AGU100 ADVANCING EARTH AND SPACE SCIENCE

Space Weather

RESEARCH ARTICLE
10.1029/2018SW002064

Special Section:
Space Weather Capabilities Assessment

Key Points:
- We present a new validation suite for models of ground magnetic perturbations, d*B*/d*t*, 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

Correspondence to:
D. T. Welling,
daniel.welling@uta.edu

Citation:
Welling, D. T., Ngwira, C. M., Opgenoorth, H., Haiducek, J. D., Savani, N. P., Morley, S. K., et al. (2018). Recommendations for next-generation ground magnetic perturbation validation. *Space Weather*, *16*, 1912–1920. https://doi.org/10.1029/2018SW002064

Received 16 AUG 2018
Accepted 9 OCT 2018
Accepted article online 12 OCT 2018
Published online 4 DEC 2018
Corrected 12 DEC 2018

This article was corrected on 12 DEC 2018. See the end of the full text for details.

Recommendations for Next-Generation Ground Magnetic Perturbation Validation

D. T. Welling[1,2], C. M. Ngwira[3,4], H. Opgenoorth[5,6], J. D. Haiducek[1], N. P. Savani[4,7], S. K. Morley[8], C. Cid[9], R.S. Weigel[10], J. M. Weygand[11], J. R. Woodroffe[8], H.J. Singer[12], L. Rosenqvist[13], and M.W. Liemohn[1]

[1]Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI, USA, [2]Department of Physics, University of Texas at Arlington, Arlington, TX, USA, [3]Department of Physics, The Catholic University of America, Washington, DC, USA, [4]Space Weather Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, USA, [5]Swedish Institute of Space Physics Uppsala Division, Uppsala, Sweden, [6]Department of Physics and Astronomy, University of Leicester, Leicester, UK, [7]Goddard Planetary Heliophysics Institute, University of Maryland, Baltimore County, Baltimore, MD, USA, [8]Space Science and Applications, Los Alamos National Laboratory, Los Alamos, NM, USA, [9]Space Weather Research Group, Universidad de Alcala, Alcala de Henares, Spain, [10]Department of Physics and Astronomy, Space Weather Lab at George Mason University, Fairfax, VA, USA, [11]Department of Earth, Planetary, and Space Sciences, University of California, Los Angeles, CA, USA, [12]Space Weather Prediction Center, NOAA, Boulder, CO, USA, [13]Swedish Defence Research Agency, Stockholm, Sweden

Abstract Data-model validation of ground magnetic perturbation forecasts, specifically of the time rate of change of surface magnetic field, d*B*/d*t*, is a critical task for model development and for mitigation of geomagnetically induced current effects. While a current, community-accepted standard for d*B*/d*t* validation exists (Pulkkinen et al., 2013), it has several limitations that prevent more complete understanding of model capability. This work presents recommendations from the International Forum for Space Weather Capabilities Assessment Ground Magnetic Perturbation Working Team for creating a next-generation validation suite. Four recommendations are made to address the existing suite: greatly expand the number of ground observatories used, expand the number of events included in the suite from six to eight, generate metrics as a function of magnetic local time, and generate metrics as a function of activity type. For each of these, implementation details are explored. Limitations and future considerations are also discussed.

Plain Language Summary Space weather forecast models of magnetic field perturbations are important for protecting the power grid and other vulnerable infrastructure. These models must be validated by comparing their predictions to observations. This paper makes recommendations for how future models should be validated in order to best test their capabilities.

1. Introduction

An ongoing challenge of model validation, especially concerning intermodel 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 toward defining validation and 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 d*B*/d*t*, 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 (GICs), which are currents driven through long, ground-based conductors during geomagnetically active periods (Pirjola, 2000; Pulkkinen et al., 2017).

