

# A gray-box model for a probabilistic estimate of regional ground magnetic perturbations: Enhancing the NOAA operational Geospace model with machine learning

E. Camporeale<sup>1,2</sup>, M. D. Cash<sup>3</sup>, H. J. Singer<sup>3</sup>, C. C. Balch<sup>3</sup>, Z. Huang<sup>4</sup>, G. Toth<sup>4</sup>

<sup>1</sup>CIRES, University of Colorado, Boulder, CO, USA

<sup>2</sup>Center for Mathematics and Computer Science (CWI), Amsterdam, Netherlands

<sup>3</sup>NOAA, Space Weather Prediction Center, Boulder, CO 80305

<sup>4</sup>Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI, USA

## Key Points:

- We present a new model to forecast the maximum value of  $dB/dt$  over 20-minute intervals at specific locations
- The model enhances the output of the physics-based Geospace model with a machine learning technique
- The model provides a probabilistic forecast of exceeding a pre-defined threshold at a given location

## Abstract

We present a novel algorithm that predicts the probability that time derivative of the horizontal component of the ground magnetic field  $dB/dt$  exceeds a specified threshold at a given location. This quantity provides important information that is physically relevant to Geomagnetically Induced Currents (GIC), which are electric currents induced by sudden changes in the Earth’s magnetic field due to Space Weather events. The model follows a ‘gray-box’ approach by combining the output of a physics-based model with a machine learning approach. Specifically, we use the University of Michigan’s Geospace model that is operational at the NOAA Space Weather Prediction Center, with a boosted ensemble of classification trees. We discuss in detail the issue of combining a large dataset of ground-based measurements ( $\sim 20$  years) with a limited set of simulation runs ( $\sim 2$  years) by developing a surrogate model for the years in which simulation runs are not available. We also discuss the problem of re-calibrating the output of the decision tree to obtain reliable probabilities. The performance of the model is assessed by typical metrics for probabilistic forecasts: Probability of Detection and False Detection, True Skill Score, Heidke Skill Score, and Receiver Operating Characteristic curve.

## 1 Introduction

Geomagnetically induced currents (GIC) represent one of the most severe risks posed by space weather events on our infrastructure on the ground, such as high-voltage power transmission systems. GICs are caused by sudden variations of the Earth’s magnetic field that, through Faraday’s law, induce a variation of the electric field, that in turn causes electric currents to be induced in long conductors, like electric power lines or gas pipelines (Boteler et al., 1998; Pirjola et al., 2000; Lanzerotti, 2001; Pulkkinen et al., 2005; Pirjola, 2007; Schrijver & Mitchell, 2013). The induced electric fields responsible for GICs can be estimated from the amplitude of the time derivative of magnetic fluctuations, often denoted as  $dB/dt$ , when combined with information of local earth conductivity characteristics (Boteler & Pirjola, 1998; Pirjola, 2002; Viljanen et al., 2004; Ngwira et al., 2008; Horton et al., 2012). Hence, much attention has been dedicated to understanding and forecasting  $dB/dt$  (Viljanen, 1997; Viljanen et al., 2001).

Previous works on forecasting  $dB/dt$  can generally be divided into empirical and physics-based models. Empirical models exploit the statistical relationships between input quantities, such as solar wind observations recorded by satellites orbiting around the L1 (first Lagrangian points) point, and the observed  $dB/dt$  at a specific station, with a typical time-lag ranging between 15 and 60 minutes. Those statistical relationship can then be encoded into a regression model, in the form of a neural network, or a linear filter. Empirical models include Gleisner and Lundstedt (2001); Weigel et al. (2002, 2003); Wintoft (2005); Wintoft et al. (2005); Weimer (2013); Wintoft et al. (2015); Lotz and Cilliers (2015).

On the other hand, physics-based models follow the evolution in time and space of the plasma and the electromagnetic field surrounding Earth and derive the ground magnetic field perturbation from physical laws. Typically the spatial domain is divided in sub-regions, where the MHD approximation is used in the outer magnetosphere, while the inner magnetosphere and the transition to ionosphere are modeled by including kinetic processes. Examples of physics-based models that can, in principle, forecast  $dB/dt$  given the conditions of the solar wind observed at L1 are OpenGGCM (Open General Geospace Circulation Model) (Raeder et al., 1998), GAMERA (Grid Agnostic MHD for Extended Research Applications) (Zhang et al., 2019), and SWMF (Space Weather Modeling Framework) (Tóth et al., 2005). Several works have assessed the ability of physics-based models to forecast geomagnetic perturbations and more generally to recover plasma and field conditions as observed in the data (see, e.g. (Yu & Ridley, 2008; Welling & Ridley, 2010; Pulkkinen et al., 2011; Rastätter et al., 2011,

2013; Gordeev et al., 2015; Jordanova et al., 2018; Welling, 2019). The validation and comparisons of different models for predicting  $dB/dt$  was specifically tackled in Pulkkinen et al. (2013) in order to support selecting a model to transition to operations at NOAA’s Space Weather Prediction Center (SWPC). As a result of that comparison, the University of Michigan’s SWMF model, henceforth referred to as the Geospace model, was selected for transition to real-time operations.

In this paper we present a new model for predicting whether  $dB/dt$  will exceed given thresholds in a given time interval at specific locations. The model builds on the physics-based Geospace model. We show that the skill of the physics-based model can be considerably enhanced with a machine learning technique, improving all the performance metrics considered.

### 1.1 The Geospace model at NOAA/SWPC

The Geospace model that runs operationally at NOAA/SWPC is a version of the Space Weather Modeling Framework developed by the University of Michigan (Tóth et al., 2005, 2012), that couples the following three modules. A global MHD model solved by BATSRUS (BlockAdaptive Tree Solarwind Roetype Upwind Scheme) (Gombosi et al., 2004), an inner magnetosphere model represented by the Rice Convection Model (RCM) (Toffoletto et al., 2003), and an ionosphere electrodynamics module represented by the Ridley Ionosphere Model (RIM) (Ridley et al., 2004). A detailed description of the Geospace model and its modules can be found in Tóth et al. (2005); Pulkkinen et al. (2013); Tóth et al. (2014)

### 1.2 Prediction of $dB/dt$

In defining the problem, we follow the strategy introduced in Pulkkinen et al. (2013) and later adopted in Tóth et al. (2014) and Welling et al. (2017). Specifically, we define

$$dB/dt = \max_{\{t, t+\Delta t\}} \sqrt{(dB_n/dt)^2 + (dB_e/dt)^2} \quad (1)$$

as the maximum value of the time derivative of the horizontal magnetic field, over an interval  $\Delta t$ , where  $n$  and  $e$  denote the north and east components of the magnetic field, respectively. More specifically, we restrict the time interval to  $\Delta t = 20$  minutes, and we cast the problem as a classification task. Namely, our model predicts the probability that  $dB/dt$  will exceed a given threshold at a given location, in a 20-minute interval. Henceforth we simply refer to  $dB/dt$  as defined in Eq. (1).

As a proof-of-concept, we will show results for the following three magnetic stations: Fresno, California (Geomagnetic latitude:  $43.12^\circ\text{N}$ , operated by USGS, code:FRN), Ottawa, Canada (Geomagnetic latitude:  $54.88^\circ\text{N}$ , operated by GSC, code:OTT), Iqaluit, Canada (Geomagnetic Latitude:  $73.25^\circ\text{N}$ , operated by GSC, code: IQA), hence testing our new method for low, mid and high magnetic latitudes, respectively (the reported magnetic coordinates are derived from the International Geomagnetic Reference Field (IGRF) 12th generation (Thébault et al., 2015)). The extension of this method to any other station is straightforward.

The need of combining a physics-based approach with machine learning can be appreciated by analyzing the accuracy of the Geospace model in predicting  $dB/dt$ . Figures 1, 2, 3 refer respectively to the *FRN*, *OTT*, and *IQA* stations. They show the number of instances of a given value  $dB/dt$  observed in the simulation (vertical axis) versus the corresponding value observed in the real data (horizontal axis), both in logarithmic scale. Each column (i.e. a fixed real value) is normalized to its maximum value. The statistics are computed over a two year interval (see below). The solid red line represents a perfect match between predicted and observed values. Figures 1-3 show that the simulations tend to underestimate  $dB/dt$  for large values (particularly at high latitude) and overestimate it for small values (particularly at low latitude).

**Table 1.** Thresholds considered for each station (in nT/s)

Station	50%	75%	85%	95%	99%
FRN	0.01	0.018	0.025	0.049	0.154
OTT	0.025	0.043	0.06	0.11	0.24
IQA	0.16	0.342	0.52	1.05	2.2

Also, the range of observed and predicted values is dependent on the geomagnetic latitude, as expected.

The paper is divided as follows. Section 2 introduces the data used for this study and the corresponding time periods covered. Section 3 describes the methodology, including the machine learning technique, the performance metrics, and the features chosen in the model. Section 4 presents the results of the new model, comparing its performance with the output of the Geospace model alone, and emphasizes the probabilistic nature of the forecast. Finally, in Section 5 we draw conclusions and make final remarks about future directions.

## 2 Data

The magnetic field historical records have been obtained by the International Real-time Magnetic Observatory Network (INTERMAGNET). The one-minute data in IAGA-2002 format were retrieved for the period 2001-01-01 to 2019-05-05 (<ftp://ftp.seismo.nrcan.gc.ca/intermagnet/minute/variation/IAGA2002/>) for the three stations (FRN, OTT, IQA), consisting of about 9.45M valid entries per station. The output of the Geospace model used for this work covers the time period 2017-05-28 to 2019-05-05, about 1,000,000 one-minute output values. The Geospace model outputs the magnetic field at the location of the three stations at one minute resolution. However, the inner boundary of the global MHD model is at 2.5 Earth radii ( $R_E$ ). Therefore, the magnetic perturbations at the magnetometer stations are calculated from the currents using Biot-Savart integrals, taking into account the following three contributions: the currents inside the BATS-R-US domain, the field-aligned currents in the gap region between 1 and 2.5  $R_E$  radial distance, and the Pedersen and Hall currents in the ionosphere electrodynamics model RIM (Tóth et al., 2014).

All the valid 20-minutes intervals (i.e. with complete data) contained in the above described data sets have been considered for this study.

