







Global Geomagnetic Perturbation Forecasting Using Deep Learning

Vishal Upendran^{1,2} , Panagiotis Tigas^{1,3}, Banafsheh Ferdousi^{1,4} , Téo Bloch^{1,5} , Mark C. M. Cheung^{1,6} , Siddha Ganju^{1,7}, Asti Bhatt^{1,8} , Ryan M. McGranaghan^{1,9} , and Yarin Gal^{1,3}

¹Frontier Development Lab, Sunnyvale, CA, USA, ²Inter University Centre for Astronomy and Astrophysics, Pune, India, ³OATML, University of Oxford, Oxford, UK, ⁴University of New Hampshire, Durham, NH, USA, ⁵University of Reading, Reading, UK, ⁶Lockheed Martin Advanced Technology Center, Palo Alto, CA, USA, ⁷NVIDIA Corporation, Santa Clara, CA, USA, ⁸SRI International, Menlo Park, CA, USA, ⁹ASTRA LLC, Louisville, CO, USA

Key Points:

- Global high-time cadence models for forecasting geomagnetic perturbations are necessary for this technologically driven society
- We develop a grid-free model that forecasts these perturbations 30 min in the future at any spatial resolution at 1 min cadence
- The proposed model outperforms/has consistent performance against the state of the practice local (global) high (low) time cadence models

Supporting Information:

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

Correspondence to:

V. Upendran,
uvishal@iucaa.in

Citation:

Upendran, V., Tigas, P., Ferdousi, B., Bloch, T., Cheung, M. C. M., Ganju, S., et al. (2022). Global geomagnetic perturbation forecasting using deep learning. *Space Weather*, 20, e2022SW003045. <https://doi.org/10.1029/2022SW003045>

Received 24 JAN 2022
Accepted 12 MAY 2022

Abstract Geomagnetically Induced Currents (GICs) arise from spatio-temporal changes to Earth's magnetic field, which arise from the interaction of the solar wind with Earth's magnetosphere, and drive catastrophic destruction to our technologically dependent society. Hence, computational models to forecast GICs globally with large forecast horizon, high spatial resolution and temporal cadence are of increasing importance to perform prompt necessary mitigation. Since GIC data is proprietary, the time variability of the horizontal component of the magnetic field perturbation (dB/dt) is used as a proxy for GICs. In this work, we develop a fast, global dB/dt forecasting model, which forecasts 30 min into the future using only solar wind measurements as input. The model summarizes 2 hr of solar wind measurement using a Gated Recurrent Unit and generates forecasts of coefficients that are folded with a spherical harmonic basis to enable global forecasts. When deployed, our model produces results in under a second, and generates global forecasts for horizontal magnetic perturbation components at 1 min cadence. We evaluate our model across models in literature for two specific storms of 5 August 2011 and 17 March 2015, while having a self-consistent benchmark model set. Our model outperforms, or has consistent performance with state-of-the-practice high time cadence local and low time cadence global models, while also outperforming/having comparable performance with the benchmark models. Such quick inferences at high temporal cadence and arbitrary spatial resolutions may ultimately enable accurate forewarning of dB/dt for any place on Earth, resulting in precautionary measures to be taken in an informed manner.

Plain Language Summary Geomagnetically induced currents (GICs) result due to the interaction of the solar wind with Earth's magnetosphere, and are catastrophic to our technologically dependent society. Since GIC data is proprietary, the time variability of geomagnetic perturbation is used as a proxy, and forecasting these perturbation at high spatial resolution and time cadence is important. In this work we develop a deep learning-based model to forecast these perturbation measurements at arbitrary spatial resolutions and at high time cadence, using only the solar wind measurements. Our model outperforms, or has consistent performance at worse with benchmark models, and hence can provide quick, accurate forecasts at high time cadence across the whole globe.

1. Introduction

Geomagnetic storms drive a spectrum of potentially catastrophic disruptions to our technologically dependent society (UN, 2017). A cohort study of insurance claims of electrical equipment provides evidence that space weather poses a continuous threat to electrical distribution grids via geomagnetic storms and geomagnetically induced currents (GICs) (Schrijver et al., 2014). GICs also pose a threat to oil pipelines, railways, and telecommunication systems (Barlow et al., 1849; Boteler, 2001; Eastwood et al., 2018; Pulkkinen et al., 2001), potentially wiping out the backbone of economies and destroying the livelihoods of people worldwide. In the case of extreme but historically probable geomagnetic storms, the economic impact due to prolonged power outages can exceed billions of dollars per day (Oughton et al., 2017). Hence, it is imperative to monitor and forecast space weather impacts like geomagnetic storms and GICs.

© 2022 The Authors.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial License](https://creativecommons.org/licenses/by-nc/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

GICs are driven by the geoelectric field that depends on temporal changes in the horizontal component of ground magnetic field perturbation (dB/dt) and local Earth geology. Due to their proprietary nature, publicly available GIC data are limited. However, the geomagnetic perturbations can be measured using Ground magnetometer stations, and may be used as a good proxy to study GICs variations (Kozyreva et al., 2018; Lanzerotti, 2001; Ngwira et al., 2018). The challenge, however, is twofold: (a) the ground magnetometers measurements are not performed uniformly across the Earth, and are spatially sparse and (b) perturbation changes occur over timescales of minutes.

For predicting dB/dt , ground magnetic perturbations models at high spatial and temporal resolution are essential. Currently, first-principle models are used to forecast magnetic field perturbation as a part of the NOAA-Space Weather Prediction Center using the Space Weather Modeling Framework (SWMF; see, e.g., Tóth et al., 2005; Tóth et al., 2011, 2012). The models generate forecasts of the global heliosphere, while they also provide forecasts of the magnetospheric parameters as a part of SWMF. However, these models are computationally expensive and require a long run time for high-resolution forecasts, which is necessary for highly localized magnetic field fluctuations.

Data-Driven empirical models are more feasible for Space Weather forecasting due to their high speed and low computational cost (Camporeale, 2019). However, these empirical data-driven models (e.g., Weimer, 2013; Weigel et al., 2002) did not perform well under Community-wide validation of geospace model ground magnetic field perturbation (dB/dt) predictions by Pulkkinen et al. (2013). The study was performed based on the three first principle models and two empirical models as a function of upstream solar wind drivers using Heidke Skill Score (HSS) metrics over a number of ground magnetometer stations in middle- and high-latitudes. Further evaluation of the models by Welling et al. (2017) concluded that all the models underpredict dB/dt during more active times and the need for model-data comparison and model improvements.

Machine learning (ML) and deep learning (DL) are rapidly growing areas that operate on large data. These have been used with great success in various studies—right from forecasting the solar wind (Upendran et al., 2020) to correlating auroral dynamics with Global navigation satellite system scintillations (Lamb et al., 2019). Wintoft et al. (2015) develop a neural network to forecast 30 min maximum of $|dB/dt|$ at multiple stations over Europe with good success. More recently, Keese et al. (2020) developed two models—an artificial neural network model, and a Long Short Term Memory cell (LSTM); Hochreiter & Schmidhuber, 1997) model—to forecast the geomagnetic perturbations at the Ottawa station. While these studies forecast the perturbations at high temporal cadence (at ≈ 1 min cadence), they are limited to forecast at specific spatial locations on the globe.

In this work, we develop a near grid-free global geomagnetic perturbation forecasting model using DL to address the issues of near-real-time forecasts at high spatial and temporal cadence. This is performed by coupling a DL model with a spherical harmonic basis, rendering the model near grid-free. The model takes the solar wind parameters, the Interplanetary Magnetic Field (IMF) measurements, and the solar radio flux measurements as input. It generates a forecast of perturbation measurements across the Earth with a lead time of 30 min. These forecasts may then be sampled over a grid at nearly any resolution. Owing to the global nature of spherical harmonics, the perturbation forecasts may in principle be sampled at any location on the globe. The remainder of the paper is structured as follows: in Section 2, we describe the data used in this work, along with the various preprocessing steps in Section 2.3. Then, we describe the main modeling scheme with the evaluation metrics in Section 3.1, benchmark models in Section 3.2, and our proposed model **Deep leArninG Geomagnetic pErtuRbation (DAGGER)** in Section 3.3. Finally, we present the results of our model in Section 4, with a detailed analysis on two selected storms in Section 4.2, and follow it up with a summary and broader impact in Section 5.

2. Data

2.1. Perturbation Measurement Data Set

In this study, we obtain the ground magnetic perturbations measurements from the SuperMAG (Gjerloev, 2012) consortium. SuperMAG is a global network of ground stations employed in measurement of geomagnetic perturbations. The available data set comprises of measurements from around 300 magnetometer stations around the globe. These data are validated, transformed to a common coordinate system and processed with the same baseline remove methodology. From SuperMAG, we obtain perturbations in the geomagnetic field, δb_e and δb_n , from 2010 to 2019 at 1 min cadence.

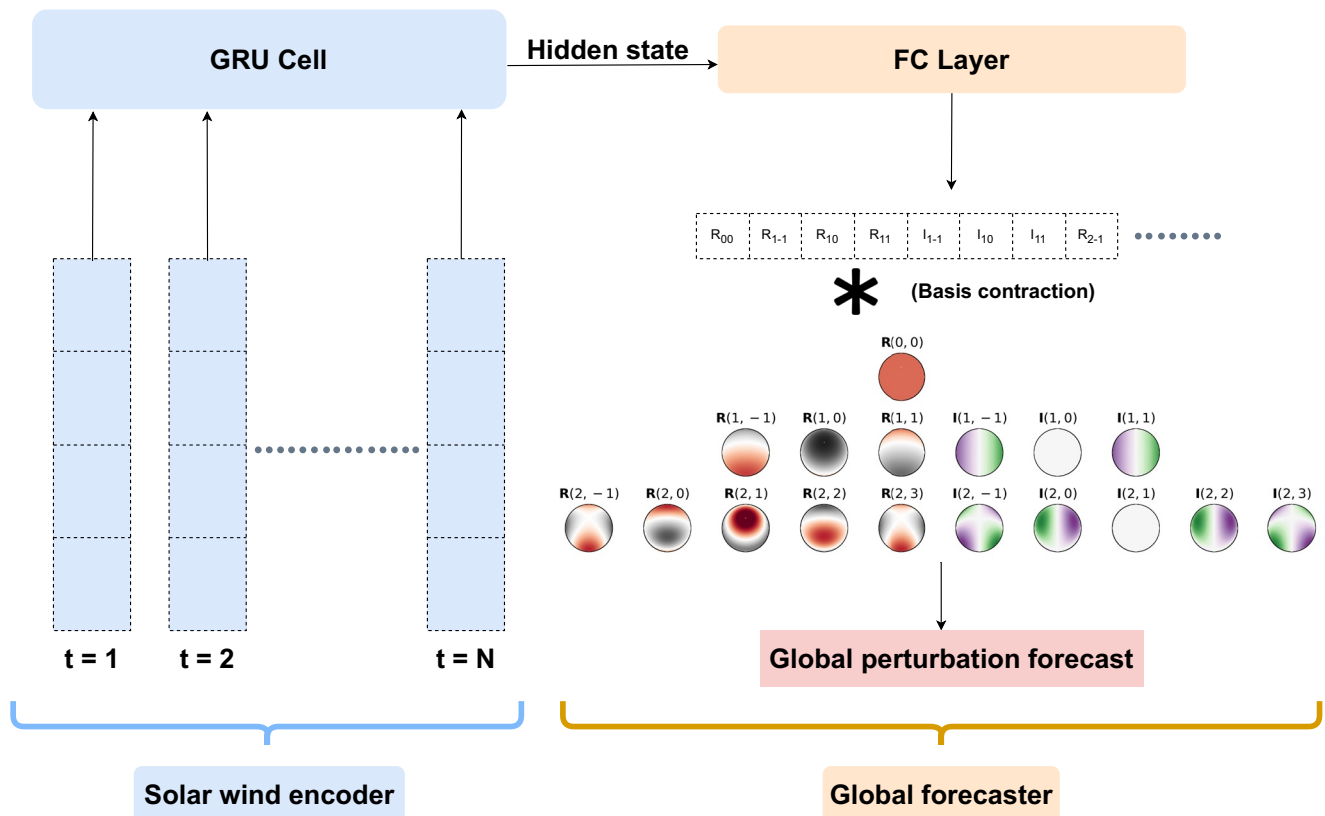


Figure 1. Architecture of Deep Learning Geomagnetic Perturbation. The model has three principle components—a time series summarizer, a coefficient generator, and a spherical harmonic constructor. The time series summarizer (Gated Recurrent Unit (GRU) cell) takes in the solar wind time series, and generates a summary hidden state. This is fed to a fully connected layer (FC Layer), which generates a vector of coefficients. These coefficients are contracted with the spherical harmonic basis to generate global forecast of perturbations.

