



## Multi-Hour-Ahead Dst Index Prediction Using Multi-Fidelity Boosted Neural Networks

A. Hu<sup>1</sup> , E. Camporeale<sup>1,2</sup> , and B. Swiger<sup>1,2</sup> 

<sup>1</sup>CIRES, University of Colorado, Boulder, CO, USA, <sup>2</sup>NOAA Space Weather Prediction Center, Boulder, CO, USA

### Special Section:

Machine Learning in Heliophysics

### Key Points:

- A new multi-hour ahead Dst prediction model developed from solar wind observations using Gated Recurrent Unit networks is proposed
- The uncertainty of the proposed Dst model is estimated by applying the Accurate and Reliable Uncertainty Estimate method
- A multi-fidelity method is developed to boost the performance of the model

### Correspondence to:

A. Hu,  
[andong.hu@colorado.edu](mailto:andong.hu@colorado.edu)

### Citation:

Hu, A., Camporeale, E., & Swiger, B. (2023). Multi-hour-ahead Dst index prediction using multi-fidelity boosted neural networks. *Space Weather*, 21, e2022SW003286. <https://doi.org/10.1029/2022SW003286>

Received 12 SEP 2022  
Accepted 27 FEB 2023

### Author Contributions:

**Conceptualization:** E. Camporeale  
**Formal analysis:** A. Hu  
**Methodology:** A. Hu  
**Project Administration:** E. Camporeale  
**Software:** A. Hu  
**Supervision:** E. Camporeale  
**Validation:** A. Hu  
**Visualization:** A. Hu  
**Writing – original draft:** A. Hu  
**Writing – review & editing:** E. Camporeale, B. Swiger

**Abstract** The Disturbance storm time (Dst) index has been widely used as a proxy for the ring current intensity, and therefore as a measure of geomagnetic activity. It is derived by measurements from four ground magnetometers in the geomagnetic equatorial region. We present a new model for predicting Dst with a lead time between 1 and 6 hr. The model is first developed using a Gated Recurrent Unit (GRU) network that is trained using solar wind parameters. The uncertainty of the Dst model is then estimated by using the Accurate and Reliable Uncertainty Estimate method (Camporeale & Carè, 2021, <https://doi.org/10.1615/int.j.uncertaintyquantification.2021034623>). Finally, a multi-fidelity boosting method is developed in order to enhance the accuracy of the model and reduce its associated uncertainty. It is shown that the developed model can predict Dst 6 hr ahead with a root-mean-square-error of 13.54 nT. This is significantly better than a persistence model or a single GRU model.

**Plain Language Summary** Geomagnetic storms pose one of the most severe space weather risks to our space-borne and ground-based electronic instruments, such as global navigation satellite systems and radio transmission systems. The Disturbance storm time (Dst) is one of the most used geomagnetic storm indicators. This study presents an innovative multi-fidelity boosted neural network method to forecast Dst 1-to-6 hours ahead. The new method improves the performance of the predictions by estimating their uncertainties.

## 1. Introduction

The Disturbance storm time (Dst) is a geomagnetic index related to the perturbation of the geomagnetic field at low latitudes (Burton et al., 1975; Rostoker, 1972). Currently, Dst is defined by using geomagnetic field measurements from four equatorial ground magnetometers: Hermanus, Honolulu, San Juan and Kakioka (Sugiura & Kamei, 1991). Dst has been widely used for monitoring geomagnetic storms, which pose one of the most severe space weather risks to our space-borne and ground-based electronic instruments, such as global navigation satellite systems and radio transmission systems (Z. Li et al., 2021; Wan et al., 2021).

A short-term prediction of Dst is produced operationally at the NOAA Space Weather Prediction Center by means of a physics-based model (the Space Weather Modeling Framework developed at the University of Michigan, (Tóth et al., 2005)). A longer lead-time operational Dst forecast is provided by Space Environment Technologies, using the Anemomilos model (Tobiska et al., 2013). Dst is also used as an essential input for forecasting thermosphere mass density and ionospheric parameters, and to parameterize several empirical models. The following is a non-exhaustive list of models that use Dst as one of their inputs: O'Brien and Moldwin (2003) empirically estimate the location of the plasmopause; Agapitov et al. (2015) derive a statistical model for the lower band chorus distribution; S. Li et al. (2016) estimate the ionospheric global electron content storm-time response; Boardsen et al. (2000) derive an empirical model of the high-latitude magnetopause; and, Zhao et al. (2018) derive a model of radiation belt electron pitch-angle distributions.