## 3 Methodology

As mentioned in the Introduction, the goal of this work is not to predict the precise value of  $dB/dt$  for any given 20 minutes interval, but rather to estimate the probability that a pre-defined threshold will be exceeded. Hence, the first task is to define such thresholds. In this paper, we slightly deviate from Pulkkinen et al. (2013), which focused on the following four thresholds: (0.3, 0.7, 1.1, 1.5) nT/s, independent of the station considered. Instead, we define thresholds specific for each location, by analyzing the overall distribution of  $dB/dt$  observed in the INTERMAGNET data ( $\sim 19$  years of data) and choosing the following percentiles as thresholds: (50%, 75%, 85%, 95%, 99%). The resulting thresholds are summarized in Table 1.

### 3.1 Metrics

The task under consideration is a probabilistic classification: for a given station the model outputs the probability that  $dB/dt$  will exceed a specified threshold value. Such a probabilistic outcome can be interpreted as a deterministic binary prediction (i.e. positive/negative) by simply assigning ‘positive’ to all predictions above a certain probability, and ‘negative’ otherwise. Once the probabilistic outcome is interpreted as a binary prediction, one can calculate the following quantities, defined over a certain number of predictions:

- $P$  = total number of observed positives (event occurrences);
- $N$  = total number of observed negatives (event non-occurrences);
- $TP$  = True Positives: number of predicted positives that are observed positives;
- $FP$  = False Positives: number of predicted positives that are observed negatives;
- $TN$  = True Negatives: number of predicted negatives that are observed negatives;
- $FN$  = False Negatives: number of predicted negatives that are observed positives;

and the following performance metrics:

- $TPR = TP/P$  = True Positive Rate (also called Probability of Detection, Sensitivity, Hit Rate);
- $FPR = FP/N$  = False Positive Rate (also called Probability of False Detection, False Alarm Rate);
- $TSS = TPR - FPR$  = True Skill Score.
- $HSS = 2(TP \cdot TN - FN \cdot FP) / (P(FN + TN) + N(TP + FP))$  = Heidke Skill Score

The  $TPR$  measures the ability to find all positive events and a perfect classifier results in  $TPR = 1$ ; the  $FPR$  measures the probability of wrongly classifying a negative as a positive, and a perfect classifier results in  $FPR = 0$ . Hence,  $TSS$  is a useful metric that combines both types of information and should be as close as possible to 1. Moreover, in a Receiver Operating Characteristic (ROC) curve,  $TPR$  and  $FPR$  are respectively on the vertical and horizontal axis, and  $TSS$  measures the distance to the diagonal (no-skill) line (Krzanowski & Hand, 2009). Finally, the  $HSS$  measures the skill of a method compared to a baseline represented by random chance.  $HSS$  has been used in Pulkkinen et al. (2013) and is used here for comparison with previous studies.

The baseline accuracy is represented by the True Skill Score and the Heidke Skill Score yielded by the Geospace model alone, that is by calculating  $dB/dt$  directly from the simulation output and comparing to observations. Figures 4 and 5 show the  $TSS$  and  $HSS$ , respectively, where blue, red, and yellow lines are for  $FRN$ ,  $OTT$ , and  $IQA$  stations, and the scores are shown for different thresholds, on the horizontal axis (the circles denote the scores obtained using the thresholds in Table 1, and a straight line is drawn to connect the circles). One can notice that both scores are latitude dependent.  $TSS$  never exceeds 0.45, getting as low as  $\sim 0.02$  for  $IQA$ , while  $HSS$  ranges approximately between 0.25 and 0.03.

### 3.2 Machine Learning classifier

A variety of methods exist in the Machine Learning arena to perform a probabilistic classification task. For this work, we have opted to use a boosted ensemble of classification trees. The method of choice is called RobustBoost (Freund, 2009). In this section we provide a short introduction and appropriate references.

Let us assume we want to assign a label  $y = \{0, 1\}$  to a data point  $\mathbf{x} = \{x_1, x_2, \dots, x_D\}$ , where  $D$  is the dimensionality of  $\mathbf{x}$ . The task is a supervised binary classification, meaning that we make use of a large dataset of labeled examples to infer a pattern between the inputs  $\mathbf{x}$  and the binary outputs  $y$ , that hopefully can be used to infer the label of new data points that have not been used to train the model. A decision tree is a simple method that recursively partitions the  $D$ -dimensional hyper-space of input variables one dimension at the time, thus creating a tree-like structure. In other words, by taking as decision boundaries hyperplanes defined by simple inequalities such as  $x_i < c$ , a decision tree divides the input space into a number of hypercubes where a given label is assigned to all the data belonging to the same hypercube. Decision trees have the great advantage of being very transparent and easily interpretable. In fact, one can simply follow the tree structure from top to bottom to understand how a label is associated to a given data point. In order to choose where to set up a decision boundary (i.e. the value of the constant  $c$ ) and along which variable, a partition criterion is followed. Two of the most popular partition criteria are the Gini Index, and the Information Gain. The Gini Index measures the reduction in class impurity, which is defined as the probability that two randomly chosen data that belong to the same partition have different labels. At a given iteration when growing a tree, the best partition is the one that reduces such impurity, or in other words that minimizes the probability of a data point being mislabeled. The Information Gain is based on the entropy measured at each node and the optimal split is the one that minimizes the global entropy (or maximizes information). A standard monograph on decision trees is Breiman (2017). The accuracy of a classification tree can be improved by using a boosting strategy. Boosting refers to a class of algorithms that makes use of an ensemble of (not very accurate) predictions to produce a much more accurate one. The members of the ensemble are called weak learners and their weighted sum is referred to as strong learner. In the context of classification trees, the weak learners are represented by trees that are grown to only a few layers. One of the most successful and widely applied boosting techniques is *Adaboost* (short for Adaptive Boosting), introduced in the seminal paper by Freund and Schapire (1997). This is an algorithm that iteratively adds members to an existing ensemble. The newest member increasingly focuses on the data points that were misclassified by the previous members, and the weights of each member are iteratively adjusted. AdaBoost is typically less prone to overfitting than other algorithms, but it is very sensitive to outliers, because it will keep focusing on the few data points that are mis-classified, eventually at the expense of the remaining dataset. A modification of Adaboost that adds robustness to the algorithm (in the sense of not being so sensitive to outliers) is *RobustBoost*, being introduced in Freund (2009). As explained in the original paper, RobustBoost can be intuitively understood as “giving up” on data points that are so far on the incorrect side of a decision boundary that they are unlikely to be correctly classified even after many iterations.

In this work we have used the MATLAB (R2019a) implementation of RobustBoost which is included in the Statistics and Machine Learning Toolbox. We have tested and compared the following boosting techniques: AdaBoost, GentleBoost, LogitBoost, RobustBoost, and Bagging, and although their results were comparable, RobustBoost is the algorithm that consistently yielded better results.

### 3.3 Features

In the machine learning jargon a feature is an explanatory variable that is used as an input for a given algorithm. The present work builds up on the idea presented in Tóth et al. (2014). The main finding was that a strong correlation exists between observed  $dB/dt$  and the observed maximum variation in the amplitude of the magnetic field, within the same 20-minute interval. In fact, the correlation is almost linear when both quantities are expressed in their logarithm. Tóth et al. (2014) argued that the

values of  $\log_{10}(\max(B) - \min(B))$  (where  $\max$  and  $\min$  are intended within a 20-minute interval) obtained by the Geospace model simulations are much more reliable than the values of  $dB/dt$  and hence the latter can be inferred from the former. We expand this idea using 6 features as input variables to our model. The 6 features are the following:

1.  $\log_{10}(\max(B) - \min(B))$  (as in the original work (Tóth et al., 2014), obtained from the Geospace simulation output)
2.  $\log_{10}(\max(B_x) - \min(B_x))$  (same as 1. but considering the  $x$ -component only, obtained from the Geospace simulation output)
3.  $\log_{10}(\max(B_y) - \min(B_y))$  (same as 1. but considering the  $y$ -component only, obtained from the Geospace simulation output)
4.  $\log_{10}(\max(B) - \min(B))$  - 1hr before (same as 1. but calculated in the time window 1 hour preceding the target window, obtained from magnetometer data)
5.  $\log_{10}(\max(dB/dt))$  - 1hr before (the same quantity as the target, but calculated in the time window 1 hour preceding the target window, obtained from magnetometer data)
6.  $\log_{10}(|V_x|)$  - 1hr before (the  $x$ -component of the solar wind speed, measured one hour before the target window, obtained from the OMNI dataset)

These six quantities are all strongly correlated with the target  $dB/dt$ . Figures 6, 7, 8 show such correlations, with the same normalization described for previous figures, respectively for the FRN, OTT, and IQA stations.