The SuperMAG stations are primarily located in the Northern Hemisphere. Thus, while the perturbations are densely sampled in the Northern Hemisphere, the sampling becomes sparser (and hence more susceptible to outliers) for lower latitudes and the Southern Hemisphere. Particularly, the coverage of SuperMAG stations is dense for Magnetic Latitude (MAGLAT) $\geq 40^\circ$. Hence, to ensure a robust forecast and as a first step in developing a grid-free model, we select stations only above a MAGLAT of 40° . Since we focus primarily on forecasting at MAGLAT $\geq 40^\circ$, our results are well constrained for the same regions. However, we emphasize that our solution formalism is generic enough to perform a forecast anywhere on the globe—the forecasts for regions with MAGLAT $\leq 40^\circ$, however, may not be expected to be as well constrained. This selection leaves us with a total of 175 magnetometer stations (at max) to constrain our forecasts.

2.2. Solar Wind, IMF, and Solar Proxy Data Set

We use the solar wind and IMF measurements at 1 min cadence from NASA/GSFC's OMNI data set (through OMNIWeb). Particularly, we use, measurements of the three components of the IMF in Geocentric Solar Magnetospheric coordinates (B_x , B_y , B_z), solar wind speed (V_{sw}), solar wind proton temperature (T), the clock angle of the IMF (θ_c), and finally the solar radio flux at 10.7 cm ($F_{10.7}$) (King & Papitashvili, 2005; Papitashvili et al., 2014).

From these basic measurements, we generate “good” features as input to our model following Weimer (2013). We perform this feature generation to ensure accelerated convergence of our model,

Table 1

Model Architecture: A Summary

Layer name	Size
GRU	8 units
FC: MLP Layer 1	16
FC: MLP Layer 2	440*2 (real and imaginary parts)
Spherical harmonic layer (NOT trainable)	—

Note. GRU, Gated Recurrent Unit.

Table 2
Hyperparameters Set Through Grid Search

Hyperparameter	Value
OMNI time series length	120 min
Maximum number of modes	20
Learning rate	5×10^{-3}
L2 regularization coefficient	5×10^{-5}
Dropout probability	0.7
Batch size	8,500
Optimizer	Adam, with default Pytorch parameters

as these features are known to be important for reconstruction of the perturbation maps (Weimer, 2013; Weimer et al., 2010).

The inputs to our model are: $B_x, B_y, B_z, B_T, V_{SW}, t$ (dipole axis angle in radians), $\theta_c, T, \sqrt{F_{10.7}}, B_T \cos(\theta_c), V_{SW} \cos(\theta_c), t \cos(\theta_c), \sqrt{F_{10.7}} \cos(\theta_c), B_T \sin(\theta_c), V_{SW} \sin(\theta_c), t \sin(\theta_c), \sqrt{F_{10.7}} \sin(\theta_c), B_T \cos(2\theta_c), V_{SW} \cos(2\theta_c), B_T \sin(2\theta_c), V_{SW} \sin(2\theta_c)$.

2.3. Data Preprocessing

In general, the data set for any ML work is split into three independent training, testing, and validation sets. The training set is used to train the model, while the validation set is used to find the best model parameters that explain both the training and validation sets well. Finally, the model is evaluated on a testing set. Since we have a continuous time series of data, which covers almost 75% of the solar cycle, a naive division of different years into the

three sets may result in bias due to prevalence of storms. Thus, in order to obtain a long enough time series to avoid edge effects, and mitigate bias from storm prevalence, we divide the whole time series into 100 buckets. Of these, we consider the two buckets with the 2011 and 2015 storms for benchmark. The remaining buckets are then split as 80% training set, 10% validation, and 10% testing set. Also, note that following Weimer (2013), we have included the F10.7 measurement, which is a widely used index of solar ultraviolet radiation levels and solar activity (Clette, 2021; Verbanac et al., 2011). While we expect F10.7 to provide some degree of information regarding the solar cycle, note that this index also shows localized variations (Tapping, 2013). However, performing a detailed, quantitative analysis of the effect of the solar cycle on our model is beyond the scope of the current work. Hence, we may only expect some effect of the solar cycle to be captured by our model at this stage.

The OMNI data at 1 min cadence have missing values at multiple times, while the SuperMAG measurements have missing data both at different times, and for different stations. Across the full data set (train + test + val + storm), the OMNI solar wind measurements have the maximum missing data ($\approx 25\%$). Similarly, for the two storm time series, the solar wind measurements again have the maximum number of missing data ($\approx 18\%$ for 2011, and $\approx 24\%$ for 2015 storm). The SuperMAG data, on the other hand, have stations that go offline. This results in no target sample at the station location. During the storm times, the stations in consideration have a median missing fraction of $\approx 5\%$. We report the median missing fraction for the missing SuperMAG measurements, as the missing stations do not contribute to our training scheme.

To make the data set uniform, we replace all missing values with 0 for both the OMNI and SuperMag data. To prevent any effect of missing measurements on our network, we replace the corresponding forecasts with 0 during training and validation time. This ensures that the “error” is zero for the particular sample, and that it does not contribute to training (and validation) of the network.

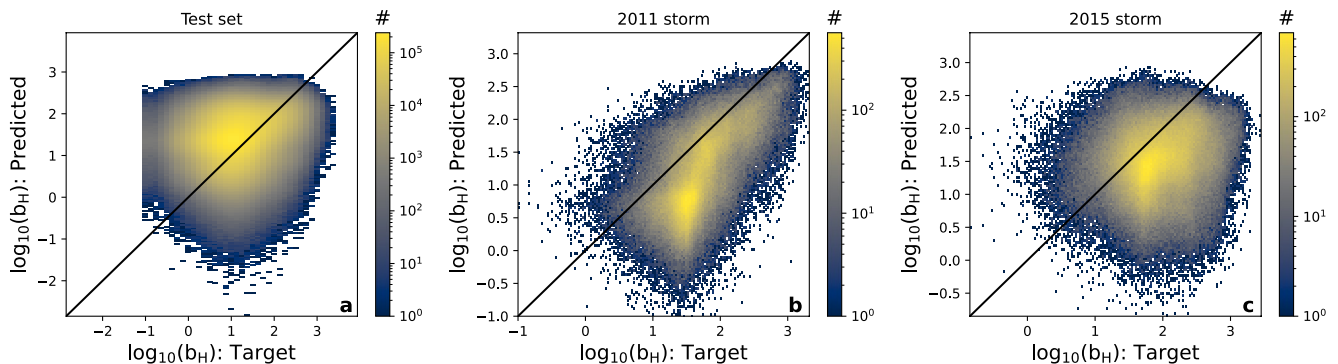


Figure 2. The joint histogram of predictions v/s target SuperMag measurements for all points in the test set (panel a), 2011 storm set (panel b), and 2015 storm set (panel c). The colors depict the number of points in each bin of the joint histogram.

Table 3
RMSE and MAE Comparison Between DAGGER, W2013, and Persistence Models

Storm	Metric	DAGGER		W2013		Persistence	
		δb_e	δb_n	δb_e	δb_n	δb_e	δb_n
2011	MAE	34.99	53.20	67.41	76.74	30.87	43.52
	RMSE	72.86	100.46	127.54	140.93	73.53	97.41
2015	MAE	61.44	104.7	104.69	121.48	47.17	67.4
	RMSE	102.45	175.37	179.97	195.52	87.78	128.90

Note. Both the metrics are in units of nT. DAGGER, Deep leArninG Geomagnetic pErturbation; MAE, Mean Absolute Error; RMSE, Root Mean Square Error.

Before feeding the data (both OMNI and SuperMag) to our model, it is good practice to standardize the data by subtracting the mean and dividing by the standard deviation of the training set for each column. Due to memory constraints and the very large number of datapoints in the data set, we generate the mean and standard deviation for 10,000 random points from the data. This “Monte Carlo” sample of points generates a mean and standard deviation, which serves as a proxy for the training set mean and standard deviation. During inference time, these values are used to scale the validation and testing sets.

3. Modeling and Methods

3.1. Metrics for Model Evaluation

We define multiple metrics to evaluate our model. For a target measurement of y and forecast of \hat{y} , the metrics are listed below:

1. Root Mean Square Error (RMSE):

$$\text{RMSE} := \sqrt{\frac{1}{N} \sum_i^N (y - \hat{y})^2},$$

where the average is taken across all samples.

2. Mean Absolute Error (MAE):

$$\text{MAE} := \frac{1}{N} \sum_i^N (|y - \hat{y}|),$$

where the average is taken across all samples.

Apart from these two metrics, we also use the Pulkkinen-Welling metrics, which are based on binary event analysis for geomagnetic storms (Pulkkinen et al., 2013). This analysis is performed only for the two storm series of 2011 and 2015, and not for the validation and testing sets. For such an analysis, we define the horizontal perturbation component as

$$\delta b_H = \sqrt{\delta b_e^2 + \delta b_n^2},$$

The time derivative $d\delta b/dt$ is approximated as:

$$\frac{d\delta b_{H,i}}{dt} \approx \sqrt{\left(\frac{\delta b_{e,i} - \delta b_{e,i-1}}{1\text{min}}\right)^2 + \left(\frac{\delta b_{n,i} - \delta b_{n,i-1}}{1\text{min}}\right)^2}.$$

Following Pulkkinen et al. (2013), we divide our target and forecast δb_e and δb_n for each station, into 20 min nonoverlapping time windows. For each window, if $d\delta b_{H,i}/dt$ crosses a specified threshold, the segment is given a value 1—else, it is given a value of 0. Thus, by comparing strings of 1 and 0 s, we can then understand how good the model is at predicting events above or below a specific magnitude. **Hits** (H) are defined as the number of correctly forecasted 1s, while **Misses** (M) correspond to the number of measured 1 s marked 0 by the model. Similarly, **False alarms** (F) correspond to observed 0 s, which are marked as 1 by the model, while **True negatives** (N) are 0 s in the observation, marked as 0 by the model. Using this contingency table, we define four standard metrics following Welling et al. (2018) to evaluate our model:

2011 storm: Comparison of performance for the top - 3 best, and worst performing stations

