Model evaluation guidelines for geomagnetic index predictions

Michael W. Liemohn[1], James P. McCollough,[2] Vania K. Jordanova,[3] Chigomezyo M. Ngwira,[4,5] Steven K. Morley,[3] Consuelo Cid,[6] W. Kent Tobiska,[7] Peter Wintoft,[8] Natalia Yu. Ganushkina,[1,9] Daniel T. Welling,[1,10] Suzy Bingham,[11] Michael A. Balikhin,[12] Hermann J. Opgenoorth,[12] Miles A. Engel,[3] Robert S. Weigel,[14] Howard J. Singer,[15] Dalia Buresova,[16] Sean Bruinsma,[17] Irina S. Zhelavskaya,[18, 19] Yuri Y. Shprits,[18, 19, 20] and Ruggero Vasile[18]

[1] Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI USA

[2]Space Vehicles Directorate, Air Force Research Laboratory, Kirtland AFB, NM USA

[3]Space Science and Applications, Los Alamos National Laboratory, Los Alamos, NM USA

[4]Department of Physics, The Catholic University of America, Washington, DC USA

[5]NASA Goddard Space Flight Center, Space Weather Laboratory, Greenbelt, MD USA

[6]Department of Physics and Mathematics, Uuniversity of Alcalá, Alcalá de Henares, Madrid, Spain

[7]Space Environment Technologies, Pacific Palisades, CA USA

[8]Swedish Institute of Space Physics, Lund, Sweden

[9]Finnish Meteorological Institute, Helsinki, Finland

[10]University of Texas at Arlington, Arlington, TX USA

[11]UK Met Office, Exeter, Devon, United Kingdom

[12]Department of Automatic Control and System Engineering, University of Sheffield, Sheffield, South Yorkshire UK

[13]Swedish Institute of Space Physics, Uppsala, Sweden

[14]Department of Physics and Astronomy, George Mason University, Fairfax, VA, USA

[15]Space Weather Prediction Center, National Oceanic and Atmospheric Administration, Boulder, CO USA

[16]Institute of Atmospheric Physics, CAS, Prague, Czech Republic

[17]Department of Space Geodesy CNES, Toulouse, France

[18]GFZ German Research Centre for Geosciences, Telegrafenberg, Potsdam, Germany

[19]Institute of Physics and Astronomy, University of Potsdam, Potsdam, Germany

[20]Department of Earth and Space Sciences, UCLA, Los Angeles, CA USA

Corresponding author: Michael Liemohn (liemohn@umich.edu)

Submitted to *Space Weather*

Special section on *Space Weather Capabilities Assessment: International Forum*

Key Points:

- We review existing practices for assessing geomagnetic index prediction models and recommend a "standard set" of metrics
- Along with fit performance metrics that use all data-model pairs in their formulas, event detection performance metrics are recommended
- Other aspects of metrics assessment best practices, limitations, uncertainties, and geomagnetic index caveats are also discussed

AGU Index Terms:

		Category
• 7924	Forecasting	in 7900 SPACE WEATHER
• 7959	Modeling	in 7900 SPACE WEATHER
• 7954	Magnetic storms	in 7900 SPACE WEATHER
• 4305	Space weather	in 4300 NATURLA HAZARDS
• 4318	Statistical analysis	in 4300 NATURAL HAZARDS

Keywords:

Space weather, geomagnetic indices, metrics, statistical analysis, forecasting, ROC curve

Abstract

Geomagnetic indices are convenient quantities that distill the complicated physics of some region or aspect of near-Earth space into a single parameter. Most of the best-known indices are calculated from ground-based magnetometer data sets, such as Dst, SYM-H, Kp, AE, AL, and PC. Many models have been created that predict the values of these indices, often using solar wind measurements upstream from Earth as the input variables to the calculation. This document reviews the current state of models that predict geomagnetic indices and the methods used to assess their ability to reproduce the target index time series. These existing methods are synthesized into a baseline collection of metrics for benchmarking a new or updated geomagnetic index prediction model. These methods fall into two categories: (1) fit performance metrics such as root mean square error (RMSE) and mean absolute error (MAE) that are applied to a time-series comparison of model output and observations; and (2) event detection performance metrics such as Heidke Skill Score and probability of detection (POD) that are derived from a contingency table that compares model and observation values exceeding (or not) a threshold value. A few examples of codes being used with this set of metrics are presented, and other aspects of metrics assessment best practices, limitations, and uncertainties are discussed, including several caveats to consider when using geomagnetic indices.

Plain Language Summary

One aspect of space weather is a magnetic signature across the surface of the Earth. The creation of this signal involves nonlinear interactions of electromagnetic forces on charged particles and can therefore be difficult to predict. The perturbations that space storms and other activity causes in some observation sets, however, are fairly regular in their pattern. Some of these measurements have been compiled together into a single value, a geomagnetic index. Several such indices exist, providing a global estimate of the activity in different parts of geospace. Models have been developed to predict the time series of these indices, and various statistical methods are used to assess their performance at reproducing the original index. Existing studies of geomagnetic indices, however, use different approaches to quantify the performance of the model. This document defines a standardized set of statistical analyses as a baseline set of comparison tools that are recommended to assess geomagnetic index prediction models. It also discusses best practices, limitations, uncertainties, and caveats to consider when conducting a model assessment.

1. Introduction

Geomagnetic indices are compilations of a set of similar measurements to produce a single parameter, a time series of the magnitude of disturbance in some part of geospace. They are highly convenient for distilling complicated phenomena down to an activity value, often being global in their integrative nature of the underlying physical processes. Because they are systematically calculated with well-known methodologies, they are comparable between events, even ones separated by decades. While the original motivation was summarizing observations and reducing data volume (e.g., Mayaud, 1980), they are now used as a proxy for some aspect of geomagnetic activity.

Most geomagnetic indices are derived from ground-based magnetometer observations. For instance, the polar cap index, PC (Troshichev et al., 1988; see also Stauning, 2013), is known

as an estimate of the electric field across the polar cap. The auroral electrojet indices AL and AU, and their difference, AE, are all distilled from a dozen or so high-latitude stations (e.g., Davis & Sugiura, 1966; see also Kamide & Kokubun, 1996; Gjerloev, 2012; Kamide & Rostoker, 2004), providing an estimate of the plasma flows and electric currents in this part of the ionosphere from the closure of field-aligned currents (region 1, region 2, or the substorm current wedge). Kp is derived from 3-h intervals of 13 midlatitude magnetometer measurements (Bartels et al., 1939) and is a measure of global geomagnetic activity. As Kp strongly responds to the motion of the inner edge of the plasma sheet, it is often used as an estimate of convection strength (e.g., Volland, 1975; see also Thomsen, 2004). Dst and SYM-H (Sugiura, 1964; see also Iyemori, 1990; Iyemori et al., 1992), often used interchangeably (see, e.g., the comparisons by Katus & Liemohn, 2013; Love & Gannon, 2009; Wanliss & Showalter, 2006), are derived from 4 to 10 low-latitude magnetometer stations, and is an index that captures the dynamics of inner magnetospheric current systems and large-scale magnetospheric currents. Please see the reviews by Rostoker (1972), Mayaud (1980), Murayama (1982), and Menvielle et al. (2011) for a complete description of geomagnetic indices. Dst if often used to define the geomagnetic storms and their phases (Sugiura & Chapman, 1960), while other indices or combinations of them may be used depending on the considered magnetospheric phenomenon (Borovsky, 2014; Borovsky & Shprits, 2017).

Because of their convenience as a single time series, these indices are often cited as measures of space weather activity. In fact, they are regularly used as input values to drive some numerical models. For example, Kp has been used in several different ways as an input to inner magnetosphere models, such as for the large-scale electric field description in drift physics models (e.g., Maynard & Chen, 1975; see also Fok et al., 1995; Jordanova et al., 1996; Liemohn et al., 1999, 2001; Ganushkina et al., 2001), for plasmapause locations (e.g., Carpenter & Anderson, 1992; see also Moldwin et al., 2002; O'Brien & Moldwin, 2003), for ULF wave activity (e.g., Brautigam & Albert, 2000; Brautigam et al., 2005; Ozeke et al, 2014), and whistler mode chorus and hiss wave activity (e.g., Agapitov et al., 2015; Orlova et al., 2014, 2016; Spasojevic et al., 2015). Even though they are used as a crude proxy to unmodeled physical processes, they are part of our understanding of space physics and an integral aspect of space weather modeling and forecasting.

Much time and effort has been devoted to the prediction of geomagnetic indices. The output of each new model is, of course, tested against an index for one or more intervals. These studies, however, take different approaches to that validation task. That is, while many papers have assessed the performance of a given geomagnetic index prediction model, there is no standard for this assessment. It is proposed here to establish a baseline set of statistical analysis metrics for benchmarking a geomagnetic index prediction model. This metrics set will be useful from both scientific and operational perspectives. For science, it will be useful for assessing model capabilities and identifying where and under what circumstances model improvements are needed. For operations, it will be useful for assessing model skill for serving those affected by space weather conditions.

In early 2017, the space weather community organized into working groups to address this issue of metrics for space weather models. This effort culminated in a CCMC-LWS workshop in April 2017 (Community Coordinated Modeling Center – Living With a Star International Forum for Space Weather Capabilities), at which many hours of discussion led to community consensus on various issues of space weather forecasting capabilities (see the

assessment website). One of the working groups focused on the metrics related to geomagnetic indices. This document presents the output of that working group, presenting a review of existing geomagnetic index models, a baseline set of metrics for assessing new or updated index models, and a few examples of this statistical toolkit applied to geomagnetic index prediction models. Many acronyms are used throughout this paper and a full list of definitions is provided in Table S1.

2. Prior Assessment of Index Prediction Models

There are essentially three main groupings of "global" geomagnetic indices from ground-based magnetometers. The first set is the low-latitude indices, specifically Dst and SYM-H, responding to the large-scale current systems in geospace. The second class is the mid-latitude Kp index, in a class by itself because it is a unique index with a distinct calculation scheme, yet has been demonstrated to be useful as an organizer of geomagnetic activity. The third category is the high-latitude indices, most notably AL, AU, and AE, which are measures of ionospheric current systems in the auroral region. In the following subsections, the history of models that predict these indices is briefly presented and discussed.

2.1. Dst and SYM-H

Table 1 lists the studies, grouped by model in order of the year of their first publication, that included a predictive model for Dst/SYM-H and a quantitative assessment of the accuracy of the comparison. The second column gives a very brief description of the numerical approach used to calculate the index and the third column lists some of the key metrics discussed in the papers for the model performance in reproducing one of these indices. In this last column, HWHM is half-width at half maximum of a distribution of data-model differences, R is the Pearson linear correlation coefficient, RMSE is the root mean square error, ARV is the average relative variance, PE is prediction efficiency, HSS is Heidke Skill Score, NRMSE is normalized root mean square error, POD is probability of detection, and ME is mean error.

One of the first studies to predict the low-latitude magnetic disturbance was Burton et al. (1975), who didn't actually predict Dst but a similarly comprised collection of magnetometer signals from around the globe. We refer to the ordinary differential equation they adopted as the Burton Equation, and numerous other prediction schemes have followed this methodology (Fenrich & Luhmann, 1998; O'Brien & McPherron, 2000a; Temerin & Li, 2002; Wang et al., 2003). A more advanced version of this approach was presented by Horton and Doxas (1996), who expanded it to a full "circuit diagram" set of 8 differential equations. Two of the ouputs from this model are analogous to Dst/SYM-H and AL, and have been successfully used to predict these indices (e.g., Mays et al., 2009). Also in this category is the severe space weather event determination model of Balan et al. (2017), who based their model on the same solar wind input parameters as were used in the Burton Equation.

Neural networks have been used for Dst/SYM-H prediction. This is a broad category and there are several different algorithms within this category. For example, both Lundstedt and Wintoft (1994) and Bala et al. (2009) used a time delay neural network algorithm while Wu and Lundstedt (1997) adopted the Elman neural network approach. Revallo et al. (2014) also used the Elman neural network method but instead of feeding the solar wind values straight into the code, they filtered them first with a time-integrative function. As seen in the metrics column of Table 1, most of these approaches are very good at reproducing indices.

188 Table 1. Dst and SYM-H prediction models and key metrics of the comparison

References	Description	Select Metrics and Results
Burton et al. (1975), Lindsay et al. (1999), O'Brien and McPherron (2000b)	The Burton equation, $dDst^*/dt = - Q + Dst^*/\tau$	HWHM Error = 18 nT
Lundstedt and Wintoft (1994)	Time delay neural network, inputs of B, n, and v	Qualitative comparisons against several storm intervals
Wu and Lundstedt (1997), Lundstedt et al (2002)	One-layer Elman neutral network, inputs of B, n, and v	R=0.88, HWHM Error = 11 nT
Horton and Doxas (1996, 1998), Spencer et al. (2007, 2009), Mays et al (2009)	WINDMI model, low-dimensional (8 differential equations) description of geospace, predicts Dst	ARV=0.54, R=0.80, RMSE=9.8 nT
Fenrich and Luhmann (1998)	Burton-equation model with SW P and Ey	HWHM Error = 17 nT
Klimas et al. (1998)	Local-linear autoregressive moving average method	R=0.80, RMSE=22 nT
O'Brien and McPherron (2000a, 2000b)	Updated Burton equation model with variable loss lifetime	Single-step: PE=0.97, HSS=0.37; Multi-step: PE=0.88, 20 nT; Real-time: HWHM Error = 18 nT
Boaghe et al. (2001), Wei et al 2004, 2007, and Boynton et al. (2011)	NARMAX model orthogonal least squares-error reduction ratio method	Single-step: R=0.99, NRMSE=0.14; Multi-step: R=0.84, NRMSE=0.34
Temerin and Li (2002, 2006)	Triple Burton-equation model with dozens of free parameters	R=0.96, PE=0.91, RMSE=6.7 nT
Liemohn and Jazowski (2008), Liemohn et al. (2010), Liemohn and Katus (2012)	HEIDI modeling of all intense storms from solar cycle 23,	Dst_min: R=0.70; all SYMH CME: R=0.85, RMSE= 29 nT; all SYMH CIR: R=0.71, RMSE=43 nT
Saiz et al. (2008)	Dst predictor with just IMF Bz south magnitude and duration	Intense storms: POD=0.24 to .48; Moderate storms: POD=0.52
Tsubouchi and Kubo (2010)	Probabilistic Dst prediction model based on waiting times between storms	Observed frequency and forecast probability close to unity slope
Bala et al (2009) and Bala and Reiff (2012, 2014)	Artificial neural network scheme	6-h lead-time: R-0.80, RMSE=10.3 nT
Rastätter et al (2013)	Comparison of 30 different models against Dst	PE, log spectral distance, R, modeling yield, and timing error
Tobiska et al (2013)	Dst prediction using the Anemomilos solar flare-Dst correlation method	Now-mean: R=0.995; 3-day forecast: R=0.6
Revallo et al. (2014)	Neural network algorithm	R=0.74, PE = 0.44
Zhang and Moldwin (2015)	Probabilistic forecast of SYM-H based on previous 12 hours of Dst values	Cumulative probability distributions for ICME, SIR, and Alfvenic SW inputs
Balan et al. (2017)	Severe Dst prediction scheme based on $\Delta VxBz$ threshold	Nearly 100% success for Dst < -200 nT storms
Haiducek et al. (2017)	SYM-H prediction from SWMF for all of Jan 2005	R=0.84, RMSE= 17 nT, ME=4 nT
Liemohn et al. (2018)	Dst prediction from SWMF in real-time mode	R=0.69, PE=0.41, HSS=0.57, RMSE=13 nT
Morley et al. (2018)	SYM-H prediction for the 5 April 2010 storm from an ensemble run varying solar wind input	Probability distributions of MAE, ME, and RMSE

189

Another numerical approach is the autoregressive moving average model of Billings and Voon (1986), of which the NARMAX version of this technique (Nonlinear Autoregressive Moving Average Model With Exogenous Inputs) was applied to predict geospace indices like Dst by Boaghe et al. (2001). This uses an equation set of specified combinations of the input variables, back one or more time intervals (again, specified), and then iteratively determines the optimal coefficients for each term. The initial equation can have dozens of free parameters but, usually, there are only a few dominant terms in the final model. A related method is that of Klimas et al. (1998), who used a local-linear prediction analogue method to forecast Dst.

