Geospace Environment Modeling 2008–2009 Challenge: Ground magnetic field perturbations

A. Pulkkinen,[1,2] M. Kuznetsova,[2] A. Ridley,[3] J. Raeder,[4] A. Vapirev,[4] D. Weimer,[5] R. S. Weigel,[6] M. Wiltberger,[7] G. Millward,[8] L. Rastätter,[2] M. Hesse,[2] H. J. Singer,[8] and A. Chulaki[2]

Received 11 June 2010; revised 27 November 2010; accepted 10 December 2010; published 24 February 2011.

[1] Acquiring quantitative metrics-based knowledge about the performance of various space physics modeling approaches is central for the space weather community. Quantification of the performance helps the users of the modeling products to better understand the capabilities of the models and to choose the approach that best suits their specific needs. Further, metrics-based analyses are important for addressing the differences between various modeling approaches and for measuring and guiding the progress in the field. In this paper, the metrics-based results of the ground magnetic field perturbation part of the Geospace Environment Modeling 2008–2009 Challenge are reported. Predictions made by 14 different models, including an ensemble model, are compared to geomagnetic observatory recordings from 12 different northern hemispheric locations. Five different metrics are used to quantify the model performances for four storm events. It is shown that the ranking of the models is strongly dependent on the type of metric used to evaluate the model performance. None of the models rank near or at the top systematically for all used metrics. Consequently, one cannot pick the absolute "winner": the choice for the best model depends on the characteristics of the signal one is interested in. Model performances vary also from event to event. This is particularly clear for root-mean-square difference and utility metric-based analyses. Further, analyses indicate that for some of the models, increasing the global magnetohydrodynamic model spatial resolution and the inclusion of the ring current dynamics improve the models' capability to generate more realistic ground magnetic field fluctuations.

Citation: Pulkkinen, A., et al. (2011), Geospace Environment Modeling 2008–2009 Challenge: Ground magnetic field perturbations, *Space Weather*, *9*, S02004, doi:10.1029/2010SW000600.

1. Introduction

[2] Ground magnetic field perturbations are of interest to a number of industries impacted by space weather. For example, magnetic field perturbations induce potentially harmful electric currents to long conductor systems such as buried pipelines and high-voltage power transmission systems [e.g., *Boteler et al.*, 1998; *Pirjola*, 2005]. Also directional drilling used by the oil industry for drilling oil and gas offshore can be impacted by the unexpected geomagnetic field changes [e.g., *Reay et al.*, 2005]. In their attempt to mitigate possible space weather-related problems, these industries benefit from accurate modeling-based quantification of the geomagnetic field perturbations at the sites of their interest. Quantitative evaluation of the ground magnetic field perturbation model performance, which is the topic of the paper at hand, helps the potential users of the model products to better understand the capabilities of the models and to choose the approach that suits the best their specific needs.

[3] The Geospace Environment Modeling (GEM) community has recognized that, due to the increasing need for physics-based space weather modeling products and maturity and the increasing complexity of the state-of-the-art global space weather models, there is a great need for a

[1]Institute for Astrophysics and Computational Sciences, Catholic University of America, Washington, D. C., USA.
[2]NASA Goddard Space Flight Center, Greenbelt, Maryland, USA.
[3]Department of Atmospheric, Oceanic, and Space Sciences, University of Michigan, Ann Arbor, Michigan, USA.
[4]Space Science Center and Physics Department, University of New Hampshire, Durham, New Hampshire, USA.
[5]Center for Space Science and Engineering Research, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA.
[6]Department of Computational and Data Sciences, George Mason University, Fairfax, Virginia, USA.
[7]High Altitude Observatory, National Center for Atmospheric Research, Boulder, Colorado, USA.
[8]Space Weather Prediction Center, NOAA, Boulder, Colorado, USA.

Table 1. Geospace Events Studied in the Challenge[a]

Event	Date and Time	Min (Dst)	Max (Kp)
1	29 October 2003 0600 UT to 30 October 0600 UT	−353 nT	9
2	14 December 2006 1200 UT to 16 December 0000 UT	−139 nT	8
3	31 August 2001 0000 UT to 1 September 0000 UT	−40 nT	4
4	31 August 2005 1000 UT to 1 September 1200 UT	−131 nT	7

[a]The last two columns give the minimum Dst index and the maximum Kp index of the event, respectively.

systematic and quantitative evaluation of different geospace circulation modeling approaches. To respond to the need, the GEM Global Geospace Circulation Modeling (GGCM) Metrics and Validation Focus Group organized a modeling Challenge focusing on the inner magnetospheric dynamics and ground magnetic field perturbations. The new activity follows the series of earlier GEM Challenges [*Lyons*, 1998; *Birn et al.*, 2001; *Raeder and Maynard*, 2001]. The 2008–2009 Challenge is a natural next step to GEM efforts described by *Lyons* [1998] and *Raeder and Maynard* [2001] as instead of ionospheric convection or isolated substorm events, full storm events containing great variety of different geospace states are studied. Further, to facilitate unambiguous and objective interpretation of the Challenge results, a particular focus is now placed on systematic metrics-based analyses. The primary goals of the evaluations carried out in the 2008–2009 Challenge are to address differences between various modeling approaches, evaluate the current state of GGCM models, demonstrate effects of model coupling and grid resolution, encourage collaborations, and facilitate further model improvements.

[4] The Challenge was initiated at the summer GEM workshop 2008 in Midway, Utah and announced in September 2008. Model result submissions received by 1 September 2009 are included in this paper. The submission were made via Community Coordinated Modeling Center's (CCMC) online submission system, which also enables online model comparisons (see http://ccmc.gsfc.nasa.gov). Further, a number of model submissions were generated via CCMC's runs-on-request system. The corresponding simulations are publicly available for analysis via CCMC's visualization interface.

[5] Preliminary analysis of a subset of the Challenge model submissions were presented by *Pulkkinen et al.* [2010]. In this paper, the final metrics-based Challenge results pertaining to the ground magnetic field perturbations are reported. In a companion paper by L. Rastätter et al. (Geospace Environment Modeling 2008–2009 Challenge: Geosynchronous magnetic field, submitted to *Space Weather*, 2010), results pertaining to the geostationary magnetic field perturbations are reported. For brevity, no detailed scientific analysis of the model submissions is carried out in this work. Focused scientific analyses will be reported in follow-up studies. It is emphasized that the individual model comparisons can be examined in further detail by using the online metric tool available at the CCMC's website.

2. The Challenge Setup

[6] Four geospace storm events listed in Table 1 were chosen for the study. Solar wind bulk plasma and the interplanetary magnetic field observations carried out by SWEPAM and MAG instruments onboard Advanced Composition Explorer (ACE) for the events are shown in Figure 1. Note that due to problems with data from the SWEPAM instrument during the October 2003 event (event 1), only low temporal resolution plasma velocity data could be constructed [*Skoug et al.*, 2004]. Further, the plasma density data for the event was obtained from the Geotail Plasma Wave Instrument. Events 1 and 2 are well-known coronal mass ejection-related major storm events while events 3 and 4 are less active periods associated with much more subtle changes in the solar wind driving. Only Lagrange 1 solar wind observations were used to drive the models discussed in this paper.

[7] For each event in Table 1, the model performance is evaluated by means of model versus observations comparisons for the following geospace parameters: (1) parameter 1, magnetic field at the geosynchronous orbit; (2) parameter 2, magnetopause crossings by geosynchronous satellite; (3) parameter 3, plasma density/temperature at the geosynchronous orbit; (4) parameter 4, ground magnetic field perturbations; and (5) parameter 5, Dst index.

[8] One minute geomagnetic observatory recordings were used to provide the unambiguous high-precision "ground truth" for parameter 4 discussed in this paper. 12 geomagnetic observatories (magnetometer stations) listed in Table 2 and shown in Figure 2 were selected based on the global spatial and temporal coverage. One minute temporal resolution magnetic field recordings were downloaded via INTERMAGNET (www.intermagnet.org). The data were transformed from geographic coordinates, as provided by INTERMAGNET, into geomagnetic dipole coordinates. IGRF 2000 coefficients were used to compute the coordinate transformation matrices as given by *Hapgood* [1992]. The quiet time baseline level was determined visually for each station and for each event and the baseline was removed from the magnetic field data to obtain the disturbance field. Small data gaps with length of no more than few minutes

Figure 1. Solar wind bulk plasma and the interplanetary magnetic field observations for the studied storm events ((a-d) events 1–4) given in Table 1. See the text for details. Adopted from *Pulkkinen et al.* [2010].

