Recommendations for Next-Generation Ground Magnetic Perturbation Validation

D. T. Welling [1,2], C. M. Ngwira [3,4], H. Opgenoorth [9], J. D. Haiducek [1], N. P. Savani[4,5], S. K. Morley[6], C. Cid[7], R.S. Weigel[8], H.J. Singer[10], L. Rosenqvist[11], M.W. Liemohn[1]

[1]University of Michigan Department of Climate and Space Sciences and Engineering, Ann Arbor, Michigan, United States
[2]University of Texas at Arlington Department of Physics, Arlington, Texas, United States
[3]The Catholic University of America, Department of Physics, Washington, DC, United States
[4]NASA Goddard Space Flight Center, Space Weather Laboratory, Greenbelt, MD, United States
[5] University of Maryland Baltimore County, Goddard Planetary Heliophysics Institute, Baltimore, MD, USA
[6]Space Science and Applications, Los Alamos National Laboratory, Los Alamos, NM, United States
[7]Space Weather Research Group, Universidad de Alcala, Alcala de Henares, Madrid, Spain
[8]Space Weather Lab at George Mason University, Department of Physics and Astronomy
[9]Whatever Hermann's Affil is.
[10]Space Weather Prediction Center, NOAA, Boulder, CO, United States
[11]Swedish Defence Research Agency, Stockholm, Sweden

Key Points:

- We present a new validation suite for models of ground magnetic perturbations, dB/dt, of interest for geomagnetically induced currents.
- The existing standard remains useful but provides limited information, so an expanded set of metrics is defined here.
- This work is a result of the International Forum for Space Weather Capabilities Assessment and represents a new community consensus.

Corresponding author: Daniel Welling, dwelling@umich.edu

Abstract

= enter abstract here =

1 Introduction

An ongoing challenge of model validation, especially concerning inter-model comparisons and tracking of model progress over time, is creating a validation suite that achieves community-wide acceptance and use. The goal of the International Forum for Space Weather Capabilities Assessment (https://ccmc.gsfc.nasa.gov/assessment/forum-topics.php), organized and led by NASA's Community Coordinated Modeling Center (CCMC), is to overcome this challenge by bringing the community together to achieve consensus on validation techniques. The Forum defined several focused evaluation topics, spanning space weather domains from the sun to the ionosphere. Working teams were then formed to begin work towards defining validation & metric suites that could be leveraged by the entire community. The effort of the Forum continues today to address community validation obstacles.

This work reports on the progress made by the *Ground Magnetic Perturbation* working team, whose goal is to advance validation approaches for predictions of values observed by ground-based magnetometer stations. The value of interest is dB/dt, or the rate of change of the magnetic field as measured on the Earth's surface. This value is especially relevant to geomagnetically induced currents (GIC), which are currents driven through long, ground-based conductors during geomagnetically active periods [*Pirjola*, 2000; *Pulkkinen et al.*, 2017].

Unlike many other space weather subtopics, a contemporary, community-created dB/dt validation suite both exists and continues to be employed. This suite, detailed by *Pulkkinen et al.* [2013], was created with community input via a partnership between CCMC and NOAA's Space Weather Prediction Center (SWPC). The goal of this suite was to help identify an operationally viable predictive model of dB/dt. This study stands as a baseline suite on which to improve upon: while it indeed provides insight into model performance, the information it yields is quite limited. The goal of the Ground Magnetic Perturbation team was to therefore identify the logical next-steps to improve this validation suite without over-complicating its implementation.

This paper presents the recommendations of the team for a next-generation dB/dt validation suite. The contemporary *de facto* standard is first reviewed, with strengths and weaknesses explored. The new approach is then introduced and explained in full. Outstanding issues not yet addressed by the Forum are also discussed. The recommendations are then briefly summarized in the final section.

2 Current Validation Approach

The contemporary *de facto* validation suite in use today is detailed by *Pulkkinen et al.* [2013]. This study evaluated five different models, both numerical and first-principles-based, using six ground-based magnetometers over six real-world events. The selected six events are listed in Table 1 and span very weak to extreme geomagnetic storms. The magnetometer data used began with the perturbation of the background field from a quiet reference, ΔB. For each event, data were collected from six real-world stations, whose positions are shown in Figure 1 as dark black stars. Station names and coordinates are given in Table 2 in*Pulkkinen et al.* [2013]. Geomagnetic dipole coordinates were used: two components are tangent to the surface of the Earth (geomagnetically north-south and east-west), the third is the vertical component. A 60 s sampling frequency was used, yielding a data set that was not overly dense but is unlikely to degrade the data-model comparison significantly [*Pulkkinen et al.*, 2006]. The precise definition of dB/dt used is given by,

