

# Space Weather®

## RESEARCH ARTICLE

10.1029/2023SW003446

### Special Section:

Machine Learning in  
Heliophysics

### Key Points:

- Machine learning based probabilistic predictions can provide risk forecasts of extreme  $dB/dt$  events at mid-latitude magnetometer stations
- Models utilizing both solar wind and past ground magnetometer data outperform solar wind only and  $dB/dt$  based persistence models
- Input feature analysis shows solar wind velocity and IMF  $B_z^{GSM}$  strongly influence model output, demonstrating alignment with current understanding

### Supporting Information:

Supporting Information may be found in the online version of this article.

### Correspondence to:

M. Coughlan,  
[mike.k.coughlan@gmail.com](mailto:mike.k.coughlan@gmail.com)

### Citation:

Coughlan, M., Keesee, A., Pinto, V., Mukundan, R., Marchezi, J. P., Johnson, J., et al. (2023). Probabilistic forecasting of ground magnetic perturbation spikes at mid-latitude stations. *Space Weather*, 21, e2023SW003446. <https://doi.org/10.1029/2023SW003446>

Received 31 JAN 2023

Accepted 18 MAY 2023

### Author Contributions:

**Conceptualization:** Michael Coughlan, Amy Keesee, Victor Pinto, Raman Mukundan, José Paulo Marchezi, Jeremiah Johnson

**Data curation:** Michael Coughlan, Victor Pinto, Raman Mukundan

**Formal analysis:** Michael Coughlan, Victor Pinto, Raman Mukundan, José Paulo Marchezi

© 2023. The Authors.

This is an open access article under the terms of the [Creative Commons Attribution License](#), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

## Probabilistic Forecasting of Ground Magnetic Perturbation Spikes at Mid-Latitude Stations

Michael Coughlan<sup>1,2</sup> , Amy Keesee<sup>1,2</sup> , Victor Pinto<sup>3,4</sup> , Raman Mukundan<sup>1,2</sup> , José Paulo Marchezi<sup>1,2</sup> , Jeremiah Johnson<sup>5</sup> , Hyunju Connor<sup>6</sup> , and Don Hampton<sup>7</sup> 

<sup>1</sup>Department of Physics & Astronomy, University of New Hampshire, Durham, NH, USA, <sup>2</sup>Institute for the Study of Earth, Oceans, & Space, University of New Hampshire, Durham, NH, USA, <sup>3</sup>Departamento de Física, Universidad de Santiago de Chile, Santiago, Chile, <sup>4</sup>Center for Interdisciplinary Research in Astrophysics and Space Exploration (CIRAS), Universidad de Santiago de Chile, Santiago, Chile, <sup>5</sup>Department of Electrical & Computer Engineering, University of New Hampshire, Manchester, NH, USA, <sup>6</sup>NASA Goddard Space Flight Center, Greenbelt, MD, USA, <sup>7</sup>Geophysical Institute, University of Alaska Fairbanks, Fairbanks, AK, USA

**Abstract** The prediction of large fluctuations in the ground magnetic field ( $dB/dt$ ) is essential for preventing damage from Geomagnetically Induced Currents. Directly forecasting these fluctuations has proven difficult, but accurately determining the risk of extreme events can allow for the worst of the damage to be prevented. Here we trained Convolutional Neural Network models for eight mid-latitude magnetometers to predict the probability that  $dB/dt$  will exceed the 99th percentile threshold 30–60 min in the future. Two model frameworks were compared, a model trained using solar wind data from the Advanced Composition Explorer (ACE) satellite, and another model trained on both ACE and SuperMAG ground magnetometer data. The models were compared to examine if the addition of current ground magnetometer data significantly improved the forecasts of  $dB/dt$  in the future prediction window. A bootstrapping method was employed using a random split of the training and validation data to provide a measure of uncertainty in model predictions. The models were evaluated on the ground truth data during eight geomagnetic storms and a suite of evaluation metrics are presented. The models were also compared to a persistence model to ensure that the model using both datasets did not over-rely on  $dB/dt$  values in making its predictions. Overall, we find that the models using both the solar wind and ground magnetometer data had better metric scores than the solar wind only and persistence models, and was able to capture more spatially localized variations in the  $dB/dt$  threshold crossings.

**Plain Language Summary** The interaction between the Sun and the Earth's magnetic field can cause the magnetic field measured on the Earth's surface to change rapidly. These changes can cause damaging currents to run through critical infrastructure such as power lines. Furthermore, the magnetic field can change a large amount in a small distance, meaning that stations that record the magnetic field can have drastically different readings even if they are located close to one another. We can use machine learning algorithms to forecast the risk of these changes in the magnetic field measured on the ground. Using data from stations that measure the magnetic field in addition to data from satellites in between the Sun and the Earth could improve forecast. Here, we created two different models, one using just the data from in between the Sun and the Earth, and one that also used data from the stations that measure the magnetic field on the Earth's surface, to see how they compare in their ability to perform risk assessments.

## 1. Introduction

The interaction between the solar wind and the Earth's magnetic field can cause rapid and large fluctuations in the horizontal component of the geomagnetic field observed on the ground (Bolduc, 2002; Kappenman, 2003; Ngwira et al., 2015; Oliveira & Ngwira, 2017; Rogers et al., 2020). These fluctuations enhance the geoelectric field at ground level which in turn causes currents to flow through conducting infrastructure (Beggan, 2015; Gannon et al., 2017; Ngwira et al., 2015; Viljanen, 1997). These currents, known as geomagnetically induced currents (GICs), can damage transformers, corrode pipelines, and interfere with train signals (Boteler, 2001, 2019; Pirjola, 2000; Pulkkinen et al., 2017; Rosenqvist & Hall, 2019). The ever increasing technological dependency of our society means that predicting the risk of GIC occurrence is essential for preventing the worst of the damage. A known problem is that GIC measurements are rarely available, either because they tend to be proprietary information or because the different systems of interest are not always monitored. However, GICs are known to

**Funding acquisition:** Amy Keesee, Hyunju Connor, Don Hampton

**Investigation:** Michael Coughlan, Amy Keesee, Raman Mukundan, Jeremiah Johnson

**Methodology:** Michael Coughlan, Amy Keesee, Victor Pinto

**Project Administration:** Amy Keesee, Victor Pinto, Jeremiah Johnson, Hyunju Connor, Don Hampton

**Resources:** Michael Coughlan, Amy Keesee, Hyunju Connor

**Software:** Michael Coughlan, Victor Pinto, Jeremiah Johnson

**Supervision:** Amy Keesee, Victor Pinto, José Paulo Marchezi, Jeremiah Johnson, Hyunju Connor, Don Hampton

**Validation:** Michael Coughlan

**Visualization:** Michael Coughlan, Amy Keesee, Victor Pinto, Raman Mukundan, José Paulo Marchezi

**Writing – original draft:** Michael Coughlan

**Writing – review & editing:** Michael Coughlan, Amy Keesee, Victor Pinto, Raman Mukundan, José Paulo Marchezi, Jeremiah Johnson, Hyunju Connor, Don Hampton

correlate well with ground magnetic field fluctuations ( $dB/dt$ ), providing us with a viable proxy for GIC measurements as magnetic field recordings tend to be readily available (Huttunen et al., 2002; Thomson et al., 2011; Viljanen, 1997).

In the past, there have been several attempts to directly forecast  $dB/dt$  (Blandin et al., 2022; Keesee et al., 2020; Pinto et al., 2022; Pulkkinen et al., 2013; Tóth et al., 2014). However, the extreme variability of the ground magnetic field can prove challenging to forecast and the rapid changes are difficult to capture perfectly. An additional problem is that for data-driven models, the volume of low level fluctuations even during storm time bias the models in such a way that they tend to under-predict the magnitude of the largest spikes in  $dB/dt$ . In response to those complications, several studies have moved away from attempting direct predictions of  $dB/dt$  and instead have focused on creating probabilistic forecasts to determine if  $dB/dt$  will cross a threshold value (Camporeale et al., 2020; Smith et al., 2021). A different problem arises from the fact that most studies look at the effects that large spikes in  $dB/dt$  at high latitudes produce. However, mid and low-latitudes are also susceptible to extreme events resulting in high  $dB/dt$ , especially as the auroral electrojet (AE) moves equatorward during geomagnetic storms (Thomson et al., 2011). Because studies tend to group high and mid-latitude together, sometimes the effects at mid-latitude can be severely downplayed (e.g., Pulkkinen et al., 2013).

An advantage for forecasting models is that advanced warnings of space weather events that cause large  $dB/dt$  can be provided by the solar wind monitors sitting at the first Lagrange (L1) point between the Sun and Earth. Depending on the speed of the solar wind, the events have a travel time of 20–90 min between the L1 point and the Earth's bow shock (Baumann & McCloskey, 2021). If predictions can be made using near-real time data from the satellites at the L1 point, this lead time would allow for preparations to prevent the worst of the damage. With the field of space weather entering a new data-driven era, due in part to advancements in state-of-the-art machine learning techniques and computer processing power (Camporeale, 2019), the satellites at the L1 point may prove crucial, as they have been in operation for decades, thus providing a rich set of data that can be used to train machine learning forecasting models. The fluctuations in the ground magnetic field, measured using magnetometers, contain data from stations all around the world, spanning back to the 1970s, providing another source of high-quality, long-term data for this work. The SuperMAG network (Gjerloev, 2012), one of the largest repositories for magnetometer data, is also critical when trying to forecast  $dB/dt$ .

The exclusion of ground magnetometer data in forecasting models is often done to prevent the models from over-relying on ground magnetic field data and becoming no better than a simple persistence model. However, using only solar wind data from the L1 monitors as input to forecasting models (Boteler & Jansen Van Beek, 1999; Dimmock et al., 2020; Ngwira et al., 2015, 2018; Pulkkinen et al., 2015; Viljanen et al., 2015) presents an issue related to differences in  $dB/dt$  measured at stations in close geographic proximity. In their study on Region-to-Specific Difference (RSD), Dimmock et al. (2020) was able to find correlations between solar wind parameters and differences in  $dB/dt$  between a single magnetometer station and the average  $dB/dt$  value of nearby stations. Specifically, they found that when  $B_z$  was southward and the solar wind velocity is elevated, RSD could reach more than 100 nT/min, meaning that the differences in the  $dB/dt$  measured at stations close to one another could differ by a similar magnitude during storm time conditions. In creating forecasting models for stations close to one another using only solar wind parameters, we are supplying the models with data that correlates with high RSD. If the differences between the station values are consistent with one of the stations always having higher values the models may be able to learn this behavior and make appropriate forecasts for each station. However, Ngwira et al. (2018) found that localized  $dB/dt$  can be attributed to magnetospheric and ionospheric processes, indicating that RSD cannot always be attributable to one station consistently having higher  $dB/dt$  values. If this is the case we must seek to include additional inputs to forecasting models to capture localized information. Here, we include ground magnetometer data from the SuperMAG network as its data overlaps well with the solar wind monitors, and it represents the final stage in the chain from sun to ground magnetic field fluctuations. Additionally, the use of ground magnetometer data will allow the models to capture the second order effect of internal magnetotelluric currents, which in some areas can contribute up to tens of percent to the total  $dB/dt$  value measured at a magnetometer station (Tanskanen et al., 2001). Future work will seek to include additional data from sources in between the sun and earth's surface. To evaluate how the additional ground magnetometer input parameters affect the models, we compare a model using only solar wind inputs to one using solar wind and ground magnetometer inputs (hereafter referred to as the combined model).