With the Gonzalez et al. (1994) classification of driver parameters for storms, as defined by Dst, models have been developed that predict Dst active times with these criteria. Saiz et al. (2008) employed several modified versions of the Gonzalez et al. (1994) thresholds, and Zhang and Moldwin (2015) created a probabilistic forecast technique for activity. Tsubouchi and Kubo (2010) also used these criteria to determine storm start and end times, then developing a probabilistic forecasting model for when the next storm should occur. Not only the occurrence or the severity of a storm was considered as relevant in the forecasting process, but also the remaining time for quiet state after a storm. This phase was commonly modeled as an exponential recovery, but during severe storms Dst often recovers faster (e.g., Dasso et al., 2002; Liemohn & Kozyra, 2005). The model of Aguado et al (2010) proposed an analytical expression for the recovery phase of intense storms based in a hyperbolic function.

A rather different approach is the Anemomilos method of Tobiska et al. (2013). This technique correlated solar flare intensity and location of the flare on the solar disk to the average Dst perturbations up to a few days later. Because many intense storms are driven by interplanetary coronal mass ejecta launched from the Sun along with a flare (e.g., Zhang et al., 2007), this simplistic method works quite well at capturing the daily mean changes of Dst.

A final group of modeling approaches to be mentioned here are the first-principles-based numerical models of geospace that compute a synthetic Dst/SYM-H time series. These include regional models, such as the Hot Electron and Ion Drift Integrator code (e.g., Liemohn et al., 2004; Ilie et al., 2012) that solves the gyration- and bounce-averaged kinetic equation for the phase space density of hot (~keV) charged particles in the inner magnetosphere. HEIDI has been run for all of the intense storms of solar cycle 23 (1996-2005), from which comparative metrics have been calculated (e.g., Liemohn & Jazowski, 2008, Liemohn & Katus, 2012). There are several other models like HEIDI that also calculate Dst/SYM-H from an integral of the particle phase space densities (e.g., Jordanova et al., 1998; Khazanov et al., 2003; Ganushkina et al., 2012; Fok et al., 2014), but the Dst values from these codes have only been qualitatively compared against the observed values. Another approach is with a set of coupled codes, such as the Space Weather Modeling Framework (SWMF, see Toth et al., 2012), that includes a magnetohydrodynamic model for the global magnetospheric structure, an inner magnetospheric drift physics model, and an ionospheric electrodynamics solver. Haiducek et al. (2017) used this code to simulate the entire month of January 2005, conducting a set of metrics comparisons against SYM-H, AE, and Kp as calculated from the SWMF model suite. Similarly, Liemohn et al. (2018) have assessed the output from the experimental real-time SWMF simulations being run at the Community Coordinated Modeling Center (CCMC), for which are now several years of output available. Yet another study of this kind is Morley et al. (2018), who varied upstream inputs to the SWMF to assess ground-based magnetometer comparisons with respect to solar wind uncertainties. These first-principles codes are, in general, not as good at reproducing the

low-latitude index time series as the other codes mentioned above, which are especially formulated and optimized for index prediction. They produce a far richer output set, however, that includes plasma and field parameters across a large spatial domain.

Note that a "Dst challenge" was conducted by CCMC (Rastätter et al., 2013) as part of the 2008–2009 GEM Metrics Challenge. They presented results of 30 different model configurations for four storm events (ranging from a minor storm to a super storm). Specifically, these models were: 1) three-dimensional (3-D) MHD models of the magnetosphere coupled to an ionosphere electrodynamics solver such as the SWMF (Tóth et al. 2005), the Open Geospace General Circulation Model (OpenGGCM) (Raeder et al., 2001), and the Coupled Magnetosphere-Ionosphere-Thermosphere (CMIT) model (Lyon et al., 2004; Wiltberger et al., 2004; Merkin and Lyon, 2010); 2) kinetic ring current models such as the Ring Current-Atmosphere Interactions Model with Self-Consistent Magnetic Field (RAM-SCB) (Jordanova et al., 1994; 2010; Zaharia et al., 2006) and the Rice Convection Model (RCM) (Harel et al., 1981; Wolf et al., 1991; Toffoletto et al., 2003); and 3) Dst-specification models such as the Impulse Response Function with 96 lags (IRF96) of Weigel (2010), an analytic formula called BFM (Burton et al. 1975; Feldstein 1992; Murayama 1982); and the University of Sheffield (NARMAX) algorithm (Billings et al., 1989). Rastätter et al. (2013) considered a number of different metrics, including prediction efficiency (PE), log spectral distance, correlation coefficient (R), modeling yield, and timing error. Different models and settings performed the best in each of these categories. To visualize the model performance, the scores for each run for the individual events were shown in 2-D plots (i.e., PE - R) space). It was found that the magnetosphere model runs filled a large area in PE-R space ($PE > -11$, $R > -0.15$), while most ring current model runs were clustered much closer to the ideal PE score ($PE > -2$) with a smaller range in R ($R > 0.2$). The Dst specification models were very close to perfect in PE and R except for the weakest, isolated-substorm event that proved difficult for all the models. Model outputs from this study, together with the observational data, are available on the CCMC web site (http://ccmc.gsfc.nasa.gov, under “Metrics and Validation” and then “GEM Challenge”).

The metrics quoted in Table 1 are not always directly comparable because the studies might have used different forecast windows for the comparison. Some are nowcast or even historical event reanalysis studies, others are one time step ahead, while some studies predict the index up to days in the future. In particular, models that include past observed values of the predictand will result in high scores for most performance metrics for one-step-ahead predictions if the auto-correlation is high, like for Dst. Therefore, caution should be taken in reading Table 1 and making judgments about the performance of any particular model.

2.2. Kp

Table 2 lists studies that have presented models reproducing the Kp index. The list of such models is significantly shorter than that for Dst/SYM-H. As in Table 1, the second column gives a brief description of the numerical approach and the third column lists some key metrics from the comparison. There are a few new metrics in this table that were not used in Table 1. Specifically, Gilbert SS is the Gilbert Skill Score, MAE is the mean absolute error, FAR is the false alarm ratio, and TSS is the True Skill Score.

Like for Dst, neural networks have been used for Kp models. Boberg et al. (2000) used a neural network with time delays, the Wing et al. (2005) model used two methods, the

multilayered feedforward network and a recurrent network, and Bala et al. (2009) used a feedforward neural network. These are among the best at predicting Kp several hours ahead.

Another type of model is to use a small number, perhaps even just one, ground-based magnetometer station to nowcast the global Kp index value. This was done by Takahashi et al. (2001), finding high correlation values even for a prediction based on a single station.

A version of the NARMAX model has been applied to the Kp index by Ayala Solares et al. (2016). They found that the simplified version, the NARX model, without the moving average input values but rather with direct input of single-time solar wind values, performed slightly better for Kp than the NARMAX version of the code.

Table 2. Models predicting Kp and key metrics of the comparison

References	Description	Metrics
Boberg et al. (2000)	Time delay neural network	RMSE=0.98, R=0.77
Takahashi et al (2001)	Kp estimation from one or several individual station values	Single station: R between 0.85 and 0.9; 9 stations: R=0.94
Wing et al (2005)	Feedforward backpropagation and recurrent neural network prediction schemes	R=0.94, Gilbert SS=0.2-0.5 for Kp 2 through 6, depending on year
Bala et al (2009) and Bala and Reiff (2012, 2014)	Feedforward backpropagation neural network scheme	3-h lead-time: R=0.77, RMSE=0.8, HSS for KP>6=0.964
Devos et al. (2014)	Prediction of local K-index from Chambon-la-Forêt	R=0.53, ME~0, MAE=0.3, HSS=0.52
Ayala Solares et al. (2016)	Kp with NARX, with both a "sliding window" and a "direct approach" for the input values	3-h ahead: RMSE=0.76, R=0.87, PE=0.76; 24-h ahead: RMSE=0.87, R=0.83, PE=0.68
Wintoft et al. (2017)	Ensemble of time delay neural networks	RMSE=055, R=0.92 (function of year and Kp)
Savani et al. (2017)	Kp prediction from predicted solar wind based on a coupling function empirical formula	POD=0.67, FAR=0, TS=0.6, TSS=0.6
Haiducek et al. (2017)	Kp prediction from SWMF for all of Jan 2005	RMSE=1.1, ME=0.7

Another approach is an empirical model for the relationship between Kp and solar wind input values. This was done by Savani et al. (2017), who coupled the output from a solar wind prediction model to this Kp prediction formula. The model does reasonably well at capturing high-Kp space weather events, with no false alarms in their test interval.

There is one first-principles model that has produced Kp and for which metrics have been calculated, the SWMF. For the month of January 2005, Haiducek et al. (2017) assessed the ability of three versions of the SWMF to reproduce Kp.

As with Dst/SYM-H, the metrics listed in Table 2 might not be directly comparable with each other. Some of the studies are historical reanalysis assessments, others are nowcasts, and

300 the prediction models could be one time step (3 hours, in the case of Kp) or more. Care should be
301 taken in judging one model against another in this table.

302 ### 2.3. AE, AL, and AU

303 Table 3 lists the studies that have produced a model for predicting the high-latitude
304 indices of AE, AL, and/or AU.

305

306 Table 3. Models that predict AE, AL, or AU and their key metrics

References	Description	Metrics
Clauer et al. (1981)	Linear inpulse response function for AL and AU	PE histogram peaks at 0.6 for AL, 0.3 for AU
Baker et al. (1981)	Correlating AE with espilon and VBs	R=0.54 for ε, R= 0.60 for VBs
Holzer and Slavin (1982)	Time-integral of SW VxB with AL	R=0.97 for BsxV^2, 0.92 for BsxV, and 0.82 for Bs^2xV
Goertz et al (1993)	AL predictor based on magnetotail electron data	For a 2-day interval, R>0.9
Horton and Doxas (1996, 1998), Spencer et al. (2007, 2009), Mays et al (2009)	WINDMI model, low-dimensional (8 differential equations) description of geospace, predicts AL	ARV=0.41, RMSE=111 nT, R=0.64
Takalo and Timonen (1997)	Backpropagation neural network precition of AE	2-minutes ahead: NMSE=0.04, R=0.98; 1-h ahead: NMSE=0.56
Gleisner and Lundstedt (2001)	One-layer Elman neutral network for AE	R^2=0.7, RMSE=184 nT
Morley et al (2007)	AL magnitude distribution from the Minimal Substorm Model	Cumulative distribution of AL matches observations
Li et al (2007); Luo et al. (2013)	Empirical model for AU, AL, and AE with dozens of pree parameters	AU: R=0.85, PE=0.72, RMSE=39 nT; AL: R=0.85, PE=0.72, RMSE=82 nT; AE: R=0.89, PE=0.79, RMSE=96 nT
Pallocchia et al. (2008)	AE prediction with an Elman artificial neural network	NRMSE=0.4 for AE between 400 and 1000 nT
Amariutei and Ganushkina (2012)	ARMAX model prediction for AL	1-minute ahead: PE=0.98, NRMSE=0.11; 1-h ahead: PE=0.63, NRMSE=0.61
Bala et al (2009) and Bala and Reiff (2012, 2014)	Artificial neural network scheme	3-h lead-time: R=0.75, RMSE=113 nT
Zhang and Moldwin (2015)	Probabilistic forecast of AE based on previous 12 hours of Dst values	Cumulative probability distributions for ICME, SIR, and Alfvenic SW inputs
Haiducek et al. (2017)	AL prediction from SWMF for all of Jan 2005	RMSE=230 nT, ME=15 nT

307

308

There were a number of prediction algorithms created for these indices in the early 1980s. Clauer et al. (1981) used a linear impulse response function for AL and AU, Baker et al. (1981) correlated AE against two solar wind coupling functions, and Holzer and Slavin (1982) compared time-integrals of the solar wind coupling functions with AL. This last study produced the largest correlation coefficients, indicating that an hour or two of integrated input is all that is needed to accurately predict this index.

Goertz et al. (1993) created an AL prediction model from magnetotail observations. While they only tested it on a small interval, the correlation was high, indicating that such measurements have potential for the prediction of this index.

Several of the models mentioned above also predict one or more of these indices. The 8-differential-equation model of Horton and Doxas (1996) produces a output that can be considered a synthetic AL value. Gleisner and Lundstedt (2001) adopted their neural network model for AE prediction, Bala et al. (2009) used their neural net for AE forecasts, Amariutei and Ganushkina (2012) used the ARMAX model for predicting AL, Zhang and Moldwin (2015) included AE in their probabilistic forecast of geomagnetic activity, and Haiducek et al. (2017) computed AL from the SWMF model results.

There are a few similarly formulated but different models listed in Table 3. Two of these include neural network approaches for AE (Takalo & Timonen, 1997; Pallocchia et al., 2008). Another is the Minimal Substorm Model (Morley et al., 2007), which calculates AL based on solar wind inputs distilled into two components, an unloading DP1 portion and a directly-driven DP2 part. Finally, there is the AL prediction model of Li et al. (2007), which is based on the Temerin and Li (2002) Dst prediction model approach.

As with the other tables, the metrics listed in Table 3 might not be directly comparable against each other. Caution is advised in assessing one model against another based on the listings in Table 3.

2.4. Other Indices

Prediction methods have also been developed for a few other geomagnetic indices that do not fit into the three categories listed above. For example, both Cade et al. (1995) and Shen et al. (2002) calculated a relationship between Dst and AL/AE, finding relatively high correlation between these indices. Boyle et al. (1997) developed a prediction scheme for the cross polar cap difference of the ionospheric electric potential, basing it on solar wind input values. Borovsky (2014) used canonical correlation analysis for geospace system prediction. This uses several geospace system parameters, including Kp and SYM-H, and several solar wind input parameters, to determine a set of best-fit linear combinations of both the solar wind input and the geospace output. A solar parameter, the F10.7 solar radio flux, is regularly used as a proxy for the extreme ultraviolet photon flux from the Sun to the Earth. It is especially useful for the ionosphere-thermosphere research community, and Henney et al. (2012) developed an F10.7 prediction scheme that yields forecasts up to 7 days in advance.

Some of the studies mentioned above also calculated other geomagnetic indices and computed data-model comparison metrics. Specifically, along with Kp, Devos et al. (2014) includes a prediction algorithm for F10.7. Using the SWMF model suite, Haiducek et al. (2017) simulated the northern and southern hemisphere cross polar cap potential and compared with an observation-based estimate of this value.

3. The Baseline Assessment Metrics

As seen from the above-listed studies, there is no single set of metrics used by geomagnetic index predictive-model developers to benchmark their codes. Model verification and validation is an important aspect of development; Jolliffe and Stephenson (2012) give three main reasons for conducting quantitative assessments of models. The first is administrative – documenting the improvement of modeling capabilities over time. The second is economic – users of models want to optimize the return on their product development investment and offer the best service (in this case, predictions of various aspects of geomagnetic activity, as captured by indices) to their clients. The third is scientific – understanding the input conditions and expected output values for which a model has high or low performance capabilities reveals strengths and weaknesses of the underlying methodology, and possibly also about the physical processes governing index response.

Because the model is producing an output that, ideally, should exactly match an observed index time series, the Pearson correlation coefficient has been used extensively. One metric alone, however, is not enough to assess the accuracy of a model, especially given the fact that different users of the same model might want different performance capabilities and standards (e.g., Halford et al., 2018). For an index predictor, the general desire of both the model developer and user is an improvement of the existing model's performance. The modeler, however, has made choices in creating the prediction scheme: what input parameters to use, what functional form to assume for the causal relationship, what statistical methods to employ to get coefficients, even what time intervals to use for training and validation. For example, the user of a model's prediction may care about one or more of the following: its ability to predict extreme events; its long-lead-time forecasting ability; its accuracy for reanalysis of past events; or its ability to minimize false alarms. That is, each user will want a model that works for a particular comparison at an accuracy standard they have specified.

