

JGR Space Physics






RESEARCH ARTICLE

10.1029/2024JA032762

Special Collection:

Recent Discoveries in Substorm Research

On the Association of Substorm Identification Methods

C. J. Lao¹ , C. Forsyth¹ , M. P. Freeman^{1,2} , A. W. Smith³ , and M. K. Mooney⁴ 

¹UCL, Mullard Space Science Laboratory, Dorking, UK, ²British Antarctic Survey, Cambridge, UK, ³Department of Mathematics, Physics and Engineering, Northumbria University, Newcastle upon Tyne, UK, ⁴School of Physics and Astronomy, University of Leicester, Leicester, UK



Key Points:

- Less than 50% of the events from each list analyzed were associated with onsets identified by another signature
- The Chu et al. (2015, <https://doi.org/10.1002/2015JA021104>) and Newell and Gjerloev (2011, <https://doi.org/10.1029/2011JA016779>) lists show the highest levels of event coincidence with other onset lists
- Substorm onsets identified using particle injection signatures showed poor association with other onset signatures

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

C. J. Lao,
christian.lao.21@ucl.ac.uk

Citation:

Lao, C. J., Forsyth, C., Freeman, M. P., Smith, A. W., & Mooney, M. K. (2024). On the association of substorm identification methods. *Journal of Geophysical Research: Space Physics*, 129, e2024JA032762. <https://doi.org/10.1029/2024JA032762>

Received 19 APR 2024

Accepted 6 SEP 2024

Author Contributions:

Conceptualization: C. Forsyth, M. P. Freeman

Data curation: C. Forsyth

Formal analysis: C. J. Lao

Methodology: C. J. Lao, C. Forsyth,

M. P. Freeman, A. W. Smith,

M. K. Mooney

Supervision: C. Forsyth, M. P. Freeman

Visualization: C. J. Lao

Writing – original draft: C. J. Lao

©2024. The Author(s).

This is an open access article under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Abstract Substorms are a rapid release of energy that is redistributed throughout the magnetosphere-ionosphere system, resulting in many observable signals, such as enhancements in the aurora, energetic particle injections, and ground magnetic field perturbations. Numerous substorm identification techniques and onset lists based on each of these signals have been provided in the literature, but often with no cross-calibration. Since the signals produced are not necessarily unique to substorms and may not be sufficiently similar to be identified for each and every substorm, individual event lists may miss or misidentify substorms, hindering our understanding and the development and validation of substorm models. To gauge the scale of this problem, we use metrics derived from contingency tables to quantify the association between lists of substorms derived from SuperMAG SML/SMU indices, midlatitude magnetometer data, particle injections, and auroral enhancements. Overall, although some degree of pairwise association is found between the lists, even lists generated by applying conceptually similar gradient-based identification to ground magnetometer data achieve an association with less than 50% event coincidence. We discuss possible explanations of the levels of association seen from our results, as well as their implications for substorm analyses.

1. Introduction

Magnetospheric substorms are a major energy unloading process in the solar wind-magnetosphere-ionosphere system that are capable of processing $\sim 10^{15}$ J of stored solar wind energy through a complete cycle (Tanskanen et al., 2002). In broad terms, energy is built up and stored in the magnetosphere during substorm growth phases and released during substorm expansion and recovery phases with the start of the expansion phases typically known as substorm onset. The energy transport and conversion between different parts of the magnetosphere occur as a result of various plasma physical processes and at a variety of timescales and time lags. As with any natural phenomenon, the success of a substorm theory or model lies in its ability to describe not only its characteristic features but also its variability. Although considerable progress has been made in identifying and understanding the characteristic features of a substorm (see Section 2), quantitative explanation or prediction of substorm variability such as size or timing is more limited (e.g., Freeman & Morley, 2004; Morley et al., 2007; Maimaiti et al., 2019).

A necessary requirement in developing a model to explain or predict substorm variability is a sufficiently extensive catalog of substorms to validate the model over the widest possible range of circumstances. The first substantial catalog of 1,373 substorm onset times was produced by Borovsky et al. (1993) based on visual inspection of energetic particle data at geostationary orbit and was used to examine the variability of waiting times between substorm onsets. Subsequently, another major catalog of 2,437 substorm onsets was produced by Frey et al. (2004) based on visual inspection of auroral imaging data and used to analyze the variability in substorm onset location in the ionosphere. The data for the two catalogs were not contemporaneous and therefore could not be directly compared, but they both identified ≈ 1500 substorms per year. As discussed in Section 2 below, in the last decade many more catalogs of substorm onset times have been produced based on a variety of data sources and using both visual inspection and automated detection methods. Notably, however, they differ significantly with the number of substorms differing by a factor of five over a common 2.5-year time interval (see Table 1) and with very different distributions of waiting times between onsets (Forsyth et al., 2015; McPherron & Chu, 2018).

Consequently, those seeking to validate their models and theories of substorms using such catalogs are left with the open question of whether any validation is quantifying the model's ability to replicate the substorm phenomenon or its ability to replicate a property of a catalog of events that may or may not all be substorms (Haiducek et al., 2020). Any conclusions could turn out to be specific to the catalog being analyzed. As an example, Forsyth

Writing – review & editing: C. J. Lao,
C. Forsyth, M. P. Freeman, A. W. Smith,
M. K. Mooney

Table 1
Onset Lists Used for Analysis

Substorm identification phenomena	Onset identification method	Data set	Number of substorms
Auroral Enhancements	Frey et al. (2004)	Fr2004 IMAGE	2,437
	Forsyth et al. (2015)	SOPHIE SML	1,724–10,035
Auroral Zone	Newell and Gjerloev (2011)	NG2011 SML	4,793
Ground Magnetic Perturbations	Borovsky and Yakymenko (2017)	BY2017 SML	1,960
	McPherron and Chu (2018)	MC2018 SML	7,908
	Ohtani and Gjerloev (2020)	OG2020 SML	2,265
Midlatitude	Chu et al. (2015)	C2015 MPB	3,418
Ground Magnetic Perturbations	McPherron and Chu (2018)	MC2018 MPB	8,601
Geosynchronous Particle Injections	Borovsky and Yakymenko (2017)	BY2017 PI	2,149

et al. (2015) showed distinct differences both in the average evolution of an auroral magnetic index using different onset lists as the reference and in the distribution of waiting times between onsets, which might support or otherwise corresponding models such as Weimer (1994) and Freeman and Morley (2004), respectively.

Although the signature used for each substorm catalog may be considered to be a plausible identifier of substorm onset, and the substorm concept suggests that the various signatures should all occur at onset, there are various practical and physical reasons why the catalogs may nevertheless differ: Firstly, instruments may not be available or in the right place to observe a substorm signature, which may be relatively localized. For example, spacecraft are continually moving, the spatial distribution of ground magnetometers is not uniform (Shore et al., 2018), an auroral brightening might not be visible in the sunlit summer hemisphere, or a particle injection might not reach geostationary orbit (Boakes et al., 2011). Secondly, signatures may not be sufficiently similar from substorm to substorm to be uniquely defined. For example, substorm identification based on an amplitude and/or rate of change threshold of some measure will generally be affected by the variability of substorm energy loss (Tanskanen et al., 2002), and the ground magnetic perturbations will be affected by variable ionospheric conductance from seasonally dependent solar illumination (Forsyth et al., 2018). Thirdly, substorm identification methods may focus solely on onset (Borovsky & Yakymenko, 2017; Chu et al., 2015; Newell & Gjerloev, 2011; Ohtani & Gjerloev, 2020) rather than a signature of the whole substorm cycle (Forsyth et al., 2015). Fourthly, the observational features that are leveraged for substorm identification may also be attributed to other magnetospheric phenomena. For example, sudden changes in the ground magnetic field may be caused by enhanced magnetospheric convection (Kissinger et al., 2011; Sergeev et al., 1996; Walach & Milan, 2015), pseudobreakups (Kullen & Karlsson, 2004; Pulkkinen et al., 1998), and global sawtooth oscillations (Belian et al., 1995; Cai et al., 2006; Henderson et al., 2006; Lee et al., 2004).

In summary, the absence of an onset in one signature may not indicate the absence of a substorm, and conversely, a singular occurrence may not truly indicate a substorm. Yet, the validity of any study of the physics of substorms or the effect of substorms on the magnetosphere-ionosphere-thermosphere system requires consensus on what constitutes a substorm event and a consistent catalog of events. To make progress toward this, in this study, we examine the extent to which various substorm onset lists identify events at similar times by quantifying the levels of association between pairs of onset lists. A binary forecast verification framework is used, applying pairwise comparison to the onset lists from a wide variety of instrumentation including ground-based magnetometers (Borovsky & Yakymenko, 2017; Chu et al., 2015; Forsyth et al., 2015; McPherron & Chu, 2018; Newell & Gjerloev, 2011; Ohtani & Gjerloev, 2020), spacecraft auroral imagers (Frey et al., 2004) and in-situ spacecraft (Borovsky & Yakymenko, 2017) between 18 May 2000 and 31 December 2002, a common time period in which all onset lists based on different phenomena can contribute. We used the Heidke Skill Score (HSS) to examine the extent to which events in one list have a corresponding event in another within a conservative estimation of an expansion phase lifetime.