$$|dB/dt|_H = \sqrt{(dB_{North}/dt)^2 + (dB_{East}/dt)^2} \tag{1}$$

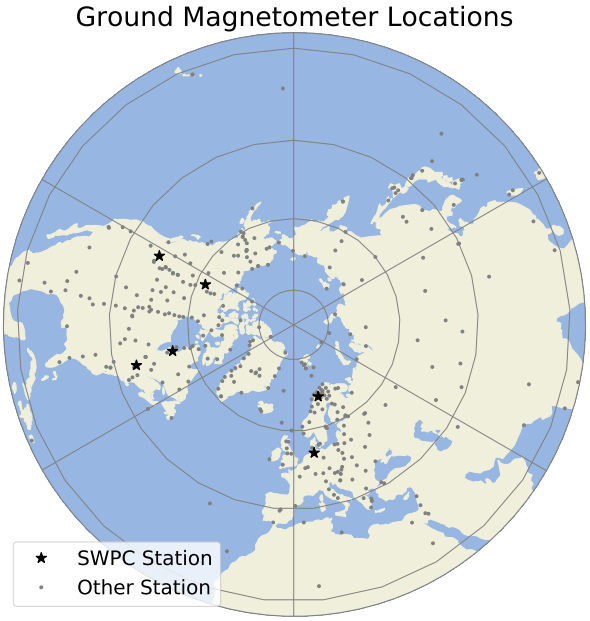


Figure 1. Locations of magnetometer stations used in the original validation suite (black stars), stations with 10 *s* data available (red dots), and other stations (grey dots).

This definition was chosen to investigate the horizontal field fluctuations (i.e., components tangent to the Earth's surface), which are associated with GIC hazards [*Viljanen et al.*, 2001; *Pulkkinen et al.*, 2017]. A simple forward-difference method was used to obtain derivatives; this simple approximation is adequate for the given time resolution [*Tóth et al.*, 2014].

To quantify the data-model comparisons, binary event analysis was employed [*Jolliffe and Stephenson*, 2012]. This approach first divides a time series into non-overlapping time windows; 20 minute windows were used in the existing validation suite. Each window is then categorized based on whether or not the observed and/or modeled dB/dt value crossed a given threshold. A "hit" signifies that both crossed the threshold; a "miss" indicates that the observation crossed but the model did not; a "false positive" occurs when the model predicts a threshold crossing that was not observed, and a "true negative" is when neither observation nor model crosses within the time window. Four thresholds were leveraged: 0.3, 0.7, 1.1 and 1.5 nT/s. Metrics can be constructed from the number of events in each category. Three are used presently: the *probability of detection* (POD) which is the fraction of observed threshold crossings predicted by the model, also called hit rate; *probability of false detection* (POFD) which is the fraction of non-event periods when a crossing was forecast, also called false alarm rate; and finally the *Heidke Skill Score* (HSS).

The probability of detection is defined as

$$\mathrm{POD} = \frac{a}{a+c} \tag{2}$$

where a is the number of hits, b is the number of false positives, c is the number of misses and d is the number of true negatives. POD gives the probability of an event being correctly predicted given that an event occurred. The probability of false detection is defined as

$$\mathrm{POFD} = \frac{b}{b+d} \tag{3}$$

Table 1. List of events in the current dB/dt test suite (1-6), new events recommended for inclusion by the working group (7-8), and other events considered by the working group (9-13). For each, the start time, duration over which data-model comparisons should be made, maximum F10.7 solar flux, Kp, AE, and minimum Sym-H values are shown in each column from left to right, respectively.

#	Event Start	Extent (hours)	F10.7 (sfu)	Kp	AE (nT)	Sym-H (nT)
1	29 Oct 2003 06:00 UT	24	275.4	9^o	4056.0	-391.0
2	14 Dec 2006 12:00 UT	36	90.5	8^+	2284.0	-211.0
3	31 Aug 2001 00:00 UT	24	203.0	4^o	959.0	-46.0
4	31 Aug 2005 10:00 UT	26	86.0	7^o	2063.0	-119.0
5	05 Apr 2010 00:00 UT	24	79.0	8^-	2565.0	-67.0
6	05 Aug 2011 09:00 UT	24	113.0	8^-	2611.0	-126.0
7	17 Mar 2015 02:00 UT	34	116.0	8^-	2298.0	-234.0
8	22 Jul 2004 06:00 UT	162	178.4	9^-	3632.0	-208.0
9	07 Nov 2004 00:00 UT	60	138.1	9^-	3360.0	-394.0
10	30 Mar 2001 12:00 UT	48	257.2	9^-	2407.0	-437.0
11	17 Mar 2013 00:00 UT	48	124.5	7^-	2689.0	-132.0
12	06 Apr 2000 12:00 UT	48	178.1	9^-	2481.0	-320.0
13	15 May 2005 00:00 UT	24	105.2	8^+	2051.0	-305.0

and considers the number of intervals in which a threshold crossing was predicted but did not occur. POFD gives the probability of an event being incorrectly predicted given that an event did not occur. Smaller values of POFD indicate a better model performance.

