AGU100 ADVANCING EARTH AND SPACE SCIENCE

Space Weather

RESEARCH ARTICLE
10.1029/2018SW002000

Special Section:
Space Weather Capabilities Assessment

Perturbed Input Ensemble Modeling With the Space Weather Modeling Framework

S. K. Morley[1], D. T. Welling[2,3], and J. R. Woodroffe[1]

[1]Space Science and Applications, Los Alamos National Laboratory, Los Alamos, NM, USA, [2]Climate and Space Sciences and Engineering Department, University of Michigan, Ann Arbor, MI, USA, [3]Department of Physics, University of Texas at Arlington, Arlington, TX, USA

Key Points:
- A new nonparametric method for drawing different realizations of solar wind data to drive magnetospheric models is derived
- The new method is used to obtain uncertainties on predicted geophysical indices from the operational Space Weather Modeling Framework
- Model skill can be improved by considering the uncertainty on model input

Correspondence to:
S. K. Morley,
smorley@lanl.gov

Citation:
Morley, S. K., Welling, D. T., & Woodroffe, J. R. (2018), Perturbed input ensemble modeling with the space weather modeling framework. *Space Weather*, *16*, 1330–1347. https://doi.org/10.1029/2018SW002000

Received 3 JUL 2018
Accepted 8 AUG 2018
Accepted article online 23 AUG 2018
Published online 12 SEP 2018

Abstract To assess the effect of uncertainties in solar wind driving on the predictions from the operational configuration of the Space Weather Modeling Framework, we have developed a nonparametric method for generating multiple possible realizations of the solar wind just upstream of the bow shock, based on observations near the first Lagrangian point. We have applied this method to the solar wind inputs at the upstream boundary of Space Weather Modeling Framework and have simulated the geomagnetic storm of 5 April 2010. We ran a 40-member ensemble for this event and have used this ensemble to quantify the uncertainty in the predicted Sym-H index and ground magnetic disturbances due to the uncertainty in the upstream boundary conditions. Both the ensemble mean and the unperturbed simulation tend to underpredict the magnitude of Sym-H in the quiet interval before the storm and overpredict in the storm itself, consistent with previous work. The ensemble mean is a more accurate predictor of Sym-H, improving the mean absolute error by nearly 2 nT for this interval and displaying a smaller bias. We also examine the uncertainty in predicted maxima in ground magnetic disturbances. The confidence intervals are typically narrow during periods where the predicted dB_H/dt is low. The confidence intervals are often much wider where the median prediction is for enhanced dB_H/dt. The ensemble also allows us to identify intervals of activity that cannot be explained by uncertainty in the solar wind driver, driving further model improvements. This work demonstrates the feasibility and importance of ensemble modeling for space weather applications.

Plain Language Summary Forecasts of space weather usually rely on spacecraft measurements of the solar wind from about a million miles away from Earth. Like water flowing toward a rock in a stream, measurements at a single point upstream may not reflect exactly what will hit the Earth. Forecasts that are driven by these measurements have uncertainty due to the uncertainty in the measurements driving the forecast models. We have developed a technique to estimate the uncertainty on space weather predictions using 7 years of solar wind measurements from two satellites. We have performed computer simulations of the same geomagnetic storm 41 times. In each simulation, the inputs were modified slightly each time to reflect the uncertainty in the measurements. By considering the set of simulations as a whole, we have shown that space weather forecasts can be improved by accounting for the uncertainty in the input data. We have also shown that accounting for uncertainty in the data driving, the model can highlight where incorrect forecasts are due to the uncertainty, as well as where they are due to inadequacies in the model itself. This work shows the importance of ensemble methods and accounting for uncertainties in space weather simulation and forecasting.

1. Introduction

Most space weather modeling consists of applying deterministic equations to an assumed initial condition and subsequently calculating a single predicted value for each output parameter. For example, typical models to predict relativistic electron flux at geosynchronous orbit (e.g., Osthus et al., 2014) use measurements from an upstream solar wind monitor to specify the solar wind state and an estimate of the relativistic electron flux at a previous time, before applying a set of deterministic equations to predict the flux at the following time step. Similarly, predictions of geomagnetic indices such as the Kp index typically take a set of inputs including solar wind data and use models of varying complexity to predict a single value of the required index per time step (e.g., Haiducek et al., 2017; Wing et al., 2017).

Ensembles of model output are widely used for assessing uncertainties in model predictions (Slingo & Palmer, 2011). Ensemble modeling has a rich history across weather and climate research (e.g., Epstein, 1969; Kay et al., 2015; Murphy et al., 2004; Owen & Palmer, 1987) but is relatively recent in its application to space weather (e.g., Andriyas et al., 2012; Cash, Biesecker, Pizzo, et al., 2015; Knipp, 2016; Murray, 2018; Riley et al., 2013). Approaches to ensemble modeling include multimodel ensembles (Guerra et al., 2015), single-model perturbed physics ensembles (Murphy et al., 2004; Smithtro & Sojka, 2005), and perturbed initial condition ensembles (Kay et al., 2015; Morley, 2008).

Multimodel ensembles combine predictions from different models, often using some sort of weighted averaging (Barnston et al., 2003; Murray, 2018); Guerra et al. (2015) used a linear combination of results from four different probabilistic flare prediction models to develop a better performing ensemble forecast. Perturbed physics ensembles use the same model, but parameter values within the model are varied to produce different simulation results. An example of this approach is given by Cash, Biesecker, Pizzo, et al. (2015) who varied the parameters used in fitting a coronal mass ejection (CME), including initial CME speed and angular width, to study the uncertainty of the predicted CME arrival time. The perturbed initial condition ensemble method explores the problem identified by Lorenz (1963), namely, that small perturbations in the definition of the model's initial state can lead to different temporal evolution in the simulation. Kay et al. (2015) ran 30 different climate simulations using the same model and external forcings, where the difference between ensemble members was numerical differences, at the scale of floating-point roundoff, in the atmospheric initial condition.

Boundary conditions are particularly important in driven systems like the magnetosphere (e.g., Borovsky & Valdivia, 2018; Vassiliadis et al., 1995). Recently, Chen et al. (2018) studied an ensemble of inner magnetosphere simulations using the Rice Convection Model (RCM)-Equilibrium (Lemon et al., 2004) where the electric field boundary condition was varied using a statistical model of errors in the cross-polar cap potential drop. Using this perturbed boundary condition, Chen et al. (2018) determined that uncertainty in the applied electric field boundary condition was of secondary importance compared to inadequately capturing the physics of particle loss within the model. Given a sufficient number of ensemble members to adequately describe the probability density function of a predictand, such as the Dst index, this approach allows a direct determination of the uncertainty in the model output that results from uncertainty in the boundary condition.

In this work we consider the uncertainty in the output of a model driven by upstream solar wind data due to the uncertain specification of the true state of the solar wind interacting with the magnetosphere. We specifically consider the case of a space weather model that uses, as input, solar wind data from a monitor orbiting the first Lagrangian point (L1). We will first describe some of the issues leading to an uncertain specification of the solar wind properties that interact with the Earth. We will then describe a nonparametric resampling approach to estimating possible realizations of the solar wind interacting with Earth. Given a resampling model of perturbed solar wind time series, we use an ensemble of different realizations of the solar wind to drive a perturbed-input ensemble of simulations using the Space Weather Modeling Framework. We then assess, for the first time, the uncertainty in the modeling due to the uncertainty in the solar wind input.

2. Uncertainties in Specifying the Solar Wind State for Magnetospheric Modeling

Measurements of the solar wind plasma and of the interplanetary magnetic field (IMF) from an L1 solar wind monitor, such as the Advanced Composition Explorer (ACE; Stone et al., 1998) or Deep Space Climate Observatory (DSCOVR; Cash, Biesecker, Reinard, et al., 2015), are point measurements in a turbulent medium that has varying correlation scales. To use these data to drive a model of the geospace environment, an estimate is made of the solar wind plasma and IMF arriving at the bow shock nose. A variety of methods is used to propagate the upstream measurements to the bow shock nose, all of which can be shown to have errors in the arrival time based on observed structures in the solar wind.

Figure 1 shows a schematic of the solar wind as it propagates toward Earth's magnetosphere, in the Geocentric Solar Ecliptic coordinate system (e.g., Fränz & Harper, 2002), based on the illustrations of Mailyan et al. (2008). The orbit of the upstream monitor is that of the ACE spacecraft. The orbit of Geotail, which we use as a near-Earth monitor, is also shown. The equatorial locations of the magnetopause and bow shock are also shown, using the Shue et al. (1997) and Chao et al. (2002) models, respectively. For this schematic, we used nominal input conditions for each model. It can be clearly seen that the size of the L1 halo orbit is larger than

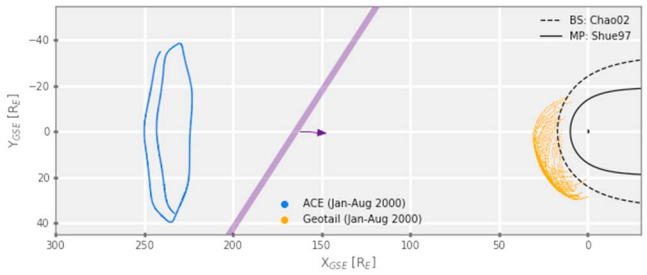

Figure 1. Schematic showing, in the *X*-*Y* GSE plane, the projection of the first Lagrangian point halo orbit of ACE in blue, and of the orbit of Geotail whenever it is upstream of the bow shock in yellow. The *X* axis in GSE coordinates is defined by the vector from the Earth to the Sun. The *Y* axis lies in the ecliptic plane and is positive in Earth's antiorbit direction. Nominal locations for the bow shock and magnetopause are shown by the black dashed and solid lines, respectively. The purple bar indicates a planar structure in the solar wind that is propagating toward Earth with the solar wind velocity, in the direction indicated by the purple arrow. The orbital projections are based on 8 months of data from 2000. GSE = geocentric solar equatorial; ACE = Advanced Composition Explorer; BS = Bow shock; MP = Magnetopause.

the width of the magnetosphere. The transparent purple bar is intended to illustrate a planar front in the IMF, perpendicular to the Parker spiral, that propagates radially outward toward Earth. As the solar wind is not homogeneous along the front illustrated here, we can identify a number of key sources of uncertainty in our solar wind measurement as a driver for a space weather model:

1. Our upstream monitor orbits around the L1 point but is rarely sampling a ballistic trajectory that would reach the nose of the bow shock (e.g., Borovsky, 2017).
2. Solar wind propagation methods assume a certain homogeneity in the solar wind that is being propagated, while observations suggest that the plasma and magnetic field are not homogeneous (e.g., Kessel et al., 1999; Borovsky, 2008, 2017).
3. The solar wind properties are discontinuous across boundaries between regions with scale sizes approaching the cross section of the magnetosphere (e.g., Borovsky, 2012, 2017).
4. The propagation method itself is not perfect and can introduce some uncertainty in the parameters projected to be arriving at the bow shock (e.g., Case & Wild, 2012; Cash et al., 2016).