In this work, we use input data at time  $t$  to predict the probability of  $dB/dt$  exceeding a threshold value in a time window set from  $t + 30$  to  $t + 60$ . In Section 2 we describe the datasets that were utilized. In Section 3 the forecasting methodology is discussed. In Section 4 we discuss the results and compare the models, as well as

**Table 1**

*Ground Magnetometer Stations Used in This Study, Their Geomagnetic Location (Mean Value From 1998 to 2017), the 99th Percentile dB/dt Value in nT/min, and the Continental Location of Each Station Used for Reference in Section 4*

Station name	Code	MLat	MLon	99th percentile (nT/min)	Location
Brorfelde	BFE	52.20	89.48	6.801	Europe
Wingst	WNG	50.16	86.67	5.587	Europe
Lerwick	LER	58.08	80.95	8.052	Europe
Eskdalemuir	ESK	52.76	77.22	6.525	Europe
St. Johns	STJ	53.00	31.42	4.940	North America
Ottawa	OTT	55.35	1.96	7.145	North America
Newport	NEW	54.74	-55.59	6.648	North America
Victoria	VIC	53.69	-62.82	5.456	North America

input parameter analysis. Finally in Section 5 we present our conclusions and proposals for future work.

## 2. Data Sources

For this work we utilized solar wind data from the Advanced Composition Explorer (ACE) satellite (Stone et al., 1998). ACE is an L1 monitor that has been in operation since 1998. Because of the need for operational forecasting, ACE data was chosen for the solar wind parameters in this study instead of OMNI (Papitashvili & King, 2020) which is time-propagated to the bow shock. By choosing ACE data we expect the model to learn the relationship between measurements at the L1 point and the effects observed by the ground magnetometers at a later time. The ACE data requires re-sampling to fit with the 1 min resolution of the magnetometer data. ACE plasma data is up-sampled from its 64 s resolution using a backward filling method, and the ACE magnetic field data is down-sampled using a rolling average of the 16 s resolution data points. Data was downloaded from CDAWeb for the years 1998–2017. From the ACE data we extracted the total interplanetary

magnetic field ( $B_{Total}^{GSM}$ ), the Y and Z components in the GSM coordinate system ( $B_y^{GSM}, B_z^{GSM}$ ), the three components of the solar wind velocity ( $V_x, V_y, V_z$ ), solar wind density ( $\rho_{sw}$ ), and temperature ( $T$ ). It should be noted that we are using ACE science data here as opposed to the ACE near real time (NRT) data despite the desire to eventually perform real time forecasting of large  $dB/dt$  events. Working with real time data presents its own unique challenges when creating machine learning models with curated datasets. In particular when considering ACE data, Smith et al. (2022) found significant differences between the NRT and the post-processed science data. For the solar wind density and temperature they found that only ~20%–30% of the NRT data falls within 10% of the science values. They also report a large divergence between the NRT and science  $B_y^{GSM}$  and  $B_z^{GSM}$  values. They also find that 68% of the NRT  $B_{Total}^{GSM}$  falls within 10% of the science values, as do the vast majority (96%) of  $V_x$  values. Addressing these issues for real time forecasting, including the need for rapid processing and release of data, deserves its own body of work which we hope to address in the near future while building on the work presented here.

To train, validate, and test our models, eight magnetometer stations were selected from the SuperMAG ground magnetometer database. Their descriptions can be found in Table 1. The data from each of these stations is of 1-min resolution and had the baseline removed as described in Gjerloev (2012). These stations were chosen because their magnetic latitude fell between 50 and 60°, placing them within the mid-latitude region. Additionally, these stations span several Magnetic Local Time (MLT) sectors and capture the ground magnetic field response in different locations during storm time. While analysis of the MLT dependence will not be completed in this work, this criteria was used to build a framework for future work. The stations were also chosen due to their data availability overlap with ACE data, with all stations having data available from 1998 to 2017. From the SuperMAG stations we extracted the North and East components of the ground magnetic field ( $B_N, B_E$ ), the solar zenith angle (SZA), and the MLT which was transformed into  $\sin(\text{MLT})$  and  $\cos(\text{MLT})$  so the model could learn the dependence without the hard transition from 23 MLT to 0 MLT. The SZA was used as an indicator of the seasonal variations in the Earth's susceptibility to large spikes in  $dB/dt$  (Russell & McPherron, 1973). SZA can be calculated well in advance for a given time resolution. The  $B_N$  and  $B_E$  components of the geomagnetic field were used to calculate the total horizontal field ( $B_H$ ) and the  $dB/dt$  values for each station using Equations 1 and 2 below.

$$B_H = \sqrt{(B_N)^2 + (B_E)^2} \quad (1)$$

$$\frac{dB}{dt} = \sqrt{\left(\frac{dB_N}{dt}\right)^2 + \left(\frac{dB_E}{dt}\right)^2} \quad (2)$$

In addition to the ACE and SuperMAG data, the AE Index and the SYM-H index were taken from the OMNI database. The SYM-H was used to define the storm time data used for model training (see Section 3) and was not explicitly used as training data. The AE Index is input to the model so it can learn the dependency of  $dB/dt$

**Table 2***Input Parameters for the Solar Wind Model and the Combined Solar Wind + Ground Magnetometer Model*

Model	Input parameters
Solar wind	$B_y^{GSM}$ , $B_z^{GSM}$ , $B_{total}^{GSM}$ , $V_x$ , $V_y$ , $V_z$ , $\rho_{sw}$ , $T$ , $\sin(\text{MLT})$ , $\cos(\text{MLT})$
Combined	$B_y^{GSM}$ , $B_z^{GSM}$ , $B_{total}^{GSM}$ , $V_x$ , $V_y$ , $V_z$ , $\rho_{sw}$ , $T$ , $\sin(\text{MLT})$ , $\cos(\text{MLT})$ , $B_N$ , $B_E$ , $B_H$ , $dB/dt$ , AE Index, SZA

on current geomagnetic activity. While the AE Index is not available conventionally in real time, it is available graphically in near-real time from the Data Analysis Center for Geomagnetism and Space Magnetism at Kyoto University. It is included in the inputs to our models in the hope that it can be transitioned to a real time feed that could be used for forecasting. The same is true for the SuperMAG data, which is not currently available in real time and would be needed in near-real time to enable this type of forecasting. The input parameters for both models can be found in Table 2.

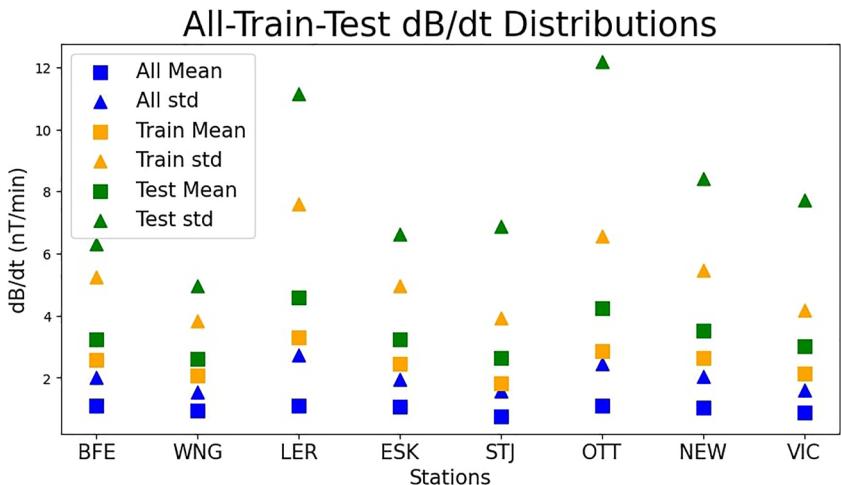
Because  $dB/dt$  is used to create the target data, its use as an input feature for the models raises concerns that the model would become too dependent on this one feature. To ensure that the model is not just propagating  $dB/dt$  values from the input array to determine if there is a threshold crossing in the prediction window, we created a persistence model for comparison. This persistence model has an output of 1 at time  $t$  if  $dB/dt$  is above the threshold during the input window (time  $t$  to  $t - 30$ ), and 0 otherwise. Using the entire input window to create the persistence model enables better comparison with the rolling 30 min forecasting window applied to the threshold crossings.

### 3. Methodology

#### 3.1. Data Preparation

The volume of quiet time measurements heavily biases the  $dB/dt$  values at any particular station. For instance, the median  $dB/dt$  value for the OTT station for the years 1998–2017 is  $\sim 0.91$  nT/min, significantly lower than its 99th percentile value of 7.145 nT/min. To address this bias, a similar method for restricting training data to storm time used in Pinto et al. (2022) was implemented. Times where  $\text{SYM-H} < -50$  nT for a minimum of 2 hr were identified. 12 hours of lead time and 24 hr of recovery time were added to the point of minimum  $\text{SYM-H}$  for each segmented period. Using this method, 397 storms were identified from the years 1998–2017, resulting in  $\sim 9 \times 10^5$  training samples, about 9% of the total possible training samples in the 20 year period being examined. After separating out the storm time data, the median  $dB/dt$  for the OTT station increases by more than 50% to  $\sim 1.43$  nT/min. For evaluation, previous studies (Blandin et al., 2022; Keesee et al., 2020; Pinto et al., 2022; Smith et al., 2021; Tóth et al., 2014) have used standardized threshold crossings defined in Pulkkinen et al. (2013) to evaluate the performance of their models. However, these threshold values were chosen to garner a statistically significant number of threshold crossings while looking at both high and mid-latitudes. Because we are looking at just mid-latitude stations, these thresholds are too high for such a purpose and lower values must be used. We chose the 99th percentile  $dB/dt$  value at each station as calculated from the SuperMAG data from the years 1998–2017. We then evaluate our model on a selection of geomagnetic storms. A breakdown of the  $dB/dt$  profiles for each station for all available data, storm time training data, and testing storm data can be found in Figure 1. The list of storms that were used for testing and model evaluation can be found in Supporting Information S1. Storms are a subset from the list of suggested storms from Welling et al. (2018), chosen to create a set of storms with varying intensity across a large span of the available ACE data set.