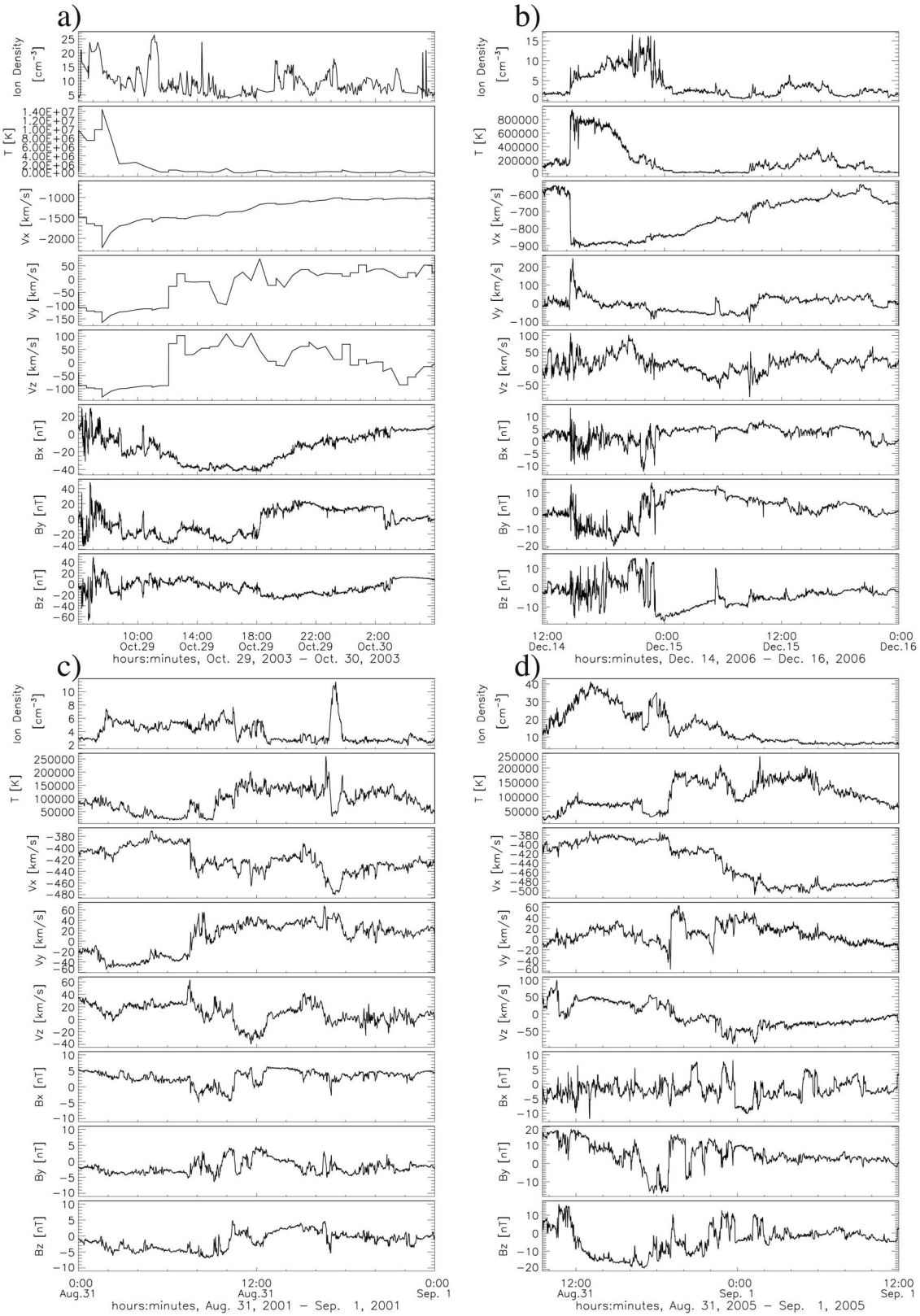
a)
Ion Density [cm⁻³]
T [K]
Vx [km/s]
Vy [km/s]
Vz [km/s]
Bx [nT]
By [nT]
Bz [nT]
10:00 Oct.29
14:00 Oct.29
18:00 Oct.29
22:00 Oct.29
2:00 Oct.30
hours:minutes, Oct. 29, 2003 – Oct. 30, 2003
b)
12:00 Dec.14
0:00 Dec.15
12:00 Dec.15
0:00 Dec.16
hours:minutes, Dec. 14, 2006 – Dec. 16, 2006
c)
0:00 Aug.31
12:00 Aug.31
0:00 Sep. 1
hours:minutes, Aug. 31, 2001 – Sep. 1, 2001
d)
12:00 Aug.31
0:00 Sep. 1
12:00 Sep. 1
hours:minutes, Aug. 31, 2005 – Sep. 1, 2005

Figure 1

Table 2. The Locations of the Geomagnetic Observatories Used in the Study

Station Code	Geomagnetic Latitude	Geomagnetic Longitude
YKC	68.9	299.4
MEA	61.6	306.2
NEW	54.9	304.7
FRN	43.5	305.3
IQA	74.0	5.2
PBQ	65.5	351.8
OTT	55.6	355.3
FRD	48.4	353.4
HRN	73.9	126.0
ABK	66.1	114.7
WNG	54.1	95.0
FUR	48.4	94.6

were patched by means of linear interpolation. The modeled magnetic field data were resampled by means of spline interpolation to match the time stamps of the observations.

[9] Note that although the Dst index addresses the low-latitude geomagnetic phenomena, no low-latitude magnetometer stations were included in the Challenge. This is particularly important for first-principles models as due to the inner boundary, which is located typically at 3–4 Earth radii, the global magnetohydrodynamic (MHD) approach can provide ionospheric output only at high latitudes. This constraint can be alleviated by coupling global MHD models to inner magnetospheric models capturing the ring current dynamics and providing ionospheric output also at lower latitudes [e.g., *De Zeeuw et al.*, 2004; *Toffoletto et al.*, 2004] [see also *Yu et al.*, 2010]. It is noted that although some of the model submissions included in the Challenge use coupling to an inner magnetospheric model, only high-latitude ionospheric currents are used in the computation of the ground magnetic field perturbations.

[10] Due to the computational constraints, in model submissions that were generated via CCMC's runs-on-request system, only horizontal and field-aligned ionospheric currents within 1000 km radius about the location of the magnetometer station are taken into account (see *Pulkkinen et al.* [2010] for details). While integration over only part of the ionosphere is acceptable when studying the horizontal components of the ground magnetic field, the vertical component may be poorly estimated [*Yu and Ridley*, 2008]. Further, some of the Challenge submissions included only the horizontal components of the magnetic field. Consequently, only horizontal components of the ground magnetic field are used in the analyses carried out in this paper.

3. Model Performance Metrics

[11] Formally, the term *metric* is used here to refer to functions mapping two elements of a set (e.g., time series of the observed and the modeled ground magnetic fields) into a single real number. In model evaluations, the number characterizes the performance of the model with respect to the observations. Further, there is no metric that is universally applicable to all situations. The metric needs to be selected or designed carefully based on the characteristics of the studied signal one is mostly interested in. Based on the GEM community input, five different metrics are used in evaluating the model performances in this paper. These five metrics are described below.

3.1. Root-Mean-Square Difference

[12] One of the classic means to quantify the difference between two elements of a set is to compute the root-mean-square difference defined as

$$RMS = \sqrt{\left\langle (x_{obs} - x_{mod})^2 \right\rangle_i} \quad (1)$$

where x_{obs} and x_{mod} are the observed and the modeled signals, respectively, $\langle .. \rangle_i$ indicates arithmetic mean (all means/averages taken in this work are arithmetic means) taken over i. Throughout this work i corresponds to the time series over individual events. $RMS = 0$ indicates perfect model performance. It should be noted that in contrast to other metrics used in this work, RMS has a dimension, which is equal to that of signal x. Further, it should also be noted that since RMS is not normalized, comparisons between events having large differences in the amplitude of the signal can be somewhat problematic, as will be seen below.