### 3.4 Constructing surrogate data

A major problem with using the output of the Geospace model simulations as features of a machine learning algorithm (i.e. Features 1-3 in Sec. 3.3) is the relative scarcity of data and the fact that a variety of physical conditions may not be well represented in such a limited set of model runs. In fact, it is well known that machine learning has a good chance of working well only when the training set is large enough to contain most possible scenarios. As mentioned before, we have worked with a little less than 2 years of data from the archived Geospace output, covering the period from May 2017 to May 2019. This was a time of relatively quiet geomagnetic condition, that cannot possibly be considered general enough to train a robust machine learning algorithm. Moreover, to test how well our method performs, we need to put aside a portion of the data (the so-called test set) so that the performance metrics can be calculated using data that has not been used for training, reducing even more the data at our disposal. On the other hand, we have at our disposal a much larger dataset of magnetometer measurements from the INTERMAGNET archive (more than 9,000,000 points), covering more than 18 years of data. However, to be able to train our model on this large period of time, we need to be able to estimate the output of the Geospace model for this time period. The core idea is to use a generative model that creates surrogate Geospace data for the years in which we do not have simulation runs. In practice, we need to create surrogate data only for Features 1-3, and not for the actual values of magnetic field. We have employed the following simple strategy. Figures 9, 10, 11 (for FRN, OTT and IQA, respectively) show the distribution of data for Features 1-3, where the Geospace simulation data is on horizontal axis, and the corresponding real data (from magnetometers) is on vertical axis. Superimposed to the heat-map we show the contour lines of a 2D Gaussian distribution, with same mean and covariance. One can notice that all of these distributions are approximately Gaussian. The same is true for any 2D distribution of any pairs of Features 1-6. Moreover, graphical inspection of such marginal distributions indicates that they all are uni-modal (i.e. single peaked). Therefore, we assume that the multidimensional data comprising Features 1-6, the same quantities as in Features 1-3 but obtained from magnetometers, and  $dB/dt$  (10 dimensions in total), can be approximated reasonably



well by a multivariate Gaussian distribution. Let us recall that a multivariate Gaussian distribution in the  $D$ -dimensional variable  $\mathbf{x}$  can be written as

$$p(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}, \quad (2)$$

where  $\boldsymbol{\mu}$  is a  $D$ -dimensional mean vector,  $\boldsymbol{\Sigma}$  is a  $D \times D$  covariance matrix, and  $|\boldsymbol{\Sigma}|$  denotes the determinant of  $\boldsymbol{\Sigma}$ . A nice property of multivariate Gaussian distributions is that their conditional distributions (that is, fixing given values for one or more component of  $\mathbf{x}$ ) are also Gaussian and can be derived analytically. Following (Bishop, 2006), suppose that we partition  $\mathbf{x}$  into two disjoint subsets  $\mathbf{x}_a$  and  $\mathbf{x}_b$  so that the first 3 components of  $\mathbf{x}$  are in  $\mathbf{x}_a$  and the remaining 7 in  $\mathbf{x}_b$ . The mean vector and the covariance matrix can be equally decomposed:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad (3)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \quad (4)$$

Then, the conditional distribution  $p(\mathbf{x}_a | \mathbf{x}_b)$  has the following mean and covariance

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab} \boldsymbol{\Sigma}_{bb}^{-1} (\mathbf{x}_b - \boldsymbol{\mu}_b) \quad (5)$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab} \boldsymbol{\Sigma}_{bb}^{-1} \boldsymbol{\Sigma}_{ba} \quad (6)$$

In conclusion, we use the 2017-2019 data (for which we have Geospace outputs) to calculate the 10-dimensional mean vector  $\boldsymbol{\mu}$  and the 10x10 covariance matrix  $\boldsymbol{\Sigma}$ . Then, we use Equations 5 and 6 on each 7-dimensional data point  $\mathbf{x}_b$  in the period 2001-2016 (for which we do not have Geospace output) to construct a 3-dimensional Gaussian distribution of the Features 1-3. Finally, we draw a random sample from this distribution to generate surrogate values of the Features 1-3 for that data point.

## 4 Results

In order to assess the goodness of a trained machine learning model, it is important that the performance metrics are calculated on a portion of a data that has not been used for training (so-called unseen data). Moreover, when dealing with temporal dataset, it is equally important that the training and test sets are temporally disjoint so to minimize the temporal correlations between the two and to ensure that the machine learning algorithm does actually learn some patterns and does not merely memorizes the training data. Here, we have used the period 2001-2016 for training the model and all the performance metrics shown in the following have been evaluated on the period 2017-2019. As a reminder, in the training period no Geospace data is available, and therefore Features 1-3 have been generated with the method described in Section 3.4, while the test period coincides with the time for which archived Geospace simulations are available.

We present the probability of detection (TPR), the Probability of False Detection (FPR), the True Skill Score (TSS), and the Heidke Skill Score (HSS) respectively in Figures 12, 13, 14, 15 for the three stations and as functions of the different threshold levels (see Table 1). The legend is as follows. 'real B/real B' (blue lines) means that a model has been trained using real (INTERMAGNET) data for Features 1-3 and tested also using the real data as input. This gives us the upper bound in the accuracy of the method, that would be achieved if the Geospace model would yield perfect results



for the Features 1-3. On the other hand, one might wonder if the effort of producing surrogate data described in Section 3.4 was worthwhile, or if one could have trained the model by using Features 1-3 as observed on the the real data, without worrying that the Geospace model outputs slightly different results. That is plotted as red lines (real B/sim B), meaning that a model has been trained using real (INTERMAGNET) data and then tested using Geospace outputs as inputs, for Features 1-3. One can see that in this case the accuracy of the predictions strongly degrades. Finally, the model described in this paper is shown with yellow lines (surrogate B/sim B) where the surrogate data described in Section 3.4 have been used for training and the Geospace outputs for testing. In all cases, Features 4-6 (Section 3.3) are available from real data. It is worth pointing out that each point in the previous figures (i.e. for a given station and given threshold) corresponds to an independently trained model. In conclusion, the machine learning algorithm properly corrects the systematic errors of the Geospace model predictions.

#### 4.1 Re-calibration

As anticipated in the introduction, the goal of this work is not to provide a binary classification, but rather to estimate the probability of exceeding pre-defined thresholds. In principle, classification trees can output probabilities, which are simply calculated as the observed ratio between positives and negatives on a given leaf (the final node on a decision tree) calculated over the whole training set. A well-known problem with classification trees is that such probabilities are often mis-calibrated (Niculescu-Mizil & Caruana, 2005). Calibration refers to the consistency between the predicted probability assigned to an event and the actual frequency observed for that event. For instance, in the binary classification setting, if we collect all the instances in which a model predicts a probability  $p$  for a 'positive' outcome (in our case, exceeding a threshold), that model is well-calibrated if on average a positive is actually observed with frequency  $p$  (the frequency being calculated over all those instances). One way to visualize the relationship between predicted probabilities and observed frequency is through a reliability diagram (DeGroot & Fienberg, 1983). To construct such diagram for binary classification, one discretizes the predicted probabilities in bins. For each bin, the average predicted frequency (horizontal axis) is plotted against the true fraction of positive cases in that bin (vertical axis). A perfect calibration will result in a diagonal straight line. Figures 16, 17, 18 show the reliability diagrams for FRN, OTT, and IQA, respectively. Each panel refers to a different threshold (see Table 1), and the blue circles represent the calibration of the boosted ensemble models, as trained by the MATLAB routine. One can clearly see that such predictions are mis-calibrated. We apply a simple calibration strategy, where a mapping between old and new probabilities is derived by simply interpolating linearly the blue circles. For instance, a probability of 40% might be re-calibrated to a new value of 30%. To perform re-calibration fairly, we have derived the reliability diagram and the corresponding calibration map from the training set only. Figures 16, 17, 18 show the reliability diagram calculated over the test set. The red diamonds represent the re-calibrated reliability diagrams, that clearly suggest that all the models have been properly re-calibrated.

#### 4.2 Receiver operating characteristic (ROC) curve

Another important diagnostic for a probabilistic model is the so-called ROC curve. As we mentioned, in order to interpret a probabilistic prediction in terms of true/false positives/negatives (see Sec. 3.1), a probability threshold needs to be used to separate the predicted positives from the negatives. In the limit that such threshold is pushed to 0%, all the predictions become positives, which means that both the true positive rate (TPR) and the false positive rate (FPR) are equal to 1 (all positives are

correctly predicted, but all negative are mis-classified). In the opposite limit, when the threshold is 100% and all predictions are negative both TPR and FPR become equal to 0 (no positives are predicted, but all negatives are correctly predicted). The ROC curve is a continuous curve in the (FPR,TPR) space that connects these extreme scenarios (TPR=FPR=1 and TPR=FPR=0) by gradually changing the threshold from 0% to 100%. Since the optimal prediction is TPR=1 and FPR=0, the optimal threshold is the point on the ROC curve with minimal distance from the upper-left corner. ROC curves for FRN, OTT, and IQA stations are shown in Figures 19, 20, 21, respectively. Different colors denote the five different thresholds, and a filled circle represents the optimal values (that have been used in previous Figures). Note that the True Skill Score (TSS) is the vertical distance between the ROC curve and the diagonal line (TPR=FPR), which represents no skill (i.e. a climatological forecast). The ROC curves demonstrate the general tendency of the models to improve their skill for higher threshold, as already shown in previous Figures. Moreover, it is important to realize that the re-calibration described in the previous Section does not affect the ROC curve. In fact, by mapping old to new probabilities, the points on a given ROC curve get shifted along the same curve. In other words, what changes through re-calibration is the value of the optimal threshold, but not the corresponding values of TPR, FPR, and Skill Scores. In practice, because the un-calibrated models tend to be overconfident (i.e. below the diagonal line in the reliability diagram), re-calibration changes the optimal threshold from 50% to larger values. For instance, it can be that for a given model one needs to interpret as positives predictions with probabilities larger than 80% rather than 50%.

## 5 Conclusions

We have developed a model that estimates the probability of  $dB/dt$  exceeding a given threshold, for three stations ranging from low, to mid and high latitudes (FRN, OTT, and IQA). Five different thresholds were chosen for each station, by calculating the 0.5, 0.75, 0.85, 0.95, 0.99 percentiles on a long-span historic dataset. One of the crucial points of this work is that it combines a physics-based prediction provided by the Geospace model to a machine learning algorithm for binary classification, effectively following what is known as a gray-box approach (Camporeale et al., 2018; Camporeale, 2019). Indeed, we have shown that the Geospace model alone provides limited skills for predicting  $dB/dt$ , although we expect the model to improve over time by better capturing properties of the physical system. However, as already noted in Tóth et al. (2014), the maximum perturbation of the magnetic field within a 20-minutes interval correlates very strongly with  $dB/dt$  and hence it can be used as a predictor in a machine learning algorithm.

Because the archived simulations of the operational Geospace model run at NOAA/SWPC span only a few years, while observations are available for almost 20 years, we have applied a generative model to create surrogate data for the years in which we do not have Geospace runs, by following the simple assumptions that the joint distribution of 10 variables of interest (3 of which needed to be inferred) is Gaussian. In this way, one can analytically derive the 3-dimensional conditional distribution and generate samples from such multivariate Gaussian.

The chosen machine learning algorithm is an ensemble of classification trees, adaptively boosted via RobustBoost (Freund, 2009), and the performance metrics that we have analyzed are the True Positive Rate (TPR, or probability of detection), False Positive Rate (FPR, or probability of false detection), True Skill Score (TSS), and Heidke Skill Score (HSS). Finally, we have discussed the issue of re-calibration and the ROC curve relative to all models.