2. Substorm Identification Methods

The term substorm was first published in Akasofu and Chapman (1961), where it was defined as a magnetic disturbance in the auroral zones caused by currents flowing in the ionosphere characterized by fluctuations faster than those of a geomagnetic storm. Akasofu (1964) defined the auroral substorm, providing much of the framework we use to study substorm dynamics, describing its onset, as well as the expansion and recovery phases of evolving auroral morphology that coincide with the polar magnetic disturbance. The auroral and magnetic observations were joined by Akasofu (1968) in a framework called the magnetospheric substorm, although previous studies had noted a link between changes in auroral activity and magnetic disturbances (Heppner, 1954). Subsequent studies have further developed the concept of substorms to include injections of energetic particles into the inner magnetosphere (Kamide & McIlwain, 1974; Lezniak et al., 1968), the rapid reconfiguration of the nightside magnetic field to one that is more dipolar (Cummings & Coleman, 1968), and the formation and ejection of a plasmoid (Hones, 1984; Hones et al., 1986). These events coincide with the reduction of magnetic flux in the magnetotail lobes and the diversion of the tail current along magnetic field lines into the auroral ionosphere to form a field-aligned current system known as the substorm current wedge (SCW) (Kepko et al., 2015; McPherron et al., 1973). The field-aligned currents of the SCW cause a characteristic ground magnetic perturbation at midlatitudes and the closure of the SCW in the westward auroral electrojet result in the Disturbance Polar type 1 (DP1) ground magnetic deflection at higher latitudes (Akasofu et al., 1965; Forsyth et al., 2018; Heppner, 1954; Murphy et al., 2022).

2.1. Auroral Signatures

Methods using the aurora as a proxy for substorms still rely on the Akasofu (1964) definition of the onset of the expansion phase as a sudden brightening of the equatorward edge of a quiet auroral arc, followed by a rapid poleward and azimuthal expansion of this arc. Using this definition, Frey et al. (2004), Frey and Mende (2006), Nishimura et al. (2010), Liou et al. (2001) have visually identified onsets from successive auroral images. This relies on the observer's judgment, introducing a degree of subjectivity and lack of reproducibility to the lists created. The lists also differ in the instruments used: Nishimura et al. (2010) used the ground-based THEMIS All-Sky Imager (ASI) array (S. B. Mende et al., 2008) in which each imager takes an image every 3 s with a spatial resolution of 1 km per pixel over an all-sky field of view. The aurora can only be observed in the absence of clouds, and the array covers a limited portion of the auroral oval. Frey et al. (2004), Frey and Mende (2006) used the space-based Imager for Magnetopause-to-Aurora Global Exploration (IMAGE) Far Ultraviolet (FUV) global auroral imager (S. Mende et al., 2000) which can view the entire northern auroral oval at a spatial resolution of up to 50 km per pixel every 2 min. Liou et al. (2001) used the space-based Polar Ultraviolet Imager (UVI) (Torr et al., 1995) with similar capabilities. Based on the superior coverage of space-based imagers and the overlap in time with other catalogs below, we have selected the Frey et al. (2004) onset list for our analysis.

2.2. Ground Magnetic Perturbations

Another class of identification methods are those that make use of ground-based magnetometer arrays. These methods use the Midlatitude Positive Bay (MPB) (Chu et al., 2015; McPherron & Chu, 2018) or SuperMAG AL (SML) (Borovsky & Yakymenko, 2017; Forsyth et al., 2015; McPherron & Chu, 2018; Newell & Gjerloev, 2011; Ohtani & Gjerloev, 2020) indices to identify substorms, both at 1 min cadence. The SML index is the generalization to many stations of the AL index developed by Davis and Sugiura (1966), leveraging the SuperMAG amalgamation of magnetometer arrays. This index traces the minimum deflection of the northward component of the geomagnetic field across a range of auroral latitude magnetometer stations. The motivation behind the development of the index was to study auroral zone activity by maximizing the auroral electrojet (DP1) contribution to the index while reducing any contributions from the ring current. The MPB index was developed by Chu et al. (2015) using the INTERMAGNET magnetometer chain of stations between 20° and 53° latitudes in both hemispheres, positioned further away from the auroral oval at midlatitudes in order to detect the disturbance caused by the large-scale field-aligned currents of the SCW (McPherron et al., 1973).

The methods applied to the MPB and SML lists can be split into two schemes: those leveraging a significant change of the raw index for identification (Borovsky & Yakymenko, 2017; Chu et al., 2015; Forsyth et al., 2015; Newell & Gjerloev, 2011; Ohtani & Gjerloev, 2020), and techniques applying a standardization procedure on the index before inspecting the new index for deviations that could be related to substorms (McPherron & Chu, 2018).

The raw index methods identify substorm onsets when the SML or MPB index exceeds a threshold gradient or integrated value. Henceforth, we refer to the subset of methods that utilize changes in the gradient of SML for identification as SML gradient lists. Some of these methods include an explicit minimum possible recurrence time in the procedure, such that another onset cannot be identified in the vicinity of an already selected event (Borovsky & Yakymenko, 2017; Chu et al., 2015; Newell & Gjerloev, 2011) or identify only isolated events with a clear growth phase (Ohtani & Gjerloev, 2020). The standardization method developed by McPherron and Chu (2018) is applied to both the SML and MPB indices to create separate onset lists. The indices are standardized by subtracting the mean of a rolling 4 hr interval and dividing by its standard deviation before identifying onsets as significant deviations from this normalized index. It should be noted that for the SML-based methods, it is possible for the contributing station to SML at the time of substorm onset to be located at dayside magnetic local times (MLTs), such that it is highly improbable that the detected magnetic perturbation is directly linked to the substorm phenomenon. It is possible to obtain substorm-related SML enhancements away from the nightside and toward dawn; however, caution should be taken when using lists using this magnetic index. One list that attempts to filter out these possible misidentifications is the Ohtani and Gjerloev (2020) list, which filters for events occurring within 20-03 MLT only.

2.3. Particle Injections

The final type of signature that we use for the identification of substorms is the Earthward injection of energetic electrons and protons observed at geosynchronous orbit (Lezniak et al., 1968), which previous studies have temporally associated with auroral signatures of substorms (Kamide & McIlwain, 1974; Weygand et al., 2008; Yeoman et al., 1994). Borovsky et al. (1993) visually identified substorm onsets as the sudden increase in energetic particle flux after a “drop-out” that is associated with the growth phase observed by at least two geosynchronous satellites located nearest to magnetic midnight. A more recent study has applied an automated method to identify substorm onset-related energetic electron injections using the specific entropy of the hot electron population observed at geosynchronous orbit as a proxy (Borovsky & Yakymenko, 2017). The onset list applying this automated method is produced at a 30 min cadence, in comparison to 1 min for the ground magnetometer array lists and 2 min for the Frey et al. (2004) IMAGE list.

3. Data and Method

3.1. Substorm Onset Lists

We studied the association between the different substorm onset signatures by applying pairwise dichotomous forecast verification comparisons to onsets identified by the above methods between the dates of 18 May 2000 to 31 December 2002, which is the interval of the Frey et al. (2004) auroral onset list. The published lists are Frey et al. (2004), Newell and Gjerloev (2011), Chu et al. (2015), Forsyth et al. (2015), Borovsky and Yakymenko (2017), McPherron and Chu (2018), and are summarized in Table 1.

For the Forsyth et al. (2015) Substorm Onsets and Phases From Indices of the Electrojet (SOPHIE) method, substorm onsets are identified when a negative change in the low-pass filtered SML index exceeds an Expansion Percentile Threshold (EPT), defined as a given percentile of the distribution of observed negative rates of change of SML in each year. Onset lists were published for three different EPT values—50%, 75%, and 90%. In this study, we extend the range between 5% and 95%, in order to (a) evaluate the EPT value for which the SOPHIE onsets are best associated with events in each of the other onset lists and (b) thereby infer the relative sensitivity of the other methods. The discrete EPT thresholds used for this study are shown in Table 2, with the corresponding number of onsets ranging from 1724 to 10,035. For the non-SOPHIE substorm lists analyzed, there are also a varying number of onsets observed for the period analyzed, from 1960 (BY2017 SML) to 8601 (MC2018 MPB) events, as can be seen in Table 1.

3.2. Windowing Event Lists for Contingency Table Analysis

As described in Section 3.3, the analysis carried out here made use of contingency tables (or truth tables) to evaluate the coincidence between events identified by different substorm identification techniques. To create the contingency tables, we divided our analysis period into consecutive windows, 30 min in length. This setup is shown in Figure 1a, where two example time series of events are displayed, one labeled the Reference list and another the Comparison list, with events from each highlighted as triangles. At this timescale, we populated a

Table 2
Number of Onsets Identified Using Different Thresholds of the SOPHIE Technique

Expansion percentile threshold	Number of onsets
05	10,035
10	10,141
15	10,094
20	10,073
25	9,943
30	9,769
35	9,608
40	9,349
45	9,160
50	8,738
55	8,404
60	7,986
65	7,433
70	6,754
75	6,012
80	5,306
85	4,343
90	3,245
95	1,724

contingency table for a pairwise comparison by noting whether an event was seen only in a Reference list, only in a Comparison list, in neither or in both within each 30 min time window. The 30 min timescale was chosen because it is the identification cadence of the slowest method analyzed—the particle injections identified by Borovsky and Yakymenko (2017), and is the typical time for a substorm expansion phase to develop and for the onset signatures to spread through the magnetosphere.

