SPACE WEATHER, VOL. 11, 187–205, doi:10.1002/swe.20036, 2013

Geospace environment modeling 2008–2009 challenge: D_{st} index

L. Rastätter,[1] M. M. Kuznetsova,[1] A. Glocer,[2] D. Welling,[3] X. Meng,[3] J. Raeder,[4] M. Wiltberger,[5] V. K. Jordanova,[6] Y. Yu,[6] S. Zaharia,[6] R. S. Weigel,[7] S. Sazykin,[8] R. Boynton,[9] H. Wei,[9] V. Eccles,[10] W. Horton,[11] M. L. Mays,[12] and J. Gannon[13]

Received 3 October 2012; revised 25 February 2013; accepted 27 February 2013; published 11 April 2013.

[1] This paper reports the metrics-based results of the D_{st} index part of the 2008–2009 GEM Metrics Challenge. The 2008–2009 GEM Metrics Challenge asked modelers to submit results for four geomagnetic storm events and five different types of observations that can be modeled by statistical, climatological or physics-based models of the magnetosphere-ionosphere system. We present the results of 30 model settings that were run at the Community Coordinated Modeling Center and at the institutions of various modelers for these events. To measure the performance of each of the models against the observations, we use comparisons of 1 hour averaged model data with the D_{st} index issued by the World Data Center for Geomagnetism, Kyoto, Japan, and direct comparison of 1 minute model data with the 1 minute D_{st} index calculated by the United States Geological Survey. The latter index can be used to calculate spectral variability of model outputs in comparison to the index. We find that model rankings vary widely by skill score used. None of the models consistently perform best for all events. We find that empirical models perform well in general. Magnetohydrodynamics-based models of the global magnetosphere with inner magnetosphere physics (ring current model) included and stand-alone ring current models with properly defined boundary conditions perform well and are able to match or surpass results from empirical models. Unlike in similar studies, the statistical models used in this study found their challenge in the weakest events rather than the strongest events.

Citation: Rastätter, L., et al. (2013), Geospace environment modeling 2008–2009 challenge: D_{st} index, *Space Weather*, *11*, 187–205, doi:10.1002/swe.20036.

1. Introduction

[2] As an increasing number of applications specify and predict space weather conditions, it becomes more important to quantitatively assess the performance of the underlying statistical and physics-based models. With quantifiable metrics, users of space weather modeling products will be able to better understand the strengths and weaknesses of each modeling approach and select the approach best suited for their application. In addition to serving the user, modelers gain insight into how different modeling parameters influence the performance of a given model and how different versions of a model are improving over time.

[3] A metrics challenge for state-of-the-art global magnetospheric space weather models has been discussed for years in the Geospace Environment Modeling (GEM) community. The GEM Global Geospace Circulation Modeling (GGCM) Metrics and Validation Focus Group organized a modeling challenge to focus on the dynamics of the inner magnetosphere and ground magnetic field perturbations. The 2008–2009 challenge was defined at the 2008 GEM workshop in Midway, Utah, and was broadly announced in September 2008. Model result submissions were accepted through the Community Coordinated Modeling Center

[1]Community-Coordinated Modeling Center, Code 674, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA.
[2]Code 673, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA.
[3]Department of Atmospheric, Oceanic and Space Sciences, College of Engineering, University of Michigan, Ann Arbor, Michigan, USA.
[4]Department of Physics, Institute for the Study of Earth, Oceans and Space, University of New Hampshire, Durham, New Hampshire, USA.
[5]National Center for Atmospheric Research, Boulder, Colorado, USA.
[6]Los Alamos National Laboratory, Los Alamos, New Mexico, USA.
[7]Department of Computational and Data Sciences, George Mason University, Fairfax, Virginia, USA.
[8]School of Physics, Astronomy, and Computational Sciences, Rice University, Houston, Texas, USA.
[9]ACSE, University of Sheffield, Sheffield, UK.
[10]Space Environment Corporation, Providence, Utah, USA.
[11]Institute for Fusion Studies, University of Texas, Austin, Texas, USA.
[12]Code 670, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA.
[13]United States Geological Survey, Golden, Colorado, USA.

Corresponding author: L. Rastätter, Space Weather Laboratory, Code 674, NASA Goddard Space Flight Center, Greenbelt, MD 20770, USA. (lutz.rastaetter@nasa.gov)

1542-7390/13/10.1002/swe.20036

Table 1. Event Numbers with Dates, Minimum D_{st}, and Maximum Kp Values

Event #	Date and UT time	Min(D_{st}) [nT]	max(Kp)
1	29 Oct 2003 06:00–30 Oct 2003 06:00	–353	9
2	14 Dec 2006 11:30–16 Dec 2006 00:00	–139	8
3	31 Aug 2001 00:00–01 Sep 2001 00:00	–40	4
4	31 Aug 2005 10:00–01 Sep 2005 12:00	–131	7

(CCMC) and submissions received through 31 March 2011 are included in this paper. Besides the online submission system, an online model comparison tool is available on the CCMC Web site to compare existing submissions to observations.

[4] This study is a collaborative effort by a large group of modelers and follows the first studies in this GEM 2008 series [*Pulkkinen et al.*, 2010; *Rastätter et al.*, 2011].

[5] The focus of this study is the widely used geomagnetic activity index D_{st}, which is derived from perturbations of the horizontal component B_H of the geomagnetic field at mid-latitude stations. In this study, we use the index from two sources: the hourly index as issued by the World Data Center (WDC) of Geomagnetism at the University of Kyoto, Japan and the 1 minute index now issued by the United States Geological Survey (USGS, *Gannon and Love* [2011]).

[6] This 2008–2009 Challenge follows a series of earlier GEM Challenges [*Lyons*, 1998; *Birn et al.* 2001; *Raeder and Maynard*, 2001], but extends the focus from ionospheric convection events and isolated substorms to geomagnetic storms and observations on the ground and in geosynchronous orbit. This study is based on four events that contain a large range of geomagnetic states including three storms of various strength and one interval with an isolated substorm. The primary goals of this challenge are to evaluate differences between the available modeling approaches, study effects of model couplings and uncover the influence of model resolution. This challenge is the first in a series of challenges that can be used by anyone to track the performance measures as models improve. The ongoing comparison of observations and models will also encourage collaboration between modelers and data analysts.

[7] The detailed analysis of the models' performance to calculate the D_{st} index is a first step in assessing the models' ability to track geomagnetic variations on the longer 1 hour time scale as well as on the short 1 minute time scale. In the future, we are planning to extend the analysis of models' ability to predict geomagnetic variations on the regional scale (auroral zone, sub-auroral zone and low latitudes) as is permitted by the design of the models.

[8] The scope of this paper is to report on the overall performance of the submitted model setups and improved techniques developed both by modelers and the CCMC to obtain D_{st} (and the magnetic perturbation at any station location) from model outputs. A central tool used in the analysis is the metrics evaluation tool that is available online at CCMC. All time series plots and skill scores in this paper have been generated using the online tool. This tool will continue to be upgraded as the variety of metrics challenges grow and the time scales and type of parameters analyzed change.

2. Setup of the Challenge

[9] Four geospace events were selected. Two events represent highly disturbed times: event 1 from 29 Oct 2003 6:00 to 30 Oct 2003 6:00 UT, known as part of the "Halloween storm" and event 2 from 14 Dec 2006 11:30 UT to 16 Dec 2006 0:00 UT, known as the "AGU storm". The other two events represent quieter times: event 3 from 31 Aug 2001 0:00 UT to 01 Sep 2001 0:00 UT and event 4 from 31 Aug 2005 9:30 UT to 01 Sep 2005 12:00 UT. All events with their start and end dates and times, minimum D_{st}, and maximum Kp index values are listed in Table 1.

[10] For each of the events, the solar wind magnetic field and plasma parameters obtained by the MAG and SWEPAM instruments on the Advanced Composition Explorer (ACE) satellite are shown in Figure 1. All events were covered by the ACE measurements except event 1, the "Halloween storm", for which plasma velocity data could be reconstructed only with low time resolution [*Skoug et al.*, 2004]. Plasma density data were constructed from the Plasma Wave Instrument on the Geotail satellite. Events 1 and 2 are large CME-related storms whereas events 3 and 4 are in less active periods.

[11] The D_{st} index measures general geomagnetic activity and the strength of the inner magnetospheric currents. Currents that create a magnetic perturbation on the ground are located in the ionosphere and the near-Earth magnetosphere and consist mainly of the ring current. When calculated from magnetospheric and ionospheric current systems, the D_{st} index is approximated by the magnetic perturbation at the center of the Earth. All current systems in the ionosphere and magnetosphere [*Yu et al.*, 2010] are equally important.

3. Models Used

[12] Models used for this challenge fall into four groups that reflect the different physics and numerical approaches taken by the models.

3.1. Three-Dimensional Magnetosphere Models

[13] These models are three-dimensional (3-D) magnetohydrodynamic (MHD) models of the magnetosphere that are coupled to an ionosphere electrodynamics solver. Models of this category are the Space Weather Modeling Framework (SWMF) [*Tóth et al.* 2005], the Open Geospace General Circulation Model (OpenGGCM) [*Raeder et al.*, 2001a], and the Coupled Magnetosphere-Ionosphere-Thermosphere (CMIT) model, also referred to by the magnetospheric Lyon-Fedder-Mobarry (LFM) component [*Lyon et al.*, 2004; *Wiltberger et al.*, 2004; *Merkin and Lyon*, 2010]. These models are run on clusters of computers and require large amounts of computing time and storage space for output

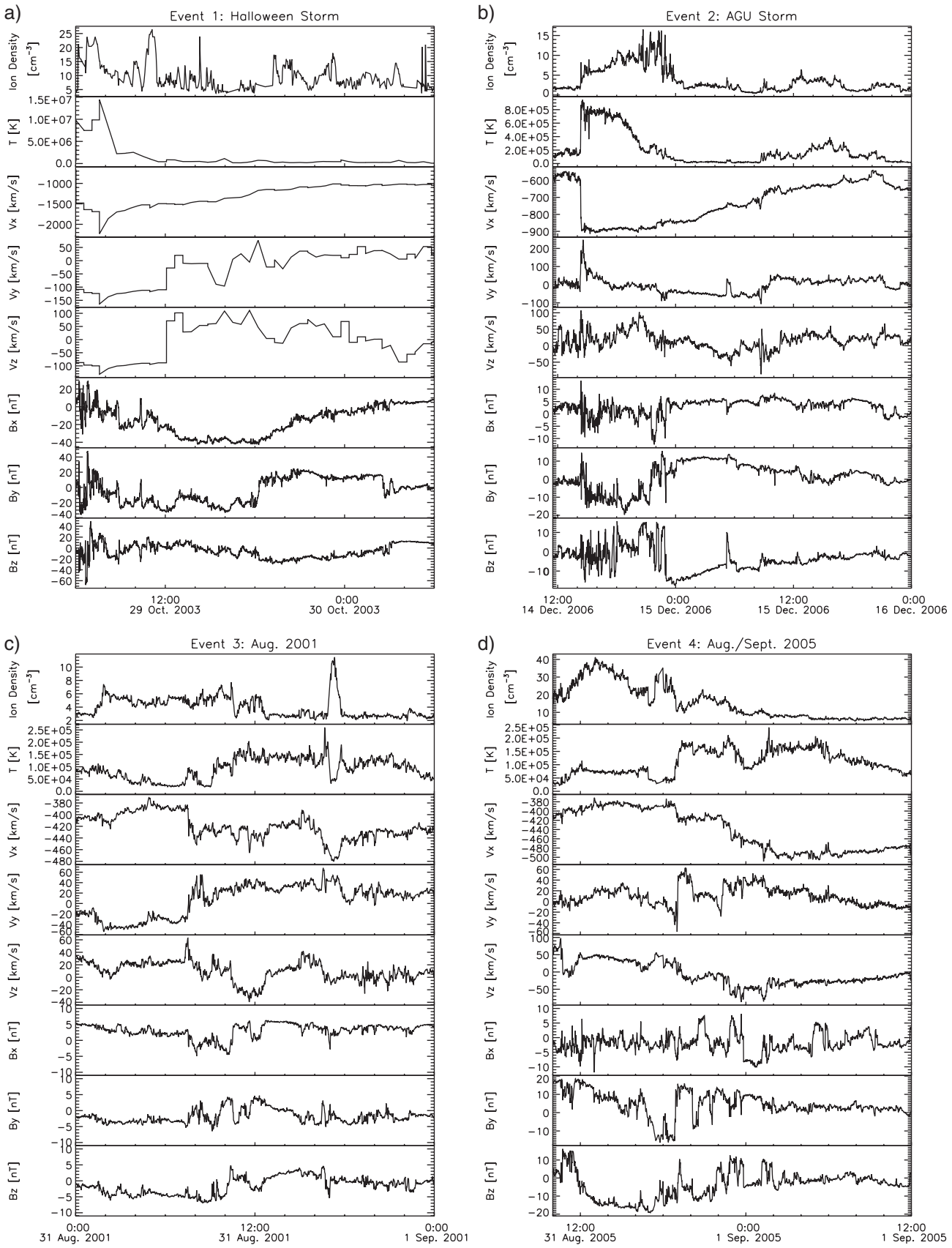

Figure 1. Solar wind bulk plasma and magnetic field observations for the four events listed in Table 1.

data. The parameter sets for these models are identical to the previously reported runs [*Rastätter et al.*, 2011] (the IDs are listed in Table 1). The D_{st} index is calculated from two contributions:

[14] ***(a) Magnetosphere currents:*** Electric currents from the magnetosphere outside of the "current pickup radius" of the magnetosphere model are used to calculate the magnetic perturbation $\delta\mathbf{B}$ by using the Biot-Savart formula.

$$\delta\mathbf{B} = \frac{\mu_0}{4\pi}\sum\frac{\mathbf{J}\times\mathbf{R}}{R^3}\mathrm{d}V \tag{1}$$

with **R** being the radius vector to a grid cell with current density **J** and volume dV. The "current pickup radius" is the radial distance from the Earth's center at which electric currents calculated by the magnetospheric model component are sampled to obtain magnetic field-aligned currents (FAC). For SWMF and OpenGGCM, this radius is $3R_E$, one or two grid cell layers away from the inner boundary (at 2.5 R_E). For LFM, the inner boundary of the magnetosphere grid (cell centers) is located at about $2.2R_E$ and currents that flow into the ionosphere are picked up from near that boundary. The calculation of currents in the LFM magnetosphere is done in a post-processing step using single precision magnetic field data stored in the LFM output. This results in significant errors near the inner boundary due to the strong dipole field gradient. As a consequence, we exclude currents calculated from within the first four grid layers, roughly 3.5 R_E. This is different from the calculations done with the OpenGGCM and SWMF models, which only show significant errors within the first grid layer adjacent to the inner boundary. Only volume elements (magnetosphere grid cells) dV that are centered at positions $\mathbf{R} = [x, y, z]$ beyond this "inner boundary" radius from the Earth's center are considered.