Data gaps are a continuing challenge for using machine learning for space weather forecasting. Methods of dealing with these missing data points are varied but the most common is linear interpolation (Blandin et al., 2022; Keesee et al., 2020; Pinto et al., 2022; Smith et al., 2021). Smith et al. (2022) examined how many gaps in Near-Real Time solar wind data can be filled by limiting the linear interpolation to 5 and 15 min, finding that approximately 90% of data is continuous for 30 min of time history when applying the 15 min limit. We find between 70% and 80% data available for all stations for 30 min of time history when linearly interpolating gaps of up to 15 min and restricting the analysis to the storm time science quality ACE data. A figure containing these results for all stations can be found in Supporting Information S1. The drop from the results in Smith et al. (2022) is most likely due to the post processing of the ACE data discarding data with lower quality flags and the tendency of data



**Figure 1.** The  $dB/dt$  data distributions for all data at each station in blue, the training data in orange, and the test storms in green. The squares indicate the mean  $dB/dt$  value and the triangles the standard deviation value.

to be missing more often during storm times due to instrument saturation. For this work we use up to 15 min of linear interpolation to maximize the available training data.

### 3.2. Convolutional Neural Network

We used a 2-Dimensional Convolutional Neural Network (CNN) (LeCun et al., 1990) to perform the probabilistic predictions. Performing this type of forecast turns the problem of directly predicting  $dB/dt$  into a binary classification problem. This is a particular strength of the CNN which is commonly used for tasks like photo or hand-writing identification. The CNN has also shown promise in its ability to utilize input parameter time history to make predictions (Pinto et al., 2022; Smith et al., 2021). To obtain probabilistic predictions, a softmax activation function is applied component-wise to the outputs of the last layer of the model. Equation 3 shows the softmax activation function where  $x_i$  is the output of the  $i$ th node. This type of activation normalizes the binary class outputs to be between 0 and 1.

$$\text{softmax}(x_i) = \frac{\exp x_i}{\sum_{j=1}^n \exp x_j} \quad (3)$$

Limited hyperparameter tuning of the depth of the network, the number of filters/nodes, filter window size, and time history were performed. The final model has only one CNN layer using 128 filters, a  $2 \times 2$  window, and “same” padding. A learning rate of  $1e-6$  and ReLu activation were used for training, with the exception of the final layer which used the softmax activation. The CNN layer is followed by a maxpooling layer and three dense layers with dropout layers in between. The input dimensions are  $(N, 30, P, 1)$  where the  $N$  indicates the number of training samples, 30 corresponds to the time history,  $P$  the number of input features (16 for the combined solar wind and magnetometer model, and 10 for the solar wind only model), and 1 the number of channels. The final output of the model is of dimensions  $(N, 2)$ , with  $N$  again being the number of input arrays as before, and the second dimension corresponding to the two possible classes (crossing or no crossing). Model results presented in Section 4 represent outputs from the positive (crossing) node. The feature inputs for the solar wind driven model and the combined model are shown in Table 2. The  $\sin(\text{MLT})$  and  $\cos(\text{MLT})$  are included in the Solar Wind model because we wanted to examine the affect of adding the ground magnetic field data, not the position of the magnetometer which can see larger  $dB/dt$  at particular MLT locations. For a more detailed explanation of CNNs see Siciliano et al. (2021).

Because machine learning algorithms and the methods used to prepare data are subject to unavoidable bias and stochastic randomness, we seek to attach a confidence level to our predictions by performing 100 randomly shuffled splits of the training and validation data where each split of the data uses 80% for training and 20% for validation. Unique models were trained on each of the randomly shuffled training-validation pairs and the 95th

percentile of the 100 model outputs is used as the confidence level. This provides us with a measure of how the differences in the training-validation pairs can affect the predictions.

### 3.3. Model Explainability

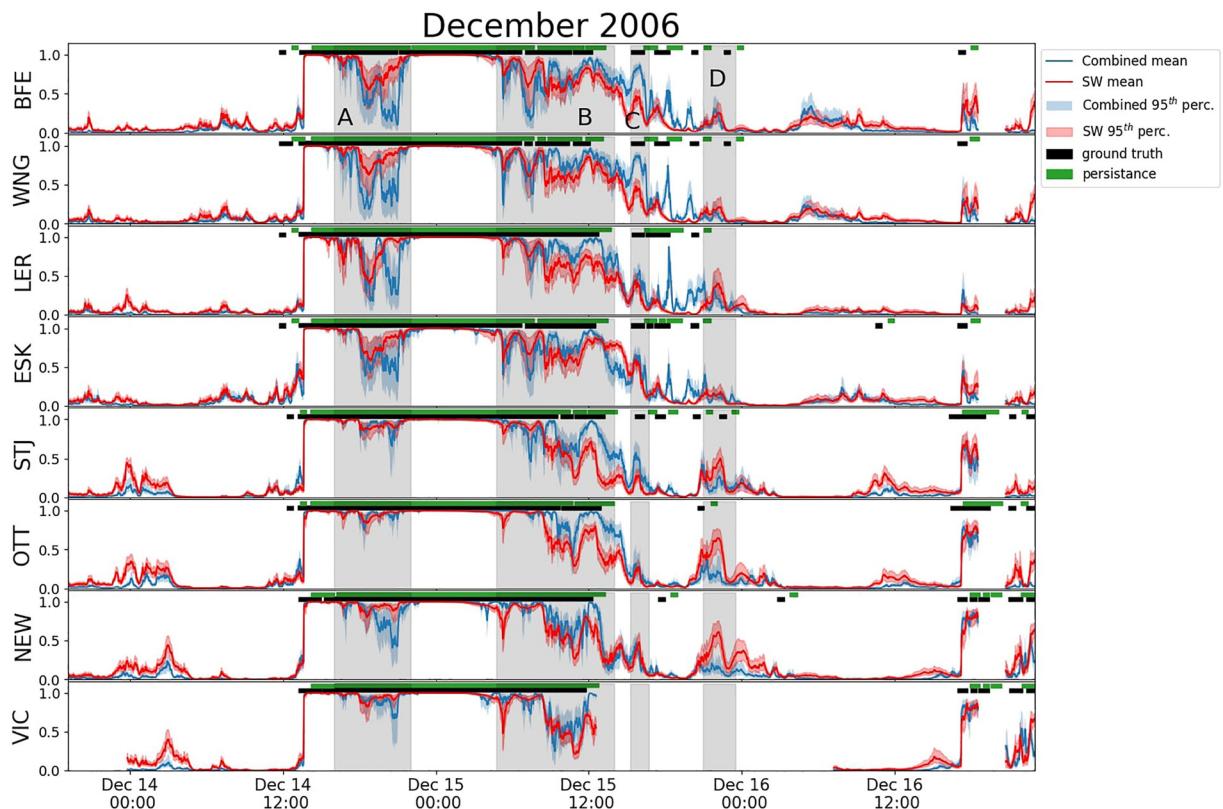
Machine learning algorithms have enormous potential for space weather forecasting due to their deployment speed once trained and their ability to continually improve as new data becomes available via online/offline learning. One of the drawbacks of utilizing neural networks models such as the one featured in this work is determining the driving factors in the model's output. The so called black box or opaque box problem presents a practical issue for space weather forecasters who need to understand why a model may output an elevated risk, but more fundamentally it prevents us from determining whether these data-driven models align with our physical understanding of underlying processes. This could lead to ineffective or misleading models and could inhibit our ability to make data driven discoveries. In response to the increased use of machine learning in space weather there has been a movement toward using methods more attuned to explainability (such as the many tree based models), or to utilize more in-depth methods of model interpretation (Iong et al., 2022; Reddy et al., 2022) such as SHapley Additive exPlanation (SHAP).

The SHAP method (Lundberg & Lee, 2017) for determining feature importance is based on Shapley values (Shapley, 1953) which use cooperative game theory to determine the marginal contribution of model input parameters. In theory it does this by using an additive feature method that retrains the model for all input feature subsets. In assigning an importance value to each feature the method compares a model retrained on a subset of possible features including the feature being examined, and one trained with the same subset but with the examined feature withheld. To account for the effect of the other features in the set on withholding the examined feature it repeats this process for all possible subsets and input features. The final importance value for the examined feature is determined by taking the weighted average of all the individual importance values. In practice this is prohibitively expensive as it requires retraining  $2^N$  models where N is the number of input features, which in our case would be either 1,024 for the solar wind model or 65,536 for the combined model. Thankfully a Shapley Sampling method can be used which applies sampling for the feature importance calculations and approximates the effect of withholding features by integrating over samples from the training data set. In doing this we both avoid having to retrain the models and reduce the computational burden of performing  $2^N$  replacements.

Feature importance can be global or local. Global feature importance provides a general picture of the influence of a feature on the model over the entire training set, where a local feature importance determines the feature's contributions to a single prediction (Iong et al., 2022). SHAP is a model agnostic method that determines local feature importance using the game theory approach of Shapley values. A note must be made here about how the feature independence assumption of the SHAP method could affect our feature importance results. If two input features have a high correlation or dependency then the calculations of their marginal contributions to one another can be skewed. For instance, if feature A and B are highly correlated, and feature A is being withheld from a subset containing feature B, the output of the model may not change significantly due to the addition of feature A because the model already had the required information from feature B. In this case, feature A may not receive any marginal importance value from this model output difference calculation. This can lead to feature A having an artificially low or feature B an artificially high global and specific feature importance. Alternatively the marginal contributions can be shared between the features through permutations of the different subsets, creating relatively equal feature importance values where one should be significantly higher. The sets of parameters that are most highly correlated in this work are  $B_y^{GSM}/B_{total}^{GSM}$ ,  $B_E$  and  $B_N$  with  $B_H$  and  $B_H$  with  $dB/dt$ . Special consideration will be paid to these sets of parameters when their contributions are being examined.