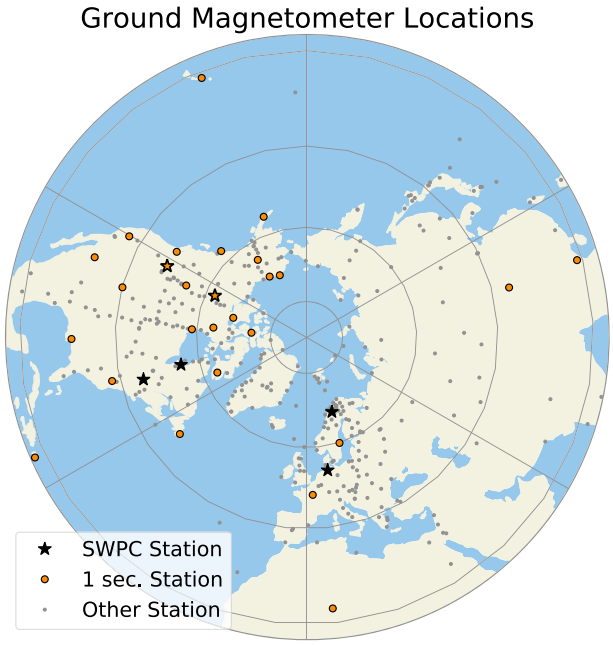

Figure 1. Locations of magnetometer stations used in the original validation suite (black-bordered stars), stations with 1-s data available (orange dots and orange stars) and other stations (gray dots). SWPC = Space Weather Prediction Center.

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 National Oceanic and Atmospheric Administration'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 overcomplicating 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 in three latitudinal chains over six real-world events. The selected six events are listed in Table 2 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 the six real-world stations, whose positions are shown in Figure 1 as black-bordered 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 radial 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

$$|\mathrm{d}B/\mathrm{d}t|_H = \sqrt{(\mathrm{d}B_{\mathrm{North}}/\mathrm{d}t)^2 + (\mathrm{d}B_{\mathrm{East}}/\mathrm{d}t)^2}. \quad (1)$$

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 (Pulkkinen et al., 2017; Viljanen et al., 2001). A simple

Table 1
An Illustrative Contingency Table Used in Binary Event Analysis

	Observed?	
Forecast?	Yes	No
Yes	**a**	**b**
	Hits	False alarms
No	**c**	**d**
	Misses	True negatives

Note. Bold letters indicate labels used in equations (2)–(5).

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 & Stephenson, 2012). This approach first divides a time series into nonoverlapping time windows; 20-min windows were used in the existing validation suite. Each window is then categorized based on if 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 alarm" 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. Table 1 shows an example contingency table that is formed by tallying up the categories for each window. Four thresholds were chosen, 0.3, 0.7, 1.1, and 1.5 nT/s, to provide a range of activity and yield a meaningful number of events to study. 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 nonevent periods when a crossing was forecast, also called false alarm rate; and finally, the *Heidke Skill Score* (HSS).

The POD 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, as illustrated in Table 1. POD gives the probability of an event being correctly predicted, given that an event occurred. The POFD is defined as

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

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 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 HSS 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 0, 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., Austin & Savani, 2018; Ganushkina et al., 2015; Lopez et al., 2007; Pulkkinen et al., 2013; Welling & Ridley, 2010; Yu & Ridley, 2008).

The metrics were calculated for three subsets: one that combined all stations and all events, one for high-latitude (>60°) stations only, and one for midlatitude (<60°) stations only. Note that low-latitude regions were not considered; the lowest latitude station currently included is 54.1° geomagnetic latitude. The end result is a handful of numbers that were used to rank the evaluated models.

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 intermodel comparison for

Table 2
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)

#	Event start	Extent (hr)	F10.7 (sfu)	Kp	AE (nT)	SYM-H (nT)
1	29 Oct 2003 06:00 UT	24	275.4	9°	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°	959.0	−46.0
4	31 Aug 2005 10:00 UT	26	86.0	7°	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

Note. AE = auroral electrojet. 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.

ground magnetic perturbations. By focusing on d*B*/d*t*, 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-min windows provides a built-in margin 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 National Oceanic and Atmospheric Administration 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 d*B*/d*t* 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. Four areas of focus were selected: 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. We also recommend an additional minor change: expanding the number of metrics by one. Each of these is 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 2, events 1–6) all occur during periods of high Kp index, four of the six events have middling SYM-H signatures that are less than 150 nT in magnitude (Table 2, 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.