[15] ***(b) Ionospheric currents:*** Height-integrated currents in the ionosphere (assumed to be centered at an altitude of 110 kilometers [*Yu and Ridley*, 2008]) that close the field-aligned currents are considered using equation 1. For dV in equation 1, we here use 2-D surface elements instead of 3-D volume elements. We employ the currents in cartesian coordinates (J_x, J_y, J_z) reported by the ionospheric electrodynamic solvers on the respective mode's grid for a full Biot-Savart summation. Latitudinal resolution ranges between 0.5 degrees (OpenGGCM), 1.4 degrees (SWMF) and 2 degrees (LFM). Longitudinal resolution is 2 degrees for all models.

[16] ***(c) Field-aligned currents:*** In this study, the contribution of FAC to D_{st} is exactly zero due to the fact that D_{st} is being calculated as the north (axial) component of the magnetic perturbation at the Earth's center. Currents that are picked up from the magnetosphere and mapped as field-aligned currents along dipolar field lines are purely poloidal in SM spherical coordinates (r, θ, ϕ) since the dipole field has no ϕ component. Any contribution to the magnetic perturbation at the Earth's center is proportional to the cross product of two poloidal vectors (the position **r** and current element **J** that have components in the r and θ direction only) and thus is purely toroidal (i.e., in the ϕ direction in spherical coordinates). Since the polar axis is poloidal (radial direction for $\theta = 0$), the sum of toroidal contributions projected along the axis remains zero regardless of the spatial distribution of the FAC.

[17] Using the contributions from magnetospheric and ionospheric currents, D_{st} is approximated by the North-South component of the perturbation magnetic field δB_Z in SM coordinates at the Earth's center location. The magnetic perturbation is obtained by transforming contributions from the magnetosphere to SM coordinates and adding the contribution from the ionospheric currents that already comes in SM coordinates from each of the models.

[18] At CCMC, the magnetohydrodynamic (MHD) models of the global magnetosphere, SWMF, OpenGGCM and CMIT are run routinely for Runs-on-Request and thus were run for the four events.

3.1.1. SWMF

[19] SWMF is run here as a combination of the magnetosphere MHD component (SWMF/GM/BATSRUS) coupled to the Ridley Ionosphere Model (SWMF/IE/RIM) electrodynamic solver [*Tóth et al.* 2005]. Some runs include an inner magnetospheric component: the Rice Convection Model (SWMF/IM/RCM2) or the Comprehensive Ring Current Model (SWMF/M/CRCM). We performed four runs for each event with SWMF, spanning the setup for real-time simulation (755,000 grid cells) to higher-resolutions Run-on-Request grids (two and three million grid cells, respectively). One run was performed without a ring current model (4_SWMF), two others (5_SWMF, 8_SWMF) were run with the Rice Convection Model (RCM), and one setting (7_SWMF) was run with the Comprehensive Ring Current Model (CRCM), which was recently developed at NASA GSFC [*Buzulukova et al.*, 2010]. The CRCM model is coupled into SWMF in a similar manner as the RCM model *DeZeeuw et al.* [2004]. Both coupled ring current models require the same information (magnetic fields, plasma density and temperature) from the magnetosphere MHD model and return the same modifications (plasma pressure) back to the MHD model. SWMF runs with one million grid cells at 0.25 R_E resolution (8_SWMF) can be executed in real time on a cluster with 64 processors. Larger grids used in 4_SWMF, 5_SWMF, and 7_SWMF take proportionally longer. Finer resolution (i.e., 0.125 R_E used in 5_SWMF) takes twice as much time in explicit time stepping. Run 7_SWMF took 2.7 times longer than real time on 200 processors (or about eight times longer on 64 processors) using a serial version of CRCM. With a parallelized version of CRCM (A. Glocer, submitted paper), this run may run close to or faster than real time on 200 processors.

3.1.2. OpenGGCM

[20] The OpenGGCM magnetospheric MHD model was run with the Coupled Thermosphere-Ionosphere Model (CTIM) [*Fuller-Rowell et al.*, 1996] and the OpenGGCM model's ionospheric potential solver as described in *Raeder et al.* [2001b]. Runs performed with model version 3.1 have a fixed geomagnetic dipole orientation. Runs with model version 4.0 include an updating dipole orientation. OpenGGCM was run with a medium-resolution grid of 6.55 million cells with minimum cell size of 0.25 R_E for 2_OPENGGCM and 3.88 million cells with a minimum cell size of 0.25 R_E for 4_OPENGGCM. Runs of 2_OPENGGCM were two times slower than real-time on 64 processors whereas 4_OPENGGCM performed in real-time.

3.1.3. CMIT

[21] The Coupled Magnetosphere, Ionosphere and Thermosphere (CMIT) model [*Wiltberger et al.*, 2004] consists of the Lyon Fedder Mobarry (LFM) magnetosphere [*Lyon et al.*, 2004], the MIX ionosphere electrodynamics solver [*Merkin and Lyon*, 2010] and the TIE-GCM ionosphere-thermosphere model [*Richmond et al.*, 1992]. The CCMC employed standard settings of two model versions: LTR-2_1_1, available in 2011, for run 2_LFM-MIX with the LFM and MIX components, and LTR-2.1.5, issued in 2012, for run 2_CMIT that also includes the TIE-GCM ionosphere model to specify the ionospheric conductances. The LFM grid with 53 × 48 × 64 cells has radial spacing of about $0.4R_E$ in the dayside magnetosphere (within $10R_E$). CMIT runs execute in real-time on 24 processors.

3.2. Low-Dimensional Models of the Magnetosphere-Ionosphere

[22] This category includes a single model WINDMI.

3.2.1. WINDMI

[23] In WINDMI, the magnetosphere, ring current and ionosphere system is represented by a low-dimensional computational system using energy fluxes computed from empirically determined coupling parameters [*Horton and Doxas*, 1998; *Mays et al.*, 2009]. D_{st} is one of the model's outputs and is obtained from the ring current energy using the Dessler-Parker-Sckopke relation (see equation 2, which is discussed in the ring current model section below). The WINDMI model solves a set of ordinary differential equations with parameters that can be physically motivated. For this reason, WINDMI has been grouped with the 3-D magnetosphere MHD models in plots and listings of skill scores in the remainder of this paper. The model was run with a nominal parameter set as listed in Table 1 in *Mays et al*. [2009] and three different solar wind coupling functions. WINDMI runs very fast and takes less than a minute on a single processor to produce a day's worth of D_{st} at a 1 minute cadence.

3.3. Kinetic Ring Current Models

[24] Ring current models are run either in stand-alone mode or coupled to one of the magnetospheric models mentioned above. Models included here are the Ring Current-Atmosphere Interactions Model with Self-Consistent Magnetic Field (RAM-SCB) and the Rice Convection Model (RCM).

[25] The total energy in the ring current is converted to D^*_{st} using the Dessler-Parker-Sckopke relation [*Dessler and Parker*, 1959; *Sckopke*, 1966] in equation 2. D^*_{st}, in turn, is converted to D_{st} with the correction of magnetopause current using the dynamic pressure of the solar wind in equation 3:

$$D^*_{st} = -3.98 \cdot 10^{-30} E_{RC} \quad (2)$$

$$D^*_{st} = \frac{D_{st}}{1.5} + 0.2\sqrt{P_{dyn,sw}} - 20\,. \quad (3)$$

The ring current energy E_{RC} is measured in kilo-electron-volt (1 keV = 1.602×10^{-16}J) and solar wind dynamic pressure $P_{dyn,sw} = \rho V_x^2$ is measured in eVm^{-3}.

3.3.1. RAM-SCB

[26] Runs with the Ring-current Atmosphere interaction Model with Self-Consistent 3D magnetic (B) field (RAM-SCB) [*Jordanova et al.*, 1994, 2010; *Cheng*, 1995; *Zaharia et al.*, 2004, 2006; *Yu et al.*, 2011] were performed in stand-alone mode using two models for the equatorial electric field (*Kp*-dependent Volland-Stern, *Volland* [1973]; *Stern* [1975]; *Burke* [2007], and the interplanetary plasma and magnetic field dependent Weimer 2000 [*Weimer*, 2001]). The plasma boundary conditions were specified after geosynchronous LANL satellite data, while the Tsyganenko-89 [*Tsyganenko*, 1989] and the SWMF models were used to specify the magnetic field conditions.

[27] Two runs of RAM-SCB (4_RAMSCB and 5_RAM-SCB) were coupled to SWMF. Plasma boundary conditions and magnetic fields are provided by the BATSRUS magnetosphere model, and the electric field is provided by the RIM ionosphere electrodynamics model. For 5_RAMSCB, SWMF was also run with the Polar Wind Output Model that describes the mass loading of the magnetosphere from the ionosphere [*Welling et al.*, 2011]. The various runs with RAM-SCB have been described in detail in comprehensive validation studies by *Welling et al.* [2011] and *Yu et al.* [2011]

[28] In each RAM-SCB model run, the self-consistent magnetic field determination (SCB) takes most of the computer time. Without SCB (run 1_RAMSCB) the stand-alone RAM-SCB model can run up to 18 times faster than real-time. With SCB (2_RAMSCB, 3_RAMSCB), the model runs between one and five times faster than real time. Time discretization that is mandated by the dynamics of an event affects total run time. Larger storms require smaller time steps and slow down computations considerably. The two runs coupled with SWMF (4_RAMSCB, 5_RAMSCB) require a cluster with 200 processors and run between two and four times slower than real time. Due to the computational expense and the cutting-edge development required, 4_RAMSCB and 5_RAMSCB were not run for all events. Results were submitted for illustration and benefit of the GEM community.

3.3.2. Rice Convection Model

[29] The Rice Convection Model (RCM) is a well-established and extensively used model of the plasma electrodynamics in the inner magnetosphere and its coupling to the ionosphere [*Harel et al.*, 1981; *Wolf et al.*, 1991; *Sazykin*, 2000; *Toffoletto et al.*, 2003]. The model describes the $\mathbf{E}\times\mathbf{B}$ and gradient-curvature drifts of an isotropic plasma on closed magnetic field lines in parts of the inner magnetosphere [*Wolf*, 1983]. Field-aligned currents are calculated from the magnetospheric plasma pressure gradients using the Vasyliunas equation [*Vasyliunas*, 1970]. The FAC determine the ionospheric potential and electric fields that are mapped back into the magnetosphere to close the computational loop. The RCM is driven by input functions that include the magnetic field, the electric potential distribution on the high-latitude boundary, the ionospheric conductance, and the influx of particles across the high-latitude boundary.

[30] The RCM was run with the Hilmer and Voigt magnetic field description [*Hilmer and Voigt*, 1995] and with either the Siscoe-Hill [*Siscoe* 1982; *Hill*, 1984; *Burke et al.* 2007] or the Weimer [*Weimer*, 2005] electric field models. Plasma sheet boundary conditions were specified by the Tsyganenko and Mukai [*Tsyganenko and Mukai*, 2003] model, by plasma transport conditions specified by *Borovsky et al.* [1998], or by the Magnetosphere Specification Model

(MSM) [*Freeman et al.*, 1994; *Tascione et al.*, 1988]. The stand-alone RCM model runs in real time on a single-processor workstation.

3.4. D_{st}-Specification Models

[31] Models in this class include the Impulse Response Function with 96 lags (IRF96), an analytic formula after Burton, Feldstein and Murayama (BFM), the University of Sheffield (UOS) NARMAX-RJB (NARMAX), the RiceDST model, and the RDST real-time specification of D_{st}. D_{st} is derived directly through an analytic or iterative formula or a neural-network based algorithm without modeling the intrinsic energy flow through the magnetosphere-ionosphere system. All specifications except RDST use recent and current solar wind conditions to obtain a value for D_{st}. RDST uses data from four magnetometers similar to the Kyoto D_{st} determination. RiceDST and NARMAX estimate real-time D_{st} from values of D_{st} obtained in previous iterations of the model in addition to current solar wind conditions. All these models run very fast (a 24 hour period is modeled within a few minutes on a single-processor workstation) and generate D_{st} as their only output.

3.4.1. IRF96

[32] The IRF96 forecast model is an impulse response function (IRF) model with 96 coefficients. The coefficients, h, were derived by creating an over-determined matrix using

$$D_{st}(t) = h_{\Delta} + \sum_{t'=-N_a}^{N_c} vB_s(t-t')h(t'), \tag{4}$$

based on historical 1 hour KYOTO D_{st} values and v_Bs taken from the OMNI2 data set in the time range of 1963 through 2006. Further details on the procedure are detailed in *Weigel* [2010]. Any time intervals (t) with a gap in 1 hour measurements of $D_{st}(t)$ and $vB_s(t)$, $vB_s(t-1)$, ..., $vB_s(t-95)$ were omitted from the matrix. The resulting matrix has approximately 105,000 rows. Any measurements from time intervals in the test storms were omitted so that the predictions were out-of-sample and N_a was set to zero to enable a forecast.

3.4.2. BFM

[33] The analytic representation after Burton, Feldstein and Murayama [*Burton et al.* 1975; *Murayama* 1982; *Feldstein* 1992] was implemented at CCMC and uses the solar wind condition at a 1 minute resolution throughout the intervals studied. The description uses the electric field $E_y = V_x B_z$ $\left(\text{in } \frac{\text{mV}}{\text{m}}\right)$ and the dynamic pressure of the solar wind $P_{dyn} = \rho V_x^2$ (with ρ being the solar wind plasma density and V_x the solar wind bulk speed in the x direction) to compute the coupling function (with $d = 0.0015$ nT/(mVs/m)):

$$F(E_y) = \begin{cases} d(E_y - 0.5) & \text{for } E_y > 0.5 \\ 0 & \text{for } E_y \le 0.5 \end{cases} \tag{5}$$

and a ring current decay time

$$\tau = \begin{cases} 7.7h & \text{for } E_y < 4 \\ 3h & \text{for } E_y \ge 4\,. \end{cases} \tag{6}$$

The coupling function and decay rate then yield the strength of the ring current D_{st}^* via:

$$\frac{\mathrm{d}}{\mathrm{d}t} D_{st}^* = F(E_y) - \frac{D_{st}^*}{\tau}. \tag{7}$$

The iterative method starts from an initial D_{st}^* value that may be obtained as a quiet time driver multiplied by the inverse of the decay rate $D_{st,0}^* = F(E_y(t_0))/\tau)$ at a time t_0 several hours before the time interval of interest. The numerical formulation then uses solar wind data available every minute. D_{st} is computed from D_{st}^* using equation 3.