## 4. Results

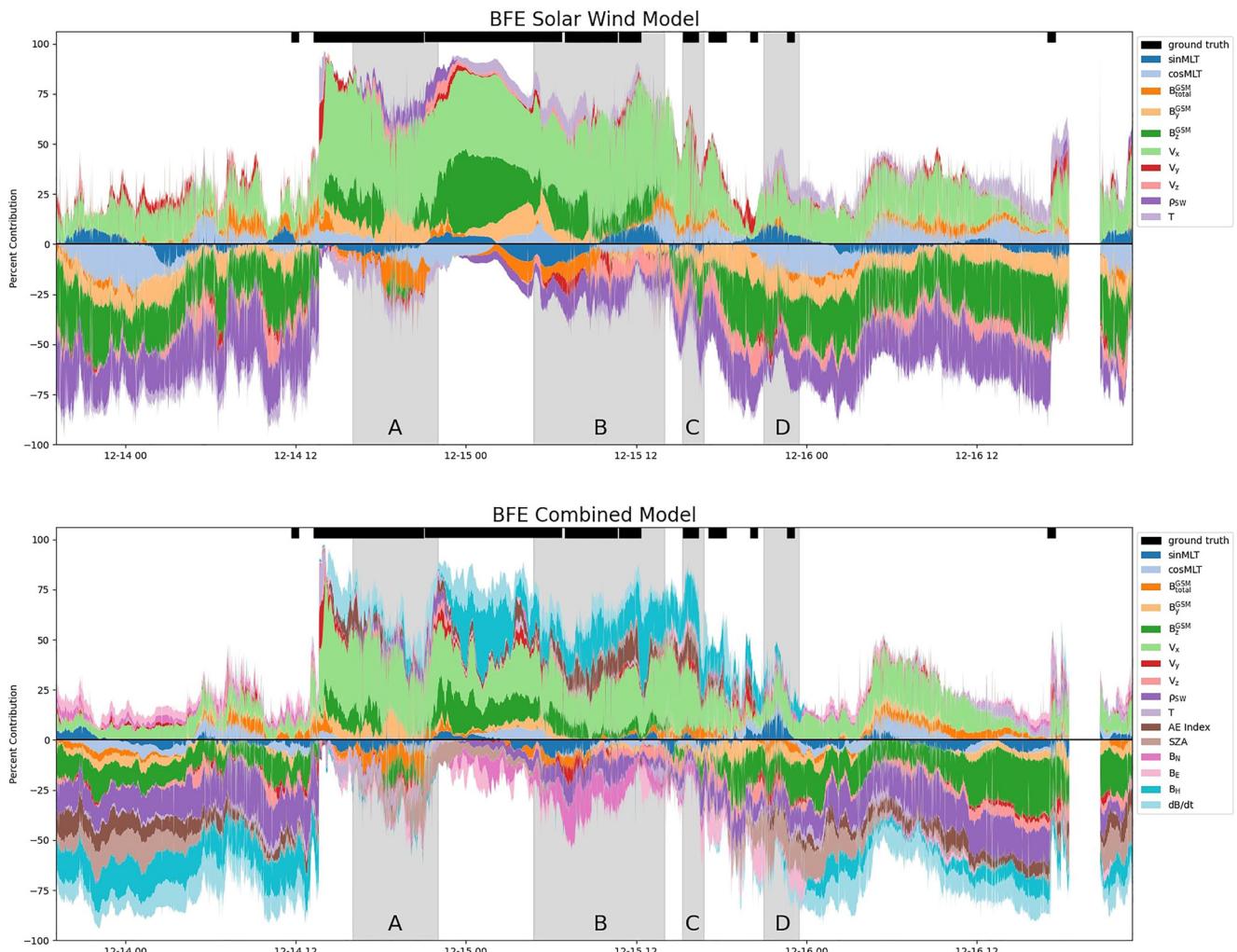
The results for two of the test storms, the December 2006 and March 2015 storms, can be seen in Figures 2 and 4 below. Model results for each station can be seen in their respective panel, with the easternmost station (BFE) in the top panel and the westernmost (VIC) in the bottom panel. Figures 3 and 5 are the parameter SHAP values for the December 2006 storm at the BFE station and the March 2015 storm at the LER station, respectively, as a percentage of the total contribution to the model's output. Positive percentages push the model's probabilistic output toward 1 (a threshold crossing) and negative percentages toward 0 (no crossing). The top panel is the



**Figure 2.** Model outputs for the December 2006 storm. Dark blue line indicates the mean model output for the combined model and the light blue shading indicates the 95th percentile. The red line is the mean output for the solar wind only model and the light red shading the 95th percentile. The black bar at the top of each panel indicates where there is a threshold crossing 30–60 min into the future (positive model class). The green bar is the persistence model. The y-axis is the model's positive class output ranging from 0 to 1. The model will output a high probability in this class if it thinks there will be a threshold crossing in the future prediction window. The gray shading highlights periods of interest.

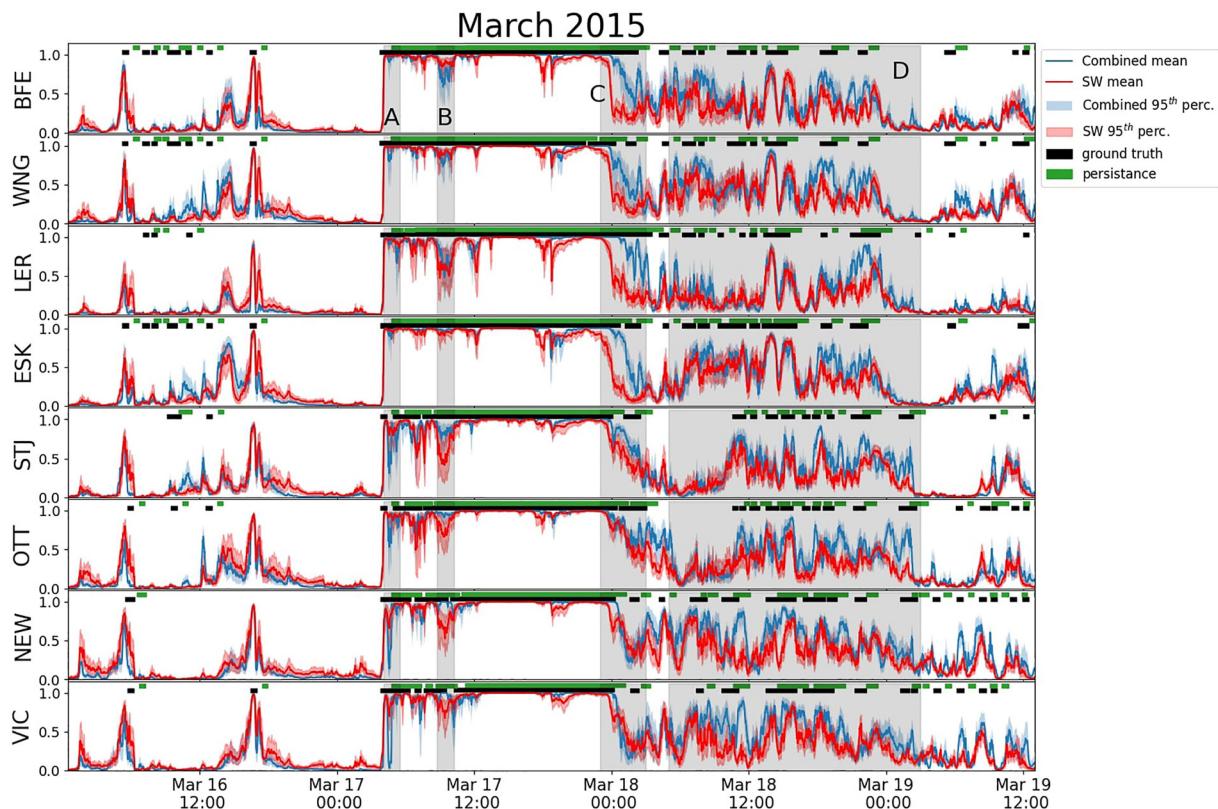
values for the solar wind model and the bottom the combined model. The rest of the storms in the test set, as well as the SHAP value percentage contributions for the other stations and storms, can be found in Supporting Information S1.

Overall, both models are able to reliably forecast the initial effect of the onset and the main phase of the storms on the  $dB/dt$  at each station. In Figure 2 the models show a period of elevated probabilities around 00:00 on December 14th which doesn't correspond with any threshold crossings, with the solar wind models showing higher outputs than the combined models. This coincides with small fluctuations in  $B_z$ ,  $V_x$ ,  $V_y$ , and  $T$  in the solar wind, displaying the solar wind model's heavier reliance on these parameters than the combined model. This is evident in Figure 3 with  $V_x$  (light green) showing a significant positive contribution. Plots of solar wind and magnetometer data during the testing storm periods can be found in Supporting Information S1. The European stations (BFE, WNG, LER, and ESK), OTT and the STJ station have a short period where the ground truth is positive right before the main period of threshold crossings. Though the models don't perfectly capture these first threshold crossings, their probability outputs become raised above baseline, indicating they are picking up on elevated geomagnetic activity. The NEW and VIC stations do not have a crossing and show no such elevated output. All model outputs jump at approximately 13:00 UT just after the onset of the main period of threshold crossings. The model probabilities then become more variable before returning to high values during the main phase of the storm, after which the probabilities drop back closer to zero. During the later half of the main phase (Figure 2 block B) many of the solar wind parameters are gradually returning to baseline values, and the solar wind models begin to under-predict the combined models while the ground truth is still positive. The models output higher probabilities starting at about 16:00 UT on December 16th where all stations except for LER see a small period of threshold crossings. The North American stations (STJ, OTT, NEW, and VIC) see a more prolonged period of threshold crossings and elevated model probabilities.



**Figure 3.** BFE model percentage contributions of each parameter's SHAP values for the solar wind model (top panel) and combined model (bottom panel) during the December 2006 storm. Positive percentage contributions influence the model's output to be closer to one and negative percent contributions push the model output to be closer to zero. The width of the percent contributions is 100%. Black bar at the top of the panels indicates a threshold crossing 30–60 min into the future. Gray blocks indicate areas of interest corresponding to those in Figure 2.