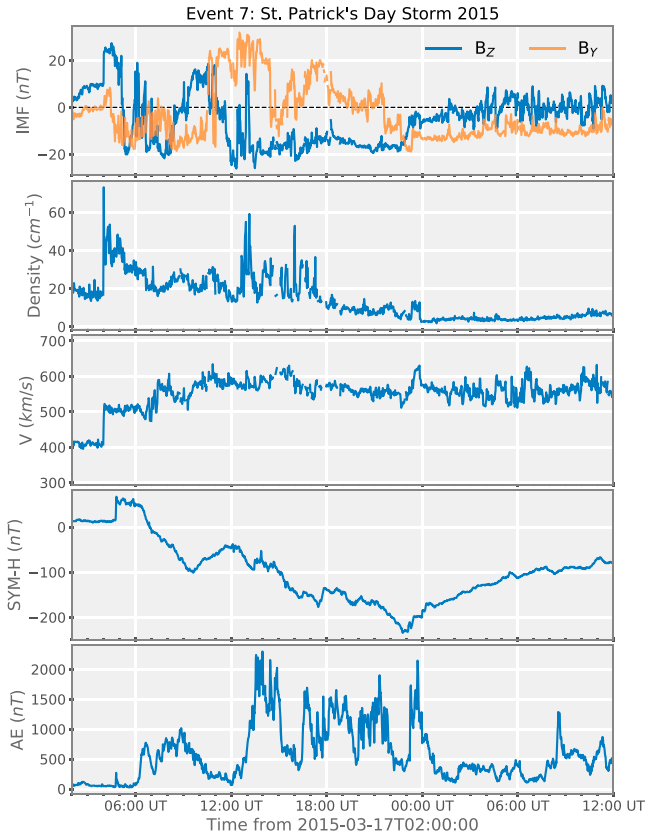

Figure 2. Summary of event 7 in terms of interplanetary magnetic field (IMF; top frame), solar wind density and earthward velocity (second and third frames from the top), and the geomagnetic response in terms of SYM-H and auroral electrojet (AE) indexes (bottom two frames).

Many events were suggested, and a short list of seven potential new events was constructed. The short list is shown in Table 2 as items 7–13. For comparison to the existing events, peak F10.7 radio flux, Kp index, and auroral electrojet (AE) index 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 geocentric solar magnetospheric *Y* and *Z* components of the interplanetary magnetic field, solar wind density, and earthward velocity. Values are taken from the Wind spacecraft via NASA's CDAWeb and the OMNI database. Observations from the ACE spacecraft are also available for this time period (not shown), but these contain several coverage gaps in the density and velocity values. The bottom two frames summarize the magnetospheric response via the SYM-H and AE geomagnetic indexes. As this storm is widely studied (e.g., Carter et al., 2016; Divett et al., 2018; Guerrero et al., 2017; Lotz et al., 2017; Ngwira et al., 2018), it provides ample opportunity for further validation outside of ground magnetic perturbations. With an 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 coronal mass ejection (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 subevents drives a stronger response from the magnetosphere, both in terms of SYM-H and AE. The final subevent drives the third strongest SYM-H and second strongest AE signature among 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 hr (6 days), it is 4 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 60-s 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-min data. These choices are limitations of the study. The spatial coverage is poor, leaving large gaps uncovered (e.g., considering only the stations marked with stars in 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 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 to as many stations as available in the northern hemisphere. Ten-second frequency will be adopted for both observations and model output for all real-world stations that have this sampling rate available. While 1 s is desirable, 10-s output will improve the comparisons without greatly taxing forecast models. Rather than just six stations, all magnetometer observatories that report 60 s ΔB data will be included. Data reporting at 10 s or lower will be downsampled to 10 s. Stations available via the SuperMAG database (Gjerloev, 2012) are shown in Figure 1 as gray dots; stations available from the INTERMAGNET website with ≤10-s data available are indicated with orange dots or orange stars (the latter being part of the original validation suite). At current, there are just over