3.2. Prediction Efficiency

[13] Another commonly used metric is the prediction efficiency (PE) defined as

$$PE = 1 - \frac{\left\langle (x_{obs} - x_{mod})^2 \right\rangle_i}{\sigma_{obs}^2} \quad (2)$$

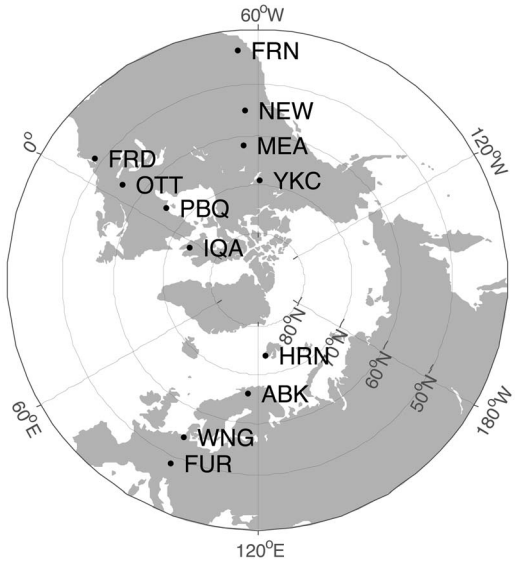

Figure 2. The locations and the station codes of the geomagnetic observatories used in the study. Geomagnetic dipole coordinates are used.

where x_{obs} and x_{mod} are the observed and the modeled signals, respectively, $\langle..\rangle_i$ indicates arithmetic mean taken over i and σ^2_{obs} is the variance of the observed signal. Note that $PE = 1$ indicates a perfect prediction while $PE = 0$ means that the model predicts the signal equally well to a model that uses the mean value of the signal as a predictor.

3.3. Log-Spectral Distance

[14] Geomagnetically induced currents (GIC) flowing, for example, in high-voltage power transmission systems are one of the major ground-based space weather hazards. Consequently, the third applied metric was developed to quantify the models' capability to reproduce the GIC-related features of the magnetic field fluctuations. For this, it was assumed that the horizontal geoelectric field ($\tilde{E}_x$, $\tilde{E}_y$) can be calculated from the observed or modeled ground magnetic field accurately by means of the plane wave approach [see, e.g., *Pirjola*, 2004]. In the plane wave approach, the spectral domain electric field is obtained by computing

$$\tilde{E}_{x,y} = (-)\frac{\tilde{Z}(\omega)}{\mu_0}\tilde{B}_{y,x} \tag{3}$$

where μ_0 is the vacuum permeability, ($\tilde{B}_x$, $\tilde{B}_y$) the spectral domain magnetic field (tilde denotes quantities in the spectral domain and ω denotes the angular frequency) and $\tilde{Z}(\omega)$ the surface impedance. In equation (3), the minus sign applies for the ($\tilde{E}_y$, $\tilde{B}_x$) pair. Further, it was assumed that GIC can be obtained from the relation

$$\tilde{GIC}(\omega) = a\tilde{E}_x + b\tilde{E}_y = \frac{a}{\mu_0}\tilde{Z}\tilde{B}_y - \frac{b}{\mu_0}\tilde{Z}\tilde{B}_x \tag{4}$$

where a and b are the system parameters that depend on the topology and the electrical properties of the investigated conductor system. For power spectra, the relation (4) takes the form of a triangle inequality

$$|\tilde{GIC}(\omega)| \le |\frac{a}{\mu_0}||\tilde{Z}||\tilde{B}_y| + |\frac{b}{\mu_0}||\tilde{Z}||\tilde{B}_x| \tag{5}$$

Motivated by the inequality (5), one then computes the logarithm of the ratios of the right-hand side of equation (5) and assumes $|a| = |b|$ to introduce m_s defined as

$$m_s(\omega) = \log\left[\frac{|\tilde{B}_x|_{obs} + |\tilde{B}_y|_{obs}}{|\tilde{B}_x|_{mod} + |\tilde{B}_y|_{mod}}\right] \tag{6}$$

which is a dimensionless quantity characterizing model's ability to reproduce the GIC-related magnetic field fluctuations. To obtain a single number, one finally computes

$$M_s = \sqrt{\frac{1}{N}\sum_{\omega} m_s^2} \tag{7}$$

where the sum is over N frequencies. M_s in equation (7) is called *log-spectral distance* measuring the positive definite distance from the perfect ($M_s = 0$) model performance. The sum in equation (7) is carried out over the periods 2–120 min in this work.

[15] The power spectra for individual time series in equation (6) were computed as follows. First the time series was divided into 2 h long segments with 50% overlap between neighboring segments. The data within each segment was multiplied with the Hann window function [*Press et al.*, 1992, p. 554] prior to the fast fourier transform and the final spectrum was obtained by taking an average over the segments.

3.4. Utility

[16] The fourth metric used in the Challenge is the so-called utility metric (for details, see *Weigel et al.* [2006, and references therein]). In using the utility, one quantifies the performance in terms of a model's capability to predict *events*. An event is defined here as follows: within a forecast window $0 \le t \le t_f$, the absolute value of the parameter of interest exceeds event threshold $|x_{thres}|$ (note that in more general cases the absolute value is not necessarily required). The windows are moved over the time series in nonoverlapping parts and events for given t_f and $|x_{thres}|$ are recorded for both the measured and the modeled x. The utility of the forecast is defined as

$$U_f = BN_H - CN_{\overline{H}} \tag{8}$$

where N_H is the number of correct forecasts, $N_{\overline{H}}$ is the number of false alarms, C is the cost of taking mitigating action and B is the benefit from having taken mitigating action when an event occurred. In using this particular metric, one assumes that 1.) the user takes the same mitigating action following each forecast of an event, 2.) an "always mitigate" strategy yields a net monetary loss for the user and 3.) the user seeks to maximize the monetary gain U_f. Note that one is considering utility with respect to a system that is never mitigated. In another words, the difference between losses/gains experienced by the two systems is considered. It follows that missed events need not to be considered; a missed event will cause the same monetary loss for both the reference system and the system using mitigation actions.

[17] As values for B and C in equation (8) are system-dependent and are estimated by the user, rather than computing U_f, the forecast ratio $R_f = N_H/N_{\overline{H}}$ is reported here, which is an approach suggested by *Weigel et al.* [2006]. It is easily seen that the utility U_f is positive if $R_f > C/B$ and thus by reporting R_f, the user can determine if the forecasts can be used to provide positive utility once the values B and C are known. In model comparisons, a model with a larger R_f will have a greater utility U_f. It is noted that there are a number of ways to carry out the event-based analysis, the utility of the forecast being only one alternative to characterize the model performance. For example, event-based metrics such as probability of detection (POD) and false alarm ratio (FAR) have also been used to evaluate space physics model performance [*Lopez et al.*, 2007].

Table 3. Model Submissions Analyzed in the Challenge[a]

Identifier	Model	Grid (Number of Cells, Min. Res.)
1_CMIT	CMIT 2.0, LFM coupled to TIEGCM	40,000, 0.5 R_e
1_LFM	LFM	160,000, 0.3 R_e
1_OPENGGCM	OpenGGCM v3.1 coupled to CTIM	3 million, 0.3 R_e
2_OPENGGCM	OpenGGCM v3.1 coupled to CTIM	6.5 million, 0.25 R_e
1_SWMF	SWMF v7.73, BATS-R-US	2 million, 0.25 R_e
2_SWMF	SWMF v7.73, BATS-R-US	700,000, 0.25 R_e
3_SWMF	SWMF v8.01 BATS-R-US coupled to RCM	2 million, 0.25 R_e
4_SWMF	SWMF v8.01, BATS-R-US	3 million, 0.125 R_e
5_SWMF	SWMF v8.01, BATS-R-US coupled to RCM	3 million, 0.125 R_e
6_SWMF	SWMF v20090403, BATS-R-US coupled to RCM	900,000, 0.25 R_e
1_WEIMER	*Weimer* [2005]	4 min output interpolated into 1 min
2_WEIMER	New empirical model by D. Weimer	4 min output interpolated into 1 min
1_WEIGEL	*Weigel et al.* [2003]	30 min output

[a]Each model is assigned a unique model identifier given in the first column. The table indicates the model version, and if applicable, the number of cells and the minimum spatial resolution used in the global MHD part of the model. Note that different model setups are referred as different "models."

[18] From the GIC viewpoint, the time derivative of the magnetic field is typically a better indicator of the activity than the mere amplitude of the magnetic field [e.g., *Viljanen et al.*, 2001]. Consequently, in this work the forecast ratios are computed both for the predictions of the horizontal magnetic field amplitude $|\mathbf{B}_h| = \sqrt{B_x^2 + B_y^2}$ and the amplitude of the time derivative of the horizontal magnetic field vector $|d\mathbf{B}_h/dt|$.