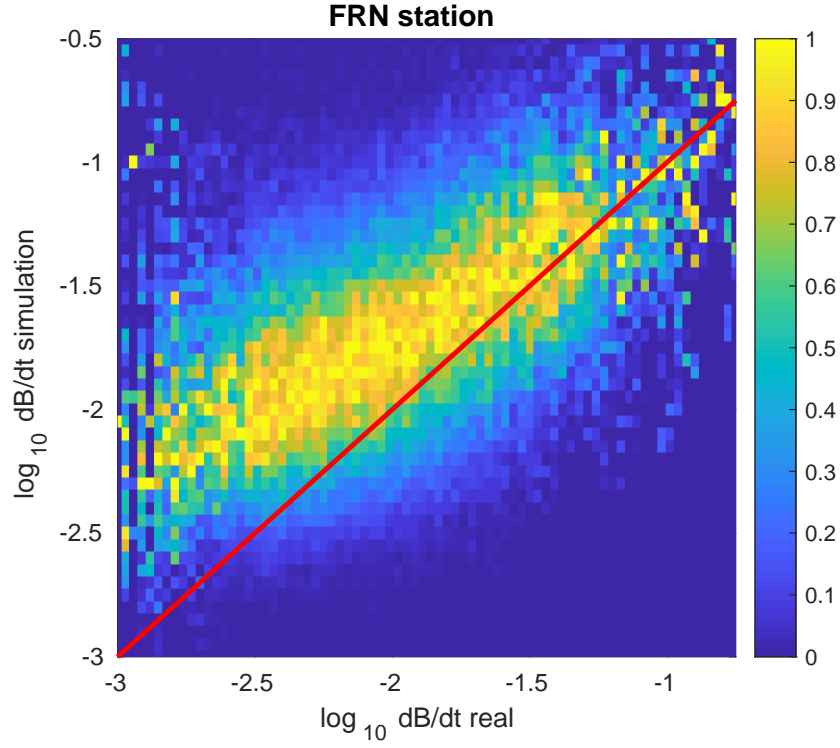
Overall the gray-box approach proposed in this paper consistently enhances the results of the corresponding ‘white box’ approach, where one would directly take the results of the Geospace model as predictors of  $dB/dt$ . Indeed, Figure 22 summarizes

the findings of previous Figures by comparing the True Skill Score (left panel) and the Heidke Skill Score (right panel) of the Geospace model alone (horizontal axis) against the corresponding results applying machine learning (vertical axis). Different symbols are for the three different stations, and the region above the diagonal black solid line denotes an improvement.

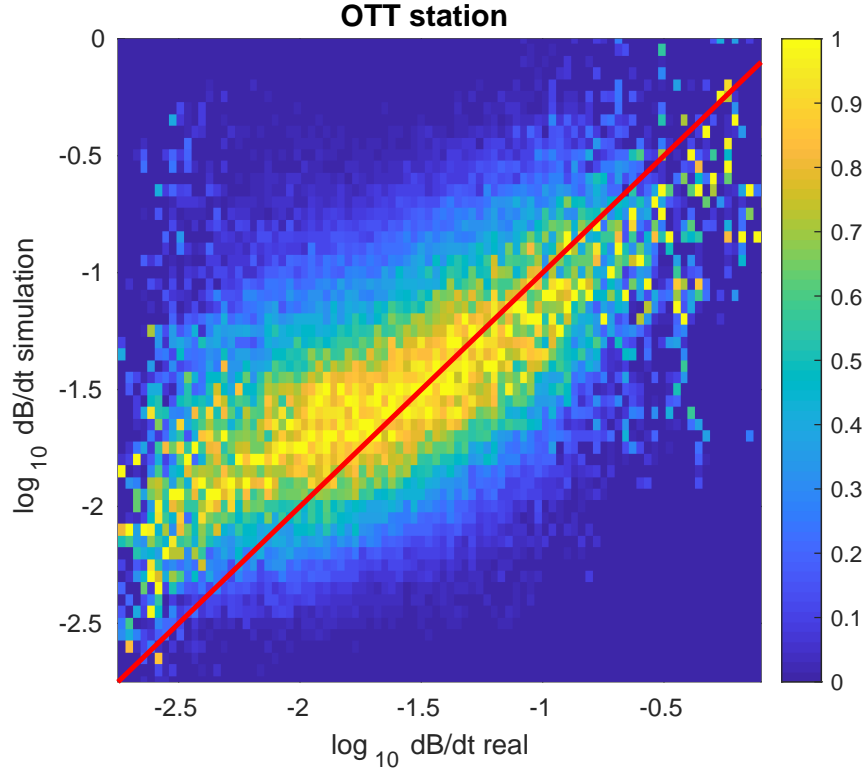
### Acknowledgments

The results presented in this paper rely on the data collected at Fresno (FRN), Ottawa (OTT), and Iqaluit (IQA) geomagnetic observatories. We thank the U.S. Geological Survey and Natural Resources Canada Geomagnetism Programs for supporting their operation and INTERMAGNET for promoting high standards of magnetic observatory practice ([www.intermagnet.org](http://www.intermagnet.org)). The INTERMAGNET data used for this study is publicly available on <ftp://seismo.nrcan.gc.ca/intermagnet/minute/variation/IAGA2002/>.

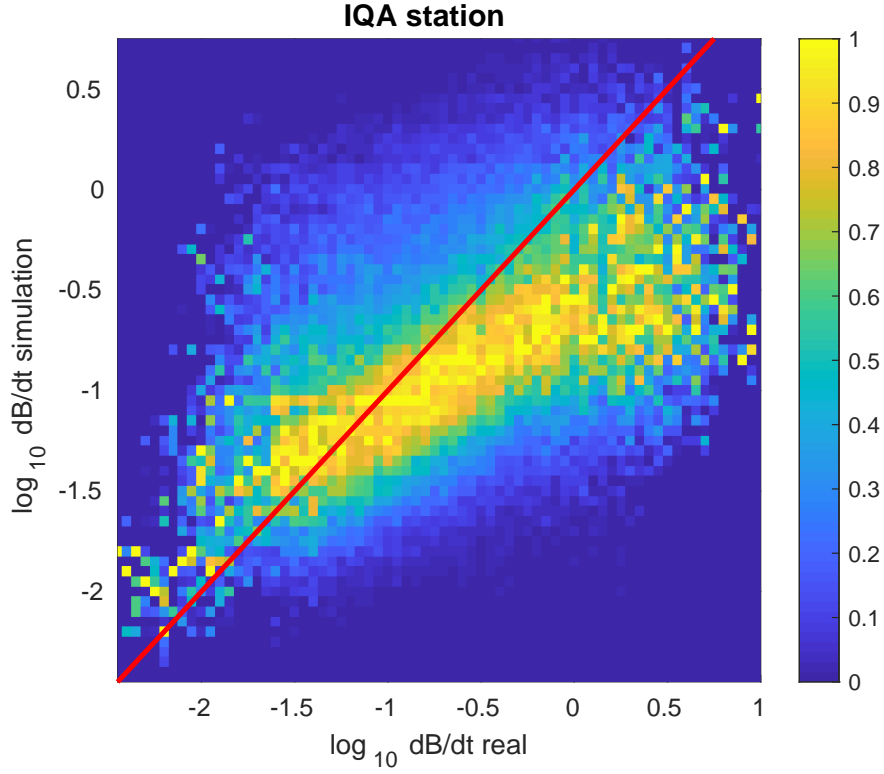
All the data and codes will be made available as a Zenodo/Github repository, after the manuscript is accepted for publication.



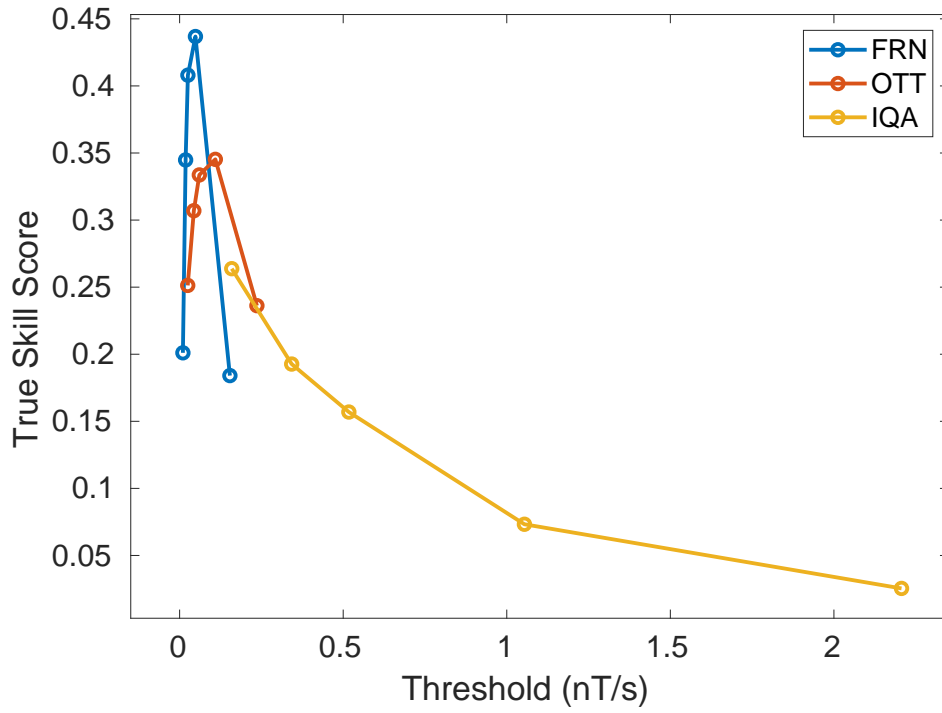
**Figure 1.** FRN station: 2D histogram of the counts of  $dB/dt$  as obtained from the Geospace simulation (vertical axis) vs the corresponding measured values (horizontal axis). Both axes are in logarithmic scale, and the heat-map is normalized column-wise with respect to the maximum value for each column, for better visualization.



**Figure 2.** OTT station: 2D histogram of the counts of  $dB/dt$  as obtained from the Geospace simulation (vertical axis) vs the corresponding measured values (horizontal axis). Both axes are in logarithmic scale, and the heat-map is normalized column-wise with respect to the maximum value for each column, for better visualization.

