PERGAMON

Journal of Atmospheric and Solar-Terrestrial Physics 61 (1999) 85–100

Journal of ATMOSPHERIC AND SOLAR-TERRESTRIAL PHYSICS

The evolving concept of a magnetospheric substorm

Gordon Rostoker*

Department of Physics, University of Alberta, Edmonton, Alberta, Canada T6G 2J1

Received 26 January 1998; received in revised form 28 August 1998; accepted 19 October 1998

Abstract

A magnetospheric substorm is an episode of energy transport and dissipation in the Earth's ionosphere and magnetosphere which takes place in response to a time limited increase in energy input from the solar wind to the magnetosphere. For the past few decades, scientists have tried to understand the physical processes which take place that are responsible for the substorm disturbances of the geospace environment. In this paper, The development of the substorm concept is reviewed from its origins at the beginning of the 20th century to the present time. The theoretical framework in which substorm physics is normally presented is then discussed, and an outline is given of how that framework has changed in recent times. This paper concludes by posing two questions which need to be answered if further progress is to be made in solving the substorm problem.

1. Introduction

Since the beginning of the 20th century, there has been an ever increasing interest in the origin of the aurora and the often large magnetic field perturbations that accompany displays of the the northern and southern lights. Near the end of the 19th century, it had already become clear that electromagnetic fields were involved in the process whereby the auroras were created; however, it was the work of the great Norwegian scientist Kristian Birkeland that really marked the start of the modern age of solar-terrestrial studies using the techniques of mathematics and physics. In his pioneering studies, Birkeland (1908) recognized that significant electrical currents flowed in the upper atmosphere in the region of bright auroras. He further correctly attributed the source of energy to ionized particles coming from the Sun and he carried out detailed modelling studies which demonstrated that the current system responsible for the observed magnetic perturbations was of the form shown in Fig. 1. The reader will recognize the form of what, today, is known as the substorm current wedge.

During the next three decades there was not a great deal of progress in understanding the way in which the Sun provided the energy for auroral disturbances. In fact,

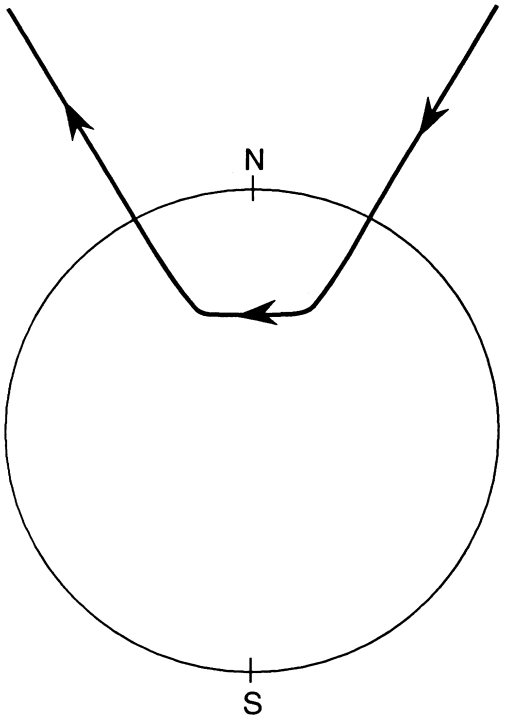

Fig. 1. Three dimensional current system proposed for the polar elementary storm by Birkeland (1908). The polar elementary storm was later named the polar magnetic substorm by Akasofu (1968) and, to this day, is thought to be associated with a three-dimensional current system of the type shown here.

* Tel.: +1-403-492-1061; fax: +1-403-492-4256.
E-mail address: rostoker@space.ualberta.ca. (G. Rostoker)

S1364–6826/99/$ - see front matter
PII: S1364—6826(98)00119—9

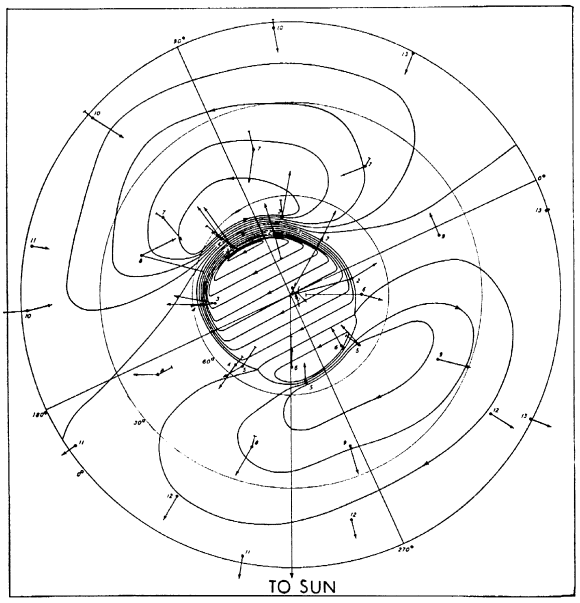

Fig. 2. Equivalent current system for geomagnetic bays inferred from ground magnetometer data by Silsbee and Vestine (1942). The equivalent current vectors are 3-h averages and therefore pertain only to the low frequency component of substorm activity.

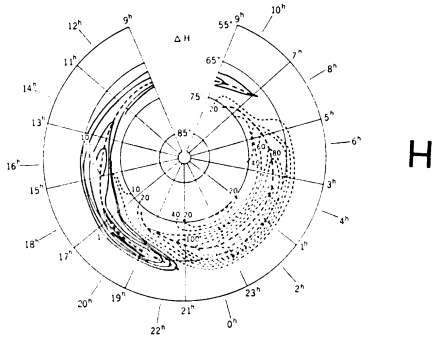

Fig. 3. Magnetic perturbation pattern for high latitude magnetic disturbances obtained by Harang (1946). The evening (morning) sector eastward (westward) electrojet produces regions of positive (negative) H-component. These plots were made with hourly averaged data and hence are dominated by the low frequency component of substorm activity, although not to the same extent as the equivalent current system of Silsbee and Vestine (cf Fig. 2).

it was generally believe that space was a vacuum and that auroral outbursts only took place when ejecta from the Sun arrived at Earth orbit. Even at the beginning of the third decade of the 20th century, this concept was still prevalent when Chapman and Ferraro (1931) carried out their study of the interaction of a plasma stream with the terrestrial magnetic field. Nonetheless, the fact that plasma physics had developed to the stage that it could be used in studies of the Sun–Earth environment was an important step forward which was to have a significant impact several decades later.

During the 1930s and 40s, further progress in understanding the solar-terrestrial interaction came from analysing data from irregularly distributed arrays of ground based magnetometers. The work of Silsbee and Vestine (1942) began the era of the use of equivalent current systems to portray the two dimensional distribution of horizontal magnetic perturbation vectors measured at each observing site. The idea here was that it was assumed that all currents flowed only in the ionosphere and that a infinite sheet current approximation could be used to evaluate the strength and direction of the current flow at each site. Thus, Fig. 2 from Silsbee and Vestine was obtained by rotating the magnetic perturbation vectors by 90° and drawing the current configuration that best fitted the data. From this figure it was inferred that there was a westward electrojet flowing in the morning sector auroral region and a weaker eastward electrojet flowing in the afternoon sector. A similar picture was obtained by Harang (1946), shown in Fig. 3, although here the magnetic perturbation data are shown and the reader is asked to infer that negative (positive) H-component perturbations reflect the presence of a westward (eastward) electrojet. It is extremely important to recognize that the figure of Harang was obtained using hourly averaged values of the magnetic perturbation vectors while Silsbee and Vestine used 3-h averaged values. This will have important implications for substorm controversies, as we shall see shortly.

2. The question of the true nature of the substorm current system

The work of Fukushima (1953) began the era of controversy as regards the character of the equivalent current systems which accompanied high latitude auroral activity. He presented several examples of equivalent current systems taken at times of geomagnetic bay activity which did not conform to the two cell picture of Silsbee and Vestine and resembled more what one would expect from the three dimensional current system proposed by Birkeland (1908). Ultimately, Sugiura and Heppner (1965) highlighted the problem by asking the question of whether or not geomagnetic bays were best described by a two-cell equivalent current system or a one cell system. Around this time, the term, 'substorm,' was coming into use after its introduction by S. Chapman and S.-I. Akasofu in the early 1960s and its use to describe the auroral breakups near midnight known to occur in conjunction with geomagnetic bays (viz the auroral substorm). So it was about this time that arguments began to be made in

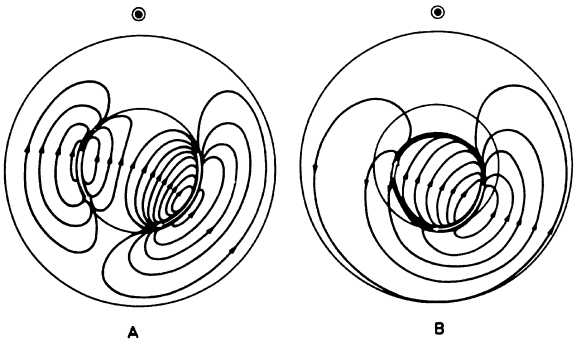

Fig. 4. Schematic diagrams of the two equivalent current systems proposed for the substorm disturbance (after Rostoker, 1996). We now understand that there are disturbances during the time of a substorm, one of which has a two cell equivalent current system and the other of which features the one cell system. The one cell system is more appropriate for the shorter period component of the substorm disturbance.

terms of substorm properties rather than the properties of geomagnetic bays.