3.5. Ratio of the Maximum Amplitudes

[19] Sometimes only the forecasted maximum amplitude of the space weather event is of interest to the end user. Consequently, the fifth metric used is the ratio of the maximum amplitudes $R_{\max}$:

$$R_{\max} = \frac{\max(|x_{mod}|_i)}{\max(|x_{obs}|_i)} \tag{9}$$

where x_{obs} and x_{mod} are the observed and the modeled signals, respectively, and the maximum is taken over i. Clearly, $R_{\max} = 1$ indicates perfect model performance while $R_{\max} > 1$ and $R_{\max} < 1$ indicate that model overestimates and underestimates, respectively, the maximum amplitude of the signal. The ratios of the maximum amplitudes are computed both for the predictions of the horizontal magnetic field amplitude $|\mathbf{B}_h| = \sqrt{B_x^2 + B_y^2}$ and the amplitude of the time derivative of the horizontal magnetic field vector $|d\mathbf{B}_h/dt|$.

4. Models

[20] Table 3 summarizes the model submissions analyzed in the Challenge. In the following each model is described briefly. One should see the given references for more detailed descriptions of individual models. Models are referred below by using the identifiers given in Table 3.

[21] 1. Identifier 1_CMIT, Coupled Magnetosphere-Ionosphere-Thermosphere (CMIT) [*Wiltberger et al.*, 2004] version 2.0 consists of coupled Lyon-Fedder-Mobarry (LFM) global MHD model of the magnetosphere [*Lyon et al.*, 2004] and Thermosphere-Ionosphere Electrodynamics General Circulation Model (TIEGCM) [*Richmond et al.*, 1992]. Ionospheric currents from TIEGCM were used to compute the ground magnetic field perturbations. In contrast to the other Challenge global MHD models that use Cartesian grids, CMIT uses a distorted spherical grid in the magnetosphere. Event 1 was not submitted for 1_CMIT.

[22] 2. Identifier 1_LFM is LFM without coupling to TIEGCM. Event 1 was not submitted for 1_LFM.

[23] 3. Identifers 1_OPENGGCM and 2_OPENGGCM are OpenGGCM version 3.1 global MHD model [*Raeder et al.*, 2001] coupled to Coupled Thermosphere Ionosphere Model (CTIM) [*Fuller-Rowell et al.*, 1996]. OpenGGCM uses a stretched Cartesian grid. Event 1 was not submitted for 1_OPENGGCM, and events 1 and 3 were not submitted for 2_OPENGGCM.

[24] 4. Identifiers 1_SWMF to 6_SWMF are Space Weather Modeling Framework (SWMF) versions 7.73, 8.01 and 20090403 including BATS-R-US global MHD model [*Powell et al.*, 1999] with and without coupling to Rice Convection Model (RCM) of the inner magnetosphere [*De Zeeuw et al.*, 2004; *Tóth et al.*, 2005]. BATS-R-US uses a block-adaptive Cartesian grid.

[25] 5. Identifier 1_WEIMER is an empirical model by *Weimer* [2005] of ionospheric electric fields and field-aligned currents. This model is based on measurements of the electric and magnetic fields with the Dynamics Explorer 2 satellite. The *Weimer* [2005] publication describes the method by which the ionospheric fields are used for predicting the geomagnetic variations. As it takes of the order of 20 min for the entire ionosphere to reconfigure in response to the interplanetary magnetic field changes, for these predictions the input to the model uses averages of the interplanetary magnetic field in 15 min intervals, stepping forward at 4 min increments. These results were interpolated to 1 min steps, as required for the Challenge comparisons.

[26] 6. Identifier 2_WEIMER is a new empirical model by D. Weimer. Partial description of the model is available in the work by *Weimer et al.* [2010]. The model is based on measurements from 105 ground magnetometers at geomagnetic latitudes above 35°N and simultaneous measurements from the ACE spacecraft. This database spans the time period from 1998 through 2001. This model also uses spherical harmonics, with coefficients that vary according to the interplanetary magnetic field and dipole tilt angle. A very preliminary version was completed just prior to the deadline for the Challenge submissions. For these metric tests, this model also used output calculated at 4 min intervals. These results were interpolated to 1 min steps, as required for the Challenge comparisons.

[27] 7. Identifier 1_WEIGEL is an empirical model by *Weigel et al.* [2003]. The model generates output with a 30 min temporal resolution, which was used in the model submissions.

[28] All model submissions except 1_WEIGEL had a 1 min temporal resolution. 1_WEIGEL was interpolated to a 1 min resolution for computing the prediction efficiency discussed in section 3.2. Submissions 1_OPENGGCM, 2_OPENGGCM and 1_SWMF to 5_SWMF were generated at CCMC and corresponding simulations are publicly available for analysis via CCMC's visualization interface. 1_OPENGGCM and 1_SWMF to 3_SWMF were studied by *Pulkkinen et al.* [2010]. For submissions generated at CCMC, the ground magnetic field perturbations were computed by using the approach described by *Pulkkinen et al.* [2010]. In addition to the models in Table 3, also an ensemble model was generated (identifier ENSEMBLE). The ensemble was formulated by computing the average over all model predictions. The averaging was carried out separately for each station and for both horizontal magnetic field components. It is noted that due to a number of SWMF submissions, the ensemble model generated by simple averaging may be weighted slightly toward that specific framework.

5. Results

[29] Figure 3 shows an example of the observed and modeled ground magnetic field perturbations for the event 2 for one of the meridional chains of magnetometer stations used in the study. Note that one can generate similar plots via CCMC's online metrics tool (see http://ccmc.gsfc.nasa.gov). Although the data is shown only to provide a qualitative visual impression on how models compare to the observations for this particular event, one can identify some differences between the models. For example, empirical models 2_WEIMER and 1_WEIGEL seem to give the amplitude of the negative magnetic field perturbation at the lower-latitude stations OTT and FRD better than the first-principles models. It is also clear from Figure 3 that storm time field fluctuations are a challenge to predict accurately. However, it is difficult to make definite conclusions based on a large number of partially overlapping curves. For more definite comparisons between the models, one needs to quantify the model performance by using metrics, which is the main objective of the work at hand.

[30] Figures 4, 5, 6, and 7 show the results of the metrics-based analyses. In Figures 4–7, results for individual events and the average over the events are shown. It should be noted that not all models had data for all four events and thus some caution should be taken when comparing the averages taken over the events.

[31] Figure 4 provides the model ranking based on the root-mean-square difference, prediction efficiency and the log-spectral distance described in sections 3.1, 3.2 and 3.3, respectively. Figure 4a shows that the root-mean-square differences rather systematically increase as a function of geomagnetic activity: differences are much larger for the Kp index 9 event 1 than for the Kp index 4 event 3. It should be noted that models 1_LFM and 1_CMIT are ranked favorably in terms of mean root-mean-square difference partly because the corresponding models had no data for the event 1 that is associated with the largest root-mean-square differences. With this notion, models 1_LFM, 2_WEIMER and 1_CMIT are the top ranking models in terms of the root-mean-square difference. As is seen from Figure 4b, the empirical models 2_WEIMER, 1_WEIGEL and the ensemble model ENSEMBLE are the top ranking models in terms of the prediction efficiency. The prediction efficiencies observed are consistent with those found by *Weigel et al.* [2003], who showed that in the auroral zone, prediction efficiencies varied from 0 to 0.6. First-principles-based models fail to provide positive average prediction efficiencies, which is yet another indication that storm time ground magnetic field perturbations are a challenge to predict accurately. On the other hand, empirical models fail to generate the rapid changes in the magnetic field that are necessary to perform well in log-spectral distance-based comparisons as seen from Figure 4c. One possible reason for the empirical models' good performance in terms of the prediction efficiency and poor performance in terms of the log-spectral distance is that the parameters for the models were selected by minimizing an error function that is related to the root-mean-square difference and prediction efficiency. Due to the low-cadence 30 min output, 1_WEIGEL is not included in the spectral or $|d\mathbf{B}_h/dt|$-based analyses.