A large amount of literature has been devoted to Dst prediction, notably using empirical and machine learning techniques (Camporeale, 2019). Lundstedt et al. (2002) first attempted to implement a multi-layer perception neural network to forecast Dst 1 hour ahead using interplanetary magnetic field (IMF) data. Several researchers presented models to extend the Dst forecast up to 6 hr in advance (Bala & Reiff, 2012; Lazzús et al., 2017; Saiz et al., 2008). A Gaussian Process model was introduced by Chandorkar et al. (2017) and combined with a long short-term memory (LSTM) architecture in Gruet et al. (2018) to provide probabilistic predictions up to 6 hours in advance. An ensemble learning algorithm was applied by Xu et al. (2020) for the same purpose. Laperre et al. (2020) evaluated the performance of a LSTM model based on a Dynamical Time Warping metric as the cost function.

© 2023. The Authors.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs License](https://creativecommons.org/licenses/by/4.0/), which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

**Table 1**  
*List of Variables Used to Train a Dst Forecast Model*

Reference	Variable	Symbol
Gruet et al. (2018)	Proton density	$n$
	Solar wind velocity	$V$
	Norm of IMF magnetic field vector	$ B $
	IMF magnetic field vector in $z$ direction	$B_z$
	Dst	Dst
Weimer (2013)	Clock angle from tangential IMF in the GSM Y-Z plane	$\theta_c$
	Dipole tilt angle	<i>dipole</i>

*Note.* The most recent 6 hr of each variable is used for training.

In this study, we first train a model using a machine learning (ML) technique called Gated Recurrent Unit (GRU), which is a flavor of a recurrent neural network (RNN, Hu and Zhang (2018)), to forecast Dst during strong storm periods (Dst < -100 nT) 1–6 hr ahead. The corresponding uncertainties associated to the predictions, which we refer to as  $\Delta$ Dst, are then estimated by using the ACCRUE method (Camporeale & Carè, 2021), also using a GRU network. The multi-fidelity boosting method proposed here works as follows. The accurate estimate of the uncertainty,  $\Delta$ Dst, for a given trained model, can inform us about the input conditions under which the model does not perform well. Hence, we can identify a subset of the original training set that can be used for training a different, independent model. Such a strategy can be iterated a number of times. The final result will be a collection of models, each working very well for a specific subset of input conditions (hence the built-in multi-fidelity). The crucial point, though, is that since each Dst model comes with its own estimate of uncertainty, that information can be used as a weighting factor when optimally combining (in a linear fashion) a large number of models.

The paper is divided as follows. Section 2 introduces the data used for this study, the criteria to define storm events, and the corresponding time periods covered. The methodology, including the designed uncertainty quantification-based machine learning architecture and developed multi-fidelity boosting method, is also described. Section 3 presents the results of the developed model, and discusses the advantages of the proposed model. Finally, in Section 4, we draw conclusions and make final remarks about future directions.

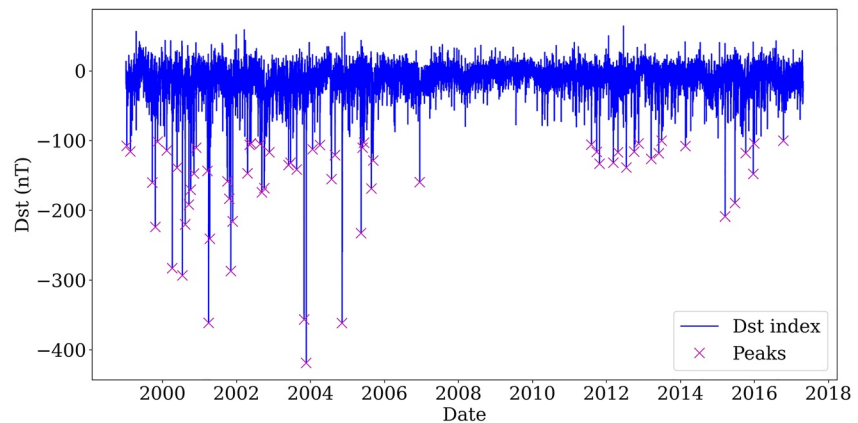
## 2. Data and Methods

### 2.1. Data

Previous studies have shown that various solar wind parameters have some predictive power with respect to future Dst values. Gruet et al. (2018) selected the proton density  $n$ , the solar wind velocity  $V$ , IMF  $|B|$ , and IMF  $B_z$ . The same variables are considered in this study. They convey information both about incoming high speed streams and Coronal Mass Ejections that directly influence Dst. In addition, we use the clock and dipole tilt angles, as defined in Weimer (2013). All variables are shown in Table 1. The models will be trained by using this variable set defined between  $t - \text{delay} - 6$  and  $t - \text{delay}$ , where  $t$  is the time stamp of a prediction, and *delay* ranges between 1 and 6 hr.