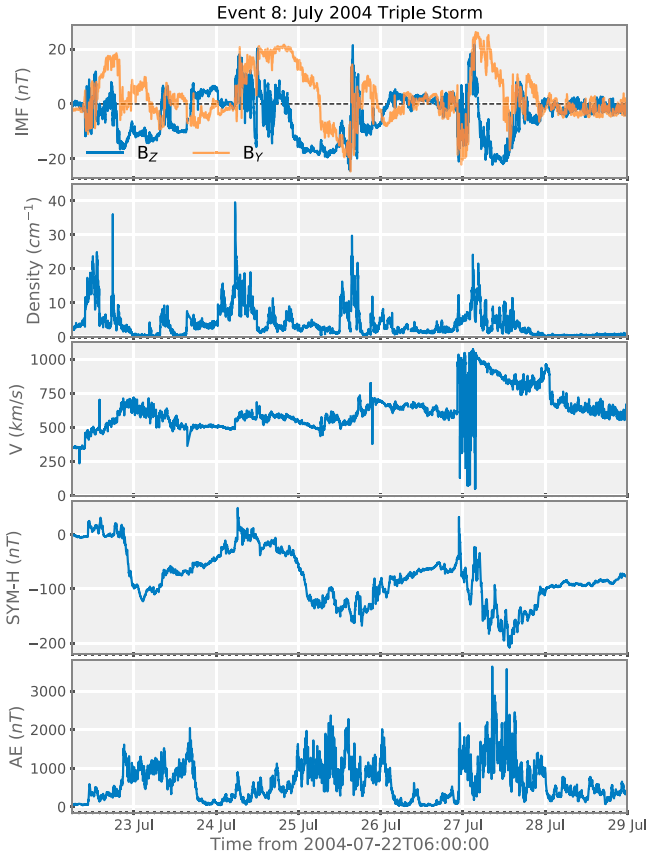

Figure 3. Summary of event 8; same format as Figure 2. AE = auroral electrojet; IMF = interplanetary magnetic field.

400 stations available in the northern hemisphere; 31 of which report data at a 1-s frequency. Expanding the suite in this way will both improve the quality of the d*B*/d*t* comparisons while growing the statistical strength of the reported metrics.

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 d*B*/d*t* peak is predicted correctly temporally but at the wrong location, the model will be penalized. Localized surface disturbance peaks are not unexpected (Pulkkinen et al., 2015). Temporal near misses are already accounted for via the 20-min windows employed by the binary event analysis. To improve the validation suite without overcomplicating 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 magnetometers report a d*B*/d*t* threshold crossing?" This will create contingency tables and metrics as a function of MLT quadrants instead of on a per-station basis. The expanded MLT binning should be implemented alongside the existing latitudinal segregation currently used in the Pulkkinen et al. (2013) study.

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., dayside vs. nightside). Further, discrepancies between the per-station and regional analysis will help inform users of spatial near misses. For example, if the regional analysis' HSS 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 subevents, 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 (SSCs), 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 subevent time windows is challenging, as there are 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:

- 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 SSC, a broad time window is defined starting 10 min before the event epoch and lasting

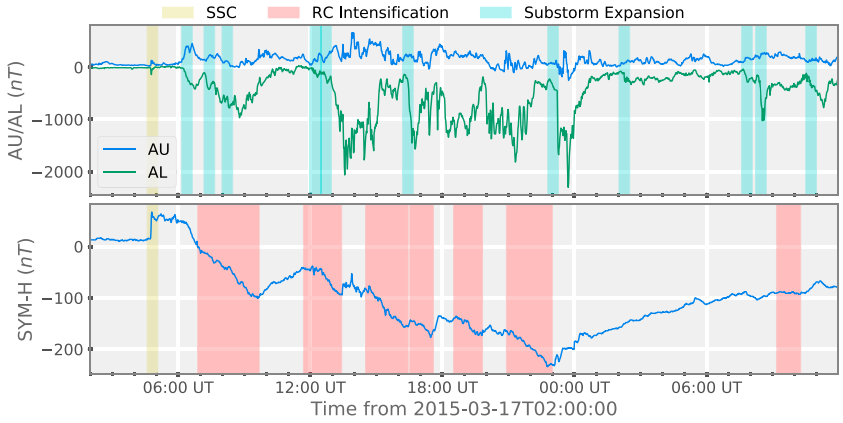

Figure 4. AU/AL (top frame) and SYM-H (bottom frame) indexes for validation event 7. Storm sudden commencements, ring current intensifications, and substorm periods are marked by yellow, red, and cyan boxes, respectively. RC = ring current; SSC = storm sudden commencements.

20 min after. The time window range allows the metrics to capture SSC-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, this criteria will be narrowed to periods where both SYM-H and the time derivative of SYM-H are less than zero. To remove small timescale 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 AE indexes, specifically AL, is 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 & 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 2). 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, respectively. 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.