Skill scores are measures of accuracy relative to a reference model [*Wilks*, 2011]. The Heidke Skill Score (HSS) uses the proportion correct (PC) as the accuracy measure, which is defined as

$$\mathrm{PC} = \frac{a+d}{a+b+c+d} \tag{4}$$

and measures the fraction of predictions that obtained the correct result. The reference model used in calculating the HSS is the PC that would be obtained for random predictions that are statistically independent of the observations [*Wilks*, 2011]. The Heidke Skill Score is then defined as

$$\mathrm{HSS} = \frac{\mathrm{PC} - \mathrm{PC}_{\mathrm{ref}}}{1 - \mathrm{PC}_{\mathrm{ref}}} = \frac{2(ad - bc)}{(a+c)(c+d) + (a+b)(b+d)} \tag{5}$$

For random predictions and constant predictions HSS is zero indicating that the prediction is unskilled. Predictions that outperform random chance have a positive HSS, while a perfect prediction has an HSS of 1. These metrics are frequently employed in space weather applications [e.g., *Lopez et al.*, 2007; *Yu and Ridley*, 2008; *Welling and Ridley*, 2010; *Pulkkinen et al.*, 2013; *Ganushkina et al.*, 2015; *Austin and Savani*, 2018].

Although relatively simple, the SWPC-CCMC test suite is both important and useful today. Because of the community involvement in defining the suite, it stands as an agreed-upon approach for inter-model comparison for ground magnetic perturbations. By focusing on dB/dt, the suite is highly relevant to operations. Though limited in number, the metrics yield a good description of overall performance by showing the user the balance between hits, false positives, and overall skill. The use of binary event analysis with 20 minute windows provides a built-in way to account for slight discrepancies in timing between the models and data. More broadly, the validation suite was a critical step in selecting a model to transition to operations at NOAA SWPC. The suite continues to be used today to track the progress of the operational model as it is further developed.

3 Recommendations for Improvement

Despite the strengths of the SWPC-CCMC suite, it remains limited in the amount of information that it provides to the user. Only a handful of events are tested with a limited number of stations. This limits the statistical power of the study. Values are combined to give metrics that very broadly describe performance across a variety of locations and types of activity. Large spatial gaps exist between the six stations, meaning much dB/dt activity can be missed. Results from the validation suite are used to tell a developer if a model is deficient, but where and how it is deficient remain unanswered.

There are many possible ways to improve the original validation suite to increase its utility. Rather than seek complicated and labor intensive solutions, the Ground Magnetic Perturbation team sought improvements that are powerful, relatively simple to employ, and widely agreed upon by team members. Another aspect of model performance that can be captured with a trivial expansion of the metrics suite is the tendency of the model to either over- or under-predict. This is captured by the frequency bias, which is calculated as

$$\text{Bias} = \frac{a+b}{a+c} \tag{6}$$

and gives the ratio of event forecasts to event observations. A bias of 1 indicates that the same number of events were forecast as were observed, If the model predicts too many events then the bias will be greater than one.

Four additional areas of focus were selected by the working team: increasing the number of validation events; increasing the number and fidelity of observations; implementing a regional analysis scheme; and segregating results by type of activity. Each of these are described briefly below.

3.1 New Validation Events

An immediate concern of the Ground Magnetic Perturbation Working Team was to expand the number of events included in the validation suite. While the currently included events (Table 1, events 1-6) all occur during periods of high K_P index, four of the six events have middling SYM-H signatures that are less than 150 nT in magnitude (Table 1, rightmost column). The only true "super storm" is Event 1, which is the well-known Halloween Storm of 2003. Expanding the event list will also help improve the number of threshold crossings, improving the statistical significance of overall test. It is clear that one of the easiest ways to improve this validation suite would be to expand the event list and, therefore, the amount of time over which the models were tested.

Many events were suggested, and a short list of seven potential new events was constructed. The short list is shown in Table 1 as items 7-13. For comparison to the existing events, peak F10.7 radio flux, K_P index, and Auroral Electrojet index (AE) are shown as is minimum SYM-H (fourth through seventh columns, respectively). A preference was given to strong and extreme storms; contemporary storms were also sought to yield events with excellent coverage from modern missions and data campaigns. Members of the working group voted and narrowed the list to two new events.