Since the position in time of our windows is arbitrary, it is possible for events that are within 30 min of each other to appear in different windows (see, for example, the last events in both lists in Figure 1a). This will result in close events coming up as misses, returning windows as false positives and false negatives. To mitigate this, we applied a pseudo-bootstrap method. We implemented this by incrementing the start time by 1 min steps up to the 30 min length of our windows. Figure 1b shows an example of pseudo-bootstrap method applied to the procedure set up in Figure 1a, where we have stepped by half of the comparison window length. As can be seen from the figure, the resultant window labeling varies. As a result of this method, there will be 30 versions of the contingency table created since we use a window length of 30 min. To apply the metrics stated in Section 3.3, we took the median values of the contingency tables for the final pairwise comparison. We evaluated the median numbers of true positives, false positives, false negatives and true negatives of the 30 realizations to be used in the metrics describing event association rather than simply summing them to avoid multiple counting event coincidence or disagreement between the lists.

For comparisons with the Fr2004 IMAGE list, adjustments of the analysis were made to compensate for IMAGE having sight of the entire northern auroral oval for only 8–10 hr of every orbit (Frey et al., 2004). The time

windows in the pairwise comparison when IMAGE was unable to observe the northern auroral oval were filtered out, which still left ~32,000 of the ~46,000 30 min windows. Additionally, the BY2017 PI list indicates a 30 min interval in which an injection is observed, where they have marked the beginning of the interval as their epoch times (Borovsky and Yakymenko, Personal Communication), this was compensated for by adding 15 min to the epoch times to not bias toward the beginning or end of their identified intervals.

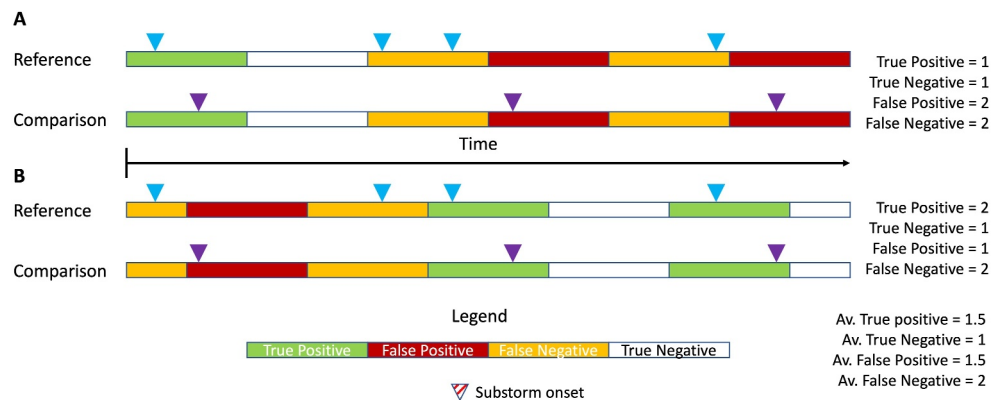


Figure 1. Schematic of time series windowing method used to populate the contingency tables of the pairwise comparisons of association. Here, the blue triangles indicate an event in the Reference “time series” and purple an event in the Comparison “time series.” In the pairwise comparison shown by (a), we see the classification of the windows dependent on whether an event from either is observed in them or not. The rows denoted by (b) shows our bootstrap method to account for events that coincide within 30 min but may have been classed in different windows by the segmentation of the time series. The median of the 30 min-stepped window labellings are what are used as the elements of the contingency tables.

Table 3
2 × 2 Contingency Table for Binary Forecasts (Hogan & Mason, 2011)

		Reference	
		Yes	No
Comparison	Yes	<i>a</i> (True Positive)	<i>b</i> (False Positive)
	No	<i>c</i> (False Negative)	<i>d</i> (True Negative)

3.3. Contingency Tables and Verification Metrics

In order to evaluate the association between the different onset lists, we make use of metrics predominantly used in the verification of dichotomous meteorological forecasts. In recent years, these techniques have been more regularly used in the field of space physics, particularly in the forecasting of space weather events and evaluation of space weather models (Haiducek et al., 2020; Mooney et al., 2021; Smith et al., 2021). This method of analysis is displayed in a two-by-two contingency table, an example of which is shown

in Table 3, from which the metrics are assembled from the elements, *a*, *b*, *c* and *d*. We created the contingency tables by windowing the time series as described in Section 3.2. In the upper left corner of the contingency table are the True Positives (*a*): the number of windows in which both a reference (List A) substorm onset and a comparison (List B) onset are observed. In the same row are the False Positives (*b*): the number of windows in which no reference onset was observed, but a comparison onset was detected. In the bottom right corner are the True Negatives (*d*): the number of windows in which no substorm onset was observed in either list. Finally, there are the False negatives (*c*): the number of windows in which a substorm was observed in the reference list (List A) only.

From the elements of the contingency table, it is possible to construct several metrics that summarize the level of association between the two lists that are compared. In this study, we use the HSS to quantify the association between the onset lists analyzed; also known as Cohen's Kappa outside the meteorology literature (Cohen, 1960). This metric is derived from the Proportion Correct (*PC*), a measure of the correct predictions compared to the total number of predictions, given by

$$PC = \frac{a + d}{a + b + c + d}. \quad (1)$$

A proportion correct of 1 indicates a perfect association between two onset lists. The HSS is derived from *PC* by comparing *PC* to a reference value, PC_{ref} , such that PC_{ref} is the Proportion Correct that would be obtained by random coincidence of the events populating each list if they were statistically independent of each other, and is given by:

$$PC_{ref} = p(\text{Reference} = \text{Yes})p(\text{Comparison} = \text{Yes}) + p(\text{Reference} = \text{No})p(\text{Comparison} = \text{No}). \quad (2)$$

Its maximum likelihood estimate from the elements of the contingency table is:

$$PC_{ref} = \left(\frac{a + b}{n}\right)\left(\frac{a + c}{n}\right) + \left(\frac{b + d}{n}\right)\left(\frac{c + d}{n}\right), \quad (3)$$

where $n = a + b + c + d$, is the sum of the elements of a given contingency table.

The HSS is then formulated as:

$$HSS = \frac{PC - PC_{ref}}{PC_{perfect} - PC_{ref}} = \frac{PC - PC_{ref}}{1 - PC_{ref}}, \quad (4)$$

which after inputting our expressions for *PC* and PC_{ref} gives:

$$HSS = \frac{2(ad - bc)}{(a + b)(b + d) + (a + c)(c + d)}. \quad (5)$$

The HSS has a range of -1 to 1 , with 1 indicating a perfect association, 0 equivalent to random or no association, and -1 anti-association between two onset lists (where the comparison perfectly identifies events in windows that the reference has not and vice versa). Note that the HSS is the same no matter which of the two lists is the reference or comparison, that is, the score is the same when *b* and *c* are swapped in expression 5.

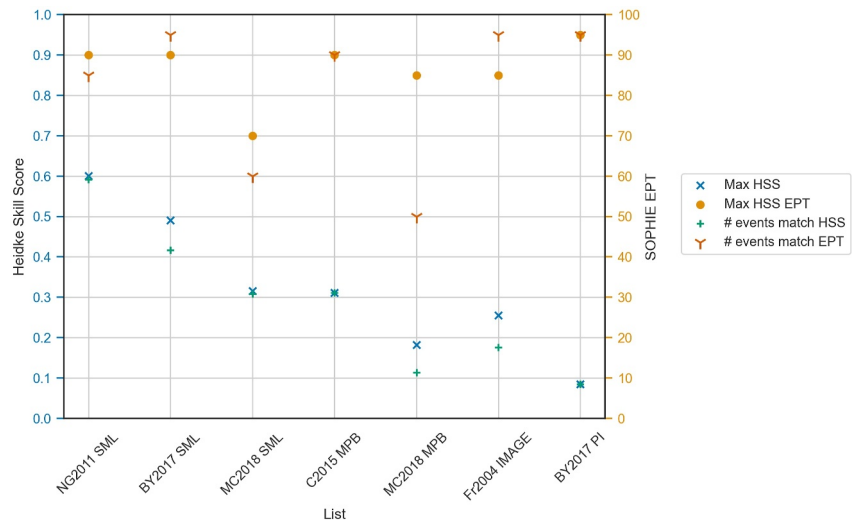


Figure 2. The results of comparing various onset lists against various sensitivities of the SOPHIE technique. The blue crosses indicate the maximum Heidke Skill Score achieved by the SOPHIE technique, and light orange dots indicate which tuning of the technique achieved this maximum score. The dark orange tri-markers denote the Expansion Percentile Threshold (EPT) that has the closest correspondence to the event count of the compared list, with the green pluses being the score achieved by that EPT.

In addition to the HSS, we also make use of the True Positive Rate (TPR), also called Sensitivity. The TPR is the probability that a window containing an onset identified by the reference method is also populated by an event from the comparison method. In terms of the elements of the contingency table, it is given by:

$$\text{TPR} = \frac{a}{a + c}. \quad (6)$$

4. Results

4.1. Intercalibration of the SOPHIE Technique With Other Substorm Onset Identification Methods

We first examine the effect of SOPHIE detection sensitivity on the association between the SOPHIE technique and other lists. Figure 2 shows the results of this analysis, where we have plotted the maximum HSS achieved by the SOPHIE technique against the other event lists, as well as the EPT that achieves this score. As explained in Section 3.3, the HSS quantifies the improvement over the random chance that events in two lists coincide.

Figure 2 shows that the SOPHIE technique is able to identify substorm onsets associated with those observed by different methods with some degree of skill, achieving Heidke skill scores ranging from 0.08 with BY2017 PI to 0.60 with NG2011 SML. The best associations are obtained with detection methods using the same SML data set as SOPHIE, obtaining scores of 0.31–0.60. Associations of SOPHIE with MPB and auroral images are lower, with scores between 0.18 and 0.31, whilst the association of SOPHIE with BY2017 PI is only marginally better than random chance. For the methods using either the SML index or the MPB index, the skill score is among the lowest for the MC2018 method that applies a normalization procedure, presumably because SOPHIE is based on the unnormalized SML index.