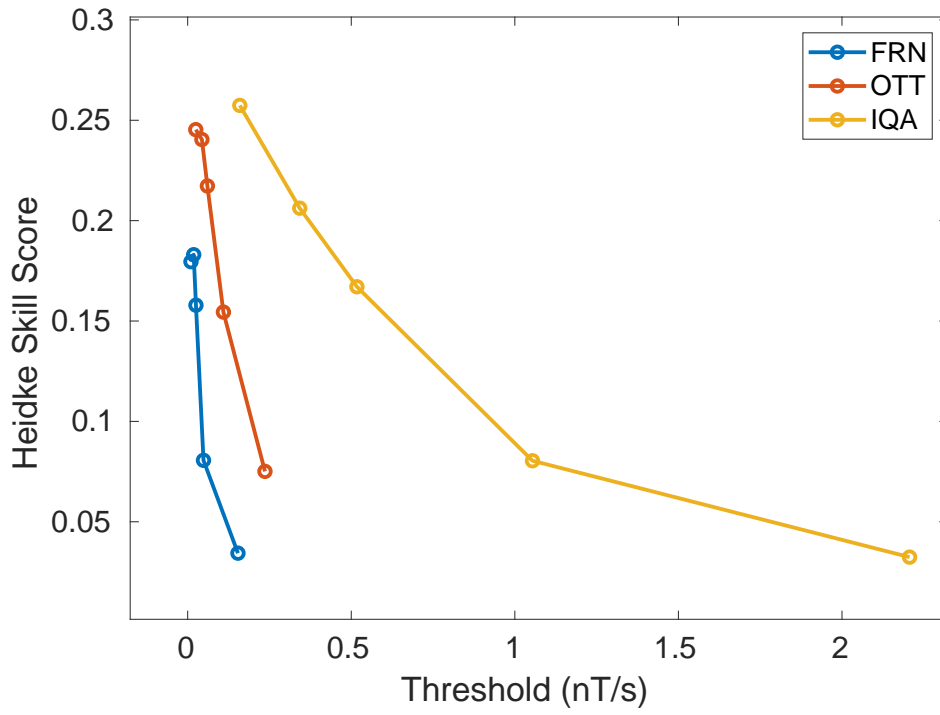


**Figure 3.** IQA station: 2D histogram of the counts of  $dB/dt$  as obtained from the Geospace simulation (vertical axis) vs the corresponding measured values (horizontal axis). Both axes are in logarithmic scale, and the heat-map is normalized column-wise with respect to the maximum value for each column, for better visualization.

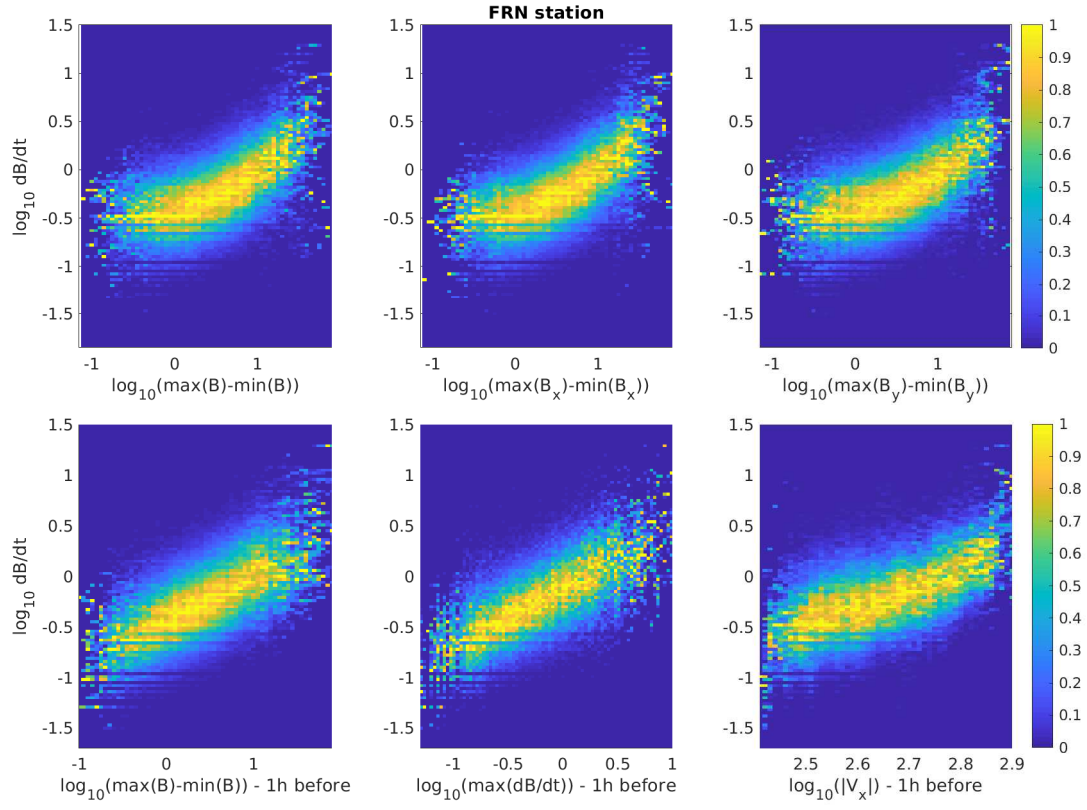


**Figure 4.** True Skill Score obtained from the predictions of the Geospace model, for different stations (in blue for FRN, red for OTT, and yellow for IQA), as functions of the different thresholds (defined in Table 1).

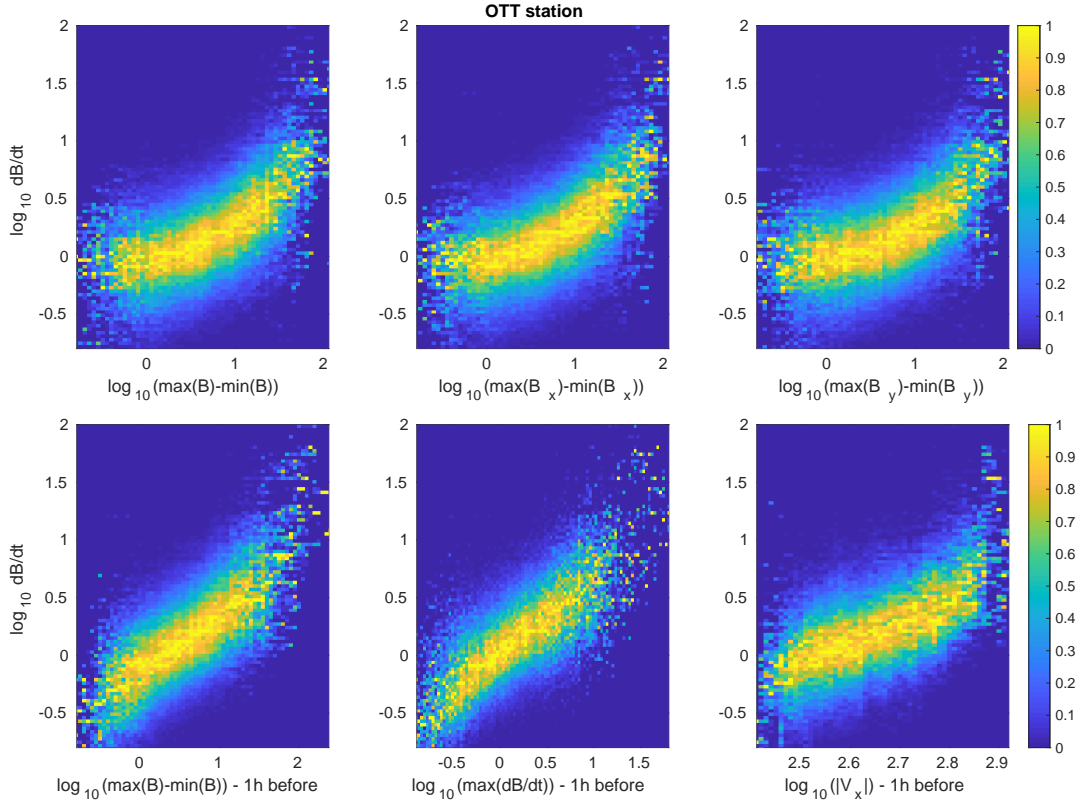




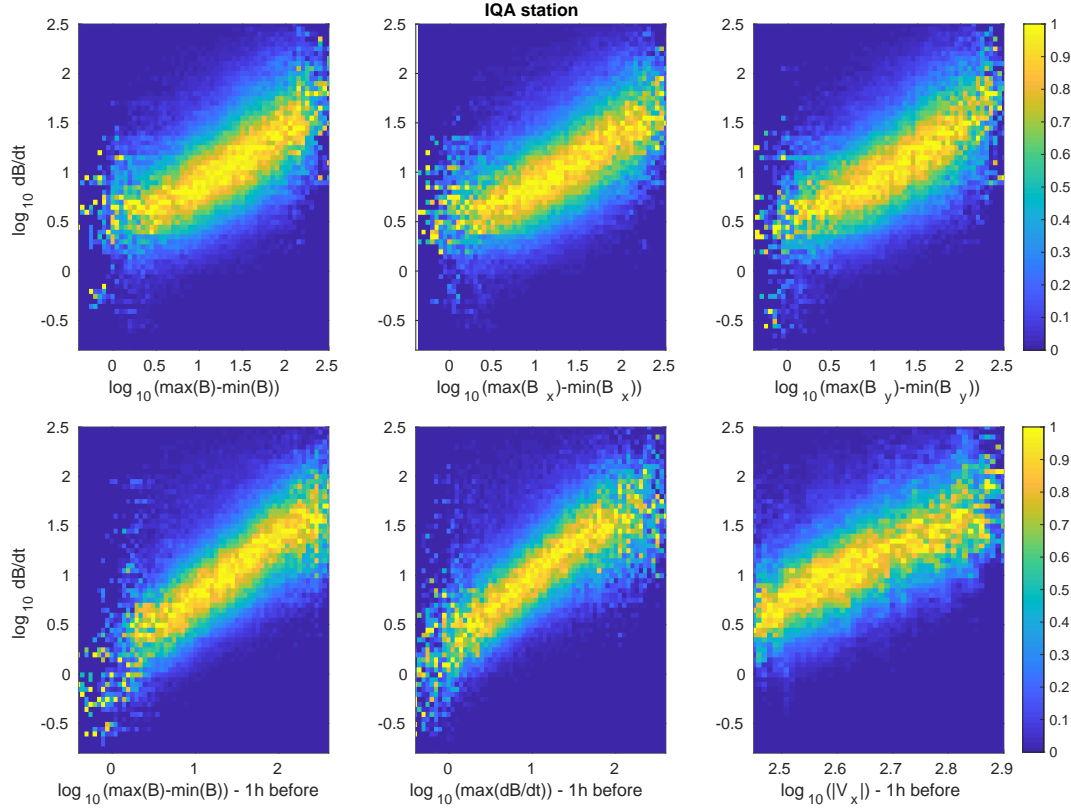
**Figure 5.** Heidke Skill Score obtained from the predictions of the Geospace model, for different stations (in blue for FRN, red for OTT, and yellow for IQA), as functions of the different thresholds (defined in Table 1).