3.4.3. NARMAX

[34] The NARMAX algorithm [*Billings et al.*, 1989] is an advanced system identification technique, similar to neural networks. A NARMAX model is able to represent a wide class of linear and non-linear systems with physically interpretable parameters [*Leontaritis and Billings*, 1985a, 1985b]. The output $y(t)$ of the model at time t is a polynomial function F of the previous values of inputs u, outputs y, and error terms e as described by equation 8.

$$\begin{aligned} y(t) = F\big[&y(t-1), \ldots, y(t-n_y), \\ &u_1(t-1), \ldots, u_1\left(t-n_{u_1}\right), \ldots, \\ &u_m(t-1), \ldots, u_m\left(t-n_{u_m}\right), \\ &e(t-1), \ldots, e(t-n_e)\big] + e(t)\,. \end{aligned} \tag{8}$$

Index m is the number of inputs to the system and n_y, $n_{u_1}, \ldots, n_{u_m}$ are the maximum time lags of the output and the m inputs, respectively. The NARMAX algorithm was first applied to magnetosphere predictions by [*Boaghe et al.* 2001]. Here, a model of the D_{st} index was derived using vB_s as the input. This model was shown to have a high correlation and coherency with the measured D_{st} index. For the D_{st} NARMAX models, the function F is a quadratic polynomial, in which the monomials comprise all the possible quadratic cross-coupled combinations of past inputs, outputs and error terms. Here, the output is the D_{st} index, and the inputs are the solar wind parameters. The NARMAX algorithm consists of three stages: model structure selection, parameter estimation and model validation. The first stage is aimed at reducing the large number of possible monomials by determining the most significant monomials using the Error Reduction Ratio (ERR) [*Billings et al.*, 1989]. The monomials with a small ERR are deemed negligible, while the monomials with a high ERR are carried on to the second stage, the parameter estimation, where the coefficients for each of these monomials are calculated. During the last stage, the model is validated by exploiting both dynamic and statistical approaches [*Billings and Voon*, 1986]. More recently, *Boynton et al.* [2011a] also employed the NARMAX algorithm to deduce a model for the D_{st} index using a different coupling function ($p^{1/2}v^{4/3}B_T \sin\theta/2$) as an input. This coupling function was shown to be the best coupling function for the D_{st} index using an ERR analysis [*Boynton et al.*, 2011b]. The model by *Boynton et al.* [2011a] was shown to have a higher correlation and coherency with the measured D_{st} than that of *Boaghe et al.* [2001]. A startup of 50 hours is sufficient to allow the system to reach a state

Table 2. Model Run Settings Used in the Challenge

Model Description	Identifier
SWMF v8.01, BATSRUS, 3M cells, min. res. 0.125 R_E (CCMC)	4_SWMF
SWMF v8.01, BATSRUS with RCM, 3M cells, min. res. 0.125 R_E (CCMC)	5_SWMF
SWMF v20110215, BATSRUS with CRCM, 1.78M cells, min res. 0.125 R_E (UMich.)	7_SWMF
SWMF v20110111, BATSRUS with RCM, 1M cells, min res. 0.25 R_E ("real time", UMich.)	8_SWMF
OpenGGCM v3.1 with CTIM, 6.55M cells, min. res. 0.25 R_E (CCMC)	2_OPENGGCM
OpenGGCM v4.0 with CTIM, 3.88M cells, min. res. 0.25 R_E ("real time", CCMC)	4_OPENGGCM
CMIT_2-1-5, LFM with 53 × 48 × 64 cells, min. res. 0.4 R_e radial, MIX, TIEGCM (CCMC)	2_CMIT
LFM-MIX_2-1-1, LFM with 53 × 48 × 64 cells, min. res. 0.4 R_e radial, MIX ("real time", CCMC)	2_LFM-MIX
WINDMI 1.0 with nominal parameters, rectified solar wind driver (CCMC)	1_WINDMI
WINDMI 1.0 with nominal parameters, Siscoe solar wind driver (CCMC)	2_WINDMI
WINDMI 1.0 with nominal parameters, Newell solar wind driver (CCMC)	3_WINDMI
RAM-SCB, RAM-SCB, driven by LANL MPA/SOPA Volland-Stern E-field, dipole B-field	1_RAMSCB
RAM-SCB, RAM-SCB, driven by LANL MPA/SOPA, Weimer-2K E-field, dipole B-field	2_RAMSCB
RAM-SCB, RAM-SCB, driven by LANL MPA/SOPA, Weimer-2K E-field, T89 B-field	3_RAMSCB
RAM-SCB, SWMF/RAM-SCB, driven by BATSRUS, Ridley Ionosphere Model (RIM)	4_RAMSCB
RAM-SCB, SWMF/RAM-SCB, driven by multi-species BATSRUS, RIM, Polar Wind	5_RAMSCB
RCM with Siscoe-Hill potential drop, Tsyganenko-Mukai bc, Hilmer & Voigt (1995) B-field	1_RCM
RCM with Siscoe-Hill potential drop, Borovsky 1998 bc, Hilmer & Voigt (1995) B-field	2_RCM
RCM with Siscoe-Hill potential drop, MSM bc, Hilmer & Voigt (1995) B-field	3_RCM
RCM with Weimer 2005 potential drop, Borovsky 1998 bc, Hilmer &Voigt (1995) B-field	4_RCM
IRF, Impulse Response Function with 96 lags (version as of 04 June 2010) (GMU)	1_IRF96
Analytic formula after Burton (1975), Feldstein (1992) and Murayama (1982) (CCMC)	1_BFM
UoS NARMAX using previous estimated D_{st} and 1 hour OMNI solar wind	1_NARMAX
UoS NARMAX including Ring Current effects, inputs as in 1_NARMAX	2_NARMAX
Rice D_{st} neural network using Boyle (1997) solar wind driver and dynamic pressure	1_RiceDST
Real-time D_{st} derivation (RDst version 2.1), Space Environment Corp.	1_RDST

that is independent of the initial D_{st} values that have to be set arbitrarily (usually zero).

3.4.4. RiceDST

[35] The RiceDST model is a neural-network-based time prediction model. The model is driven by input time histories (10 hours) of solar wind coupling function described by the Boyle function [*Boyle et al.*, 1997]. an empirical approximation that estimates the Earth's polar cap potential, and the solar wind dynamic pressure to predict the D_{st} index approximately 1 hour ahead. The model can be run in real-time through inputs from an upstream solar wind monitor such as ACE [*Bala and Reiff*, 2012] to specify *Kp*, D_{st} and *AE* indices. As of 2012, the model is operational and the real-time estimates of the indices can be obtained from http://space.rice.edu/ISTP/wind.html.

3.4.5. RDST

[36] Space Environment Corporation developed a real-time D_{st} estimator with robustness of the calculation in mind given that real time data streams are not assured. The RDST_CALC program produces the real-time D_{st} (RDST) value as the best possible estimate of D_{st} whether there are four, three, two or one magnetic observatories available for the analysis. The RDST_CALC program is currently deployed at Air Force Weather Agency (AWFA). The first difference of the RDST calculation from the traditional D_{st} determination is that RDST is based on Hermanus (HER), Honolulu (HON), San Juan (SJG), and Guam (GUA) magnetic observatories. The Guam Magnetic Observatory replaces the Kyoto Magnetic Observatory for the Pacific sector to assure a robust data stream to AFWA. The RDST algorithm also has differences from the traditional definitive calculation of D_{st} to strive for the best estimate of the definitive D_{st} with uncertain real-time data streams. Each station is used to make a best estimate of the definitive D_{st}, then the available estimates of each station are averaged to produce the RDST value. Each single-station D_{st} estimate uses similar reduction algorithms to the traditional D_{st} analysis. The first step is to provide a stable estimate of the Secular Variation (SV) within the horizontal component for each station, s, $\left(H_s^{SV}\right)$. The second step is to obtain the current Solar Quiet (SQ) variation of the horizontal component with the secular variation removed $\left(H_s^{SQ}\right)$. Removing these two components leaves the station's ring current deflection, ΔH_s.

$$\Delta H_S = H_S - H_S^{SQ} - H_S^{SV}. \tag{9}$$

The standard latitudinal dependence is applied to the ΔH_s value. There is an expected Universal Time dependence in the response of a single station to the magnetospheric currents. The single station estimate of D_{st} is obtained using a linear regression analysis for each UT hour of ΔH_s and the historic definitive D_{st}. Reduction is adjusted with a set of linear coefficients, which have a Universal Time dependence:

$$\mathrm{RDST}_i = a_i(t)\frac{\Delta H}{\cos\lambda} + b_i(t) \tag{10}$$

where λ is the magnetic latitude of Earth's tilted dipole. Finally, the four, three, or two single-station estimates of D_{st} are averaged to obtain a robust real time estimate of the definitive D_{st}. RDST values are obtained at the standard 1 hour cadence.

[37] The above four groups of models all use the solar wind data (RDST: magnetometer data) as input. None of the models use the observed D_{st} as input. The detailed model parameters are listed in Table 2.

4. The D_{st} Index

[38] Analyses and skill score calculations were performed using two implementations of the D_{st} index that measures

the strength of the ring current and is used generally as a proxy for the intensity of geomagnetic activity.

4.1. The Kyoto D_{st} Index

[39] The D_{st} index is an averaged north-south perturbation of the geomagnetic field obtained by using observations at four stations ($N = 4$, $n = 1, \ldots, N$) located at magnetic mid-latitudes: Kakioka in Japan (geographic longitude = 140.18, latitude = 36.23), Honolulu in Hawaii (lon. = 201.98, lat. = 21.32), San Juan in Puerto Rico (lon. = 293.88, lat. = 18.11) and Hermanus in South Africa (lon. = 19.22, lat. = –34.40). The D_{st} index is defined as a 1 hour average of magnetic disturbances (H) measured at the four stations, weighed by the cosine of the respective magnetic latitude at each station [*Sugiura*, 1964; *Love and Gannon*, 2009]:

$$D_{st} = \frac{\sum_{n=1}^{N} H_n}{\sum_{n=1}^{N} \cos(MLAT_n)}. \quad (11)$$

For the comparisons, we use the final index, which is published by the World Data Center for Geomagnetism in Kyoto, Japan, between 3 and 6 years after the magnetometer observations have been gathered (final data are currently available through the end of 2008). The 1 hour averaged index has been produced since 1957. The index has been modified and improved over the years (e.g., *Sugiura and Hendricks* [1967]; *Sugiura and Kamei* [1991]; *Karinen and Mursula* [2006]). The disturbance H is obtained by subtracting the diurnal variation of the horizontal magnetic field from an average of several of the quietest days around the time of the observation. This method, however, requires a substantial amount of data before and after a measurement to determine the quiet-time baseline.

4.2. The USGS D_{st} Index

[40] The United States Geological Survey (USGS) has recently developed a system to determine D_{st} on a 1 minute time scale [*Gannon et al.*, 2011; *Gannon and Love*, 2011]. The derivation of the baseline involves frequency domain analysis and can be performed in near-real-time as opposed to the Kyoto D_{st} index that uses long-time series to determine the baseline. Thus nearly definitive D_{st} values are available almost immediately using 1 year of data to determine the *Sq*-harmonics. The real-time USGS D_{st} is very similar to the definitive USGS D_{st} (99.5% correlation). We use definitive 1 minute USGS D_{st} data for the events studied, with baseline removal based on the analysis of 23 years of nearly continuous magnetic observatory data. Since the USGS-D_{st} is a 1 minute index, it allows us to better study the faster time evolution of the magnetic perturbation in the inner magnetosphere. The same time cadence can be obtained by physics-based models of the magnetosphere and ring current and the BFM D_{st} specification model. It is important to note that USGS D_{st} index is derived using a different normalization compared to the Kyoto D_{st} to correct for the four stations' magnetic latitudes:

$$D_{st} = \frac{1}{N} \sum_{n=1}^{N} \frac{H_n}{\cos(MLAT_n)}. \quad (12)$$

Love and Gannon [2009] compared Kyoto and USGS D_{st} values for years from 1957 through 2007. The comparison showed that Kyoto D_{st} values are on average 8.60 nT lower than the USGS D_{st} values. The root-mean square difference between the indices is 11.01 nT. Several strong storm events were found to have substantial difference in terms of minimum D_{st} values. While we do see a difference in the baseline between the two implementations of the D_{st} index, the D_{st} minima are very similar in the storms used in this study.

5. Types of Skill Scores

[41] We employed the Prediction Efficiency (*PE*) and the Log-Spectral Distance (M_s) as used in *Rastätter et al.* [2011]. In addition, we add the Correlation Coefficient (*CC*), Model Yield (*YI*) and the Timing Error (ΔT) to the analysis, and we also chart the models' performance using a combination of two scores at a time to assess the performance of classes of model runs. The skill scores used in this paper are described below:

5.1. Prediction Efficiency *PE*

[42] The D_{st} values provided by the models introduced in the previous sections can be evaluated by computing a Prediction Efficiency (*PE*), defined for a discrete time series as follows:

$$PE = 1 - \frac{< (x_{mod} - x_{obs})^2 >}{\sigma_{obs}^2} \quad (13)$$

with $< \ldots >$ denoting the arithmetic mean, x_{obs} the observation, x_{mod} the modeled signal, and $\sigma_{obs}^2 =< x_{obs}^2 >$ the variance of the observed signal. The numerator is often referred to as the Mean Squared Error (MSE). $PE = 1$ indicates perfect model performance and $PE = 0$ indicates performance comparable to predicting the arithmetic mean of the observed signal. *PE* can reach unlimited negative values.