The historical Dst index is available at 1-hr cadence from the NASA OMNI database. Figure 1 displays the Dst index in the period 1999–2017. The model is trained and tested on storm events with a Dst peak smaller than -100 nT, shown by magenta crosses in Figure 1. Overall, 66 such storm periods are selected for this study.

Consistently defining a storm time window is a difficult task, since it usually includes a pre-storm period, a main phase and a recovery phase (Gonzalez et al., 1994). In this study, we define a storm event by looking for the nearest positive Dst values immediately before and after each negative peak, and then extending the time window by a 24-hr buffer. An example of such a definition of a storm event is shown in Figure 2, where the Dst peak is observed on 20 November 2003. The storm period is defined as ranging between 07 November 2003 and 30 November 2003. With this procedure we make sure that the time intervals are selected in such a way that the negative Dst peaks do not always occur at the same time within the chosen storm-time window. Hence, the neural network does not memorize. The average period of selected storm events is approximately 15 days. All 66 selected storms are listed in Table 2.



**Figure 1.** Time history of Dst during the period 1999 to 2017. The time is on the horizontal axis, and the Dst values are on the vertical axis. The magenta crosses denote peak values less than  $-100$  nT that are used for defining storm events considered in this study (see Table 2).

## 2.2. Methodology

### 2.2.1. Gated Recurrent Unit (GRU) Neural Networks

Gated Recurrent Unit (GRU) networks are one of the most widely used Recurrent Neural Networks. Similar to LSTM, GRU was proposed as a solution to short-term memory and vanishing gradient problems (Mikolov et al., 2010). In most scenarios, the performance of GRU is on par with LSTM, but computationally more efficient because of a less complex structure (Kaiser & Sutskever, 2015). The architecture of GRU is shown in Figure 3. All variables in Table 1 at a certain time  $t$  are used to form  $x_t$ , where  $X$  is a time series with a 6-hr span and a 1-hr time step, and  $Y$  is the corresponding Dst with a fixed time delay, that is, 1–6 hr.

### 2.2.2. ACCRUE Method

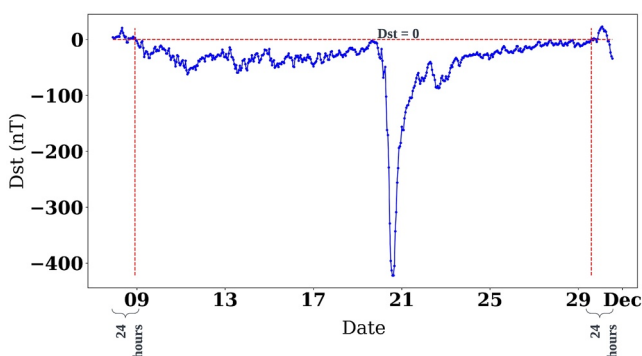
The ACCRUE method was proposed by Camporeale and Carè (2021) for assigning uncertainties to single-point predictions generated by a deterministic model that outputs a continuous variable. The problem of estimating the optimal uncertainty is set up as an optimization problem, where the ACCRUE score is defined as the combination of the continuous ranked probability score and the Reliability Score (RS) according to a certain ratio. In  $\Delta$ Dst modeling, the ACCRUE score is used as the cost function for the same GRU architecture utilized for the corresponding Dst model. The main assumption of the ACCRUE method is that the residuals of the underlying deterministic model (i.e., the difference between the model output and the ground truth) are normally distributed. The model output is taken as the mean of a Gaussian distribution, and its uncertainty is assigned by estimating the corresponding standard deviation. Note that the ACCRUE model is trained in a self-supervised regression mode, because the target (i.e., the standard deviation) is unknown, and it enters in the minimized loss function as a non-trivial, yet analytical, function.

If the ACCRUE model is successfully converged, the distribution of the so-called z-scores, that is the algebraic errors divided by the corresponding standard deviations, collected over the whole training set, should also be normally distributed. A graphical technique called quantile-quantile plots (Q-Q plots) (Wilk & Gnanadesikan, 1968) can then be applied to investigate whether the ACCRUE results indeed return a normal distribution for the z-scores.