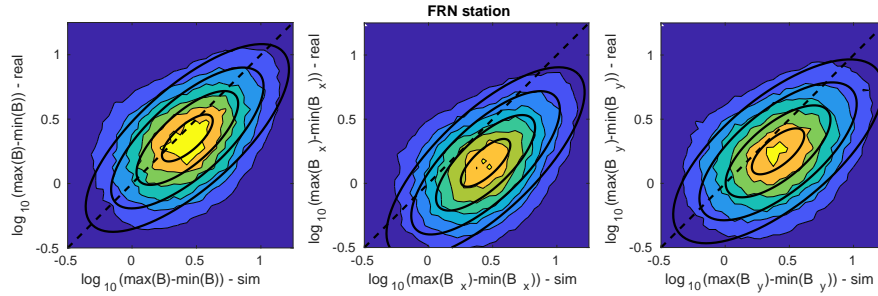
**Figure 6.** FRN station: Correlations between the target variable  $dB/dt$  (vertical axis) and the 6 features described in Sec. 3.3. Each heat-map is normalized column-wise with respect to its maximum value.



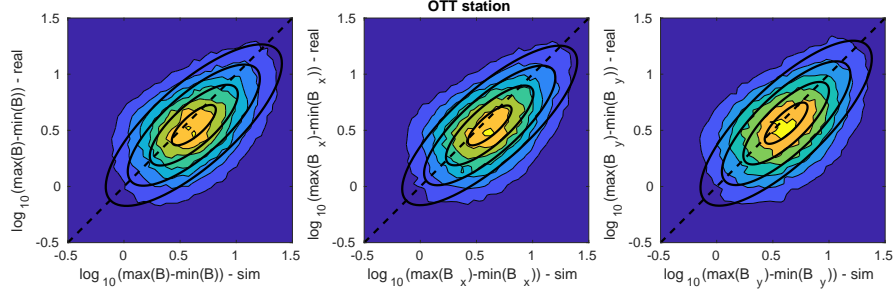
**Figure 7.** OTT station: Correlations between the target variable  $dB/dt$  (vertical axis) and the 6 features described in Sec. 3.3. Each heat-map is normalized column-wise with respect to its maximum value.



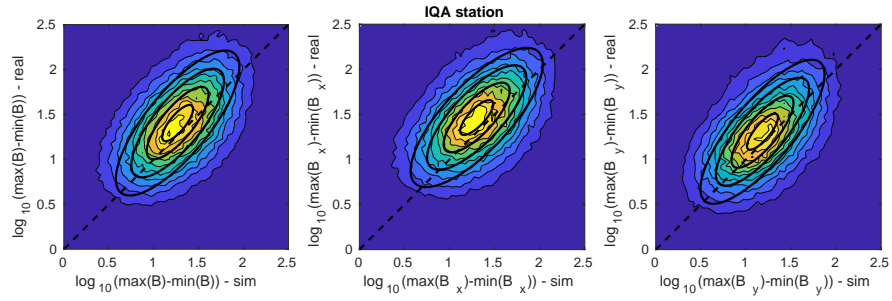
**Figure 8.** IQA station: Correlations between the target variable  $dB/dt$  (vertical axis) and the 6 features described in Sec. 3.3. Each heat-map is normalized column-wise with respect to its maximum value.



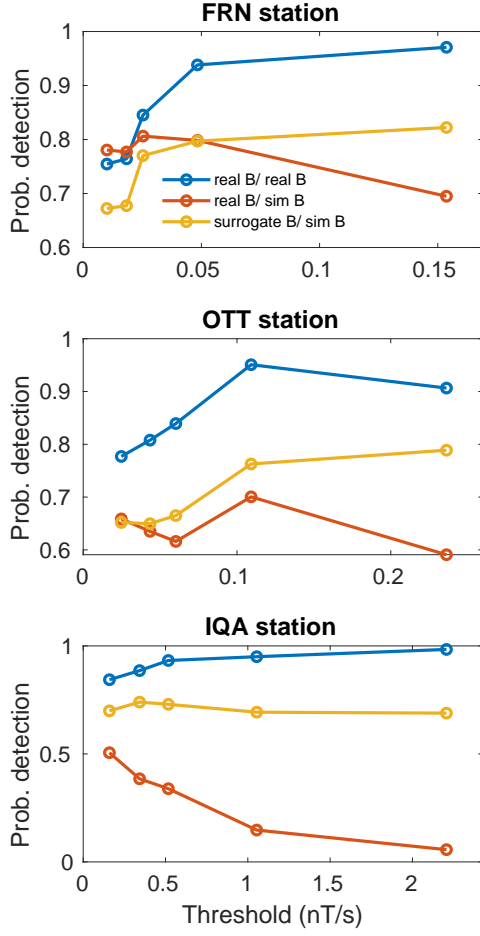
**Figure 9.** FRN station: Correlation between features 1-3 (from left to right panels) as measured from simulation results (horizontal axis) vs real observations (vertical axis). The superposed black lines denote isocontours of the corresponding two-dimensional Gaussian distribution, with same mean and covariance matrix.



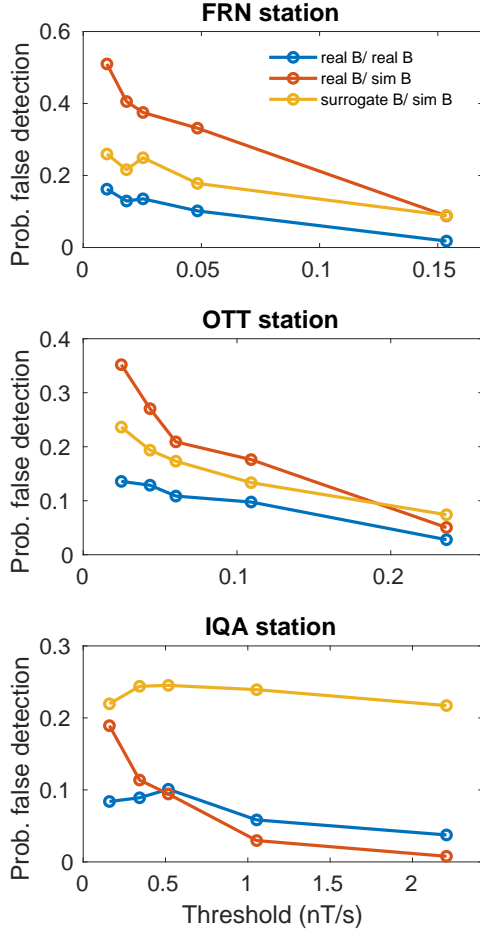
**Figure 10.** OTT station: Correlation between features 1-3 (from left to right panels) as measured from simulation results (horizontal axis) vs real observations (vertical axis). The superposed black lines denote isocontours of the corresponding two-dimensional Gaussian distribution, with same mean and covariance matrix.



**Figure 11.** IQA station: Correlation between features 1-3 (from left to right panels) as measured from simulation results (horizontal axis) vs real observations (vertical axis). The superposed black lines denote isocontours of the corresponding two-dimensional Gaussian distribution, with same mean and covariance matrix.

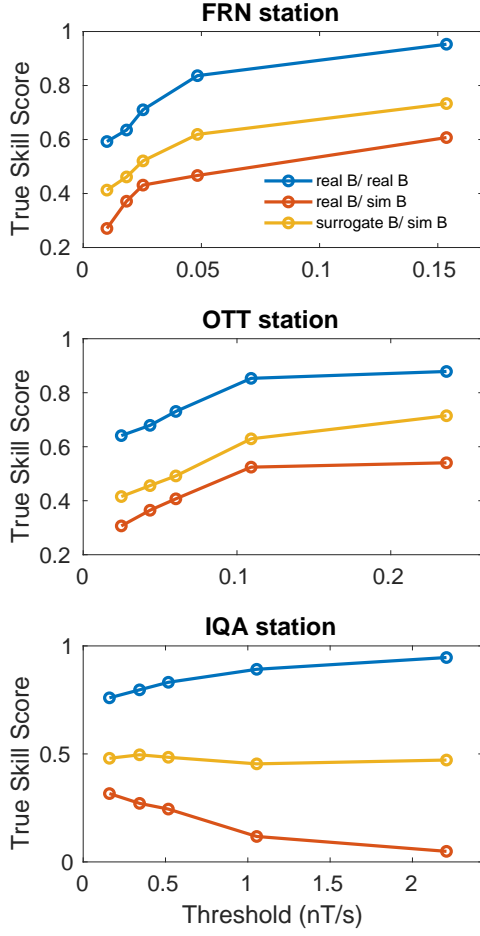


**Figure 12.** Probability of detection (vertical axis) vs different threshold (horizontal axis). Blue, red, and yellow lines denote respectively: a model trained on real data and tested using real data as input ('real B/real B'), a model trained on real data but tested using simulation data as input ('real B/sim B'), a model trained on the surrogate data and tested using simulation data as input ('surrogate B/sim B'). The results for three stations FRN, OTT, and IQA are represented on the top, middle and bottom panels.