5.2. Log-Spectral Distance M_s

[43] The spectral power determines the level of disturbance on different time scales that can be produced by the models in comparison to the observed level of fluctuations. The analysis of the Log-Spectral Distance evaluates the spectral distribution of fluctuations in a given spectral range. This is accomplished by computing a single number that measures the distance between the observed spectral distribution from that obtained by a model. The comparison of spectral distributions in the model outputs compared to the observations (and solar wind inputs) indicates how well a model preserves activity levels in various frequency ranges. A model would perform perfectly if the spectral distribution in the observations matched the modeled spectrum. To compute the Log-Spectral Distance, the logarithm of the ratio of the spectral power of the observed ($(|\tilde{x}_{obs}|)$) and modeled variable ($|\tilde{x}_{mod}|$)

$$m_s = \log\left[\frac{|\tilde{x}_{mod}|}{|\tilde{x}_{obs}|}\right] \quad (14)$$

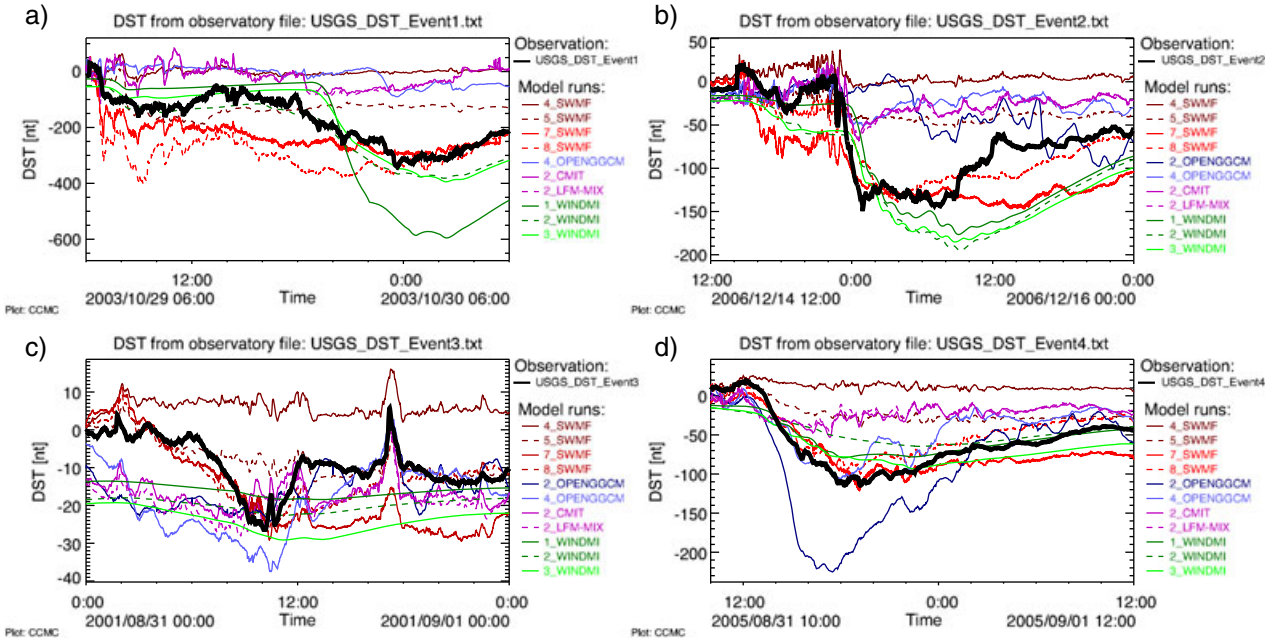

Figure 2. One minute USGS D_{st} data and magnetosphere model results. Magnetosphere models listed in the first section in Table 2 are shown in color: SWMF in brown and red, OpenGGCM in blue, CMIT/LFM in purple and WINDMI in green colored traces. The panels show the individual events: (a) event 1 (29–30 Oct 2003), (b) event 2 (14–16 Dec 2006), (c) event 3 (31 Aug–1 Sep 2001) and (d) event 4 (31 Aug–1 Sep 2005).

is calculated for each frequency. The root mean square of m_s over the N frequencies f yields the Log-Spectral Distance M_s:

$$M_s = \sqrt{\frac{1}{N}\sum_f m_s^2}. \quad (15)$$

The score M_s is equal or larger than zero. $M_s = 0$ is a perfect score. To perform the spectral analysis, 2 hour length windows are selected from the 1 minute data and model results, yielding a Fourier spectrum for periods between 2 minutes (f_1 = 1/120 *hertz*) and 120 minutes (f_2 = 1/7200 hertz). A 75% overlap between adjacent windows is allowed. Spectra from all valid windows (those that have no missing data) are averaged to form the spectra from the observation data ($\tilde{x}_{obs}$) and model outputs ($\tilde{x}_{mod}$).

[44] The computation of spectra does not make sense for the Kyoto D_{st}, which is defined on an hourly basis. We can, however, use the spectral analysis for the USGS D_{st} values that are available on a 1 minute scale as long as the model output time resolution is comparable.

5.3. Correlation Coefficient

[45] The Correlation Coefficient (*CC*) is the cross-correlation computed with zero lag:

$$CC = \frac{< x_{obs} x_{mod} >}{\sigma_{obs}\sigma_{mod}}. \quad (16)$$

A *CC* of 1 is perfect correlation and –1 is perfect anti-correlation (model values with same shape but with opposite sign compared to observations). Both *PE* and *CC* require the observation signal in the denominator to have nonzero variances $(\sigma^2_{obs} > 0)$. To obtain a finite *CC*, the modeled signal also needs to have a non-zero variance $(\sigma^2_{mod} > 0)$. Quiet events with low variance in the observation signal may amplify modeling errors in these scores.

5.4. Modeling Yield

[46] The Modeling Yield *YI* compares the largest change seen in the modeled index value versus the observed index. This skill score is suitable for a quantity that starts from a constant baseline (e.g., zero) and for events that are characterized by a single peak or minimum.

$$YI = \frac{\max(x_{mod}) - \min(x_{mod})}{\max(x_{obs}) - \min(x_{obs})}. \quad (17)$$

A modeling Yield of 1 is an ideal score. Yields near 1 may indicate a good model performance, but only if they are accompanied by good values of the Correlation Coefficient *CC*.

5.5. Timing Error

[47] The Timing Error ΔT applies to events where a time series shows distinct extreme values (minima in the case of D_{st}). ΔT is the time difference between the time of the observed minimum and the time of the modeled minimum of D_{st}. A ΔT of zero hours is ideal.

[48] Many geomagnetic storms show a single minimum followed by a monotonic recovery phase. Often, however, geomagnetic storms feature multiple onsets of activity.

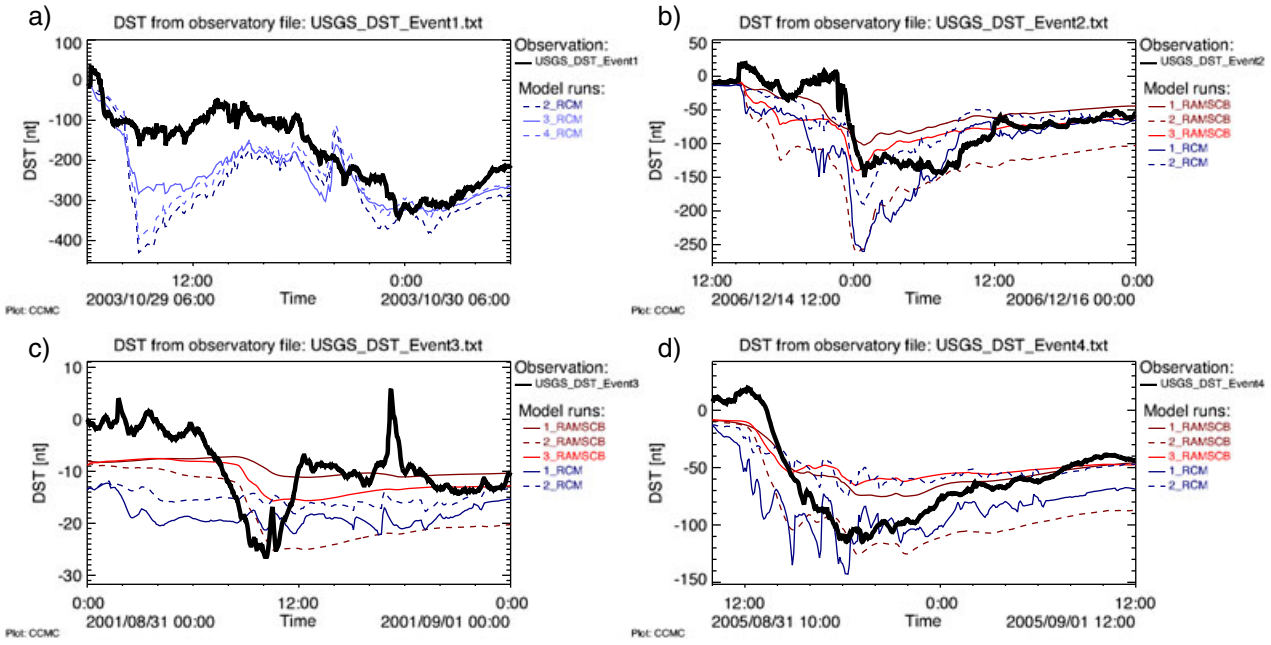

Figure 3. One minute USGS D_{st} data (black line) and ring current model results (colors). In the same format as Figure 2, this figure shows the results of ring current models (listed in the middle section in Table 2): RCM runs are shown in blue and RAM-SCB in brown and red colors. The vertical scales of the panels are not the same as that in the respective panels in Figure 2. Note that only within each figure (Figures 2–4, for each class of models) the color and line style combination is unique for each model run.

In fact, among the events of this study, we find two distinct minima in event 1 and a long period of near-minimum D_{st} values during event 2 that can be considered as having two minima as well. In the analysis, we will consider two parts of the time series for events 1 and 2 to better analyze model performance with respect to ΔT.

6. Results

[49] Using USGS data, *CC*, *PE* and M_s were derived by interpolating model outputs written on longer cadences to the 1 minute cadence of the USGS data. We calculate the Timing Error and the Yield using 1 hour Kyoto D_{st} observations and also calculate *CC* and *PE* with Kyoto data. Model results with a smaller time resolution than 1 hour were averaged to the 1 hour time cadence of the Kyoto D_{st} to perform these comparisons.

[50] Global magnetospheric models were run to provide D_{st} as a snapshot of the effects of the global current system at a 1 minute cadence (7_SWMF and 8_SWMF generate disturbances at a 5 second cadence that were subsampled at 1 minute intervals). Magnetosphere MHD models other than 7_SWMF and 8_SWMF use a post-processing step developed at CCMC to compute D_{st} from magnetospheric and ionospheric currents contained in model outputs written every minute. Ring Current models (RCM, RAM-SCB) provide outputs every 5 minutes. The D_{st} specification models output 1 hour D_{st} values, except for the BFM formula. BFM produces output at the data rate (ACE-L2, interpolated to a 1 minute cadence) that is sufficient for comparison with USGS 1 minute D_{st} data to obtain spectral information.

[51] Figures 2, 3 and 4 show the time series data used for the study. Figure 2 shows the USGS D_{st} data values and results obtained from the magnetosphere models (SWMF, OpenGGCM, CMIT, LFM-MIX) and WINDMI. Figure 3 shows the USGS D_{st} data values and results obtained from the ring current models RCM and RAM-SCB. Figure 4 shows the Kyoto D_{st} data values and outputs from the specification models (IRF, BFM, NARMAX, RiceDST, RDST). To analyze the Timing Error, two events that had two minima during the time period under consideration were split in two: Event 1 had its first D_{st} minimum at 10:00 UT on 29 Oct 2003 and the second minimum at 1:00 UT on 30 Oct 2003. The event was split at 14:00 UT on 29 Oct 2003 (shown as vertical line in Figure 4a). Event 2 has the first minimum at 2:00 UT on 15 Dec 2006 and the second at 09:00 UT. The split was done at 03:00 UT on 15 Dec 2006 (Figure 4b).

[52] In the following sections, we describe the results obtained with the skill scores. Rankings shown in plots described below were obtained for model settings that were run in at least three of the four events. This excludes 3_RCM, 4_RCM (both run for only one event), 4_RAMSCB and 5_RAMSCB (run for two events).

6.1. Correlation Coefficient and Prediction Efficiency

[53] We compute *CC* and *PE* using 1 minute USGS and 1 hour Kyoto D_{st} observations and model results interpolated to 1 minute snapshots when using USGS data or averaged to the 1 hour cadence when using Kyoto data.

[54] Figure 5a, shows the ranking of all runs using the Correlation Coefficient (*CC*) based on comparison with 1 minute USGS D_{st} data and Figure 5c shows the ranking from the comparison with Kyoto D_{st} data. In Figure 5a, the specification models (shown in black) appear at the left of the plot. RiceDST suffered weak performance for two

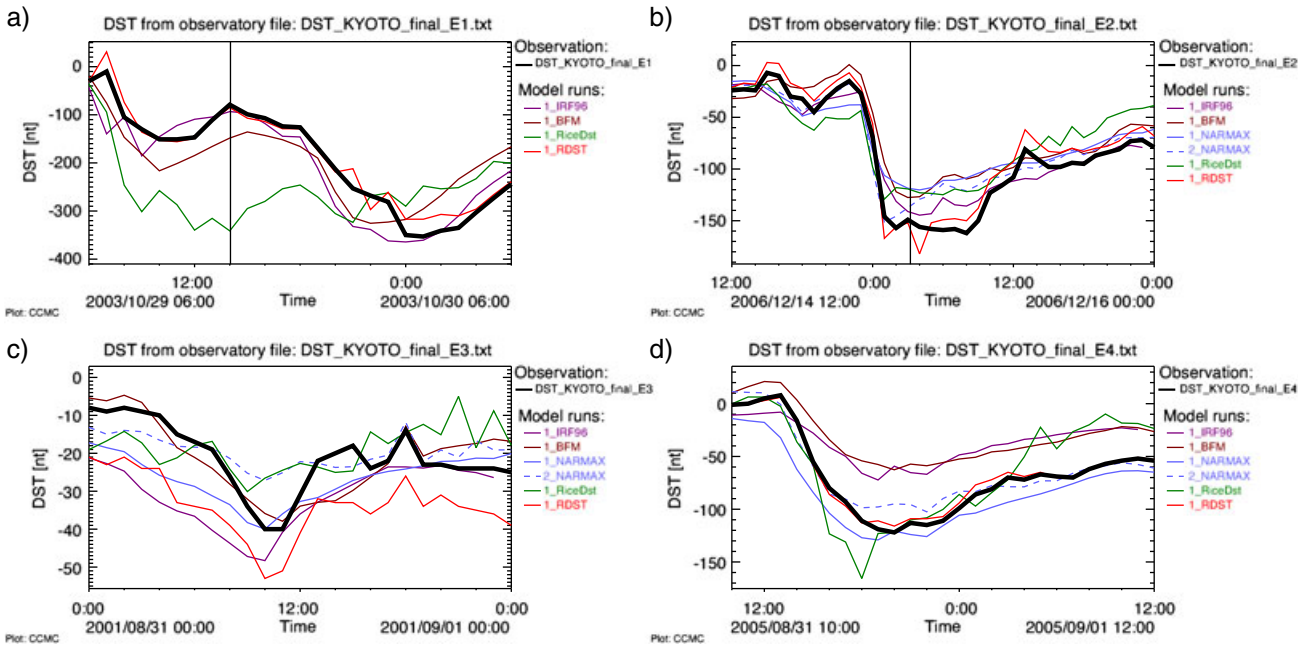

Figure 4. One hour KYOTO D_{st} data and D_{st}-specification model results. Kyoto D_{st} is shown in black and the D_{st}-specification model results are shown in colored traces. Simulation runs are listed in the bottom section in Table 2. IRF96 is shown in purple, BFM in brown, NARMAX in blue, RiceDST in green and RDST in red. The first two events feature two minima of D_{st} and are split in two for Timing Error metric analysis along the vertical lines shown in Figures 4a and 4b. In comparison with Figures 2 and 3, one may note the difference in the baseline between the Kyoto and the USGS D_{st} index. This is especially apparent for the weakest event shown in Figures 2c and 3c.

events and fell outside of the range occupied by the other statistical models. Event 3 resulted in weak performance for most models as can be seen from the diamonds appearing far below the averages. A few magnetosphere models, BFM and RiceDST, do not perform well for event 1, shown as squares.