In the absence of three-dimensional observations of the solar wind as it propagates from L1 toward Earth, it is difficult to disentangle these sources of uncertainty. Some authors have explored the differences between propagation methods (e.g., Cash et al., 2016; Mailyan et al., 2008), and Pulkkinen and Rastätter (2009) have examined the differences in predicted ground magnetic disturbances using different propagation methods. Accurate prediction of the solar wind conditions just upstream of the Earth, based on measurements near L1, is further complicated by nonplanarity of solar wind phase fronts and the fact that the solar wind evolves between L1 and the Earth (Kessel et al., 1999; Tsurutani et al., 2005).

3. Error Model for Solar Wind Inputs

For this work we assume that a solar wind monitor close to, and upstream of, Earth's bow shock provides a better representation of the solar wind that is interacting with the magnetosphere. We use Geotail (Nishida et al., 1992) as our near-Earth monitor and use its plasma and IMF measurements as a ground truth. We then use the point measurements from ACE as our estimate of the solar wind state. Prior to estimating the error in the estimated state, we account for propagation of the solar wind by using the spacecraft-specific OMNI data set (King & Papitashvili, 2005; Papitashvili et al., 2014). Both data sources are lagged to the location of the bow shock nose using the same method. The error between our upstream measurement and our near-Earth measurement is then given by the difference between the propagated ACE data and the propagated Geotail data. For this work we use data from January 1999 through December 2005.

$$\varepsilon_X = X_{\mathrm{ACE}} - X_{\mathrm{Geotail}}. \tag{1}$$

A major source of uncertainty is the structure within the solar wind. That is the upstream solar wind monitor may not be measuring the same plasma that eventually interacted with the magnetosphere. Systematic differences due to structure in the solar wind plasma and magnetic field will result in significant persistence in the time series of errors in any given parameter. In other words, the error $\varepsilon_X(t)$ will be correlated with the error at a previous time $\varepsilon_X(t - \Delta t)$. Additionally, errors in components of the IMF are likely to be correlated. For example, the orientation of the field may be the same, but the observed magnitude differs between ACE and Geotail, leading to a correlated error in each component. Alternatively, the measured magnitude may be the same, but the observed clock angle may be different, again leading to correlation between the errors in the components of the IMF.

To model the expected solar wind parameters observed at Geotail, given only measured data from ACE, we need to apply errors that are consistent with those observed, as described above. Several approaches can be taken here, and we briefly describe initial approaches taken in the preliminary stages of this work, followed by the method chosen for this application. While our initial methods have caveats that limited their utility for this particular work, they may well be suitable for perturbed input ensemble modeling of different systems.

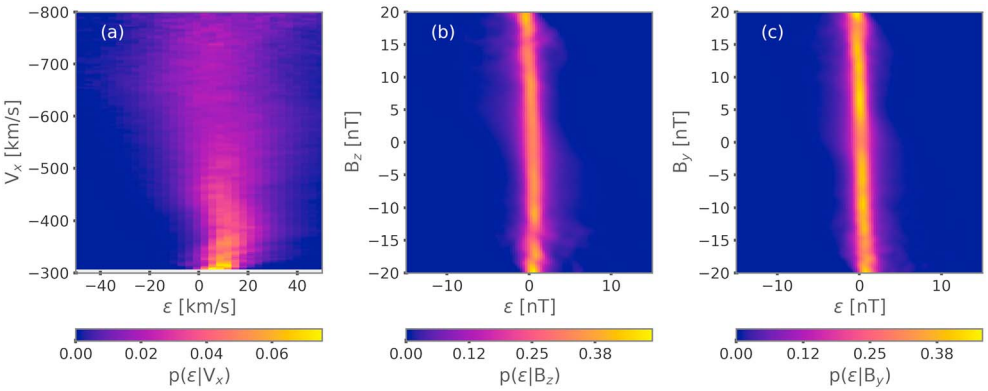

Figure 2. Bivariate probability density functions where the ordinate in each panel is the variable as measured at Advanced Composition Explorer (ACE) and the abscissa in each panel is the error between ACE and Geotail. The color in each case shows the conditional probability $p(\varepsilon|X)$. (a) shows the probability of an error given a conditioning value of solar wind V_x. (b) and (c) show the same for IMF B_z and B_y respectively.

The exploration of these methods does, however, provide important information about how the errors vary and are correlated.

3.1. Conditional Probability Distributions of Errors

Hassan et al. (2015) explored the differences in solar wind speed between ACE and Geotail and showed the distributions of solar wind speed for discrete ranges of speed measured at ACE. This work led to the realization that the difference (error) should be the quantity of interest. We therefore initially aimed to characterize the probability density functions of the errors such that new realizations of the solar wind could be drawn by sampling from the error distributions and adding the errors to the upstream measurements.

Following Hassan et al. (2015), we use kernel density estimates (KDEs) to characterize the probability density functions of the observed errors. For each of the parameters that we wish to perturb, we calculate the errors using equation (1) and then fit bivariate KDEs for the joint probability $p(X, \varepsilon_X)$. The form of these probability density functions is dominated by the distribution of the solar wind parameter itself. To understand the distribution of errors at a given value of the upstream parameter, we need to estimate the conditional probability by

$$f(\varepsilon_X|X = x_{\mathrm{ACE}}) = \frac{f(X_{\mathrm{ACE}}, \varepsilon_X)}{f(X_{\mathrm{ACE}})}, \tag{2}$$

where f represents a probability density function, X represents the variable, and x represents a realization of X. We refer to the distributions using upper case and to individual values or variates using lower case. To restate equation (2) in words, each slice of the bivariate joint probability density function is normalized by the probability of that value of X, such that the area under each slice sums to 1.

Figure 2 shows the bivariate KDEs for the conditional probabilities. The ordinate in each panel is the variable as measured at ACE, and the abscissa in each panel is the error between ACE and Geotail. The color in each case shows the conditional probability $p(\varepsilon|X)$. Figure 2a shows that the distribution of errors between ACE and Geotail is narrower at low solar wind speed and broader at high solar wind speed. By contrast, Figures 2b and 2c show that the distribution of errors in the transverse magnetic field components between ACE and Geotail is very narrow and does not vary significantly with the magnitude of the component. That is, these bivariate distributions support an assumption that the errors in the transverse components of the IMF are conditionally independent of the magnitude of the component.

In the absence of autocorrelation but in the presence of conditional dependence, we can empirically determine the probability density function of the error, given the value measured at our upstream monitor, following the method given above. We can then sample directly from that probability density function using, for example, Monte Carlo rejection sampling (e.g., Mackay, 1998). Specifically, for each time step, we would find $f(\varepsilon_X|X = x_{\mathrm{ACE}})$ and draw a random variate from the conditional probability distribution. We can then add the randomly drawn error to the upstream measurement. Thus, at each time step, given an upstream value x, we can draw an ensemble of likely alternate states given by $x + \varepsilon_X$. As the differences between upstream and

near-Earth measurements arise, at least in part, from structure, we expect autocorrelation in the time series of errors. This approach does not capture any autocorrelation, and we require that our error model adequately captures temporal correlations between errors.

As noted previously, Figures 2b and 2c demonstrate that the errors in the transverse components of the IMF (B_y and B_z) are largely independent of the magnitude of the components; therefore, we can treat these variables as conditionally independent. Assuming that the errors can be described by a first-order autoregressive model, we can then estimate the conditional probability of ε at time t given the value of ε at time $t-1$

$$f(\varepsilon_t|\varepsilon_{t-1}) = \frac{f(\varepsilon_{t-1}, \varepsilon_t)}{f(\varepsilon_{t-1})}. \tag{3}$$

As before these conditional probability density functions can then be directly sampled to draw an ensemble of different realizations of X, given our error model. That is, the error at time t is drawn as a random variate from the distribution of errors specified by $f(\varepsilon_t|\varepsilon_{t-1})$. However, $\varepsilon(t)$ has longer-range autocorrelations than implied by the conditional probability model, and this assumption leads to a large high-frequency variability in the different realizations of the solar wind parameter that is unrealistic. Attempts to use first-order autoregressive models, either by empirically specifying the conditional probability distributions or by fitting a Gaussian AR(1) model, did not adequately capture the correlative structure of the error time series. Additionally, the approach described above treats each parameter independently; correlations between the errors in V_X, B_y, and B_z are not accounted for. Fitting a multidimensional parametric autoregressive model could potentially account for this, as could adopting a sampling method that accounts for both autocorrelations and correlations between variables. We account for these factors by using a block resampling method. The application of a nonparametric method, rather than fitting a parametric model, mitigates the errors associated with both model selection and model fitting (e.g., Vogel & Shallcross, 1996) while still preserving the correlations in time and between variables.

3.2. Block-Resampled Error Model

The bootstrap (Efron, 1979; Efron & Tibshirani, 1986) is a nonparametric method for estimating the uncertainty of a sample statistic using random samples of the same size as the original sample, drawn with replacement from the original sample. This technique is commonly used for estimating errors or confidence intervals (e.g., Kawano & Higuchi, 1995; Morley & Freeman, 2007). A known limitation of resampling with replacement is that correlations between points in the sample are lost (Solow, 1985). The moving block bootstrap (Kunsch, 1989) approach modifies the bootstrap to capture serial dependence in time series by resampling blocks of values rather than individual values. Our sampling methodology is derived from the sampling for the moving block bootstrap, with some minor differences as described below.

Our block-resampled error model uses the time series of observed errors ($\varepsilon_X(t)$) and resamples, with replacement, to draw errors with which to model different realizations of the likely solar wind state near Earth. To capture the observed autocorrelation, we use block resampling. That is, instead of drawing a single value, we randomly select a starting index and draw a contiguous block of errors from $\varepsilon_X(t)$. The block length used in this study is 1 hr (60 samples), and the total number of samples in each error series is 884,658, corresponding to 14,743 possible blocks. By drawing blocks with a length much greater than the correlation scale, the autocorrelations in the error series are preserved. The selected block length is consistent with the rule-of-thumb that the block length should be approximately $N^{1/a}$ where a is between 3 and 4 (Niehof & Morley, 2012, and references therein). For typical solar wind speeds of 300 to 800 km/s, this corresponds to scale lengths of 169–452 R_E, several times larger than typical flux tube diameters in the solar wind (Borovsky, 2017).