If the ACCRUE model is successfully converged, the distribution of the so-called z-scores, that is the algebraic errors divided by the corresponding standard deviations, collected over the whole training set, should also be normally distributed. A graphical technique called quantile-quantile plots (Q-Q plots) (Wilk & Gnanadesikan, 1968) can then be applied to investigate whether the ACCRUE results indeed return a normal distribution for the z-scores.

### 2.2.3. Multi-Fidelity Boosting Method

The multi-fidelity boosting strategy is used to develop a collection of models trained on a subset of the original training set. For a given model, the ACCRUE method is applied in order to estimate the associated uncertainties

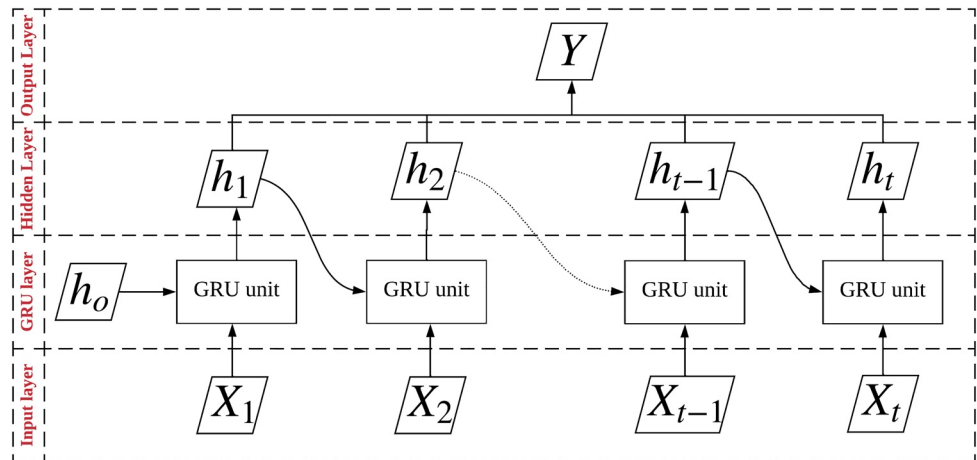


**Figure 2.** An example of the selection criterion used to define the time range for one storm event. The Dst peak occurs on 20 November 2003. The nearest positive Dst values before and after the peak occur on November 8 and November 29, respectively. The entire storm range is defined to be between 7 November 2003 and 30 November 2003, which includes a 24-hr buffer zone before and after the positive Dst occurrences. The list of selected storm events can be found in Table 2.