The value of SOPHIE EPT that provides the maximal HSS is mostly relatively high at 80%–95%. This implies that non-SOPHIE onsets associate best with the largest changes in the low-pass filtered SML used by SOPHIE, and specifically in the upper 20% of all such changes. This is true for all but the MC2018 SML onset list, where a HSS of 0.32 is achieved with a lower EPT = 70%. The MC2018 SML list has the second-highest number (=7,908) of identified events, and so it may be thought that in this case the best associating EPT is low in order that the number of SOPHIE events matches the high number of MC2018 SML events. By comparing Tables 1 and 2, we see that to within the 5% EPT resolution, the closest number of SOPHIE events (=7,986) to that of MC2018 SML (=7,908) is indeed found for a low EPT = 60%, but this is lower still than the EPT = 70% that maximizes the HSS, for

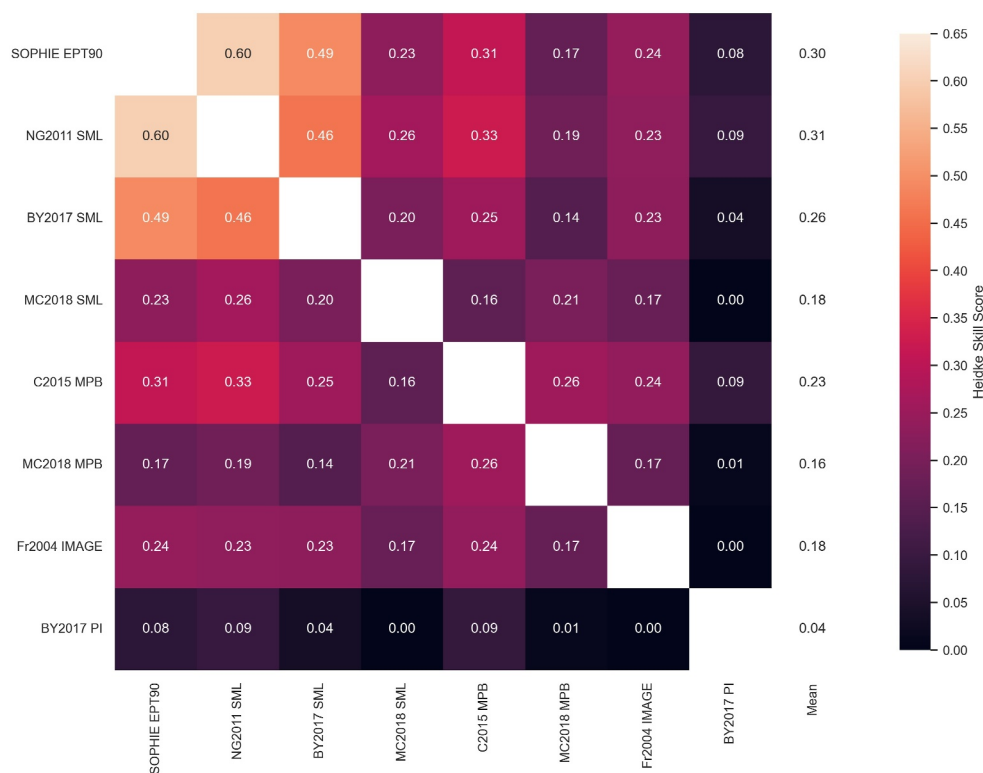


Figure 3. A heatmap of the Heidke Skill Scores achieved by every cross-comparison. Lighter colors indicate a higher level of event coincidence between the two onset lists. It is symmetric along the main diagonal by a property of the HSS formulation. The last column is the mean HSS achieved by that onset lists for all its comparisons.

which SOPHIE identifies 6,754 events, some 1,154 or 15% fewer events than MC2018 SML. To investigate this further, we plot in Figure 2 the EPT that has the closest number of events to each of the other compared lists, and the HSS this EPT achieved. We see that in general the scores are not simply maximized by minimizing the difference in the number of events between the two lists, as for five of the seven list comparisons the EPT that scores best does not have the most similar number of events. That is, it is not only the number of events in each list, but also their relative timing that matters. Indeed, if the reference and comparison lists had the same number of events but the events in one of the lists were purely random in time, then the events would be statistically independent of each other and the HSS would be zero by definition.

4.2. Relationship Between Different Onset Methods

Figure 3 shows a “heatmap” of the Heidke skill scores obtained when comparing all the onset lists against one another. It should be noted that this analysis now uses a single SOPHIE sensitivity in order not to bias the results toward the SOPHIE technique by including multiple SOPHIE onset lists with different EPTs. EPT 90 was selected as it was the version that scored highest on average when compared against the other non-SOPHIE onset lists. (The scores achieved by each SOPHIE threshold with the other onset lists is provided in the Supplementary Information.) We have quoted scores to the 2nd significant figure as the uncertainty in the Heidke skill scores shown is in the order of 10^{-3} , assuming that event occurrence is approximately independent, stationary, and rare such that the standard error of the count in each element of the truth table is the square root of the count based on Poisson counting statistics. As the HSS is invariant to which event list is used as the comparison or reference list, the heatmap produced is symmetric about the main diagonal. The mean of the Heidke Skill Scores achieved by one onset list with each of the others is shown on the rightmost column of the heatmap.

The highest scores are obtained when comparing the SML gradient lists that do not filter out events by MLT (SOPHIE EPT90, NG2011 SML and BY2017 SML) with one another (0.46–0.60) although none of them are an exact match due to the differing nuances of each algorithm, for example, smoothing of the raw SML signal,

gradient threshold, hard-coded delay between identification of successive onsets, etc. Furthermore, these SML gradient lists that do not filter by MLT also achieve the highest mean scores of 0.26–0.31, although these means are biased because they contain cross-comparisons with other SML gradient lists, a problem that we address in Section 4.3.

The next highest mean HSS of 0.22 is achieved by the C2015 MPB method, which is another gradient method but based on the MPB index. It performs particularly well when compared with the SOPHIE technique and NG2011 SML, with scores of 0.31 and 0.33 respectively. The comparison of C2015 MPB with NG2011 is the highest score achieved outside of comparisons between gradient-SML lists.

The normalized method lists of MC2018 show a weaker association with other lists than the gradient-based method, for both the SML and MPB indices. The average HSS is 0.17 for MC2018 SML and 0.18 for MC2018 MPB.

A similar mean HSS of 0.19 is found for the Fr2004 IMAGE event list based on visual identification of auroral onset. Notably, Fr2004 IMAGE performs better when compared against the gradient-based methods.

The OG2020 SML list performs similarly to the other SML gradient methods when compared with Fr2004 IMAGE achieving a HSS of 0.24. However, it achieves a lower mean score compared to the other SML gradient methods, 0.18, similar to the performance of the MC2018 normalized index method. This reduction could be due to the filtering of this method for isolated onsets occurring at 20-03 MLT only. An MLT filter was also applied to the NG2011 SML onset list, using the same MLTs as the OG2020 SML, resulting in a reduced mean HSS, although achieving the same score when compared with Fr2004 and BY2017 PI (results not shown).

The BY2017 PI list obtains the lowest average score with an average HSS of 0.04, with no comparison with the BY2017 PI events achieving a HSS greater than 0.09, showing very little improvement over chance coincidence of events identified from particle injections and those identified by other methods and data. Indeed, comparisons between BY2017 PI and both MC2018 SML or Fr2004 IMAGE obtain a score of 0.00 implying a purely random association. Referring to the previous section, we note that this despite BY2017 PI and Fr2004 IMAGE having almost the same number of onsets (=2,149 and 2,437, respectively).

4.3. Relationship Between Common Index Methods and Other Methods

As SML-based onset lists tend to perform well against each other, including multiple SML-based lists in Figure 3 biases their overall mean skill score compared to that of the other lists. Figure 4 shows a heatmap of the Heidke skill scores achieved by each SML-based list compared to the other non-SML lists. The SML-based lists achieve the best scores with the C2015 MPB list, followed by the Fr2004 IMAGE list. MC2018 MPB achieves its highest score when compared to its equivalent algorithm on the SML data set, MC2018 SML. The BY2017 PI achieves a mean score of 0.05, which implies that, in general, there is little improvement over random chance of agreement between the BY2017 PI and the SML-based lists. We also note that there is comparable performance for the SOPHIE, NG2011 and BY2017 gradient-SML lists when compared to the Fr2004 auroral list.

Comparing the performance of the SML detection algorithms against the other signatures, we find that the best performing list on average is NG2011 SML, with a mean HSS of 0.21. This is only a marginal improvement over SOPHIE EPT90's score of 0.20. The lowest mean HSSs are achieved by the normalization detection method, MC2018 SML, and the MLT-subselected gradient method by OG2020, with a score of 0.13. We also applied the same MLT filter on the NG2011 onset list as was applied for OG2020, only including onsets that the contributing station to SML was within 20-03 MLT. In addition, we tested various values for the cutoffs used to restrict possible onset locations within the NG2011 onset list and found that there was no singular value that increased the scores with other onset lists. As such, we present the results of NG2011 using the same cutoffs as used by OG2020 SML for best comparison. We find that there is a slight reduction in the HSS obtained by a constrained version of NG2011 with the MPB lists of C2015 and MC2018, while the scores with Fr2004 and BY2017 PI remain the same. Thus, restricting the possible MLTs that an NG2011 SML onset can occur at not only removes false negatives and positives but also true positive coincidences. This is evidence that the lower mean score of OG2020 SML is not only due to the restriction in possible MLTs of identification but also due to the other criteria used, such as the requirement for a clear observable growth phase, although it should be noted that (Ohtani & Gjerloev, 2020) does not assert that their onset list is a complete list of substorm events.