The first event that should be added to the validation suite is summarized in Figure 2. This is the well-known St. Patrick's Day storm of March, 2015. The top three frames of Figure 2 show the solar drivers in terms of GSM Y and Z components of the interplanetary magnetic field (IMF), solar wind density, and Earthward velocity. The bottom two frames summarize the magnetospheric response via the SYM-H and Auroral Electrojet (AE) geomagnetic indexes. As this storm is widely studied [e.g., *Carter et al.*, 2016; *Lotz et al.*, 2017; *Ngwira et al.*, 2018; *Divett et al.*, 2018], it provides ample opportunity for further validation outside of ground magnetic perturbations. THIS STORM HAS BEEN STUDIED A TON. WE NEED SOME REFERENCES HERE FOR THIS STORM. With a SYM-H minimum at -234 nT, it would become the second strongest storm in the validation suite.

The second storm selected is actually a triple-CME event occurring in late July, 2004. The solar wind conditions for this event and the corresponding geomagnetic indexes are shown in Figure 3. Each of the sub-events drives a stronger response from the magnetosphere, both in terms of SYM-H and AE. The final sub-event drives the third strongest SYM-H and second strongest AE signature amongst all events in the validation suite. Inclusion of this event will test models in very unique ways. Because there are three distinct storm intensifications and recoveries, the ability of the models to properly capture the hysteresis of the system will be tested. At 162 hours (6 days), it is four times longer than any other event. Models will need to robustly simulate this extended period in order to obtain positive skill scores. These challenges increase the operational relevance of the validation suite overall.

3.2 Increased Coverage and Resolution in Observations

The original validation suite compared model results against only six magnetometers, each reporting ΔB with a 60s sampling rate. This made the initial study straightforward to perform because only a small number of stations were included and because most magnetometer stations release 1-minute data. These choices are limitations of the study. The spatial coverage is poor, leaving large gaps uncovered (e.g., Figure 1). The data-model statistics are thin; a problem that intensifies as comparisons are segregated by latitude. While a 60 s sampling rate captures most GIC-pertinent fluctuations, a 1 s resolution is optimal [*Pulkkinen et al.*, 2006]. The lower time resolution observations also limit the quality of the numerical derivative of ΔB [e.g., *Tóth et al.*, 2014] More stations and with higher sampling rates are simple ways to improve the fidelity of the validation suite.

For the improved validation suite, the observational comparison set will be expanded both by the number of stations and in terms of the sampling frequency. A 10 s frequency will be adopted for both observations and model output. While 1 s is desirable, 10 s output will improve the comparisons without reducing the available real world observatories or greatly slowing model execution. Rather than just six stations, all magnetometer observatories that report 10 s ΔB data will be included. WE NEED TO KNOW HOW MANY 10s TIME RESOLUTION MAGS ARE AVAILABLE. DTW WILL ADD THIS TO TABLE 1 AND FIGURE 1. Table 1 reports on the number of magnetometers available for each event given this criterion. Stations with 10 s data available are indicated on Figure 1, illustrating the expanded spatial coverage. Expanding the suite in this way will both improve the quality of the dB/dt comparisons and while growing the statistical strength of the reported metrics.

An example of 60s vs. 10s ΔB and the approximate derivative would be powerful to demonstrate the need here.

3.3 Regional Analysis

Another limitation of the current validation approach is one of location and proximity. The results provided by the *Pulkkinen et al.* [2013] study segregated results into two latitude groups, but did not provide information about model performance as a function of magnetic local time (MLT). Further, if a dB/dt peak is predicted correctly temporally but at the wrong location, the model will be penalized. Temporal near-misses are already accounted for via the 20-minute windows employed by the binary event analysis. We could use some references discussing how spatial scales for dB/dt can be small, demonstrating the need to compensate for spatial near-misses. To improve the validation suite without over-complicating its implementation, a simple MLT binning method is recommended. First, a set of virtual magnetometers is included as part of the model results that do not correspond to real world observatories. Rather, these are regularly spaced at 5° latitude and longitude intervals across the entire globe. Such output is currently produced by the operational SWPC Geospace model at present. An alternate version of the binary event study will then be used. For each MLT quadrant, the question will be asked, “do *any* real observatories or *any* virtual magnetome-

ters report a dB/dt threshold crossing?" This will create contingency tables and metrics as a function of MLT quadrants instead of on a per-station basis.

The results of this additional metric calculation will be used to provide more information than the per-station metrics alone. Regional analysis will help modelers understand where their codes perform the best and where they perform the worst (e.g., day side vs. night side). Further, discrepancies between the per-station and regional analysis will help inform users of spatial near-misses. For example, if the regional analysis' Heidke Skill Score is considerably higher than the traditional per-station results, it is likely that the model is frequently predicting threshold crossings that correspond to real crossings but at the wrong location. Adding this portion to the validation suite grows its utility.

3.4 Segregation by Activity Type

The SWPC-CCMC validation suite is activity agnostic, meaning that skill scores are calculated across all time periods. Geomagnetic storms are the net effect of many sub-events, including substorms, sudden commencements, and many other categories of processes. The question naturally arises, "under what types of activity does a certain model do best or worst?" The current validation suite is incapable of answering such inquiries.