Here, we define and describe a standard list of statistical analysis metrics that is recommended for any geomagnetic index prediction model. While this is a limited and tractable set, it covers a broad range of possible metrics choices (see, for instance, Hogan and Mason, 2012; Morley et al., 2018). Each one has been selected because it assesses a certain aspect of the data-model comparison. Note that this is a minimum set for everyone to use; additional statistics can and should be used depending on the specific application for which the model is being developed.

The baseline set of metrics proposed here is a combination of two categories of values. The first set consists of "fit performance assessments" that include each data-model pair in the considered time interval. The second set is the "event performance assessments" that measure how well the model reproduces the timing and intensity of geomagnetic activity across a range of thresholds.

3.1. Fit performance metrics

The metrics in this category are as follows: linear fit parameters of intercept and slope, A and B; R, the Pearson correlation coefficient; root mean square error, RMSE; mean absolute error, MAE; the mean error, ME; and the prediction efficiency, PE. The modeled and observed time series are one-dimensional comparisons that do not require specialized multi-dimensional

394 comparison algorithms. Let us quickly define each of these and defend their selection in the
395 baseline set.

396 Because the model (M) is predicting an observed index (O), the relationship should be
397 linear and thus the intercept (A) and slope (B) are direct measures of the performance of the
398 model. While the relationship should be checked visually by plotting M versus O, the equation of
399 interest is this:

400 $$M_i = A + B \times O_i \tag{1}$$

401 and nearly all calculation software includes functions for computing the *A* and *B* coefficients. A
402 perfect prediction should have a zero offset and a unity slope. The offset *A* reveals a model bias
403 at the lowest observational values (specifically, when the observational value is zero) and the
404 slope *B* quantifies whether the trend of the model results with increasing observational values
405 keeps pace with the observed increase or under- or overshoots it. Uncertainties can and should be
406 calculated on *A* and *B* (e.g., Taylor, 1997, Chapter 8; Sheskin, 2007, pp. 1241-1243), like this:

407 $$s_A = s_M \sqrt{\frac{\sum O_i^2}{N \sum O_i^2 - \left(\sum O_i\right)^2}} \qquad s_B = s_M \sqrt{\frac{N}{N \sum O_i^2 - \left(\sum O_i\right)^2}} \tag{2}$$

408 where σ_M is the standard deviation of the model values and N is the number of data-model pairs.
409 These are often converted to fractional or percent uncertainties with a division by *A* and *B*,
410 respectively. Note that these uncertainty values in equation (2) assume that the error distribution
411 is Gaussian and that each error source is independent. If this is not the case, then a bootstrap
412 method (e.g., Reiff, 1990) can be used by randomly selecting a subset of data, calculating *A* and
413 B, and repeating this hundreds of times to generate a distribution of *A* and *B* values, from which
414 a spread can be calculated.

415 The Pearson linear correlation coefficient, *R*, is commonly used to indicate how well the
416 model predicts the trends of the index. It is calculated as the data-model covariance divided by
417 the standard deviations of each set:

418 $$R = \frac{\mathrm{cov}\left(M,O\right)}{s_M s_O} \tag{2}$$

419 The value ranges between -1 and 1, which indicate perfect anticorrelation or correlation where
420 all of the data-model pairs lie along a straight line. The significance of an *R* value is dependent
421 on N, with a probability of an *R* value occurring by chance of less than 0.05 being called
422 significant and a probability less than 0.01 called highly significant. The significance of this
423 probability statistic is necessary but not sufficient for a high-quality linear fit, because for large
424 N these probabilities are met even for *R* values close to zero. In addition to the significance
425 check, *R* should also be above a user-defined threshold that means the specified requirement for
426 the application. This is usually at least 0.5, perhaps even 0.7 or even 0.9, to convince users that
427 the model is performing well.

428 The next two metrics, root mean square error, RMSE, and mean absolute error, MAE,
429 reveal how well the model captures the range of values of the index. The RMSE (e.g., Wilks,
430 2006, chapter 8) is

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(M_i - O_i\right)^2} \tag{3}$$

Because of the square term inside the summation, highlights the data-model pairs with larger differences contribute more to this error than in the MAE. This is often during active times when the index and, presumably, the error, are farther from zero than during quiet times. The MAE, however, defined as

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|M_i - O_i\right| \tag{4}$$

does not include this square term and therefore emphasizes the "usual state" of the index (i.e., no extra weighting to the active times). Note that the MAE is sometimes referred to as the absolute relative error, ARE. Each reveals something important about the data-model comparison, one weighting the active times when the errors are often larger and the other weighting the quiet times for which there are usually far more data-model pairs. Depending on the user's final application of the model, either of these could be the more valuable metric.

Another to go along with these two is the mean error, ME, which is a difference of the means of the observed and modeled values, including the sign (e.g., Wilks, 2006, chapter 8):

$$ME = \frac{1}{N}\sum_{i=1}^{N}\left(M_i - O_i\right) \tag{5}$$

This tells you the bias between the two number sets. ME above zero shows that, on average, the model overpredicts the data, while a score below zero shows that the model underpredicts the observed values, on average.

The final metric in this set is the prediction efficiency, PE (this the "Case I" skill score considered by Murphy, 1988). Skill scores are defined by comparing the model against a specified reference forecast. In the case of prediction efficiency, the reference model is the average of the data:

$$PE = 1 - \frac{\sum_{i=1}^{N}\left(M_i - O_i\right)^2}{\sum_{i=1}^{N}\left(O_i - \bar{O}\right)^2} \tag{6}$$

The PE is related to the Average Relative Variance (ARV) by PE=1-ARV, and the ARV represents the fraction of the variance in the data that is predicted by the model. Because active times create a long one-sided tail on the measured geomagnetic indices, models that are even somewhat capable of reproducing this activity will have a positive PE score. That said, the PE quantifies the model's overall accuracy at reproducing the time variation of the observed index, weighting the active times more heavily than the quiet times in this assessment.

In addition, the PE formula is valuable because the observational mean in the denominator can be swapped out for any reference model time series. It is no longer PE at that point, which has the specific meaning of defining the model's capability relative to the observed

climatological average. Rather, when this value is swapped out, equation (6) becomes the prediction efficiency relative to an existing modeling capability (Murphy, 1988).

It is useful to discuss normalization of the above-mentioned quantities. Some of these metrics, in particular RMSE, MAE, and ME, are reported not as calculated above but instead as a value relative to a parameter of observed index set. The common choices for normalization are the observed index mean or standard deviation, but it could also be a function of the data range, such as the median, the interquartile distance (the 75% quartile value minus the 25% quartile value), or even the full range of the data (maximum minus minimum value). This kind of normalization puts the metric in the context of the observed values in the chosen interval. If the data span the typical range of the index values, then this extra calculation is not particularly helpful. Normalization is sometimes useful, however, when the observed values cover an exceptionally large or unusually small range of index values. When this is the case, then normalization can help put the data-model comparisons in the proper perspective.

3.2. Event detection performance metrics

Across a month, year, or solar cycle, the time series of a geomagnetic index value is far more often near the quiet-time average than perturbed into an active state. For example, for Dst, a histogram of values shows that only 5% are below -50 nT. That is, storm intervals are a small part of the total database and so the quiet time state dominates the curve fitting, including for data-model comparison metrics in the previous subsection. It is the active state intervals, however, that are often the times when users want a model to perform well. In fact, the user might not care about the quantitative difference between the modeled and observed values, as long as the model output indicated that an event was occurring. An analysis based on when the data and/or model values have reached an active state, therefore, overcomes the issue of quiet time dominance in the fit performance statistics.

This type of assessment is called event detection performance and is based on the formation of a contingency table. By defining an index value as an "event threshold," both the observed and model index time series can be compared against this threshold to determine if either was in an event state. Sometimes these are considered at the highest time cadence available and other times the event state determination is done over a longer window of time, checking for event status among a set of values, declaring event detection if one of the values is beyond the threshold (or, depending on the application, some proportion of the values). The value pairs (or windows) are then classified as hits, misses, false alarms, and correct negatives (defined here as H, M, F, and N, respectively). These are also called, in the same order, true positives, false negatives, false positives, and true negatives. Various quantities can be calculated from these, and a few of these quantities have been selected as the baseline set for geomagnetic index model assessment. Specifically, the chosen metrics are the Heidke Skill Score, the Probability of Detection, the Probability of False Detection, the False Alarm Ratio, and the Contingency Table Bias. By varying the threshold from a very low/quiet value to a very high/disturbed value, you get a set of scores for each of these quantities, which reveal how well the model captures the "events" in the observed index across a wide range of "event" definition.

The Heidke Skill Score (HSS), from *Heidke* (1926), condenses the entire contingency table into a single measure of the performance with the exclusion of predictions from random chance:

$$HSS = \frac{2\left[\left(H \cdot N\right) - \left(M \cdot F\right)\right]}{\left[\left(H + M\right)\left(M + N\right) + \left(H + F\right)\left(F + N\right)\right]} \tag{7}$$

HSS has a perfect score of one, when all of the values are either hits or correct negatives (i.e., when M=F=0). Values of zero or below indicate that the model has no skill in predicting events of that threshold. The lowest value for HSS is -1, which occurs when no time is correctly modeled and all of the values are evenly distributed between misses and false alarms (i.e., when H=N=0 and M=F). While there are several other contingency table skill scores available, this one has many useful features (see, e.g., Hogan and Mason, 2012). First off, it is truly equitable, meaning that a random forecast or constant forecast will have a score of zero. It also has the added benefits of being bounded, linear, and transpose symmetric. Finally, it is devised so that a biased model cannot obtain a perfect score. In short, it is a commonly used distillation of the entire contingency table into a meaningful single quantity.

The next is the Probability of Detection (POD). POD, only using half of the contingency table, gives the fraction of observed events that were captured by the model:

$$POD = \frac{H}{H + M} \tag{8}$$

It is sometimes referred to as the hit rate. POD ranges from 0 to 1, with higher values being better. If the user is concerned about reproducing all of the real events, then POD is the quantity to maximize.

A related metric is the Probability of False Detection (POFD), which uses the other half of the contingency table. It gives the fraction of the times when the observed index was not in the event state but the model was in event state.

$$POFD = \frac{F}{F + N} \tag{9}$$

Like POD, POFD ranges from 0 to 1. Because F is one of the two off-diagonal table entries that represent a poor prediction, low POFD numbers are better. If the user is concerned about never "crying wolf" then POFD model development should focus on minimizing this parameter. POFD is sometimes called the false alarm rate, but that name will not be used because it has the same acronym as the false alarm ratio, to be discussed next. A related metric used in the space weather literature is the forecast ratio, R_F, which is simply the ratio of hits to false alarms (Weigel et al., 2006); this metric is intended for users interested in the economic utility of a forecast and is related to the value score (Wilks, 2001).

A metric that combines these two terms but still uses only half of the contingency table is the False Alarm Ratio (FAR). It is defined like this:

$$FAR = \frac{F}{F + H} \tag{10}$$

Like POFD, it ranges from 0 to 1 with values near zero being better. Because N can be quite large for geomagnetic indices, which spend a lot of time at quiet levels and only occasionally exhibit excursions to active values, the FAR highlights the false alarms relative to the correct hits rather than the correct negatives. The denominator is often much smaller for the FAR compared

to the POFD, so this value is usually the larger of the two. Designing a model to minimize POFD will also minimize FAR.

A final metric to discuss here is the Frequency Bias (FB). FB, or sometimes just "bias," is defined like this:

$$FB = \frac{H + F}{H + M} \tag{11}$$

It is a measure of the contingency table that ranges from zero (no model values classified as events) to infinity (no data values classified as events). FB values above one show that false alarms are more prevalent, indicating the model overpredicts the data for this threshold, while values under 1 shows that misses are more prevalent, revealing that the model underpredicts the observed values for a given threshold. FB does not yield any information about the skill of the model for the given threshold, rather it quantifies the diagonal asymmetry of the contingency table.

For all of the parameters discussed above, they should be calculated not just for a single activity threshold choice but for a range of threshold values. This will reveal the model performance at capturing any kind of event interval, whether the threshold is a low, medium, or high one. Calculating at least 10 thresholds yields a curve that quantifies this relationship for each of the metrics described here.

Another plot that is part of the baseline set of calculations to perform as part of the event detection assessment is the Receiver Operating Characteristic (ROC) curve (ROC can also stand for relative operating characteristic), first used in Britain in 1940 by the Royal Air Force for radar signal processing (Carter et al., 2016). The ROC curve plots POD on the y axis and POFD on the x axis, for all threshold values. The unity slope line represents no skill for the model, so a ROC curve above this line is desirable. In fact, the model can be optimized to move the ROC curve towards the upper-left corner of the plot space; that is, better models will maximize the "area under the curve." A ROC curve below the unity slope line means that the model is worse than random sampling of the index at event detection (for those event threshold values that fell below the unity slope line). More on the history of the ROC curve can be found in Ekelund (2011) while Berrar and Flach (2012) provide additional caveats to ROC curve interpretation.

Uncertainties can be placed on these contingency table values. Both Agresti and Coull (1998) and Hogan and Mason (2012) provide thorough discussions of uncertainties on performance measures, including a reasonable set of parameter variances, S^2. The uncertainties in Hogan and Mason (2012) rely on the assumption that the time series is the model time series does not have any significant discontinuities or secular trends within the time interval of interest (true for most models) and that successive model outcomes are independent (which is not the case with many models). Stephenson (2000) also provides a robust discussion of confidence intervals and uncertainties for forecast metrics, arguing that, for many skill scores, the sampling distribution is nearly impossible to determine analytically and therefore analytical uncertainty estimates are also challenging. The bootstrap method described above is a alternative method to determining uncertainties, sampling with replacement and recalculating the metrics many times (typically more than 1000 iterations). Note that there are other confidence interval calculations that can be performed, such as the Wald interval or the Agesti-Coull interval.

583 ### 3.3. Additional Performance Assessment Best Practices

584 For the metrics discussed above, this does not have to dominate a new study's results
585 section. In its most compact form, it is simply a "benchmarking" subsection within a longer
586 study. There are several additional points that should be brought up about implementing the
587 standard set of metrics defined above for new geomagnetic index models, discussed below.

588 One choice that all modelers must make is the set of observations against which the
589 model should be tested. No predetermined event list or interval is specified as part of this
590 metrics definition. Such selections are often interesting to the community only for a few years,
591 after which new events and intervals become the preferred comparison set. In addition, some
592 researchers might want to assess their model against only quiet times, or only storm times, only
593 substorm times, or other requirements based on expected usage. Furthermore, some may argue
594 for active-time event lists as the preferred comparison set while others think that a long-time-
595 span interval, one that includes both quiet and active periods, is more appropriate. In short,
596 mandating specific times would not be helpful unless we covered all possible activity parameter
597 combinations for all possible geomagnetic indices. It is proposed that the baseline metrics set
598 matters more than the specific interval. Researchers should discuss why they chose the interval
599 they are using and the geomagnetic activity qualities of that interval. A good alternative to using
600 just one specific time interval for testing or validation is to use the K-fold cross-validation
601 procedure (e.g., Jonathan et al., 2000), which ensures that the training and the validation sets
602 have a similar distribution of events in terms of geomagnetic activity (the training set is used to
603 build a model, and the validation set is used to validate or test a model).

604 Regardless of the event list or intervals chosen for the test comparison, there is a specific
605 requirement that should be met. Specifically, the comparison set should be large enough to
606 contain hundreds, if not thousands, of data values. A minimum cutoff is that there should be at
607 least 10 values in both the hits and correct negatives bins for all threshold values used in defining
608 the ROC curve, and there should be at least 10 distinct threshold levels along the ROC curve, for
609 which the number of hits and correct negatives changes by at least one, if not several, per level.
610 So, an absolute minimum is ~100 values in the comparison set. However, several hundreds or
611 even thousands of data-model pairs would be better, to allow for more threshold settings and a
612 smoother curve set for the event performance.