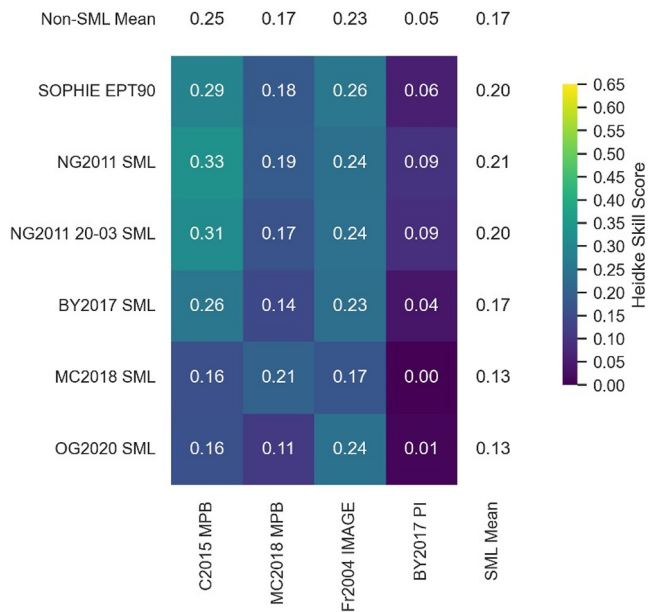


Figure 4. A heatmap of the Heidke Skill Scores achieved by the SuperMAG AL-based onset lists versus lists derived from other data sets. Lighter colors indicate a higher level of event coincidence.

We note that when MC2018 SML was compared with MC2018 MPB, this gave the highest score of all comparisons with MC2018 MPB, however, the improvement in this score over NG2011 SML with MC2018 MPB was marginal. In comparison, there is a much greater difference in score between MC2018 SML and C2015 MPB compared to NG2011 SML and C2015 MPB. As NG2011 was the best performing SML event list, it was the list used for the cross-comparison of distinct signatures only in Section 4.4 below.

As we also have two methodologies applied to the MPB index, we examine their performance too compared to the other lists. To avoid biasing the overall average HSS, we compare these to NG2011 (SML gradient list), MC2018 SML (SML normalized list) as well as Fr2004 IMAGE and BY2017 PI. Figure 5 shows a heatmap of the score achieved by C2015 MPB and MC2018 MPB with the algorithms of the other substorm signatures. C2015 MPB outperforms MC2018 MPB, with an average score of 0.20 compared to 0.14. The scores for C2015 MPB are consistently higher, apart from the comparison with MC2018 SML, where C2015 MPB scores 0.16 compared to MC2018 MPB's score of 0.21. Notably, C2015 MPB outperforms MC2018 MPB in comparison with features not based on either electrojet or field-aligned currents (Fr2004 IMAGE and BY2017 PI). As such, for further comparison, we used the C2015 MPB as the best performing MPB list.

4.4. Relationship Between Distinct Methods

Figure 6 shows the HSS (A) and True Positive Rate (B) heatmaps of distinct substorm identification signatures only. This was done by selecting the highest scoring methods of the SML and MPB data sets, shown in the prior figures, to be compared against each other and with the auroral image event list of Fr2004 IMAGE and the particle injection event list of BY2017 PI. We find very similar performance between NG2011 SML and C2015 MPB, both with a mean HSS of 0.22. The Fr2004 IMAGE list and BY2017 PI lists show some skill in event coincidence with the other signatures, obtaining mean scores of 0.16 and 0.06 respectively. However, our analysis shows that the event coincidence of the auroral onsets and particle injections achieve a score of 0.00, implying that any event coincidence between these two lists is due to random chance.

When looking at the associated TPR's of these comparisons in Figure 6b, we can infer the approximate number of events the pairwise comparison of lists have in common by multiplying the TPR by the total number of events in the reference list (provided in the row headers). From this Figure 6b, we see that less than 50% of the events in each list are associated with an event in the other lists. The list with the highest proportion of events associated

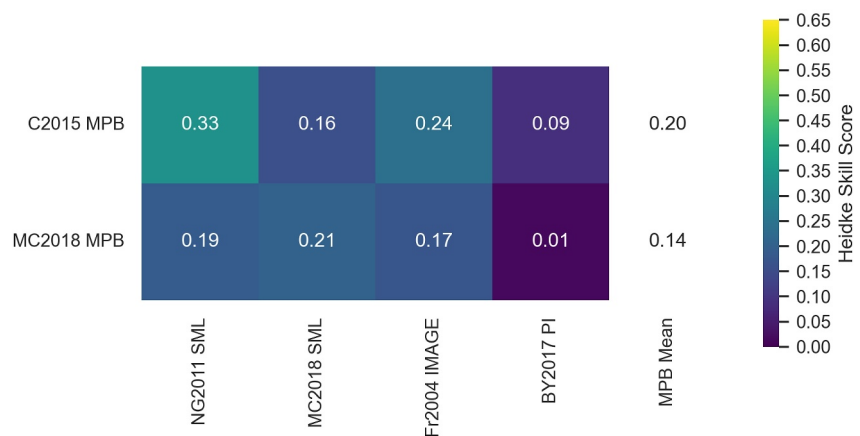


Figure 5. A heatmap of the Heidke Skill Scores achieved by the Midlatitude Positive Bay Index-based onset lists versus lists derived from other data sets. Lighter colors indicate a higher level of event coincidence.

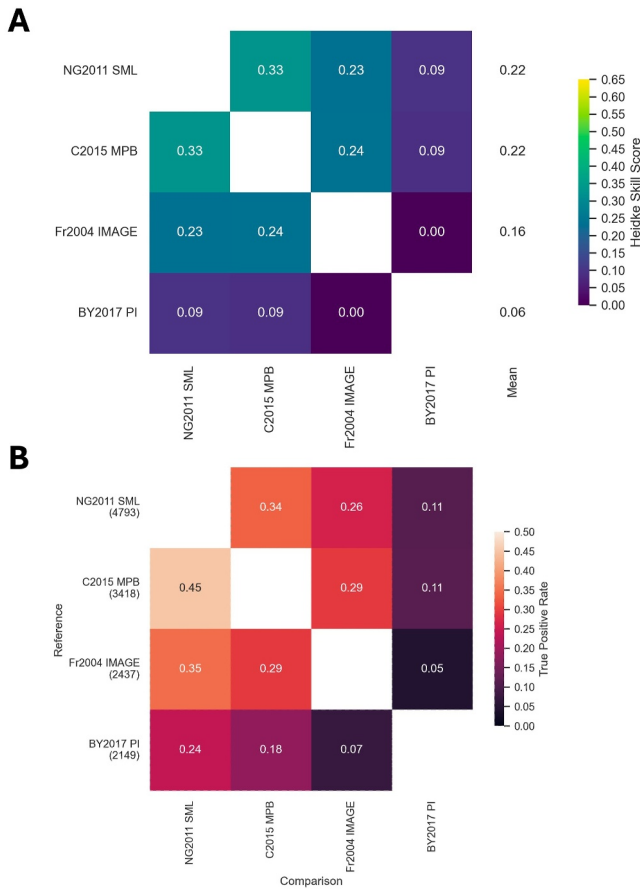


Figure 6. (a) Heidke Skill Score heatmap from distinct data sets only (SML, MPB, Auroral Images, Particle Fluxes at Geosynchronous Orbit). (b) The True Positive Rates (TPR) for each cross comparison. The number in parentheses underneath the row labels is the number of events that populate each list. Note that this is asymmetric, as directionality matters in the formulation of TPR.

with another list is C2015 MPB, for which 45% of the 30 min windows were associated with an event in the NG2011 SML list. In contrast, only 34% of 30 min intervals with a NG2011 SML event were associated with a C2015 MPB event. For the Fr2004 IMAGE list, approximately one third of the 30 min windows were associated with an SML or MPB event, while less than a third of 30 min windows with an SML or MPB event in were associated with an auroral event. Finally, we see that approximately 10% or less of 30 min windows with an auroral, MPB, or SML event are associated with a particle injection event.

5. Discussion

In summary, we have evaluated the association between SML-based, MPB-based, auroral-based, and injection-based substorm onset lists using metrics used in dichotomous forecast verification. In general, we found that lists based on SML or MPB with onsets identified from gradients in these indices showed the greatest association. In contrast, the particle injection list showed a relatively poor association with the other lists, with little or no improvement over random chance.

Typically, we found that less than 50% of the events in each list were associated with events in another list using a different substorm signature, that is, they occurred within the same 30 min time window. This is broadly consistent with a different type of analysis by Forsyth et al. (2015), who examined the probability distribution of the time difference between a SOPHIE onset for various EPT and the nearest onset from another list. They found that only about a third of the onsets of EPT90 occurred within ± 30 min of the onsets on the Frey and Mende (2006) list and about two-thirds of EPT90 onsets occurred within ± 30 min of the onsets in the Newell and Gjerloev (2011) list.