[32] Figures 5 and 6 show the results of the utility metric analysis described in section 3.4. The forecast ratios were computed by integrating the numbers of correct forecasts and false alarms over all stations. The thick dashed line in Figure 5 was obtained by integration over both stations and events. It should be noted that statistics for individual events may be poor especially with large amplitude thresholds and in some cases no correct forecasts or false alarms are recorded. In such cases the forecast ratio is zero or infinite and the corresponding values are not present in Figures 5 and 6. Since the integration over both stations and events takes into account correct forecasts or false alarms associated with possible zero or infinite forecast ratios for individual events, the thick dashed lines in Figures 5 and 6

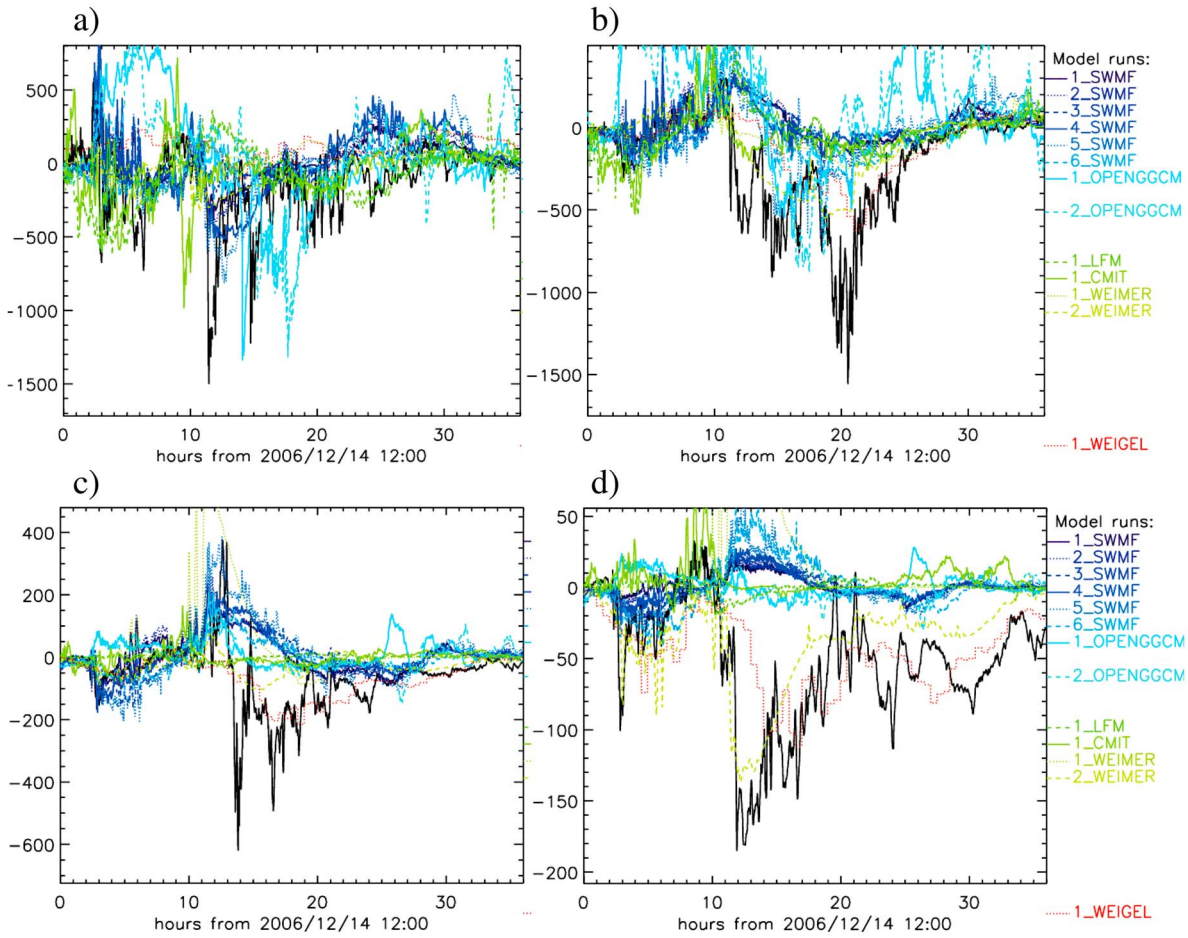

Figure 3. Observed (black line) and modeled geomagnetic north-south component of the magnetic field (in nT) for the event 2 at four stations: (a) IQA, (b) PBQ, (c) OTT, and (d) FRD. See Figure 2 for locations of the stations. The coding of each modeled trace is indicated on the right-hand side of Figures 3b and 3d.

do not necessarily go through values associated with lone individual events seen in the panels. Forecast window length of 45 min, which is comparable to the forecast lead time obtained with the Lagrange 1 observations-based modeling approaches, was used in the utility metric analyses.

[33] In Figure 5, model forecast ratios are shown for four different thresholds of the horizontal magnetic field amplitude: 100, 300, 500 and 650 nT. These represent the midrange amplitudes for the four storm events. Similarly to Figure 4a, the empirical models 2_WEIMER, 1_WEIGEL and the ensemble model ENSEMBLE are seen to rank in the top for all threshold levels. It can also be observed that there are considerable differences between different events: forecast ratios associated with the event 1 tend to be clearly higher for most models especially at the thresholds of 100 and 300 nT.

[34] In Figure 6 model forecast ratios are shown for four different thresholds of the amplitude of the time derivative of the horizontal magnetic field vector: 0.3, 0.7, 1.1 and 1.5 nT/s. Also these represent the midrange amplitudes for the four storm events. Although forecast ratios vary significantly between different events and general interpretations should thus be made with caution, 2_WEIMER and ENSEMBLE are again seen to rank in the top for all threshold levels. This is somewhat surprising given the fact that the model 2_WEIMER lacked the rapid fluctuations needed to rank in the top in Figure 4b. The high ranking in terms of the forecast ratio can be attributed to the generally low amplitudes of $|d\mathbf{B}_h/dt|$ and consequently very low rate of false alarms generated by 2_WEIMER. Similar to Figure 5, it can be observed from Figure 6 that the forecast ratios associated with the event 1 tend to be higher for most models.

[35] Finally, Figure 7 shows the ratios of the maximum amplitudes of both the horizontal magnetic field and the time derivative of the horizontal magnetic field vector. It is noted that since the ratio of the maximum amplitudes can take values on both sides of the optimal $R_{max} = 1$, a model that both overestimates and underestimates the maximum amplitudes can perform optimally in an average sense.