To address this, the recommendation of the Working Team is to calculate additional values corresponding to periods of certain types of activity. To make this immediately feasible, three activity types are recommended: storm sudden commencements, substorm expansions, and ring current intensifications. There are many more types of activity, and becoming more granular in definitions may be beneficial for future work. These initial three classifications are enough to expand the informative power of the validation suite without making implementation exceedingly difficult to accomplish.

Defining sub-event time windows is challenging, as there many ways to define classes of activity based on different observations and different criteria. The goal for this revised validation suite is to use definitions that are easy to implement, have a reasonable level of community agreement, and are likely to create a meaningful signal in the chosen metrics. For the three activity classes selected, the following criteria are used:

- Storm Sudden Commencements (SSCs) are well defined in literature and easily identified via a sharp increase in the SYM-H index corresponding to the arrival of a solar wind dynamic pressure pulse. The epoch of the event is defined as the start of the Sym-H rise. For each SSI, a broad time window is defined starting ten minutes before the event epoch and lasting twenty minutes after. The time window range allows the metrics to capture SSI-driven activity while compensating for small timing discrepancies between the model and real system.
- Ring current intensifications can be identified as periods of decreasing SYM-H index. For the revised validation suite, all times where both SYM-H and the time derivative of SYM-H are less than zero. To remove small time scale features and deviations not likely related to the ring current, a median filter is applied to SYM-H and only windows of at least an hour in length are considered.
- Auroral substorm expansions are a critical source of dB/dt but also the most challenging to quickly identify in a reliable manner. Use of auroral electrojet indexes, specifically, AL, are a popular, simple, but imperfect way to identify substorms. Several automated methods exist. For this study, the methodology of *Borovsky and Yakymenko* [2017] is employed. This is chosen over the more well established Supermag AL index algorithm [*Newell and Liou*, 2011] because it is far less sensitive to weaker auroral activity. The focus is therefore on moderate to strong substorms that are more relevant to GIC applications.

Figure 4 illustrates the above criteria as applied to Validation Event 7 (row 7 in Table 1). The top frame shows AU and AL indices for the entire event; the bottom frame shows SYM-H. Yellow, red, and blue windows show the SSC, ring current intensification, and substorm validation windows. Binary-event based metrics would be made using each color region separately in order to best characterize model performance as a function of the type of activity. With the expanded observational set and new events added to the validation suite, there will be enough data-model comparisons to produce meaningful activity-dependent metrics.

4 Future Considerations

Consistent with the approach taken in *Pulkkinen et al.* [2013], the current recommendation defines time intervals on the order of days during which significant geomagnetic events occur and to test model performance during these time intervals.

This approach has the advantage of limiting the amount of model run time and the amount of data that needs to be processed. In addition, the performance results apply only to active periods, which are of most interest to the end-user.

The ultimate objective of forecast model development is to have predictions available in real-time or near-real-time and to have the models run continuously. Therefore, future time intervals will include a long and continuous time interval (on the order of a year). In addition to allowing the estimation of prediction performance under realistic use conditions, such a long interval will allow additional features of model performance to be considered, including magnetic local time and day-of-year.

A second consideration is the scaling of the number of events to allow error bars to be generated for the model performance metrics. With 13 events, we will have the ability to calculate meaningful error bars on the aggregate model performance; additional events will allow a better characterization of the error and will allow the end-user to determine if the reliability of the model performance is sufficient to allow decisions to be made based on a forecast (*Thomson* [2000];*Weigel et al.* [2006]).

5 Summary and Conclusions

Acknowledgments

F10.7 data was obtained from the LASP Interactive Solar Irradiance Data Center (http://lasp.colorado.edu/lisird). Geomagnetic index data was obtained from the World Data Center for Geomagnetism, Kyoto (http://wdc.kugi.kyoto-u.ac.jp). The authors thank the WDC and their many data providers (http://wdc.kugi.kyoto-u.ac.jp/wdc/obslink.html) who make this data publicly available.

References

Austin, H. J., and N. P. Savani (2018), Skills for forecasting space weather, *Weather*, *0*(0), doi:10.1002/wea.3076.

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

Carter, J. A., S. E. Milan, J. C. Coxon, M.-T. Walach, and B. J. Anderson (2016), Average field-aligned current configuration parameterized by solar wind conditions, *J. Geophys. Res. Sp. Phys.*, *121*(2), 1294–1307, doi:10.1002/2015JA021567.

Divett, T., G. S. Richardson, C. D. Beggan, C. J. Rodger, D. H. Boteler, M. Ingham, D. H. Mac Manus, A. P. Thomson, and M. Dalzell (2018), Transformer-Level Modeling of Geomagnetically Induced Currents in New ZealandâĂŹs South Island, *Space Weather*, *16*,

718–735, doi:10.1029/2018SW001814.