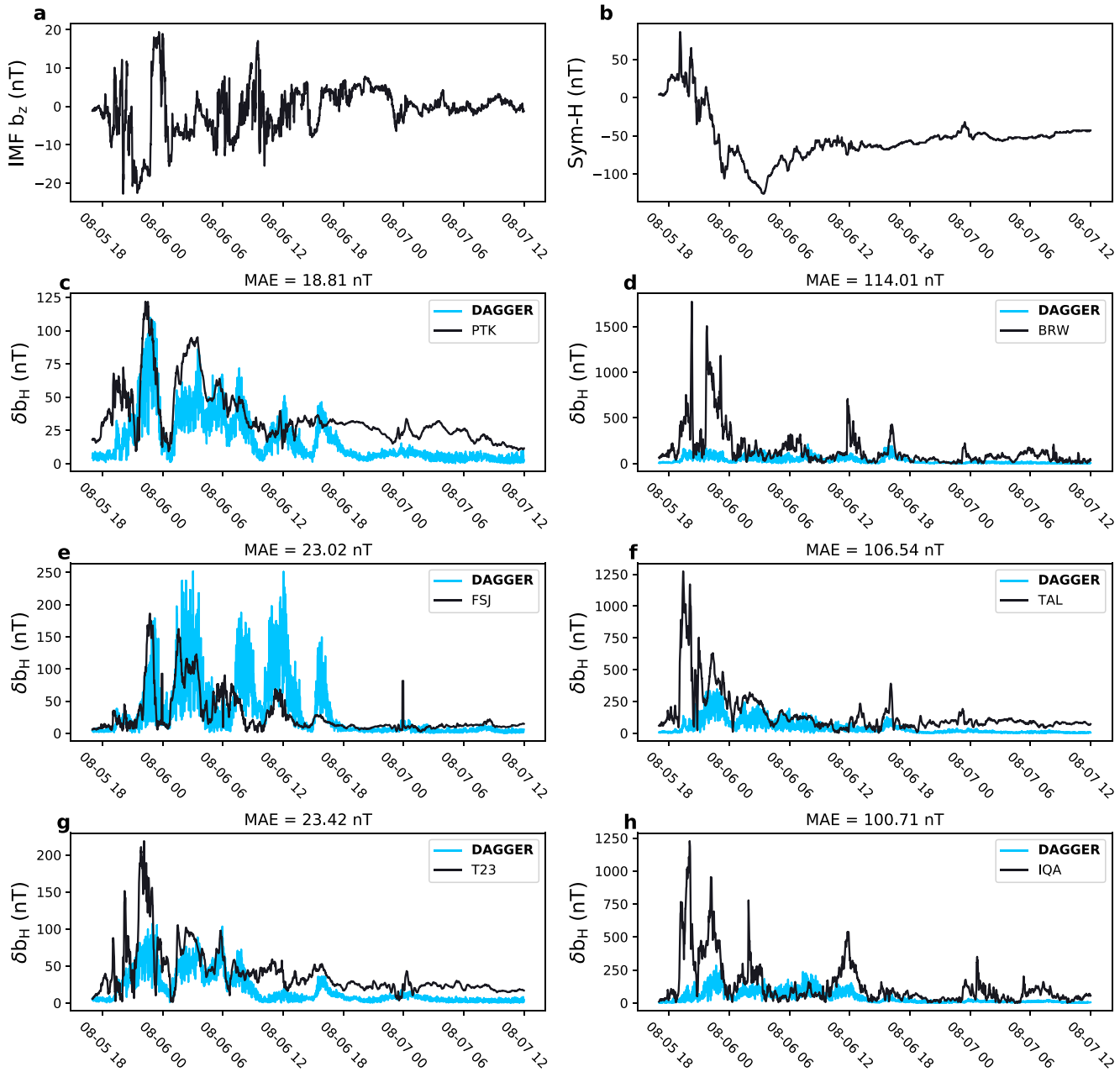


Figure 3. The IMF B_z (panel a), Sym-H (panel b) and top three best (panels c, e, g) and worst (panels d, f, h) performing stations for the 2011 storm. The blue color indicates forecast from Deep leArninG Geomagnetic pErtuRbation (DAGGER), while the black color indicates measurements at different stations (in the legend of each figure), with the Mean Absolute Error (MAE) reported on top.

1. Probability of Detection (POD):

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

2. Probability of False Detection (POFD):

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

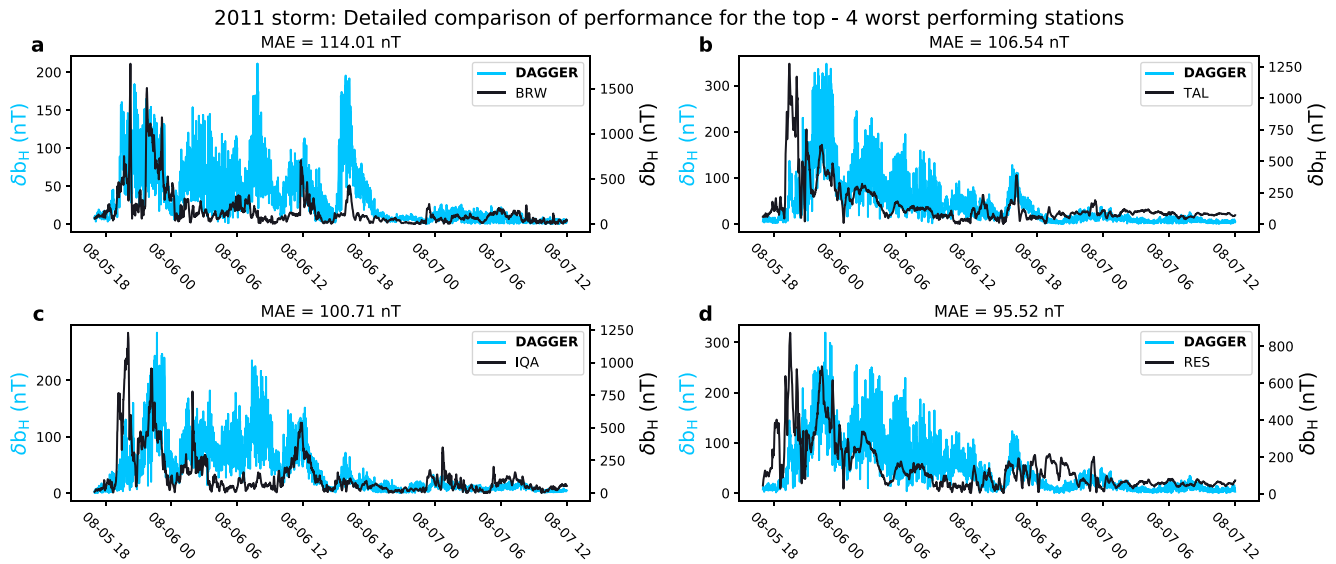


Figure 4. The measurements at different stations (black) and forecast (blue) of Deep leArninG Geomagnetic pErTuRbation (DAGGER) for the 2011 storm, with different Y-axis scales to bring out the detailed features from Figure 3.

3. Proportion Correct (PC):

$$PC = \frac{H + N}{H + N + F + M}$$

4. HSS is a measure of correctly predicted results after accounting for those which may be correct, purely due to chance. The HSS is defined as:

$$HSS = \frac{2(HN - MF)}{(H + M)(M + N) + (H + F)(F + N)}$$

In this work, we select four different thresholds of 18, 42, 66, and 90 nT/min following Pulkkinen et al. (2013).

3.2. Benchmark Models

We use the 2011 and 2015 storm data sets, at 1 min cadence as benchmark. Thus, the results presented here may be directly compared with other models evaluated on the same data (e.g., with the models proposed by Keese et al., 2020). However, we also have two self-consistent benchmark models operated on the same data set.

The first and the simplest model is a persistence model. In our formulation, this model propagates the target SuperMAG measurement at time T to $T + LAG$, where our LAG time is the forecasting horizon of our model. This propagation is performed for each station. Such a persistence model imposes a strong constraint on the utility of any proposed modeling scheme on “how much” new information is captured. For each target measurement, we also compute all the metrics for the persistence model.

Our second benchmark model is the empirical fitting scheme of Weimer (2013, henceforth called **W2013**). This is an empirical fitting scheme that decomposes the perturbation measurements into spherical harmonics, assuming the coefficients depend only on the solar wind parameters. This is a much stronger constraint over the persistence model, for it actually generates a map between the solar wind and perturbation measurements. Note that the **W2013** metrics are generated only for the two storm times, since we do not have the forecast for all times in our data set.

2015 storm: Comparison of performance for the top - 3 best, and worst performing stations

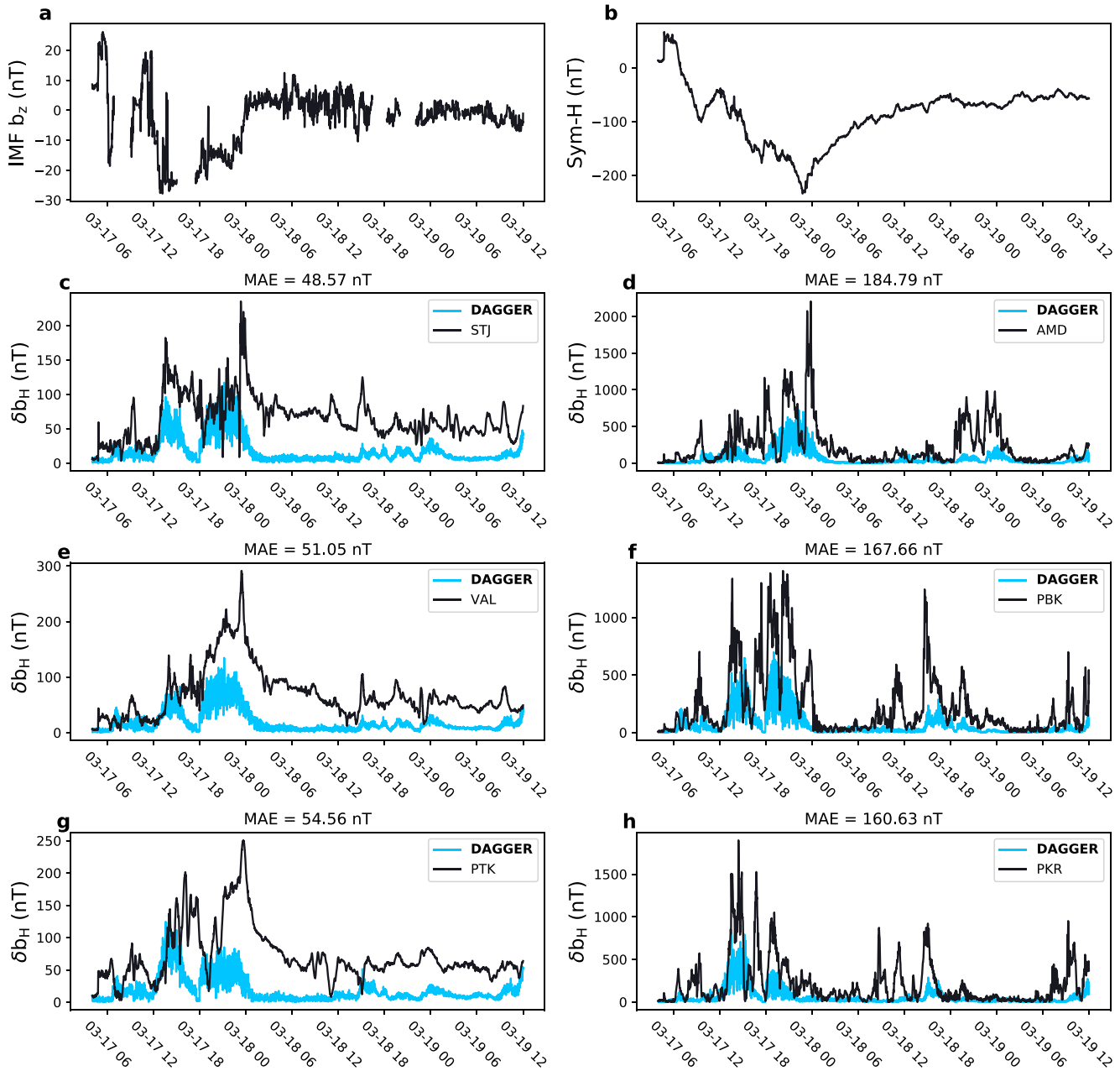


Figure 5. Same as Figure 3, but for the 2015 storm.

3.3. Proposed Deep Learning Model: DAGGER

The **DAGGER** model is a DL model. We use T hours of OMNI data at 1 min cadence as input, and forecast the geomagnetic perturbations LAG minutes from the final input. The length of OMNI data and the LAG value are free parameters, which are set through a hyperparameter search—this is explained later in Section 3.4. The model has three parts: a time series summarizer, a coefficient generator, and a spherical harmonic constructor. We first describe the spherical harmonic formulation, and then explain the full model.