To ensure that correlations between errors on different variables are preserved, we use the same starting index to draw errors for all variables. This resampling approach has previously been used for bootstrap confidence intervals for bivariate data, called pairwise-moving block bootstrap resampling (Ólafsdóttir & Mudelsee, 2014), and our approach ensures that correlations between errors in any solar wind parameters we wish to resample are captured. To illustrate, we wish to perturb $V_x(t)$, $B_y(t)$, and $B_z(t)$ where the series has M elements. We begin by selecting an integer, i, from a random uniform distribution with the same length as the set of errors minus the block length (L). This integer is used as the starting index of the block and errors for each variable are then given by $X(i, i+L-1)$, where the term in brackets indicates an inclusive range of numbered elements. This

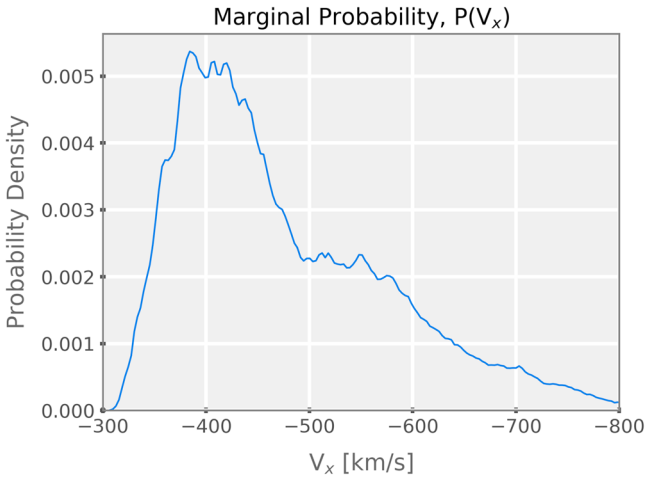

Figure 3. Marginal probability of the solar wind velocity in the X_{GSE} direction.

range of indices is used to draw errors for each variable in turn. We then repeat this process until the entire series has been perturbed.

We note that our block resampling method implicitly assumes that the the errors are conditionally independent. This arises because the start time of each block is randomly chosen and hence the errors within any block are assumed to be representative of all times. As shown in section 3.1, the errors in the B_y and B_z are conditionally independent, but the same is not true for V_x. To qualitatively assess the likely impact of this assumption on our model, we examine the marginal probability $P(Vx)$. Figure 3 shows the probability density function of V_x. The bulk of the distribution lies below 500 km/s where the error distribution can be seen to be narrower (cf. Figure 2a). This suggests that for periods with fast solar wind (>500 km/s), the block resampling might tend to underestimate the errors, and the periods with slow solar wind (<500 km/s), this method is likely to overestimate the errors by occasionally sampling from an error distribution that is broader.

Some further caveats should be noted for the block-resampled error model. First, we assume that the data from both the upstream monitor and the near-Earth monitor are adequately calibrated. This work shows that there are systematic differences between the parameters measured by ACE and by Geotail. For example, the solar wind speed measured at Geotail is typically of order 10 km/s slower than the corresponding measurement at ACE for slow solar wind. The bias in solar wind speed appears to be smaller for faster solar wind, but the differences can also be very much larger. Similarly, the distributions of errors for the transverse IMF components are not centered at exactly 0, and the offset varies slightly with the value of the magnetic field. Our analysis ignores these effects, effectively assuming that any systematic errors are real. As the systematic offsets in the IMF data are small, ignoring them should have minimal effect on our results. Future work should assess the effect of systematic errors and apply any necessary corrections to the data from the solar wind monitors. Second, both the time series of errors and their temporal correlations are likely to vary with the type of solar wind. Some preliminary analysis of the differences in the distributions of solar wind speed between ACE and Geotail was presented by Hassan et al. (2015). Further refinement of our method will be required to account for this type of effect.

Finally, we note that we do not include solar wind number density in this work. While extending our resampling method to include the error in number density between ACE and Geotail would be trivial, the number densities measured by Geotail require additional work to be able to reliably include them in this analysis. The spacecraft-specific OMNI data do not include a cross calibration of the number density, and the Geotail data show systematic differences in number density, relative to upstream monitors, that vary as a function of the number density. Using the radial component of the velocity and the transverse magnetic field components is sufficient to give a good estimate of the variability due to uncertainty in the solar wind state and to demonstrate the methodology. As noted previously, cross calibrations are important for this approach, and we restrict our initial work to parameters that do not display substantial systematic differences.

4. Application: Simulations of Geospace Driven by Solar Wind Inputs

We demonstrate the utility of perturbed input ensemble modeling by running a set of large-scale simulations of the magnetosphere and assessing the uncertainty in the predictions that arise from characterizing the uncertainty in the inputs. For this we have chosen to use the Space Weather Modeling Framework (SWMF; e.g., Tóth et al., 2005, 2012). The SWMF couples together component models to simulate a variety of domains in a self-consistent manner. Here we use a configuration that is the same as the Operational Geospace model currently in use at National Oceanic and Atmospheric Administration's Space Weather Prediction Center (SWPC). Analysis and plotting was performed using the open-source SpacePy (Morley et al., 2010, 2011) and PyForecastTools (Morley, 2018) libraries.

The *operational geospace* configuration of the SWMF couples: (1) the Block-Adaptive-Tree Solar Wind, Roe-Type Upwind Scheme (De Zeeuw et al., 2000; Powell et al., 1999); (2) the RCM (e.g., Toffoletto et al., 2003); and (3) the Ridley Ionosphere Model (Ridley et al., 2003, 2004). A schematic of the coupling is shown in Figure 4. The Block-Adaptive-Tree Solar Wind, Roe-Type Upwind Scheme is an adaptive-mesh magnetohydrodynamic

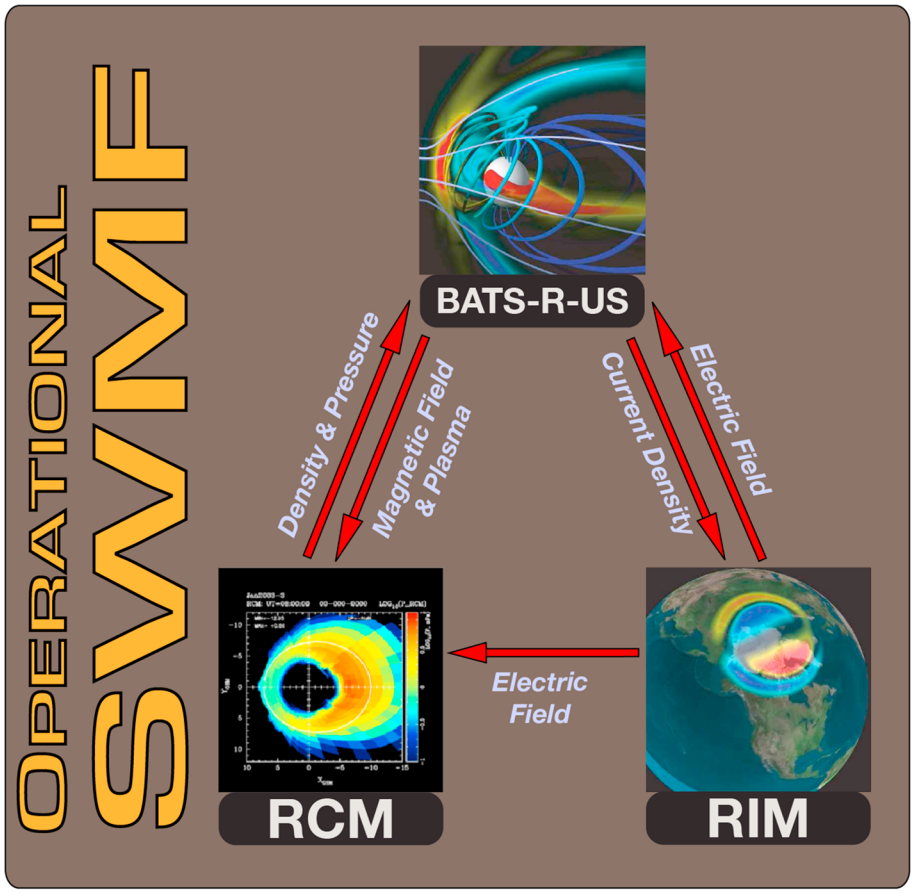

Figure 4. Diagram of the intermodel couplings used by the operational configuration of the Space Weather Modeling Framework (SWMF). RCM = Rice Convection Model; RIM = Ridley Ionosphere Model; BATS-R-US = Block-Adaptive-Tree Solar Wind, Roe-Type Upwind Scheme.

(MHD) solver that solves the ideal MHD equations throughout the magnetosphere. RCM models the inner magnetosphere, and Ridley Ionosphere Model simulates ionospheric electrodynamics. Further details of the operational configuration and its components are given by Pulkkinen et al. (2013) and Haiducek et al. (2017). At the time of writing, the operational forecasts use a single, deterministic simulation and do not provide estimates of the uncertainty of predicted quantities.

4.1. SWPC Challenge Event 5: April 2010 Storm

The event we simulate here is event 5 (hereafter referred to as *event 5*) from the "SWPC Challenge" as described by Pulkkinen et al. (2013). The event covers the interval from midnight on 5 April 2010 through midnight on 6 April 2010, and each simulation was started at 19:00 universal time coordinated (UTC) on 4 April 2010. The minimum Dst in the interval was −73 nT and the maximum Kp was 8^-. This event was selected from the set studied by Pulkkinen et al. (2013) as it had a very strong response in Kp, complete solar wind coverage, and atypically large currents in the nightside ionosphere (Connors et al., 2011).

We applied the error model described in section 3.2 to the solar wind input data used for event 5, such that 40 different realizations of the input solar wind data were generated. The simulations were run on the *Wolf* institutional computing cluster at Los Alamos National Laboratory. Each simulation used approximately 2,500 core hours to complete. In addition to the 40 ensemble members driven with perturbed solar wind inputs, we also ran the unperturbed simulation as a reference. As noted in section 3.2, we perturbed V_x, B_y, and B_z. The number density was not modified from the propagated ACE data. The IMF B_x component was set to 0 to reduce the divergence of the magnetic field in the simulation. This is consistent with the mode of operation used for the study of Pulkkinen et al. (2013).