**Table 2**  
*List of First 66 Storm Events*

Start date/time	End date/time	Min. Dst (nT)	Start date/time	End date/time	Min. Dst (nT)
2000-02-10T01	2000-02-19T06	-135.0	2004-01-20T23	2004-01-29T23	-130.0
2000-03-31T05	2000-04-12T13	-292.0	2004-04-02T09	2004-04-05T00	-117.0
2000-05-22T19	2000-06-02T09	-147.0	2004-07-23T05	2004-08-06T12	-170.0
2000-07-14T09	2000-07-19T12	-300.0	2004-08-28T20	2004-09-03T19	-129.0
2000-08-08T23	2000-08-19T01	-234.0	2004-11-06T14	2004-11-19T11	-374.0
2000-09-16T13	2000-09-24T08	-201.0	2005-01-15T15	2005-01-21T11	-103.0
2000-09-23T11	2000-10-12T16	-181.0	2005-05-13T23	2005-05-19T22	-247.0
2000-11-02T22	2000-11-10T00	-159.0	2005-05-28T16	2005-06-04T05	-113.0
2000-11-25T16	2000-12-02T23	-119.0	2005-06-11T11	2005-06-16T00	-106.0
2001-03-18T05	2001-03-22T08	-149.0	2005-08-23T02	2005-08-30T21	-184.0
2001-03-29T21	2001-04-04T09	-387.0	2005-09-08T08	2005-09-24T21	-139.0
2001-04-07T06	2001-04-21T10	-271.0	2006-12-04T17	2006-12-31T20	-162.0
2001-04-20T18	2001-04-25T08	-102.0	2011-08-04T14	2011-08-13T15	-115.0
2001-08-16T10	2001-08-19T06	-105.0	2011-09-25T09	2011-09-30T17	-118.0
2001-09-24T15	2001-10-07T02	-166.0	2011-10-23T15	2011-10-30T04	-147.0
2001-10-18T04	2001-10-25T03	-187.0	2012-03-07T07	2012-03-22T11	-145.0
2001-11-04T13	2001-11-14T18	-292.0	2012-04-22T11	2012-05-05T06	-120.0
2001-11-23T00	2001-11-30T08	-221.0	2012-07-13T19	2012-07-20T07	-139.0
2002-03-22T09	2002-03-28T15	-100.0	2012-09-29T08	2012-10-05T02	-122.0
2002-04-16T05	2002-04-27T01	-149.0	2012-11-12T11	2012-11-17T03	-108.0
2002-05-10T05	2002-05-18T14	-110.0	2013-03-16T00	2013-03-23T17	-132.0
2002-05-22T05	2002-05-30T14	-109.0	2013-05-30T19	2013-06-08T23	-124.0
2002-07-31T17	2002-08-06T21	-102.0	2013-06-26T19	2013-07-04T12	-102.0
2002-08-17T16	2002-08-25T04	-106.0	2014-02-17T07	2014-02-23T01	-119.0
2002-09-02T17	2002-09-16T01	-181.0	2015-03-16T00	2015-03-25T06	-223.0
2002-09-28T20	2002-10-23T09	-176.0	2015-06-21T05	2015-07-02T22	-204.0
2002-11-16T04	2002-12-06T10	-128.0	2015-09-30T07	2015-10-17T14	-124.0
2003-05-20T08	2003-06-07T14	-144.0	2015-12-18T22	2015-12-24T07	-155.0
2003-06-15T03	2003-06-26T06	-141.0	2015-12-30T05	2016-01-04T05	-110.0
2003-07-09T19	2003-07-23T01	-105.0	2016-10-12T00	2016-10-20T01	-104.0
2003-08-16T12	2003-08-31T17	-148.0	2017-05-26T16	2017-05-31T13	-125.0
2003-10-23T20	2003-11-04T00	-383.0	2017-09-05T22	2017-09-14T06	-142.0
2003-11-07T22	2003-11-29T14	-422.0	2018-08-24T11	2018-09-01T05	-174.0

(trained on the errors evaluated over the whole training set) (Camporeale & Carè, 2021). All training samples are then sorted according to a chosen accuracy criterion. The best 10% samples are discarded in subsequent iterations. Meanwhile, a new subset is defined by using the worst 50% samples as the next training set, that is then used to train an independent model. Such a strategy can be iterated a number of times to form an ensemble of multi-fidelity models. Finally, by exploiting the knowledge about the uncertainties associated with each model, the final predictions, which are expected to outperform the predictions of each single model, are defined as a weighted combination of each ensemble member. Three criteria are tested in this study. They are readily defined for each sample in the training set, and they are: the absolute error, the standard deviation as estimated from ACCRUE ( $\sigma$ ), and their z-score defined as  $\frac{error}{\sigma}$ . When developing a model that predicts  $N$  hours ahead, we use the final model previously developed for  $N - 1$  hr as our baseline model, from which the iterative methods starts, as shown in Figure 5.



$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z)$$

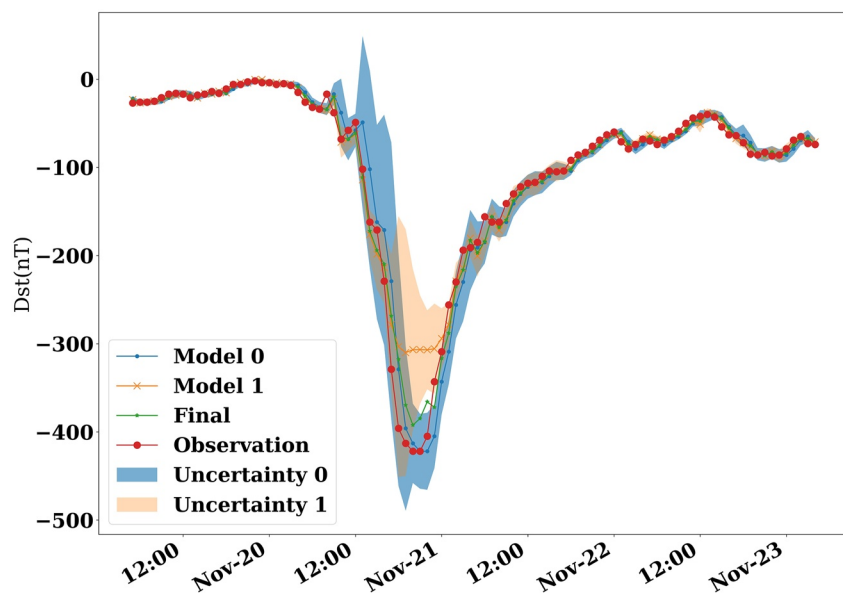
$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r)$$