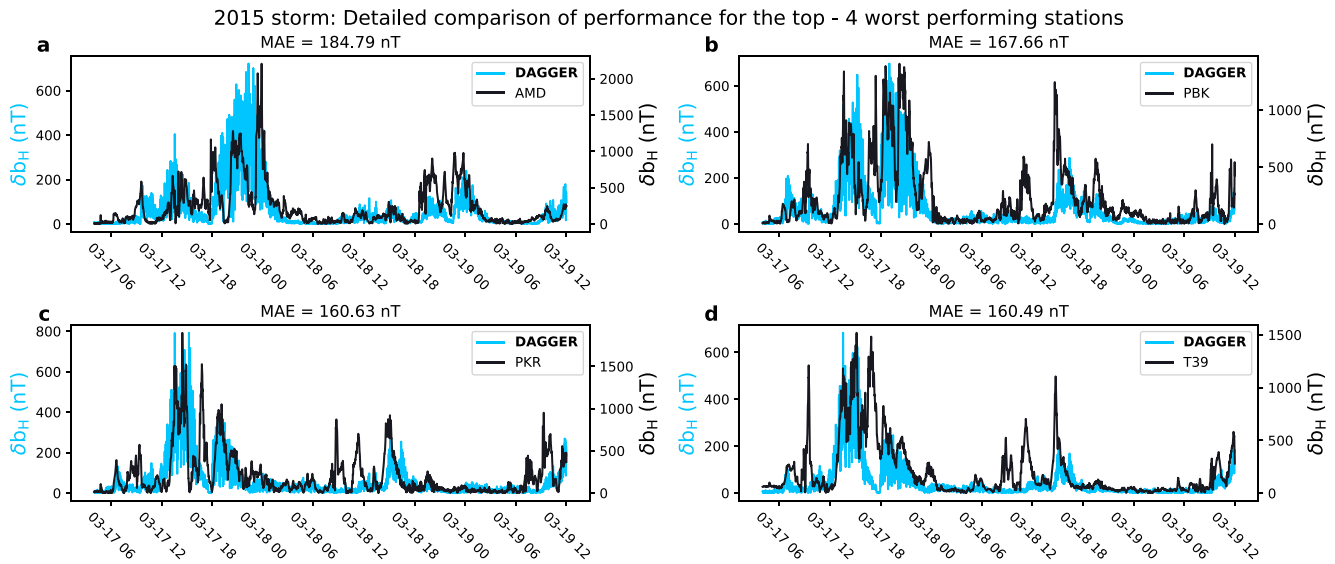


Figure 6. Same as Figure 4, but for the 2015 storm.

3.3.1. Spherical Harmonic Formulation

Since we seek to develop a forecast model with continuous spatial coverage, we develop an almost “grid free” approach to forecast using Spherical Harmonics. Spherical harmonics assume a continuous and differentiable functional form of any field sought to be decomposed over a spherically symmetric manifold. Since we expect the perturbation fields to be largely smooth and devoid of localized peaks, we forecast the spherical harmonic coefficients, which can be easily transformed to the perturbations depending on the grid.

Any scalar field over the unit sphere can be expressed as

$$f(\theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^n a_{nm} Y_{nm}(\theta, \phi),$$

where

$$Y_{nm}(\theta, \phi) := \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} e^{im\theta} P_n^m(\cos(\phi)),$$

and $P_n^m(\cos(\phi))$ are the associated Legendre polynomials.

These functions $Y_{nm}(\theta, \phi)$ are solutions to Laplace equation in a spherically symmetric coordinate system. If the sum is truncated at a maximum harmonic degree N , $f(\theta, \phi)$ is approximated as

$$\tilde{f}(\theta, \phi) = \sum_{n=0}^N \sum_{m=-n}^n a_{nm} Y_{nm}(\theta, \phi). \quad (1)$$

Defining $i = n^2 + n + m$, we may rewrite Equation 1 as

$$f(\theta, \phi) \approx \tilde{f}(\theta, \phi) = \sum_{i=0}^{(N+1)^2-1} a_i Y_i(\theta, \phi). \quad (2)$$

If the 2-D fields over θ, ϕ are unrolled as one-dimensional arrays, we have

$$\tilde{f} = B\vec{a},$$

Maps for timestep with min MAE Target SuperMag

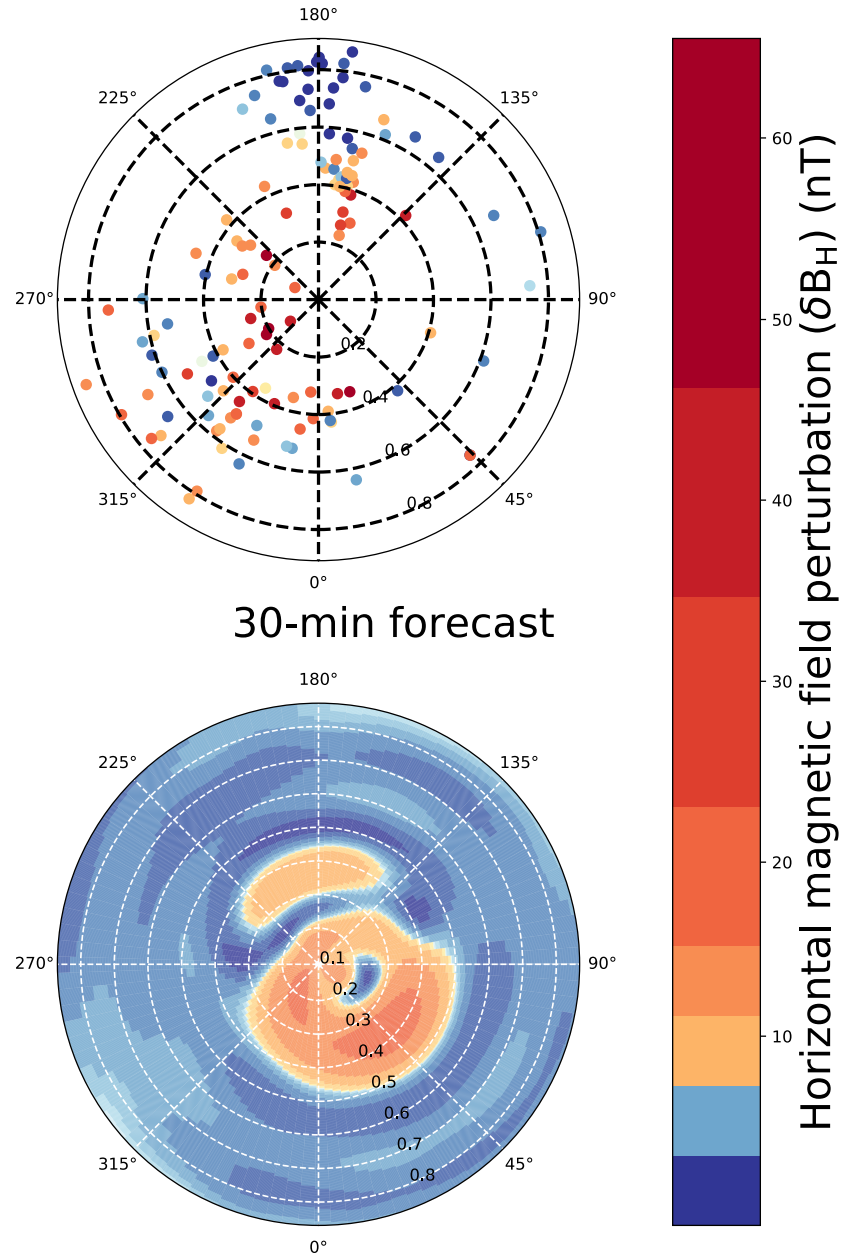


Figure 7. Maps of the measurement (top row) and forecast (bottom row) for the 2011 storm at times with minimum (left), mean (center) and maximum (right) mean absolute error (MAE).

where $\vec{a} = (a_i)$ is a vector of spherical harmonic coefficients, and $\mathcal{B} = (\vec{b}_i)$ is the basis matrix wherein column vector \vec{b}_i corresponds to the set of basis functions $Y_{nm}(\theta, \phi)$. The maximum harmonic degree, or the number of modes N is a free parameter, which is fixed by hyperparameter tuning. This is explained in Section 3.4.

We forecast both δb_e and δb_n in this work. Hence, we generate coefficients for both the parameters with the same code.

Map of SuperMag target and FDL forecast for timestep with mean MAE
Target SuperMag

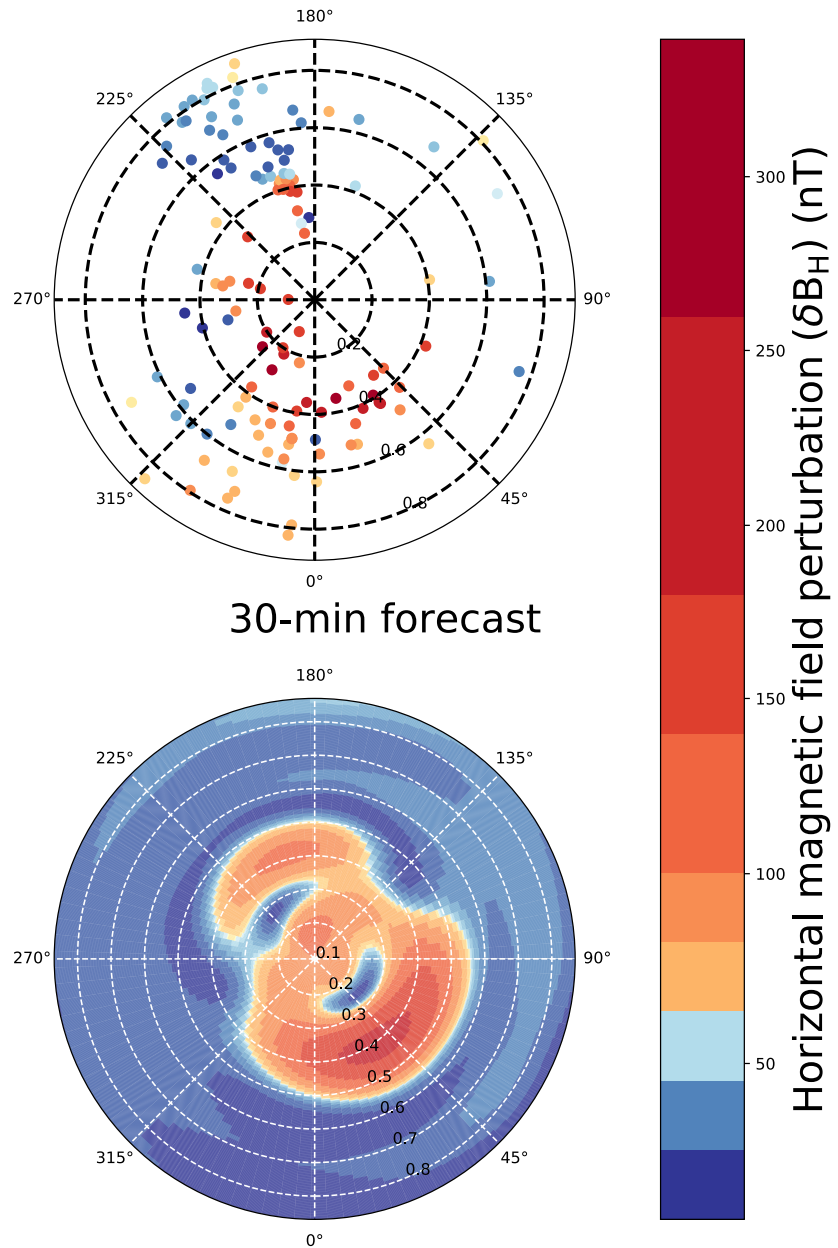


Figure 7. (Continued)

3.3.2. Model Architecture

We use a Gated Recurrent Unit (GRU) cell (Cho et al., 2014) as a time series summarizer. A GRU cell is a variant of the Recurrent Neural Network (Rumelhart et al., 1985). The GRU cell has an internal memory in the form of a “hidden state,” which is updated as inputs are given to it. This update happens through a sequence of nonlinear projection and shift operations (see Cho et al., 2014, for details). Thus, the input time series is used to update the hidden state, encoding the information content of the input time series.