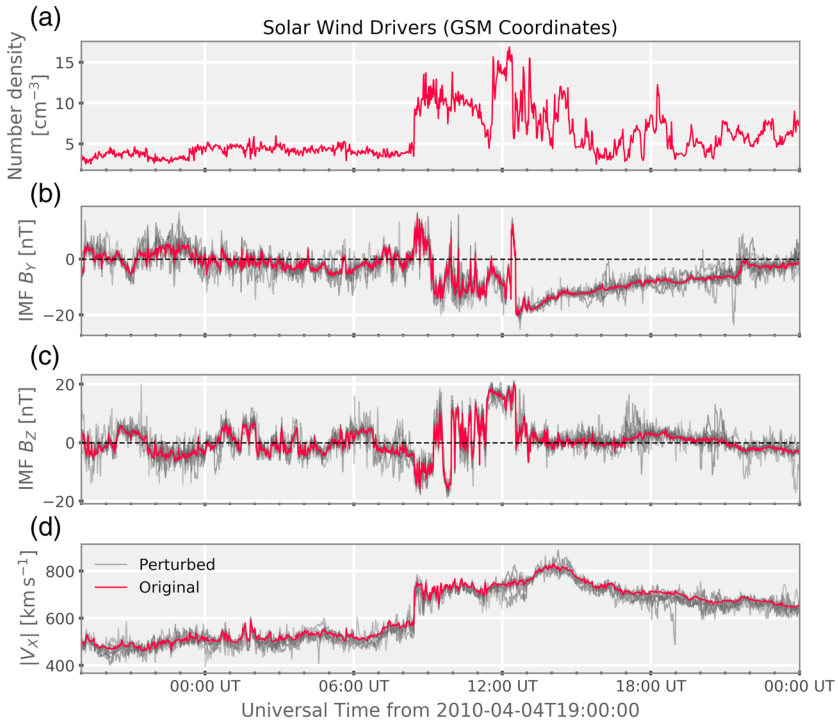

Figure 5. Plots of the key solar wind input parameters for SWPC event 5. All unperturbed inputs from Advanced Composition Explorer are shown in red. Perturbed ensemble members are shown in gray, where the color of each line is varied slightly to help distinguish between ensemble members. For clarity, we only show eight randomly selected ensemble members. (a) shows the solar wind number density, which we did not perturb for this investigation. (b) and (c) shows the IMF B_y and B_z, respectively. (d) shows the magnitude of the x component of the solar wind velocity ($|V_x|$). IMF = interplanetary magnetic field; GSM = geocentric solar magnetospheric.

Analysis of a perturbed-input ensemble allows us to investigate the uncertainty in the model output that arises from uncertainty in the solar wind input. This is conceptually similar to the recent study by Chen et al. (2018) of the effect of uncertain electric field boundary conditions on inner magnetosphere simulations. The aim and the approach are slightly different, however. We use a nonparametric method to perturb our solar wind boundary condition, and we use twice as many ensemble members. This allows us to estimate the probability distribution of model outputs such as the Sym-H and Kp indices to quantify the uncertainty, as well as quantifying the uncertainty on the model skill at predicting threshold crossings in dB/dt. To reiterate, the uncertainty captured by this study is due to imperfect specification of the solar wind that drives the simulation. Any uncertainty due to imperfectly specified physical processes like empirical ionospheric conductance models (Welling et al., 2016) or insufficient grid resolution (Haiducek et al., 2017) is not captured here, though these effects can manifest as observations occurring well outside the expected range of uncertainty estimated in this study.

Figure 5 shows the key solar wind inputs used to drive the SWMF simulations. From top to bottom, the panels show the solar wind number density, the y and z components of the IMF (in geocentric solar magnetospheric coordinates), and the magnitude of the radial component of the solar wind velocity, respectively. The red lines show the observations propagated from ACE to the front of the SWMF simulation domain, and the gray lines show the perturbed solar wind input. For clarity, we here only show eight randomly selected members of the ensemble.

4.2. Geophysical Indices

The Sym-H index can be thought of as a high-resolution version of the Dst index (Wanliss & Showalter, 2006) and, as such, measures the intensity of the ring current. Kp is a 3-hourly range index (Mayaud, 1980) that provides a good measure of general geomagnetic activity and is a good proxy for the strength of magnetospheric convection (Thomsen, 2004). Figure 6a shows the 1-min resolution simulated Sym-H index from SWMF and the observed Sym-H index (at 1-min resolution) for comparison. Results from eight randomly selected ensemble members are shown as gray lines, the ensemble mean is shown by the magenta line, and the simulation

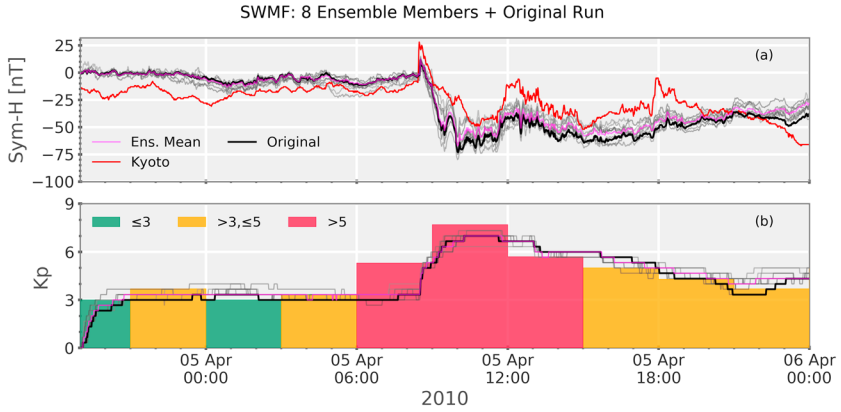

Figure 6. A comparison of modeled and observed geomagnetic indices, Sym-H and Kp, for SWPC event 5. (a) shows the Sym-H index from observation (red), the model run using unperturbed inputs (black), individual ensemble members (gray), and the mean Sym-H calculated from the full ensemble (magenta). (b) shows the observed Kp index in the colored step plot, and the Kp calculated from SWMF is shown in black for the unperturbed run and in gray for the perturbed ensemble members. The ensemble average, calculated from the full ensemble, is shown in magenta. For clarity, we only show eight randomly selected ensemble members in each panel. SWMF = Space Weather Modeling Framework.

result from driving SWMF with just the ACE data is shown in black. The Sym-H index reported by the World Data Center at Kyoto is shown in red. Figure 6b shows the observed Kp as a color-coded bar chart, and the simulated Kp is shown by the plotted lines. As before, the individual ensemble members are shown in gray, the ensemble mean is shown in magenta, and the result from the unperturbed run is shown in black.

We can assess the uncertainty in the SWMF predictions by constructing probability distributions of the predicted quantities. Figure 7 has the same basic layout as Figure 6, but the results from individual ensemble members have been replaced by blue bands marking different confidence intervals. The central, darker blue band marks the central 50% of the probability distribution at each time step, and the broader, light blue band marks the central 95% of the predicted Sym-H. To obtain these intervals, Gaussian KDE to the distribution of Sym-H in each time bin and find the 2.5, 25, 75, and 97.5 percentiles. These are found by integrating the fitted KDE from a large negative value to a target value and calculating the cumulative probability $F(x)$. The value at which the cumulative probability corresponds to the desired percentile (q) is found by using Brent's method (e.g., Press et al., 1992) to locate the root of $F(x) - q$. For comparison of the observed Kp to the simulated Kp, it is important to note that the SWMF calculates the Kp index for a user-configurable time window, at a user-configurable cadence. The operational geospace configuration (as used in this work) uses a window

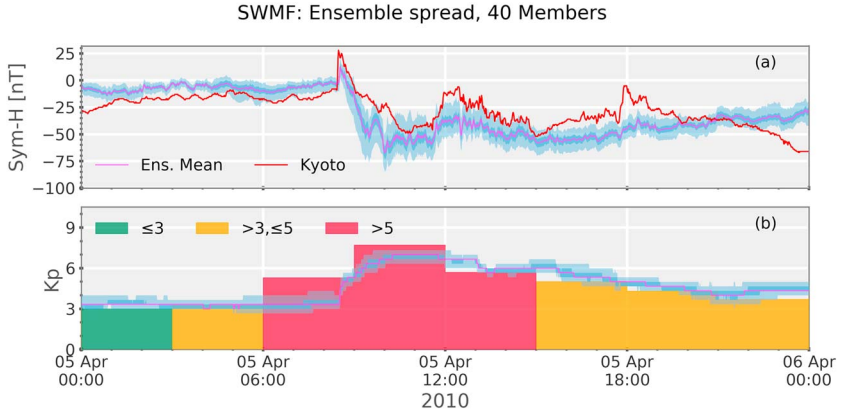

Figure 7. A comparison of modeled and observed geomagnetic indices, Sym-H and Kp, for SWPC event 5. Similar to Figure 6a, (a) shows the Sym-H index from observation (red), and the mean Sym-H calculated from the full ensemble (magenta). The 50% and 95% confidence intervals for the Sym-H prediction are shown by the blue bands. (b) shows the observed Kp index in the colored step plot and ensemble average in the same format as Figure 6b. The 50% and 95% confidence intervals for the Kp prediction are shown by the blue bands. SWMF = Space Weather Modeling Framework.