[55] Several of the magnetosphere model runs (WINDMI, 7_SWMF and 8_SWMF, shown in red) and one ring current model (3_RCM, shown in blue) closely match the performance of the statistical models (shown in black), followed by the bulk of ring current models (all three RAM-SCB settings that were run for three events and 2_RCM, shown in blue). On the right half of the plot, the remaining magnetosphere models (6_SWMF, 2_LFM-MIX, 2_CMIT, and the OPENGGCM runs) exhibit a wide variation of performance between the events with average scores near zero (no skill).

[56] We also note that event 3 (diamonds) results in poor performance for nearly all models (many physics-based and all statistical models). Several models also show poor performance for event 1 (square well below the average for 1_BFM, 8_SWMF, 5_SWMF, 6_SWMF, 1_RiceDST and 4_OPENGGCM). Results for event 4 are always above average and mostly above average for event 2. The ranking is only slightly changed when comparing against Kyoto D_{st} in Figure 5c.

[57] Figure 5b shows how the models perform using the Prediction Efficiency *PE* using USGS D_{st} data and Figure 5d shows *PE* results using Kyoto D_{st} data. The ranking is completely different from the ranking derived from the *CC* scores: Three of the RAM-SCB runs (1_RAMSCB, 3_RAMSCB, 5_RAMSCB) are followed by two SWMF runs (6_SWMF, 8_SWMF) then followed by 3_RCM, 1_NARMAX and 2_RCM. The center is dominated by magnetosphere models mixed with statistical models.

[58] The Prediction Efficiency *PE* takes into account the variance of the observed signal during an event in relation to the difference of observations and model results. In contrast to the Correlation Coefficient, *PE* includes the effects of biases and the amplitude of the modeled signal in addition to the shape of the time series. Anti-correlated signals and signals with good correlation but incorrect amplitudes may result in negative *PE* scores.

[59] Between Figures 5b and 5d, we can see two differences in *PE* between USGS and Kyoto D_{st}: The first difference is that the 1 hour averaging results in better *PE* scores for all models. The spread of scores between the events for each model is much smaller. The second difference is that *PE* scores derived from USGS D_{st} in Figure 5b for event 3 are far below the average for almost all models. In Figures 5d, *PE* scores derived from Kyoto D_{st} for event 3 agree better with *PE* scores from other events. This has the largest effect on the D_{st}-specification models that rank near the top of all models when averaged over all events. This change in the performance of the specification models can be explained by the fact that they were developed to match the Kyoto D_{st} including its baseline. Scores for event 3 with its small D_{st} amplitude are most sensitive to the baseline, which is different in USGS data compared to Kyoto data.

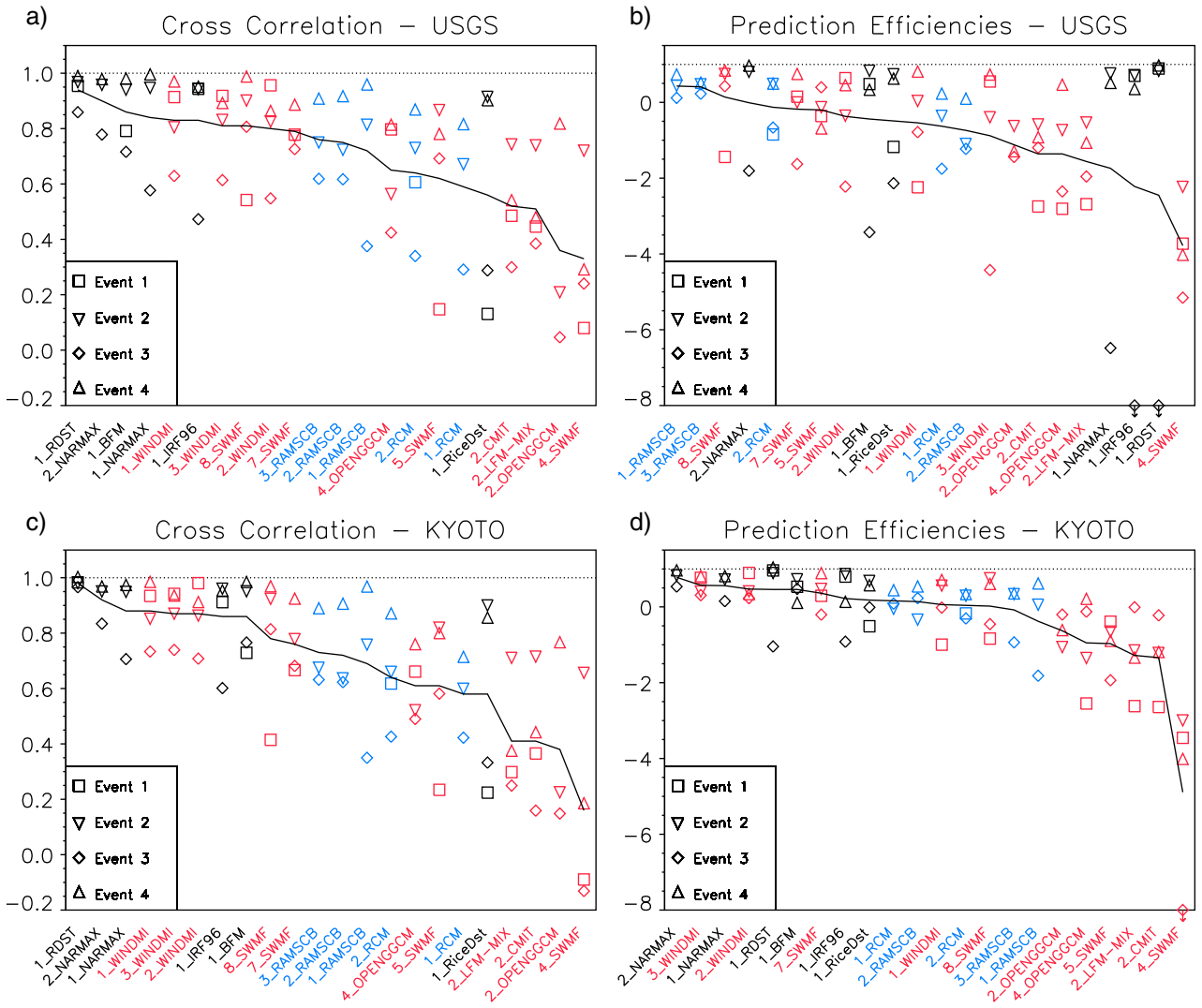

Figure 5. Model ranking using Correlation Coefficient (*CC*) and Prediction Efficiency (*PE*). Ranking with respect to *CC* for (a) USGS and (c) Kyoto D_{st}, respectively. Models performing best on average are to the left. Besides the average score (solid line), the scores for the events are plotted as symbols for each run: Event 1: square, event 2: triangle with point down, event 3: diamond, and event 4: triangle with point up. (b, d) Models ranked by Prediction Efficiency. Runs performed with the different model types are shown in different colors (magnetosphere models in red, ring current models in blue and D_{st}-specification models in black). All panels also contain the ideal score (*CC* = 1, *PE* = 1) as the horizontal dashed line near the top. Note that the vertical range for *CC* starts at –0.2 and ends at 1.1, and for *PE*, the range starts at –8 and ends at 1.5. Two *PE* scores for 1_IRF96 (–10.6) and 1_RDST (–12.5) lie below the vertical plot range in Figure 5b and one score for 4_SWMF (–9.1) is below the plot range in Figure 5d, indicated by the symbols with downward arrows at the bottom.

6.2. Model Yield

[60] Figures 6a and 6c show the model runs sorted by Model Yield (*YI*). Low values are on the left and high values on the right. Best-scoring models (*YI* ∼ 1) are near the right of each panel. Figure 6a shows the ranking when using USGS D_{st} data and Figures 6c shows the ranking when using Kyoto D_{st} data: The D_{st}-specification models all score between 0.5 and 1.3 with the RDST model being the closest to 1 on average and with the smallest spread between the events. Magnetosphere runs 7_SWMF and 8_SWMF are performing equally well. A few ring current models (1_RCM, 2_RAMSCB, 4_RAMSCB) perform well on average but with a considerably larger variance among the events. Magnetosphere model runs with the exception of 7_SWMF, 8_SWMF and the two OpenGGCM runs yield an average score below 1, indicating that the models only weakly reproduce the changes of D_{st} or that they do not see the signal at all. The models' average *YI* values are very similar in the comparisons with USGS and Kyoto D_{st} and only a few models change places between the two panels. The model that changes position most is OpenGGCM, which experienced the largest short-term fluctuations. Averaging

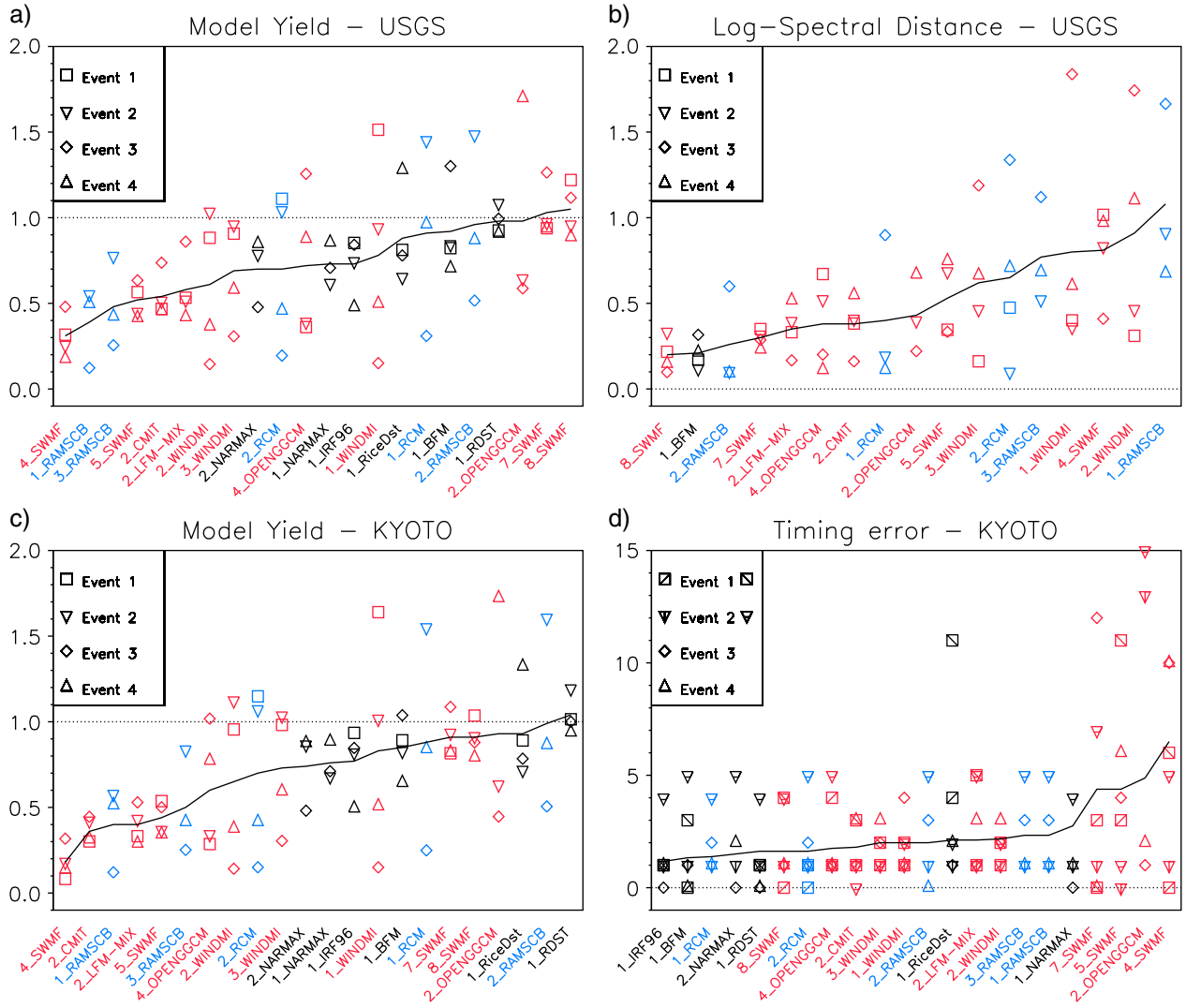

Figure 6. Model ranking using Yield (*YI*), Log-Spectral Distance M_s and Timing Error ΔT. For *YI* in Figures 6a and 6c (USGS and Kyoto D_{st}, respectively), the best performing models are near the right (nearest the score of $YI = 1$, shown as the dashed near the middle of the vertical range of each panel). Figure 6b shows the log-spectral distance M_s based on USGS D_{st} with best-performing models on the left. For the Timing Error ΔT, based on Kyoto D_{st}, in Panel d), events 1 and 2 have been split and two values have been obtained. These values are indicated by the two squares with a diagonal line for event 1 and two triangles with a vertical and horizontal line for event 2. The best models (near $\Delta T = 0$) are on the left of the plot. Ideal scores ($M_s = 0$, $\Delta T = 0$) are shown as dashed lines near the bottom of the vertical ranges in both panels.

reduces the *YI* values considerably. The spread between events for individual models is similar in Figures 6a and 6c.

6.3. Log-Spectral Distance and Timing Error

[61] Figure 6b shows the model runs sorted by Log-Spectral Distance M_s with the best-scoring models (lowest values of M_s) on the left. Of the specification models, only BFM is included here since it produces output at the 1-minute cadence for a fair comparison in this skill score. The magnetospheric models 8_SWMF and 7_SWMF rank best together with the BFM model, featuring a low average and small variation between the events. The ring current model 2_RAMSCB matches the performance on average but Event 3 posed a challenge. 2_RAMSCB (as all the other RAM-SCB settings) was not run for Event 1.