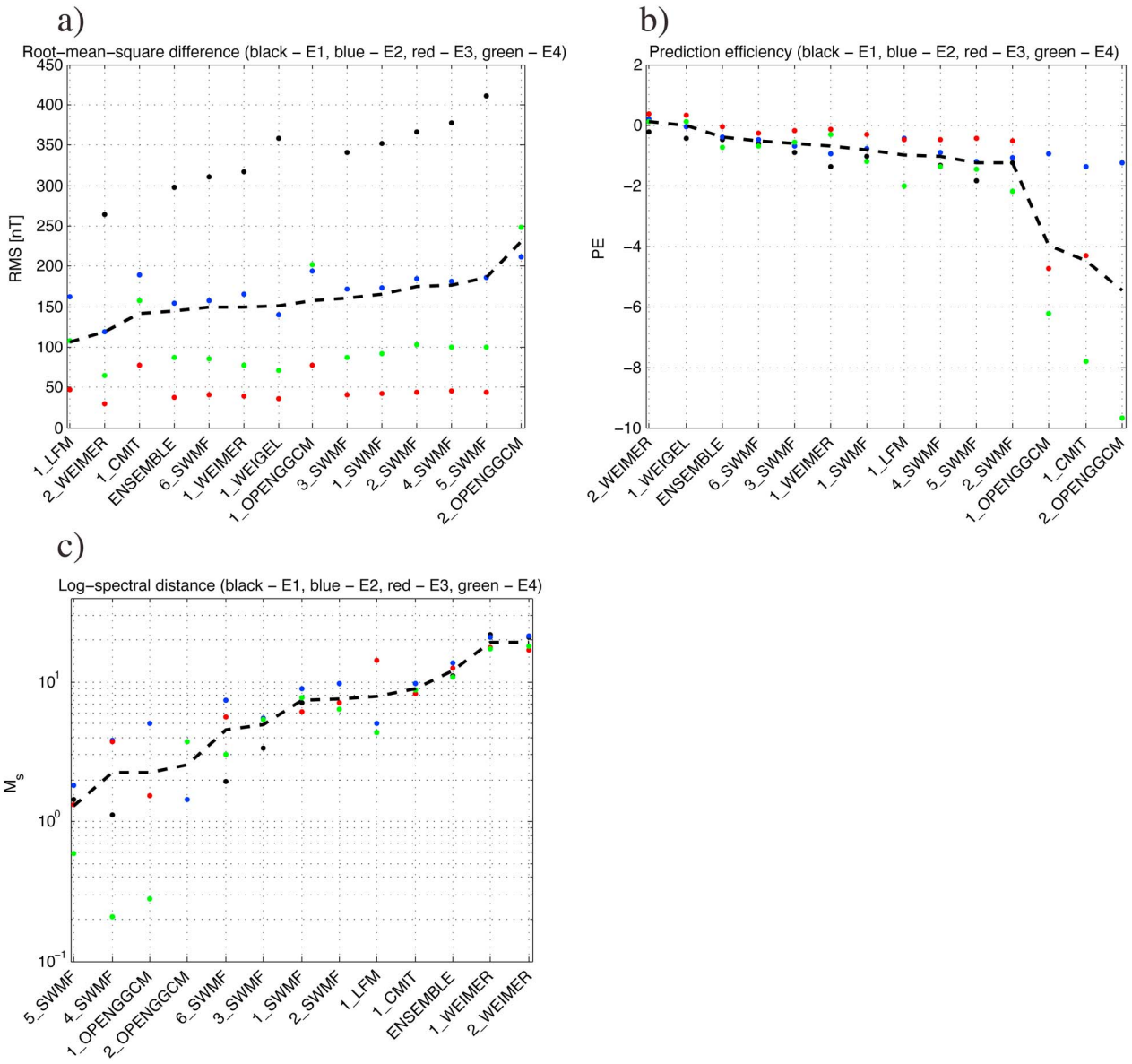

Figure 4. Ranking of the models according to (a) average (average taken over stations and horizontal field components) root-mean-square difference, (b) the average prediction efficiency (average taken over stations and horizontal field components), and (c) the average log-spectral distance (average taken over stations). In Figures 4a–4c the best performing model is located in the extreme left. Dots with different colors correspond to different events: black, event 1; blue, event 2; red, event 3; green, event 4. The thick dashed line gives the model average taken over different events. The ranking is based on the averages taken over the events. See Table 3 for model identifiers on the horizontal axis. Identifier 1_WEIGEL is not included in the analysis of Figure 4c.

Consequently, no model ranking is given in Figure 7. The results are used to provide only general indication about models' capability to generate the extreme amplitudes associated with the four events.

[36] It is seen from Figure 7 that most of the models tend to underestimate the maximum $|\mathbf{B}_h|$ and $|d\mathbf{B}_h/dt|$. For example, as seen from Figure 7b, 2_WEIMER gives systematically only about 20%–30% of the maximum $|d\mathbf{B}_h/dt|$ amplitudes. Only models 1_CMIT, 1_OPENGGCM and 2_OPENGGCM are able to generate comparable and higher maximum $|\mathbf{B}_h|$ in comparison to the observations. Models 3_SWMF to 6_SWMF, 1_OPENGGCM and 2_OPENGGCM are able to

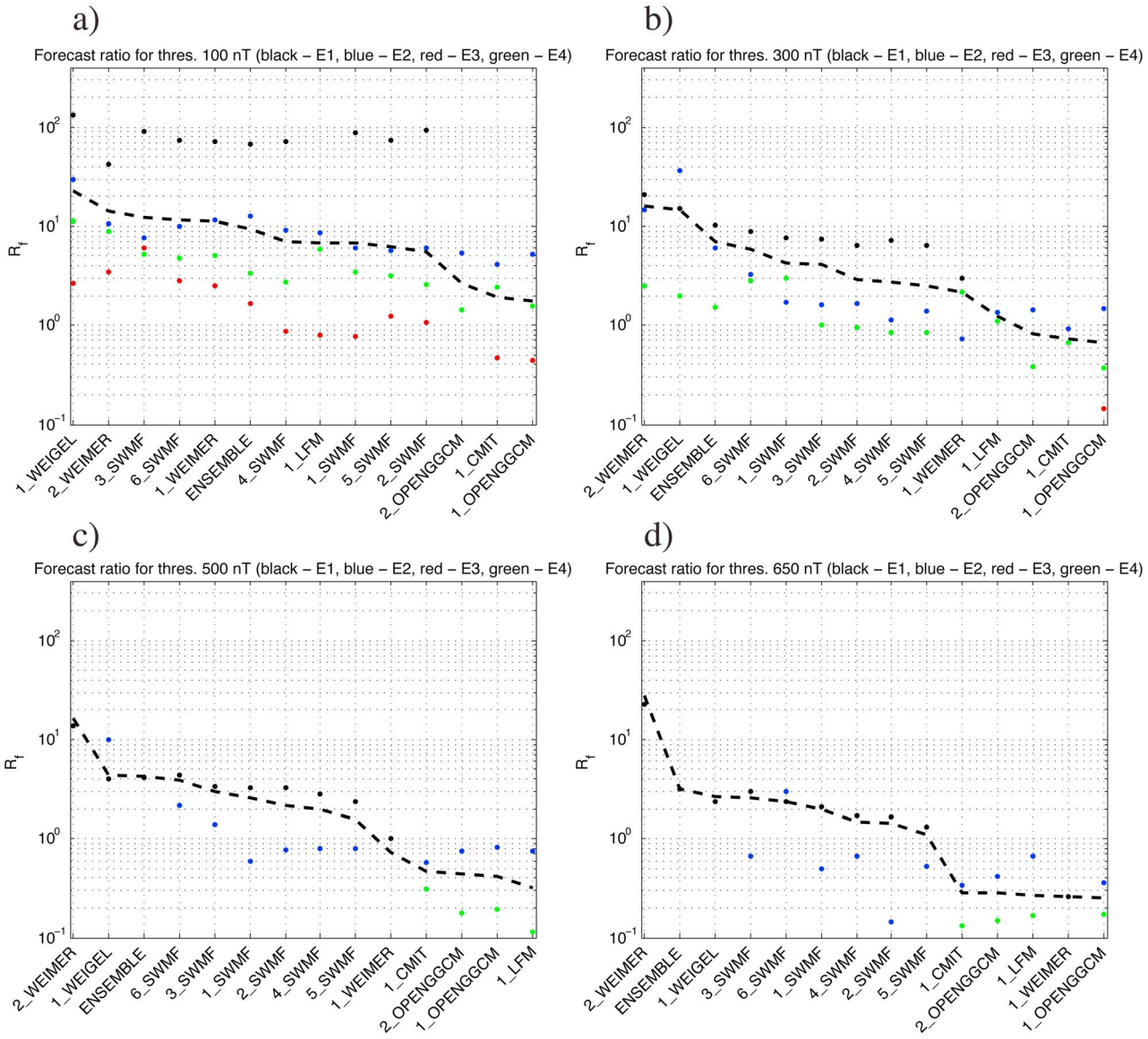

Figure 5. Ranking of the models according to the integrated (integration over stations) forecast ratio for different thresholds of the horizontal magnetic field amplitude. (a) Threshold 100 nT, (b) threshold 300 nT, (c) threshold 500 nT, and (d) threshold 650 nT. The best performing model is located in the extreme left of each panel. Dots with different colors correspond to different events: black, event 1; blue, event 2; red, event 3; green, event 4. The thick dashed line gives the forecast ratio integrated over different stations and events. The ranking is based on values integrated over stations and events. See Table 3 for model identifiers on the horizontal axis. Forecast window length of 45 min was used in the analysis.

generate comparable and higher maximum $|d\mathbf{B}_h/dt|$ in comparison to the observations.

6. Discussion

[37] In this paper, the metrics-based results of the ground magnetic field perturbation part of the GEM 2008–2009 Challenge were reported. Predictions made by 14 different models, including an ensemble model, were compared to geomagnetic observatory recordings from 12 different northern hemispheric locations. Five different metrics were used to quantify the model performances.