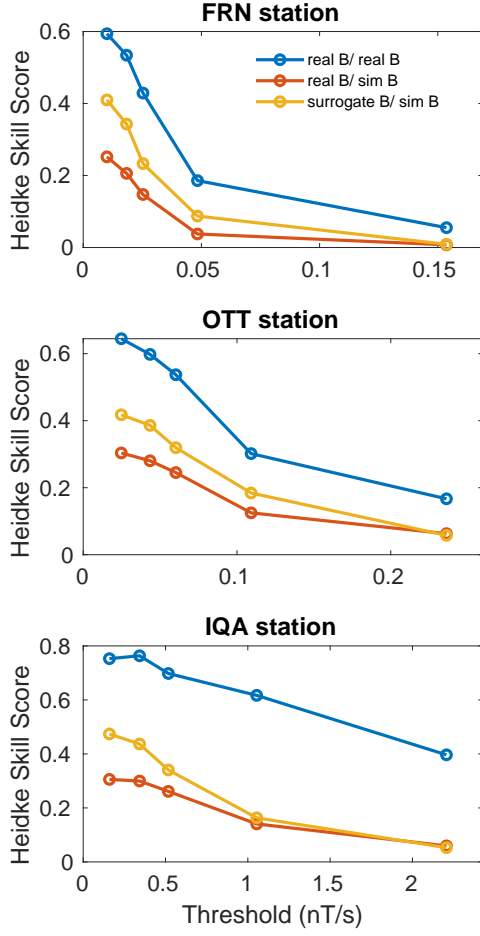


**Figure 13.** Probability of false detection (vertical axis) vs different threshold (horizontal axis). Blue, red, and yellow lines denote respectively: a model trained on real data and tested using real data as input ('real B/real B'), a model trained on real data but tested using simulation data as input ('real B/sim B'), a model trained on the surrogate data and tested using simulation data as input ('surrogate B/sim B'). The results for three stations FRN, OTT, and IQA are represented on the top, middle and bottom panels.

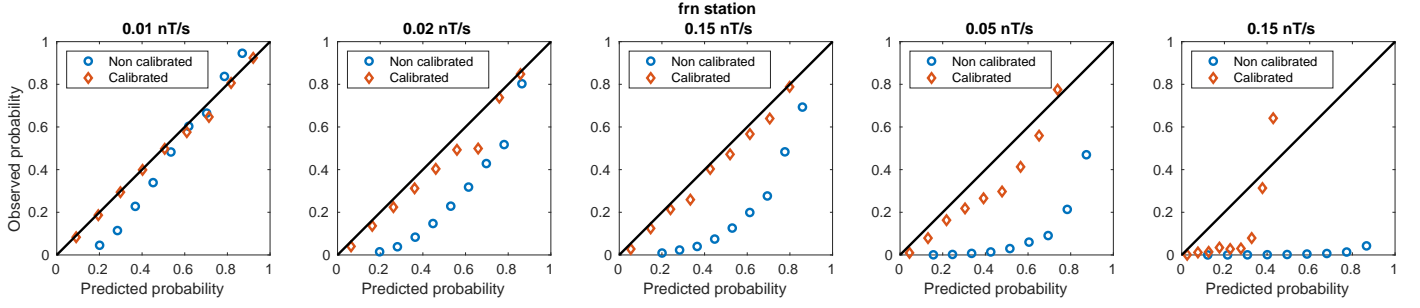




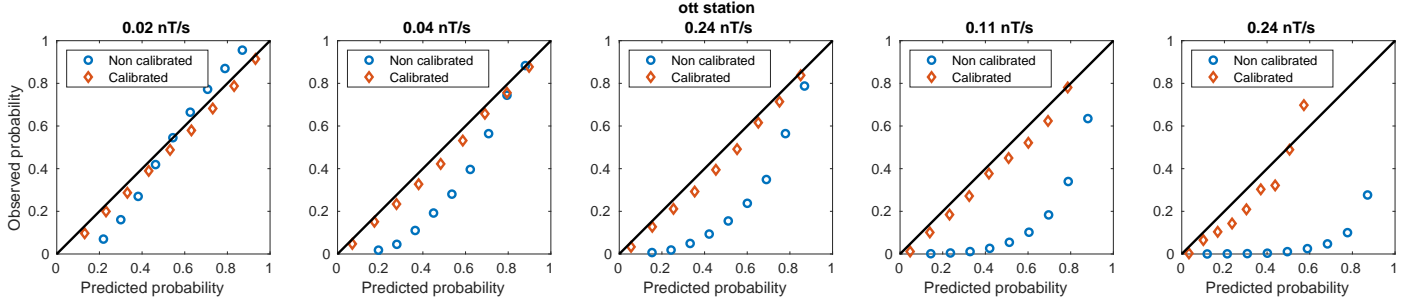
**Figure 14.** True Skill Score (vertical axis) vs different threshold (horizontal axis). Blue, red, and yellow lines denote respectively: a model trained on real data and tested using real data as input ('real B/real B'), a model trained on real data but tested using simulation data as input ('real B/sim B'), a model trained on the surrogate data and tested using simulation data as input ('surrogate B/sim B'). The results for three stations FRN, OTT, and IQA are represented on the top, middle and bottom panels.



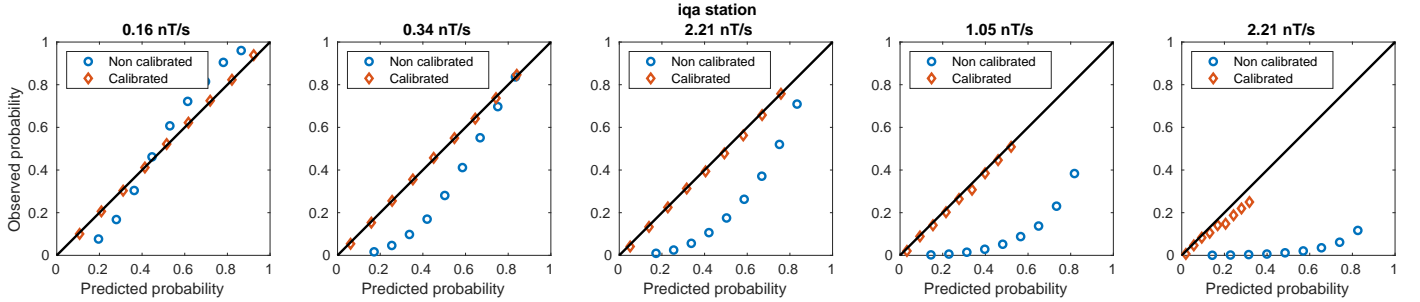
**Figure 15.** Heidke Skill Score (vertical axis) vs different threshold (horizontal axis). Blue, red, and yellow lines denote respectively: a model trained on real data and tested using real data as input ('real B/real B'), a model trained on real data but tested using simulation data as input ('real B/sim B'), a model trained on the surrogate data and tested using simulation data as input ('surrogate B/sim B'). The results for three stations FRN, OTT, and IQA are represented on the top, middle and bottom panels.



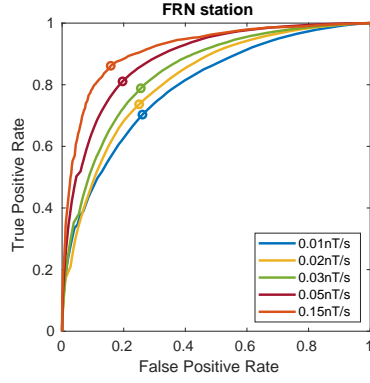
**Figure 16.** FRN station: Reliability diagrams for different thresholds (increasing from left to right panels). Blue circles indicates the result of the non-calibrated models, and the red diamonds indicate the reliability achieved after re-calibration.



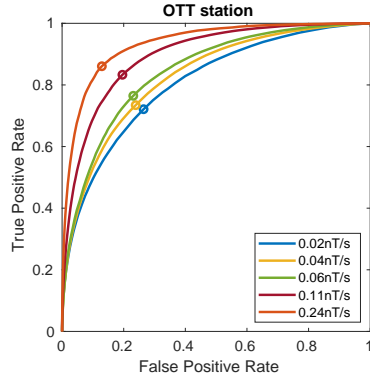
**Figure 17.** OTT station: Reliability diagrams for different thresholds (increasing from left to right panels). Blue circles indicates the result of the non-calibrated models, and the red diamonds indicate the reliability achieved after re-calibration.



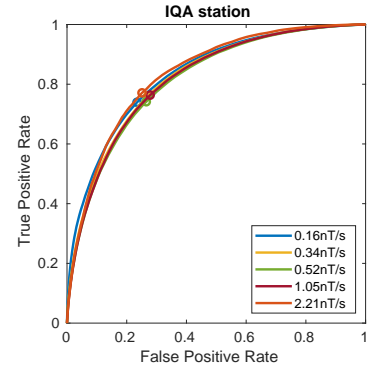
**Figure 18.** IQA station: Reliability diagrams for different thresholds (increasing from left to right panels). Blue circles indicates the result of the non-calibrated models, and the red diamonds indicate the reliability achieved after re-calibration.



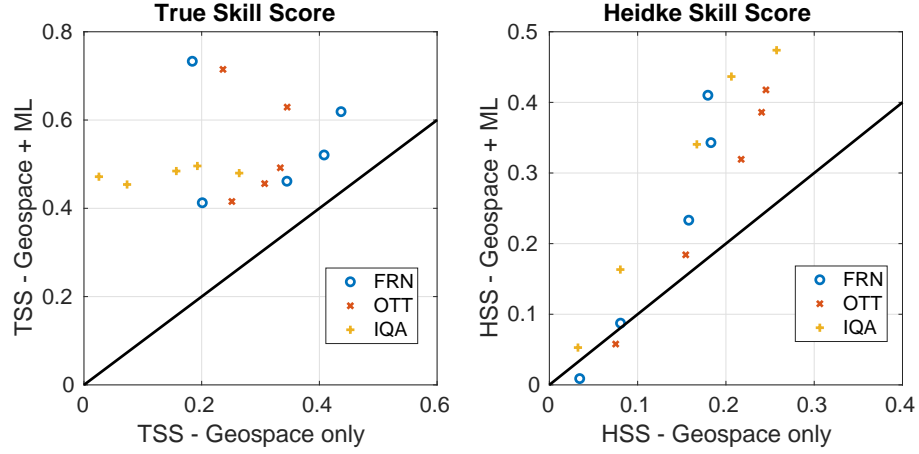
**Figure 19.** FRN station: ROC curves (TPR vs FPR) for different thresholds. Filled dots indicate the optimal points along a given ROC curve.