Figure 4 displays the results for the 17 March 2015 storm. The models for all stations show a spike in probability early in the testing period, seemingly due to a rapid change in the solar wind density (dark purple in Figure 5), and a small but rapid change in IMF  $B_y$  and  $B_z$ . Another spike in solar wind density appears to cause the second main peak in probabilities seen across all station models near 16:00 UT on 16 March. All station models return to near zero with the solar wind models remaining slightly more elevated until the main period of threshold crossings begins between 04:00 and 05:00 on 17 March, when the models all approach a probability of one. The models output high probabilities matching the ground truth data with some variability until 00:00 UT 18 March (Figure 4 block C). After this point the uniformity of positive ground truth data across the stations breaks, and threshold crossings become more intermittent, most notably at the STJ and OTT stations which have a large gap in crossings. During this gap both STJ models respond almost immediately, with outputs falling below 0.5 and remaining low for the duration of the gap. The solar wind OTT models exhibit a similar behavior to the STJ models, though they remain slightly more elevated. However, the combined OTT models don't see immediate output drops and instead remain elevated for more than an hour after the beginning of the threshold crossing gap. This difference is likely due to the OTT station seeing a solitary spike in  $dB/dt$  of larger intensity than present at the STJ station and a correspondingly elevated  $B_H$  value during this period. The threshold crossings become more intermittent in the later half of the testing period, and the model outputs become more variable, with the solar wind and combined models following similar trends.



**Figure 4.** Model outputs for the March 2015 storm (St. Patrick's Day storm). Same format as Figure 2.

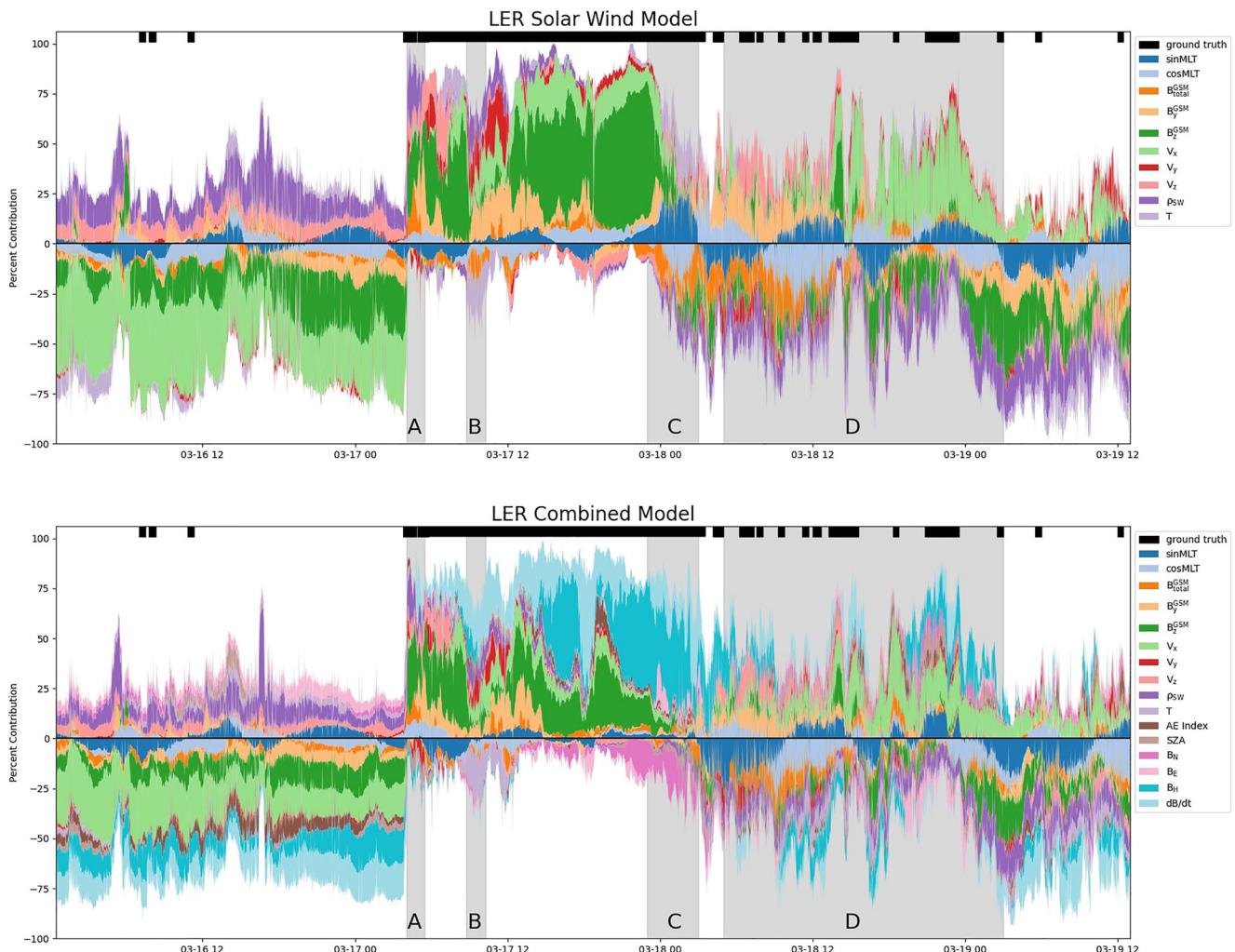
#### 4.1. Qualitatively Evaluating Differences in Model Behaviors

##### 4.1.1. December 2006 Storm

Beginning with the December 2006 storm in Figure 2 we can see the solar wind models for stations close to one another behaving similarly, indicating they have some dependence on the  $\sin(\text{MLT})$  and  $\cos(\text{MLT})$  (the light and dark blue contributions in Figure 3). During the dip in probabilities around 17:00 UT on 14 December (Figure 2 block A) both models for the European stations show a larger drop in probabilities than the models for the North American stations. However, the solar wind model only has one of these large drops, while the combined models have two consecutive periods of depressed outputs. This appears to be due to the ground magnetic field returning to baseline between the first smaller spike in  $dB/dt$  and the main phase of the storm for these stations, while the North American stations have more consistent  $dB/dt$  values.

The beginning of block B sees a northward turning that drops the outputs of many of the models. Interestingly it appears that the change is most drastic for the BFE and WNG combined models and the solar wind models for the other stations with the exception of ESK where both models appear equally affected. Shortly after this first drop in probabilities we see a second drop. This drop appears to be location specific, with the drops more pronounced in locations where  $\sin(\text{MLT})$  is negatively contributing (see Supporting Information S1 for all station SHAP plots). In the later half of block B we see the largest benefit of the combined model which remains elevated throughout the main phase of the storm when the solar wind models begin to under-predict due to the solar wind parameters returning to baseline.

At around 16:00 on 15 December (block C) the combined models for the BFE and WNG stations output much higher probabilities during a positive threshold crossing than the solar wind models. Figure 3 shows this is due mostly to the additional contribution of  $B_H$  (dark cyan) which remains more elevated at these stations. The solar wind and combined models for LER and ESK show similar elevated probabilities for the crossings at this point. The solar wind model at the STJ and OTT stations show equivalent small elevations during this period, though there is a crossing at the STJ station that is better captured by the combined model, and no threshold crossing at



**Figure 5.** LER model percentage contributions of each parameter's SHAP values for the solar wind model (top panel) and combined model (bottom panel) during the March 2015 storm. Gray blocks indicate periods of interest corresponding to those in Figure 4. Same format as Figure 3.

the OTT station. The solar wind models return to baseline after this point and do not reflect the spike in probabilities by the combined model at the European stations that occurs around 18:00 on the 15th between blocks C and D which is mostly driven by a significant increase in  $dB/dt$ .

The solar wind models at all stations except ESK and VIC (which is missing data during this period) show probabilities elevated above the combined models at approximately 21:00 that same day when there is a period of threshold crossings at BFE, WNG, STJ, and OTT (block D). Of these crossings, the solar wind model for the STJ station is the only one that is correctly timed, with the elevated probabilities at the other stations occurring between crossings, though the solar wind models for the OTT station see a response to the crossing even if the peak in the outputs occurs after the ground truth. These increases appear driven by an increase in  $V_x$  and the additional contribution of temperature, though there is no obvious change in that parameter.

Overall during the December 2006 storm we can see the strength of the combined models is their ability to maintain a higher probabilistic output during a period of positive ground truth in the later half of the storm when the solar wind models begin to under-predict. Additionally, we can see the solar wind models appear to be much more reactive to changes in the solar wind conditions, while the combined models' changes are less responsive to the solar wind inputs due to the additional contributions of the ground magnetometer inputs. However, the combined models show their susceptibility to the magnetometer inputs in block A, where their outputs are depressed due to decreases in  $B_H$ . The combined models appear to better capture the threshold crossings in block C, while the solar

wind models have mistimed but elevated response in block D. During the main phase  $V_x$  appears to be the input feature with the largest contribution for the solar wind model, and a combination of  $V_x$  and  $B_H$  for the combined model.

#### 4.1.2. March 2015 Storm

Moving to the March 2015 storm seen in Figure 4 we can see a drop in probabilities for the combined models at the North American stations just after the beginning of the main period of threshold crossings while the solar wind models remain elevated (block A). This appears to be driven by a negative contribution from  $B_{total}^{GSM}$ , several of the positional parameters ( $\sin(\text{MLT})$ , SZA), and marginal contributions from the ground magnetic field parameters. We also see a drop in most solar wind model outputs beginning around 09:00 on 17 March (block B). Figure 5 shows this drop being due to the negative contributions from the temperature and  $B_{Total}^{GSM}$ , and a drop in positive contributions in all of other solar wind parameters, all of which show strong fluctuations during this period. The solar wind models at STJ, OTT, NEW, and to a lesser extent VIC show much larger and more extended drops in probabilities compared to their combined models. The LER combined model also shows a drop in probability, displaying the models' heavier reliance on solar wind parameters at this time.

At 00:00 on 18 March (block C) we see a sharp drop in probabilities for the solar wind models at the European stations. This occurs about an hour before a similar drop in the combined model for the BFE, WNG, and ESK stations, and an hour before a period of large variability in the combined model for the LER station which drops in probability between 03:00 and 04:00. Looking at block C in Figure 5, we can see that the top panel shows this decline in positive contributions for the solar wind model, while the bottom panel shows that the combined model outputs higher probabilities at this time due to positive contributions from  $B_H$  and  $dB/dt$ .

Both the combined and the solar wind models show rapidly changing probabilities that follow a similar pattern during a period of intermittent threshold crossings (block D). During this time the solar wind models for the North American stations chronically under-predict the combined models. The solar wind models at the European stations more closely align with the combined model outputs until approximately 18:00 on 18 March when they also begin to under-predict the combined model outputs. As can be seen in Figure 5, the periods of under-prediction by the solar wind models in block D coincide with a larger positive contribution of ground magnetic field parameters. The same pattern holds for the SHAP contributions found in Supporting Information S1.