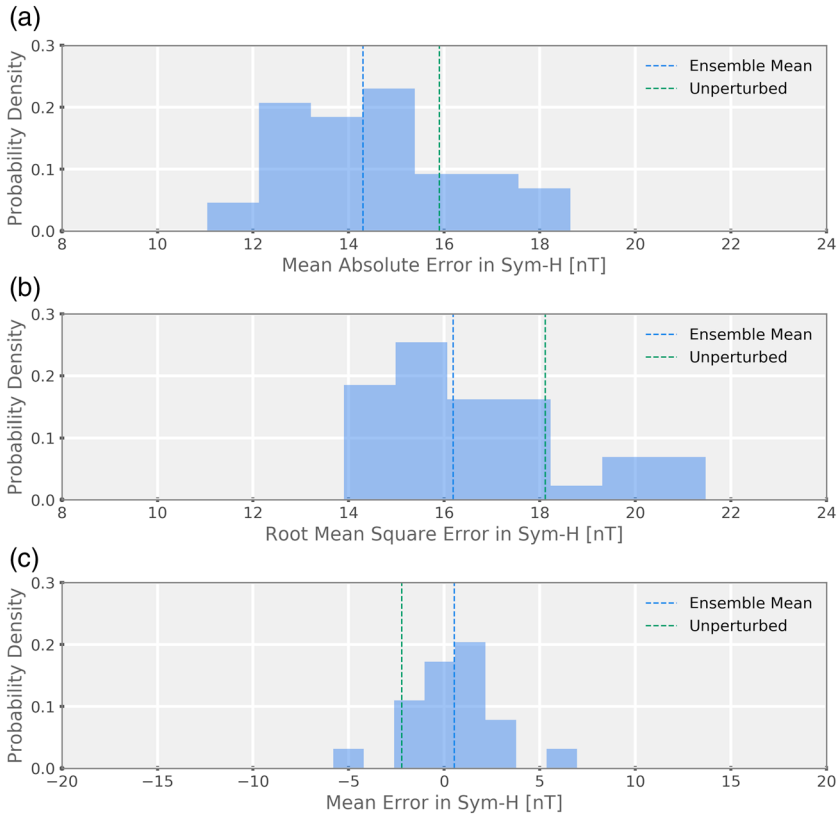

Figure 8. A statistical overview of model performance at predicting Sym-H for SWPC event 5. (a) shows the distribution of mean absolute error in Sym-H for all ensemble members (filled), and the vertical bars mark the mean absolute error for the ensemble mean of Sym-H (blue dashed) and the unperturbed run (green dashed). (b) follows the same format but for the root mean square error of Sym-H. (c) follows the same format but shows the mean error.

length of 3 hr, consistent with the derivation of the observed Kp index, and the cadence is 1 min. Thus, the time window for the SWMF-calculated Kp is only identical to the observed Kp at the end of each Kp block plotted in Figures 6b and 7b. The confidence intervals on Kp are calculated using the same method as for Sym-H, but we discretize the mean and quantiles of Kp by rounding them to the nearest valid Kp value.

Qualitatively, Figure 7 shows that the SWMF predictions are sensitive to both errors in the solar wind drivers and internal sources of error. For example, the observed Sym-H tends to frequently lie within the ensemble 95% confidence interval, demonstrating that differences between the model and observation can be explained via uncertainty in the solar wind drivers. However, there are periods where observed and modeled Sym-H diverge well beyond the confidence intervals. Many factors may contribute to this, including the resolution of the MHD model, poor specification of plasma sheet density and composition, or others (e.g., Welling et al., 2011; Welling & Ridley, 2010). The performance of the Kp forecast is overall better and less sensitive to solar wind uncertainty. Much of this arises from the pseudo-logarithmic nature of the index (Rostoker, 1972): broad ranges of activity can produce the same Kp value. Still, expanding the forecast to include the confidence intervals helps improve data-model agreement.

To quantify the performance of the ensemble prediction of Sym-H, we examine two accuracy metrics and one bias metric (see, e.g., Morley, Brito, et al., 2018). To characterize the accuracy, we use the mean absolute error (MAE) and the root mean square error (RMSE). To characterize the bias, we use the mean error (ME). These metrics are defined as (Morley, Brito, et al., 2018)

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|, \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2\right)}. \tag{5}$$

Table 1
Contingency Table of the Comparison Between Predictions and Observations

		Observed (x)	
		Yes	No
Predicted (y)	Yes	a	b
	No	c	d

Note. The letters a–d represent the number of cases in each category.

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i\right), \tag{6}$$

where y is the predicted value, and x is the observation.

Figure 8 shows model performance metrics for the Sym-H predictions. The distribution of MAE in the Sym-H prediction, from all ensemble members, is shown by the normalized histogram in the top panel. The blue vertical bar gives the MAE for the ensemble mean Sym-H prediction, and the green vertical bar gives the MAE for the unperturbed run. The middle panel presents RMSE, and the lower panel presents ME. The ensemble mean shows better predictive performance in all three metrics, suggesting that accounting for the uncertainty in the prediction due to the solar wind driver can help improve the prediction of the Sym-H index.

While we have shown the uncertainty in the Kp prediction, estimated from the spread in the ensemble, we do not present any quantitative analysis of the accuracy or bias in the Kp predictions. As described above, the SWMF-calculated Kp is only identical to the observed Kp at the end of each 3-hr interval. This means that a quantitative comparison for the 24 hr of the event interval would contain only 8 data points.

4.3. Ground Magnetic Perturbations

The quantity we will examine in our assessment of ground magnetic perturbations is $\mathrm{d}B_H/\mathrm{d}t$, which, following Pulkkinen et al. (2013), we define as

$$\frac{\mathrm{d}B_H}{\mathrm{d}t} = \sqrt{\left(\frac{\mathrm{d}B_N}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}B_E}{\mathrm{d}t}\right)^2}, \tag{7}$$

where B_N and B_E represent the North and East (horizontal) components of the magnetic field, respectively. In our analysis we calculate the time derivative of each component of the geomagnetic field using a central difference and second-order forward and backward differences at the endpoints. This gives the derivatives on the same set of time stamps as the original magnetic field perturbations.

To assess the model performance at predicting the magnetic perturbations at specific locations on the ground, we follow Pulkkinen et al. (2013) and use threshold crossings in 20-min time windows. That is, if the $\mathrm{d}B_H/\mathrm{d}t$ exceeds a given threshold in a 20-min interval, it is marked as a predicted event. Similarly, if the observed $\mathrm{d}B_H/\mathrm{d}t$ exceeds that threshold in the same 20-min interval, it is marked as an observed event. Pulkkinen et al. (2013) tested the skill of the model at predicting threshold crossings at combined sets of ground stations. To illustrate the behavior of the ensemble, we will focus on individual stations.

4.3.1. Metrics for Quantifying Model Performance

First, we briefly introduce the metrics we use to quantify the performance of our event prediction. Defining an event as any 20-min window where the peak $\mathrm{d}B_H/\mathrm{d}t$ exceeds a given threshold, we can construct a contingency table of (a) true positives, (b) true negatives, (c) false positives, and (d) false negatives. Such a contingency table is shown in Table 1.

We use the three metrics employed by Pulkkinen et al. (2013) as well as one additional metric. The employed metrics are probability of detection (POD), probability of false detection (POFD), Heidke Skill Score (HSS), and bias. For all reported metrics, we also calculate a 95% confidence interval. While confidence intervals can be easily estimated from the contingency table for metrics based on rates (e.g., Stephenson, 2000; Wilks, 2006), the confidence intervals on the HSS or bias cannot. We therefore use bootstrap estimates of the 95% confidence intervals for each reported metric. Surrogate series of events and nonevents are generated by drawing (with replacement) pairs of prediction and observation. From these surrogate series of *predicted* and *observed*

events, we then construct a contingency table and calculate the metric in the usual way. We repeat this procedure 2,000 times then define our 95% confidence interval as the interval containing the central 95% of bootstrapped values.

POD and POFD are measures of *discrimination* (Wilks, 2006). POD is defined as

$$\mathrm{POD} = \frac{a}{a+c}, \tag{8}$$

and this gives the probability of an event being correctly predicted given that an event occurred. If the model predicts all observed events then it will have a POD of 1. POFD is defined as

$$\mathrm{POFD} = \frac{b}{b+d}. \tag{9}$$

POFD considers the number of intervals in which a threshold crossing was predicted but did not occur. Describing this as a conditional probability, we see that POFD gives the probability of an event being incorrectly predicted given that an event did not occur. Smaller values of POFD indicate a better model performance, and a model with no false predictions will have a POFD of 0.

Skill scores are measures of relative accuracy (e.g., Wilks, 2006). The HSS is a commonly used skill score for categorical event predictions across space weather and is in widespread use in magnetospheric physics. The specific accuracy measure that it uses is the proportion correct (PC), which is defined as

$$\mathrm{PC} = \frac{a+d}{a+b+c+d} \tag{10}$$

and simply measures the fraction of predictions that obtained the correct result. A perfect prediction has a PC of 1. The reference used in the HSS is the PC that would be obtained for random predictions that are statistically independent of the observations (Wilks, 2006). HSS is defined as

$$\mathrm{HSS} = \frac{\mathrm{PC} - \mathrm{PC}_{ref}}{1 - \mathrm{PC}_{ref}} = \frac{2(ad - bc)}{(a+c)(c+d) + (a+b)(b+d)}. \tag{11}$$

For random predictions, the HSS is 0, and the model is deemed unskilled. Constant predictions, that is, the model always predicts no event, also have an HSS of 0 and are deemed unskilled. Predictions that underperform relative to chance have negative HSS, while predictions that outperform random chance have positive HSS, and a perfect prediction has HSS of 1. By constructing the reference from the contingency table, the HSS is constrained to lie in the interval [−1, 1].

Bias measures the correspondence between the average prediction and the average observation. In the case of a 2 × 2 contingency table, this measure provides information about whether the model predicts the right number of events or whether it predicts too few (underpredicts) or too many (overpredicts). Bias is defined as

$$\mathrm{Bias} = \frac{a+b}{a+c} \tag{12}$$

and is the ratio of the number of forecast events to the number of observed events. An unbiased forecast has a bias of 1. If more events are forecast than are observed, the bias will be greater than 1. Similarly, the bias will be below 1 if the model underpredicts.

A wide variety of other metrics can be calculated to highlight different aspects of model performance (see, e.g., Stephenson, 2000). All simulated magnetometer outputs and geomagnetic indices from the set of model runs presented in this work have been archived in an open access repository (Morley, Welling, et al., 2018), so that additional analysis can be performed or comparisons made with new work, using any metrics.