Sugiura and Heppner claimed that the substorm disturbance was best described by a two cell system conflicting with the argument by Akasofu et al. (1966) that it was best described by a one cell system. The question was answered by Rostoker (1969) who produced evidence to show that, in fact, disturbances featuring both types of equivalent current system co-existed during substorm activity. Figure 4 shows schematically the two proposed equivalent current configurations, while Fig. 5 shows a set of three magnetograms from northern Scandinavia which demonstrate the coexistence of disturbances which are characterized by those equivalent current systems. It turns out that the short lived (<1-h time scale) disturbances are best represented by a one-cell equivalent current system while the longer lived (shown by the dashed curve on the Tromsø magnetogram) disturbance is best represented by the two-cell equivalent current system. This is why the fact that the equivalent currents inferred by Silsbee and Vestine (1942) and later by Harang (1942) were of the two cell type. Both analyses involved averaging of the data (3 h for Silsbee and Vestine and 1 h for Harang) which would produce results dominated by the lower frequency disturbance.

At the beginning of the 1970s, substorm research received somewhat of a setback for two separate reasons. The first reason related to the beginning of the serious use of indices of auroral zone activity to study individual events. Davis and Sugiura (1966) introduced the AE index as a way to track the level of geomagnetic activity on a global basis. The index involved establishing the upper (AU) and lower (AL) envelopes of the magnetograms traces of the North–South components of the disturbance field from several stations distributed as uni-

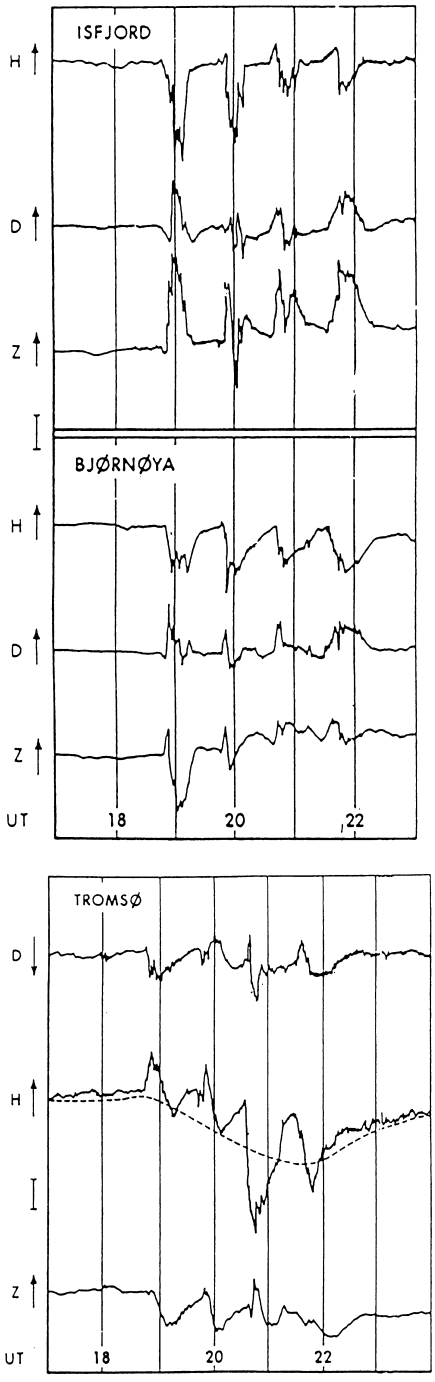

Fig. 5. Magnetograms from the high latitude stations of Isfjord (74.5° corrected geomagnetic latitude CGL), Bjørnøya (70.9° CGL) and Tromsø (66.30° CGL) lying approximately along a common geomagnetic meridian (after Rostoker, 1996). The four disturbances visible between ~1830–2230 UT would be called substorms by most researchers and indeed are characteristic of the polar elementary storm of Birkeland (1908) and the polar magnetic substorm of Akasofu (1968). The longer period disturbance on which the polar magnetic substorms are riders is best represented by a two cell equivalent current system.

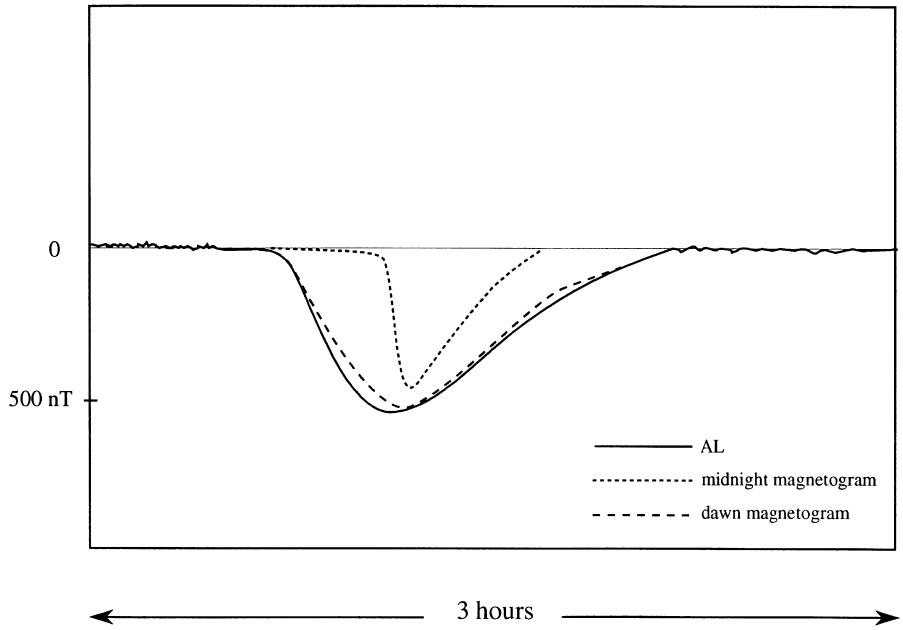

Fig. 6. Schematic diagram of the envelope of auroral oval H-component disturbances representing a typical AL index variation during a period of substorm activity (solid curve). Also shown is an H-component magnetogram from a midnight sector station whose data were used in the construction of this theoretical AL index (dotted line). In this case the peak disturbance of the substorm expansive phase portrayed here is slightly smaller than the contribution to the H-component from the directly driven activity which maximizes a a dawn sector station (shown by the dashed curve). The midnight sector substorm would not be detected using the AL index in this case.

formly around the world at auroral zone latitudes. (The value of AE is the sum of the absolute values of AU and AL at any instant of time.) Although AE, AU or AL are very useful for statistical studies of auroral zone disturbances, researchers began to use the indices in the study of individual events primarily as an indicator of onset times of substorm expansive phases. In doing so, they were studying the combined signatures of disturbances described by the two-cell and one-cell equivalent current systems in a situation where it was impossible to decouple the contributions of the two systems. Since expansive phase effects are best described by the one-cell equivalent current system, increases in the two-cell current system could easily be misinterpreted as expansive phase onsets. More importantly, as can be seen from Fig. 6, an expansive phase onset could be missed completely if the maximum perturbation associated with that expansive phase was less than the maximum perturbation associated with the two-cell equivalent current system. Thus, in the early years of the use of the AE, AU and AL indices it was often not possible to establish from studies that utilized these indices which increases of the index were due to expansive phases of substorms and which were due to the disturbance which featured a two-cell equivalent current system.

A second reason for the setback came from the establishment by McPherron et al. (1973) of the substorm current wedge as a real (rather than equivalent) current system. This approach succeeded in drawing attention to the possible physical mechanisms which might be responsible for the substorm expansive phase, but had the unfortunate side effect of leaving to the side the other component of substorm activity, viz, the disturbance responsible for the two-cell equivalent current system. However, the concentration on the expansive phase disturbance that was best represented by the substorm current wedge led to an enhanced search for knowledge about the phenomenology of that component of substorm activity. During this time it was discovered that the expansive phase current system was not a monolithic large scale current system, but rather that it was a superposition of more local current elements which expanded in a step-wise fashion both poleward (Kisabeth and Rostoker, 1974) and westward (Wiens and Rostoker, 1975) as the substorm developed.

The importance of the two cell equivalent current system re-emerged in the late 1970s in the work of Perreault and Akasofu (1978) who studied the response of the magnetosphere to solar wind input as quantified by the ε parameter defined by

$$\varepsilon = 1_0^2 v_s B^2 \sin^4(\theta/2)$$

where v_s is the solar wind speed, B is the magnitude of the interplanetary magnetic field (IMF), θ is the polar angle measured from the northward geomagnetic axis of the component of the IMF normal to the Sun–Earth line and 1_0 (~ 7 R_E) is a constant with the dimensions of distance. The comparison of $\varepsilon(t)$ with the energy dissipation in the magnetosphere-ionosphere system $U(t)$ led Akasofu (1980) to argue that a large portion of the energy entering the magnetosphere from the solar wind was dissipated in what he called directly driven activity. It turned out that his directly driven activity was manifested in the two-cell equivalent current system observed during substorm activity, and once again some attention was paid to the large scale electrojet currents that flow in the auroral oval (i.e., the eastward electrojet in the evening sector and the westward electrojet in the morning sector). Later, Clauer et al. (1983) used linear prediction filtering to evaluate the impulse response of the magnetosphere and from this work discovered that a substantial fraction of the variance in the AL index (41%) was related to the direct input of solar wind energy (viz directly driven activity). From this research in the late 1970s and early 80s it was established that directly driven activity constituted a significant portion of substorm energy dissipation. However, the ratio of energy dissipation in directly driven activity to that in the release process varies from event to event and from one time in a given event to another time in the same event. Exactly what determines the proportions is not fully understood at this time.