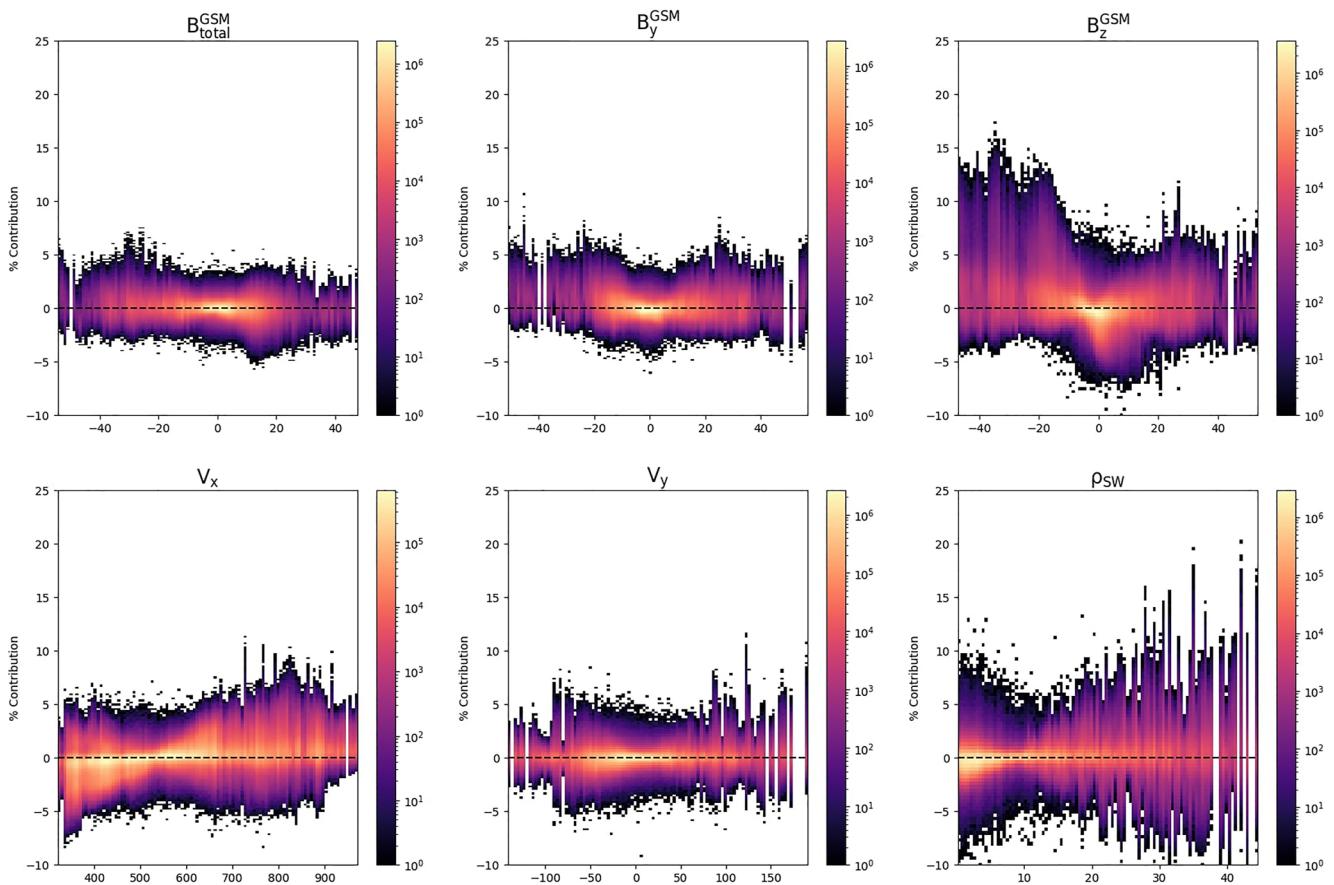
Overall we can see a pattern of the ground magnetometer inputs filling in the gap of positive prediction contributions when the solar wind parameters begin to fluctuate or return to their baseline values, most notably in the later parts of the storm. This indicates that some ground magnetometer information is necessary to capture the elevated risk during later periods of the main phase. In contrast to the December 2006 storm, the input feature with the largest contribution during the main phase of the March 2015 storm for the solar wind model is  $B_z^{GSM}$ . In addition to this,  $B_H$  and  $dB/dt$  have the highest contributions for the combined model.

#### 4.1.3. Overall Trends

Across the storms the LER, OTT, and NEW stations observe higher  $dB/dt$  than the ESK, STJ, and VIC stations. All model outputs appear to be heavily influenced by solar wind conditions and for the combined models, past local  $B_H$  and  $dB/dt$  values. Overall WNG and BFE behave similarly even when there are magnitude differences in  $dB/dt$ . All models are able to show improvement on the persistence models by outputting high probabilities near the beginning of the main phase of the storms, significantly before the persistence models show an indication of activity. During the main phase of the storms, all model outputs remain high, likely due to the continuing elevated input values for all parameters. Solar wind driven models all fall off faster toward the end of the main phase of the storm as the solar wind conditions begin to return to their baseline values. There are several periods where the combined models are better able to capture short spikes in  $dB/dt$  leading to threshold crossings, though there are notable exceptions such as just before 00:00 on 16 December (Figure 4 block D) where the solar wind model at the STJ station shows a higher probability for a positive threshold crossing.

### 4.2. Global Feature Importance

Figures 6 and 7 show histograms of the SHAP percent contribution as a function of parameter value for several input parameters for the solar wind and combined models, respectively. These result from taking a sample of 10 randomly chosen models for each station and calculating the SHAP values across all test storms. These values



**Figure 6.** Histograms of percent contribution as a function of input parameter value for a subset of Solar Wind Model inputs. Color bar indicates histogram count.

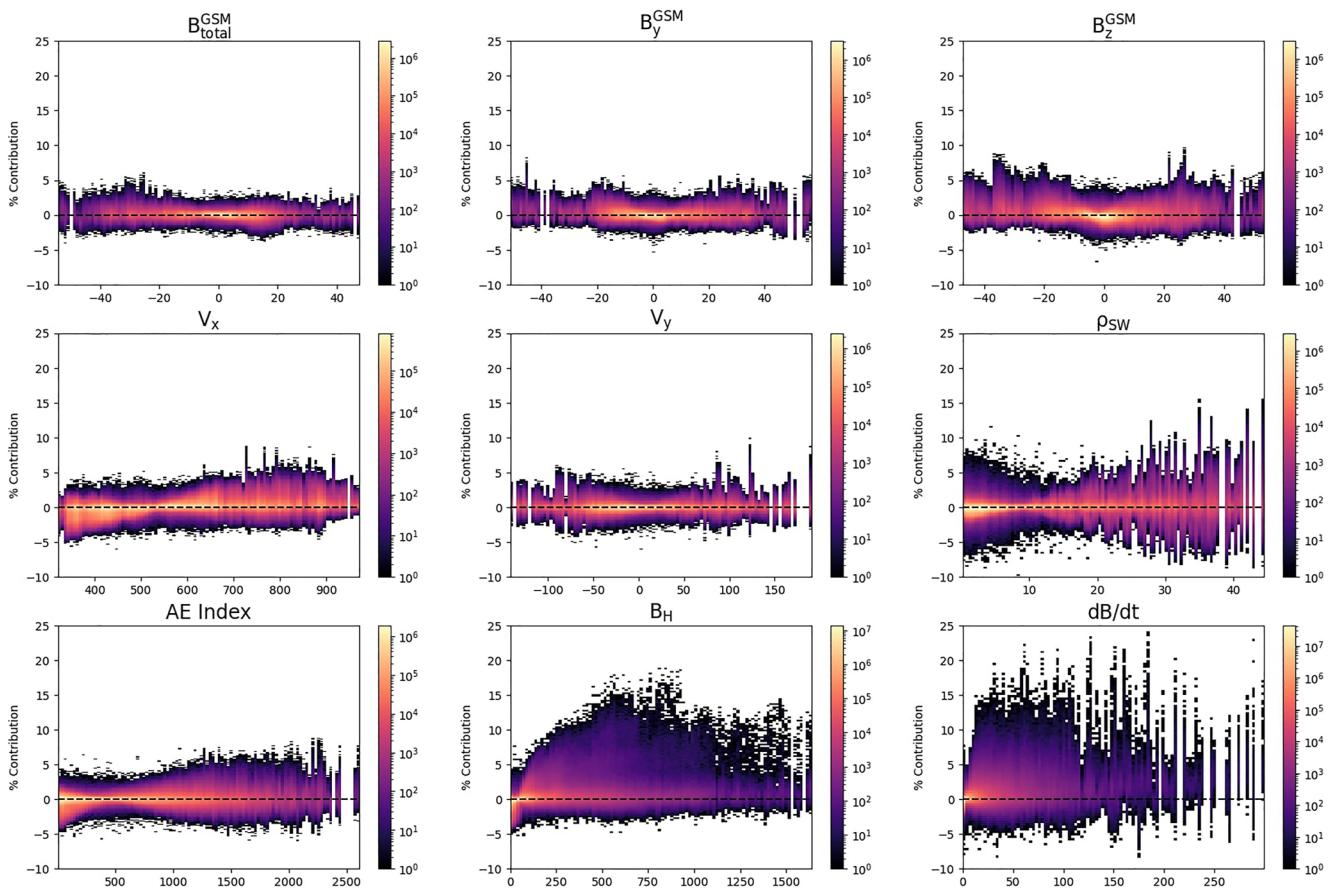
are then normalized as a percentage contribution for each input array. In Figure 6 we can see higher positive contributions for more negative  $B_z$ , and more negative contributions when  $B_z$  is closer to 0. Similarly,  $V_x$  shows more negative contributions for values closer to the  $V_x$  average and positive contributions for more extreme  $V_x$ . These confirm that the models are behaving in a way that we would expect given our current understanding of the interactions between the solar wind and magnetosphere.

Moving to Figure 7 we see the shape of the contributions for the solar wind parameters is similar to those in Figure 6 but with a smaller magnitude, which is to be expected as there are additional parameters included in the models. The inclusion of  $B_H$  and  $dB/dt$  clearly influences the model significantly, with both of these parameters having negative contributions near their baseline values and large, mostly positive contributions when their values increase. The large positive contributions of  $B_H$  for values elevated beyond its baseline reiterate the strong correlation between the change from baseline  $B_H$  and  $dB/dt$  discussed in Tóth et al. (2014).