4.3.2. Assessing Model Predictions of Ground Magnetic Perturbations

The first station we examine is Newport which has a geomagnetic latitude of 54.9° and a geomagnetic longitude of 304.7°. The observed time series of dB_H/dt is plotted as a red line in Figure 9a, and the simulated dB_H/dt from the unperturbed model run is shown as a blue line. The maximum value of dB_H/dt in each 20-min window is shown as a colored symbol, a red cross for the observations and a blue-filled circle for the simulation. The horizontal dashed lines mark the thresholds used for event determination. The modeled peaks in

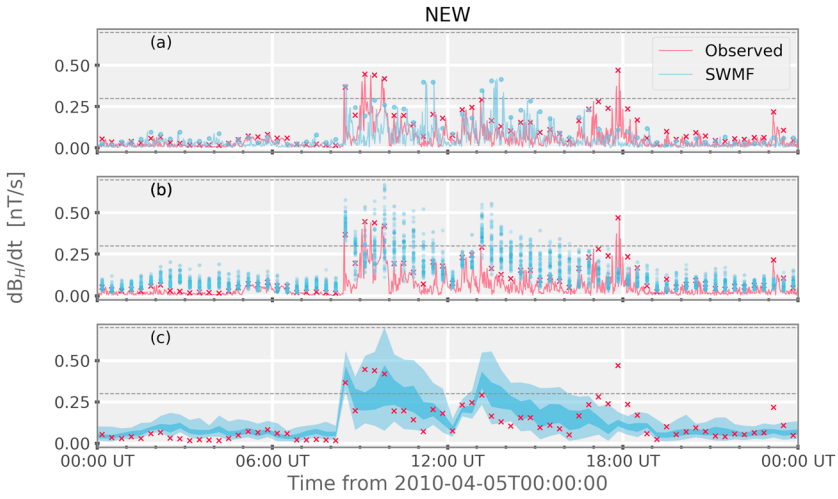

Figure 9. Observed and simulated dB_H/dt for the Newport (NEW) magnetic observatory. (a) shows the observed time series of dB_H/dt as a red line, and the simulated dB_H/dt from the unperturbed (reference) model run is shown as a blue line. The maximum values of dB_H/dt in nonoverlapping 20-min windows are shown as a colored symbols. The red crosses mark the observed bin maxima, and the blue-filled circles mark the bin maxima for the simulation. The horizontal dashed lines mark the thresholds used for event determination. (b) shows the observed time series and bin maxima in the same format as (a), and the bin maxima from each ensemble member are shown by the blue-filled circles. (c) shows the observed bin maxima and the estimated interquartile range and 95% confidence interval derived from kernel density estimate fits to the bin maxima from the ensemble. SWMF = Space Weather Modeling Framework; UT = universal time.

each bin are broadly similar to the observed maxima. From about 09:00 to 10:00 UTC, the observations indicate crossings of the 0.3-nT/s threshold, while the unperturbed model run underpredicts and fails to predict these events. Near 18:00 UTC, the observed dB_H/dt again crosses the marked threshold where the simulation does not. Comparing the unperturbed simulation to the observations, we find POD = 0.200[0, 0, 0.67], POFD = 0.075[0.02, 0.14], HSS = 0.115[−0.09, 0.46], and bias = 1.200[0.36, 5.0]. These metrics are collected in Table 2. We note that the low number of events leads to the confidence intervals on POD and HSS containing 0.

Figure 9b again shows the observed time series in red and marks the bin maxima with red crosses. The bin maxima from each of the 40 ensemble members are now plotted as blue-filled circles. The markers for the ensemble members are semitransparent so that overlapping markers appear darker. We can use the spread in the predicted maxima to quantify the uncertainty in the predicted output due to the uncertain solar wind input. To better visualize the spread of the ensemble members, Figure 9c shows the observed bin maxima with red crosses and two blue-filled regions. The central, darker blue band marks the interquartile range, and the broader, lighter blue band marks the central 95% of the predicted maxima. To obtain these intervals, we fit a Gaussian KDE to the distribution of maxima in each time bin and find the 2.5, 25, 75, and 97.5 percentiles. Inspection of Figure 9c shows that the observed threshold crossings between 09:00 and 10:00 UTC fall within

Table 2
Event Analysis Metrics for NEW and YKC Stations With Different Thresholds

	POD [$CI_{0.95}$]	POFD [$CI_{0.95}$]	HSS [$CI_{0.95}$]	Bias [$CI_{0.95}$]
		Unperturbed simulation		
NEW (0.3 nT/s)	0.200 [0.00, 0.67]	0.075 [0.02, 0.14]	0.115 [−0.09, 0.46]	1.200 [0.36, 5.00]
YKC (0.7 nT/S)	0.556 [0.36, 0.74]	0.111 [0.02, 0.21]	0.469 [0.24, 0.68]	0.741 [0.50, 1.00]
YKC (1.1 nT/s)	0.471 [0.21, 0.71]	0.055 [0.00, 0.12]	0.474 [0.21, 0.71]	0.647 [0.36, 1.00]
		Naive probabilistic classifier		
NEW (0.3 nT/s)	0.400 [0.00, 1.00]	0.015 [0.00, 0.05]	0.473 [−0.03, 0.88]	0.600 [0.00, 2.00]
YKC (0.7 nT/S)	0.593 [0.39, 0.77]	0.089 [0.02, 0.18]	0.531 [0.31, 0.72]	0.741 [0.50, 1.00]
YKC (1.1 nT/s)	0.412 [0.18, 0.65]	0.055 [0.00, 0.12]	0.417 [0.15, 0.66]	0.588 [0.30, 1.00]

Note. Station and threshold are given in the table. NEW = Newport; YKC = Yellowknife; POD = probability of detection; POFD = probability of false detection; HSS = Heidke Skill Score; CI = confidence interval.

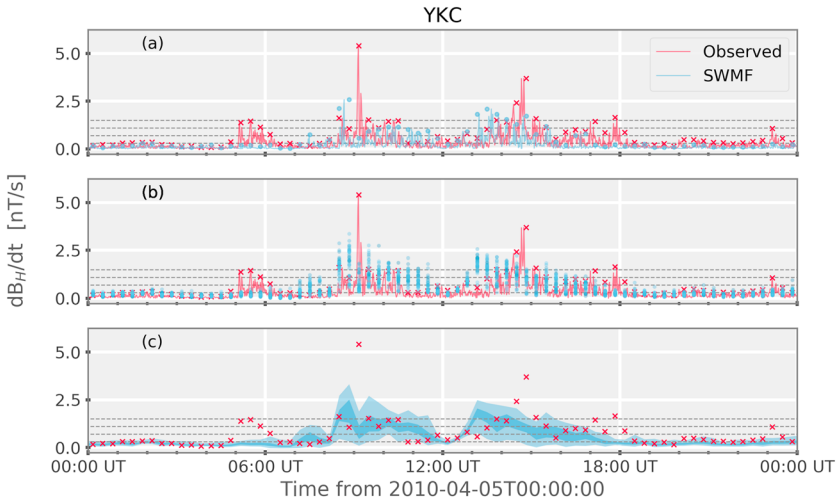

Figure 10. Same as Figure 9 but for the Yellowknife (YKC) magnetic observatory. SWMF = Space Weather Modeling Framework; UT = universal time.

the range of predicted activity consistent with the uncertainty due to the upstream boundary condition. Conversely, the brief surge in activity observed near 18:00 UTC lies well outside the 95% confidence interval , indicating that the model failure to capture this activity is either from uncertainty in the solar wind not captured by our error model (such as uncertainty in the number density) or from inadequacies in the model configuration.

We can also use the ensemble to attempt to improve the prediction. A variety of methods could be used, but as we are predicting binary events (threshold crossings), we can use our set of ensemble members to estimate the probability of exceeding the threshold. So that we can still compare these results to the deterministic case of a single prediction and observation, we have used a naive probabilistic classifier (NPC). We define the NPC as predicting an event if at least 50% of ensemble members predict an event; that is, if the NPC indicates an event probability of >50%, then we interpret this as a deterministic prediction of an event. Comparing the NPC to the observations, we find POD = 0.400[0.0, 1.0], POFD = 0.015[0.0, 0.05], HSS = 0.473[−0.03, 0.88], and bias = 0.600[0.0, 2.0]. These metrics are collected in Table 2. Although using the set of ensemble members as a classifier has increased the calculated skill of the operational geospace configuration of SWMF, the confidence interval still contains zero, and hence, neither the initial simulation nor the ensemble classifier can be said to have significant skill. We note that four of the ensemble members had HSS that were significantly different from zero and thus display significant skill.

The second ground station that we assess is Yellowknife at a geomagnetic latitude of 68.9° and a geomagnetic longitude of 299.4°. This station is at a similar longitude to Newport but is at much higher latitude and is subject to much larger variations in dB_H/dt. The data and simulation results for Yellowknife are shown in Figure 10 using the same format as Figure 9. The largest values of dB_H/dt, observed near 09:00 and 15:00 UTC, are not captured by any of the simulations although the ensemble does predict a low probability of exceeding the 1.5-nT/s threshold. Turning to the interval of activity between 06:00 and 07:00 UTC, it is clear that all ensemble members performed consistently. The activity observed cannot be attributed to uncertainty in the upstream boundary condition.

To provide a quantitative summary of the model's ability to predict dB_H/dt, we first examine a threshold of 0.7 nT/s. Comparing the unperturbed simulation to the observations, we find POD = 0.556[0.36, 0.74], POFD = 0.111[0.02, 0.21], HSS = 0.469[0.24, 0.68], and bias = 0.741[0.5, 1.0]. All summary metrics for this analysis are collected in Table 2. Comparing the NPC to the observations, we find POD = 0.593[0.39, 0.77], POFD = 0.089[0.02, 0.18], HSS = 0.531[0.31, 0.72], and bias = 0.741[0.5, 1.0]. The naive prediction using the ensemble yields an improvement in the predictive ability of the simulation. The POFD is reduced, while the POD is increased, leading to an improvement in the skill. The bias is unchanged in this case. As before, the low number of events leads to broad confidence intervals on the performance metrics, and the improvement in skill from the NPC cannot be determined to be statistically significant using this event.

Table 3
Event Analysis Metrics for All Stations (FRD, FRN, FUR, HRN, IQA, MEA, NEW, OTT, SNK/PBQ, WNG, and YKC), Using a Threshold of 0.3 nT/s

	POD [$CI_{0.95}$]	POFD [$CI_{0.95}$]	HSS [$CI_{0.95}$]	Bias [$CI_{0.95}$]
Unperturbed simulation				
All stations	0.521 [0.46, 0.58]	0.036 [0.02, 0.05]	0.543 [0.46, 0.58]	0.595 [0.53, 0.67]
Naive probabilistic classifier				
All stations	0.560 [0.50, 0.63]	0.037 [0.02, 0.05]	0.577 [0.52, 0.64]	0.638 [0.57, 0.71]

Note. POD = probability of detection; POFD = probability of false detection; HSS = Heidke Skill Score; CI = confidence interval; NEW = Newport; YKC = Yellowknife; FRD = Fredericksburg; FRN = Fresno; FUR = Furstenfeldbruck; HRN = Hornsund; IQA = Iqalut; MEA = Meanook; OTT = Ottawa; SNK = Sanikiluaq; PBQ = Poste-de-la-Baleine; WNG = Wingst.