613 Most models use as inputs measurements from the Advanced Composition Explorer,
614 ACE, or more generally, the OMNI database of upstream values, which includes measurements
615 from satellites prior to the ACE era. Usage of the new solar wind monitor, the Deep Space
616 Climate Observatory, DSCOVR, would be advantageous, not only for the sake of comparison
617 but also because of its higher time resolution plasma data. For reproducibility, model
618 assessments should be specific about the input data and time intervals used in both the training
619 and testing of the model, preferably even saving these input values with the model output at a
620 permanent data repository that provides a digital object identifier for the files.

621 Note that the range of observed index values included in both the training set and the
622 comparison is important. If users of space weather modeling tools want to understand the
623 usefulness of an index prediction model, operational code output must be placed in the context of
624 what was used to create it and test it. For example, the empirical function of O'Brien and
625 McPherron (2000a) between Dst and solar wind Ey "is restricted to Dst > -150 nT" (quoted from
626 the abstract), and therefore predictions of larger storms with this model should be understood to

be extrapolations of the model validity and therefore subject to larger uncertainty and caution in decision-making by users.

Uncertainty calculations have been given for some of the baseline metric quantities. For a few, one can even calculate a standard deviation. For others, though, the bootstrap method and cross-validation is useful for determining uncertainties (e.g., Huber 1981; Michaelsen, 1987; Efron & Tibshirani, 1993; Reiff, 1990) and used in several space physics data-model comparison studies (e.g., Jorgensen et al., 2004; Liemohn & Katus, 2012; Katus et al., 2013).

4. Application of the Standardized Assessment Set

We will show some examples of index prediction models using the standard assessment metrics. The first model assessment is of WINDMI. This is an independent simulation conducted at the CCMC, with no input from the model developers. Another example is by a user of a code, the UPOS Kp prediction model, not the original model developers. A third example is output from a physics-based ring current model, RAM, with the simulations conducted by the current set of developers for this model. Note that another study that used a very similar set of metrics for a geomagnetic index comparison is Liemohn et al. (2018), who analyzed the accuracy of the experimental real-time simulations of the Space Weather Modeling Framework being conducted by CCMC. It is also useful to note that metrics similar to those discussed in this paper, especially the "event detection performance metrics," were applied to studies that evaluated Geospace models for use in operations by NOAA's Space Weather Prediction Center (Pulkkinen et al. 2013; Glocer et al., 2016).

4.1. Dst and AL from the WINDMI low-dimensional geospace model

WINDMI, the solar wind interaction with the magnetosphere and ionosphere model (Horton & Doxas, 1996; Spencer et al., 2007), is a set of eight differential equations that characterizes geospace as a nonlinear electrical circuit. After scaling, two of those parameters are interpreted as equivalent to the Dst/SYM-H time series and the AL time series. Mays et al. (2009) assessed the performance of WINDMI for a set of substorm intervals with a few different metrics, finding that the Newell et al. (2008) solar wind-geospace coupling function works best as an input parameter for this code. This model is quick to execute and available for "instant runs" at the CCMC, making it an ideal code to use as an example model for this new standardized set of geomagnetic index performance metrics.

The code was run at CCMC for the entire year of 2014. This is the last complete year for which Dst/SYM-H final values are available (as opposed to provisional or real-time values) at the time the runs were conducted. Using the ACE Level 2 solar wind data set with the Newell coupling function, WINDMI was run for the entire year and simulated values of Dst/SYM-H and AL were produced every minute. These were compared against the SYM-H index and provisional AL index from the Kyoto World Data Center. Figure 1 shows the time series for these two comparisons, with the observed indices in black and the WINDMI results in red. It is seen that there is a systematic offset in the values, evident in both panels, with the quiet-time WINDMI output consistently lower (more negative) than the indices.

The fit performance metrics are listed in the first two data columns of Table 4. Over half a million data-model pairs were included in the calculations. The results are quite similar for both SYM-H and AL, so the comparisons will be described together. Regarding the linear fit values, the model is more negative (i.e., more active, for these two indices) for index values near

zero, but the slope of the fit is less than one, so the running average of the model values eventually crosses that of the data, with the data being more negative for large negative values of the index. The correlation coefficient is positive but only 0.66 (for both indices, coincidentally). The ME values are negative, indicating that the observations are more negative than the model. The RMSE and MAE values are slightly larger than the ME magnitudes, indicating that the bias of the model is smaller than the variation of the model around the observed values. For the selected interval, the PE values are negative for both indices. As seen in Figure 1, this is because the largest values for the modeled index values are slightly negative, around -20 nT for SYM-H and -50 for AL. These offsets make the comparisons during quiet times quite poor, which is seen in these fit performance metrics that take into account all values across the entire time interval.

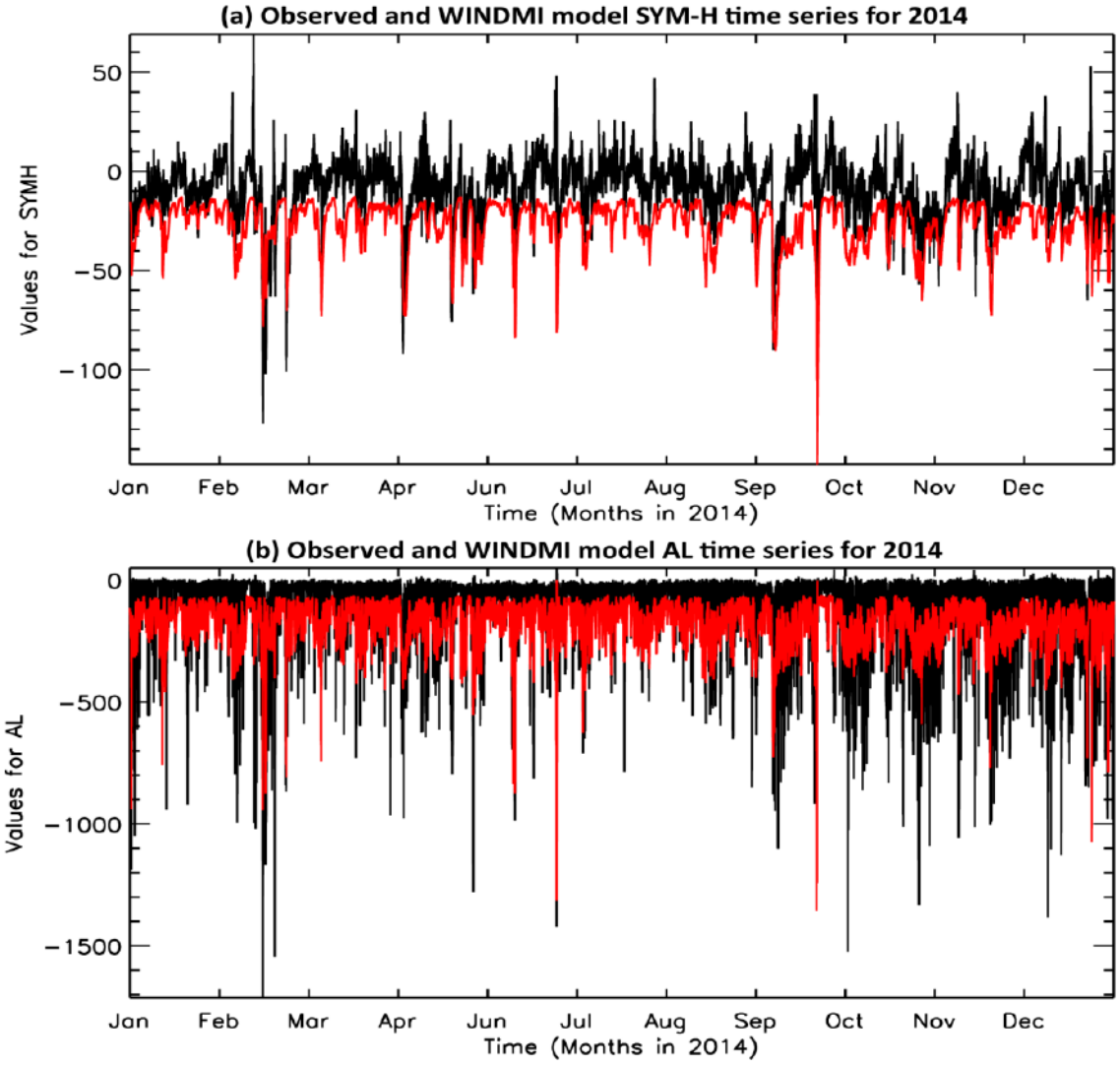


Figure 1. Times series values for the WINDMI model (red curves) against the (a) SYM-H and (b) AL indices (black curves) for the year 2014. The units of the y axes are in nanoTeslas.

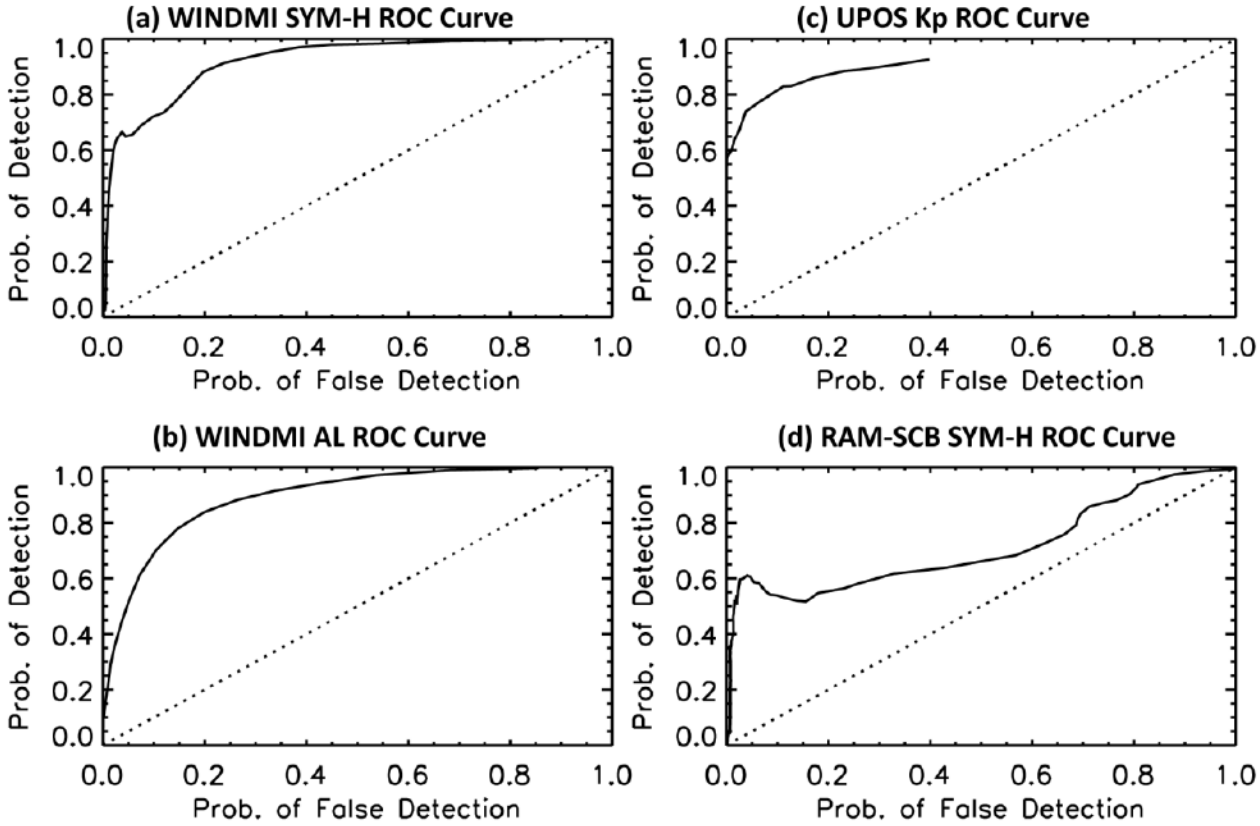
686 **Table 4.** Fit performance statistics of the example comparisons

	WINDMI SYM-H	WINDMI AL	UPOS Kp Estimation	RAM-SCB SYM-H
Number of values in comparison	525,600	525,600	99,842	44,639
Intercept of the linear fit	-21.5 nT	-135 nT	0.35	-7.8 nT
Slope of the linear fit	0.55 nT/nT	0.48 nT/nT	0.85	0.54 nT/nT
Pearson correlation coefficient (R)	0.66	0.66	0.86	0.68
Root mean square error (RMSE)	20.9 nT	127 nT	0.73	15.8 nT
Mean absolute error (MAE)	18.3 nT	108 nT	0.54	12.2 nT
Mean error (ME, or bias)	-17.6 nT	-87.4 nT	-0.08	1.56 nT
Prediction efficiency (PE)	-1.08	-0.10	0.73	0.45

687

688 Figures 2a and 2b show the ROC curves for these two indices. The ROC curves are well
689 above the unity slope line, indicating that the model is much better than random chance at
690 reproducing events (large negative excursions) in the observed time series. This is in contrast to
691 the relatively low PE score; the model does fairly well as predicting active time events.

692



693

694 Figure 2. ROC curves for the comparisons of (a) WINDMI SYM-H, (b) WINDMI AL, (c) UPOS
695 Kp, and (d) RAM-SCB SYM-H.

696

Figure 3a-3j give the HSS, POD, POFD, FAR, and FB for the WINDMI comparisons against the observed SYM-H and AL indices. For both indices, HSS hovers near zero for most threshold values but the POD is above 0.5 for most thresholds and the POFD is below 0.5 for most threshold values. The FAR is relatively high, indicating that there are more false alarms than hits when the model predicts an event. The frequency bias is large for both indices for near-zero thresholds, but for AL it drops to below one for the active-time thresholds (indicating more misses than false alarms for these thresholds).

This is an interesting comparison because the ROC curves show that the model has some skill at capturing events in SYM-H and AL, but the prediction efficiency, which is a skill score against the mean value of the observations, is not particularly good for either index. This touches on the issue of what a user might want from a prediction model and the need to examine more than one metric when assessing model performance.

4.2. Kp from the UPOS Kp Estimation Model

The UPOS Kp Estimation model was developed as part of the University Partnering for Operational Support (UPOS) project by the Applied Physics Laboratory of Johns Hopkins University following the method of Takahashi et al. (2001). This model produces an estimate of Kp every hour from magnetometer observations. For model assessment, we use definitive Kp values produced by GFZ Potsdam. Definitive Kp is produced every three hours and the Kp analysis tool produces output every hour. Thus, the question of how to relate the two quantities must be considered. Kp is intrinsically only defined over a three-hour window (see Section 2.2), so the approach taken here is to assign the Kp value for a given three-hour period to each hour within the period.

We performed analysis of model outputs from 1 October 2001 through 29 July 2013 allowing coverage of a complete solar cycle. At a 3-h cadence, this results in almost a hundred thousand data-model pairs in the comparison. Table 1 provides the fit performance values and Figures 2c and 3k-3o show the event performance for the model. The values for r and PE are high, at 0.86 and 0.73, respectively. Both the RMSE and MAE are below one, i.e., the variation of the model around the data is usually within one Kp unit increment. The discrete nature of Kp makes the linear fit more qualitative than for other indices, but they still convey performance information, which for the UPOS model appears to be very reasonable. In Figure 2c, the ROC curve for this model is well above the unity slope line. All of the other event statistics (in the third column of Figure 3) are quite good across most of the threshold values, but they start to deviate to slightly worse values near a threshold value above Kp of 8.