Ganushkina, N. Y., O. A. Amariutei, D. Welling, and D. Heynderickx (2015), Nowcast model for low-energy electrons in the inner magnetosphere, *Sp. Weather*, *13*(1), 16–34, doi:10.1002/2014SW001098.

Jolliffe, I. T., and D. B. Stephenson (2012), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 288 pp., John Wiley & Sons.

Lopez, R. E., S. Hernandez, M. Wiltberger, C. L. Huang, E. L. Kepko, H. Spence, C. C. Goodrich, and J. G. Lyon (2007), Predicting magnetopause crossings at geosynchronous orbit during the Halloween storms, *Sp. Weather*, *5*(1), n/a–n/a, doi: 10.1029/2006SW000222.

Lotz, S. I., M. J. Heyns, and P. J. Cilliers (2017), Regression-based forecast model of induced geo-electric field, *Space Weather*, *15*, 180–191, doi:10.1002/2016SW001518.

Newell, P. T., and K. Liou (2011), Solar wind driving and substorm triggering, *J. Geophys. Res.*, *116*(A3), A03,229, doi:10.1029/2010JA016139.

Ngwira, C. M., D. Sibeck, M. V. D. Silveria, M. Georgiou, J. M. Weygand, Y. Nishimura, and D. Hampton (2018), A study of intense local dB/dt variations during two geomagnetic storms, *Space Weather*, *16*, doi:10.1029/2018SW001911.

Pirjola, R. (2000), Geomagnetically induced currents during magnetic storms, *IEEE Trans. Plasma Sci.*, *28*(6), 1867–1873.

Pulkkinen, A., A. Viljanen, and R. Pirjola (2006), Estimation of geomagnetically induced current levels from different input data, *Sp. Weather*, *4*(8), n/a–n/a, doi: 10.1029/2006SW000229.

Pulkkinen, A., L. Rastätter, M. Kuznetsova, H. Singer, C. Balch, D. Weimer, G. Toth, A. Ridley, T. Gombosi, M. Wiltberger, J. Raeder, and R. Weigel (2013), Community-wide validation of geospace model ground magnetic field perturbation predictions to support model transition to operations, *Sp. Weather*, *11*(6), 369–385, doi:10.1002/swe.20056.

Pulkkinen, A., E. Bernabeu, A. Thomson, A. Viljanen, R. Pirjola, D. Boteler, J. Eichner, P. J. Cilliers, D. T. Welling, N. P. Savani, R. S. Weigel, J. J. Love, C. Balch, C. M. Ngwira, G. Crowley, A. Schultz, R. Kataoka, B. Anderson, D. Fugate, J. J. Simpson, and M. MacAlester (2017), Geomagnetically induced currents: Science, engineering, and applications readiness, doi:10.1002/2016SW001501.

Thomson, A. W. P. (2000), Evaluating space weather forecasts of geomagnetic activity from a user perspective, *Geophys. Res. Lett.*, *27*, 4049–4052, doi:10.1029/2000GL011908.

Tóth, G., X. Meng, T. I. Gombosi, and L. Rastätter (2014), Predicting the time derivative of local magnetic perturbations, *J. Geophys. Res. Sp. Phys.*, *119*(1), 310–321, doi: 10.1002/2013JA019456.