Increasing the threshold to 1.1 nT/s has a slightly different outcome. Comparing the unperturbed to the observations, we find POD = 0.471[0.21, 0.71], POFD = 0.055[0.0, 0.12], HSS = 0.474[0.21, 0.71], and bias = 0.647[0.36, 1.0]. Comparing the NPC to the observations, we find POD = 0.412[0.18, 0.65], POFD = 0.055[0.0, 0.12], HSS = 0.417[0.15, 0.66], and bias = 0.588[0.3, 1.0]. That is, the NPC tends to underpredict, while examination of the 95% confidence interval in Figure 10c shows that the majority of predicted events at this threshold are within the expected range of values. The selection of the fraction of ensemble members to use to define an event is known as calibration. We leave the issue of calibration for future work, but note that using ensembles of model runs brings the opportunity to significantly improve the skill of the predictions.

While these results are encouraging, the length of the interval that we use for identifying threshold crossings has too few events to allow us to draw many definitive conclusions about the skill of the model and of the NPC. The presented methodology and results represent a first step toward perturbed input ensemble modeling with solar wind-driven simulations such as the SWMF and demonstrate some ways in which ensemble forecasts could be used to help improve the forecast and estimate the uncertainty in an operational setting.

It is instructive to note that Pulkkinen et al. (2013) used six events and combined the predictions from stations in bands of geomagnetic latitude. While they do not present confidence intervals for the derived metrics, their skill scores are calculated from samples approximately 18 times larger and will have much narrower confidence intervals than the results we present. For illustrative purposes, we also present the model performance metrics for a prediction that combines all 11 stations and uses a threshold of 0.3 nT/s; combining all stations and selecting a low threshold maximizes the number of events.

Combining the 11 magnetometer stations used in this study and repeating this analysis gives an overall measure of model performance at predicting threshold crossings in dB_H/dt. The model performance metrics are given in Table 3. The naive classifier displays a higher POD and a lower POFD and correspondingly a higher HSS. Again, although the NPC outperforms the unperturbed simulation, we are unable to say that the improvement is statistically significant given the short time period and low number of events. Although not directly comparable, we refer the reader to Figures 7a and 7c of Pulkkinen et al. (2013) where the POD, POFD, and HSS for event 5 are given, using a threshold of 0.3 nT/s, aggregated over midlatitude and high-latitude stations separately. For this event, the SWMF outperformed the other tested models, with HSS of 0.366 (midlatitude) and 0.326 (high latitude). As shown in Table 3, our NPC, aggregated over all stations, has a HSS of 0.577 and improves on the performance of the unperturbed simulation.

5. Conclusions

We have developed a nonparametric method for generating multiple possible realizations of the solar wind just upstream of the bow shock based on observations near L1. We have applied our perturbation model to the solar wind inputs for the SWMF and have simulated the geomagnetic storm that occurred on 5 April 2010. This event was selected as event 5 in the set of challenge events used by Pulkkinen et al. (2013).

We ran a 40-member ensemble for this event and have used this ensemble to quantify the uncertainty in the model output due to the uncertainty in the upstream (driving) boundary conditions. We have further examined the performance of naive models derived from the ensemble and compared them to the simulation with unperturbed inputs. For parameters where we predict the value (Sym-H and Kp), we use the ensemble

mean as our naive model. For parameters where we predict a threshold crossing (dB_H/dt), we use a naive classifier in which we predict an event if at least half of the ensemble members predict an event.

Both the ensemble mean and the unperturbed simulation tend to underpredict the magnitude of Sym-H in the quiet interval before the storm and overpredict the magnitude of the disturbance in the storm itself, consistent with the results of Haiducek et al. (2017). The ensemble mean is a more accurate predictor of Sym-H than the result from the unperturbed simulation, improving the MAE by nearly 2 nT for this interval. The ensemble average is closer to unbiased than the unperturbed run, but this summary measure masks the systematic behavior described previously.

Using an ensemble of predictions, we have shown the uncertainty of the predicted maxima of dB_H/dt given the uncertainty in the solar wind boundary condition. The estimated 95% confidence intervals can be broad compared to the spacing between the thresholds that Pulkkinen et al. (2013) selected for study. The confidence intervals are typically narrow during periods where the dB_H/dt is predicted to be low. The confidence intervals are often much wider where the median prediction is for enhanced dB_H/dt.

The ensemble of simulations allows us to identify intervals of activity that can not be explained by uncertainty in the solar wind driver. Routine calculation of a small ensemble could help model developers improve predictions by identifying phenomenology that a given model configuration cannot capture. Operationally, we suggest that ensembles of deterministic models should be run where possible to enable probabilistic forecasts and communicate uncertainty in the forecast to the customer.

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy and was funded by the Laboratory Directed Research and Development program (grant 20170047DR). This work used the SWMF/BATSRUS tools developed at The University of Michigan Center for Space Environment Modeling (CSEM). The SWMF and embedded models can be obtained via http://csem.engin.umich.edu/. Analysis used the SWMF tools in the SpacePy package and the PyForecastTools package. SpacePy is available at https://github.com/spacepy/spacepy, and PyForecastTools is available at https://github.com/drsteve/PyForecastTools. The satellite-specific solar wind data used in this study are provided by NASA's Space Physics Data Facility (SPDF) and are available via FTP at ftp://spdf.gsfc.nasa.gov/pub/data/omni/high_res_omni/sc_specific/. The magnetometer data used in this study are available from the Community Coordinated Modeling Center at https://ccmc.gsfc.nasa.gov/RoR_WWW/pub/dBdt/out/deltaB. Input files and magnetometer output files for all simulations are archived at https://zenodo.org/record/1324562. S. K. M. acknowledges John Steinberg (LANL) and Ehab Hassan (Ain Shams University) for useful discussions and initial work examining the errors in propagated solar wind data that led to this work. S. K. M. also thanks Matthew Hoffman (LANL) for useful discussions.

References

Andriyas, T., Spencer, E., Raj, A., Sojka, J., & Mays, M. L. (2012). Forecasting the Dst index during corotating interaction region events using synthesized solar wind parameters. *Journal of Geophysical Research*, *117*, A03204. https://doi.org/10.1029/2011JA017018

Barnston, A. G., Mason, S. J., Goddard, L., DeWitt, D. G., & Zebiak, S. E. (2003). Multimodel ensembling in seasonal climate forecasting at IRI. *Bulletin of the American Meteorological Society*, *84*(12), 1783–1796. https://doi.org/10.1175/BAMS-84-12-1783

Borovsky, J. E. (2008). Flux tube texture of the solar wind: Strands of the magnetic carpet at 1 AU? *Journal of Geophysical Research*, *113*, A08110. https://doi.org/10.1029/2007JA012684

Borovsky, J. E. (2012). Looking for evidence of mixing in the solar wind from 0.31 to 0.98 AU. *Journal of Geophysical Research*, *117*, A06107. https://doi.org/10.1029/2012JA017525

Borovsky, J. E. (2017). The spatial structure of the oncoming solar wind at Earth and the shortcomings of a solar-wind monitor at L1. *Journal of Atmospheric and Solar-Terrestrial Physics*. https://doi.org/10.1016/j.jastp.2017.03.014

Borovsky, J. E., & Valdivia, J. A. (2018). The Earth's magnetosphere: A systems science overview and assessment. *Surveys in Geophysics*, *39*(5), 817–859. https://doi.org/10.1007/s10712-018-9487-x

Case, N. A., & Wild, J. A. (2012). A statistical comparison of solar wind propagation delays derived from multispacecraft techniques. *Journal of Geophysical Research*, *117*, A02101. https://doi.org/10.1029/2011JA016946

Cash, M. D., Biesecker, D. A., Pizzo, V., Koning, C. A., Millward, G., Arge, C. N., et al. (2015). Ensemble modeling of the 23 July 2012 coronal mass ejection. *Space Weather*, *13*, 611–625. https://doi.org/10.1002/2015SW001232

Cash, M. D., Biesecker, D. A., Reinard, A., & de Koning, C. A. (2015). DSCOVR: Real-time solar wind data and operational products (AGU Fall Meeting Abstracts #SM31E-03). San Francisco, CA.

Cash, M. D., Hicks, S. W., Biesecker, D. A., Reinard, A. A., Koning, C. A., & Weimer, D. R. (2016). Validation of an operational product to determine L1 to Earth propagation time delays. *Space Weather*, *14*, 93–112. https://doi.org/10.1002/2015SW001321

Chao, J., Wu, D., Lin, C.-H., Yang, Y.-H., Wang, X., Kessel, M., et al. (2002). Models for the size and shape of the Earth's magnetopause and bow shock. In J. Chao, et al. (Eds.), *Space Weather Study Using Multipoint Techniques, COSPAR Colloquia Series* (Vol. 12, pp. 127–135), Taipei, Taiwan: Pergamon. https://doi.org/10.1016/S0964-2749(02)80212-8

Chen, M. W., O'Brien, T. P., Lemon, C. L., & Guild, T. B. (2018). Effects of uncertainties in electric field boundary conditions for ring current simulations. *Journal of Geophysical Research: Space Physics*, *123*, 638–652. https://doi.org/10.1002/2017JA024496

Connors, M., Russell, C. T., & Angelopoulos, V. (2011). Magnetic flux transfer in the 5 April 2010 Galaxy 15 substorm: An unprecedented observation. *Annales Geophysicae*, *29*(3), 619–622. https://doi.org/10.5194/angeo-29-619-2011

De Zeeuw, D. L., Gombosi, T. I., Groth, C. P. T., Powell, K. G., & Stout, Q. F. (2000). An adaptive MHD method for global space weather simulations. *IEEE Transactions on Plasma Science*, *28*(6), 1956–1965. https://doi.org/10.1109/27.902224

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*(1), 54–75. https://doi.org/10.1214/ss/1177013815

Epstein, E. S. (1969). The role of initial uncertainties in predicion. *Journal of Applied Meteorology*, *8*(2), 190–198. https://doi.org/10.1175/1520-0450(1969)008<0190:TROIUI>2.0.CO;2

Fränz, M., & Harper, D. (2002). Heliospheric coordinate systems. *Planetary and Space Science*, *50*(2), 217–233. https://doi.org/10.1016/S0032-0633(01)00119-2

Guerra, J. A., Pulkkinen, A., & Uritsky, V. M. (2015). Ensemble forecasting of major solar flares: First results. *Space Weather*, *13*, 626–642. https://doi.org/10.1002/2015SW001195

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

Hassan, E., Morley, S. K., & Steinberg, J. T. (2015). A statistical ensemble for solar wind measurements: A step toward forecasting. In M. M. Cowee (Ed.), *2015 Los Alamos Space Weather Summer School Research Reports* (pp. 17–31). Los Alamos, NM: LA-UR-15-29127. https://doi.org/10.2172/1227256

Kawano, H., & Higuchi, T. (1995). The bootstrap method in space physics: Error estimation for the minimum variance analysis. *Geophysical Research Letters*, *22*(3), 307–310. https://doi.org/10.1029/94GL02969

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., et al. (2015). The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, *96*(8), 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

Kessel, R. L., Quintana, E., & Peredo, M. (1999). Local variations of interplanetary magnetic field at Earth's bow shock. *Journal of Geophysical Research*, *104*(A11), 24869–24878. https://doi.org/10.1029/1999JA900230

King, J. H., & Papitashvili, N. E. (2005). Solar wind spatial scales in and comparisons of hourly Wind and ACE plasma and magnetic field data. *Journal of Geophysical Research*, *110*, A02104. https://doi.org/10.1029/2004JA010649

Knipp, D. J. (2016). Advances in space weather ensemble forecasting. *Space Weather*, *14*, 52–53. https://doi.org/10.1002/2016SW001366

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, *17*(3), 1217–1241.