To conclude this section, we reflect on the definitions of the various phases of substorm activity trying to relate the original definitions to how we presently view the substorm phenomenon. The original definition of a substorm by Akasofu (1964) involved an expansive phase in which the auroras moved poleward and a recovery phase in which they drifted equatorward to their pre-expansive phase location. Subsequently, McPherron (1970) introduced the concept of a growth phase during which energy was stored in the magnetotail to be released sometime later in the expansive phase. While the growth phase concept introduced by McPherron is accepted to this day, the actual signatures (which were disturbances measured by ground based magnetometers) were not actually signatures of storage of energy in the tail but rather signatures of the growth of the directly driven electrojets. Fortunately, an increase in the energy input from the solar wind into the magnetosphere most often leads to both an increase in directly driven activity and storage of the energy in the magnetotail. Thus the ground based magnetic signatures identified by McPherron were indeed effective proxy measures for the storage of energy in the tail, viz, the growth phase. Finally, we should note that the concept of recovery phase as it was originally defined by Akasofu (1964) really pertained to the local behaviour of the auroral arcs during a substorm (i.e., the development of the auroras in the field of view of an allsky camera) as shown in Fig. 7. Here recovery was understood as the period of time after the substorm disturbed region had expanded to its maximum poleward position till the arcs had drifted equatorward to their pre-substorm position. In the more modern global view of the magnetospheric substorm shown in Fig. 8, one sees that growth is characterized by an equatorward expansion of the auroral oval while recovery is characterized by the contraction of the auroral oval to its original position (i.e., recovery on a global scale is characterized by poleward motion in contrast to recovery on a local scale which is characterized by equatorward motion). The auroral substorm as defined by Akasofu (1964) best describes the latitudinally and longitudinally localized regions of auroral enhancements that have surgelike form and which expand poleward and then die out in a relatively short time ($\sim$15–30 min) compared to the lifetime of a typical magnetospheric substorm. In fact, on a global scale equatorward motion of arcs is more likely to be associated with growth in the sense of more energy entering the magnetosphere from the solar wind and being stored in the magnetotail.

3. Role of the magnetotail in the substorm

In the previous section we have concentrated on the ionospheric signatures of substorm activity that dominated early research in this subject area. In recent times, research has concentrated on the physical mechanisms responsible for the various disturbances observed in the time frame of the magnetospheric substorm. It is clear that the magnetic field lines threading the ionosphere (where auroral substorm disturbances are observed) map into the magnetotail. Therefore, a great deal of effort has been expended in trying to define the behaviour of the particles and fields in the near-Earth, middle and distant magnetotail where the source regions for the observed ionospheric disturbances are likely to be located. In this section we shall explore the evolution of the the near-Earth neutral line (NENL) model for substorms which is thought, by most members of the substorm community, to be the most likely framework in which to develop an understanding of the substorm process.

The present day model for substorms owes its origin to the proposal by Dungey (1961) that energy could be transferred from the interplanetary medium to the magnetosphere through a process in which the solar wind magnetic field merged with the terrestrial magnetic field at the dayside magnetopause when there was a component of the IMF anti-parallel to the Earth's magnetic field (i.e., when the IMF Bz component was approximately southward). Following the dayside merging, the

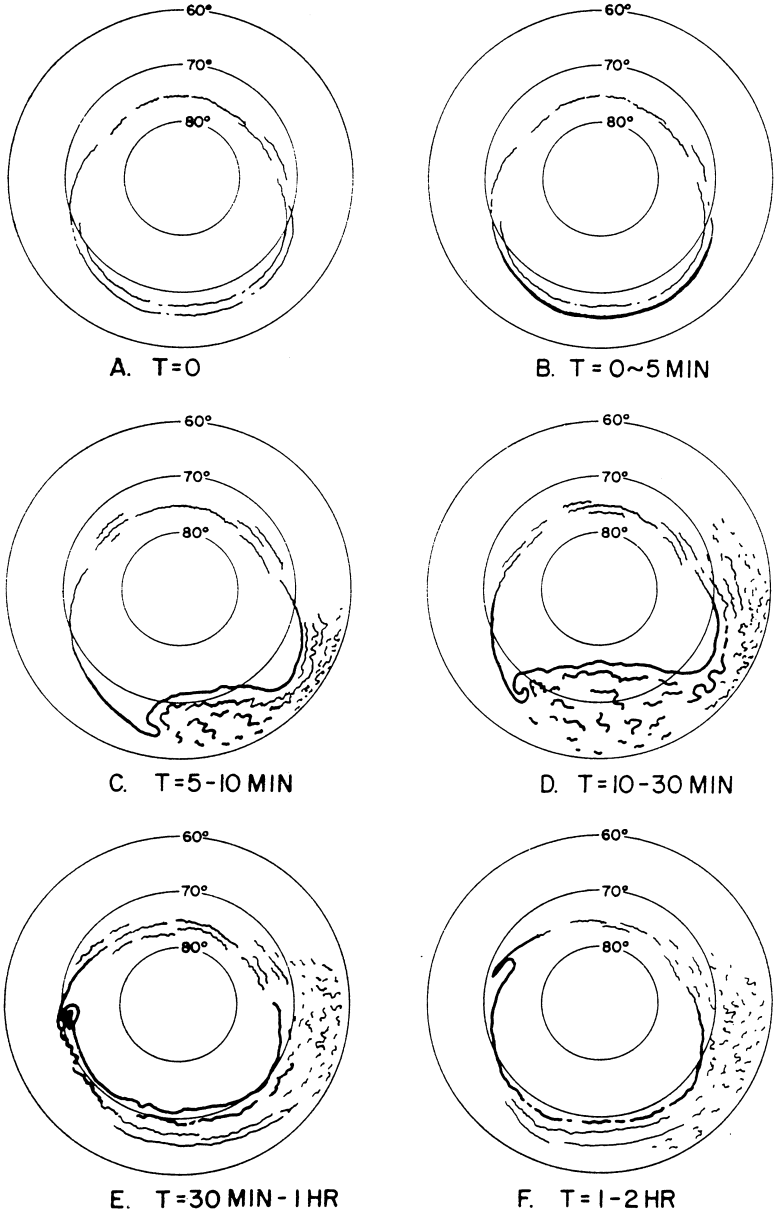

Fig. 7. The auroral substorm as inferred from allsky camera data by Akasofu (1964). The expansive phase involves poleward motion of the region of auroral disturbance while the recovery phase involves equatorward motion of the remaining arc structures.

field lines would then be transported over the poles and would reconnect in the lee of the Earth with the reconnected field lines contracting back towards the Earth as shown in Fig. 9.

Camidge and Rostoker (1970) used IMP A and IMP B magnetometer data to show the response of the magnetotail to substorm activity and reached the conclusion that substorms were associated with the formation of a neutral line normally somewhere beyond $\sim -21\ R_E$. Rostoker and Camidge (1971) further concluded that the disturbed region of the tail only occupied a limited azimuthal extent for each substorm intensification. Coroniti and Kennel (1972) explored the concept of the association of substorms with tail reconnection in more detail and reached the conclusion that there would normally be an imbalance between the frontside merging rate and the reconnection rate in the magnetotail. To begin with, the frontside merging rate would exceed the tail reconnection rate and magnetic flux would pile up in the magnetotail. From time to time there would be sud-