4.3. SYM-H from the RAM-SCB drift physics model

The ring current-atmosphere interactions model (RAM) developed by Jordanova et al. (1994, 1996) was first employed to simulate the effects of adiabatic drifts and collisional losses on the major ring current ions H^+, O^+, and He^+ using a centered dipole magnetic field model and the analytical Volland-Stern (VS) (Volland, 1973; Stern, 1975) convection and corotation potential model. The 4-dimensional simulation domain of RAM is specified by radial distance in the equatorial plane, magnetic local time (MLT), energy, and equatorial pitch angle. RAM can couple with the 3-dimensional self-consistent magnetic field (SCB) (Zaharia et al., 2004; Zaharia, 2008) as well as having an implementation of a self-consistent electric field coupling (RAM-SCBE; Yu et al., 2017). As noted by Jordanova et al. (2018), a simplified version of

RAM with the same components as its early implementation has been developed for near-real time operations, using a dipole magnetic field and the VS electric field model, with the particle flux at the outer boundary being driven by data when available and by a statistical model (Denton et al. 2015, 2016) when data are not available. This model configuration is robust and computationally inexpensive. To demonstrate the robustness of the model we simulated the month of January 2005, following Haiducek et al. (2017), using data from the LANL (Los Alamos National Laboratory) geosynchronous satellites to specify the outer flux boundary.

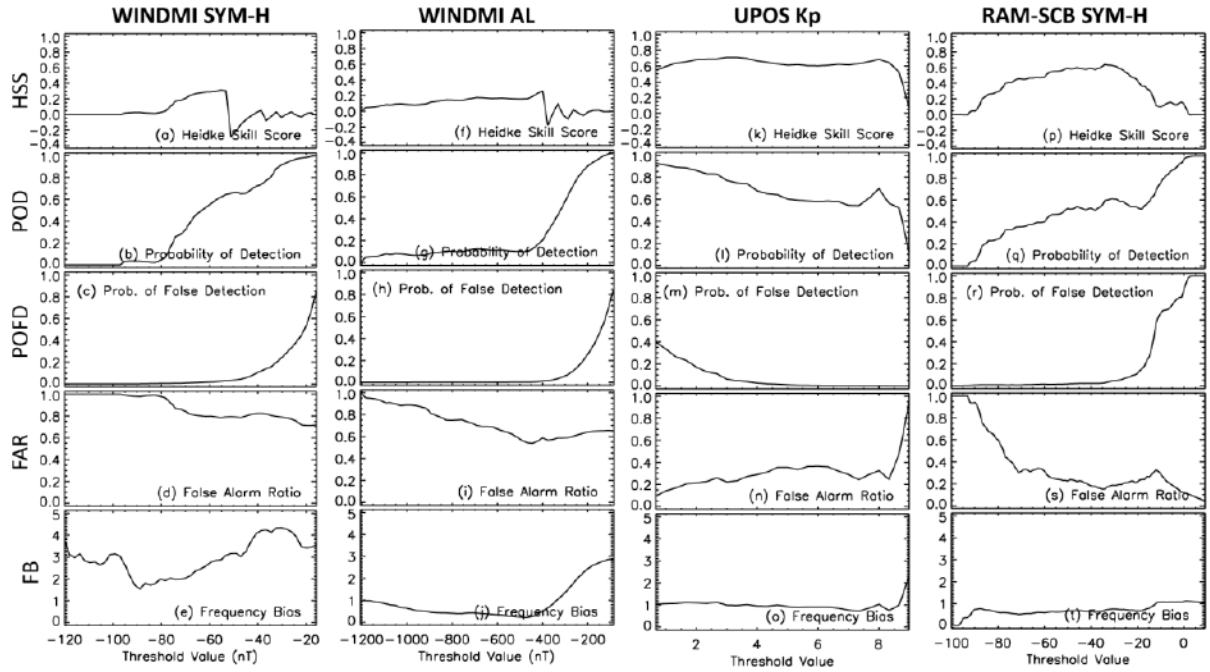


Figure 3. Event performance metrics for the comparisons of (first column, a-e) WINDMI SYM-H, (second column, f-j) WINDMI AL, (third column, k-o) UPOS Kp, and (fourth column, p-t) RAM-SCB SYM-H.

The set of metrics for assessment given in section 3 have been calculated for the SYM-H index. The simulation is as described above, where the SYM-H is calculated using a Biot-Savart integration, and the SYM-H is provided by the World Data Center for Geomagnetism in Kyoto. Both series are given at 1-minute resolution, giving us 44639 data points in each series. We perform a linear regression using ordinary least squares to obtain the linear fit parameters giving a slope of 0.538, an intercept of -7.77 and a Pearson correlation coefficient of 0.684. The accuracy of the model is measured by MAE and RMSE, giving 12.2 nT and 15.8 nT, respectively. The model tends to slightly over predict Sym-H, with an ME of 1.56 nT. The prediction efficiency is 0.452, representing a 45.2% improvement in skill over a prediction of the sample mean. These metrics are summarized in Table 1. We note that the reported accuracy of this SYM-H prediction is comparable to the operational configuration of the SWMF reported for this same month by Haiducek et al. (2017) and that the RAM predictions are less biased.

The event performance metrics are shown in Figures 2 and 3. The ROC curve for the RAM comparison is in Figure 2d, which, like the other models, is above the unity slope line,

767 indicating that the model has some skill in reproducing active times. The other event metrics are
768 shown in Figures 3p – 3t. Of note are that the HSS peaks for moderate storm events, reaching a
769 value above 0.5 and that FB hovers close to unity for nearly all threshold levels.

770 **5. Discussion**

771 Thus far, a summary of existing geomagnetic index prediction models has been
772 presented, a standardized set of metrics has been defined, and three models have undergone
773 calculations of these metrics for different intervals.

774 As discussed in section 3, these metrics were chosen because they each assess a particular
775 aspect of model performance. We encourage all new and updated models to undergo the full set
776 described above, and then discuss the performance of the model with respect to each of these
777 metrics. This is a recommendation, not a requirement, and while the full set of metrics is
778 encouraged for all new or improved index prediction models, there are certainly some metrics
779 that will be more suitable for particular needs than others and perhaps not all models need to be
780 evaluated with the full set.

781 That is, models should be created with potential users in mind, perhaps even identified.
782 Each of those potential or real users will have specific needs for index prediction performance.
783 One example is that a user might only care about accuracy during the extreme events and not
784 during quiet times. In this case, RMSE is more important than r, MAE, or PE; the event
785 performance is, in general, more relevant for the user than the fit performance; and even within
786 that, the metric values for the "big event" thresholds are more the assessments of higher interest
787 than the rest of the curves. Maximizing this subset of the standard set of metrics is what best
788 suits that user's needs, even if the model is not particularly good for other metrics. It is therefore
789 recommended that developers and users follow the Application Usability Level (AUL) procedure
790 defined by Halford et al. (2018), starting with a particular purpose in mind and using this metrics
791 set to help define what is important for each user. This will best inform how to direct the model
792 development (or selection among several existing models).

793 An example of this is that some geomagnetic indices are suitable as input drivers for
794 understanding and predicting ionospheric disturbances. Users interested in this application
795 should tailor their performance assessment of an index prediction model for this purpose. One
796 factor to consider is how precisely the indices are able to indicate magnitude of expected
797 ionospheric disturbances. For example, Borovsky and Denton (2006) summarize different
798 geospace responses depending on the type of solar wind structure causing the activity. While Dst
799 and SYM-H are good indices for monitoring intense storm activity, other geomagnetic indices
800 are better for less intense disturbance (see, for example, Borovsky & Shprits, 2017). Specifically
801 for ionospheric disturbances, Buresova and Lastovicka (2017) noted a shift in which
802 geomagnetic index is most relevant for ionospheric prediction. Because of this usage of indices
803 as drivers, it is recommended that discussions occur between the ionospheric community and
804 those developing models to predict geomagnetic indices. This would be very useful and
805 important for improving both forecasts of geomagnetic indices and ionospheric disturbances.

806 Regarding model development, O'Brien (2006) discusses the limits on complexity of
807 geomagnetic index predictor models. He lays out the situation as an example application of
808 Occam's razor – only add complexity to a model (e.g., a new parameter) if it significantly
809 improves the fit. There is also a robust discussion in Osthus et al. (2014) on parameter estimates

for regression models and multicollinearity. The main point is that when input variables are correlated with each other the interpretation of the model parameters gets difficult. This is something to consider when developing or modifying a code.

In assessing a new or improved model, it should be remembered that the input parameters to the model have uncertainties associated with each data stream. These uncertainties might vary with time, usually being larger during more active solar wind conditions. The uncertainties can also be larger during very quiet conditions, when the signal starts approaching the noise level of the instrument. It is also important to note how the measurements are propagated from the upstream spacecraft to the Earth's magnetopause, including an understanding of input ambiguity due to the spacecraft distance from the Sun-Earth line. This input uncertainty is in addition to the uncertainties mentioned in section 3 above, and should be propagated through the calculation (e.g., Taylor, 1997, ch. 3). While this error propagation can be done mathematically, systematically or randomly varying inputs around the observed data stream can quantify the sensitivity of the prediction model to uncertainties in specific input parameters.

It should be noted that hemispheric bias exists in most ground-based geomagnetic indices. Compared to the southern hemisphere, the northern hemisphere has a higher land coverage percentage and a larger population, which has resulted in far more ground-based magnetometer observatories in this half of the world. Therefore, there is a northern-hemisphere bias to most indices derived from ground-based magnetometers. While these metrics do not directly address this issue, the point should be acknowledged and index users should consider themselves cautioned about inferring physical processes from such times series. It is also important to note that ground-based magnetic indices are sensitive to the location of the magnetometer stations. For example, the auroral electrojet moves in latitude, so a set of stations at even a slightly different latitude would result in a different times series for these indices that represent the strength of auroral currents (see, e.g., Newell and Gjerloev, 2011).

Similarly, there is a systematic bias implicit in ground-based magnetometer data from the local induced currents just below the Earth's surface. This is different around each observatory, yet only some indices take this influence into account when combining data from the stations. Again, this metrics set does not directly address this issue; it something about which geomagnetic index users should be aware.

The timing of the model value relative to the observed index value is important. Specifically, Dst and Kp have a 1-hour and 3-hour cadence to their time series and represent variation of the magnetic field on the ground within fixed, not sliding 1-hour and 3-hour intervals, respectively. Sometimes, however, modelers assign a specific time to each of the values of a given index, rather than considering indices as corresponding to an interval in time. Furthermore, some models generate index predictions at a much higher cadence than the index time series. The choice of this timestamp can cause ambiguity in data-model comparisons, since the information used as an input to a predictive model depends on which part of the 1-hour (for Dst) or 3-hour interval (for Kp) is chosen as a timestamp (e.g., the beginning, the middle, or the end of the interval). Care must be taken when comparing model with observation when the index is compiled over a relatively long (~hour or more) interval.

Figure 4 illustrates this ambiguity on an example of a) nowcast and b) forecast 3 hours ahead of the Kp index. If the timestamp of Kp is chosen at the beginning of the 3-hour interval, then to issue a prediction for the interval of 0-3 hours ahead (Figure 4a) the solar wind

information available until the beginning of that interval should be used (indicated by the dark-grey shaded region). However, if the timestamp of Kp is chosen at the end of the interval, the information during the current Kp interval in addition to the information until hour 0 should be used for the prediction (indicated by the light-grey shaded region). While both models may be referred to as a nowcast, they are different in their predictive capabilities since they use different information to issue predictions. Figure 4b illustrates this issue for the case of 3 hours ahead prediction. To avoid that ambiguity in the definition of prediction horizons, models for the same prediction horizon should use input information available until the same point in time. It should be clearly indicated for which specific time interval in the future or past the prediction is made and what information (prior to which part of the 1-hour or 3-hour interval for Dst and Kp respectively) is used to issue predictions. An example of possible nomenclature for the prediction of Kp for the intervals of 0-3 and 3-6 hours ahead is shown in Figure 4c and 4d, respectively. Here, a model that uses information prior to the current time (hour 0) to predict Kp for 0-3 hours ahead is called "a model predicting the Kp index for the interval of 0-3 hours ahead" (Figure 4c), and a model that uses information prior to the current time to predict Kp for the 3-6 hours interval is called "a model predicting Kp for the interval of 3-6 hours ahead" (Figure 4d). The same can be applied to any t-t+3 hours ahead prediction. In these terms, a model that uses the information shown by the light-grey shaded region in Figure 4a would be called "a model predicting Kp for the interval -3-0". In summary, a model, including its input values, should align with the time cadence and intervals of the index so that values are truly comparable.

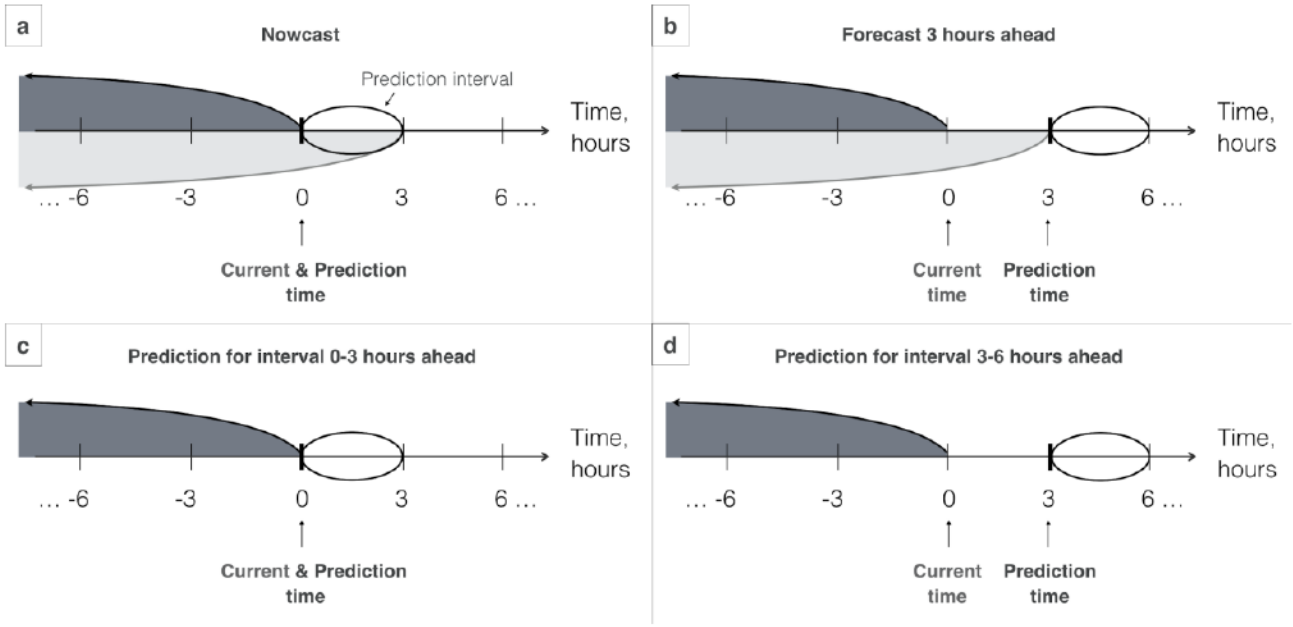


Figure 4. Examples of possible ambiguity in the definitions of prediction horizons caused by selecting different timestamps of the model relative to the timing of the Kp index for a) nowcast and b) forecast for 3 hours ahead, and suggested definitions of prediction horizons based on intervals for c) 0-3 hours ahead and d) 3-6 hours ahead prediction.

6. Conclusion

Geomagnetic indices provide single-value distillations of expansive data sets and complex physics. While they are not particularly useful for deciphering specific processes or for informing decision-making at the local level, they can be very helpful in understanding general activity levels in different regions of geospace. Many researchers have undergone the task of developing models for predicting these indices, as summarized in section 2 above. While developers and users are usually quite careful in their quantitative assessments of each model, there is no accepted set of metrics for benchmarking a new code that seeks to reproduce the time series of a geomagnetic index.

Section 3 presents a baseline set of metrics that quantify the fit performance and event detection abilities of a model. The parameters are easily calculated and examine a number of different aspects about the model. It is recommended as a minimum collection of metrics that should be calculated and analyzed for each new model or model upgrade.

A few best practice procedures were discussed for conducting a performance assessment of a geomagnetic index prediction model. No set time interval is specified for testing a new model. This is left up to the developer or user, depending on their planned implementation of the code. For statistical robustness, it is advised that at least hundreds, if not thousands, of data-model value pairs be used in the comparison. It was suggested that uncertainties be calculated and examined, to understand the possible variation in each performance metric due to systematic or random errors in the observations or modeling approach.