Lemon, C., Wolf, R. A., Hill, T. W., Sazykin, S., Spiro, R. W., Toffoletto, F. R., et al. (2004). Magnetic storm ring current injection modeled with the rice convection model and a self-consistent magnetic field. *Geophysical Research Letters*, *31*, L21801. https://doi.org/10.1029/2004GL020914

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, *20*(2), 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2

Mackay, D. J. C. (1998). Introduction to Monte Carlo methods. In *Learning in graphical models* (pp. 175–204). https://doi.org/10.1007/978-94-011-5014-9_7

Mailyan, B., Munteanu, C., & Haaland, S. (2008). What is the best method to calculate the solar wind propagation delay? *Annales Geophysicae*, *26*(8), 2383–2394. https://doi.org/10.5194/angeo-26-2383-2008

Mayaud, P. N. (1980). *Derivation, meaning and use of geomagnetic indices, Geophysical Monograph Series* (Vol. 22). Washington, DC: American Geophysical Union.

Morley, S. K. (2008). Observations of magnetospheric substorms during the passage of a corotating interaction region. In *Proceedings of 7th Australian Space Science Conference 2007* (pp. 118–129). Sydney, Australia.

Morley, S. (2018). drsteve/PyForecastTools: PyForecastTools: Version 1.0. https://doi.org/10.5281/zenodo.1256922

Morley, S. K., Brito, T. V., & Welling, D. T. (2018). Measures of model performance based on the log accuracy ratio. *Space Weather*, *16*(1), 69–88. https://doi.org/10.1002/2017SW001669

Morley, S. K., & Freeman, M. P. (2007). On the association between northward turnings of the interplanetary magnetic field and substorm onsets. *Geophysical Research Letters*, *34*, L08104. https://doi.org/10.1029/2006GL028891

Morley, S. K., Koller, J., Welling, D. T., Larsen, B. A., Henderson, M. G., & Niehof, J. T. (2011). Spacepy—A Python-based library of tools for the space sciences. In *Proceedings of the 9th Python in Science Conference (SciPy 2010)* (pp. 39–45). Austin, TX: Astrophysics Source Code Library.

Morley, S. K., Koller, J., Welling, D. T., Larsen, B. A., & Niehof, J. T. (2010). Spacepy. Retrieved from https://sourceforge.net/p/spacepy. [Published: 20 May 2010; Accessed: 3 July 2018].

Morley, S. K., Welling, D. T., & Woodroffe, J. R. (2018). Space Weather Modeling Framework ensemble simulations [Data set]. https://doi.org/10.5281/zenodo.1324562

Murphy, J., Sexton, D., Barnett, D., Jones, G., Webb, M., Collins, M., & Stainforth, D. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, *430*(7001), 768–772. https://doi.org/10.1038/nature02771

Murray, S. A. (2018). The importance of ensemble techniques for operational space weather forecasting. *Space Weather*, *16*, 777–783. https://doi.org/10.1029/2018SW001861

Niehof, J., & Morley, S. (2012). Determining the significance of associations between two series of discrete events: Bootstrap methods (Tech. Rep. LA-14453-MS). Los Alamos, NM: Los Alamos National Laboratory.

Nishida, A., Uesugi, K., Nakatani, I., Mukai, T., Fairfield, D. H., & Acuna, M. H. (1992). Geotail mission to explore Earth's magnetotail. *Eos, Transactions American Geophysical Union*, *73*(40), 425–429. https://doi.org/10.1029/91EO00314

Ólafsdóttir, K. B., & Mudelsee, M. (2014). More accurate, calibrated bootstrap confidence intervals for estimating the correlation between two time series. *Mathematical Geosciences*, *46*(4), 411–427. https://doi.org/10.1007/s11004-014-9523-4

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

Owen, J. A., & Palmer, T. N. (1987). The impact of El Niño on an ensemble of extended-range forecasts. *Monthly Weather Review*, *115*(9), 2103–2117. https://doi.org/10.1175/1520-0493(1987)115<2103:TIOENO>2.0.CO;2

Papitashvili, N., Bilitza, D., & King, J. (2014). OMNI: A description of near-Earth solar wind environment (40th COSPAR Scientific Assembly, Abstract #C0.1-12-14), COSPAR Meeting, Held 2–10 August 2014, in Moscow, Russia.

Powell, K. G., Roe, P. L., Linde, T. J., Gombosi, T. I., & Zeeuw, D. L. D. (1999). A solution-adaptive upwind scheme for ideal magnetohydrodynamics. *Journal of Computational Physics*, *154*(2), 284–309. https://doi.org/10.1006/jcph.1999.6299

Press, W. H., Teukolsky, S. A., Vettering, W. T., & Flannery, B. P. (1992). *Numerical recipes in C: The art of scientific computing*. New York: Cambridge University Press.

Pulkkinen, A., & Rastätter, L. (2009). Minimum variance analysis-based propagation of the solar wind observations: Application to real-time global magnetohydrodynamic simulations. *Space Weather*, *7*, S12001. https://doi.org/10.1029/2009SW000468

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

Ridley, A. J., Gombosi, T. I., & DeZeeuw, D. L. (2004). Ionospheric control of the magnetosphere: Conductance. *Annales Geophysicae*, *22*(2), 567–584. https://doi.org/10.5194/angeo-22-567-2004

Ridley, A. J., Richmond, A. D., Gombosi, T. I., Zeeuw, D. L. D., & Clauer, C. R. (2003). Ionospheric control of the magnetospheric configuration: Thermospheric neutral winds. *Journal of Geophysical Research*, *108*, 1328. https://doi.org/10.1029/2002JA009464

Riley, P., Linker, J. A., & Mikić, Z. (2013). On the application of ensemble modeling techniques to improve ambient solar wind models. *Journal of Geophysical Research: Space Physics*, *118*, 600–607. https://doi.org/10.1002/jgra.50156

Rostoker, G. (1972). Geomagnetic indices. *Reviews of Geophysics*, *10*(4), 935–950. https://doi.org/10.1029/RG010i004p00935

Shue, J., Chao, J. K., Fu, H. C., Russell, C. T., Song, P., Khurana, K. K., & Singer, H. J. (1997). A new functional form to study the solar wind control of the magnetopause size and shape. *Journal of Geophysical Research*, *102*(A5), 9497–9511. https://doi.org/10.1029/97JA00196

Slingo, J., & Palmer, T. (2011). Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *369*(1956), 4751–4767. https://doi.org/10.1098/rsta.2011.0161

Smithtro, C. G., & Sojka, J. J. (2005). A new global average model of the coupled thermosphere and ionosphere. *Journal of Geophysical Research*, *110*, A08305. https://doi.org/10.1029/2004JA010781

Solow, A. R. (1985). Bootstrapping correlated data. *Journal of the International Association for Mathematical Geology*, *17*(7), 769–775. https://doi.org/10.1007/BF01031616

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

Stone, E., Frandsen, A., Mewaldt, R., Christian, E., Margolies, D., Ormes, J., & Snow, F. (1998). The advanced composition explorer. *Space Science Reviews*, *86*(1), 1–22. https://doi.org/10.1023/A:1005082526237

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

Toffoletto, F., Sazykin, S., Spiro, R., & Wolf, R. (2003). Inner magnetospheric modeling with the Rice Convection Model. *Space Science Reviews*, *107*(1), 175–196. https://doi.org/10.1023/A:1025532008047

Tóth, G., Sokolov, I. V., Gombosi, T. I., Chesney, D. R., Clauer, C. R., Zeeuw, D. L. D., et al. (2005). Space Weather Modeling Framework: A new tool for the space science community. *Journal of Geophysical Research*, *110*, A12226. https://doi.org/10.1029/2005JA011126

Tóth, G., van der Holst, B., Sokolov, I. V., Zeeuw, D. L. D., Gombosi, T. I., Fang, F., et al. (2012). Adaptive numerical algorithms in space weather modeling. *Journal of Computational Physics*, *231*(3), 870–903. https://doi.org/10.1016/j.jcp.2011.02.006

Tsurutani, B. T., Guarnieri, F. L., Lakhina, G. S., & Hada, T. (2005). Rapid evolution of magnetic decreases (MDs) and discontinuities in the solar wind: ACE and Cluster. *Geophysical Research Letters*, *32*, L10103. https://doi.org/10.1029/2004GL022151

Vassiliadis, D., Klimas, A. J., Baker, D. N., & Roberts, D. A. (1995). A description of the solar wind-magnetosphere coupling based on nonlinear filters. *Journal of Geophysical Research*, *100*(A3), 3495–3512. https://doi.org/10.1029/94JA02725

Vogel, R. M., & Shallcross, A. L. (1996). The moving blocks bootstrap versus parametric time series models. *Water Resources Research*, *32*(6), 1875–1882. https://doi.org/10.1029/96WR00928

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

Welling, D. T., Anderson, B. J., Crowley, G., Pulkkinen, A. A., & Rastätter, L. (2016). Exploring predictive performance: A reanalysis of the geospace model transition challenge. *Space Weather*, *15*, 192–203. https://doi.org/10.1002/2016SW001505

Welling, D. T., Jordanova, V. K., Zaharia, S. G., Glocer, A., & Toth, G. (2011). The effects of dynamic ionospheric outflow on the ring current. *Journal of Geophysical Research*, *116*, A00J19. https://doi.org/10.1029/2010JA015642

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

Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences* (2nd). London, UK: Academic Press.

Wing, S., Johnson, J. R., Jen, J., Meng, C., Sibeck, D. G., Bechtold, K., et al. (2017). Kp forecast models. *Journal of Geophysical Research: Space Physics*, *110*, A04203. https://doi.org/10.1029/2004JA010500