3.5. New Metric: Frequency Bias

Overall, the number of metrics employed by the suite is descriptive but minimal. This is somewhat by design, as users can quickly compare and assess model performance. We recommend expanding the metrics by one to include the frequency bias (FB), defined as the ratio of threshold crossing forecasts (including false positives) to the observed crossing forecasts:

$$\mathrm{FB} = \frac{a+b}{a+c}. \tag{6}$$

This metric reports the bias of the model in terms of the frequency of predicting threshold crossings. A value greater than 1 means that the model predicts crossings far more frequently than the real-world rate; a value less than 1 means that it is predicting at a frequency less than the real-world rate. This minor change improves the completeness of the metric suite.

4. Future Considerations

The recommendations here represent immediate next steps for improving the existing validation suite. With any such effort, there will always be shortcomings, both obvious and hidden, that must be addressed. Several important shortcomings identified by the working team should be addressed in the next iterations of the validation suite.

While we recommend expanding the number of events, expanding further is still necessary. Consistent with the approach taken in Pulkkinen et al. (2013), the current recommendation defines time interval on the order of days during which a significant geomagnetic event occurred 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 needed 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 should include a long and continuous time interval (on the order of months to 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 MLT and season. A second consideration is the scaling of the number of events to allow error bars to be generated for the model performance metrics. With eight 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).

As GIC forecasting evolves, further thought will be required concerning the value being compared to observations. At current, dB/dt, as defined in equation (1), is used because it is directly relevant to GICs. The magnitude of the horizontal components of dB/dt is used because it offers a simple, single-value time series over which to test. The long-term goal for GIC forecasting is geoelectric field, not merely dB/dt. Obtaining geoelectric field will require the separate components of dB/dt, requiring validation of each individually. Further, magnetotelluric models that can calculate the geoelectric field (e.g., Kelbert et al., 2017) require the geomagnetic field, ΔB, not its time derivative. The metrics and validation approach may need to be revisited and reexamined to best serve these changing needs.

5. Summary

The Ground Magnetic Perturbation working team of the International Forum for Space Weather Capabilities Assessment recommends that the validation methodology of Pulkkinen et al. (2013) be expanded in the following ways:

- Two new events should be added to the validation suite: the 17 March 2015 storm and the 22 July 2004 triple storm.
- The observational data set to be compared against should be expanded to include as many magnetometers as possible. As available, comparisons should be made at a 10-s sampling frequency.
- Regional analysis should be expanded to include MLT bins that test against any dB/dt threshold crossing within each bin.
- Metrics should be calculated as a function of substorm, ring current intensification, and sudden commencement activity.
- FB should be included as an additional metric.

The end result of this suite applied to a model is 10 sets of metrics: a set calculated over all stations at all times, a set calculated for high-latitude and then midlatitude stations, a set for each of four MLT bins, and a set for each of three activity types. Each set contains four values: POD, POFD, FB, and HSS. The results of this validation suite will better inform developers and users of a model's performance compared to the existing suite (Pulkkinen et al., 2013).

These recommendations represent the immediate next steps for improving ground perturbation validation. The validation suite as described here should not be considered a final product but an advancement of the Pulkkinen et al. (2013) efforts. As the capabilities and needs of the research and operational community evolve, so should the methodology for ground validation.

Acknowledgments
F10.7 data were obtained from the LASP Interactive Solar Irradiance Data Center (http://lasp.colorado.edu/lisird). Geomagnetic index data were 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 these data publicly available.

References

Austin, H. J., & Savani, N. P. (2018). Skills for forecasting space weather. *Weather*, 362–366. https://doi.org/10.1002/wea.3076

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

Carter, B. A., Yizengaw, E., Pradipta, R., Weygand, J. M., Piersanti, M., Pulkkinen, A., et al. (2016). Geomagnetically induced currents around the world during the 17 March 2015 storm. *Journal of Geophysical Research: Space Physics*, *121*, 10,496–10,507. https://doi.org/10.1002/2016JA023344

Divett, T., Richardson, G. S., Beggan, C. D., Rodger, C. J., Boteler, D. H., Ingham, M., et al. (2018). Transformer-level modeling of geomagnetically induced currents in New Zealand's South Island. *Space Weather*, *16*, 718–735. https://doi.org/10.1029/2018SW001814

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

Gjerloev, J. W. (2012). The SuperMAG data processing technique. *Journal of Geophysical Research*, *117*, A09213. https://doi.org/10.1029/2012JA017683

Guerrero, A., Palacios, J., Rodríguez-Bouza, M., Rodríguez-Bilbao, I., Aran, A., Cid, C., et al. (2017). Storm and substorm causes and effects at midlatitude location for the St. Patrick's 2013 and 2015 events. *Journal of Geophysical Research: Space Physics*, *122*, 9994–10,011. https://doi.org/10.1002/2017JA024224

Jolliffe, I. T., & Stephenson, D. B. (2012). *Forecast verification: A practitioner's guide in atmospheric science*. UK: John Wiley & Sons. 288 pp.