Three examples were given of different geomagnetic index models undergoing this regimen of metrics assessment. No conclusions about the quality of these models are drawn from these values; these are simply examples that others can repeat. There are, however, significant differences in the performance of these models that highlight the need for a broad mix of metrics when assessing a prediction model.

A number of caveats and limitations to geomagnetic index prediction and usage were discussed. One is that models should be developed with potential users in mind and design the tool to produce output that best suits the requirements for that application. It is noted that there are several known issues with geomagnetic indices, in particular their northern hemispheric bias and possible offsets due to ground conductivity. There is also the issue of timing when making a data-model comparison against a geomagnetic index, especially Kp with its 3-hour cadence.

The selected metrics, best practice advice, and caveats are summarized as follows:

- Recommended fit performancemetrics: linear fit intercept and slope, Pearson correlation coefficient, root mean square error, mean absolute error, mean error, and prediction efficiency

- Recommended event detection performance: Heidke Skill Score, probability of detection, probability of false detection, false alarm ratio, frequency bias (all as a function of threshold setting), and a receiver operating characteristic curve
- Recommended interval selection: no set interval, but hundreds, if not thousands of data values should be used in the metrics assessment
- Recommended solar wind input values: none but specify which satellite and data product version is used for repeatability, and note the uncertainty in these input values and propagate the error through the model results
- Recommended uncertainty calculations: encouraged but not demanded
- Recommended emphasis among the metrics: each metric quantifies only a particular aspect of the data-model comparison, so keep the end-use in mind when conducting and interpreting a model assessment
- Recommended model development philosophy: only add complexity to a model if it significantly improves the metrics of particular interest
- Recommended caveat to geomagnetic index interpretation: systematic bias exists in ground-based magnetometer data – northern hemispheric bias, influences of local induced currents in the Earth, and the relative timing of observed and modeled index production – which could confound interpretation of results

This standard set of metrics can be used in a number of ways. The first is that a model developer can run their geomagnetic index prediction tool through this set of metrics to provide a baseline performance assessment of the model. A second use would be for a user of a particular model to conduct these tests, independent of the developer, to understand the accuracy, applicability, and limitations of the chosen model for their specific needs. A third possibility is that a user without a preselected model could use some or all of these metrics to select the most appropriate tool for their application. We hope that this standard set of metrics is useful for the space weather research and operations communities.

Acknowledgments and Data

This paper is the product of the Geomagnetic Indices Working Group of the International CCMC-LWS Working Meeting on Space Weather Metrics. The authors would like to thank the organizers of the workshop for their time and effort to rally the community into action on devising assessment standards for space weather models. We would also like to thank others that contributed but declined authorship, specifically Lutz Rastätter and Leila Mays at NASA and Joshua Rigler at USGS. The projects leading to these results have received funding from the European Union Seventh Framework Programme (FP7/ 2007-2013) under grant agreement No 606716 SPACESTORM and from the European Union's Horizon 2020 research and innovation program under grant agreement No 637302 PROGRESS. Work in the US was conducted under Work at the University of Michigan was supported by NASA grants NNX14AF34G, NNX17AI48G, NNX17AB87G, 80NSSC17K0015, and NNX14AC02G, and NSF grant 1663770. The Catholic University of America effort was performed under the CUA-NASA Cooperative Agreement supported by NASA Grant NNG11PL10A 670.135. Funding at the

957 University of Sheffield was provided by STFC UK grant ST/R000697/1. The work done at the
958 University of Alcala was supported by grant from MINECO AYA2016-80881-P. SKM
959 acknowledges support from the US Department of Energy's Laboratory Directed Research and
960 Development program (grant number 20170047DR). The work at GFZ Potsdam was supported
961 by Geo.X, the Research Network for Geosciences in Berlin and Potsdam, under Grant No
962 SO_087_GeoX, and by the European Union's Horizon 2020 research and innovation program
963 under grant agreement No 776287 SWAMI. Work at Los Alamos was supported through the
964 Laboratory Directed Research and Development program by the US Department of Energy under
965 contract DE-AC52-06NA25396. Work at the Institute of Atmospheric Physics was supported by
966 the H2020 COMPET-2017 TechTIDE Project (776011). Work at IRF-Lund was supported by
967 ESA Contract SSA- SWE-P2-1.5

968

969 Data used in the metrics assessments in this paper were obtained from the Space Physics
970 Data Facility at http://cdaweb.gsfc.nasa.gov/, Supermag at http://supermag.jhuapl.edu/, WDC-
971 Kyoto at http://wdc.kugi.kyoto-u.ac.jp/. Model output and the code used to create the figures and
972 calculate the metrics is available at the University of Michigan Deep Blue Data repository,
973 https://deepblue.lib.umich.edu/data/?locale=en . We have uploaded a temporary version here
974 and will "mint a DOI" to finalize and freeze the data brick upon acceptance.

975

976 **References**

977 Agapitov, O. V., Artemyev, A. V., Mourenas, D., Mozer, F. S., & Krasnoselskikh, V. (2015).
978 Empirical model of lower band chorus wave distribution in the outer radiation belt.
979 *Journal of Geophysical Research: Space Physics, 120*, 10,425–10,442.
980 https://doi.org/10.1002/2015JA021829

981 Agresti, A. & Coull, B.A. (1998). Approximate is better than 'exact' for interval estimation of
982 binomial proportions. *The American Statistician, 52*, 119–126.

983 Aguado, J., C. Cid, E. Saiz, & Y. Cerrato (2010). Hyperbolic decay of the Dst index during the
984 recovery phase of intense geomagnetic storms. *Journal of Geophysical Research Space*
985 *Physics, 115*, A07220, doi:10.1029/2009JA014658.

986 Amariutei, O. A. & N. Yu. Ganushkina (2012). On the prediction of the auroral westward
987 electrojet index. *Annales Geophysicae, 30*, 841-847, doi:10.5194/angeo-30-841-2012.

988 Ayala Solares, J. R., H.-L. Wei, R. J. Boynton, S. N. Walker, & S. A. Billings (2016). Modeling
989 and prediction of global magnetic disturbance in near-Earth space: A case study for Kp
990 index using NARX models. *Space Weather, 14*, doi: 10.1002/2016SW001463.

991 Baker, D. N., E. W. Hones, Jr., J. B. Payne, & W. C. Feldman (1981). A high time resolution
992 study of interplanetary parameter correlations with AE. *Geophysical Research Letters, 8,*
993 1971.

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

996 Bala, R., & P. Reiff (2014). Validating the Rice neural network and the Wing Kp real-time
997 models. *Space Weather, 12*, 417-425, doi: 10.1002/2014SW001075.

Bala, R., P. H. Reiff, & J. E. Landivar (2009). Real-time prediction of magnetospheric activity using the Boyle index. *Space Weather*, 7, S04003, doi:10.1029/2008SW000407.

Balan, N., Y. Ebihara, R. Skoug, K. Shiokawa, I. S. Batista, S. Tulasi Ram, Y. Omura, T. Nakamura, & M.-C. Fok (2017). A scheme for forecasting severe space weather. *Journal of Geophysical Research Space Physics, 122*, doi: 10.1002/2016JA023853.

Bartels, J., Heck, N.H. & Johnston, HF. (1939). The three-hour range index measuring geomagnetic activity. *Journal of Geophysical Research*, *44*, 411–454.

Berrar, D., & P. Flach (2012). Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Briefings in Bioinformatics*, *13*(1), 83–97, https://doi.org/10.1093/bib/bbr008

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

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

Boaghe, O. M., M. A. Balikhin, S. A. Billings, & H. Alleyne (2001). Identification of nonlinear processes in the magnetospheric dynamics and forecasting of *Dst* index. *Journal of Geophysical Research Space Physics*, 106(A12), 30,047–30,066, doi:10.1029/2000JA900162.

Boberg, F., P. Wintoft, & H. Lundstedt (2000). Real time *Kp* predictions from solar wind data using neural networks. *Physics and Chemistry of the Earth, Part C*, 25, 275–280, doi:10.1016/S1464-1917(00)00016-7.

Borovsky, J. E. (2014). Canonical correlation analysis of the combined solar wind and geomagnetic index data sets. *Journal of Geophysical Research: Space Physics*, *119*, 7, 5364-5381.

Borovsky, J. E., & M. H. Denton (2006). Differences between CME-driven storms and CIR-driven storms. *Journal of Geophysical Research, 111*, A07S08, doi:10.1029/2005JA011447.

Borovsky, J. E., & Shprits, Y. Y. (2017). Is the Dst index sufficient to define all geospace storms? *Journal of Geophysical Research: Space Physics, 122*, 11,543–11,547. https://doi.org/10.1002/2017JA024679

Boyle, C. B., P. H. Reiff, & M. R. Hairston (1997). Empirical polar cap potentials. *Journal of Geophysical Research Space Physics*, *102*, 111–125, doi:10.1029/96JA01742.

Boynton, R. J., M. A. Balikhin, S. A. Billings, A. S. Sharma, & O. A. Amariutei (2011). Data derived NARMAX Dst model, *Annales Geophysicae, 29*, 965-971, doi: 10.5194/angeo-29-985-2011.

Brautigam, D. H., & J. M. Albert (2000). Radial diffusion analysis of outer radiation belt electrons during October 9, 1990, magnetic storm. *Journal of Geophysical Research Space Physics, 105*, 291– 309.

1037 Brautigam, D. H., G. P. Ginet, J. M. Albert, J. R. Wygant, D. E. Rowland, A. Ling, & J. Bass
1038 (2005). CRRES electric field power spectra and radial diffusion coefficients. *Journal of*
1039 *Geophysical Research Space Physics, 110*, A02214, doi:10.1029/2004JA010612.

1040 Buresova, D. & J. Lastovicka (2017). Differences in Midlatitude Ionospheric response to
1041 magnetic Disturbances at Northern and Southern Hemispheres and Anomalous response
1042 During the Last Extreme Solar Minimum. In *Ionospheric Space Weather: Longitude and*
1043 *Hemispheric Dependences and Lower Atmosphere Forcing*, Geophysical Monograph, ed.
1044 by Timothy Fuller-Rowell, Endwoke Yizengaw, Patricia H. Doherty and Sunanda Basu,
1045 American Geophysical Union, Chapter 4, 41-58, Published by John Wiley&Sons, Inc.

1046 Burton, R. K., R. L. McPherron, & C. T. Russell (1975). An empirical relationship between
1047 interplanetary conditions and *Dst*, *Journal of Geophysical Research Space Physics,*
1048 80(31), 4204–4214, doi:10.1029/JA080i031p04204.

1049 Cade, W. B., III, J. J. Sojka, & L. Zhu (1995). A correlative comparison of the ring current and
1050 auroral electrojets using geomagnetic indices. *Journal of Geophysical Research Space*
1051 *Physics, 100*(A1), 97–105, doi:10.1029/94JA02347.

1052 Carpenter, D. L., & R. R. Anderson (1992). An ISEE/Whistler model of equatorial electron
1053 density in the magnetosphere. *Journal of Geophysical Research Space Physics, 97*, 1097.

1054 Carter, J. V., J. Pan, S. N. Rai, & S. Galandiuk (2016). ROC-ing along: Evaluation and
1055 interpretation of receiver operating characteristic curves. *Surgery, 159*(6), 1638-1645,
1056 https://doi.org/10.1016/j.surg.2015.12.029

1057 Clauer, C. Ro, R. L. McPherron, C, Searls, & M. G. Kivelson (1981). Solar windcontrol of
1058 auroral zone geomagnetic activity. *Geophysical Research Letters, 8*, 915.

1059 Dasso, S., D. Gómez, & C. H. Mandrini (2002). Ring current decay rates of magnetic storms: A
1060 statistical study from 1957 to 1998. *Journal of Geophysical Research*, *107*(A5), doi:
1061 10.1029/2000JA000430.

1062 Davis, T. N., & M. Sugiura (1966). Auroral electrojet activity index *AE* and its universal time
1063 variations. *Journal of Geophysical Research Space Physics*, *71*, 785–801,
1064 doi:10.1029/JZ071i003p00785.

1065 Denton, M.H., Thomsen, M.F., Jordanova, V.K., Henderson, M.G., Borovsky, J.E., Denton, et al.
1066 (2015). An empirical model of electron and ion fluxes derived from observations at
1067 geosynchronous orbit. *Space Weather. 13* https://doi.org/10.1002/2015SW001168.

1068 Denton, M.H., Henderson, M.G., Jordanova, V.K., Thomsen, M.F., Borovsky, J.E., Woodroffe,
1069 J., et al. (2016). An improved empirical model of electron and ion fluxes at
1070 geosynchronous orbit based on upstream solar wind conditions. *Space Weather, 14*
1071 https://doi.org/10.1002/2016SW001409.

1072 Devos A, Verbeeck C & Robbrecht E (2014). Verification of space weather forecasting at the
1073 Regional Warning Center in Belgium. *Journal of Space Weather and Space Climate, 4*,
1074 A29.

1075 Efron, B., & R. J. Tibshirani (1993). *An Introduction to the Bootstrap*, 436pp., Chapman and
1076 Hall, New York.

1077 Ekelund, S. (2011). ROC Curves – What are They and How are They Used?, *Point of Care,*
1078 *11*(1), 16-21, doi: 10.1097/POC. 0b013e318246a642.

1079 Feldstein, Y. (1992). Modelling of the magnetic field of the magnetospheric ring current as a
1080 function of interplanetary medium parameters. *Space Science Reviews, 59*, 83–165,
1081 doi:10.1007/BF01262538.

1082 Fenrich, F. R., & J. G. Luhmann (1998). Geomagnetic response to magnetic clouds of different
1083 polarity. *Geophysical Research Letters, 24*, 2999-3002, doi: 10.1029/98GL51180

1084 Fok, M.-C., T. E. Moore, J. U. Kozyra, G. C. Ho, & D. C. Hamilton (1995). Three-Dimensional
1085 Ring Current Decay Model. *Journal of Geophysical Research*, *100*(A6), 9619–9632, doi:
1086 10.1029/94JA03029.

1087 Fok, M.-C., N. Y. Buzulukova, S.-H. Chen, A. Glocer, T. Nagai, P. Valek, & J. D. Perez (2014).
1088 The Comprehensive Inner Magnetosphere-Ionosphere Model. *Journal of Geophysical*
1089 *Research Space Physics*, *119*, 7522–7540, doi: 10.1002/2014JA020239.

1090 Ganushkina, N. Y., T. I. Pulkkinen, V. F. Bashkirov, D. N. Baker, & X. Li (2001). Formation of
1091 intense nose structures. *Geophysical Research Letters, 28*, 491-494.

1092 Ganushkina, N. Yu., M. W. Liemohn, & T. I. Pulkkinen (2012). Storm-time ring current: Model-
1093 dependent results. *Annales Geophysicae, 30*, 177, doi: 10.5194/angeo-30-177-2012.

1094 Gleisner, H., & H. Lundstedt (2001). Auroral electrojet predictions with dynamic neural
1095 networks. *Journal of Geophysical Research Space Physics*, *106*(A11), 24,541–24,549,
1096 doi:10.1029/2001JA900046.
1097 Glocer, A., L. Rastätter, M. Kuznetsova, A. Pulkkinen, H. J. Singer, C. Balch, et al. (2016).
1098 Community-wide validation of geospace model local K-index predictions to support model
1099 transition to operations. *Space Weather*, 14,469–480, doi:10.1002/2016SW001387.

1100 Gjerloev, J. W. (2012). The SuperMAG data processing technique. *Journal of Geophysical*
1101 *Research Space Physics*, 117, A09213, doi: 10.1029/2012JA017683.

1102 Goertz, C. K., L.-H. Shan, & R. A. Smith (1993). Prediction of geomagnetic activity. *Journal of*
1103 *Geophysical Research Space Physics*, *98*(A5), 7673–7684, doi:10.1029/92JA01193.