[62] Figure 6d shows the model runs sorted by Timing Error ΔT with the best score (zero) on the left. To be acceptable, the Timing Error should be zero or 1 hour for events with a well-defined minimum of the D_{st}. The D_{st} specification model (and 2_RAMSCB) in the left third of the plot performed best, predicting the D_{st} minimum within 2 hours, except for the second section of event 2. Moderately successful models (WINDMI, 8_SWMF and 5_SWMF, 1_RAMSCB and 3_RAMSCB, 2_CMIT, and 2_RCM in the center) predicted the minimum within 5 hours. Models on the right, third of Figure 6d show a very large variance between the events. Scores of six or larger indicate that the models did not predict the minimum at all. The modeled minimum is assumed at a random place anywhere in the time frame of the event. All magnetosphere models (except

Table 3. Average and Standard Deviations for Prediction Efficiency (*PE*), Model Yield (*YI*), Correlation Coefficient (*CC*), Timing Error (ΔT) and Log-Spectral Distance (M_s) for Comparisons with KYOTO and USGS D_{st} Values[a]

DST	Skill Scores							
	Kyoto				USGS			
Model ID	PE	YI	CC	ΔT	PE	YI	CC	M_s
4_SWMF	–4.89 ± 2.85	0.18 ± 0.10	0.16 ± 0.37	6.50 ± 4.04	–3.78 ± 1.23	0.31 ± 0.13	0.33 ± 0.28	0.81 ± 0.28
5_SWMF	–0.97 ± 0.68	0.44 ± 0.10	0.61 ± 0.27	4.38 ± 2.87	–0.20 ± 0.48	0.52 ± 0.10	0.62 ± 0.32	0.53 ± 0.22
7_SWMF	0.36 ± 0.44	0.91 ± 0.13	0.76 ± 0.12	4.38 ± 5.34	–0.18 ± 1.01	1.03 ± 0.16	0.79 ± 0.06	0.30 ± 0.05
8_SWMF	0.02 ± 0.79	0.91 ± 0.10	0.78 ± 0.25	1.62 ± 0.75	0.14 ± 1.07	1.05 ± 0.15	0.81 ± 0.19	0.20 ± 0.10
2_OPENGGCM	–0.63 ± 0.40	0.93 ± 0.69	0.38 ± 0.33	4.88 ± 6.12	–1.12 ± 0.47	0.98 ± 0.63	0.36 ± 0.40	0.43 ± 0.23
4_OPENGGCM	–0.95 ± 1.24	0.60 ± 0.35	0.61 ± 0.12	1.75 ± 0.96	–1.36 ± 1.49	0.72 ± 0.43	0.65 ± 0.19	0.38 ± 0.26
2_CMIT	–1.34 ± 0.86	0.36 ± 0.07	0.41 ± 0.20	1.80 ± 1.15	–1.36 ± 0.96	0.54 ± 0.13	0.52 ± 0.19	0.38 ± 0.16
2_LFM-MIX	–1.28 ± 1.07	0.40 ± 0.11	0.41 ± 0.21	2.12 ± 1.03	–1.56 ± 0.96	0.58 ± 0.19	0.51 ± 0.16	0.35 ± 0.15
1_WINDMI	0.06 ± 0.77	0.83 ± 0.65	0.88 ± 0.11	2.00 ± 1.35	–0.54 ± 1.30	0.78 ± 0.59	0.83 ± 0.15	0.80 ± 0.70
2_WINDMI	0.47 ± 0.30	0.65 ± 0.47	0.87 ± 0.12	2.17 ± 0.76	–0.37 ± 1.30	0.61 ± 0.42	0.80 ± 0.18	0.91 ± 0.66
3_WINDMI	0.57 ± 0.23	0.73 ± 0.34	0.87 ± 0.09	2.00 ± 1.00	–0.88 ± 2.41	0.69 ± 0.30	0.81 ± 0.14	0.62 ± 0.43
1_RAMSCB	–0.38 ± 1.27	0.40 ± 0.25	0.69 ± 0.31	2.33 ± 1.15	0.43 ± 0.29	0.39 ± 0.23	0.72 ± 0.30	1.08 ± 0.52
2_RAMSCB	0.14 ± 0.39	0.99 ± 0.56	0.72 ± 0.15	2.00 ± 1.73	–0.74 ± 0.69	0.96 ± 0.49	0.75 ± 0.15	0.26 ± 0.29
3_RAMSCB	–0.08 ± 0.74	0.50 ± 0.30	0.73 ± 0.13	2.33 ± 1.15	0.41 ± 0.16	0.48 ± 0.27	0.76 ± 0.14	0.77 ± 0.31
4_RAMSCB	–1.59 ± 2.82	1.09 ± 0.86	0.72 ± 0.04	2.50 ± 2.12	–3.54 ± 2.85	1.09 ± 0.84	0.78 ± 0.09	0.78 ± 0.23
5_RAMSCB	–0.81 ± 1.42	0.27 ± 0.20	0.78 ± 0.01	3.50 ± 2.12	0.29 ± 0.11	0.26 ± 0.19	0.78 ± 0.09	1.24 ± 1.01
1_RCM	0.16 ± 0.20	0.88 ± 0.65	0.58 ± 0.14	1.38 ± 1.11	–0.63 ± 1.01	0.91 ± 0.57	0.59 ± 0.27	0.40 ± 0.43
2_RCM	0.04 ± 0.33	0.70 ± 0.49	0.64 ± 0.18	1.62 ± 1.11	–0.13 ± 0.73	0.70 ± 0.45	0.64 ± 0.22	0.65 ± 0.52
3_RCM	0.50 ± NaN	0.91 ± NaN	0.82 ± NaN	0.50 ± NaN	0.13 ± NaN	0.85 ± NaN	0.81 ± NaN	0.12 ± NaN
4_RCM	0.14 ± NaN	1.05 ± NaN	0.62 ± NaN	2.00 ± NaN	–0.27 ± NaN	1.02 ± NaN	0.61 ± NaN	0.41 ± NaN
1_IRF96	0.22 ± 0.84	0.77 ± 0.19	0.86 ± 0.17	1.17 ± 1.26	–2.22 ± 5.62	0.73 ± 0.17	0.83 ± 0.24	not calculated
1_BFM	0.46 ± 0.30	0.85 ± 0.16	0.86 ± 0.13	1.33 ± 1.53	–0.44 ± 2.01	0.92 ± 0.26	0.86 ± 0.12	0.21 ± 0.08
1_NARMAX	0.56 ± 0.36	0.76 ± 0.11	0.88 ± 0.15	2.75 ± 3.33	–1.74 ± 4.12	0.73 ± 0.12	0.84 ± 0.23	not calculated
2_NARMAX	0.78 ± 0.21	0.74 ± 0.23	0.92 ± 0.07	1.50 ± 1.29	–0.01 ± 1.56	0.70 ± 0.20	0.90 ± 0.11	not calculated
1_RiceDST	0.18 ± 0.56	0.93 ± 0.27	0.58 ± 0.35	2.12 ± 1.31	–0.49 ± 1.41	0.88 ± 0.27	0.56 ± 0.41	not calculated
1_RDST	0.46 ± 1.01	1.04 ± 0.11	0.98 ± 0.01	1.62 ± 1.97	–2.45 ± 6.70	0.98 ± 0.08	0.94 ± 0.06	not calculated

[a]The models are listed in the order they are introduced in Table 2, starting with magnetosphere models, then ring current models, and, finally, D_{st}-specifications. Standard error values of NaN (Not-A-Number) appears where a model was only run for a single event.

WINDMI), RCM, and RiceDST have runs that fall into this category for at least one event.

[63] Almost all events posed challenges to the models that stem from two effects: First, for two events, the D_{st} values remained near the minimum value for an extended period of time, which potentially makes it harder for models to hit the correct time when the actual minimum was reached. In event 2, the minimum D_{st} (–150 nT) was reached at 1:00 UT on 15 Dec 2006 (Figure 4b), but remained within 20 nT for eight more hours. As a consequence, event 2 (triangle with downward point) was split at 2:00 UT on 15 Dec 2006 to examine the first minimum at 1:00 UT separately from the later minimum at 8:00 UT. In event 4, the minimum (–120) was reached at 18:00 UT (8th hour in the plots of Figure 4d) on 31 Aug 2005 in USGS D_{st}, but at 20:00 UT in KYOTO D_{st} observations. The D_{st} value remains close to that minimum for 3 hours (18:00 UT to 21:00 UT). Models reached an acceptable score if they predicted the minimum anywhere within the range. The second effect was that the weakest event (event 3) challenged all models including the statistical models due to the weak D_{st} signal. The baselines of many model runs were considerably different from the baselines of both the USGS and Kyoto D_{st} indices. The baseline difference between the Kyoto and USGS D_{st} was about 20 nT, a substantial fraction of the overall strength of the event (–50 nT).

[64] Table 3 summarizes the skill scores used for the rankings presented in this section (average over all events and standard deviation). All skill scores were computed for Kyoto D_{st} values (1 hour intervals) and USGS D_{st} (1 minute intervals). Some models were run for only a single event and "NaN" (Not-A-Number) appears as the standard deviation. For the plots and rankings in this chapter, *PE* values derived from Kyoto data as listed in Table 3 were not used.

6.4. Two-Dimensional Scores

[65] A visual impression of the distribution of the different model classes in the multi-dimensional skill score space can be obtained by plotting each model run's position with respect to two skill scores. In Figure 7, we present the locations of runs in *PE*-*CC* space (left column) and ΔT-*YI*-space (right column). The top row shows results of magnetosphere model runs, the middle row the ring current model runs and the bottom row shows the D_{st}-specification models. Dashed lines in each panel illustrate the ideal scores of each of the two skill scores with the intersection denoting the ideal combined score. A run is identified by symbol size and color, events are identified by the different shapes of the symbols.

[66] We note that the different types of models show distinct distribution patterns: Magnetosphere models fill a wide area in the *PE*-*CC*-space with most model runs with $PE > -6$ and $CC > 0$ (only 4_OPENGGCM and 2_LFM-MIX show one score outside that range). WINDMI runs (green, dark green) are characterized by higher *CC* values ($CC > 0.5$). WINDMI runs are joined by runs of SWMF (7_SWMF, 8_SWMF, red) near the ideal score ($PE > 0$, $CC > 0.8$). In ΔT-*YI* space, most model runs exhibit low Yield values $0.1 < YI < 0.9$. A few (mostly OpenGGCM

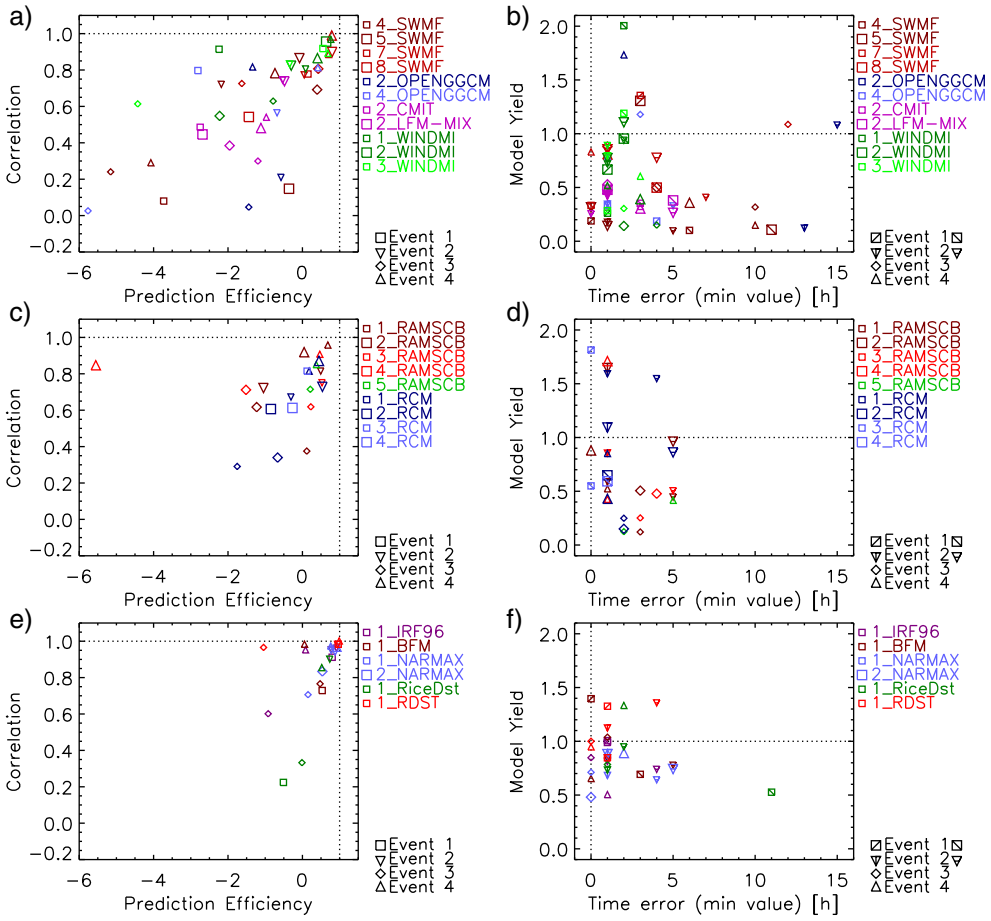

Figure 7. Model runs in 2-D skill score space. (a, c, e, left column) Prediction Efficiency (*PE*) and Correlation Coefficient (*CC*), with Figures 7a and 7c using USGS D_{st} values and Figure 7e using KYOTO D_{st} values. (b, d, f, right column) Timing Error (ΔT) of minimum D_{st} value and Model Yield (*YI*) based on Kyoto D_{st}. The Figures 7a and 7b, show magnetosphere models, Figures 7c and 7d, show ring current models, and Figures 7e and 7f, show D_{st}-specification models. Model runs are listed on the top right in each panel with a sample symbol for event 1 shown. The color scheme is the same as in Figure 2. Runs that share a color are distinguished by symbol size. Each model setting may be run for up to four events, symbolized by a square for event 1, a triangle pointing down for event 2, a diamond for event 3 and a triangle pointing up for event 4. Dashed lines in the plot (vertical for the *X*-axis score, horizontal for the *Y*-axis score) indicate perfect scores. Events 1 and 2 have been split to determine ΔT.

runs) have too large Yields (up to $YI = 3.25$). Scores below $YI = 1$ may be attributed to a weak dynamic response of MHD models due to inevitable damping, diffusion and the lack of inner-magnetospheric physics if a ring current model is not coupled into the MHD model of the magnetosphere. The Timing Errors vary widely (up to 18 hours, comparable to the entire length of a simulation, typically 24 hours). One might expect that runs that score low in *YI* may also exhibit large ΔT values. This was not the case as we found both good and bad scores in ΔT regardless of the Yield score.