We feed in the T hours of the solar wind measurements to the cell, which are summarized into a “hidden state” of the cell. Note again that this length of the time series is fixed through the hyperparameter search described in

Map of SuperMag target and FDL forecast for timestep with max MAE
Target SuperMag

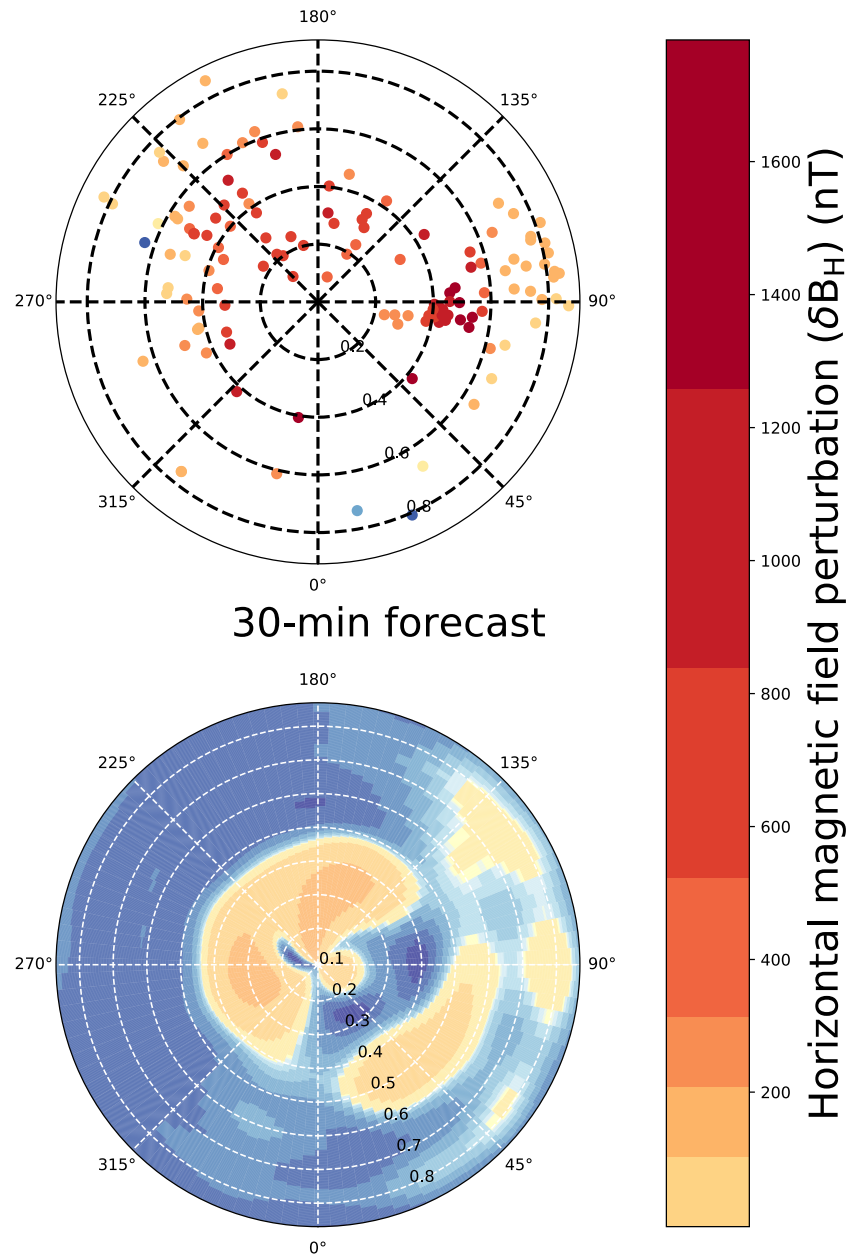


Figure 7. (Continued)

Section 3.4. The hidden state vector has a size of 8 units. This state vector acts as a proxy for all the solar wind information needed for our forecast.

The hidden state is then fed into a fully connected layer, which transforms the hidden state to a vector of coefficients. The number of coefficients is determined by the largest mode we seek to forecast from the code.

Finally, the output from the fully connected layer is then contracted with the spherical harmonic basis, giving out the forecast of perturbation measurements at any required spatial location. This basis, which enforces our GRU hidden state to be the spherical harmonic coefficients, is called the Spherical harmonic basis layer. Since the basis

Maps for timestep with min MAE Target SuperMag

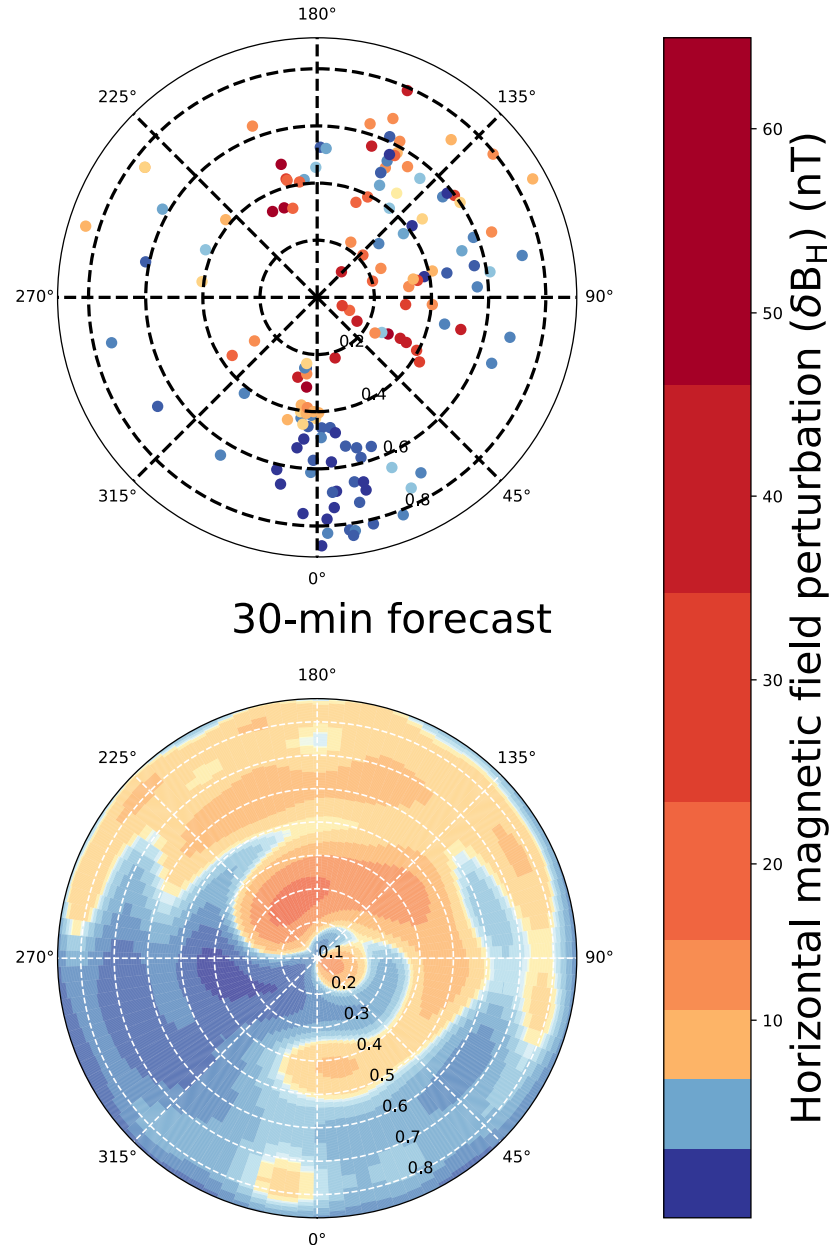


Figure 8. Same as Figure 7, but for the 2015 storm.

functions are computed using their analytical form (and not learned through data), the basis layer is not trainable. The model architecture is summarized in Figure 1, while the layer sizes are provided in Table 1.

The MLTs of various SuperMAG stations change with time. Hence, during training and inference time, the B are evaluated during every forward pass for the (MAGLAT,MLT) of the stations where the measurements are made. Hence, the spherical harmonic coefficients are constructed during each forward pass. Also note that the spherical harmonic formulation presented in Section 3.3.1 has the azimuth origin at the North Pole. Hence, we transform the MAGLAT into Magnetic colatitude.

Map of SuperMag target and FDL forecast for timestep with mean MAE
Target SuperMag

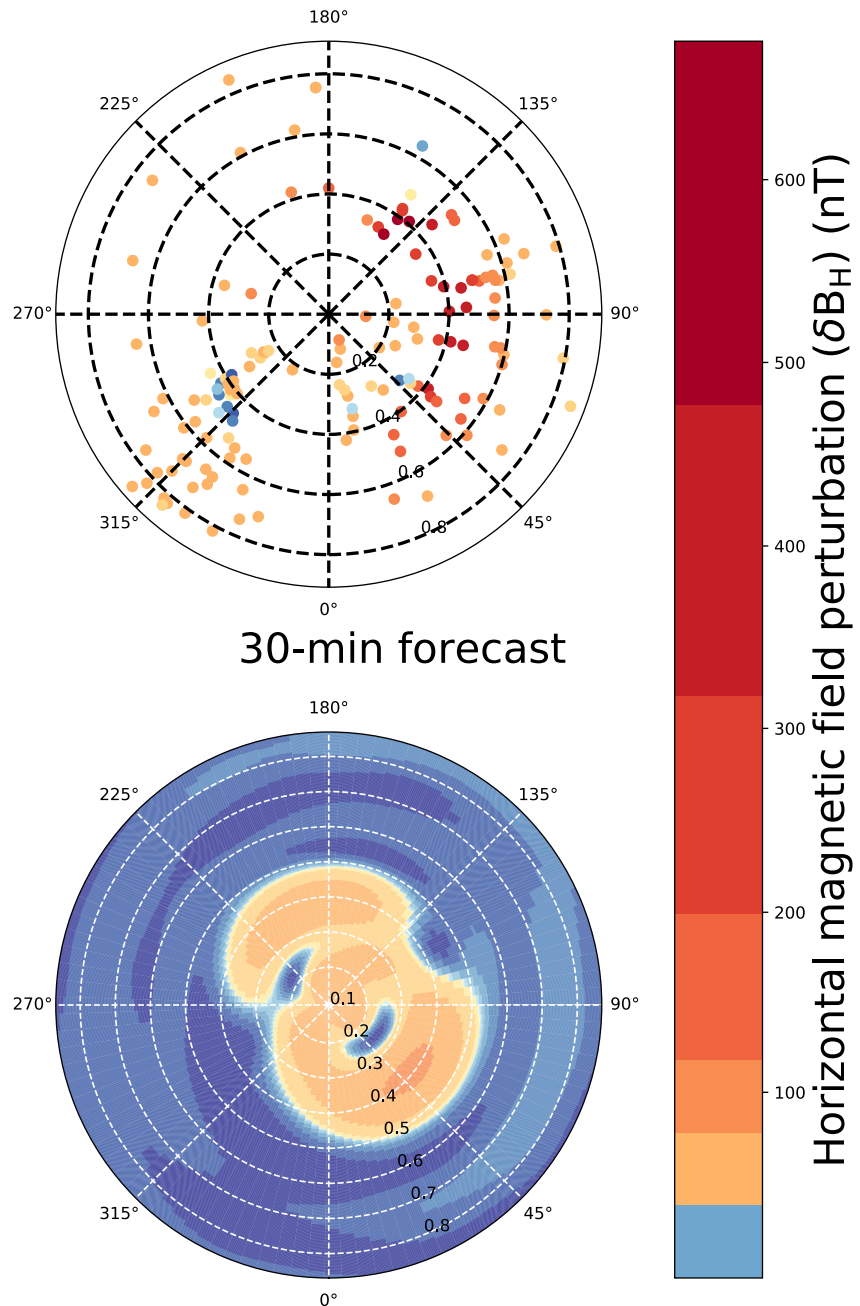


Figure 8. (Continued)

3.4. Hyperparameters

Deep Learning models generally have trainable parameters (which we shall henceforth call weights), and free parameters, which must be set manually (called hyperparameters). We monitor the validation set performance for different combinations of the hyperparameters, and use a Bayesian grid search to select the hyperparameters, which give the best validation set performance as the final model. A Bayesian grid search is a more informed search over a random search, which updates the next to-be-tested hyperparameter combination conditioned on the previous samples and validation set performance. We performed the hyperparameter search using Weights and

Map of SuperMag target and FDL forecast for timestep with max MAE
Target SuperMag