[38] It was seen that the ranking of the models is strongly dependent on the type of metric used to evaluate the model performance. None of the models ranked to the top systematically for all used metrics. For example, empirical

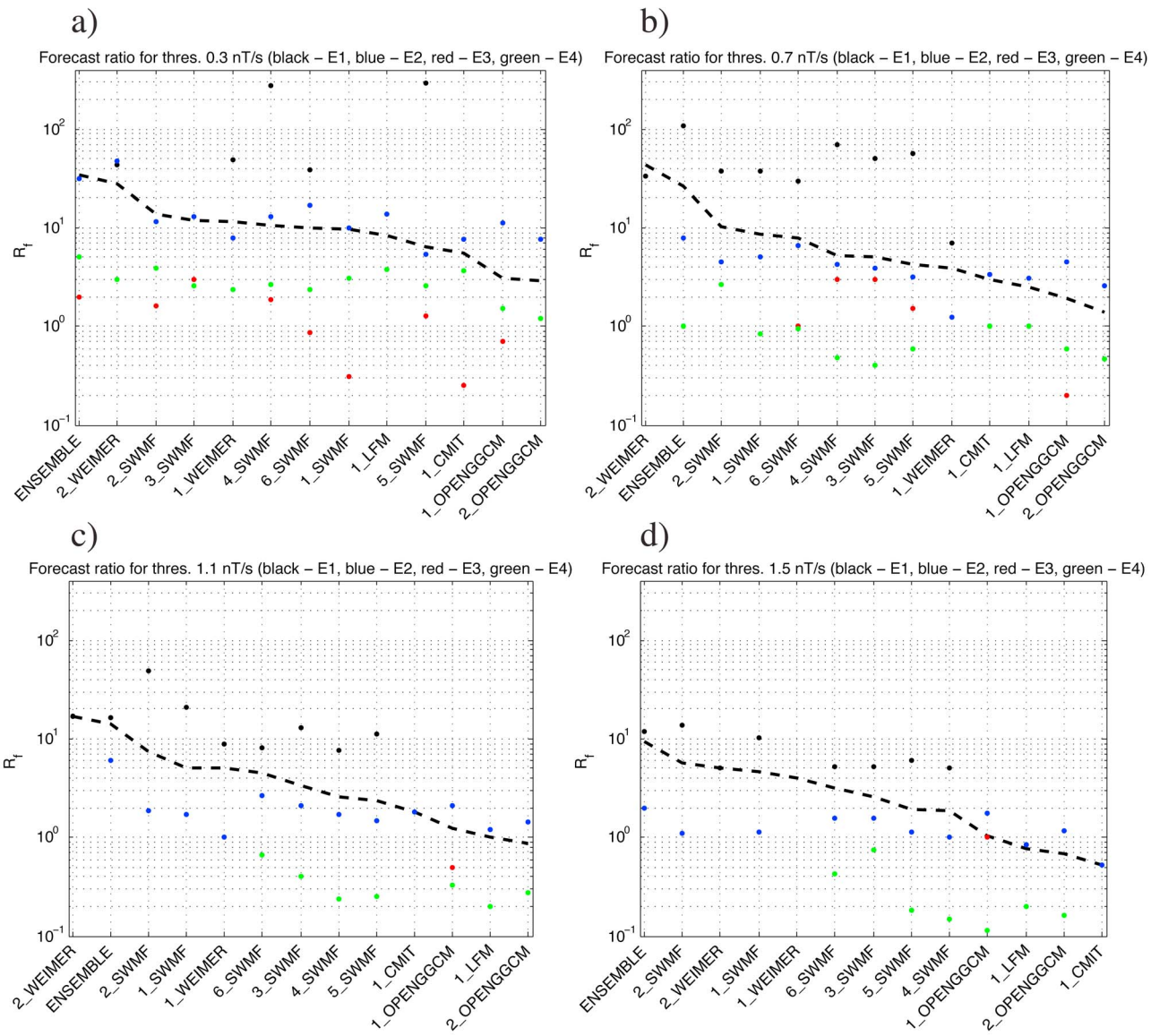

Figure 6. Ranking of the models according to the integrated (integration over stations) forecast ratio for different thresholds of the amplitude of the time derivative of the horizontal magnetic field vector. (a) Threshold 0.3 nT/s, (b) threshold 0.7 nT/s, (c): threshold 1.1 nT/s, and (d) threshold 1.5 nT/s. The best performing model is located in the extreme left of each panel. Dots with different colors correspond to different events: black, event 1; blue, event 2; red, event 3; green, event 4. The thick dashed line gives the forecast ratio integrated over different stations and events. The ranking is based on values integrated over stations and events. See Table 3 for model identifiers on the horizontal axis. Identifier 1_WEIGEL is not included in the analysis. Forecast window length of 45 min was used in the analysis.

models 2_WEIMER and 1_WEIGEL and the ensemble model ENSEMBLE were the top models in terms of the prediction efficiency and the forecast ratio while first-principles models performed better in terms of the log-spectral distance. The first-principles models were also more successful in generating the maximum $|\mathbf{B}_h|$ and $|d\mathbf{B}_h/dt|$ associated with the studied storm events.

[39] Model performances varied also from event to event and in many cases the scatter was larger than the difference between individual models. The varying model performance between events was particularly clear for utility metric-based analysis where the event 1; that is, the Halloween storm event was seen to be associated with the largest forecast ratios. This result emphasizes the need for

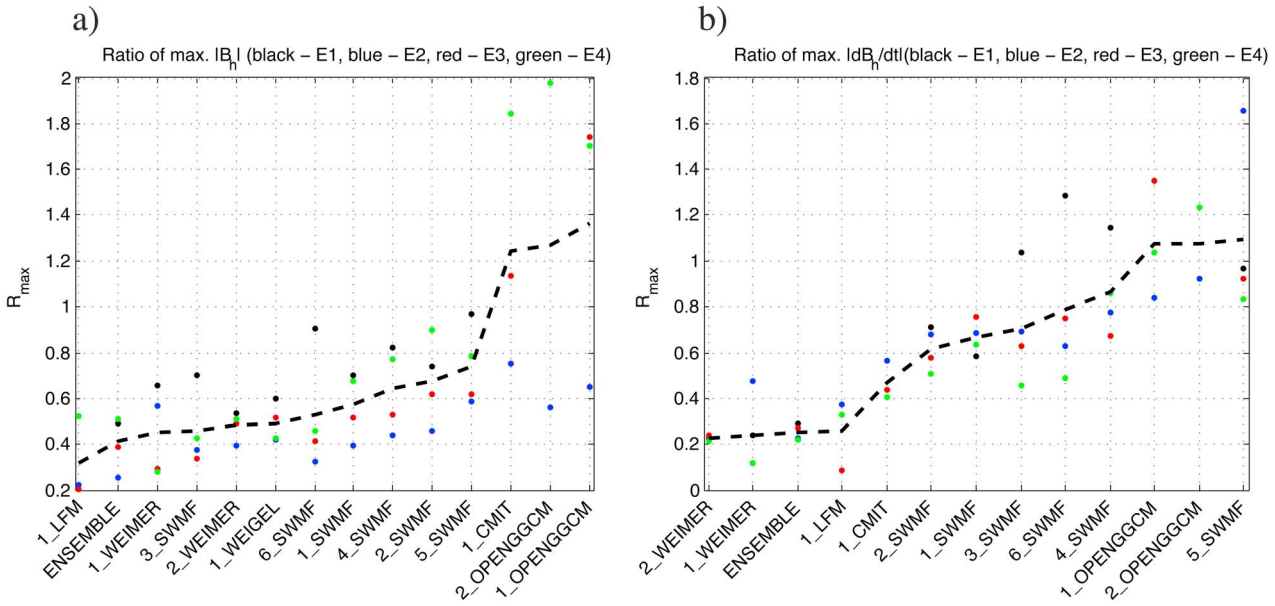

Figure 7. Average (average over stations) ratios of (a) the maximum horizontal magnetic field amplitudes and (b) the maximum amplitude of the time derivative of the horizontal magnetic field vectors. Dots with different colors correspond to different events: black, event 1; blue, event 2; red, event 3; green, event 4. The thick dashed line gives the model average taken over different events. See Table 3 for model identifiers on the horizontal axis. Identifier 1_WEIGEL is not included in the analysis of Figure 7b.

model evaluation for varying interplanetary conditions. It is conceivable that different modeling approaches may be preferable for different types of events. An "optimal model" could in fact be a collection of models tuned and used separately for predicting specific characteristics of a signal of interest for specific types of interplanetary conditions.

[40] Another question of general interest is the effect of model complexity, namely spatial resolution of the global MHD and coupling to inner magnetospheric models, on the model performance. The results in the work by *Pulkkinen et al.* [2010] indicated that inclusion of ring current physics and better spatial resolution in global MHD improved the model performance. Such improvement is not observable in the modest increase of the spatial resolution from 1_OPENGGCM to 2_OPENGGCM. While 1_CMIT predicts a more realistic maximum $|\mathbf{B}_h|$, inclusion of the thermospheric physics does not systematically improve the model performance from 1_LFM to 1_CMIT. However, it should be noted that the spatial resolution of the global MHD in 1_LFM is higher than in 1_CMIT, which can also affect the results.

[41] Although no systematic improvement can be observed in terms of the prediction efficiency or the forecast ratio for an increasing model complexity sequence 2_SWMF, 1_SWMF, 3_SWMF to 5_SWMF, a clear improvement is seen from Figures 4c and 7b in terms of the log-spectral distance and the ratio of the maximum amplitude of the time derivative of the horizontal magnetic field vector. Increasing the global MHD model spatial resolution and the inclusion of the ring current dynamics is seen to improve the models' capability to generate more realistic ground magnetic field fluctuations.