[67] SWMF runs are consistently below or near $YI = 1$, with the ones having coupled CRCM or RCM (7_SWMF and 8_SWMF, red) scoring close to or slightly above 1. Even runs with a good Yield (such as 7_SWMF and 8_SWMF) may score poorly in ΔT if they predicted the global minimum at the time when a smaller, temporary minimum was reached (such as in event 1). WINDMI runs, in general, represented the time history of D_{st} well, resulting in good scores in terms of Timing Error ($\Delta T \leq 3$). Yields, however, were more varied than for runs of SWMF ($0.15 < YI < 1.6$). CMIT and LFM runs scores fairly in terms of Timing Error ($2 \leq \Delta T \leq 7$) and also fairly in terms of Yield ($0.3 \leq YI \leq 0.7$). OpenGGCM runs tend to over-estimate variability and often resulted in $YI \gg 1$. However, the time series of D_{st} were seldom reproduced. We saw low correlations and minima that were reached at random times giving poor scores (ΔT up to 13 hours). Sometimes, model runs (e.g., 7_SWMF, 8_SWMF, and WINDMI runs) would continue to predict decreasing values of D_{st} although the observation indicates a recovery (e.g., event 2 in Figure 2b).

[68] Ring current models populate a smaller area in *PE*-*CC* space ($PE > -2$, $CC > 0.2$) than the magnetosphere models. RAMSCB features higher scores ($CC > 0.6$). RCM run scores are in the middle (0.3–0.9). Best-performing

combined scores have been achieved with RAM-SCB, but several RCM runs come near the ideal score as well. In terms of ΔT and *YI*, ring current models show a smaller scatter in ΔT ($\leq$ 13) than the magnetosphere models and a similar range in Yield ($0.1 < YI < 21.8$). All RAM-SCB runs and all RCM runs have $\Delta T \leq 5$. RAM-SCB *YI* scores are below 1, except one for event 4 at $YI = 1.7$ and RCM *YI* scores can reach values up to 1.8.

[69] Most D_{st} specification models do very well in *PE*-*CC*-space compared to first-principles models. All but one of the runs for events 1, 2 and 4 have $PE > 0.1$ and $CC > 0.7$. All models, however, score poorly for event 3 (diamonds scattered between $-1 < PE < 0$ and $0.3 < CC < 0.85$). The relatively weak D_{st} minimum (–40) and the shape of the time series of D_{st} for that event proved difficult to predict by all of the models. Many models showed a bias of up to 15 nT at the beginning of the time period shown in Figure 4, which severely impacted their *PE* score.

[70] In ΔT-*YI* space, all specification models are close to the ideal Yield and many show good to moderate Timing Errors ($\Delta T < 5$). Event 2 proved challenging, with D_{st} fluctuating near the minimum for nearly 9 hours resulting in all models to score between 4 and 7 hours in ΔT. RiceDST performed poorly for event 1 on all scores. It predicted the global minimum near the time of the weaker minimum seen early in the event time period but missed the larger minimum altogether.

7. Discussion

[71] The large number of models were categorized into magnetosphere (MHD) models, ring current (kinetic) models and D_{st}-specification models: Magnetosphere models are coupled magnetosphere-ionosphere models that may include a ring current model component. Most magnetosphere models are—DMHD models with the exception of WINDMI, which describes the energy flow from solar wind into the magnetosphere, ring current and ionosphere in a low-dimensional manner. Ring current models are kinetic models of the drift physics in the inner magnetosphere that were either run in stand-alone mode driven by statistical plasma sheet, electric field and magnetic field models or were run coupled to magnetosphere MHD models. D_{st}-specification models run instantly off the solar wind data and (mostly) predict D_{st} on a 1 hour cadence to replicate the Kyoto-D_{st} index.

[72] Five skill scores were used in the investigation:

[73] ***Correlation Coefficient*:** In terms of Correlation Coefficient (*CC*), we found that specifications of D_{st} performed best (led by 1_RDST, followed by 2_NARMAX 1_BFM and 1_NARMAX runs) with several coupled magnetosphere-ring-current-ionosphere models (all WINDMI runs, 7_SWMF and 8_SWMF) following not far behind. In the middle were stand-alone ring current models (all RAM-SCB runs and three RCM runs) and an older version of a coupled magnetosphere-ring-current-ionosphere model (5_SWMF). RiceDST fared worst among the specification models, scoring in the trailing third of all models, just ahead of magnetosphere-ionosphere models not containing a ring current component, i.e., the two OpenGGCM runs, 2_CMIT and the 2_LFM-MIX runs (listed not in order of performance).

[74] ***Prediction Efficiency*:** If we look at Prediction Efficiency, which also takes into account model biases and the variance of the modeled signal, the ranking is considerably different. Prediction Efficiencies change when the model comparisons are made against the Kyoto D_{st} instead of the USGS D_{st} data. Event 3 yielded far worse model scores with USGS D_{st} compared to Kyoto D_{st}, especially affecting the specification models. All models experience a larger variation of scores between the events when compared against USGS 1 minute data, reflecting the challenge faced by all models to specify the data on short time scales. Like *CC* scores, *PE* scores were worst for magnetosphere-ionosphere models that do not contain a ring-current component.

[75] ***Model Yield*:** In terms of Yield (*YI*), best-scoring models are the versions of SWMF that have ring current models coupled into the magnetosphere (7_SWMF with CRCM, 8_SWMF with RCM), several RCM runs (1_RCM and 3_RCM, on average) and RAM-SCB runs (2_RAM-SCB, 4_RAMSCB, on average) and 1_RDST. Specification models other than RDST show yields below unity as well as most magnetosphere magnetohydrodynamic models without coupled ring current models. OpenGGCM and WINDMI runs suffer from a very large variability in *YI* between the events, often exceeding unity for some events while remaining far below unity for other events.

[76] ***Log-Spectral Distance*:** The log-spectral distance was computed using USGS data and excluded most specification models because of their 1 hour output resolution. The one specification model (1_BFM), the two coupled magnetosphere-ring-current-ionosphere models (8_SWMF, 7_SWMF), and 2_RAMSCB did best. CMIT/LFM and OpenGGCM runs followed close behind. Most models in trailing positions (WINDMI, other RAMSCB, RCM and 4_SWMF) had large (poor) scores coming from the substorm event (event 3).

[77] ***Timing Error*:** Timing errors were derived using the Kyoto D_{st} and were measured in full hours. Events 1 and 2 had to be split to account for separate D_{st} minima that occurred. Models that scored well in *CC* or *PE* did best here as well: Specification models are leading together with some of the stand-alone ring current models and the SWMF runs that include a ring current component. In event 3, 7_SWMF failed to follow the recovery of D_{st} after the isolated substorm and thus scores a large error. WINDMI runs, as well as many of the ring current models, did not see the substorm at all.

[78] In this study, we found that each skill score by itself is not a very reliable measure of model performance. We found that D_{st}-specification models performed very well for the stronger events, but failed in terms of Prediction Efficiency due to differences in the baseline (the Kyoto and USGS D_{st} values at the start of the interval for event 3 differed by 8 nT). Prediction Efficiency is very sensitive to this type of bias for weak events. Another part of the difference between Correlation Coefficient and Prediction Efficiency can be explained by the Yield. Imperfect Yields paired with

good correlations result in worse performance in terms of Prediction Efficiency. The D_{st}-specification models consistently show $YI < 1$ and as a group show worse rankings in *PE* than in *CC*. The Timing Error, together with the Model Yield and Correlation Coefficient adds another dimension to the analysis of model performance to reproduce a time series of an index value.

[79] To visualize the performance of groups of different models, we plotted scores for each run for the individual events in 2-D plots (one on *PE-CC* space, the other in ΔT-*YI* space). We plotted the magnetosphere models, the ring current models and the D_{st}-specification models separately and found that the groups of models filled the skill score space quite differently. Magnetosphere model runs fill a large area in *PE-CC* space ($PE > -11$, $CC > -0.15$). The scatter was also large in ΔT-*YI* space ($\Delta T \leq 12$, $0.1 < YI < 1.7$). Larger timing errors were encountered more often for lower-yielding model runs. Most ring current model runs were clustered much closer to the ideal *PE* score ($PE > -2$, except for one run with $PE = -5.5$) with a smaller range in *CC* ($CC > 0.2$) Ring current models showed a smaller scatter in Timing Error ($\Delta T \leq 5$ hours) and a little less in Yield ($YI < 1.8$). Most Yields were well below unity as was the case with magnetosphere runs. D_{st} specification models were very close to perfect in Prediction Efficiency and Correlation Coefficient except for event 3 that challenged all the models. Timing errors were small to moderate ($\Delta T \leq 5$ hours, except for one run at 11 hours) and yields were closer to unity ($0.5 \leq YI \leq 1.4$) for all runs and events compared to the other model groups.

[80] No single model scores best in all the skill scores. The study included several sets of similar model settings: four SWMF, two OpenGGCM, two LFM (2_CMIT, 2_LFM-MIX), three WINDMI, five RAMSCB, four RCM, and two NARMAX runs. We found best-performing model settings for some models, while other model run sets did not show any preference for any one run:

[81] ***Magnetosphere Model Runs*:** Among the SWMF magnetosphere model runs, 8_SWMF (with CRCM) and 7_SWMF (with RCM) scored best, followed by 5_SWMF with an older implementation of the coupling of the magnetosphere MHD to the RCM model, and then 4_SWMF without any ring current model. Out of eight cases (*YI*, *CC*, *PE*, M_s computed using USGS, and *YI*, *CC*, *PE*, ΔT computed using Kyoto D_{st}), 8_SWMF scores best six times and 7_SWMF two times. 5_SWMF is second once and 4_SWMF always scores lowest among the SWMF runs. The four SWMF runs do differ in terms of resolution near the Earth: 7_SWMF and 5_ SWMF have a finer grid than 8_SWMF and 4_SWMF. The increased resolution does not seem to benefit a model run: 8_SWMF with $0.25R_E$ resolution leads or performs similarly to 7_SWMF with $0.125R_E$ resolution. Among the two runs with coupled RCM, 8_SWMF always scores better than 5_SWMF, although the latter has the finer resolution. The role of the grid resolution in the quality of the magnetosphere ring-current coupling needs to be investigated in a separate study.

[82] WINDMI results were mixed with 1_WINDMI (rectified solar wind driver) leading in four of the eight cases, and 2_WINDMI (Siscoe solar wind driver) and 3_WINDMI (Newell solar wind driver) leading in two other scores each. CMIT (the newer model version 2-1-5 with TIE-GCM ionosphere) scored practically identical to 2_LFM-MIX (model version 2.1-1). There was no clear preference among the OpenGGCM runs (2_OPENGGCM scored best in five out of eight cases).

[83] ***Ring Current Model Runs*:** None of the RAM-SCB runs were executed for Event 1, indicating that solar wind magnetic fields and plasma velocities exceeded the valid range of the Volland-Stern and Weimer electric field models as well as the LANL-SOPA plasma sheet models. 4_RAMSCB and 5_RAMSCB were not run for events 1 and 2 due to their computational cost as they were coupled to the SWMF magnetosphere MHD model. Large values of B_z in the solar wind during event 1 also exceeded the valid range of the statistical Tsyganenko-Mukai plasma sheet model driving 1_RCM. Large values of V_x exceeded the range of the Borovsky et al. plasma sheet model driving 2_RCM. Special settings in runs 3_RCM and 4_RCM were introduced for this event. None of the RCM run settings performed for all events. The four events represented very different solar wind driving conditions and responses of the magnetosphere. Stand-alone ring current models depend on their drivers (statistical plasma sheet and field models) and several settings need to be run to have one match the observed behavior. A thorough assessment of the performance of the RCM drivers represented here is in preparation by S. Sazykin.

[84] Among ring current models, 2_RAMSCB ranked best among the RAM-SCB runs. The RAM-SCB runs driven by first-principles models (4_RAMSCB and 5_RAMSCB) were executed for only two events (3 and 4), which did not provide enough data for a valid comparison with all models based on all events. On average, 4_RAMSCB scored second for two metrics (*YI*, ΔT) and fourth for *PE*. 5_RAMSCB scored last for *PE* and *YI* and first for ΔT. All runs score very similarly in *CC*.

[85] ***D_{st}-specification Model Runs*:** Among the D_{st}-specification models, only the NARMAX model had multiple runs: 2_NARMAX outscored 1_NARMAX in *CC* and *PE* but 1_NARMAX leads in terms of *YI* and ΔT, favoring none.

[86] We had model runs that scored nearly perfectly in the timing of the minimum D_{st} value, but performed poorly in terms of other scores such as Yield or Correlation Coefficient. Many model runs poorly estimated the time of the D_{st} minimum. In event 1, several models predicted a single minimum early and then predicted a slow recovery while the observation shows a weak minimum that was followed by a stronger minimum several hours later. In this case, models may score well on the Yield, but poorly in Timing Error and fairly well in terms of Correlation Coefficient. In event 2, the time of the D_{st} minimum was hard to predict since the D_{st} fluctuated near the minimum value for an 8 hour period. Models that fail to predict a pronounced minimum usually have a large Timing Error (the minimum may be at any random time during the time intervals), a low Yield, poor Correlation Coefficient and Prediction Efficiency scores.

[87] In general, the D_{st} specification models perform best but some first-principles models come close for some events, especially during the stronger storms. The weakest event (event 3, the isolated-substorm event) poses a particular challenge for the physics-based models. The D_{st}-specification models struggle to the get the baseline right even if they managed to predict the substorm during the event. Since the USGS D_{st} features a baseline that is different from the Kyoto D_{st}, the specification models that were developed to match the Kyoto D_{st} do much worse when compared to USGS D_{st} data. RDST scores best since it is a real-time implementation of the D_{st} index using magnetometer data from a set of four magnetometers that is similar to the set used for the Kyoto D_{st} index (RDST uses Guam instead of Kyoto). Unlike the first-principles models and the other specification models, RDST cannot be run using solar wind measurements to obtain a prediction ahead of actual observations on the ground.

[88] The results of this challenge provide a baseline for future validation studies using new models and improved models. Model outputs used in this study together with the observation data are available on the CCMC web site (http://ccmc.gsfc.nasa.gov under "Metrics and Validation" and then "GEM Challenge") for use by the space science community. Skill scores as presented in this paper can be obtained through the online visualization tool as well.

[89] **Acknowledgments.** Hourly D_{st} data were obtained from the World Data Center of Geomagnetism, Kyoto, Japan and 1 minute data were obtained from the United States Geological Survey (USGS). Both index values include magnetic data from the following stations: KAK: Kakioka Magnetic Observatory, Japan Meteorological Agency, Japan, HON, SJG: Honolulu and San Juan magnetic observatories, USGS, HER: Hermanus Magnetic Observatory, South African National Space Agency (SANSA). Solar wind input data for the models (magnetic field and plasma parameters) were obtained from OMNI (http://omniweb.gsfc.nasa.gov) and CDAweb (cdaweb.gsfc.nasa.gov) databases. This work was supported by the Center for Integrated Space Weather Modeling, which is funded by the Science and Technology Centers program of the National Science Foundation under agreement number ATM-0120950. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

References

Bala, R., and P. Reiff (2012), Improvements in short-term forecasting of geomagnetic activity, *Space Weather*, *10*, S06001, doi:10.1029/2012SW000779.