#### 4.3. Metrics

To evaluate the predictive capacity of our models we require quantitative metrics. Choosing which metrics should be used for model evaluation in space weather forecasting is often a topic of debate, with each metric having benefits and limitations (Liemohn et al., 2021). Here, we choose to calculate a suite of metrics to gain a more complete understanding of our model's ability to determine the risk of  $dB/dt$  threshold crossings in the  $t + 30$  to  $t + 60$  time window.

Some of these metrics utilize a confusion matrix in calculating their scores. The confusion matrix uses the binary classification nature of our model to count the number of true positives (hits) ( $T_p$ ), false positives ( $F_p$ ), false negatives (misses) ( $F_n$ ), and true negatives ( $T_n$ ) that result from the comparison of our model's predictions and the ground truth. Because we trained 100 of each model for every station, we calculated the metric scores discussed



**Figure 7.** Histograms of percent contribution as a function of input parameter value for a subset of Combined Model inputs. Top two rows represent the same parameters as those in Figure 6. Bottom row contains a subset of parameters not present in the Solar Wind model.

below for each model and present both the mean and the 95th percentile. The precision and recall are calculated using Equations 4 and 5 below.

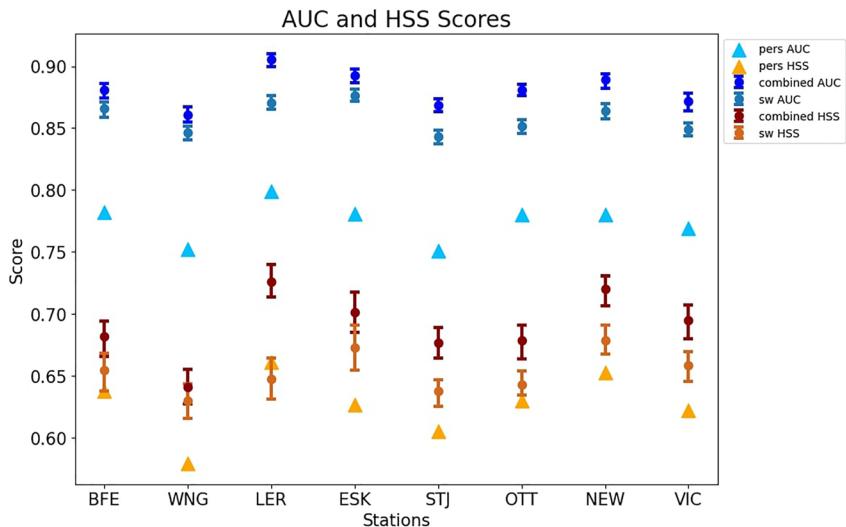
$$\text{Precision} = \frac{T_p}{T_p + F_p} \quad (4)$$

$$\text{Recall} = \frac{T_p}{T_p + F_n} \quad (5)$$

The precision-recall curve is created by defining a decision boundary, any probabilities above which count as a positive result and any below count as negative. The precision and recall metrics are calculated using this decision boundary. The boundary is changed and new precision and recall values are calculated. The decision boundary is swept from zero to one, calculating as many precision and recall values as there are unique model outputs. The values are plotted on a precision versus recall axis and the area under the resulting curve gives us the AUC score. This metric was used instead of the often used ROC score because the precision and recall do not take into account the rate of true negatives, which due to the quiet time bias in  $dB/dt$  measurements even during storm time can give seemingly good scores to models that chronically under-predict active times if they also output low probabilities during the more common quiet times. A static decision boundary of 0.5 is used to calculate the elements of the confusion matrix for the Heidke-Skill Score (HSS) (Equation 6).

$$\text{HSS} = \frac{2(T_p T_n - F_p F_n)}{(T_p + F_n)(F_n + T_n) + (T_p + F_p)(F_p + T_n)}. \quad (6)$$

Figure 8 shows the AUC and HSSs for the solar wind and combined models, as well as the scores of the persistence models. We can see that the combined models achieve higher AUC scores than the solar wind only models.



**Figure 8.** Area Under the Precision-Recall Curve (AUC) and Heidke-Skill Score (HSS) scores calculated across all tested storms. Dark blue are the AUC scores for the combined models and the lighter blue the solar wind only AUC scores. The dots show the mean AUC scores and the error bars are the 95th percentile of model outputs. The lighter blue triangles are the AUC scores for the persistence models. The dark red is the combined models HSS and the orange the solar wind only results. The light orange triangles indicate the persistence HSS scores.

Additionally the lower bound of the 95th percentiles for all of the solar wind and combined model AUC scores is higher than the persistence scores, indicating that the models are not over-relying on  $dB/dt$  as an input feature. The LER station has the highest combined model AUC scores and ESK the highest solar wind model scores. WNG has the lowest combined model scores and STJ the lowest, for the solar wind only model. The HSS scores follow a similar pattern to the AUC scores, with the mean values of the combined models still out-performing the mean of the solar wind only models. For the HSS the LER station still shows the highest scores for the combined model but has a solar wind model falling below the persistence model. Additionally, there is overlap in the 95th percentile values between the two models for the ESK, BFE, and WNG stations, though the mean values for the combined models still outperform the solar wind mean values. LER is the highest latitude station and has the highest metric scores for the combined models and WNG, the lowest latitude station, has some of the lowest metric scores, with no apparent correlation for the stations in between. In addition to the AUC and HSS metrics, the root mean square error (RMSE) and Bias were calculated and showed similar trends to the metrics above. The full results can be found in Supporting Information S1.

## 5. Conclusions

Here, we have trained two different CNN based models for eight mid-latitude ground magnetometer stations to predict the probability that  $dB/dt$  will exceed the 99th percentile 30–60 min in the future. One model used only solar wind data from the ACE satellite and  $\sin(\text{MLT})$  and  $\cos(\text{MLT})$  of the ground magnetometer from SuperMAG as inputs. The other model adds the AE Index and Solar-Zenith Angle data, and ground magnetic field data to the inputs. The training data consisted of only storm time data to reduce the quiet time bias present in real world  $dB/dt$  measurements. A bootstrapping method used a random split of the training and validation data to train 100 unique versions of each model type for each station with the 95th percentile of the results was used as a measure of uncertainty. The models were evaluated on eight storms selected as a subset of the storms in Welling et al. (2018) and modified to include additional quiet time at the beginning and ending of the storms.

The models perform well overall, with all models scoring at or above 0.83 on the area under the precision-recall curve (AUC) and above 0.61 on the HSS. The solar wind models were outperformed in all metric scores by the combined models. The models that included  $dB/dt$  in the inputs also outperformed the persistence models, alleviating fears that the combined models would rely too heavily on the  $dB/dt$  inputs. While the solar wind models have similar outputs for stations close to one another, the combined models showed variances between these stations, indicating the additional inputs to the combined model are necessary for more localized predictions. The

combined model outputs more closely followed the threshold crossing patterns in the real data, hence their higher overall metric scores. However this trend was not universal, with notable cases where the solar wind models more accurately captured future threshold crossings than the combined models. All models are limited in their ability to predict threshold crossings that happen significantly before the main phase of the testing storms, and become less accurate during the recovery phase of the storms where the threshold crossings become more intermittent.

The use of the SHAP values to approximate the influence of particular input parameters shows that these models are behaving in a way we would expect given our understanding of the solar wind—magnetosphere system. Percent contributions are elevated for solar wind and magnetometer parameter values that are indicative of storm time, when we expect the highest  $dB/dt$  values. Additionally, the SHAP values allow us to see how the contribution of  $B_H$  and  $dB/dt$  fills the gap left by the solar wind parameters as they return to their baseline values in the later half of the storm. This allows the combined model to remain elevated above the solar wind models during these later periods of positive ground truth.

Future work should seek to utilize data from sources in between the solar wind monitors and the ground magnetometers to aid in the localized prediction of  $dB/dt$  threshold crossings as well as to improve the metric scores and overall forecasting capability of these models, especially outside of the main phase of the testing storms. Additionally, times and locations of previously studied RSD will be examined using the methods presented here to further determine the necessity of including parameters in addition to solar wind data to make localized predictions.

## Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

## Data Availability Statement

ACE data can be found at <http://www.srl.caltech.edu/ACE/ASC/level2/index.html>. Ground magnetometer data from the SuperMag project can be found at <https://supermag.jhuapl.edu/>. The OMNI data used in this paper can be found at <https://cdaweb.gsfc.nasa.gov/pub/data/omni/>. Software written for this work can be found at (Coughlan, 2023).

## Acknowledgments

This work was supported by NSF EPSCoR Award OIA-1920965. HC was supported by the Space Precipitation Impacts project at Goddard Space Flight Center through the Heliophysics Internal Science Funding Model. VP thanks ANID Grant SA772100112. We thank all members of the MAGICIAN team at UNH and UAF that participated in the discussions leading to this article. We also thank the OMNIWeb, SuperMAG and ACE teams for providing the data. Additionally, the authors would like to thank Mayowa Adewuyi for discussions about model output visualizations.