In our analysis, we show that having event lists with comparable numbers is not necessarily indicative of good agreement between them. To illustrate this, we observe the comparisons of the SOPHIE technique for variable EPT and hence number of onsets with the SML and MPB lists based on the MC2018 method. MC2018 SML achieves its highest HSS with SOPHIE EPT 70, but should the HSS be only dependent on comparable numbers, we would expect

the best agreement with EPT 60. Furthermore, MC2018 MPB has a similar number of events as MC2018 SML but best associates with SOPHIE EPT 85 which identifies approximately half the events that MC2018 MPB does. The extreme example of this is that if we create a list of random onsets (i.e., Poisson process with constant probability of occurrence per unit time) with the same number as any observed list, then the occurrence of the random events is independent of the events in the observed list and the HSS is then zero by definition.

We have taken substorm onset lists using techniques and instruments with different data cadences, with the ground magnetometer techniques using the fastest data cadence of 1 min, the IMAGE spacecraft taking snapshots of the aurora every 2 min and the particle injections identifying 30 min intervals. This can induce artificial delays when identifying events; for example, between an identification method with a 1 min cadence data source and another method with a 2 min data source, there can be up to a minute difference in identification even if the event occurred concurrently in both data sources. As such, we experimented using different data cadences with the highest scoring comparison from two different data sources, NG2011 SML and C2015 MPB. The results of this numerical experiment are shown in Table 4, the top row of this table is the original score achieved by the comparison, now shown to 3 decimal places. As evidenced by the experiments, reducing the data cadence to 2 min has a minimal effect on the scores achieved using our method, $\pm 0.3\%$. However, the difference when the data cadence is reduced to 30 min is more significant, capable of reducing the score achieved by 10%. This reduction affects the comparisons with the particle injections, the lowest scores achieved in our analysis. However, even when accounting for a reduction of 10% in the HSS, these scores would remain the poorest and as such the data cadence differences have less of an effect on possible misidentification than other factors.

Table 4
Results of the Pairwise Comparison Between NG2011 SML and C2015 MPB at Different Data Cadences

NG2011 data cadence	C2015 data cadence	Heidke skill score	% Change
1 min	1 min	0.327	0.0
1 min	2 min	0.326	0.3
1 min	30 min	0.294	10.1
2 min	1 min	0.328	0.3
2 min	30 min	0.293	10.4
30 min	1 min	0.296	9.5
30 min	2 min	0.292	10.7

The strongest association that we determined was between SML and MPB-based lists, when the identification of events leveraged gradients in these indices. For these lists (C2015 MPB and NG2011 SML), we found ~1500 coincident events (45% of the MPB list and 34% of the SML list). That these lists showed the strongest associations is to be expected, as they effectively attempt to identify different aspects of the current systems in the SCW (Kepko et al., 2015). The fact that nearly two-thirds of the SML events are not associated with an apparent enhancement in MPB is intriguing. This may arise from a number of considerations. Firstly, the auroral zone magnetometers are much closer to the current system they are attempting to characterize, since the electrojets are flowing in the ionosphere at ~100 km whereas the field-aligned currents that provide the MPB signature are thousands of kilometres from their detecting stations. As such, small-scale features and enhancements may be unresolved by MPB stations, resulting in them missing

SML “events.” However, the fact that over half of the MPB events do not have an associated SML event means that this cannot be the only explanation.

The index normalization method by McPherron and Chu (2018) obtains weaker associations compared to gradient identification methods on either index, even when isolating scores against another substorm signature such as Fr2004 IMAGE. The MC2018 method identifies many more events compared to the other non-SOPHIE list, with 7,908 and 8,601 events for the SML and MPB index, respectively. The next closest method, NG2011 SML only identifies 4,793 events during this period, a reduction of ~40–45%. The lower levels of association seen then arise from the increased number of false positives in comparison to the other methods. This could be due to the normalization procedure identifying small perturbations during particularly quiet periods in either index as events, such that non-substorm fluctuations are misclassified as substorm onsets. Further evidence that substorm identification of smaller magnetic fluctuations is a possible cause of the weak association compared to the other lists is that from Figure 2, we observe that MC2018 SML's best agreement is with a lower SOPHIE threshold than the other lists, corresponding to requiring smaller gradients in SML to identify substorm onset. It should be noted that the normalization procedure could also have the effect of suppressing substorm-like perturbations in either index during particularly active times, resolving these intervals as false negatives compared to the other lists, causing the observed reduction in association.

Identification of the substorms by eye from auroral data has been the basis for much of the statistical analysis of substorm phenomenon, and is considered by some to be the most robust identification. Overall, there are far fewer auroral onsets than MPB or SML onsets, even when one considers the viewing restrictions of the IMAGE spacecraft. However, we still find that over two thirds of the identified auroral onsets do not correspond to an event in the MPB or SML lists. This may imply that many of the Fr2004 events are relatively small substorms. In fact, Forsyth et al. (2015) found that the average SML profile of the extended IMAGE onset list (Frey & Mende, 2006) was one of the weakest of the event lists they tested (reaching one of the least negative SML values at peak activity), second only to onsets identified from the THEMIS ASIs (Nishimura et al., 2010) (see Figure 4 in Forsyth et al., 2015). However, we also find that there are comparatively large SML and MPB perturbations that were not associated with auroral events in the Frey et al. (2004) list. These may be expansion phase onsets during or immediately after substorm recovery phases, which may be tricky to identify by eye because of the complex auroral activity already occurring.

Our results show that, in general, there is a poor association between particle injections and auroral, MPB, or SML signatures. This is somewhat surprising, given that substorm particle injections are considered a key driver in radiation belt enhancements (Forsyth et al., 2016; Jaynes et al., 2015). Borovsky and Yakymenko (2017) discuss that the onset times, which mark the beginning of the 30 min interval in which an injection is observed, may be systematically delayed by 30 min from the events they identified from the SML index. This is because particle injections must be detected in situ, whereas all other signatures are detected remotely. More specifically, if there is no Los Alamos National Laboratory (LANL) spacecraft within vicinity of the injection region and there is an injection at geosynchronous orbit, the injected electrons must drift toward eastward until a spacecraft is encountered. Furthermore, when using the end of the 30-min interval as the epoch time for the onset of the substorm, the Heidke skill scores for BY2017 PI were improved, in agreement with the hypothesis that there may be a systematic delay between these injections and other signatures. However, the scores achieved were still lower

than those attained by other methods, even when accounting for the data cadence effects as described above. As such, spacecraft spacing and electron drifts are not the only considerations for misidentification using injections. It is conceivable that injection events are occurring within the magnetosphere but are sufficiently localized that they pass between spacecraft and are injected further into the inner magnetosphere, or do not reach geosynchronous orbit and thus are not detected by the spacecraft. Rodger et al. (2022) observes that precipitating flux tends to penetrate further into the inner magnetosphere for larger geomagnetic events (see Figure 1 in Rodger et al., 2022), thus leveraging Low Earth Orbit spacecraft as an additional data source for injections may be required, as they can cross multiple L-shells quickly.

Most of the onset lists that we have used in this study were developed in isolation to one another. As such, there was little in the way of intercalibration to indicate that the thresholds used were appropriate and comparable to other substorm identifications. The exception to this was the study by Forsyth et al. (2015), which explicitly compared the results of multiple thresholds with similar results from existing onset lists. We have taken this one step further by determining the thresholds for the SOPHIE algorithm that provide lists with the greatest association to the other lists we have considered. Overall, we find that well-performing SML lists and the C2015 MPB list have the greatest association with the SOPHIE EPT 90 list. However, the Fr2004 IMAGE list has the greatest association with the SOPHIE EPT 85 list, somewhat higher than the EPT 75 found by Forsyth et al. (2015). Although this may imply that there is a certain degree of accidental intercalibration, the differences in peak HSS indicate that different subsets of the event lists agree with the SOPHIE EPT 90 events. Adjusting thresholds to identify a similar set of events will improve agreement quantified with verification metrics, and the exploration of differences in either onset number of the events that are misidentified by either list could improve understanding of substorm and magnetotail dynamics.

As part of our analysis, we examined the associations between similar techniques applied to the same data sets and the associations of different techniques applied to the same data sets with lists from different signatures. Typically, we found that the lists that used the normalization method of McPherron and Chu (2018) had a lesser association with the auroral image and the particle injection lists than the lists that used gradients of SML and MPB. We found that the lists that achieved the highest associations were NG2011 SML and C2015 MPB, as such, we would recommend the use of these onset lists when conducting statistical studies of magnetospheric dynamics surrounding substorm onset. It should be noted that the C2015 MPB list is only available from 1981 to 2015, thus covering missions such as IMAGE, Polar, and Cluster II but missing a large part of the lifetimes of Van Allen Probes, MMS, and any future missions. However, we note that both these lists only provide substorm onset times, whereas the Forsyth et al. (2015) SOPHIE algorithm, which scores only marginally lower than NG2011 SML when compared to non-SML methods, provides complete identification of the time series of substorm phases, in addition to leveraging the SMU data set to filter out large SML fluctuations that may be due to enhanced convection.

6. Summary

In this study, we have evaluated the event coincidence of various substorm identification methods using a technique derived from contingency tables for the period between May 2000 and December 2002. We found that while there was skill in event association, this varied depending on the method and signature used. The best scoring techniques were gradient methods applied to ground magnetometer data. However, we found that even the methods that obtain the highest levels of association, Newell and Gjerloev (2011) and Chu et al. (2015), have fewer than 50% events in common to both of them within the lifetime of the expansion phase. If one were to study the dynamics of the substorm using only one onset list, we recommend using either of the lists by Newell and Gjerloev (2011) or Chu et al. (2015) if the authors are only interested in onset timing, and the SOPHIE technique by (Forsyth et al., 2015) using an EPT of 90 when considering all phases of the substorm. Although it has previously been viewed as a robust signature of substorms, particle injections at geosynchronous orbits obtained the lowest scores in our analysis, showing poor associations with the other lists analyzed.