Billings, S., S. Chen, and M. Korenberg (1989), Identification of MIMO non-linear systems using a forward-regression orthogonal estimator, *Int. J. Control*, *49*(6), 2157–2189, doi:10.1080/00207178908559767.

Billings, S. A., and W. S. F. Voon (1986), Correlation based model validity tests for non-linear models, *Int. J. Control*, *44*, 235–244.

Birn, J., et al. (2001), Geospace environmental modeling (GEM) magnetic reconnection challenge, *J. Geophys. Res.*, *106*(A3), 3715–3719, doi:10.1029/1999JA900449.

Boaghe, O. M., M. A. Balikhin, S. A. Billings, and H. Alleyne (2001), Identification of nonlinear processes in the magnetospheric dynamics and forecasting of Dst index, *J. Geophys. Res*, *106*, 30,047–30,066, doi:10.1029/2000JA900162.

Borovsky, J. E., M. F. Thomsen, R. Elphic, T. E. Cayton, and D. J. McComas, (1998), The transport of plasma sheet material from the distant tail to geosynchronous orbit, *J. Geophys. Res.*, *103*, 20,297–20,331, doi:10.1029/97JA03144.

Boyle, C. B., P. H. Reiff, and M. R. Hairston (1997), Empirical polar cap potentials, *J. Geophys. Res.*, *102*, 111–125, doi:10.1029/96JA01742.

Boynton, R. J., M. A. Balikhin, S. A. Billings, A. S. Sharma, and O. A. Amariutei (2011a), Data derived NARMAX Dst model, *Ann. Geophys.*, *29*(6), 965–971, doi:10.5194/angeo-29-965-2011.

Boynton, R. J., M. A. Balikhin, S. A. Billings, H. L. Wei, and N. Ganushkina (2011b), Using the NARMAX OLS-ERR algorithm to obtain the most influential coupling functions that affect the evolution of the magnetosphere, *J. Geophys. Res.*, *116* (A5), A05218, doi:10.1029/2010JA015505.

Burke, W. J., L. C. Gentile, and C. Y. Huang (2007), Penetration electric fields driving main phase Dst, *J. Geophys. Res.*, *112*, A07208, doi:10.1029/2006JA012137.

Burke, W. T. (2007), Penetration electric fields: A Volland-Stern approach, *J. Atmos. Sol.-Terr. Phys.*, *69*, 1114–1126, doi:10.1016/j.jastp.2006.09.013.

Burton, R. K., R. L. McPherron, and C. T. Russell (1975), An empirical relationship between interplanetary conditions and Dst, *J. Geophys. Res.*, *80*, 4204–4214, doi:10.1029/JA080i031p04204.

Buzulukova, N. M. C. F., A. Pulkkinen, M. Kuznetsova, T. E. Moore, A. Glocer, and L. Rastätter (2010), Dynamics of ring current and electric fields in the inner magnetosphere during disturbed periods: CRCM-BATS-R-US coupled model, *J. Geophys. Res.*, *115*, A05210, doi:10.1029/2009JA014621.

Cheng, C. Z. (1995), Three-dimensional magnetospheric equilibrium with isotropic pressure, *Geophys. Res. Lett.*, *22*, 2401–2404, doi:10.1029/95GL02308.

Dessler, A. J., and E. N. Parker (1959), Hydromagnetic theory of geomagnetic storms, *J. Geophys. Res.*, *64*, 2239–2252, doi:10.1029/JZ064i012p02239.

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

Feldstein, Y. (1992), Modelling of the magnetic field of the magnetospheric ring current as a function of interplanetary medium parameters, *Space Sci. Rev.*, *59*, 83–165, doi:10.1007/BF01262538.

Freeman, J., R. Wolf, R. Spiro, B. Hausman, B. Bales, and R. Lambour, (1994), A real-time magnetospheric specification model: Magnetospheric specification and forecast model (MSFM), Final Technical Report and Software Documentation, *Report for USAF contract F19628-90-K-0012*, Rice University, Houston, TX.

Fuller-Rowell, T., D. Rees, S. Quegan, R. J. Moffett, M. V. Codrescu, and G. H. Millward (1996), A coupled thermosphere-ionosphere model (CTIM), in *STEP Report*, edited by R. W. Schunk, p. 217, Scientific Communications on Solar Terrestrial Physics, Boulder, Colorado.

Gannon, J., J. Love, P. Friberg, D. Stewart, and S. Lisowski, (2011), U.S. Geological Survey near real-time Dst index, *U.S. Geological Survey Open File Report 2011-1030*, USGS, Reston, Virginia.

Gannon, J. L., and J. J. Love (2011), USGS 1-min Dst index, *J. Atmos. Solar-Terr. Phys.*, *73*, 323–334.

Harel, M., R. A. Wolf, P. H. Reiff, R. W. Spiro, W. J. Burke, F. J. Rich, and M. Smiddy (1981), Quantitative simulation of a magnetospheric substorm 1, model logic and overview, *J. Geophys. Res.*, *86*, 2217–2241, doi:10.1029/JA086iA04p02217.

Hill, T. W. (1984), Magnetic coupling between solar wind and magnetosphere: Regulated by ionospheric conductance, *Eos Trans. Am. Geophys. Union*, *65*, 1047.

Hilmer, R. V., and G. H. Voigt (1995), A magnetospheric magnetic-field model with flexible current systems driven by independent physical parameters, *J. Geophys. Res.*, *100*, 5613–5626, doi:10.1029/94JA03139.

Horton, W., and I. Doxas (1998), A low-dimensional dynamical model for the solar wind driven geotail-ionosphere system, *J. Geophys. Res.*, *103*, 4561–4572, doi:10.1029/97JA02417.

Jordanova, V. K., J. U. Kozyra, G. V. Khazanov, A. F. Nagy, C. E. Rasmussen, and M. C. Fok (1994), A bounce-averaged kinetic-model of the ring current ion population, *Geophys. Res. Lett.*, *21*, 2785–2788, doi:10.1029/94GL02695.

Jordanova, V. K., S. Zaharia, and D. T. Welling (2010), Comparative study of ring current development using empirical, dipolar, and self-consistent magnetic field simulations, *J. Geophys. Res.*, *115*, A00J11, doi:10.1029/2010JA015671.

Karinen, A., and K. Mursula (2006), Correcting the Dst index: Consequences for absolute level and correlations, *J. Geophys. Res.*, *111*, A08207, doi:10.1029/2005JA011299.

Leontaritis, I. J., and S. A. Billings (1985a), Input-output parametric models for non-linear systems part I: Deterministic non-linear systems, *Int. J. Control*, *41*(2), 303–328, doi:10.1080/0020718508961129.

Leontaritis, I. J., and S. A. Billings (1985b), Input-output parametric models for non-linear systems part II: Stochastic nonlinear systems, *Int. J. Control*, *41*(2), 329–344, doi:10.1080/0020718508961130.

Love, J. J., and J. L. Gannon (2009), Revised D_{st} and the epicycles of magnetic disturbance: 1958–2007, *Ann. Geophys.*, *27*, 3101–3131, doi:10.5194/angeo-27-3101-2009.

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, doi:10.1016/j.jastp.2004.03.020.

Lyons, L. (1998), The geospace modeling program grand challenge, *J. Geophys. Res.*, *103*(A7), 14,781–14,785, doi:10.1029/98JA00015.

Mays, M. L., W. Horton, E. Spencer, and J. Kozyra (2009), Real-time predictions of geomagnetic storms and substorms: Use of the solar wind magnetosphere-ionosphere system model, *Space Weather*, *7*, S07001, doi:10.1029/2008SW000459.

Merkin, V. G., and J. G. Lyon (2010), Effects of the low-latitude inospheric boundary condition on the global magnetosphere, *J. Geophys. Res.*, *115*, A10202, doi:10.1029/2010JA015461.

Murayama, T. (1982), Coupling function between solar wind parameters and geomagnetic indices, *Rev. Geophys. Space Phys.*, *20*, 623, doi:10.1029/RG020i003p00623.

Pulkkinen, A., L. Rastaetter, M. 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, *Journal of Geophysical Research*, *106*(A1), 345–348, doi:10.1029/2000JA000600.

Raeder, J., R. L. McPherron, L. A. Frank, S. Kokubun, G. Lu, T. Mukai, W. R. Paterson, J. B. Sigwarth, H. J. Singer, and J. A. Slavin (2001a), Global simulation of the geospace environment modeling substorm challenge event, *J. Geophys. Res.*, *106* (A1), 381–395, doi:10.1029/2000JA000605.

Raeder, J., Y. Wang, and T. Fuller-Rowell (2001b), Geomagnetic storm simulation with a coupled magnetosphere-ionosphere-thermosphere model, in *Space Weather*, edited by Song, P., H. J. Singer, and G. L. Siscoe, 377–384, AGU Geophys. Mongr. Ser, AGU, Washington, D. C. doi:10.1029/GM125p0377.

Rastätter, L., M. M. Kuznetsova, A. Vapirev, A. Ridley, M. Wiltberger, A. Pulkkinen, M. Hesse, and H. J. Singer (2011), Geospace environment modeling 2008–2009 challenge: Geosynchronous magnetic field, *Space Weather*, *9*, S04005, doi:10.1029/2010SW000617.

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, doi:10.1029/92GL00401.

Sazykin, S. (2000), Theoretical studies of penetration of magnetospheric electric fields to the ionosphere, Ph.D. thesis, University of Utah.

Sckopke, N. (1966), A general relation between energy of trapped particles and disturbance field near earth, *J. Geophys. Res.*, *71*, 3125.

Siscoe, G. L. (1982), Polar-cap size and potential—a predicted relationship, *Geophys. Res. Lett.*, *9*, 672–675, doi:10.1029/GL009i006p00672.

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.

Stern, D. P. (1975), The motion of a proton in the equatorial magnetosphere, *J. Geophys. Res.*, *80*, 595, doi:10.1029/JA080i004p00595.

Sugiura, M. (1964), Hourly values of equatorial Dst for the IGY, *Ann. Int. Geophys.*, *35*, 9–45.

Sugiura, M., and S. Hendricks (1967), Provisional hourly values of equatorial Dst for 1961, 1962, and 1963, in *Provisional Hourly Values of Equatorial Dst for 1961, 1962, and 1963, Technical Note D-4047*, 45 pp., NASA, Washington, D.C.

Sugiura, M., and T. Kamei (1991), Equatorial Dst index 1957–1986, in *Equatorial Dst Index 1957–1986, IAGA Bull.*, vol. 40, 7–14 pp., ISGI , Publ. Office, Saint-Maur-des-Fossess, France.

Tascione, T., H. W. Kroehl, R. Creiger, J. W. Freeman, R. A. Wolf, R. W. Spiro, R. V. Hilmer, J. Shade, and B. Hausman (1988), New ionospheric and magnetospheric specification models, *Radio Sci.*, *23*, 211–222, doi:10.1029/RS023i003p00211.

Toffoletto, F. R., S. Sazykin, R. W. Spiro, and R. A. Wolf (2003), Modeling the inner magnetosphere using the Rice Convection Model (review), *Space Sci. Rev., WISER Special Issue*, *107*, 175–196, doi:10.1023/A:1025532008047.

Tóth, G., et al. (2005), Space weather modeling framework: A new tool for the space science community, *J. Geophys. Res.*, *110*(A12), A12226, doi:10.1029/2005JA011126.

Tsyganenko, N. A. (1989), A magnetospheric magnetic-field model with a warped tail current sheet, *Planet. Space Sci.*, *37*, 5–20, doi:10.1016/0032-0633(89)90066-4.

Tsyganenko, N. A., and T. Mukai (2003), Tail plasma sheet models derived from Geotail particle data, *J. Geophys. Res.*, *108*, 1136, doi:10.1029/2002JA009707.

Vasyliunas, V. M. (1970), Mathematical models of magnetospheric convection and its coupling to the ionosphere, in *Particles and Fields in the Magnetosphere*, edited by B. M. McCormac, pp. 60–71, D. Reidel, Dordrecht.

Volland, H. (1973), A semiempirical model of large-scale magnetospheric electric fields, *J. Geophys. Res.*, *78*, 171, doi:10.1029/JA078i001p00171.

Weigel, R. S. (2010), Solar wind density influence on geomagnetic storm intensity, *J. Geophys. Res.*, *115*, A09201, doi:10.1029/2009JA015062.

Weimer, D. R. (2001), An improved model of ionospheric electric potentials including substorm perturbations and application to the GEM November 24, 1996 event, *J. Geophys. Res.*, *106*, 407–416, doi:10.1029/2000JA000604.

Weimer, D. R. (2005), Improved ionospheric electrodynamic models and application to calculating Joule heating rates, *J. Geophys. Res.*, *110*, A05306, doi:10.1029/2004JA010884.

Welling, D. T., V. K. Jordanova, S. Zaharia, A. Glocer, and G. Tóth (2011), The effects of dynamic ionospheric outflow on the ring current, *J. Geophys. Res.*, *116*, A009J19, doi:10.1029/2010JA015642.

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, doi:10.1016/j.jastp.2004.03.026.

Wolf, R. A. (1983), The quasi-static (slow-flow) region of the magnetosphere, in *Solar Terrestrial Physics*, edited by R. L. Carovillano, and J. M. Forbes, pp. 303–368, D. Reidel, Dordrecht.

Wolf, R. A., R. W. Spiro, and F. J. Rich (1991), Extension of convection modeling into the high-latitude ionosphere—some theoretical difficulties, *J. Atmos. Terr. Phys.*, *53*, 817–829, doi:10.1016/0021-9169(91)90096-P.

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. 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.

Yu, Y., V. Jordanova, S. Zaharia, J. Koller, J. Zhang, and L. M. Kistler (2011), Validation study of the magnetically self-consistent inner magnetosphere model RAM-SCB, *J. Geophys. Res.*, *117*, A03222, doi:10.1029/2011JA017321.

Zaharia, S., C. Z. Cheng, and K. Maezawa (2004), 3-D force-balanced magnetospheric configurations, *Ann. Geophys.*, *22*, 251–265.

Zaharia, S., V. K. Jordanova, M. F. Thomsen, and G. D. Reeves (2006), Self-consistent modeling of magnetic fields and plasmas in the inner magnetosphere: Application to a geomagnetic storm, *J. Geophys. Res.*, *111*, A11S14, doi:10.1029/2006JA011619.