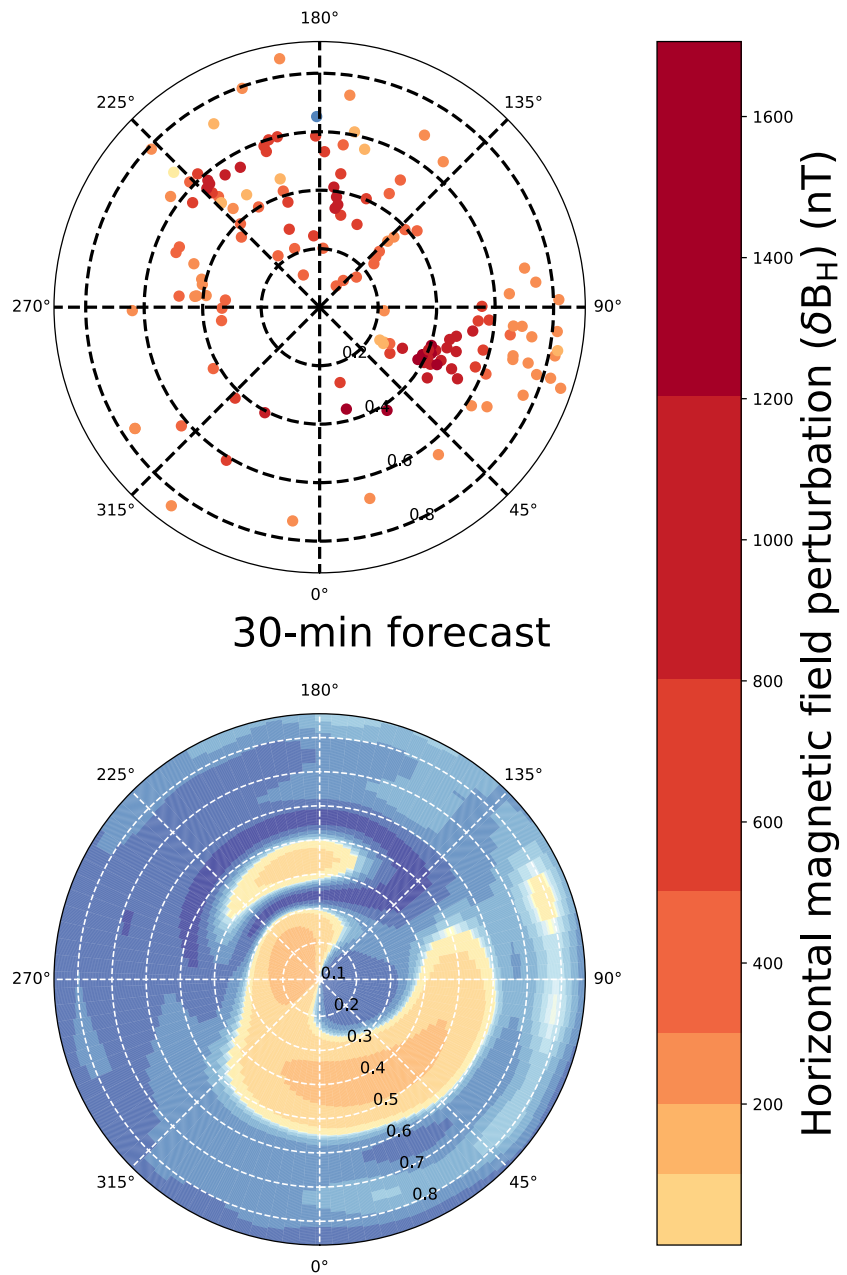


Figure 8. (Continued)

Biases (Biewald, 2020). The hyperparameters values are given in Table 2. The hyperparameter grid or bounds of the distributions are provided in Supporting Information S1.

With the model hyperparameters and architecture fixed, we train the model. We use the Mean Absolute Error (MAE) as the loss function to be optimized. The L2 regularization is a penalization term preventing the coefficients from growing too large. This penalty term serves the twofold benefit of preventing overfitting, and reducing sparsity amongst the coefficients. Since we would want as many harmonics to be captured as possible to better resolve local disturbances, we would want the “power” to be spread across as many modes as possible. Furthermore, we use dropouts (Srivastava et al., 2014) to randomly switch off neurons during the training time to enhance independent pathways within the model. This again serves to prevent overfitting in the model.

Table 4
Event-Based Metric Comparison of DAGGER With W2013 and Persistence Model, Summarized Across All Stations

Storm	Metric	DAGGER				Weimer				Persistence			
		18	42	66	90	18	42	66	90	18	42	66	90
2011	POD	0.62 ± 0.03	0.58 ± 0.03	0.42 ± 0.04	0.31 ± 0.04	0.17 ± 0.01	0.09 ± 0.01	0.05 ± 0.01	0.01 ± 0.01	0.56 ± 0.03	0.53 ± 0.02	0.38 ± 0.03	0.23 ± 0.03
	POFD	0.13 ± 0.01	0.07 ± 0.01	0.05 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.06 ± 0.01	0.03 ± 0.00	0.02 ± 0.00	0.01 ± 0.00
	PC	0.87 ± 0.01	0.92 ± 0.01	0.94 ± 0.01	0.97 ± 0.00	0.86 ± 0.01	0.94 ± 0.01	0.97 ± 0.00	0.98 ± 0.00	0.92 ± 0.01	0.95 ± 0.00	0.96 ± 0.00	0.98 ± 0.00
2015	HSS	0.37 ± 0.02	0.30 ± 0.02	0.22 ± 0.03	0.17 ± 0.03	0.18 ± 0.01	0.12 ± 0.01	0.06 ± 0.02	0.02 ± 0.01	0.47 ± 0.02	0.46 ± 0.02	0.32 ± 0.03	0.19 ± 0.02
	POD	0.69 ± 0.02	0.36 ± 0.03	0.15 ± 0.01	0.06 ± 0.01	0.11 ± 0.01	0.02 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.66 ± 0.02	0.48 ± 0.02	0.36 ± 0.02	0.30 ± 0.02
	POFD	0.57 ± 0.02	0.27 ± 0.02	0.14 ± 0.01	0.07 ± 0.01	0.04 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.21 ± 0.02	0.09 ± 0.01	0.05 ± 0.00	0.03 ± 0.00
	PC	0.66 ± 0.01	0.73 ± 0.02	0.83 ± 0.01	0.90 ± 0.01	0.68 ± 0.02	0.85 ± 0.01	0.92 ± 0.01	0.95 ± 0.00	0.85 ± 0.01	0.88 ± 0.01	0.92 ± 0.01	0.95 ± 0.00
	HSS	0.04 ± 0.01	0.00 ± 0.01	-0.02 ± 0.01	-0.03 ± 0.00	0.06 ± 0.01	0.01 ± 0.00	0.01 ± 0.00	-0.00 ± 0.00	0.41 ± 0.02	0.36 ± 0.01	0.28 ± 0.02	0.25 ± 0.02

Note. DAGGER, Deep leArminG Geomagnetic pErtuRbation; PC, Proportion Correct; POD, Probability of Detection; POFD, Probability of False Detection; HSS, Heidke Skill Score.

As already mentioned in Section 2, we train the model on ≈ 10 years of data and report the results below. We performed the training on an NVIDIA A100 GPU with 40 GB memory, with the model taking ≈ 40 hr for convergence.

4. Results

4.1. Testing Set

We report the results and performance of our model below. This is done in two ways: first, we report the statistics on the test set, and next we report the performance of our model for the two storm times described in Section 2. For both the data sets, we benchmark our model against the persistence model, while for the storm times, we also benchmark against **W2013**. Furthermore, we also report the event-based metrics for the two storm times, to enable comparison across other models and papers.

In Figure 2, we report the joint distributions of the forecast and the target δb_H for the test set (panel a), 2011 storm set (panel b), and 2015 storm set (panel c). Since there are a large number of points, the number of points in each bin is shown using the color bar. Note that the number points (and hence the color) scale logarithmically. The black line shows a slope = 1 line. For a perfect forecast, all points should lie on this line—however, this is seldom the case. From Figure 2, we see that by and large the model predictions and targets are aligned to the slope = 1 line. Furthermore, the 2011 storm is better forecasted than the 2015 storm. However, note that we can also see that **DAGGER** also has a tendency to “under-forecast”, since there exist more points below the line slope = 1 than above.

On the held-out testing set, we obtain an RMSE (MAE) of 35.28 (20.41) and 63.74 (39.36) nT for δb_e and δb_n , respectively. Hence, we see a clear effect of outlier datapoints in the computation of these metrics, resulting in higher value of RMSE over MAE. For the persistence model, we obtain an RMSE (MAE) of 26.46 (10.39) and 35.88 (13.63) nT for δb_n and δb_e , respectively. Thus, while our model shows low errors, it does not quite beat the persistence model in these metrics. Hence, significant “autocorrelation” of the perturbations seems to exist within a forecast horizon of 30 min, which results in the low RMSE and MSE of the persistence model.

However, RMSE and MAE do not quite give us any information regarding the temporal structure of the forecasts with respect to our measurements. Hence, we next validate our model performance across the two storm data sets. For these two storms, we have the nowcast from **W2013** and the persistence model to benchmark our performance.

4.2. Storm Time Performance

We now report the RMSE and MAE of our forecasts and the benchmarks, for the two storm data sets in Table 3. Note that the metrics are calculated across all times and all stations. From Table 3, we again clearly see the feature of larger RMSE over MAE due to outlier cases in the data set. The persistence model shows metrics only marginally better than **DAGGER** forecasts for the 2011 storm—infact, the RMSE in δb_e is lower for **DAGGER**. However, this is not the case for the 2015 storm. Hence, a 30 min time window still contains significant autocorrelation in the SuperMAG measurements, as we have also seen from the testing set results.

DAGGER clearly outperforms **W2013** in both the metrics for both the components of the horizontal magnetic field perturbation. Since the primary input features to our model are the same as those used by **W2013**, these results tell us that DL is able to capture a much more nonlinear association between the solar wind/IMF/solar flux and geomagnetic perturbation measurements.

To investigate “how good” RMSE and MAE are as metrics to quantify performance, we present the forecasts from our model and compare them with the measurements at different stations.

Table 5
Metric Comparison Between DAGGER, W2013, and Persistence Models for the OTT Station for the Years 2011 (Top Row) and 2015 (Bottom Row)

Metric		DAGGER		W2013		Persistence	
		2011	2015	2011	2015	2011	2015
MSE	δb_e	22.54	46.35	28.48	45.97	24.58	40.64
	δb_n	52.87	79.89	46.37	65.23	37.78	46.14
MAE	δb_e	14.68	34.18	22.51	32.33	14.63	26.12
	δb_n	32.55	57.68	35.40	54.30	20.13	27.50
18	POD	0.56	0.52	0.33	0.04	0.78	0.64
	POFD	0.14	0.42	0.00	0.04	0.03	0.08
	PC	0.84	0.57	0.95	0.82	0.95	0.88
	HSS	0.25	0.06	0.48	-0.00	0.68	0.55
42	POD	0.00	0.00	0.00	0.00	0.00	0.29
	POFD	0.00	0.07	0.00	0.01	0.01	0.04
	PC	0.99	0.89	0.99	0.95	0.98	0.93
	HSS	0.00	-0.06	0.00	-0.01	-0.01	0.21
66	POD	-	0.00	-	0.00	-	0.50
	POFD	0.00	0.01	0.00	0.00	0.00	0.01
	PC	1.00	0.98	1.00	0.99	1.00	0.99
	HSS	-	-0.01	-	0.00	-	0.49
90	POD	-	0.00	-	0.00	-	0.00
	POFD	0.00	0.01	0.00	0.00	0.00	0.01
	PC	1.00	0.99	1.00	0.99	1.00	0.98
	HSS	-	-0.01	-	0.00	-	-0.01

Note. Blanks (-) denote metrics that are unavailable due to the denominator in the metric definition going to 0. The full table for all stations is given in Supporting Information S1. DAGGER, Deep leArninG Geomagnetic pErtuRbation; PC, Proportion Correct; POD, Probability of Detection; POFD, Probability of False Detection; HSS, Heidke Skill Score; MAE, Mean Absolute Error.

Since there are ≤ 175 stations in our data set, we present results for the “best” and “worst” forecasted stations. To this end, we select forecasts for 3 stations, which show the smallest and largest MAE. Every other forecast would lie somewhere between the best and worst case scenarios. These forecasts for the two storms, along with IMF B_z and Sym-H indices are shown in Figures 3 and 5 for the two storms.

From Figure 3, we see that the dichotomy between the best and worst performing forecast is quite stark. First, we clearly see that the forecasts deemed “best” (left column) correspond to stations where the measurements are ≤ 250 nT. Second, **DAGGER** is able to clearly pick out the different peaks and troughs of the forecast—especially for the stations with the lowest MAE (panel. c). Third, the perturbation forecast and measurement values are of similar magnitudes, and in many cases match well for the stations that show the lowest errors. On the other hand, prima facie it looks like **DAGGER** is unable to forecast anything at all for the stations with large MAE. Clearly, the largest perturbation measurements from these stations are $\approx 6\times$ the largest perturbations for the stations showing low MAE. Since some salient variations seem to be captured by **DAGGER** (see panels f and h), we define different Y -axes for the forecast and measurement, to probe how good (or bad) **DAGGER** forecasts for these stations in Figure 4.