1104 Gonzalez, W. D., J. A. Joselyn, Y. Kamide, H. W. Kroehl, G. Rostoker, B. T. Tsurutani, & V. M.
1105 Vasyliunas (1994). What is a geomagnetic storm? *Journal of Geophysical Research*
1106 *Space Physics*, 99(A4), 5771–5792, doi: 10.1029/93JA02867.

1107 Haiducek, J. D., Welling, D. T., Ganushkina, N. Y., Morley, S. K., & Ozturk, D. S. (2017).
1108 SWMF global magnetosphere simulations of January 2005: Geomagnetic indices and
1109 cross-polar cap potential. *Space Weather*, *15*, 1567–1587, doi: 10.1002/2017SW001695

1110 Halford, A., A. Kellerman, K. Garcia-Sage, J. Klenzing, B. Carter, R. McGranaghan, et al.
1111 (2018). Application Usability Levels: A framework for tracking a project's progress.
1112 *Space Weather*, submitted, ms # 2018SW001977.

1113 Harel, M., R. A. Wolf, P. H. Reiff, R. W. Spiro, W. J. Burke, F. J. Rich, & M. Smiddy (1981).
1114 Quantitative simulation of a magnetospheric substorm 1, model logic and overview.
1115 *Journal of Geophysical Research, 86*, 2217–2241, doi:10.1029/JA086iA04p02217.

1116 Heidke, P. (1926). Berechnung des Erfolges und der Güte der Windstärkevorhersagen im
1117 Sturmwarnungdienst (Calculation of the success and goodness of strong wind forecasts in
1118 the storm warning service). *Geogr. Ann. Stockholm*, *8*, 301–349.

1119 Henney, C. J., W. A. Toussaint, S. M. White, & C. N. Arge (2012). Forecasting F10.7 with solar
1120 magnetic flux transport modeling. *Space Weather, 10*, S02011, doi:
1121 10.1029/2011SW000748.

1122 Hogan, R. J. & Mason, I. B. (2012). Deterministic Forecasts of Binary Events. In *Forecast*
1123 *Verification* (eds I. T. Jolliffe and D. B. Stephenson). doi:10.1002/9781119960003.ch3

1124 Holzer, R. E., & J. A. Slavin (1982). An evaluation of three predictors of geomagnetic activity.
1125 *Journal of Geophysical Research Space Physics*, *87*(A4), 2558–2562,
1126 doi:10.1029/JA087iA04p02558.

1127 Horton, W., & I. Doxas (1996). A low-dimensional energy-conserving state space model for
1128 substorm dynamics. *Journal of Geophysical Research Space Physics*, *101*(A12), 27,223–
1129 27,237.

1130 Horton, W., & I. Doxas (1998). A low-dimensional dynamical model for the solar wind driven
1131 geotail-ionosphere system, *Journal of Geophysical Research Space Physics*, *103*(A3),
1132 4561–4572.

1133 Huber, P. J. (1981). *Robust Statistics*, 308pp., John Wiley, New York.

1134 Ilie, R., M. W. Liemohn, G. Toth, & R. Skoug (2012). Kinetic model of the inner magnetosphere
1135 with arbitrary magnetic field. *Journal of Geophysical Research Space Physics, 117*,
1136 A04208, doi: 10.1029/2011JA017189.

1137 Iyemori, T. (1990). Storm-time magnetospheric currents inferred from mid-latitude geomagnetic
1138 field variations. *Journal of Geomagnetism and Geoelectricity, 42*, 1249-1265, doi:
1139 10.5636/jgg.42.1249.

1140 Iyemori, T., T. Araki, T. Kamei, & M. Takeda (1992). *Mid-latitude geomagnetic indices ASY*
1141 *and SYM (provisional) No. 1 1989*. Data Analysis Center for Geomagnetism and Space
1142 Magnetism, Kyoto Univ., Kyoto.

1143 Jonathan, P., Krzanowski, W.J. & McCarthy, W.V. (2000). On the use of cross-validation to
1144 assess performance in multivariate prediction. *Statistics and Computing, 10*, 209–229,
1145 doi: 10.1023/A:1008987426876.

1146 Jolliffe, I.T., & Stephenson, D.B. (2012). *Forecast verification: A practitioner's guide*
1147 *in atmospheric science*. Wiley-Blackwell, Hoboken, NJ.

1148 Jordanova, V. K., J. U. Kozyra, G. V. Khazanov, A. F. Nagy, C. E. Rasmussen, & M. C. Fok
1149 (1994). A bounce-averaged kinetic-model of the ring current ion population. *Geophysical*
1150 *Research Letters, 21*, 2785–2788, doi:10.1029/94GL02695.

1151 Jordanova, V. K., L. M. Kistler, J. U. Kozyra, G. V. Khazanov, & A. F. Nagy (1996). Collisional
1152 losses of ring current ions. *Journal of Geophysical Research Space Physics*, *101*(A1),
1153 111–126, doi: 10.1029/95JA02000.

1154 Jordanova, V. K., C. J. Farrugia, L. Janoo, J. M. Quinn, R. B. Torbert, K.W. Ogilvie, R. P.
1155 Lepping, J. T. Steinberg, D. J. McComas, & R. D. Belian (1998). October 1995 magnetic

cloud and accompanying storm activity: Ring current evolution, *Journal of Geophysical Research Space Physics*, *103*, 79.

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

Jordanova, V.K., G.L. Delzanno, M.G. Henderson, H.C. Godinez, C.A. Jeffery, E.C. Lawrence, et al. (2018). Specification of the near-Earth space environment with SHIELDS. *Journal of Atmospheric and Solar-Terrestrial Physics, 177*, 148-159, doi: 10.1016/j.jastp.2017.11.006.

Jorgensen, A. M., H. E. Spence, W. J. Hughes, & H. J. Singer (2004). A statistical study of the global structure of the ring current. *Journal of Geophysical Research Space Physics, 109*, A12204, doi:10.1029/2003JA010090.

Kamide, Y. (1983). Comment on 'An evaluation of three predictors of geomagnetic activity' by R. E. Holzer and J. A. Slavin. *Journal of Geophysical Research Space Physics*, *88*(A6), 4953–4954, doi:10.1029/JA088iA06p04953.

Kamide, Y., & S. Kokubun (1996). Two-component auroral electrojet: Importance for substorm studies. *Journal of Geophysical Research*, 101(A6), 13027–13046, doi: 10.1029/96JA00142.

Kamide, Y., & G. Rostoker (2004). What is the physical meaning of the AE index?, *Eos Transactions AGU*, *85*(19), 188–192, doi:10.1029/2004EO190010.

Katus, R. M., & M. W. Liemohn (2013). Similarities and differences in low- to middle-latitude geomagnetic indices. *Journal of Geophysical Research Space Physics, 118*, 5149–5156, doi:10.1002/jgra.50501.

Katus, R. M., M. W. Liemohn, D. L. Gallagher, A. Ridley, & S. Zou (2013). Evidence for potential and inductive convection during intense geomagnetic events using normalized superposed epoch analysis, *Journal of Geophysical Research Space Physics, 118*, doi:10.1029/2012JA017915.

Khazanov, G. V., T. S. Newman, M. W. Liemohn, M.-C. Fok, & R. W. Spiro (2003). Self-consistent magnetosphere-ionosphere coupling: theoretical studies. *Journal of Geophysical Research Space Physics*, *107*(A3), 1122, doi: 10.1029/2002JA009624.

Klimas, A. J., D. Vassiliadis, & D. N. Baker (1998). *Dst* index prediction using data-derived analogues of the magnetospheric dynamics. *Journal of Geophysical Research Space Physics*, *103*, 20,435–20,448, doi:10.1029/98JA01559.

Li, X., K. S. Oh, & M. Temerin (2007). Prediction of the *AL* index using solar wind parameters. *Journal of Geophysical Research Space Physics 112*, A06224, doi:10.1029/2006JA011918.

Liemohn, M. W., & J. U. Kozyra (2005). Testing the hypothesis that charge exchange can cause a two-phase decay. In *The Inner Magnetosphere: Physics and Modeling, AGU Monogr. Ser.*, vol. 155, edited by T. I. Pulkkinen, N. Tsyganenko, and R. H. W. Friedel, p. 211, Am. Geophys. Un., Washington, D. C..

1196 Liemohn, M. W., & M. Jazowski (2008). Ring current simulations of the 90 intense storms
1197 during solar cycle 23. *Journal of Geophysical Research Space Physics, 113*, A00A17,
1198 doi: 10.1029/2008JA013466, 2008.

1199 Liemohn, M. W., & R. Katus (2012). Is the storm time response of the inner magnetospheric hot
1200 ions universally similar or driver dependent?, *Journal of Geophysical Research Space*
1201 *Physics, 117*, A00L03, doi:10.1029/2011JA017389.

1202 Liemohn, M. W., J. U. Kozyra, V. K. Jordanova, G. V. Khazanov, M. F. Thomsen, & T. E.
1203 Cayton (1999). Analysis of early phase ring current recovery mechanisms during
1204 geomagnetic storms. *Geophysical Research Letters, 25*, 2845.

1205 Liemohn, M. W., J. U. Kozyra, M. F. Thomsen, J. L. Roeder, G. Lu, J. E. Borovsky, & T. E.
1206 Cayton (2001). Dominant role of the asymmetric ring current in producing the stormtime
1207 Dst*. *Journal of Geophysical Research Space Physics, 106*, 10,883.

1208 Liemohn, M. W., A. J. Ridley, D. L. Gallagher, D. M. Ober, & J. U. Kozyra (2004). Dependence
1209 of plasmaspheric morphology on the electric field description during the recovery phase
1210 of the April 17, 2002 magnetic storm. *Journal of Geophysical Research Space Physics,*
1211 *109*(A3), A03209, doi: 10.1029/2003JA010304.

1212 Liemohn, M. W., M. Jazowski, J. U. Kozyra, N. Ganushkina, M. F. Thomsen, & J. E. Borovsky
1213 (2010). CIR vs. CME drivers of the ring current during intense magnetic storms.
1214 *Proceedings of the Royal Society A*, *466*(2123): 3305-3328, doi: 10.1098/rspa.2010.0075,
1215 2010.

1216 Liemohn, M. W., N. Y. Ganushkina, D. L. De Zeeuw, L. Rstaetter, M. Kuznetsova, D. T.
1217 Welling, G. Toth, R. Ilie, T. I. Gombosi, & B. van der Holst (2018). Real-time SWMF
1218 and CCMC: assessing the Dst output from continuous operational simulations. *Space*
1219 *Weather*, in press, ms # 2018SW001953.

1220 Lindsay, G. M., C. T. Russell, & J. G. Luhmann (1999). Predictability of *Dst* index based upon
1221 solar wind conditions monitored inside 1 AU. *Journal of Geophysical Research Space*
1222 *Physics*, *104*(A5), 10,335–10,344, doi:10.1029/1999JA900010.

1223 Love, J. J., & J. L. Gannon (2009). Revised Dst and the epicycles of magnetic disturbance: 1958-
1224 2007. *Annales Geophysicae, 27*, 3101–3131.

1225 Lundstedt, H., & P. Wintoft (1994). Prediction of geomagnetic storms from solar wind data with
1226 the use of a neural network. *Annales Geophysicae*, *12*(1), 19–24, doi:10.1007/s00585-
1227 994-0019-2.

1228 Lundstedt, H., H. Gleisner, & P. Wintoft (2002). Operational forecasts of the geomagnetic *Dst*
1229 index. *Geophysical Research Letters*, *29*(24), 2181, doi:10.1029/2002GL016151.

1230 Luo, Bingxian, Xinlin Li, M. Temerin & Siqing Liu (2013). Prediction of the AU, AL, and AE
1231 indices using solar wind parameters. *Journal of Geophysical Research: Space Physics*,
1232 *118*, 12, 7683-7694.

1233 Lyon, J. G., J. A. Fedder, & C. M. Mobarry (2004). The Lyon-Fedder-Mobarry (LFM) global
1234 MHD magnetospheric simulation code. *Journal of Atmospheric and Solar-Terrestrial*
1235 *Physics, 66*, 1333–1350, doi:10.1016/j.jastp.2004.03.020.

1236 Mason, I. B. (1982). A model for assessment of weather forecasts. *Australian Meteorological*
1237 *Magazine, 30*, 291-303.

1238 Mayaud, P. N. (1980). *Derivation, Meaning, and Use of Geomagnetic Indices*, Geophys.
1239 Monogr. Ser., vol. 22, 154 pp., AGU, Washington, D. C., doi:10.1029/GM022.

1240 Maynard, N. C., & A. J. Chen (1975). Isolated cold plasma regions: Observations and their
1241 relation to possible production mechanisms. *Journal of Geophysical Research*, 80(7),
1242 1009–1013, doi: 10.1029/JA080i007p01009.

1243 Mays, M. L., W. Horton, E. Spencer, & J. Kozyra (2009). Real-time predictions of geomagnetic
1244 storms and substorms: Use of the Solar Wind Magnetosphere-Ionosphere System model.
1245 *Space Weather*, *7*, S07001, doi:10.1029/2008SW000459.

1246 McPherron, R. L., & G. Rostoker (1993). Comment on "Prediction of geomagnetic activity" by
1247 C. K. Goertz, Lin-Hua Shan, and R. A. Smith. *Journal of Geophysical Research: Space*
1248 *Physics*, *98*, A5, 7685.

1249 Menvielle, M., Iyemori T., Marchaudon A., & Nosé M. (2011). Geomagnetic Indices. In:
1250 *Geomagnetic Observations and Models. IAGA Special Sopron Book Series, vol 5.*
1251 Mandea M., Korte M. (eds) Springer, Dordrecht.

1252 Merkin, V. G., & J. G. Lyon (2010). Effects of the low-latitude ionospheric boundary condition
1253 on the global magnetosphere. *Journal of Geophysical Research, 115*, A10202,
1254 doi:10.1029/2010JA015461.

1255 Michaelsen J. (1987). Cross-validation in statistical climate models. *Journal of Climate and*
1256 *Applied Meteorology, 26*, 1589-1600, doi: 10.1175/1520-
1257 0450(1987)026<1589:CVISCF>2.0.CO;2

1258 Moldwin, M. B., L. Downward, H. K. Rassoul, R. Amin, & R. R. Anderson (2002). A new
1259 model of the location of the plasmapause: CRRES results. *Journal of Geophysical*
1260 *Research Space Physics, 107*(A11), 1339, doi:10.1029/2001JA009211.

1261 Morley, S. K., Freeman, M. P., & Tanskanen, E. I. (2007). A comparison of the probability
1262 distribution of observed substorm magnitude with that predicted by a minimal substorm
1263 model. *Annales Geophysicae, 25*, 2427-2437, https://doi.org/10.5194/angeo-25-2427-
1264 2007

1265 Morley, S. K., Brito, T. V., & Welling, D. T. (2018). Measures of model performance based on
1266 the log accuracy ratio. *Space Weather, 16*, 69–88.
1267 https://doi.org/10.1002/2017SW001669

1268 Morley, S. K., Welling, D. T., & Woodroffe, J. R. (2018), Perturbed input ensemble modeling
1269 with the space weather modeling framework. *Space Weather, 16*.
1270 https://doi.org/10.1029/2018SW002000

1271 Muller, R. H. (1944). Verification of short-range weather forecasts (a survey of the literature).
1272 *Bulletin of the American Meteorological Society, 25*, 18-27.

1273 Murayama, T. (1982). Coupling function between solar wind parameters and geomagnetic
1274 indices. *Reviews of Geophysics and Space Physics, 20*, 623,
1275 doi:10.1029/RG020i003p00623.

1276 Murphy, A. H. (1996). The Finley Affair: a signal event in the history of forecast verification.
1277 *Weather and Forecasting, 11*, 3-20.

1278 Murphy, A.H. (1988), Skill Scores Based on the Mean Square Error and Their Relationships to
1279 the Correlation Coefficient. *Monthly Weather Review, 116,* 2417–
1280 2424, https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

1281 Newell, P. T., & J. W. Gjerloev (2011). Evaluation of SuperMAG auroral electrojet indices as
1282 indicators of substorms and auroral power. *Journal of Geophysical Research Space*
1283 *Physics, 116*, A12211, doi:10.1029/2011JA016779.