Kelbert, A., Balch, C. C., Pulkkinen, A., Egbert, G. D., Love, J. J., Rigler, E. J., & Fujii, I. (2017). Methodology for time-domain estimation of storm time geoelectric fields using the 3-D magnetotelluric response tensors. *Space Weather*, *15*, 874–894. https://doi.org/10.1002/2017SW001594

Lopez, R. E., Hernandez, S., Wiltberger, M., Huang, C. L., Kepko, E. L., Spence, H., et al. (2007). Predicting magnetopause crossings at geosynchronous orbit during the Halloween storms. *Space Weather*, *5*, S01005. https://doi.org/10.1029/2006SW000222

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

Newell, P. T., & Liou, K. (2011). Solar wind driving and substorm triggering. *Journal of Geophysical Research*, *116*, A03229. https://doi.org/10.1029/2010JA016139

Ngwira, C. M., Sibeck, D., Silveria, M. V. D., Georgiou, M., Weygand, J. M., Nishimura, Y., & Hampton, D. (2018). A study of intense local dB/dt variations during two geomagnetic storms. *Space Weather*, *16*, 676–693. https://doi.org/10.1029/2018SW001911

Pulkkinen, A., Bernabeu, E., Eichner, J., Viljanen, A., & Ngwira, C. (2015). Regional-scale high-latitude extreme geoelectric fields pertaining to geomagnetically induced currents. *Earth, Planets and Space*, *67*(1), 93. https://doi.org/10.1186/s40623-015-0255-6

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

Pulkkinen, A., Bernabeu, E., Thomson, A., Viljanen, A., Pirjola, R., Boteler, D., et al. (2017). Geomagnetically induced currents: Science, engineering, and applications readiness, https://doi.org/10.1002/2016SW001501

Pulkkinen, A., Rastätter, L., Kuznetsova, M., Singer, H., Balch, C., Weimer, D., et al. (2013). Community-wide validation of geospace model ground magnetic field perturbation predictions to support model transition to operations. *Space Weather*, *11*, 369–385. https://doi.org/10.1002/swe.20056

Pulkkinen, A., Viljanen, A., & Pirjola, R. (2006). Estimation of geomagnetically induced current levels from different input data. *Space Weather*, *4*, S08005. https://doi.org/10.1029/2006SW000229

Thomson, A. W. P. (2000). Evaluating space weather forecasts of geomagnetic activity from a user perspective. *Geophysical Research Letters*, *27*, 4049–4052. https://doi.org/10.1029/2000GL011908

Tóth, G., Meng, X., Gombosi, T. I., & Rastätter, L. (2014). Predicting the time derivative of local magnetic perturbations. *Journal of Geophysical Research: Space Physics*, *119*, 310–321. https://doi.org/10.1002/2013JA019456

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

Weigel, R. S., Detman, T., Rigler, E. J., & Baker, D. N. (2006). Decision theory and the analysis of rare event space weather forecasts. *Space Weather*, *4*, S05002. https://doi.org/10.1029/2005sw000157

Welling, D. T., & Ridley, A. J. (2010). Validation of SWMF magnetic field and plasma. *Space Weather*, *8*, S03002. https://doi.org/10.1029/2009SW000494

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

Yu, Y., & Ridley, A. J. (2008). Validation of the space weather modeling framework using ground-based magnetometers. *Space Weather*, *6*, S05002. https://doi.org/10.1029/2007SW000345

Erratum

Four authors affiliation statements have been modified to remove extra affiliations. Specifically removed were: Weigel's affiliation with the University of Michigan, Weygand's affiliation with the University of Michigan, Singer's affiliation with the University of Michigan and the University of Texas at Arlington, and Rosenqvist's affiliation with the University of Michigan and the Catholic University of America. This version with updated authorship may be considered the version of record.