From Figure 4, two inferences may be made. First, **DAGGER** is able to forecast the variation of perturbation over time even for stations which have a large associated MAE. And second, **DAGGER** underpredicts the large perturbation values, which hence gives rise to a large MAE. Thus, a purely DL framework is able to assimilate the solar wind measurements and generate salient associations with the magnetic field perturbations. The exact scale of perturbations is however missed for the stations with large associated MAE.

These results are also clearly seen in Figure 5 for the 2015 storm data set. The stations having low MAE typically have a max perturbation of ≈ 300 nT, while the stations with the largest MAE are $\approx 6\times$ larger. One interesting result to be noticed for the stations with the lowest MAE for this storm is the mismatch between forecast and measurement is larger for the 2015 storm than in the 2011 storm case (compare panels c, e, and g between Figures 3 and 5). This is also consistent with larger spread in the joint histograms in

Figure 2c. To see if this is also observed for the stations with large MAE, we check the forecast and measurements on different Y -axis scales in Figure 6.

From Figure 6, we once again see that **DAGGER** is able to capture salient variation of the perturbation measurements but fails to reproduce the exact values. However, both the lowest and largest MAE for the 2015 storm are larger than those for the 2011 storm. From Figure 5a (and also from Section 2.3), we see that the 2015 storm measurements have a lot of data gaps. This is not seen for the 2011 storm (see Figure 3a). Hence, we speculate that the larger MAE for the 2015 storm arises from a lack of data (which may also depend on the imputation scheme), resulting in spurious forecast when the solar wind data is missing.

In Figure 7 and 8, we show the maps for δb_H (forecast in the bottom row, perturbations in the top) in the MLT-MCO-LAT grid, with the center being the North Pole. This is done for three cases—this time, for the time step with minimum, mean, and maximum MAE across all stations. From these plots, we clearly see that our model provides a dynamic map of the perturbations, at a cadence of 1 min. Furthermore, it also shows how underprediction gives rise to the larger MAEs. Thus, such perturbation maps for δb_n , δb_e and δb_H , changing dynamically over time scale of ≈ 1 min are made available across the two storm times as video files in the online version of the paper.

Finally, we clearly see that MAE (or MSE) are good detectors of magnitude-match of the forecast with the measurements but cannot pick out if the variations are captured with specific thresholds. Thus, we also present the event-based metrics as a measure of **DAGGER** performance in Table 4.

In Table 4, the metrics are computed for each station, and we report the mean and standard errors across all the stations. The standard error is defined as σ/\sqrt{N} , where σ is the standard deviation of metric across all stations, and N is the number of stations. The standard error reflects the uncertainty in the estimation of mean value reported in Table 4. Note that while we present the mean and uncertainty in the metric in Table 4, we report the metrics for all the stations in Supporting Information S1.

From Table 4, we first compare the performance for the 2011 storm. We see that the metrics for all the models reduce as a function of the selected threshold. First, **DAGGER** shows a larger POD than either **W2013** or the Persistence model, implying many of the events are detected well by **DAGGER**. This is in-line with **DAGGER** being able to capture the variation in peaks of the measurements well.

Next, we find that **DAGGER** shows larger POFD when compared to **W2013** or Persistence for a threshold of 18 nT/min. However, the POFD becomes small and consistent with the benchmarks for larger thresholds.

Third, we find that the PC from **DAGGER** are consistent with those from **W2013**, irrespective of the threshold value chosen. Since the POD of **DAGGER** is larger than **W2013**, this means that there are far more nonevent cases, which are captured well enough by both the models. However, the persistence model has a larger PC, which again indicates some true negatives being missed out by **DAGGER**.

Finally, **DAGGER** HSS are larger than **W2013** but smaller than the persistence model. This tells us that proportion of correct forecasts by **DAGGER** are significantly better informed than those from **W2013**. However, the forecasting horizon contains enough autocorrelation in the SuperMAG time series to give rise to a good fraction of nonrandom correct proportion of events. The HSS between **DAGGER** and Persistence become consistent only at a threshold of 90 nT/min, indicating that the large events are not very persistent, and this information from the solar wind is captured by **DAGGER**.

Interestingly, for the 2015 storm, all of our metrics—both for **DAGGER** and **W2013** are worse than the persistence model. **DAGGER** shows better metric performance when compared to **W2013**, and shows only marginally better (or worse) metrics (except HSS) when compared to persistence. However, the HSS indicates that both **DAGGER** and **W2013** are no better than a random model generating 1 and 0 s, which is consistent with the performance of similar RNN-based models (Keese et al., 2020).

Since both **DAGGER** and **W2013** do not give as good a performance as the persistence model, we get further evidence of the strong influence of missing OMNI data in giving rise to the poor performance of OMNI-based forecasting schemes.

5. Summary and Conclusion

Accurate global forecasts of geomagnetic perturbations are extremely important from the perspective of both disruptions due to GICs and to understand the modulation of Earth's global magnetic field due the streaming solar wind.

To this end, we develop a global magnetic field perturbation forecasting model in this work. The model, named **DAGGER**, has three components: a time series summarizer, a coefficient generator, and a spherical harmonic constructor. The time series summarizer takes in a time series of solar wind, IMF and solar radio flux, and generates a summary state across all variables and time. This summary state is transformed nonlinearly by a fully connected layer to generate a vector of coefficients. Finally, the spherical harmonic layer contracts with this coefficient layer, and generates perturbation forecasts at different locations on the Earth.

We find that the **DAGGER** is able to clearly capture the temporal variations of the perturbations. However, it underpredicts the perturbation values if they are $\approx 1,000$ nT, resulting in large pointwise errors. Note that **DAGGER** is trained predominantly during quiet times—since 2.1% of data consists of a SYM-H index of less

than -50 nT. Thus, the results may potentially be biased toward more quiet time than active times, resulting in underprediction of perturbations.

We benchmark our model against **W2013** and a 30 min persistence model using various metrics. We find that **DAGGER** clearly outperforms **W2013** in all metrics. However, **DAGGER** shows comparable (or slightly worse) performance than a persistence model on MAE and RMSE. On the event-based metrics, **DAGGER** shows either consistent, or worse performance than the persistence model. Clearly, a persistence model seems to possess an advantage over both **DAGGER** and **W2013**. However, the persistence can be computed only for individual stations and lacks the spatial coverage, which **DAGGER** provides.

Regardless of the exact performance measure, our results show that DL is able to capture associations between changes in the solar wind/interplanetary medium, and the Earth's magnetosphere. However, the magnetosphere seems to have enough memory over 30 min for a persistence model to show good enough performance. For the 2015 storm, **DAGGER** and **W2013** show much worse performance than the persistence model—this is not the case for the 2011 storm, where the performance is comparable. Since both **DAGGER** and **W2013** show the drastic reduction in performance on the 2015 storm, the underlying reason seems to be the missing values in the OMNI data set, resulting in a noisy input to the model. However, similarly in metrics for both the 2011 and 2015 storm for the Persistence model tells us that **DAGGER** can, in principle, perform well, given good data.

Also note that the time scale of 120 min is interestingly of the order of time to transfer information from dayside and nightside reconnection sites in the magnetosphere to the ionosphere system especially for higher latitude $\geq 40^\circ$ (Coxon et al., 2019). However, we may only speculate, and not claim an exact connection at this stage.

Our results may be compared across literature with models that benchmark on the two storms, at similar cadence. Models by Keese et al. (2020) consider the solar wind and IMF parameters as inputs, and output the geomagnetic perturbations at the Ottawa station (OTT). However, note that while **DAGGER** has a forecast horizon of 30 min (over and above the lag between OMNI and SuperMAG), such a lag is not present in Keese et al. (2020).

We can first compare the average metrics from **DAGGER** with the metrics provided by Keese et al. (2020) (compare Table 4 of this paper with Table 1 of Keese et al. (2020)). For the 2011 storm, **DAGGER** clearly outperforms the LSTM model of Keese et al. (2020) in all metrics except POFD. Similarly, **DAGGER** shows better performance than the ANN model for POD and PC, while the performance is consistent in HSS and slightly worse in POFD. For the 2015 storm, **DAGGER** shows better performance in POD, marginally worse performance in PC and HSS, while far many false detections are made by **DAGGER** for a threshold of 18 nT/min. This clearly seems to be a manifestation of the missing data and the imputation scheme deployed to tackle it. Hence, prima facie, it seems that linear interpolation is a much better imputation scheme than zeroing of inputs. Note, however, that **DAGGER** forecasts are not confined to any particular station and generates maps of forecasts.

Next, we may also pick out the specific metrics for the OTT station (presented in Table 5), and compare them with the two models of Keese et al. (2020). Here, we find for the 2011 storm that while **DAGGER** shows better performance than both the models of Keese et al. (2020) in POD and PC, the performance is marginally worse in POFD and HSS. However, note that **DAGGER** HSS is more than (or even similar to) the LSTM model of Keese et al. (2020), while it is lower than the ANN model. For the 2015 storm, **DAGGER** outperforms both the models in POD, while the performance is marginally worse in PC, POFD, and HSS. This is consistent with the average performance across all stations, and seems to again point toward a dependence on the data imputation scheme.

We also compare our result with Pulkkinen et al. (2013) study but not for a particular station or event. In general, none of the models in the community, including first-principle and empirical, can capture the high dB/dt (1.5 nT/s or 90 nT/min) threshold. This behavior is very important while forecasting particularly strong spaceweather events. If the mitigation of a storm depends on a model forecast, underprediction of the perturbation magnitude would pose a significant problem. These models are not able to reproduce point-by-point fluctuations of perturbation due to the complex waveform of the perturbation signal. Hence, this is an important issue, which would need to be mitigated in the future.

It is important to note the various caveats associated with this work. The first, and the most obvious issue is of the missing data. We have imputed the missing data with 0 s. While this is a simplistic scheme of imputation, we did not perform any interpolation as we did not find any well-motivated reason to induce artificial variations in

the data. However, addressing this issue with complex imputation schemes is far too complicated, and beyond the scope of this work.

Next, we see that the **DAGGER** forecasts follow the variations in SuperMAG measurements well but do not reproduce the exact values when the perturbations are large. The fact that **DAGGER** captures the variations but not the exact magnitude, seems to arise from a lack of “context” perturbation measurements. One can perform a nonlinear rescaling (see, e.g., Camporeale et al., 2020) to circumvent this issue. However, our overarching aim is to have a rather more self-consistent model avoiding any ad hoc scaling as much as possible. In principle, we can incorporate a proxy for the state of the Earth's magnetosphere as an input to the model. This would help provide “context” to the forecasting model, and may help it give the correct perturbation values (and not just capture the variations). Incorporating geomagnetic indices have been shown to improve the quality of magnetospheric forecasts (see, e.g., Smith et al., 2020). However, capturing a summary of the magnetosphere, given changing stations across multiple MLT and MAGLAT, is nontrivial and is a work for the future.

Third, our method assumes a smooth, continuous, and differentiable perturbation field, with power distributed amongst different modes. Furthermore, we truncate the spherical harmonics at a maximum mode due to operational constraints and hyperparameter selection. While these assumptions are physically motivated, their effect is to impose a “smooth” reconstruction, which may prevent capture of localized large peaks in data across a set of (MLT, MAGLAT). Similarly, since we have truncated the spherical harmonics at a maximum number of modes, we expect the highest frequency mode to be translated to the shortest length scale that our workflow can resolve. Hence, **DAGGER** will not be able to—in the current formulation—resolve fluctuations shorter than this “threshold” length scale. Note further that the shortest length scales of importance would also depend on local ionospheric current and local geology. We, however, expect these scales to be much smaller than the length scale corresponding to the highest harmonic mode considered (Beggan, 2015; Pulkkinen et al., 2015).

The spherical harmonic formulation performs an instantaneous decomposition of the field over the globe. However, the whole system—as a sphere—evolves dynamically over time. Hence, information propagation across different stations takes time, which must be incorporated in the basis matrix formulation itself. While this is beyond the scope of the paper, such a path is a potential future work for improvement.