1284 Newell, P. T., T. Sotirelis, K. Liou, & F. J. Rich (2008). Pairs of solar wind-magnetosphere
1285 coupling functions: Combining a merging term with a viscous term works best. *Journal*
1286 *of Geophysical Research Space Physics*, *113*, A04218, doi: 10.1029/2007JA012825.

1287 O'Brien, T. P. (2006). Limits on the complexity of empirical models of magnetic storm
1288 phenomena. *Space Weather*, *4*, S04001, doi:10.1029/2005SW000170.

1289 O'Brien, T. P., & R. L. McPherron (2000a). An empirical phase space analysis of ring current
1290 dynamics: Solar wind control of injection and decay. *Journal of Geophysical Research*
1291 *Space Physics, 105*(A4), 7707–7719, doi:10.1029/1998JA000437.

1292 O'Brien, T. P., & R. L. McPherron (2000b). Forecasting the ring current index Dst in real time.
1293 *Journal of Atmospheric and Solar Terrestrial Physics*, *62*, 1295-1299.
1294 https://doi.org/10.1016/S1364-6826(00)00072-9

1295 O'Brien, T. P., & M. B. Moldwin (2003). Empirical plasmapause models from magnetic indices.
1296 *Geophysical Research Letters, 30*(4), 1152, doi:10.1029/ 2002GL016007.

1297 Orlova, K., Spasojevic, M., & Shprits, Y. (2014). Activity-dependent global model of electron
1298 loss inside the plasmasphere. *Geophysical Research Letters*, *41*, 3744–3751.
1299 https://doi.org/10.1002/2014GL060100

1300 Orlova, K., Shprits, Y., & Spasojevic, M. (2016). New global loss model of energetic and
1301 relativistic electrons based on Van Allen Probes measurements. *Journal of Geophysical*
1302 *Research: Space Physics, 121*, 1308–1314. https://doi.org/10.1002/2015JA021878

1303 Osthus, D., P. C. Caragea, D. Higdon, S. K. Morley, G. D. Reeves, & B. P. Weaver (2014).
1304 Dynamic linear models for forecasting of radiation belt electrons and limitations on
1305 physical interpretation of predictive models. *Space Weather*, 12, 426–446, doi:
1306 10.1002/2014SW001057.

1307 Ozeke, L. G., Mann, I. R., Murphy, K. R., Rae, I. J., & Milling, D. K. (2014). Analytic
1308 expressions for ULF wave radiation belt radial diffusion coefficients. *Journal of*
1309 *Geophysical Research: Space Physics, 119*, 1587–1605.
1310 https://doi.org/10.1002/2013JA019204.

1311 Pallocchia, G., E. Amata, G. Consolini, M. F. Marcucci, & I. Bertello (2008). *AE* index forecast
1312 at different time scales through an ANN algorithm based on L1 IMF and plasma
1313 measurements, *Journal of Atmospheric and Solar Terrestrial Physics*, *70*(2–4), 663–668,
1314 doi:10.1016/j.jastp.2007.08.038.

1315 Pulkkinen, A., L. Rastätter, M. Kuznetsova, H. Singer, C. Balch, D. Weimer, et al. (2013).
1316 Community-wide validation of geospace model ground magnetic field perturbation

predictions to support model transition to operations. *Space Weather, 11*, 369–385, doi:10.1002/swe.20056.

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

Rastätter, L., M. M. Kuznetsova, A. Glocer, D. Welling, X. Meng, J. Raeder, et al. (2013). Geospace environment modeling 2008–2009 challenge: D_{st} index. *Space Weather*, *11*, 187–205, doi: 10.1002/swe.20036.

Reiff, P. H. (1990). The use and misuse of statistics in space physics. *Journal of Geomagnetism and Geoelectricity, 42*, 1145–1174, doi:10.5636/jgg.42.1145.

Revallo, M., F. Valach, P. Hejda, & J. Bochníček (2014). A neural network Dst index model driven by input time histories of the solar wind–magnetosphere interaction. *Journal of Atmospheric and Solar-Terrestrial Physics, 110–111*, 9-14, doi: 10.1016/j.jastp.2014.01.011

Rostoker, G. (1972). Geomagnetic indices. *Reviews of Geophysics and Space Physics*, *10*, 935–950.

Saiz, E., C. Cid, & Y. Cerrato (2008). Forecasting intense geomagnetic activity using interplanetary magnetic field data. *Annales Geophysicae, 26*, 3989-3998, doi:10.5194/angeo-26-3989-2008.

Savani, N. P., A. Vourlidas, I. G. Richardson, A. Szabo, B. J. Thompson, A. Pulkkinen, M. L. Mays, T. Nieves-Chinchilla, & V. Bothmer (2017). Predicting the magnetic vectors within coronal mass ejections arriving at Earth: 2. Geomagnetic response. *Space Weather, 15*, doi: 10.1002/2016SW001458.

Shen, Chao, Zhenxing Liu & Toyohisa Kamei (2002). A physics-based study of the Dst-AL relationship. *Journal of Geophysical Research: Space Physics*, *107*, A1, SMP 4-1-SMP 4-10.

Sheskin, D. J. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures*, 4th ed., Chapman and Hall/CRC, Boca Raton, Fla.

Simpson, S. (2003). From research model to forecasting tool. *Space Weather, 1*, 1009, doi: 10.1029/2003SW000029.

Spasojevic, M., Shprits, Y. Y., & Orlova, K. (2015). Global empirical models of plasmaspheric hiss using Van Allen Probes. *Journal of Geophysical Research: Space Physics, 120*, 10,370–10,383. https://doi.org/10.1002/2015JA021803

Spencer, E., W. Horton, M. L. Mays, I. Doxas, & J. Kozyra (2007). Analysis of the 3–7 October 2000 and 15–24 April 2002 geomagnetic storms with an optimized nonlinear dynamical model. *Journal of Geophysical Research Space Physics, 112*, A04S90, doi:10.1029/2006JA012019.

Spencer, E., A. Rao, W. Horton, & M. L. Mays (2009). Evaluation of solar wind-magnetosphere coupling function during geomagnetic storms with the WINDMI model. *Journal of Geophysical Research Space Physics*, *114*, A02206, doi:10.1029/2008JA013530.

1358 Stauning, P. (2013). The Polar Cap index: A critical review of methods and a new approach.
1359 *Journal of Geophysical Research Space Physics*, *118*, 5021–5038,
1360 doi:10.1002/jgra.50462.

1361 Stephenson, D. B. (2000). Use of the "odds ratio" for diagnosing forecast skill. *Weather*
1362 *Forecasting*, *15*, 221-232, https://doi.org/10.1175/1520-
1363 0434(2000)015<0221:UOTORF>2.0.CO;2

1364 Stern, D. P. (1975). The motion of a proton in the equatorial magnetosphere. *Journal of*
1365 *Geophysical Research*, *80*, 595.

1366 Sugiura, M. (1964). Hourly values of equatorial *Dst* for the IGY. *Annals of the International*
1367 *Geophysical Year*, *35*, 9–45.

1368 Sugiura, M., & Chapman, S. (1960). The average morphology of geomagnetic storms with
1369 sudden commencement. In *Abhandlungen der Akademie der Wissenschaften zu Göttingen*
1370 (pp. 1–53). Göttingen: Göttingen Math. Phys. Kl., Sondernheft Nr.4.

1371 Swets, J. A. (1973). The relative operating characteristic in psychology. *Science*, *182*, 990-1000.

1372 Takahashi, K., B. A. Toth, & J. V. Olson (2001). An automated procedure for near-real-time *Kp*
1373 estimates. *Journal of Geophysical Research Space Physics*, 106, 21,017–21,032,
1374 doi:10.1029/2000JA000218.

1375 Takalo, J., & J. Timonen (1997). Neural network prediction of *AE* data. *Geophysical Research*
1376 *Letters*, *24*(19), 2403–2406, doi:10.1029/97GL02457.

1377 Taylor, J. R. (1997). *An introduction to error analysis*. Mill Valley, CA, USA: University
1378 Science Books.

1379 Temerin, M., & X. Li (2002). A new model for the prediction of *Dst* on the basis of the solar
1380 wind. *Journal of Geophysical Research Space Physics*, *107*(A12), 1472,
1381 doi:10.1029/2001JA007532.

1382 Temerin, M., & X. Li (2006). Dst model for 1995-2002. *Journal of Geophysical Research Space*
1383 *Physics*, *111*, A04221.

1384 Thomsen, M. F. (2004). Why *Kp* is such a good measure of magnetospheric convection. *Space*
1385 *Weather*, *2*, S11004, doi:10.1029/2004SW000089.

1386 Tobiska, W. K., D. Knipp, W. J. Burke, D. Bouwer, D. Odstrcil, M. P. Hagan, J. Gannon, & B.
1387 R. Bowman (2013). The Anemomilos prediction methodology for Dst. *Space Weather*,
1388 *11*, 490-508, doi: 10.1002/swe.20094.

1389 Toffoletto, F. R., S. Sazykin, R.W. Spiro, & R. A.Wolf (2003). Modeling the inner
1390 magnetosphere using the Rice Convection Model (review). *Space Science Reviews*,
1391 *WISER Special Issue*, *107*, 175–196, doi:10.1023/A:1025532008047.

1392 Tóth, G., Igor V. Sokolov, Tamas I. Gombosi, David R. Chesney, C. Robert Clauer, Darren L.
1393 De Zeeuw, et al. (2005). Space weather modeling framework: A new tool for the space
1394 science community. *Journal of Geophysical Research*, *110*(A12), A12226,
1395 doi:10.1029/2005JA011126.

1396 Toth, G., B. van der Holst, I. V. Sokolov, D. L. De Zeeuw, T. I. Gombosi, F. Fang, et al. (2012).
1397 Adaptive numerical algorithms in space weather modeling. *Journal of Computational*
1398 *Physics, 231*, 870-903. https://doi.org/10.1016/j.jcp.2011.02.006

1399 Troshichev, O. A., V. G. Andrezen, S. Vennerstrøm, and E. Friis-Christensen (1988). Magnetic
1400 activity in the polar cap—A new index. *Planetary and Space Science*, *11*, 1095–1102.

1401 Tsubouchi, K., and Y. Kubo (2010). Quantitative assessment of the probability forecast for the
1402 geomagnetic storm occurrence. *Space Weather, 8*, S12007, doi: 10.1029/2010SW000614.

1403 Volland, H. (1973). A semiempirical model of large-scale magnetospheric electric fields.
1404 *Journal of Geophysical Research, 78*, 171.

1405 Volland, H. (1975). Differential rotation of the magnetospheric plasma as cause of the
1406 Svalgaard-Mansurov effect. *Journal of Geophysical Research*, 80(16), 2311–2315, doi:
1407 10.1029/JA080i016p02311.

1408 Wang, Y., C. L. Shen, S. Wang, & P. Z. Ye (2003). An empirical formula relating the
1409 geomagnetic storm's intensity to the interplanetary parameters: VB_z and t. *Geophysical*
1410 *Research Letters, 30*(20), 2039, doi:10.1029/2003GL017901.

1411 Wanliss, J.A., & K. M. Showalter (2006). High-resolution global storm index: Dst versus SYM-
1412 H. *Journal of Geophysical Research Space Physics, 111,* A02202,
1413 doi:10.1029/2005JA011034.

1414 Wei, H. L., S. A. Billings & M. Balikhin (2004). Prediction of the Dst index using
1415 multiresolution wavelet models, *Journal of Geophysical Research: Space Physics*, *109*,
1416 A7.

1417 Wei, H.L., D.Q. Zhu, S.A. Billings & M.A. Balikhin (2007). Forecasting the geomagnetic
1418 activity of the Dst index using multiscale radial basis function networks. *Advances in*
1419 *Space Research*, 10.1016/j.asr.2007.02.080, **40**, 12, 1863-1870.

1420 Weigel, R. S., T. Detman, E. J. Rigler, & D. N. Baker (2006), Decision theory and the analysis of
1421 rare event space weather forecasts, *Space Weather*, doi: 10.1029/2005SW000157.

1422 Weigel, R. S. (2010). Solar wind density influence on geomagnetic storm intensity. *Journal of*
1423 *Geophysical Research Space Physics*, *115*, A09201, doi: 10.1029/ 2009JA015062.

1424 Wilks, D. (2001). A skill score based on economic value for probability forecasts. Meterol.
1425 Appl., 8, 209 – 219.

1426 Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences* (2nd ed.). Oxford:
1427 Academic Press.

1428 Wiltberger, M., W. Wang, A. G. Burns, S. C. Solomon, J. G. Lyon, & C. C. Goodrich (2004).
1429 Initial results from the coupled magnetosphere ionosphere thermosphere model:
1430 Magnetospheric and ionospheric responses. *Journal of Atmospheric and Solar-Terrestrial*
1431 *Physics, 66*, 1411–1423, doi:10.1016/j.jastp.2004.03.026.

1432 Wing, S., J. R. Johnson, J. Jen, C.-I. Meng, D. G. Sibeck, K. Bechtold, J. Freeman, K. Costello,
1433 M. Balikhin, & K. Takahashi (2005). Kp forecast models, *Journal of Geophysical*
1434 *Research Space Physics, 110*, A04203, doi: 10.1029/2004JA010500.

1435 Wintoft, P., M. Wik, J. Matzka, & Y. Shprits (2017). Forecasting Kp from solar wind data: input
1436 parameter study using 3-hour averages and 3-hour range values. *Journal of Space*
1437 *Weather and Space Climate, 7*, A29, doi: 10.1051/swsc/2017027.

1438 Wolf, R. A., R. W. Spiro, & F. J. Rich (1991). Extension of convection modeling into the high-
1439 latitude ionosphere—some theoretical difficulties. *Journal of Atmospheric and*
1440 *Terrestrial Physics, 53*, 817–829, doi:10.1016/0021-9169(91)90096-P.

1441 Wu, J.-G., & H. Lundstedt (1997). Geomagnetic storm predictions from solar wind data with the
1442 use of dynamic neural networks. *Journal of Geophysical Research Space Physics*, *102*,
1443 14,255–14,268, doi:10.1029/97JA00975.

1444 Yu, Y., V. K. Jordanova, A. J. Ridley, G. Toth & R. Heelis (2017). Effects of electric field
1445 methods on modeling the midlatitude ionospheric electrodynamics and inner
1446 magnetosphere dynamics. *Journal of Geophysical Research: Space Physics*, *122*, 5,
1447 5321-5338.

1448 Zaharia, S., C. Z. Cheng, & K. Maezawa (2004). 3-D force-balanced magnetospheric
1449 configurations. *Annales Geophysicae, 22*, 251-265.

1450 Zaharia, S., V. K. Jordanova, M. F. Thomsen, & G. D. Reeves (2006). Self-consistent modeling
1451 of magnetic fields and plasmas in the inner magnetosphere: Application to a geomagnetic
1452 storm. *Journal of Geophysical Research, 111*, A11S14, doi:10.1029/2006JA011619.

1453 Zaharia, S., V. K. Jordanova, M. F. Thomsen, & G. D. Reeves (2008). Self-consistent
1454 geomagnetic storm simulation: The role of the induced electric fields. *Journal of*
1455 *Atmospheric and Solar-Terrestrial Physics, 70*, 511, doi:10.1016/j.jastp. 2007.08.067.

1456 Zhang, J., I. G. Richardson, D. F. Webb, N. Gopalswamy, E. Huttunen, J. C. Kasper, et al.
1457 (2007). Solar and interplanetary sources of major geomagnetic storms (Dst ≤ -100 nT)
1458 during 1996 – 2005. *Journal of Geophysical Research Space Physics, 112*, A10102,
1459 doi:10.1029/2007JA012321.

1460 Zhang, X.Y., & M. B. Moldwin (2015). Probabilistic forecasting analysis of geomagnetic indices
1461 for southward IMF events. *Space Weather, 13*, doi: 10.1002/2014SW001113.