistic but 'poor' initial model.



In this test, 27 parameters were used to simulate the Hall currents associated with an idealized auroral oval current system consisting of morning and evening sector electrojets and an active westward electrojet in the midnight sector. Neither a D_{st}-like parameter nor poloidal currents were used in this illustrative example. The surface magnetic fields were calculated at actual magnetic observatory locations using the target parameters. These 'observations' were then used as input to the Automated Forward Modelling routine. Since the main aim here is to demonstrate that a poor initial guess leads to convergence even with many model parameters (in other words that the AFM routine works in the ideal case), the run of χ^2 throughout convergence is shown as Figure 5.18. Much as in the single current case (see Figure 5.9), there is seen to be rapid convergence.

Figure 5.19 shows a global view of current systems put forward as a reasonable representation of the auroral oval currents, yet one which deviates considerably from the target system. The very large initial values of χ^2 seen in Figure 5.18 attest to the fact that the two systems are considerably different. Station weightings varied in the calculation of χ^2 much as in the real calculations detailed in the following chapters, with greater weight given to mid-latitude stations. As mentioned above, the scale parameter is of importance in attaining a solution: when this parameter was set initially to 0.001, no progress was made toward solution. Such a small value is suitable if one knows that the parameters of the initial forward model are very close to those of the solution [Press *et* al., 1992]. Here that is not the case and the curvature terms which are emphasized at small values of the scale parameter dominate when it is not appropriate for them to do so. Progress to solution was obtained by setting the initial scale parameter to 10. This is also the starting value for this parameter which was generally used in applying AFM to real data. The test case solution after success of AFM is presented as Figure 5.20. Currents solved for were roughly a factor of two changed from those in the initial model, and the latitudes are seen to be considerably higher than those of the initial guess (Figure 5.19). The active midnight sector has been found to be at much different longitude that has that of the initial guess, and the noon sector downward current configuration is also considerably different. The final solution's parameters differed only at the hundredth of a percent level from those of the target system, consistent with the very small χ^2 values attained. This constitutes a convincing demonstration that the AFM method works well in the mathematical sense: it is capable of inverting data supplied to it. Further, in this ideal case, the routine is seen to be efficient in terms of number of steps taken to do the inversion.

Figure 5.19 Initial current configuration for realistic global test model.