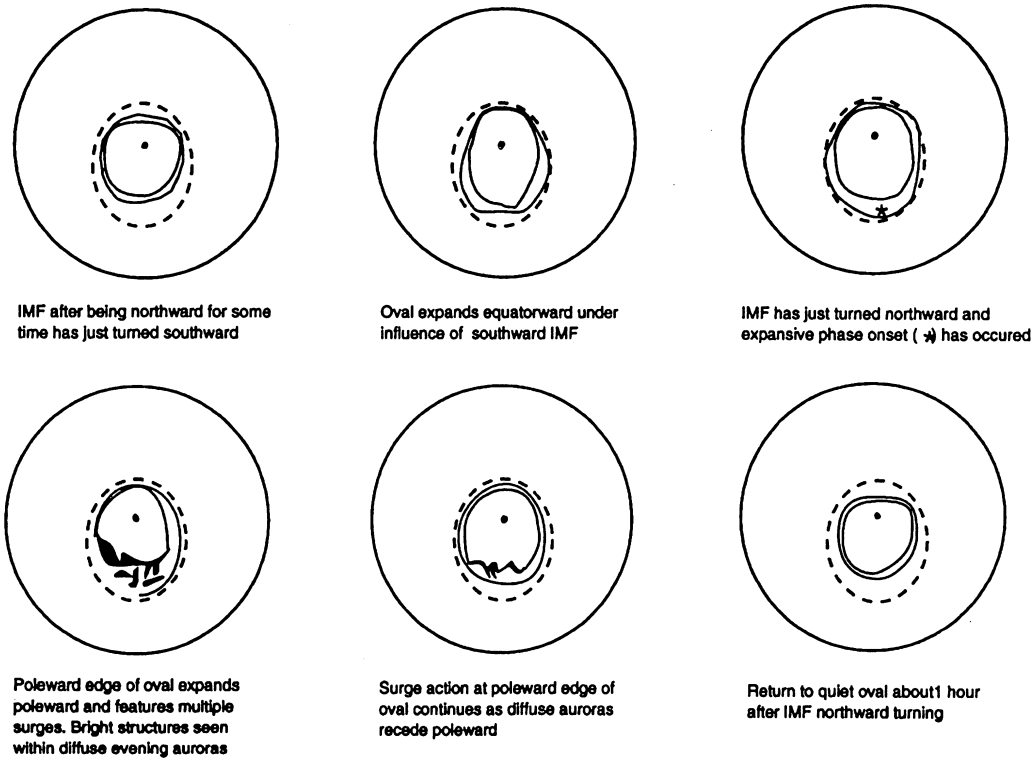

Fig. 8. Global view of auroral oval evolution during a magnetospheric substorm (after Rostoker, 1996). Storage of tail energy (i.e., growth phase) involves an expansion of the polar cap and a resultant equatorward shift of the auroral oval. Release of tail energy (i.e., recovery phase) involves a poleward shift of the auroral oval poleward border. The substorm expansive phase involves the appearance of localized bright auroral arc structures near midnight, starting at the equatorward edge of the auroral oval and progressing poleward during the course of the event. Expansive phase activity concludes with auroral arc activations at the poleward edge of the contracted oval.

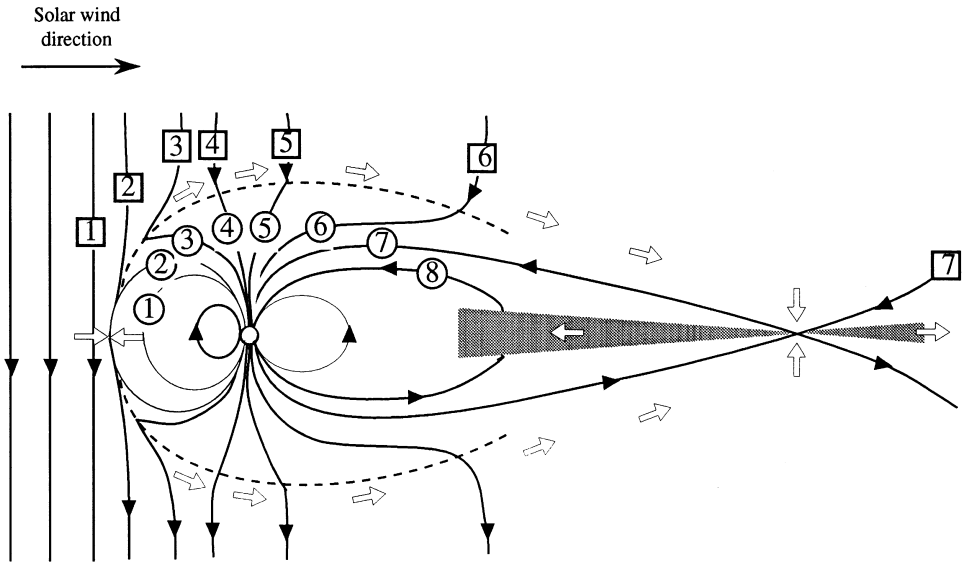

Fig. 9. The reconfiguration of magnetic field lines associated with frontside merging and nightside reconnection as proposed by Dungey (1961). In this original concept, there is only one neutral line in the magnetotail.

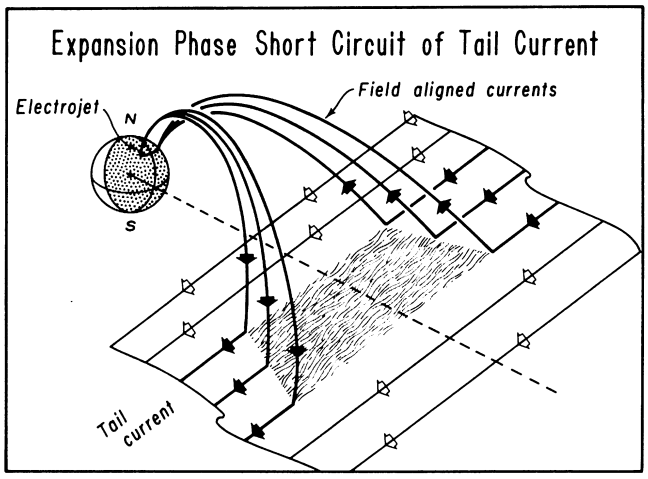

Fig. 10. The substorm current wedge (after McPherron et al., 1973). This figure shows the substorm current system as a real (rather than equivalent) current system and suggests the physical processes which might go on in the tail during a time of substorm activity.

den bursts of reconnection in the tail during which the reconnection rate would exceed the frontside merging rate and magnetic flux would be returned to the dayside. These bursts were thought to be the tail signatures of substorm expansive phases.

McPherron et al. (1973) used OGO 5 satellite data taken in the magnetotail to support the Coroniti and Kennel idea of imbalance between dayside merging and nightside reconnection described above. They further suggested that the formation of a neutral line relatively close to the Earth was associated with the formation of the substorm current wedge as shown in Fig. 10, with the reduction in crosstail current being due to shortcircuiting through the midnight sector ionosphere.

During the late 1960s and early 70s, Hones and colleagues at the Los Alamos National Laboratory carried out a series of studies of the characteristics of magnetotail particles using detectors aboard the VELA satellites whose $\sim 18\ R_E$ circular orbits allowed them to sample regions inside $\sim -20\ R_E$ behind the Earth. Those studies led to the suggestion by Hones (1976) that magnetic reconnection was occurring at almost all times in the distant magnetotail but that, at the time of substorm onset, a new neutral line formed closer to the Earth. He termed this the near-Earth neutral line (NENL), and thus was born the NENL model of magnetospheric substorms. The general character of substorm development in the magnetotail is shown in Fig. 11.

The essence of the NENL paradigm as espoused in the ~ 20 years that passed since it was proposed can be summarized using the numbers beside each of the panels in Fig. 11 that depict the evolution of the magnetotail during a substorm. Panel 1 shows a quiescent magnetotail with reconnection occurring at some distant neutral line thought (at that time) to lie several tens of R_E behind the Earth. An increase in energy input into the magnetosphere sets in motion a series of events which alters the topology of the magnetotail. The addition of energy to the magnetotail involves a thinning of the plasma sheet ultimately leading to the start of reconnection at a new near-Earth neutral line as shown in panel 2. Reconnection of closed field lines proceeds until, as shown in panel 6, the reconnection of open field lines commences. That moment was viewed as marking the onset of the substorm expansive phase (the time from the start of increased energy input to the magnetosphere to the time at which open field lines begin to reconnecting being considered as the growth phase). A large blob of plasma (termed the plasmoid) was then disconnected from the Earth and began to move downtail at high speed leaving behind a thin plasma sheet (panels 7–9). The recovery phase of the substorm was marked by a thickening plasma sheet, the thickening proceeding downtail as recovery continues (panel 10). The presence of the ISEE 3 satellite in the distant magnetotail during the early 1980s provided an opportunity to study the characteristics of the 'plasmoid' far downtail (cf Slavin et al., 1985). Figure 12 shows a 'classic' plasmoid reported in a more detailed study by Slavin et al. (1989). It carries with it the canonical signatures of a bipolar Bz magnetic field perturbation (first positive and then negative) and tailward plasma flow.

In the early 1990s, as additional evidence began to be gathered during the Solar Terrestrial Energy Program (STEP) period using both ground based and satellite borne detectors, a revised picture of the NENL model emerged. For example, both Lui (1991) and Rostoker (1991) pointed to key observational features in particle and field measurements made in space which either were inconsistent with the NENL model or which provided