Clearly, there is a need to cross-calibrate these identification methods and narrow the quantitative definition of the substorm, so as not to come to conclusions about magnetospheric dynamics when using just a subset of substorm events or intervals that may not correspond to substorms. Although there have been some efforts to calibrate these onset lists, they have been done such that a similar number of events is identified by a method to some given list, historically the auroral identification onset lists, as they contain the fewest false positives, or to reduce the time

differences of the nearest onset identified by a new method to a reference list, for example, Forsyth et al. (2015). Another possible avenue is to define the occurrence of an onset using a combination of data sets, rather than using one signature for substorm onset. The difficulty of this is to account for the differences in data cadences, coverage, and the time interval covered by the different data sets being leveraged.

Data Availability Statement

The substorm onset lists analyzed in this work were obtained from:

- Newell and Gjerloev (2011), Ohtani and Gjerloev (2020), and Frey et al. (2004): <https://supermag.jhuapl.edu/substorms/>,
- Borovsky and Yakymenko (2017): <https://doi.org/10.1002/2016JA023625>,
- Chu et al. (2015) and McPherron and Chu (2018): <https://doi.org/10.1002/2017JA024766>.

The Forsyth et al. (2015) SOPHIE substorm phase lists analyzed are available in the supplementary material and at <https://doi.org/10.5522/04/25672014.v1>.

Acknowledgments

We gratefully acknowledge the substorm timing list identified by the Newell and Gjerloev technique (Newell & Gjerloev, 2011), the substorm timing list identified by the Ohtani and Gjerloev technique (Ohtani & Gjerloev, 2020), the SMU and SML indices (Newell & Gjerloev, 2011); and the SuperMAG collaboration (Gjerloev, 2012). We acknowledge the substorm timing list identified by Frey et al. (2004). We acknowledge the substorm timing lists identified by Borovsky and Yakymenko (2017). This data is publicly available as auxiliary material to the named publications. We acknowledge the substorm timing lists identified by McPherron and Chu (2018). This data is publicly available as auxiliary material to the named publications. We acknowledge the MPB index developed and the substorm timing list identified by Chu et al. (2015). This data is publicly available as auxiliary material to the named publications. CJL was supported by the Science and Technology Facilities Council (STFC) Studentship 2573668. CF was supported by NERC grants NE/V002724/1 and NE/V002554/2. MPF was supported by NERC grants NE/V002716/1 and NE/V015133/1. AWS was supported by the NERC Independent Research Fellowship NE/W009129/1. MKM was supported by Science and Technology Facilities Council (STFC) Grant ST/S000429/1. The analysis on this paper was performed using Python including the Pandas (Wes McKinney, 2010), NumPy (Harris et al., 2020), Matplotlib (Hunter, 2007), Scikit-learn (Pedregosa et al., 2011) and Seaborn (Waskom, 2021) libraries.

References

Akasofu, S. I. (1964). The development of the auroral substorm. *Planetary and Space Science*, 12(4), 273–282. [https://doi.org/10.1016/0032-0633\(64\)90151-5](https://doi.org/10.1016/0032-0633(64)90151-5)

Akasofu, S.-I. (1968). *Polar and magnetospheric substorms* (Vol. 11). Springer Netherlands. <https://doi.org/10.1007/978-94-010-3461-6>

Akasofu, S.-I., & Chapman, S. (1961). The ring current, geomagnetic disturbance, and the Van Allen radiation belts. *Journal of Geophysical Research* (1896-1977), 66(5), 1321–1350. <https://doi.org/10.1029/JZ066i005p01321>

Akasofu, S. I., Chapman, S., & Meng, C. I. (1965). The polar electrojet. *Journal of Atmospheric and Terrestrial Physics*, 27(11), 1275–1305. [https://doi.org/10.1016/0021-9169\(65\)90087-5](https://doi.org/10.1016/0021-9169(65)90087-5)

Belian, R. D., Cayton, T. E., & Reeves, G. D. (1995). Quasi-Periodic global substorm generated flux variations observed at geosynchronous orbit. In *Space plasmas: Coupling between small and medium scale processes* (pp. 143–148). American Geophysical Union (AGU). <https://doi.org/10.1029/GM086p0143>

Boakes, P. D., Milan, S. E., Abel, G. A., Freeman, M. P., Chisham, G., & Hubert, B. (2011). A superposed epoch investigation of the relation between magnetospheric solar wind driving and substorm dynamics with geosynchronous particle injection signatures. *Journal of Geophysical Research: Space Physics*, 116(A1). <https://doi.org/10.1029/2010JA016007>

Borovsky, J. E., Nemzek, R. J., & Belian, R. D. (1993). The occurrence rate of magnetospheric-substorm onsets: Random and periodic substorms. *Journal of Geophysical Research: Space Physics*, 98(A3), 3807–3813. <https://doi.org/10.1029/92JA02556>

Borovsky, J. E., & Yakymenko, K. (2017). Substorm occurrence rates, substorm recurrence times, and solar wind structure. *Journal of Geophysical Research: Space Physics*, 122(3), 2973–2998. <https://doi.org/10.1002/2016JA023625>

Cai, X., Henderson, M. G., & Clauer, C. R. (2006). A statistical study of magnetic dipolarization for sawtooth events and isolated substorms at geosynchronous orbit with GOES data. *Annales Geophysicae*, 24(12), 3481–3490. <https://doi.org/10.5194/angeo-24-3481-2006>

Chu, X., McPherron, R. L., Hsu, T.-S., & Angelopoulos, V. (2015). Solar cycle dependence of substorm occurrence and duration: Implications for onset. *Journal of Geophysical Research: Space Physics*, 120(4), 2808–2818. <https://doi.org/10.1002/2015JA021104>

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>

Cummings, W. D., & Coleman, P. J., Jr. (1968). Simultaneous magnetic field variations at the earth's surface and at synchronous, equatorial distance. Part I. Bay-associated events. *Radio Science*, 3(7), 758–761. <https://doi.org/10.1002/rds196837758>

Davis, T. N., & Sugiura, M. (1966). Auroral electrojet activity index AE and its universal time variations. *Journal of Geophysical Research*, 71(3), 785–801. <https://doi.org/10.1029/JZ071i003p00785>

Forsyth, C., Rae, I. J., Coxon, J. C., Freeman, M. P., Jackman, C. M., Gjerloev, J., & Fazakerley, A. N. (2015). A new technique for determining substorm onsets and phases from indices of the electrojet (SOPHIE). *Journal of Geophysical Research: Space Physics*, 120(12), 10592–10606. <https://doi.org/10.1002/2015JA021343>

Forsyth, C., Rae, I. J., Murphy, K. R., Freeman, M. P., Huang, C.-L., Spence, H. E., et al. (2016). What effect do substorms have on the content of the radiation belts? *Journal of Geophysical Research: Space Physics*, 121(7), 6292–6306. <https://doi.org/10.1002/2016JA022620>

Forsyth, C., Shortt, M., Coxon, J. C., Rae, I. J., Freeman, M. P., Kalmoni, N. M. E., et al. (2018). Seasonal and temporal variations of field-aligned currents and ground magnetic deflections during substorms. *Journal of Geophysical Research: Space Physics*, 123(4), 2696–2713. <https://doi.org/10.1002/2017JA025136>

Freeman, M. P., & Morley, S. K. (2004). A minimal substorm model that explains the observed statistical distribution of times between substorms. *Geophysical Research Letters*, 31(12). <https://doi.org/10.1029/2004GL019989>

Frey, H. U., & Mende, S. B. (2006). Substorm onsets as observed by IMAGE-FUV. In *International Conference Substorms* (Vol. 8, pp. 71–75).

Frey, H. U., Mende, S. B., Angelopoulos, V., & Donovan, E. F. (2004). Substorm onset observations by IMAGE-FUV. *Journal of Geophysical Research: Space Physics*, 109(A10). <https://doi.org/10.1029/2004JA010607>

Gjerloev, J. W. (2012). The SuperMAG data processing technique. *Journal of Geophysical Research: Space Physics*, 117(A9). <https://doi.org/10.1029/2012JA017683>

Haiducek, J. D., Welling, D. T., Morley, S. K., Ganushkina, N. Y., & Chu, X. (2020). Using multiple signatures to improve accuracy of substorm identification. *Journal of Geophysical Research: Space Physics*, 125(4), e2019JA027559. <https://doi.org/10.1029/2019JA027559>

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. <https://doi.org/10.1038/s41586-020-2649-2>

Henderson, M. G., Reeves, G. D., Skoug, R., Thomsen, M. F., Denton, M. H., Mende, S. B., et al. (2006). Magnetospheric and auroral activity during the 18 April 2002 sawtooth event. *Journal of Geophysical Research: Space Physics*, 111(A1). <https://doi.org/10.1029/2005JA011111>