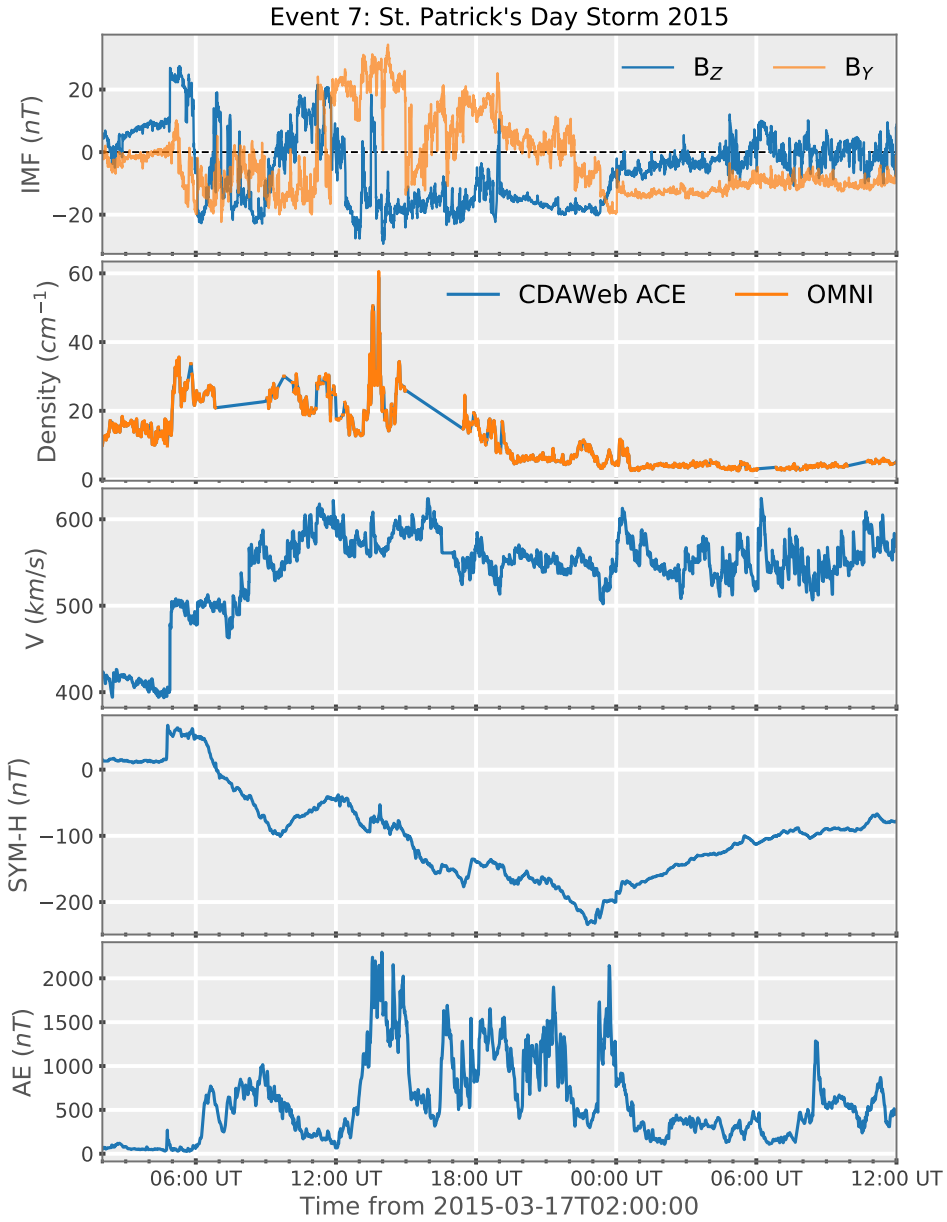
Viljanen, A., H. Nevanlinna, K. Pajunpää, and A. Pulkkinen (2001), Time derivative of the horizontal geomagnetic field as an activity indicator, *Ann. Geophys.*, *19*(9), 1107–1118.

Weigel, R. S., T. Detman, E. J. Rigler, and D. N. Baker (2006), Decision theory and the analysis of rare event space weather forecasts, *Space Weather*, *4*(5), n/a–n/a, doi: 10.1029/2005sw000157.

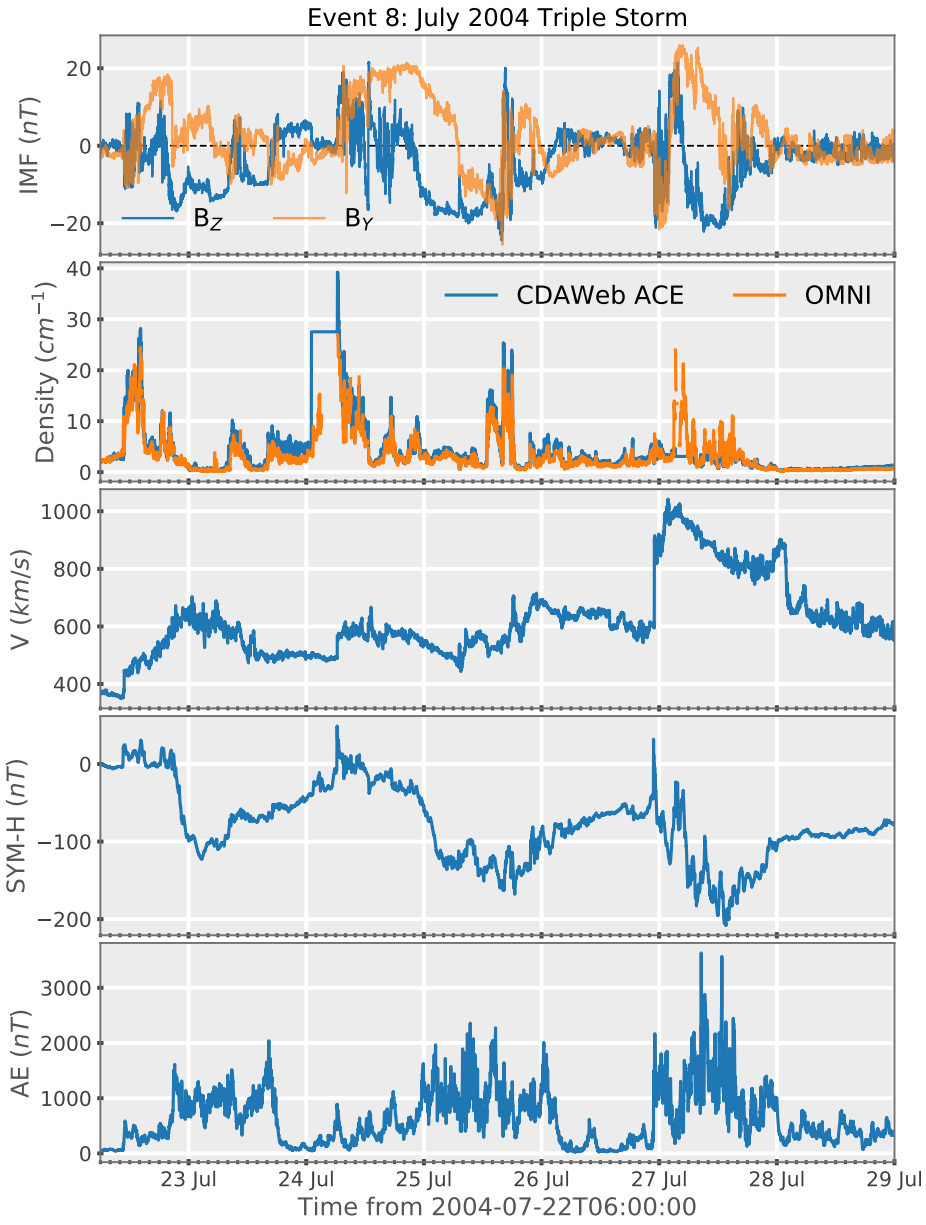
Welling, D. T., and A. J. Ridley (2010), Validation of SWMF magnetic field and plasma, *Sp. Weather*, *8*(3), n/a–n/a, doi:10.1029/2009SW000494.