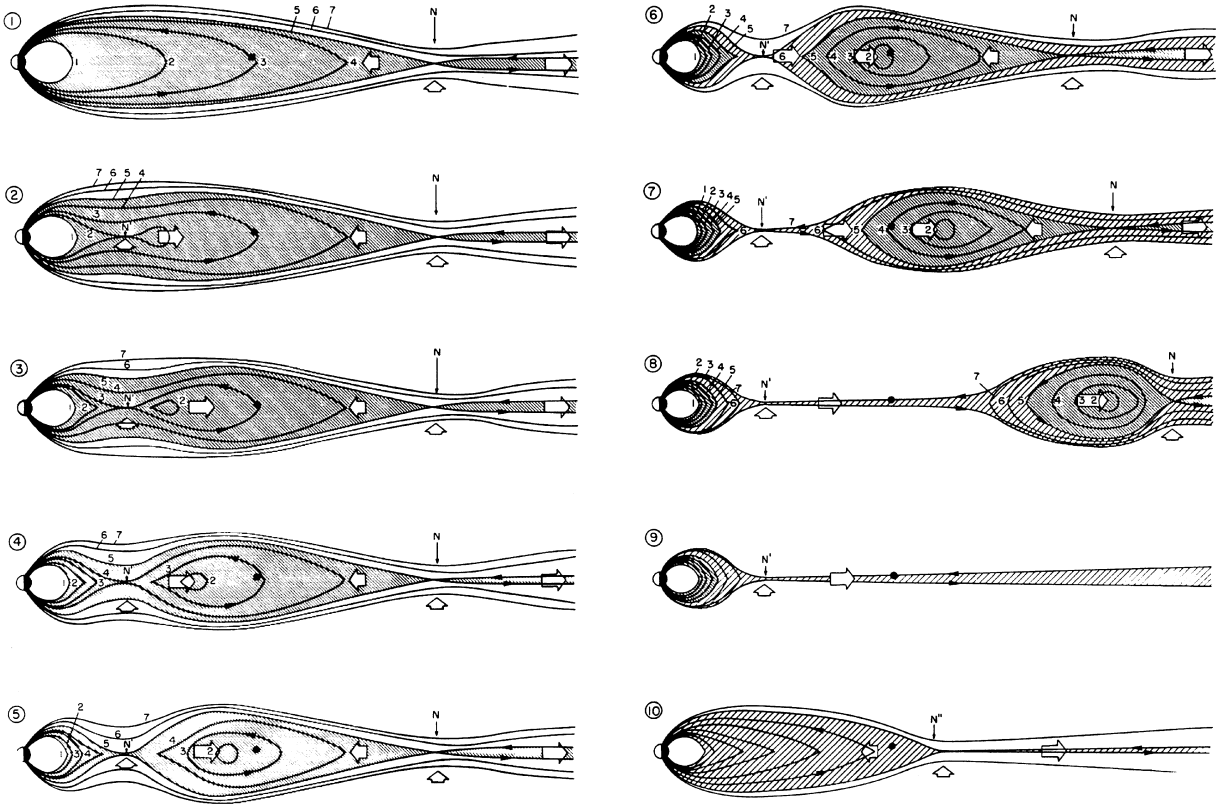

Fig. 11. The development of a magnetospheric substorm in the magnetotail according to the near-Earth neutral line hypothesis (after Hones, 1984).

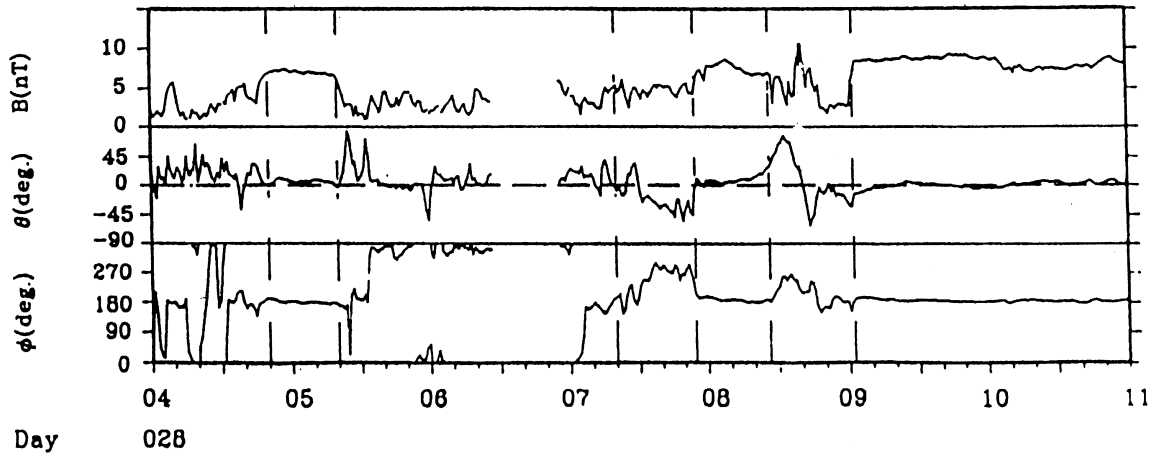

Fig. 12. Magnetic field variations associated with a structure defined as a plasmoid which was detected $\sim 200\ R_E$ behind the Earth by the ISEE 3 satellite (after Slavin et al., 1989). The variation in the θ angle between $\sim$0820–0900 UT indicates a positive Bz perturbation followed by a negative perturbation. The plasma flows tailward at $\sim$400 km/s during this transient disturbance. It is important to note that the magnetic field is strongest in the center of the structure contrary to what one would expect for a plasmoid as originally conceived but consistent with that expected for a flux rope.

compelling evidence for the initiation of expansive phase onset at a near-Earth site. However, arguably the most important discovery in the early 1990s that impacted the NENL model was the definitive placement of the region of onset within the nightside magnetosphere in the equatorial plane by Samson et al. (1992a, b). Until then, researchers had thought that the near-Earth neutral line was threaded by magnetic field lines that mapped to the region of substorm expansive phase onset in the high latitude ionosphere. (That explains why substorm expansive phase onset was considered to be marked by the start of reconnection of open (lobe) field lines, since the energy required for the expansive phase ionospheric disturbance was thought to be supplied by the lobe magnetic field through the reconnection process.) While some researchers (e.g. Lopez et al., 1990; Lui et al., 1991) had argued that expansive phase onset occurred in the near-Earth plasma sheet, the NENL model in its original form continued to hold sway in the community until the observations by Samson and colleagues made the NENL hypothesis in its original form untenable. The final proof came from two different sets of observations. First of all Samson et al. (1992a) were able to show a substorm expansive phase which was on field lines equatorward of a region of field line resonance that clearly lay on dipolar (or at the least quasi-dipolar) field lines characteristic of the inner edge of the plasma sheet. In a second study, Samson et al. (1992b) showed that the breakup arcs of a substorm expansive phase were located in virtually the same volume of space as H_β emissions associated with precipitating energetic protons. Since protons of that energy (tens of keV) are normally only found near the inner edge of the plasma sheet, this implied that substorm onset must take place in that volume of space well earthward of the region where it was thought that a near-Earth neutral line might be formed ($< -12\ R_E$).

Baker et al. (1996) recently presented a revised framework for the NENL hypothesis in which it is acknowledged that the substorm expansive phase onset lies well earthward of the near-Earth neutral line position. Based on this change in the model, they contend that substorm expansive phase onset actually starts at the time of reconnection of closed field lines (cf panel 2 of Fig. 11). Hence, the expansive phase onset is no longer marked by the start of reconnection of open (lobe) field lines and accordingly the release of the plasmoid is no longer to be associated with the substorm expansive phase onset.

A second aspect of the original NENL model (which is now being re-evaluated thanks to the better data sets acquired by satellites orbiting in the magnetotail during the STEP period) is the plasmoid itself. Originally the plasmoid was viewed as a closed loop magnetic field structure containing hot plasma sheet particles. However, this view carried with it the presumption that the magnetic field would be weakest in the center of the plasmoid. Observations of most magnetic field structures described as plasmoids do not conform to that view. Even in the 'classic' plasmoid shown in Fig. 12, it is clear that the magnetic field is in fact a maximum in the center of the plasmoid. More recently this fact has been acknowledged and the combination of bipolar Bz and tailward flow is referred to more as a flux rope. A flux rope involves a current flowing across the tail within the magnetic field structure giving a crosstail core magnetic field component. Introducing this concept has created more questions than it has answered. If current flows across the tail, does it close along the magnetopause boundary (and hence stretch across the entire tail) or does it close through field aligned currents which are connected by transverse currents flowing in the high latitude ionosphere as suggested by Kivelson et al. (1996)? This, in turn, begs the question of whether or not flux ropes are azimuthally localized giving them a property attributed to the original plasmoid. If the flux rope/plasmoid is azimuthally localized, then we must envision a channel of high speed antisunward flow near the center of the tail flanked on both sides by slow speed earthward flow as shown in Fig. 13. This, in turn speaks of two regions of very high velocity shear at the interfaces between earthward and tailward flow that must have associated with them field-aligned currents flowing into and out of the ionosphere. The resulting three dimensional current loop involving eastward closure current in the ionosphere should be detectable in the region just poleward of the substorm disturbed auroras in the midnight sector, and identification of such a current loop would be strong evidence for the proposed role of flux ropes/plasmoids in the substorm process.

If the NENL model is to explain substorms, it must satisfy the observational constraints in the near-Earth plasma sheet. Two of these pertain to the behaviour of the magnetic field in and around geostationary orbit. The first relates to the radial current density profile near the inner edge of the crosstail current sheet. It was found by Kaufmann (1987) that, in order to reproduce the stretching of the near-Earth tail magnetic field near geostationary orbit characteristic of many substorm growth phases, it was necessary to have an extremely high crosstail current density (~ 300 mA/m) over a limited radial range. Any substorm model has to be able to explain how this current density can gradually build up during the substorm growth phase. The second constraint relates to the fact that, a minute or two before the onset of an expansive phase there is a sudden explosive stretching of the tail magnetic field in the midnight sector (Ohtani et al., 1992) as shown in Fig. 14. Any substorm model must be able to explain this apparently unstable behaviour, and the explanation must bring into play both the source of energy for the sudden enhancement of near-Earth crosstail current and the the mechanism for its release in the expansive phase process. At the present time, some form of ballooning instability in the presence of either velocity shear (Voronkov et al., 1997) or a thin current sheet (A.