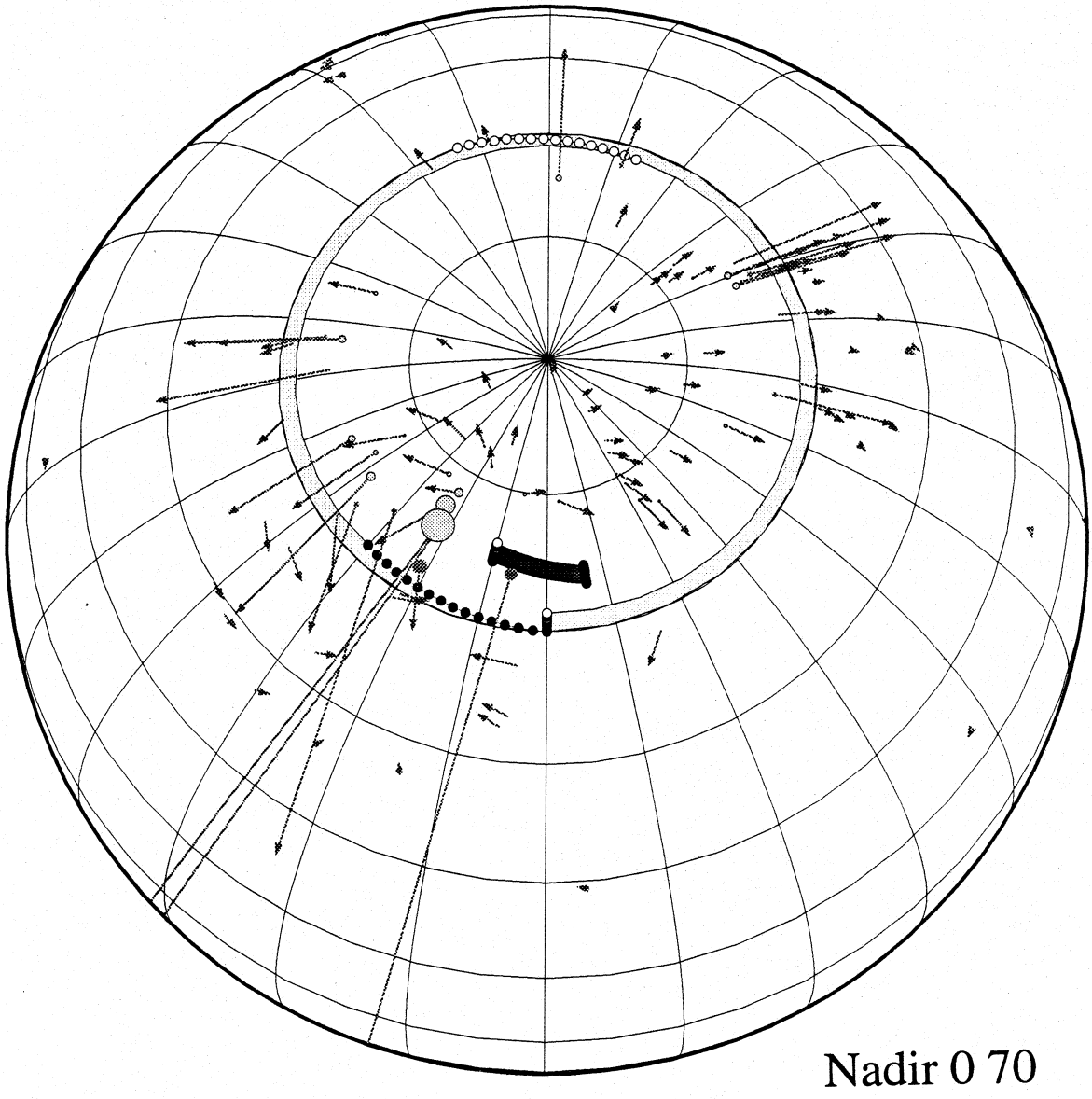
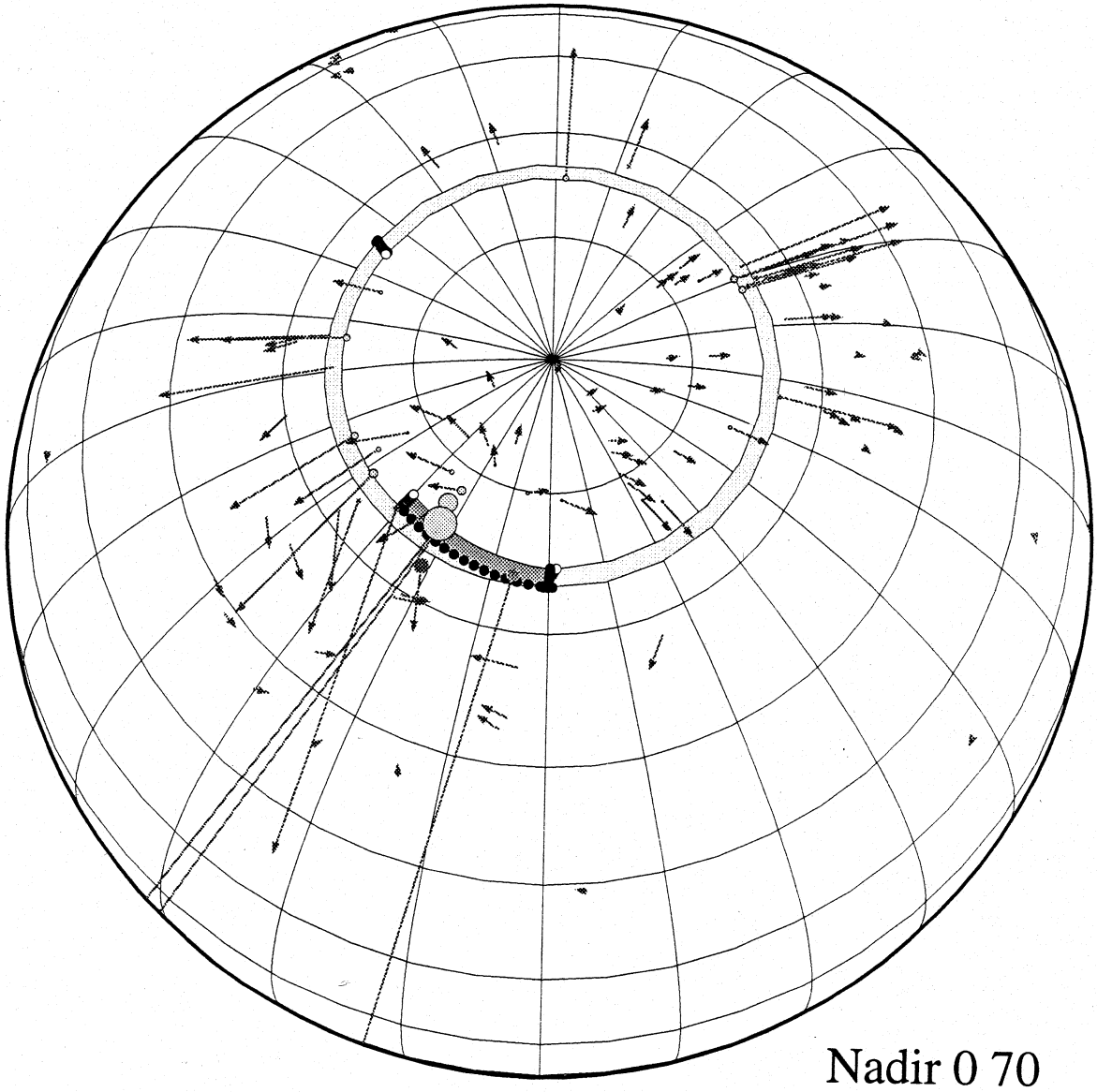


Figure 5.20 Final solution, showing perturbation magnetic field, for AFM global run.



h. Comparison with Other Data about the Events

In several cases in what follows (Chapters 6, 7, and 8), images from a spacecraft (Viking) allow rough determination of particle precipitation regions. Comparison of these regions with those deduced from the magnetic data are particularly useful. In all cases, data from various other spacecraft are available and are used to discuss the larger context of the events, and to a limited extent implications for mapping of field lines. DMSP magnetic and particle data can be used to verify results due to the known colocation of the Hall electrojets which produce most of the ground signal and the Region 1/2 field-aligned currents which produce the perturbations dominant in near-Earth space. In certain cases it is of particular interest to determine whether or not a substorm onset occurred and the deciding factor then may be other signatures of onsets such as near-tail dipolarization or particle injections.

The various types of other data available were discussed in Chapter 3. The forward models presented here are computed using only magnetic field data. Other forms of data have not been incorporated directly into the modelling, but rather used to provide a broader context to discussion of results. In principle the AMIE method (see Section 4.b) allows incorporation of certain other data since the model which is part of that inversion scheme is more physically complete than that used here.