Wilks, D. S. (2011), *Statistical methods in the atmospheric sciences*, 3rd ed., 676 pp., Academic Press.

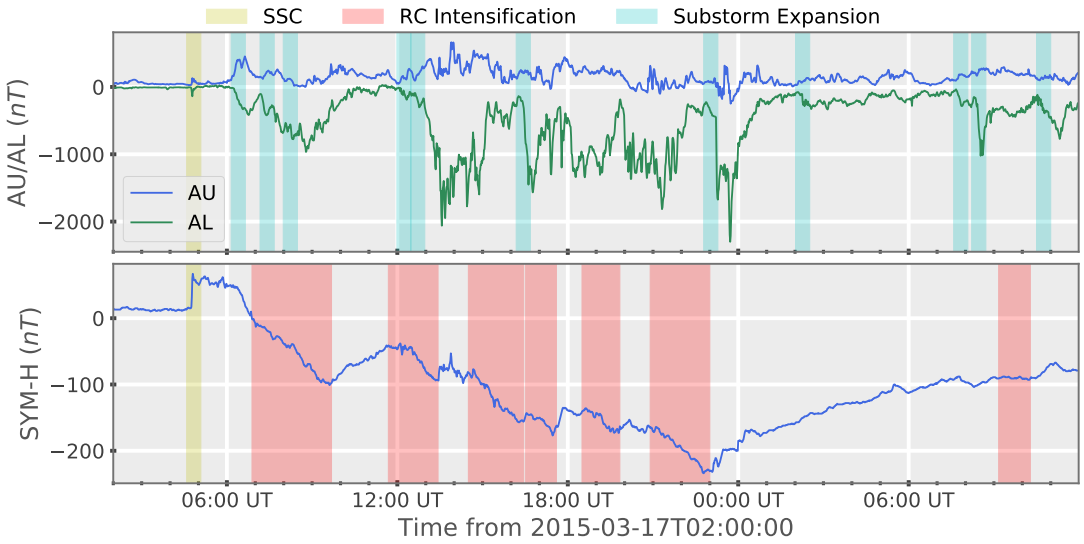
Yu, Y., and A. J. Ridley (2008), Validation of the space weather modeling framework using ground-based magnetometers, *Sp. Weather*, *6*(5), S05,002, doi:10.1029/2007SW000345.



175 **Figure 2.** Summary of Event 7 in terms of IMF (top frame), solar wind density and Earthward velocity
176 (2nd and 3rd frames from the top, and the geomagnetic response in terms of Sym-H and AE indexes (bottom
177 two frames).



188 **Figure 3.** Summary of Event 8; same format as Figure 2



284 **Figure 4.** AU/AL (top frame) and SYM-H (bottom frame) indexes for validation event 7. Storm sudden
285 commencements, ring current intensifications, and substorm periods are marked by yellow, red, and cyan
286 boxes, respectively.