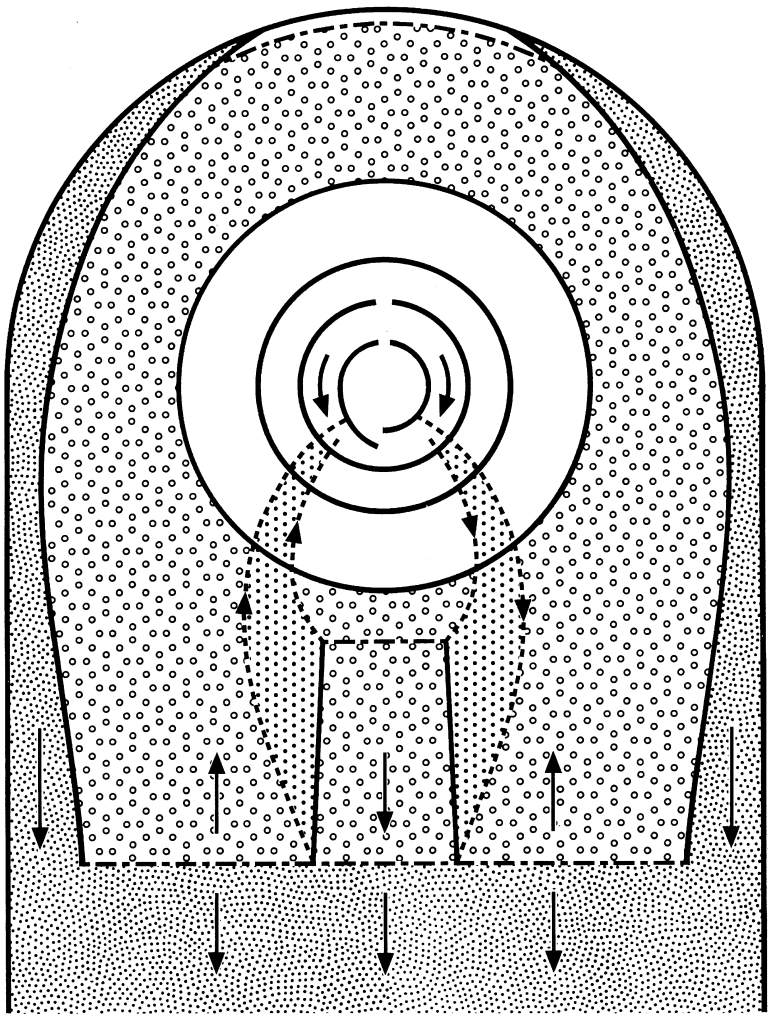

Fig. 13. Plasma flows expected in the central plasma sheet (indicated by open circles) if there is a fast anti-sunward flow tailward of a near-Earth neutral line (after Rostoker, 1991) during substorm expansive phase activity. The velocity shear zones should be associated with field aligned currents which must close in the high latitude ionosphere.

Bhatacharjee, private communication) appears attractive as a trigger for the substorm expansive phase.

In terms of how the revised NENL model explains the development of the expansive phase onset near the inner edge of the plasma sheet, the recent observations by Angelopoulos et al. (1992) provide the observational justification for the present view of this issue. Angelopoulos and colleagues, using AMPTE particle and field measurements in the region earthward of $\sim -19\ R_E$, established the presence of fast short lived earthward flows which they termed bursty bulk flows (BBFs). These flows typically last no more than a few seconds and feature speeds of hundreds of km/s against the slow background flow of a few tens of km/s. These fast flows had also been noted by Baumjohann et al. (1989) using the same AMPTE data set, and they established that almost all fast flows inside $\sim 19\ R_E$ were earthward. This observation was extremely important in that it suggested that any near-Earth neutral line must normally lie outside the AMPTE apogee of $\sim -19\ R_E$ if, indeed, the source of the fast flows was reconnection at a neutral line. The review by Baker et al. (1996) of the revised version of the NENL model does recognize the fact that any near-Earth neutral line must lie outside $\sim -20\ R_E$, and in doing so calls for some way in which reconnection at that neutral line can contribute to the substorm expansive phase process. Recent considerations of that matter (R.L. McPherron, private communication) reveal that NENL

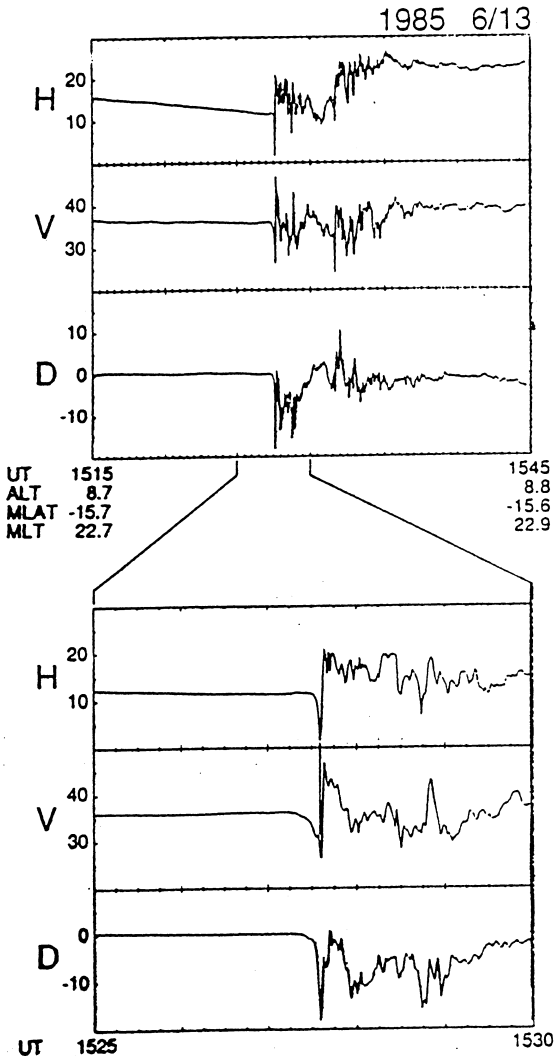

Fig. 14. Explosive growth of the crosstail current detected by the AMPTE satellite $\sim 2\ R_E$ outside geostationary orbit just in advance of a substorm expansive phase which takes place at $\sim$1527:30 UT on 13 June 1985 (after Ohtani et al., 1992).

proponents consider that the substorm expansive phase onset is caused by reconnection at a newly formed neutral line typically located between $\sim$ 20–30 R_E downtail. BBFs from this neutral line bring energy to the near-Earth plasma sheet quite rapidly in the form of kinetic energy of the earthward flowing ions. Braking of these BBFs results in the flow of the electric currents associated with the substorm current wedge (cf Shiokawa et al., 1997; Slavin et al., 1997). In order to provide sufficient energy for the substorm process in the near-Earth plasma sheet, a significant volume of the center of the magnetotail must be filled with BBFs. Future multi-satellite studies in the near-Earth plasma sheet will determine whether or not that energy constraint is met.

4. Problems to be solved in the search for the correct framework in which to describe the substorm process

In the final portion of this paper we point out a two areas in which further progress must be made before a satisfactory and minimally non-unique explanation of the substorm can be achieved.

Question #1: Is there more than one neutral line in the magnetotail, and if so where are the two neutral lines located?

To put this question in context, one must recognize that the original Dungey version of the solar wind / magnetosphere interaction involved merging on the dayside and only one neutral line on the nightside located $\sim 100\ R_E$ behind the Earth. The proposal by Hones (1976) that there was a near-Earth neutral line activated during the substorm expansive phase was based largely on observations of anti-sunward plasma flow by the VELA satellites which orbited around $-18\ R_E$ behind the Earth. The anti-sunward flows were attributed to neutral lines which lay between the satellite and the Earth. However, it is now recognized that the near-Earth neutral line rarely lies that close to the Earth and hence the tailward flows which on which the NENL hypothesis was founded may owe their origin to some other effect. Modern studies of the response of the magnetotail during substorm expansive phase activity (e.g., Nagai et al., 1998) seem to indicate that the near-Earth neutral line lies somewhere between $\sim -20\ R_E$ and $-30\ R_E$ behind the Earth. This is not far from the radial distance at which Frank and Paterson (1994) claim that the average flow in the plasma sheet changes from earthward to tailward. One is then tempted to ask the question of whether or not there might be only one neutral line whose average position is much closer to the Earth than the original conjecture of Dungey would imply. This question is made all the more complex by the claim of Nishida et al. (1996) that there are two neutral lines, one lying near the position at which Frank and Paterson claim that the average flow direction in the plasma sheet reverses and one located at $\sim -150\ R_E$. The distant neutral line position suggested by Nishida et al. is well tailward of previous estimates of the location of that important region. Thus there is enough uncertainty in proposed positions of the neutral line or neutral lines over the history of magnetotail observations to make the question of whether or not there are one or two neutral lines active in the tail during substorms worth vigorously investigating with our increasingly better data sets.

Question 2: How can one map from the ionosphere to the magnetotail so as to be able to identify source regions in the magnetotail for auroral oval disturbances?