## References

- Baumann, C., & McCloskey, A. E. (2021). Timing of the solar wind propagation delay between L1 and Earth based on machine learning. *Journal of Space Weather and Space Climate*, 11, 41. <https://doi.org/10.1051/swsc/2021026>
- Beggan, C. D. (2015). Sensitivity of geomagnetically induced currents to varying auroral electrojet and conductivity models. *Earth Planets and Space*, 67(1), 24. <https://doi.org/10.1186/s40623-014-0168-9>
- Blandin, M., Connor, H. K., ÖzTÜRK, D. S., Keesee, A. M., Pinto, V., Mahmud, M. S., et al. (2022). Multi-variate LSTM prediction of Alaska magnetometer chain utilizing a coupled model approach. *Frontiers in Astronomy and Space Sciences*, 9, 846291. <https://doi.org/10.3389/fspas.2022.846291>
- Bolduc, L. (2002). GIC observations and studies in the Hydro-Québec power system. *Journal of Atmospheric and Solar-Terrestrial Physics*, 64(16), 1793–1802.
- Boteler, D. H. (2001). Space weather effects on power systems. In *Geophysical monograph series* (pp. 347–352). American Geophysical Union. <https://doi.org/10.1029/GM125p0347>
- Boteler, D. H. (2019). A 21st century view of the March 1989 magnetic storm. *Space Weather*, 17(10), 1427–1441. <https://doi.org/10.1029/2019SW002278>
- Boteler, D. H., & Jansen Van Beek, G. (1999). August 4, 1972 revisited: A new look at the geomagnetic disturbance that caused the 14 cable system outage. *Geophysical Research Letters*, 26(5), 577–580. <https://doi.org/10.1029/1999GL000035>
- Camporeale, E. (2019). The challenge of machine learning in space weather: Nowcasting and forecasting. *Space Weather*, 17(8), 1166–1207. <https://doi.org/10.1029/2018SW002061>
- Camporeale, E., Cash, M. D., Singer, H. J., Balch, C. C., Huang, Z., & Toth, G. (2020). A gray-box model for a probabilistic estimate of regional ground magnetic perturbations: Enhancing the NOAA operational geospace model with machine learning. *Journal of Geophysical Research: Space Physics*, 125(11), e2019JA027684. <https://doi.org/10.1029/2019JA027684>
- Coughlan, M. (2023). Software for probabilistic forecasting of ground magnetic perturbation spikes at mid-latitude stations. *Zenodo*. <https://doi.org/10.5281/zenodo.7838850>
- Dimmock, A. P., Rosenqvist, L., Welling, D. T., Viljanen, A., Honkonen, I., Boynton, R. J., & Yordanova, E. (2020). On the regional variability of  $dB/dt$  and its significance to GIC. *Space Weather*, 18(8), e2020SW002497. <https://doi.org/10.1029/2020SW002497>
- Gannon, J. L., Birchfield, A. B., Shetye, K. S., & Overbye, T. J. (2017). A comparison of peak electric fields and GICs in the Pacific northwest using 1-D and 3-D conductivity. *Space Weather*, 15(11), 1535–1547. <https://doi.org/10.1002/2017SW001677>
- Gjerloev, J. W. (2012). The SuperMAG data processing technique. *Journal of Geophysical Research*, 117(A9), A09213. <https://doi.org/10.1029/2012JA017683>

- Huttunen, K. E. J., Koskinen, H. E. J., Pulkkinen, T. I., Pulkkinen, A., Palmroth, M., Reeves, E. G. D., & Singer, H. J. (2002). April 2000 magnetic storm: Solar wind driver and magnetospheric response. *Journal of Geophysical Research*, 107(A12), SMP15-1–SMP15-21. <https://doi.org/10.1029/2001JA009154>
- Iong, D., Chen, Y., Toth, G., Zou, S., Pulkkinen, T., Ren, J., et al. (2022). New findings from explainable SYM-H forecasting using gradient boosting machines. *Space Weather*, 20(8), e2021SW002928. <https://doi.org/10.1029/2021SW002928>
- Kappenman, J. G. (2003). Storm sudden commencement events and the associated geomagnetically induced current risks to ground-based systems at low-latitude and midlatitude locations. *Space Weather*, 1(3), 1016. <https://doi.org/10.1029/2003SW000009>
- Keesee, A. M., Pinto, V., Coughlan, M., Lennox, C., Mahmud, M. S., & Connor, H. K. (2020). Comparison of deep learning techniques to model connections between solar wind and ground magnetic perturbations. *Frontiers in Astronomy and Space Sciences*, 7, 550874. <https://doi.org/10.3389/fspas.2020.550874>
- LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., et al. (1990). Handwritten digit recognition with a back-propagation network. In D. Touretzky (Ed.), *Advances in neural information processing systems* (Vol. 2). Morgan-Kaufmann. Retrieved from <https://proceedings.neurips.cc/paper/1989/file/53c3bce66e43be4f209556518c2fc854-Paper.pdf>
- Liemohn, M. W., Shane, A. D., Azari, A. R., Petersen, A. K., Swiger, B. M., & Mukhopadhyay, A. (2021). RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics. *Journal of Atmospheric and Solar-Terrestrial Physics*, 218, 105624. <https://doi.org/10.1016/j.jastp.2021.105624>
- Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint*. arXiv:1705.07874[cs, stat]. <https://doi.org/10.48550/arXiv.1705.07874>
- Ngwira, C. M., Pulkkinen, A. A., Bernabeu, E., Eichner, J., Viljanen, A., & Crowley, G. (2015). Characteristics of extreme geoelectric fields and their possible causes: Localized peak enhancements. *Geophysical Research Letters*, 42(17), 6916–6921. <https://doi.org/10.1002/2015GL065061>
- Ngwira, C. M., Sibeck, D., Silveira, M. V. D., Georgiou, M., Weygand, J. M., Nishimura, Y., & Hampton, D. (2018). A study of intense local dB/dt variations during two geomagnetic storms. *Space Weather*, 16(6), 676–693. <https://doi.org/10.1029/2018SW001911>
- Oliveira, D. M., & Ngwira, C. M. (2017). Geomagnetically induced currents: Principles. *Brazilian Journal of Physics*, 47(5), 552–560. <https://doi.org/10.1007/s13538-017-0523-y>
- Papitashvili, N., & King, J. (2020). Omni 1-min data [Dataset]. NASA Space. <https://doi.org/10.48322/1shr-ht18>
- Pinto, V. A., Keesee, A. M., Coughlan, M., Mukundan, R., Johnson, J. W., Ngwira, C. M., & Connor, H. K. (2022). Revisiting the ground magnetic field perturbations challenge: A machine learning perspective. *Frontiers in Astronomy and Space Sciences*, 9, 869740. <https://doi.org/10.3389/fspas.2022.869740>
- Pirjola, R. (2000). Geomagnetically induced currents during magnetic storms. *IEEE Transactions on Plasma Science*, 28(6), 1867–1873. <https://doi.org/10.1109/27.902215>
- Pulkkinen, A., Bernabeu, E., Eichner, J., Viljanen, A., & Ngwira, C. M. (2015). Regional-scale high-latitude extreme geoelectric fields pertaining to geomagnetically induced currents. *Earth Planets and Space*, 67(1), 93. <https://doi.org/10.1186/s40623-015-0255-6>
- Pulkkinen, A., Bernabeu, E., Thomson, A., Viljanen, A., Pirjola, R., Boteler, D., et al. (2017). Geomagnetically induced currents: Science, engineering, and applications readiness: GIC overview. *Space Weather*, 15(7), 828–856. <https://doi.org/10.1002/2016SW001501>
- Pulkkinen, A., Rastätter, L., Kuznetsova, M., Singer, H., Balch, C., Weimer, D., et al. (2013). Community-wide validation of geospace model ground magnetic field perturbation predictions to support model transition to operations. *Space Weather*, 11(6), 369–385. <https://doi.org/10.1002/swe.20056>
- Reddy, S., Forsyth, C., Aruliah, A., Kataria, D., Smith, A., Bortnik, J., et al. (2022). Predicting swarm equatorial plasma bubbles via supervised machine learning. *arXiv preprint*. (arXiv:2209.13482[astro-ph, physics:physics]). Retrieved from <http://arxiv.org/abs/2209.13482>
- Rogers, N. C., Wild, J. A., Eastoe, E. F., Gjerloev, J. W., & Thomson, A. W. P. (2020). A global climatological model of extreme geomagnetic field fluctuations. *Journal of Space Weather and Space Climate*, 10, 5. <https://doi.org/10.1051/swsc/2020008>
- Rosenqvist, L., & Hall, J. O. (2019). Regional 3-D modeling and verification of geomagnetically induced currents in Sweden. *Space Weather*, 17(1), 27–36. <https://doi.org/10.1029/2018SW002084>
- Russell, C., & McPherron, R. (1973). The magnetotail and substorms. *Space Science Reviews*, 15(2–3), 205–266. <https://doi.org/10.1007/bf00169321>
- Shapley, L. (1953). A value for n-person games. In *Contributions to the theory of games* (Vol. 2, pp. 307–317).
- Stigliano, F., Consolini, G., Tozzi, R., Gentili, M., Giannattasio, F., & De Michelis, P. (2021). Forecasting SYM-H index: A comparison between long short-term memory and convolutional neural networks. *Space Weather*, 19(2), e2020SW002589. <https://doi.org/10.1029/2020SW002589>
- Smith, A. W., Forsyth, C., Rae, I. J., Garton, T. M., Bloch, T., Jackman, C. M., & Bakrania, M. (2021). Forecasting the probability of large rates of change of the geomagnetic field in the UK: Timescales, horizons, and thresholds. *Space Weather*, 19(9), e2021SW002788. <https://doi.org/10.1029/2021SW002788>
- Smith, A. W., Forsyth, C., Rae, I. J., Garton, T. M., Jackman, C. M., Bakrania, M., et al. (2022). On the considerations of using near real time data for space weather hazard forecasting. *Space Weather*, 20(7), e2022SW003098. <https://doi.org/10.1029/2022SW003098>
- Stone, E. C., Frandsen, A. M., Mewaldt, R. A., Christian, E. R., Margolies, D., Ormes, J. F., & Snow, F. (1998). The advanced composition explorer. *Space Science Reviews*, 86(1/4), 1–22. <https://doi.org/10.1023/a:1005082526237>
- Tanskanen, E. I., Viljanen, A., Pulkkinen, T. I., Pirjola, R., Häkkinen, L., Pulkkinen, A., & Amm, O. (2001). At substorm onset, 40% of AL comes from underground. *Journal of Geophysical Research*, 106(A7), 13119–13134. <https://doi.org/10.1029/2000JA000135>
- Thomson, A. W. P., Dawson, E. B., & Reay, S. J. (2011). Quantifying extreme behavior in geomagnetic activity. *Space Weather*, 9(10), e2020SW002450. <https://doi.org/10.1029/2011SW000696>
- Tóth, G., Meng, X., Gombosi, T. I., & Rastätter, L. (2014). Predicting the time derivative of local magnetic perturbations. *Journal of Geophysical Research: Space Physics*, 119(1), 310–321. <https://doi.org/10.1002/2013JA019456>
- Viljanen, A. (1997). The relation between geomagnetic variations and their time derivatives and implications for estimation of induction risks. *Geophysical Research Letters*, 24(6), 631–634. <https://doi.org/10.1029/97GL00538>
- Viljanen, A., Wintoft, P., & Wik, M. (2015). Regional estimation of geomagnetically induced currents based on the local magnetic or electric field. *Journal of Space Weather and Space Climate*, 5, A24. <https://doi.org/10.1051/swsc/2015022>
- Welling, D. T., Ngwira, C. M., Opgenoorth, H., Haidueck, J. D., Savani, N. P., Morley, S. K., et al. (2018). Recommendations for next-generation ground magnetic perturbation validation. *Space Weather*, 16(12), 1912–1920. <https://doi.org/10.1029/2018SW002064>