**Figure 20.** OTT station: ROC curves (TPR vs FPR) for different thresholds. Filled dots indicate the optimal points along a given ROC curve.



**Figure 21.** IQA station: ROC curves (TPR vs FPR) for different thresholds. Filled dots indicate the optimal points along a given ROC curve.



**Figure 22.** Comparison of the True Skill Score (left) and Heidke Skill Score (right) for models using the output of the Geospace model alone (horizontal axis) vs the model presented in this paper (combining Geospace outputs with machine learning, vertical axis). The diagonal black line indicates no improvement. Blue, red, and yellow symbols are for FRN, OTT, and IQA respectively.

## References

- Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.
- Boteler, D., & Pirjola, R. (1998). The complex-image method for calculating the magnetic and electric fields produced at the surface of the earth by the auroral electrojet. *Geophysical Journal International*, 132(1), 31–40.
- Boteler, D., Pirjola, R., & Nevanlinna, H. (1998). The effects of geomagnetic disturbances on electrical systems at the earth’s surface. *Advances in Space Research*, 22(1), 17–27.
- Breiman, L. (2017). *Classification and regression trees*. Routledge.
- Camporeale, E. (2019). The challenge of machine learning in space weather nowcasting and forecasting. *Space Weather*, 17(8).
- Camporeale, E., Wing, S., & Johnson, J. (2018). *Machine learning techniques for space weather*. Elsevier.
- DeGroot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2), 12–22.
- Freund, Y. (2009). A more robust boosting algorithm. *arXiv preprint arXiv:0905.2138*.
- Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), 119–139.
- Gleisner, H., & Lundstedt, H. (2001). A neural network-based local model for prediction of geomagnetic disturbances. *Journal of Geophysical Research: Space Physics*, 106(A5), 8425–8433.
- Gombosi, T. I., Powell, K. G., De Zeeuw, D. L., Clauer, C. R., Hansen, K. C., Manchester, W. B., ... others (2004). Solution-adaptive magnetohydrodynamics for space plasmas: Sun-to-earth simulations. *Computing in science & engineering*, 6(2), 14.
- Gordeev, E., Sergeev, V., Honkonen, I., Kuznetsova, M., Rastätter, L., Palmroth, M., ... Wiltberger, M. (2015). Assessing the performance of community-available global mhd models using key system parameters and empirical relationships. *Space Weather*, 13(12), 868–884.
- Horton, R., Boteler, D., Overbye, T. J., Pirjola, R., & Dugan, R. C. (2012). A test case for the calculation of geomagnetically induced currents. *IEEE Transactions on Power Delivery*, 27(4), 2368–2373.
- Jordanova, V. K., Delzanno, G. L., Henderson, M. G., Godinez, H. C., Jeffery, C., Lawrence, E. C., ... others (2018). Specification of the near-earth space environment with shields. *Journal of Atmospheric and Solar-Terrestrial Physics*, 177, 148–159.
- Krzanowski, W. J., & Hand, D. J. (2009). *Roc curves for continuous data*. Chapman and Hall/CRC.
- Lanzerotti, L. J. (2001). Space weather effects on technologies. *Space weather*, 125, 11–22.
- Lotz, S., & Cilliers, P. (2015). A solar wind-based model of geomagnetic field fluctuations at a mid-latitude station. *Advances in Space Research*, 55(1), 220–230.
- Ngwira, C. M., Pulkkinen, A., McKinnell, L.-A., & Cilliers, P. J. (2008). Improved modeling of geomagnetically induced currents in the south african power network. *Space Weather*, 6(11).
- Niculescu-Mizil, A., & Caruana, R. (2005). Obtaining calibrated probabilities from boosting. In *Uai* (p. 413).
- Pirjola, R. (2002). Review on the calculation of surface electric and magnetic fields and of geomagnetically induced currents in ground-based technological systems. *Surveys in geophysics*, 23(1), 71–90.
- Pirjola, R. (2007). Space weather effects on power grids. In *Space weather-physics and effects* (pp. 269–288). Springer.

- Pirjola, R., Boteler, D., Viljanen, A., & Amm, O. (2000). Prediction of geomagnetically induced currents in power transmission systems. *Advances in Space Research*, 26(1), 5–14.
- Pulkkinen, A., Kuznetsova, M., Ridley, A., Raeder, J., Vapirev, A., Weimer, D., ... others (2011). Geospace environment modeling 2008–2009 challenge: Ground magnetic field perturbations. *Space Weather*, 9(2).
- Pulkkinen, A., Lindahl, S., Viljanen, A., & Pirjola, R. (2005). Geomagnetic storm of 29–31 october 2003: Geomagnetically induced currents and their relation to problems in the swedish high-voltage power transmission system. *Space Weather*, 3(8).
- Pulkkinen, A., Rastatter, L., Kuznetsova, M., Singer, H., Balch, C., Weimer, D., ... others (2013). Community-wide validation of geospace model ground magnetic field perturbation predictions to support model transition to operations. *Space Weather*, 11(6), 369–385.
- Raeder, J., Berchem, J., & Ashour-Abdalla, M. (1998). The geospace environment modeling grand challenge: Results from a global geospace circulation model. *Journal of Geophysical Research: Space Physics*, 103(A7), 14787–14797.
- Rastätter, L., Kuznetsova, M., Gloer, A., Welling, D., Meng, X., Raeder, J., ... others (2013). Geospace environment modeling 2008–2009 challenge: D st index. *Space Weather*, 11(4), 187–205.
- Rastätter, L., Kuznetsova, M., Vapirev, A., Ridley, A., Wiltberger, M., Pulkkinen, A., ... Singer, H. (2011). Geospace environment modeling 2008–2009 challenge: Geosynchronous magnetic field. *Space Weather*, 9(4), 1–15.
- Ridley, A., Gombosi, T., & DeZeeuw, D. (2004). Ionospheric control of the magnetosphere: Conductance. In *Annales geophysicae* (Vol. 22, pp. 567–584).
- Schrijver, C. J., & Mitchell, S. D. (2013). Disturbances in the us electric grid associated with geomagnetic activity. *Journal of Space Weather and Space Climate*, 3, A19.
- Thébault, E., Finlay, C. C., Beggan, C. D., Alken, P., Aubert, J., Barrois, O., ... others (2015). International geomagnetic reference field: the 12th generation. *Earth, Planets and Space*, 67(1), 79.
- Toffoletto, F., Sazykin, S., Spiro, R., & Wolf, R. (2003). Inner magnetospheric modeling with the rice convection model. *Space Science Reviews*, 107(1-2), 175–196.
- 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.
- Tóth, G., Sokolov, I. V., Gombosi, T. I., Chesney, D. R., Clauer, C. R., De Zeeuw, D. L., ... others (2005). Space weather modeling framework: A new tool for the space science community. *Journal of Geophysical Research: Space Physics*, 110(A12).
- Tóth, G., Van der Holst, B., Sokolov, I. V., De Zeeuw, D. L., Gombosi, T. I., Fang, F., ... others (2012). Adaptive numerical algorithms in space weather modeling. *Journal of Computational Physics*, 231(3), 870–903.
- 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.
- Viljanen, A., Nevanlinna, H., Pajunpää, K., & Pulkkinen, A. (2001). Time derivative of the horizontal geomagnetic field as an activity indicator. In *Annales geophysicae* (Vol. 19, pp. 1107–1118).
- Viljanen, A., Pulkkinen, A., Amm, O., Pirjola, R., & Korja, T. (2004). Fast computation of the geoelectric field using the method of elementary current systems and planar earth models. In *Annales geophysicae* (Vol. 22, pp. 101–113).
- Weigel, R., Klimas, A., & Vassiliadis, D. (2003). Solar wind coupling to and predictability of ground magnetic fields and their time derivatives. *Journal of*



- Geophysical Research: Space Physics*, 108(A7).
- Weigel, R., Vassiliadis, D., & Klimas, A. (2002). Coupling of the solar wind to temporal fluctuations in ground magnetic fields. *Geophysical Research Letters*, 29(19), 21–1.
- Weimer, D. R. (2013). An empirical model of ground-level geomagnetic perturbations. *Space Weather*, 11(3), 107–120.
- Welling, D. (2019). Magnetohydrodynamic models of  $b$  and their use in gic estimates. *Geomagnetically Induced Currents from the Sun to the Power Grid*, 43–65.
- Welling, D., Anderson, B., Crowley, G., Pulkkinen, A., & Rastätter, L. (2017). Exploring predictive performance: A reanalysis of the geospace model transition challenge. *Space Weather*, 15(1), 192–203.
- Welling, D., & Ridley, A. (2010). Validation of swmf magnetic field and plasma. *Space Weather*, 8(3).
- Wintoft, P. (2005). Study of the solar wind coupling to the time difference horizontal geomagnetic field. In *Annales geophysicae* (Vol. 23, pp. 1949–1957).
- Wintoft, P., Wik, M., Lundstedt, H., & Eliasson, L. (2005). Predictions of local ground geomagnetic field fluctuations during the 7–10 november 2004 events studied with solar wind driven models. In *Annales geophysicae* (Vol. 23, pp. 3095–3101).
- Wintoft, P., Wik, M., & Viljanen, A. (2015). Solar wind driven empirical forecast models of the time derivative of the ground magnetic field. *Journal of Space Weather and Space Climate*, 5, A7.
- Yu, Y., & Ridley, A. J. (2008). Validation of the space weather modeling framework using ground-based magnetometers. *Space Weather*, 6(5), 1–20.
- Zhang, B., Sorathia, K. A., Lyon, J. G., Merkin, V. G., Garretson, J. S., & Wiltberger, M. (2019). Gamera: A three-dimensional finite-volume mhd solver for non-orthogonal curvilinear geometries. *The Astrophysical Journal Supplement Series*, 244(1), 20.