One of the great difficulties in studying the structure and dynamics of the magnetotail centers on the fact that most observations of plasmas and fields in that region of

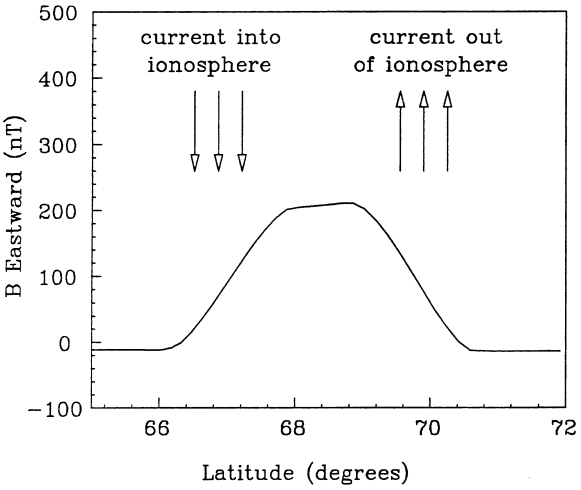

Fig. 15. Magnetic perturbation in the East–West direction typical of that which would have been measured by the polar orbiting satellite TRIAD as it crosses the auroral oval at ~800 km altitude near ~2100 MLT. Magnetic perturbations of this magnitude are normal for moderately disturbed times (after Donovan, 1993).

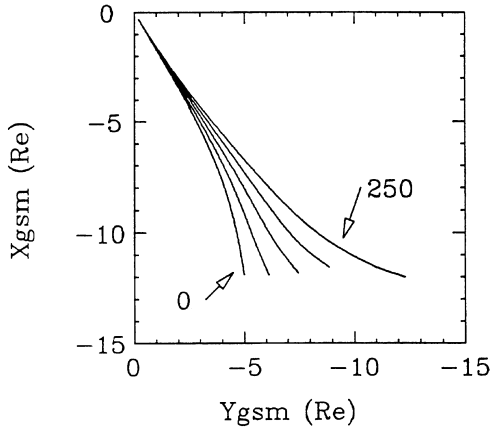

Fig. 16. Mapping of magnetic field lines from 69.25°N and 0200 LT in the ionosphere to the XY gsm equatorial plane for various strengths of Birkeland currents which would be detected at 800 km altitude. The field-aligned currents would produce 0, 62.5, 125, 187.5 and 250 nT, and lead to flankward skewing of the magnetic field lines at the interface between Regions 1 and 2 currents as shown in the figure. A modest perturbation of ~250 nT at 800 km altitude would be associated with field-aligned currents capable of skewing the field lines by up to 7 R_E for a field line with a neutral sheet crossing at ~15 R_E behind the Earth (after Donovan, 1993).

space are made by a single satellite. Since we are dealing with a magnetized plasma which is normally changing temporally while at the same time boundaries are moving with respect to a satellite, it is extremely difficult—if not impossible—to decouple the combined effects of spatial and temporal variations. For this reason, ground based observations of auroras and electric current systems (inferred from magnetic field variations) provide an important constraint for satellite observations in that they may permit the decoupling of the spatial and temporal variations to which the satellite detectors are exposed. However, to relate ground observations to satellite observations, we must have a model which allows us to map from the satellite position to the ionosphere following the magnetic field line on which the satellite is located. Establishing the satellite 'footprint' is a difficult task which calls for an appropriate model of the Earth's magnetic field including both the main field and the fields due to currents flowing in space. While considerable effort has been expended in developing such models (e.g., Tsyganenko, 1987, 1989, 1995), we still have a considerable way to go before we can be confident about mapping from the ionosphere to the magnetosphere. The major block lies in taking into account the effects of field-aligned currents, particularly those associated with the directly driven current system. Figure 15 shows the magnetic perturbation from field-aligned currents measured by a satellite at ~800 km altitude on a North–South pass. These currents extend from the ionosphere into the nightside magnetosphere where they must close through transverse currents that are driven by some sort of generator process. The important question to be answered is where, in space, this generator region is found. The problem is that the magnetic perturbations of the field-aligned currents distort the background magnetic field between the current sheets, so that those field lines can cross the equatorial plane at some considerable distance from where they might have crossed had the field-aligned currents not been present. Figure 16 shows the projection onto the XYgsm plane of magnetic field lines traced using a model developed by Donovan (1993) based on the T87 model of Tsyganenko (1987) and modified by the inclusion of field-aligned current sheets whose effects are calculated in a self-consistent fashion. It is clear that, at a distance of ~15 R_E behind the Earth, field lines at the interface between the downward and upward currents can be skewed towards the flanks of the magnetotail by as much as 7 R_E for a Birkeland current system with a modest current density capable of producing a 250 nT magnetic perturbation at an altitude of ~800 km. Thus a field line which one might have considered to cross the midplane near the center of the tail actually can cross the midplane close to the flanks. Choosing the wrong volume of space for the source region of the substorm expansive phase can have a negative impact on the task of trying to select the correct physical mechanism for explaining the disturbance. Thus it is essential for a reliable magnetic field model to be developed which will permit information obtained by ground based instrument arrays to be used

in placing constraints on satellite observations of variations of particles and fields in the magnetotail.

5. Summary and conclusion

We have attempted to review the properties of the magnetospheric substorm as measured by ground based magnetometers and have tried to relate in situ observations of particles and fields in the magnetotail to the substorm framework inferred from the ground based measurements. In the first portion of this review we have tried to emphasize that there are two components of substorm activity: directly driven and storage/release. Most of the attention over the past 30 years has been paid to the storage/release process, for which the equivalent three dimensional current system is the so-called substorm current wedge. However, directly driven activity accounts for a significant portion of the energy dissipation in a substorm, so it is rather surprising that most substorm models do not address in detail the physical mechanism for directly driven activity nor the location of the volume of space in which the energy for this activity is made available. Perhaps one of the few models addressing this question is the renovated boundary layer dynamics model proposed by Rostoker (1996); however, in the future researchers will no doubt try to find a description of directly driven activity in other frameworks (such as the NENL model). Another issue addressed in the first portion of this review was the terminology used to describe the so-called phases of a substorm. Here the important point to note is that we can look at a substorm either on a global scale or on a local scale (cf Figs 7 and 8). On a local scale the recovery of an expansive phase intensification involves the equatorward drift of auroral forms (following the poleward motion involved in the expansive phase). However, on a global scale recovery actually involves poleward motion following the equatorward expansion of the auroral oval during the growth phase. We expect a global poleward recovery to take place when the rate of energy input to the magnetosphere decreases to pre-substorm levels. It is entirely likely that the 'poleward leaps' of the aurora described by Hones (1985) simply represent the recovery from an expanded oval to a contracted oval configuration associated with a reduction of energy input into the magnetosphere as can be effected by, for example, a northward turning of the IMF.

In the second part of this review, we have looked at the evolution of the most well accepted framework in which researchers attempt to understand the substorm: the near-Earth neutral line (NENL) model. We have shown that this model is in a state of change at the present time because of the realization that the onset of the substorm expansive phase does not occur at the time that lobe field lines begin to reconnect in the magnetotail, as previously thought. One important consequence of this new development in the NENL hypothesis relates to the timing of plasmoids in the distant tail. In its original form, the NENL model featured the release of the plasmoid at the time when lobe field lines began to reconnect and the substorm expansive phase began. Thus the size of plasmoid structures and speeds of propagation downtail were established assuming the plasmoid began to propagate at the time of expansive phase onset. All these timings must now be re-examined because the present version of the NENL model does not relate the release of the plasmoid to the substorm expansive phase onset. A further question to be answered regarding the plasmoid relates to the fact that, in almost all cases, the structure has a strong core field making it more of a flux rope than a bubble of hot gas. If, indeed, the structures which have been termed plasmoids in the past feature helical currents flowing so as to generate a dawn-dusk core magnetic field within the structure, then it is important to know how that helical current closes in the magnetosphere–ionosphere system. Almost certainly, the solution to that question will be an important component to the overall problem of defining the role of the magnetotail in the substorm process. Finally, we have explained how changes in the NENL model have demanded that some way be found to transport energy from the near-Earth neutral line to the region of space near the inner edge of the plasma sheet where substorm expansive phase activity is initiated. At present, the evolving view is that this transport is achieved by bursty bulk flows which are generated through reconnection at the near-Earth neutral line with some sort of braking mechanism close to Earth acting to convert the kinetic energy of the earthward flowing ions to the electromagnetic energy for the substorm current wedge. This new concept is less than a year old, in terms of published work at the time of writing of this review, and we can expect rapid evolution of the concept as the STEP data bases are further analysed in the next few years.

In this review, we have not devoted attention to alternative models for substorms. The reader is directed to a series of papers in the June 1996 issue of the Journal of Geophysical Research for the most recent views of the proponents of the various models for substorms, including the near-Earth neutral line model.

Acknowledgements

I am grateful to my many substorm colleagues for discussions that have led, over the years, to the views expressed in this review. This research was supported by the Natural Sciences and Engineering Research Council of Canada under Operating Grant OGP 5420.

References

Akasofu, S.-I., 1964. The development of the auroral substorm. Planet. Space Sci. 12, 273.

Akasofu, S.-I., 1968. Polar and Magnetospheric Substorms. D. Reidel, Norwell, MA.