Also, note that **DAGGER** does not yet provide uncertainty estimates on the perturbation forecasts. The uncertainty estimates both provide a degree of confidence, and also inform us of ill-constrained regions of forecasts. Thus, such uncertainties may provide us with means of diagnosing the most optimum location of stations to (a) reduce uncertainty and (b) optimize the number of stations.

Finally, we emphasize that the codebase and the proposed model **DAGGER** are general enough to be suited for forecasting fields on any spherically-symmetric systems. A direct application of **DAGGER** would be transfer-learn the perturbation forecasts to magnetic field perturbation measurements in other planets. This is useful from both a spacecraft navigation and a science measurement perspective to gather data pertaining to specific locations as a study of planetary magnetospheres.

Data Availability Statement

Our model outputs are agnostic to the grid on which the basis is defined. Hence, the coefficients may be contracted with an appropriate basis to generate full-Earth maps of perturbation forecasts. The SuperMag data are available online (<https://supermag.jhuapl.edu/>), and so is the case with OMNI (https://omniweb.gsfc.nasa.gov/form/omni_min.html). To further reproducible research, and foster innovation with modeling schemes, we are making our codebase and models open source at Upendran et al. (2022). For getting researchers started with using our code, a tutorial notebook as a part of SpaceML (Koul et al., 2020) is available at <https://spaceml.org/repo/project/60c0a78d4ba8cb0012611ad4>. Our model is built in PyTorch (Paszke et al., 2019), PyTorch-lightning (Falcon et al., 2019), and Sympy (Meurer et al., 2017). We also use Numpy (Harris et al., 2020), Scikit-learn (Pedregosa et al., 2011), and Scipy (Virtanen et al., 2020) for analysis, Dask (Dask Development Team, 2016) and Pandas (development team, 2020) for data processing, and Matplotlib (Hunter, 2007), Cartopy (Met Office, 2010–2015) for plotting.

Acknowledgments

This research was conducted as a part of research sprint at Frontier Development Lab (FDL). We would like to thank Search for Extraterrestrial Intelligence, Google Cloud, Nvidia, and Lockheed Martin for funding the FDL program. We thank the SuperMAG collaboration for making the perturbation data available for the public. We also sincerely thank Prof. Daniel Weimer for providing the storm time forecasts, which served as excellent benchmarks. We thank the two reviewers for their very useful comments. U.V would like to acknowledge the Max Planck Partner Grant of Prof. Durgesh Tripathi of Inter University Centre for Astronomy and Astrophysics, Pune, for providing compute facility, and the Nvidia Academic Hardware program grant to U.V & Durgesh Tripathi. B.F was supported by NASA grant NNH19ZDA001N-LWS. P.T. is supported by the UK EPSRC CDT in Autonomous Intelligent Machines and Systems (grant reference EP/L015897/1).

References

- Barlow, W. H., Barlow, P., & Culley, R. S. (1849). Vi. on the spontaneous electrical currents observed in the wires of the electric telegraph. *Philosophical Transactions of the Royal Society of London*, 139, 61–72. <https://doi.org/10.1098/rstl.1849.0006>
- Beggan, C. D. (2015). Sensitivity of geomagnetically induced currents to varying auroral electrojet and conductivity models. *Earth Planets and Space*, 67(1), 1–12. <https://doi.org/10.1186/s40623-014-0168-9>
- Biewald, L. (2020). *Experiment tracking with weights and biases*. Retrieved from <https://www.wandb.com/> (Software available from wandb.com)
- Boteler, D. H. (2001). Space weather effects on power systems. In *Washington DC American Geophysical Union Geophysical Monograph Series* (Vol. 125, pp. 347–352). <https://doi.org/10.1029/GM125p0347>
- 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*, 125(11), e27684. <https://doi.org/10.1029/2019JA027684>
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). *Learning phrase representations using RNN encoder-decoder for statistical machine translation*. arXiv preprint arXiv:1406.1078. <https://doi.org/10.3115/v1/d14-1179>
- Clette, F. (2021). Is the $F_{10.7cm}$ - sunspot Number relation linear and stable? *Journal of Space Weather and Space Climate*, 11, 2. <https://doi.org/10.1051/swsc/2020071>
- Coxon, J. C., Shore, R. M., Freeman, M. P., Fear, R. C., Browett, S. D., Smith, A. W., et al. (2019). Timescales of birkeland currents driven by the IMF. *Geophysical Research Letters*, 46(14), 7893–7901. <https://doi.org/10.1029/2018gl081658>
- Dask Development Team. (2016). *Dask: Library for dynamic task scheduling*. Computer software manual. Dask. Retrieved from <https://dask.org/>
- development team, T. P. (2020). pandas-dev/pandas: Pandas. *Zenodo*.
- Eastwood, J., Hapgood, M., Biffis, E., Benedetti, D., Bisi, M., Green, L., et al. (2018). Quantifying the economic value of space weather forecasting for power grids: An exploratory study. *Space Weather*, 16(12), 2052–2067. <https://doi.org/10.1029/2018sw002003>
- Falcon, W. (2019). *Pytorch lightning*. *GitHub*. Note. Retrieved from <https://github.com/PyTorchLightning/pytorch-lightning>
- Gjerloev, J. (2012). The supermag data processing technique. *Journal of Geophysical Research*, 117(A9). <https://doi.org/10.1029/2012ja017683>
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- 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, 72. <https://doi.org/10.3389/fspas.2020.550874>
- King, J. H., & Papitashvili, N. E. (2005). Solar wind spatial scales in and comparisons of hourly wind and ace plasma and magnetic field data. *Journal of Geophysical Research*, 110(A2), A02104. <https://doi.org/10.1029/2004JA010649>
- Koul, A., Ganju, S., Kasam, M., & Parr, J. (2020). *SpaceML: Distributed open-source research with citizen scientists for the advancement of space technology for NASA*. arXiv e-prints, arXiv:2012.10610.
- Kozyreva, O. V., Pilipenko, V. A., Belakhovsky, V. B., & Sakharov, Y. A. (2018). Ground geomagnetic field and GIC response to march 17, 2015, storm. *Earth Planets and Space*, 70(1), 1–13. <https://doi.org/10.1186/s40623-018-0933-2>
- Lamb, K., Malhotra, G., Vlontzos, A., Wagstaff, E., Günes Baydin, A., Bhiwandiwala, A., et al. (2019). *Correlation of auroral dynamics and GNSS scintillation with an autoencoder*. arXiv e-prints, arXiv:1910.03085.
- Lanzerotti, L. J. (2001). Space weather effects on technologies. In *Washington DC Geophysical Union Geophysical Monograph Series* (Vol. 125, pp. 11–22). <https://doi.org/10.1029/gm125p0011>
- Met Office. (2010–2015). *Cartopy: A cartographic python library with a matplotlib interface*. Computer software manual. cartopy. Retrieved from <http://scitools.org.uk/cartopy>
- Meurer, A., Smith, C. P., Paprocki, M., Čertík, O., Kirpichev, S. B., Rocklin, M., et al. (2017). Sympy: Symbolic computing in python. *PeerJ Computer Science*, 3, e103. <https://doi.org/10.7717/peerj-cs.103>
- Ngwira, C. M., Sibeck, D., Silveira, M. V., Georgiou, M., Weyand, J. M., Nishimura, Y., & Hampton, D. (2018). A study of intense local d/b/d variations during two geomagnetic storms. *Space Weather*, 16(6), 676–693. <https://doi.org/10.1029/2018sw001911>
- Oughton, E. J., Skelton, A., Horne, R. B., Thomson, A. W. P., & Gaunt, C. T. (2017). Quantifying the daily economic impact of extreme space weather due to failure in electricity transmission infrastructure. *Space Weather*, 15(1), 65–83. <https://doi.org/10.1002/2016SW001491>
- Papitashvili, N., Bilitza, D., & King, J. (2014). OMNI: A description of near-earth solar wind environment. In *40th COSPAR scientific assembly* (Vol. 40, pp. 1–14).
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32, pp. 8024–8035).
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pulkkinen, A., Bernabeu, E., Eichner, J., Viljanen, A., & Ngwira, C. (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., 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>
- Pulkkinen, A., Viljanen, A., Pajunpää, K., & Pirjola, R. (2001). Recordings and occurrence of geomagnetically induced currents in the Finnish natural gas pipeline network. *Journal of Applied Geophysics*, 48(4), 219–231. [https://doi.org/10.1016/S0926-9851\(01\)00108-2](https://doi.org/10.1016/S0926-9851(01)00108-2)
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation*. (Tech. Rep.). California Univ San Diego La Jolla Inst for Cognitive Science.
- Schrijver, C. J., Dobbins, R., Murtagh, W., & Petrinec, S. M. (2014). Assessing the impact of space weather on the electric power grid based on insurance claims for industrial electrical equipment. *Space Weather*, 12(7), 487–498. <https://doi.org/10.1002/2014SW001066>

- Smith, A. W., Rae, I. J., Forsyth, C., Oliveira, D. M., Freeman, M. P., & Jackson, D. R. (2020). Probabilistic forecasts of storm sudden commencements from interplanetary shocks using machine learning. *Space Weather*, *18*(11), e2020SW002603. <https://doi.org/10.1029/2020SW002603>
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.
- Tapping, K. F. (2013). The 10.7cm solar radio flux (f10.7). *Space Weather*, *11*(7), 394–406. <https://doi.org/10.1002/swe.20064>
- Tóth, G., Sokolov, I. V., Gombosi, T. I., Chesney, D. R., Clauer, C. R., de Zeeuw, D. L., et al. (2005). Space weather modeling framework: A new tool for the space science community. *Journal of Geophysical Research*, *110*(A12), A12226. <https://doi.org/10.1029/2005JA011126>
- Tóth, G., van der Holst, B., & Huang, Z. (2011). Obtaining potential field solutions with spherical harmonics and finite differences. *Acta Pathologica Japonica*, *73*(2), 102. <https://doi.org/10.1088/0004-637X/73/2/102>
- Tóth, G., van der Holst, B., Sokolov, I. V., DeZeeuw, D. L., Gombosi, T. I., Fang, F., et al. (2012). Adaptive numerical algorithms in space weather modeling. *Journal of Computational Physics*, *231*(3), 870–903. <https://doi.org/10.1016/j.jcp.2011.02.006>
- UN. (2017). *United Nations Office of outer space affairs, International spaceweather Initiative*. Retrieved from <https://www.unoosa.org/oosa/en/ourwork/psa/bssi/iswi.html>
- Upendran, V., Cheung, M. C. M., Hanasoge, S., & Krishnamurthi, G. (2020). Solar wind prediction using deep learning. *Space Weather*, *18*(9), e02478. <https://doi.org/10.1029/2020SW002478>
- Upendran, V., Tigas, P., Ferdousi, B., Bloch, T., Cheung, M. C. M., Ganju, S., et al. (2022). *Vishal-upendran/geoeffnetnet-1: Dagger model*. Zenodo.
- Verbanac, G., Manda, M., Vršnak, B., & Sentic, S. (2011). Evolution of solar and geomagnetic activity indices, and their relationship: 1960 – 2001. *Solar Physics*, *271*(1–2), 183–195. <https://doi.org/10.1007/s11207-011-9801-y>
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- Weigel, R. S., Vassiliadis, D., & Klimas, A. J. (2002). Coupling of the solar wind to temporal fluctuations in ground magnetic fields. *Geophysical Research Letters*, *29*(19), 1–4. <https://doi.org/10.1029/2002GL014740>
- Weimer, D., Clauer, C., Engebretson, M., Hansen, T., Gleisner, H., Mann, I., & Yumoto, K. (2010). Statistical maps of geomagnetic perturbations as a function of the interplanetary magnetic field. *Journal of Geophysical Research*, *115*(A10). <https://doi.org/10.1029/2010JA015540>
- Weimer, D. R. (2013). An empirical model of ground-level geomagnetic perturbations. *Space Weather*, *11*(3), 107–120. <https://doi.org/10.1002/swe.20030>
- Welling, D. T., Anderson, B. J., Crowley, G., Pulkkinen, A. A., & Rastätter, L. (2017). Exploring predictive performance: A reanalysis of the geospace model transition challenge. *Space Weather*, *15*(1), 192–203. <https://doi.org/10.1002/2016SW001505>
- Welling, D. T., Ngwira, C. M., Opgenoorth, H., Haiducek, 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>
- 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. <https://doi.org/10.1051/swsc/2015008>