- Heppner, J. P. (1954). Time sequences and spatial relations in auroral activity during magnetic bays at College, Alaska. *Journal of Geophysical Research (1896-1977)*, 59(3), 329–338. <https://doi.org/10.1029/JZ059i003p00329>
- Hogan, R. J., & Mason, I. B. (2011). Deterministic forecasts of binary events. In *Forecast verification* (pp. 31–59). John Wiley & Sons Ltd. <https://doi.org/10.1002/9781119960003.ch3>
- Hones, E. W., Jr. (1984). Plasma sheet behavior during substorms. In *Magnetic reconnection in space and laboratory plasmas* (pp. 178–184). American Geophysical Union (AGU). <https://doi.org/10.1029/GM030p0178>
- Hones, E. W., Jr., Fritz, T. A., Birn, J., Cooney, J., & Bame, S. J. (1986). Detailed observations of the plasma sheet during a substorm on April 24, 1979. *Journal of Geophysical Research: Space Physics*, 91(A6), 6845–6859. <https://doi.org/10.1029/JA091iA06p06845>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Jaynes, A. N., Baker, D. N., Singer, H. J., Rodriguez, J. V., Loto'aniu, T. M., Ali, A. F., et al. (2015). Source and seed populations for relativistic electrons: Their roles in radiation belt changes. *Journal of Geophysical Research: Space Physics*, 120(9), 7240–7254. <https://doi.org/10.1002/2015JA021234>
- Kamide, Y., & McIlwain, C. E. (1974). The onset time of magnetospheric substorms determined from ground and synchronous satellite records. *Journal of Geophysical Research (1896-1977)*, 79(31), 4787–4790. <https://doi.org/10.1029/JA079i031p04787>
- Kepko, L., McPherron, R. L., Amm, O., Apatenkov, S., Baumjohann, W., Birn, J., et al. (2015). Substorm current wedge revisited. *Space Science Reviews*, 190(1), 1–46. <https://doi.org/10.1007/s11214-014-0124-9>
- Kissinger, J., McPherron, R. L., Hsu, T.-S., & Angelopoulos, V. (2011). Steady magnetospheric convection and stream interfaces: Relationship over a solar cycle. *Journal of Geophysical Research: Space Physics*, 116(A5). <https://doi.org/10.1029/2010JA015763>
- Kullen, A., & Karlsson, T. (2004). On the relation between solar wind, pseudobreakups, and substorms. *Journal of Geophysical Research: Space Physics*, 109(A12). <https://doi.org/10.1029/2004JA010488>
- Lee, D.-Y., Lyons, L. R., & Yumoto, K. (2004). Sawtooth oscillations directly driven by solar wind dynamic pressure enhancements. *Journal of Geophysical Research: Space Physics*, 109(A4). <https://doi.org/10.1029/2003JA010246>
- Lezniak, T. W., Arnoldy, R. L., Parks, G. K., & Winckler, J. R. (1968). Measurement and intensity of energetic electrons at the equator at 6.6 Re. *Radio Science*, 3(7), 710–714. <https://doi.org/10.1002/rds196837710>
- Liou, K., Newell, P. T., Sibeck, D. G., Meng, C.-I., Brittnacher, M., & Parks, G. (2001). Observation of IMF and seasonal effects in the location of auroral substorm onset. *Journal of Geophysical Research: Space Physics*, 106(A4), 5799–5810. <https://doi.org/10.1029/2000JA003001>
- Maimaiti, M., Kunduri, B., Ruohoniemi, J. M., Baker, J. B. H., & House, L. L. (2019). A deep learning-based approach to forecast the onset of magnetic substorms. *Space Weather*, 17(11), 1534–1552. <https://doi.org/10.1029/2019SW002251>
- McPherron, R. L., & Chu, X. (2018). The midlatitude positive Bay index and the statistics of substorm occurrence. *Journal of Geophysical Research: Space Physics*, 123(4), 2831–2850. <https://doi.org/10.1002/2017JA024766>
- McPherron, R. L., Russell, C. T., & Aubry, M. P. (1973). Satellite studies of magnetospheric substorms on August 15, 1968: 9. Phenomenological model for substorms. *Journal of Geophysical Research (1896-1977)*, 78(16), 3131–3149. <https://doi.org/10.1029/JA078i016p03131>
- Mende, S., Heeterds, H., Frey, H., Lampton, M., Geller, S., Habraken, S., et al. (2000). Far ultraviolet imaging from the IMAGE spacecraft. 1. System design. *Space Science Reviews*, 91(1), 243–270. <https://doi.org/10.1023/A:1005271728567>
- Mende, S. B., Harris, S. E., Frey, H. U., Angelopoulos, V., Russell, C. T., Donovan, E., et al. (2008). The THEMIS array of ground-based observatories for the study of auroral substorms. *Space Science Reviews*, 141(1–4), 357–387. <https://doi.org/10.1007/s11214-008-9380-x>
- Mooney, M. K., Marsh, M. S., Forsyth, C., Sharpe, M., Hughes, T., Bingham, S., et al. (2021). Evaluating auroral forecasts against satellite observations. *Space Weather*, 19(8), e2020SW002688. <https://doi.org/10.1029/2020SW002688>
- Morley, S. K., Freeman, M. P., & Tanskanen, E. I. (2007). A comparison of the probability distribution of observed substorm magnitude with that predicted by a minimal substorm model. *Annales Geophysicae*, 25(11), 2427–2437. <https://doi.org/10.5194/angeo-25-2427-2007>
- Murphy, K. R., Bentley, S. N., Miles, D. M., Sandhu, J. K., & Smith, A. W. (2022). Imaging the magnetosphere-ionosphere system with ground-based and in-situ magnetometers. In *Magnetospheric imaging* (pp. 287–340). Elsevier. <https://doi.org/10.1016/B978-0-12-820630-0.00002-7>
- Newell, P. T., & Gjerloev, J. W. (2011). Evaluation of SuperMAG auroral electrojet indices as indicators of substorms and auroral power. *Journal of Geophysical Research: Space Physics*, 116(A12). <https://doi.org/10.1029/2011JA016779>
- Nishimura, Y., Lyons, L., Zou, S., Angelopoulos, V., & Mende, S. (2010). Substorm triggering by new plasma intrusion: THEMIS all-sky imager observations. *Journal of Geophysical Research: Space Physics*, 115(A7). <https://doi.org/10.1029/2009JA015166>
- Ohtani, S., & Gjerloev, J. W. (2020). Is the substorm current wedge an ensemble of wedgelets? Revisit to midlatitude positive bays. *Journal of Geophysical Research: Space Physics*, 125(9), e2020JA027902. <https://doi.org/10.1029/2020JA027902>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pulkkinen, T. I., Baker, D. N., Wiltberger, M., Goodrich, C., Lopez, R. E., & Lyon, J. G. (1998). Pseudobreakup and substorm onset: Observations and MHD simulations compared. *Journal of Geophysical Research: Space Physics*, 103(A7), 14847–14854. <https://doi.org/10.1029/97JA03244>
- Rodger, C. J., Clilverd, M. A., Hendry, A. T., & Forsyth, C. (2022). Examination of radiation belt dynamics during substorm clusters: Magnetic local time variation and intensity of precipitating Fluxes. *Journal of Geophysical Research: Space Physics*, 127(12), e2022JA030750. <https://doi.org/10.1029/2022JA030750>
- Sergeev, V. A., Pellinen, R. J., & Pulkkinen, T. I. (1996). Steady magnetospheric convection: A review of recent results. *Space Science Reviews*, 75(3), 551–604. <https://doi.org/10.1007/BF00833344>
- Shore, R. M., Freeman, M. P., & Gjerloev, J. W. (2018). An empirical orthogonal function reanalysis of the northern polar external and induced magnetic field during solar cycle 23. *Journal of Geophysical Research: Space Physics*, 123(1), 781–795. <https://doi.org/10.1002/2017JA024420>
- Smith, A. W., Forsyth, C., Rae, I. J., Garton, T. M., Bloch, T., Jackman, C. M., & Bakrania, M. (2021). Forecasting the probability of large rates of change of the geomagnetic field in the UK: Timescales, Horizons, and thresholds. *Space Weather*, 19(9), e2021SW002788. <https://doi.org/10.1029/2021SW002788>
- Tanskanen, E., Pulkkinen, T. I., Koskinen, H. E. J., & Slavin, J. A. (2002). Substorm energy budget during low and high solar activity: 1997 and 1999 compared. *Journal of Geophysical Research: Space Physics*, 107(A6), SMP 15-1-SMP 15-11. <https://doi.org/10.1029/2001JA900153>
- Torr, M. R., Torr, D. G., Zucic, M., Johnson, R. B., Ajello, J., Banks, P., et al. (1995). A far ultraviolet imager for the international solar-terrestrial physics mission. *Space Science Reviews*, 71(1), 329–383. <https://doi.org/10.1007/BF00751335>
- Walach, M.-T., & Milan, S. E. (2015). Are steady magnetospheric convection events prolonged substorms? *Journal of Geophysical Research: Space Physics*, 120(3), 1751–1758. <https://doi.org/10.1002/2014JA020631>
- Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. <https://doi.org/10.21105/joss.03021>

- Weimer, D. R. (1994). Substorm time constants. *Journal of Geophysical Research: Space Physics*, 99(A6), 11005–11015. <https://doi.org/10.1029/93JA02721>
- Wes McKinney. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in science conference* (pp. 56–61). <https://doi.org/10.25080/Majora-92bf1922-00a>
- Weygand, J. M., McPherron, R. L., Kauristie, K., Frey, H. U., & Hsu, T. S. (2008). Relation of auroral substorm onset to local AL index and dispersionless particle injections. *Journal of Atmospheric and Solar-Terrestrial Physics*, 70(18), 2336–2345. <https://doi.org/10.1016/j.jastp.2008.09.030>
- Yeoman, T. K., Freeman, M. P., Reeves, G. D., Lester, M., & Orr, D. (1994). A comparison of midlatitude Pi 2 pulsations and geostationary orbit particle injections as substorm indicators. *Journal of Geophysical Research: Space Physics*, 99(A3), 4085–4093. <https://doi.org/10.1029/93JA03233>