Akasofu, S.-I., 1980. What is a magnetospheric substorm? In: Akasofu, S.-I. (Ed.). Dynamics of the Magnetosphere. D. Reidel Publ. Co., Dordrecht Holland, pp. 447–460.

Akasofu, S.-I., Chapman, S., Meng, C.-I., 1966. The polar electrojet. J. Atmos. Terr. Phys. 27, 1275.

Angelopoulos, V.W. et al., 1992. Bursty bulk flows in the inner central plasma sheet. J. Geophys. Res. 97, 4027–4030.

Baker, D.N., Pulkkinen, T.I., Angelopoulos, V., Baumjohann, W., McPherron, R.L., 1996. Neutral line model of substorms: Past results and present view. J. Geophys. Res. 101, 12,975–13,010.

Baumjohann, W., Paschmann, G., Cattell, C.A., 1989. Average plasma properties in the central plasma sheet. J. Geophys. Res. 94, 6597–6606.

Birkeland, K., 1908. The Norwegian Aurora Polaris Expedition 1902–1903, 1, section 1, Aschhoug and Co., Christiania.

Camidge, F.P., Rostoker, G., 1970. Magnetic field perturbations in the magnetotail associated with polar magnetic substorms. Can. J. Phys. 48, 2002–2010.

Chapman, S., Ferraro, V.C.A., 1932. A new theory of magnetic storms. Terr. Mag. Atmos. Electr. 36, 77.

Clauer, C.R., McPherron, R.L., Searls, C., 1983. Solar wind control of the low-latitude asymmetric magnetic disturbance field. J. Geophys. Res. 88, 2123.

Coroniti, F.V., Kennel, C.F., 1972. Changes in magnetospheric configuration during the substorm growth phase. J. Geophys. Res. 77, 3361–3370.

Davis, T.N., Sugiura, M., 1966. Auroral electrojet activity index AE and its universal time variations. J. Geophys. Res. 71, 785.

Donovan, E.F., 1993. Modeling the magnetic effects of field-aligned currents. J. Geophys. Res. 98, 13,529–13,543.

Dungey, J.W., 1961. Interplanetary magnetic field and the auroral zones. Phys. Rev. Lett. 6 , 47, 1961.

Frank, L.A., Paterson, W.R., 1994. Survey of electron and ion bulk flows in the distant magnetotail with the Geotail spacecraft. Geophys. Res. Lett. 21, 2963–2966.

Fukushima, N., 1953. Polar magnetic storms and geomagnetic bays. J. Fac. Sci. Tokyo Univ. 8, 293.

Harang, L., 1946. The mean field of disturbance of polar geomagnetic storms. Terr. Mag. Atmos. Electr. 51, 353.

Hones, E.W. Jr., 1976. The magnetotail: Its generation and dissipation. In: Williams, D.J. (Ed.). Physics of Solar Planetary Environments. Proceedings of the International Symposium on Solar-Terrestrial Physics. American Geophysical Union, p. 558.

Hones, E.W. Jr., 1984. Plasma sheet behaviour during substorms. In: Jones, E.W. Jr. (Ed.). Magnetic Reconnection in Space and Laboratory Plasmas. American Geophysical Union Monograph 30, Washington, DC., p. 178.

Hones, E.W. Jr., 1985. The poleward leap of the auroral electrojet as seen in auroral images. J. Geophys. Res. 90, 5333.

Kaufmann, R.L., 1987. Substorm currents: Growth phase and onset. J. Geophys. Res. 92, 7471.

Kisabeth, J.L., Rostoker, G., 1974. The expansive phase of magnetospheric substorms 1. Development of the auroral electrojets and auroral arc configuration during a substorm. J. Geophys. Res. 79, 972–984.

Kivelson, M.G., Khurana, K.K., Walker, R.J., Kepko, L., Xu, D., 1996. Flux ropes, interhemispheric conjugacy, and magnetospheric current closure. J. Geophys. Res. 101, 27,341–27,350.

Lopez, R.E., Lühr, H., Anderson, B.J., Newell, P.T., McEntire, R.W., 1990. Multipoint observations of a small substorm. J. Geophys. Res. 95, 18,897–18,912.

Lui, A.T.Y., 1991. A synthesis model for magnetospheric substorms. In: Kan, J.R., Potemra, T.A., Kokubun, S., Iijima, T. (Eds.). Magnetospheric Substorms. Amer. Geophys. Union, Washington, DC. pp. 61–72.

Lui, A.T.Y., Chang, C.-L., Mankofsky, A., Wong, H.-K., Winske, D., 1991. A cross-field current instability for substorm expansions. J. Geophys. Res. 96, 11,389–11,401.

McPherron, R.L., 1970. Growth phase of magnetospheric substorms. J. Geophys. Res. 75, 5592–5599.

McPherron, R.L., Russell, C.T., Aubry, M.P., 1973. Satellite studies of magnetospheric substorms on August 15, 1968 9. Phenomenological model for substorms. J. Geophys. Res. 78, 3131–3149.

Nagai, T. et al., 1998. Structure and dynamics of magnetic reconnection for substorm onsets with GEOTAIL observations. J. Geophys. Res. 103, 4419–4440.

Nishida, A., Mukai, T., Yamamoto, T., Saito, Y., Kokubun, S., 1996. Magnetotail convection in geomagnetically active times 1. Distance to the neutral lines. J. Geomag. Geoelectr. 48, 489–501.

Ohtani, S., Takahashi, K., Zanetti, L.J., Potemra, T.A., McEntire, R.W., Iijima, T., 1992. Initial signatures of magnetic field and energetic particle fluxes at tail reconfiguration: explosive growth phase. J. Geophys. Res. 97, 19,311–19,324.

Perreault, P., Akasofu, S.-I., 1978. A study of geomagnetic storms. Geophys. J. Roy. astron. Soc. 54, 547.

Rostoker, G., 1969. Classification of polar magnetic disturbances. J. Geophys Res. 74, 5161–5168.

Rostoker, G., 1991. Overview of observations and models of auroral substorms. In: Meng, C.-I., Rycroft, M.J., Frank, L.A. (Eds.). Auroral Physics. Cambridge Univ. Press, New York. pp. 257–271.

Rostoker, G., 1991. Some observational constraints for substorm models. In: Kan, J.R., Potemra, T.A., Kokubun, S., Iijima, T. (Eds.). Magnetospheric Substorms. Amer. Geophys. Union, Washington, DC. pp. 61–72.

Rostoker, G., 1996. Phenomenology and physics of magnetospheric substorms. J. Geophys. Res. 101, 12,955–12,973.

Rostoker, G., Camidge, F.P., 1971. The localized character of magnetotail fluctuations during polar magnetic substorms. J. Geophys. Res. 76, 6944–6951.

Samson, J.C., Wallis, D.D., Hughes, T.J., Creutzberg, F., Ruohoniemi, J.M., Greenwald, R.A., 1992a. Substorm intensifications and field line resonances in the nightside magnetosphere. J. Geophys. Res. 97, 8495–8518.

Samson, J.C., Lyons, L.R., Newell, P.T., Creutzberg, F., Xu, B., 1992b. Proton aurora and substorm intensifications. Geophys. Res. Lett. 19, 2167–2170.

Shiokawa, K., Baumjohann, W., Haerendel, G., 1997. Braking of high-speed flows in the near-Earth tail. Geophys. Res. Lett. 24, 1179–1182.

Silsbee, H.C., Vestine, E.H., 1942. Geomagnetic bays, their occurrence frequency and current systems. Terr. Mag. 47, 195.

Slavin, J.A., Smith, E.J., Sibeck, D.G., Baker, D.N., Zwickl, R.D., Akasofu, S.-I., 1985. An ISEE-3 study of average and substorm conditions in the distant magnetotail. J. Geophys. Res. 90, 10,875–10,895.

Slavin, J.A. et al., 1989. CDAW 8 observations of plasmoid signatures in the geomagnetic tail: an assessment. J. Geophys. Res. 94, 15,153–15,175.

Slavin, J.A. et al., 1997. WIND, GEOTAIL, and GOES 9 observations of magnetic field dipolarization and bursty bulk flows in the near-tail. Geophys. Res. Lett. 24, 971–974.

Sugiura, M., Heppner, J.P., 1965. The Earth's magnetic field. In: Hess, W.N. (Ed.). Introduction to Space Science. Gordon and Breach, New York. p. 5.

Tsyganenko, N.A., 1987. Global quantitative models of geomagnetic field in the cislunar magnetosphere for different disturbance levels. Planet. Space Sci. 35, 1347–1358.

Tsyganenko, N.A., 1989. A magnetospheric magnetic field model with a warped tail current sheet. Planet Space Sci. 37, 5–20.

Tsyganenko, N.A., 1993. Modeling the Earth's magnetospheric magnetic field confined within a realistic magnetopause. J. Geophys. Res. 100, 5599–5612.

Voronkov, I., Rankin, R., Frycz, P., Tikhonchuk, V.T., Samson, J.C., 1997. Coupling of shear flow and pressure gradient instabilities. J. Geophys. Res. 101, 9639–9650.

Wiens, R.G., Rostoker, G., 1975. Characteristics of the development of the westward electrojet during the expansive phase of magnetospheric substorms. J. Geophys. Res. 80, 2109–2128.