[42] From the space weather applications viewpoint, the work carried in this paper demonstrates that as results can depend heavily on the selected metric, one needs to be very careful in selecting appropriate metrics for model validation and verification. Users of the space weather products need to clearly define the characteristics of the modeled signal that are of their interest and the selection of the metric should be carried out accordingly. Further, users also need to specify what type of space weather events are of their interest. As was seen in the work at hand, the model performance can vary significantly as a function of the type of interplanetary driver and the strength of the geomagnetic disturbance.

[43] Systematic and quantitative evaluations of geospace model performances are needed to measure the progress in the field of space weather. Further, continuing evaluations can be used to guide the progress in the field, for example, by addressing the differences between various modeling approaches. From the applications viewpoint, users can utilize evaluations of collections of models to choose the approach that suits the best their specific needs. The work presented in this paper provides a benchmark for further and optimally ongoing monitoring of the space weather model performance.

[44] **Acknowledgments.** The authors wish to acknowledge the rest of the CCMC staff for their generous support throughout the work discussed in the paper. The results presented in this paper rely on data collected at geomagnetic observatories. We thank the national institutes that support them and INTERMAGNET for promoting high standards of magnetic observatory practice (www.intermagnet.org). Terry Onsager of NOAA is acknowledged for his help with selection of the ground magnetometer stations used in the study.

References

Birn, J., et al. (2001), Geospace Environmental Modeling (GEM) Magnetic Reconnection Challenge, *J. Geophys. Res.*, *106*(A3), 3715–3719.

Boteler, D. H., R. J. Pirjola, and H. Nevanlinna (1998), The effects of geomagnetic disturbances on electrical systems at the Earth's surface, *Adv. Space Res.*, *22*, 17–27.

De Zeeuw, D. L., S. Sazykin, R. A. Wolf, T. I. Gombosi, A. J. Ridley, and G. Tóth (2004), Coupling of a global MHD code and an inner magnetospheric model: Initial results, *J. Geophys. Res.*, *109*, A12219, doi:10.1029/2003JA010366.

Fuller-Rowell, T. J., D. Rees, S. Quegan, R. J. Moffett, M. V. Codrescu, and G. H. Millward (1996), A coupled thermosphere ionosphere model (CTIM), in *Handbook of Ionospheric Models*, edited by R. W. Schunk, STEP report, pp. 217–238, Utah State Univ., Logan, Utah.

Hapgood, M. A. (1992), Space physics coordinate transformations: A user guide, *Planet. Space Sci.*, *40*(5), 711–717.

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, *Space Weather*, *5*, S01005, doi:10.1029/2006SW000222.

Lyon, J. G., J. A. Fedder, and C. M. Mobarry (2004), The Lyon-Fedder-Mobarry (LFM) global MHD magnetospheric simulation code, *J. Atmos. Sol. Terr. Phys.*, *66*, 1333–1350.

Lyons, L. R. (1998), The Geospace Modeling Program Grand Challenge, *J. Geophys. Res.*, *103*(A7), 14,781–14,785.

Pirjola, R. (2004), Review on the calculation of surface electric and magnetic fields and of geomagnetically induced currents in ground-based technological systems, *Surv. Geophys.*, *23*, 71–90, doi:10.1023/A:1014816009303.

Pirjola, R. (2005), Effects of space weather on high-latitude ground systems, *Adv. Space Res.*, *36*, 2231–2240.

Powell, K. G., P. L. Roe, T. J. Linde, T. I. Gombosi, and D. L. De Zeeuw (1999), A solution-adaptive upwind scheme for ideal magnetohydrodynamics, *J. Comput. Phys.*, *154*(2), 284–309, doi:10.1006/jcph.1999.6299.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992), *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed., Cambridge Univ. Press, Cambridge, U. K.

Pulkkinen, A., L. Rastätter, M. Kuznetsova, M. Hesse, A. Ridley, J. Raeder, H. J. Singer, and A. Chulaki (2010), Systematic evaluation of ground and geostationary magnetic field predictions generated by global magnetohydrodynamic models, *J. Geophys. Res.*, *115*, A03206, doi:10.1029/2009JA014537.

Raeder, J., and N. Maynard (2001), Foreword, *J. Geophys. Res.*, *106*(A1), 345–348.

Raeder, J., Y. Wang, and T. Fuller-Rowell (2001), Geomagnetic storm simulation with a coupled magnetosphere-ionosphere-thermosphere model, in *Space Weather: Progress and Challenges in Research and Applications, Geophys. Monogr. Ser.*, vol. 125, edited by P. Song, H. J. Singer, and G. Siscoe, pp. 377–384, AGU, Washington, D. C.

Reay, S. J., W. Allen, O. Baillie, J. Bowe, E. Clarke, V. Lesur, and S. Macmillan (2005), Space weather effects on drilling accuracy in the North Sea, *Ann. Geophys.*, *23*, 3081–3088.

Richmond, A. D., E. C. Ridley, and R. G. Roble (1992), A thermosphere/ionosphere general circulation model with coupled electrodynamics, *Geophys. Res. Lett.*, *19*, 601–604.

Skoug, R. M., J. T. Gosling, J. T. Steinberg, D. J. McComas, C. W. Smith, N. F. Ness, Q. Hu, and L. F. Burlaga (2004), Extremely high speed solar wind: 29–30 October 2003, *J. Geophys. Res.*, *109*, A09102, doi:10.1029/2004JA010494.

Toffoletto, F. R., S. Sazykin, R. W. Spiro, R. A. Wolf, and J. G. Lyon (2004), RCM meets LFM: Initial results of one-way coupling, *J. Atmos. Sol. Terr. Phys.*, *66*, 1361–1370.

Tóth, G., et al. (2005), Space Weather Modeling Framework: A new tool for the space science community, *J. Geophys. Res.*, *110*, A12226, doi:10.1029/2005JA011126.

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

Weigel, R. S., A. J. Klimas, and D. Vassiliadis (2003), Solar wind coupling to and predictability of ground magnetic fields and their time derivatives, *J. Geophys. Res.*, *108*(A7), 1298, doi:10.1029/2002JA009627.

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*, S05002, doi:10.1029/2005SW000157.

Weimer, D. R. (2005), Predicting surface geomagnetic variations using ionospheric electrodynamic models, *J. Geophys. Res.*, *110*, A12307, doi:10.1029/2005JA011270.

Weimer, D. R., C. R. Clauer, M. J. Engebretson, T. L. Hansen, H. Gleisner, I. Mann, and K. Yumoto (2010), Statistical maps of geomagnetic perturbations as a function of the interplanetary magnetic field, *J. Geophys. Res.*, *115*, A10320, doi:10.1029/2010JA015540.

Wiltberger, M., W. Wang, A. G. Burns, S. C. Solomon, J. G. Lyon, and C. C. Goodrich (2004), Initial results from the coupled magnetosphere ionosphere thermosphere model: Magnetospheric and ionospheric responses, *J. Atmos. Sol. Terr. Phys.*, *66*, 1411–1423.

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

Yu, Y., A. J. Ridley, D. T. Welling, and G. Tóth (2010), Including gap region field-aligned currents and magnetospheric currents in the MHD calculation of ground-based magnetic field perturbations, *J. Geophys. Res.*, *115*, A08207, doi:10.1029/2009JA014869.

A. Chulaki, M. Hesse, M. Kuznetsova, and L. Rastätter, NASA Goddard Space Flight Center, Code 674, Greenbelt, MD 20771, USA.

G. Millward and H. J. Singer, Space Weather Prediction Center, NOAA, Boulder, CO 80305, USA.

A. Pulkkinen, Institute for Astrophysics and Computational Sciences, Catholic University of America, Washington, DC 20064, USA. (antti.a.pulkkinen@nasa.gov)

J. Raeder and A. Vapirev, Space Science Center, University of New Hampshire, Durham, NH 03824, USA.

A. Ridley, Department of Atmospheric, Oceanic, and Space Sciences, University of Michigan, Ann Arbor, MI 48109, USA.

R. S. Weigel, Department of Computational and Data Sciences, George Mason University, Fairfax, VA 22030, USA.

D. Weimer, Center for Space Science and Engineering Research, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA.

M. Wiltberger, High Altitude Observatory, National Center for Atmospheric Research, Boulder, CO 80307, USA.