$$h_t = \Phi_h(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)$$

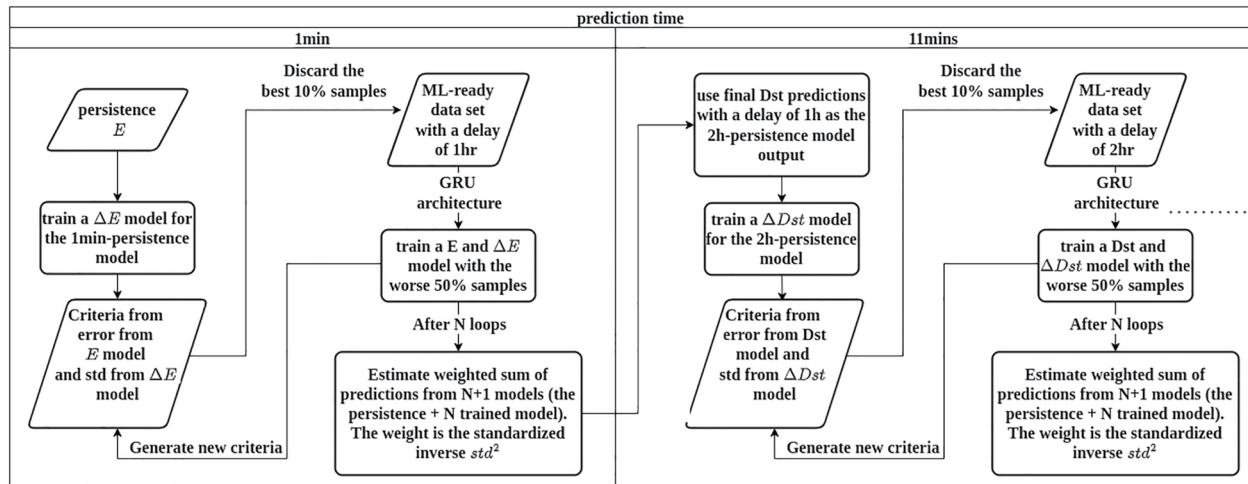
$$h_{t+1} = (1 - z_t) \odot h_t + z_t \odot \hat{h}_t$$

**Figure 3.** Structure of a Gated Recurrent Unit (GRU).  $x_t$  is the independent variable set at the time  $t$ , and  $Y$  is the target.  $h_t$  is the temporary result from the  $t$ th GRU unit and  $h_0$  is manually initialized. The connection between the hidden layer and the output layer is by means of a simple fully-connected layer. Each GRU unit can be considered as a multi-layer perceptron model.  $z_t$  is an update gate vector and  $r_t$  is a reset gate vector.  $W$ ,  $U$ , and  $b$  are the coefficients that need to be estimated during training. In addition,  $\sigma_g$  and  $\Phi_h$  are sigmoid and tanh activation functions, respectively.

An example of the proposed multi-fidelity boosting method for a given storm event is shown in Figure 4. Here, “model 0” is a simple persistence model. It is clear that “model 1” (obtained after one iteration of the boosting method) performs better during the pre-storm and recovery phase, and “model 0” performs better during the main phase of the storm. The combination of these two models can outperform each if their uncertainty can be



**Figure 4.** Illustration of multi-fidelity boosting method for the 2003-Halloween storm. Time is on the horizontal axis and Dst values are on the vertical axis. The blue and orange lines are from trained models. The blue and orange shaded areas are the corresponding standard deviations. The green line is the final prediction, and the red dotted line is the observed Dst.



**Figure 5.** Workflow of the multi-fidelity boosting method. It should be noted that, in order to mitigate overfitting, the best 10% of samples will be discarded before each training iteration.

well estimated. The final predictions are shown as Equation 1, where  $pred_f$  and  $pred_i$  denote final predictions and predictions from the  $i$ th model, and  $\sigma_i$  is the uncertainty corresponding to the  $i$ th model.

$$pred_f = \sum_{i=1}^n \frac{\frac{1}{\sigma_i^2}}{\sum_{j=1}^n \frac{1}{\sigma_j^2}} \times pred_i \quad (1)$$

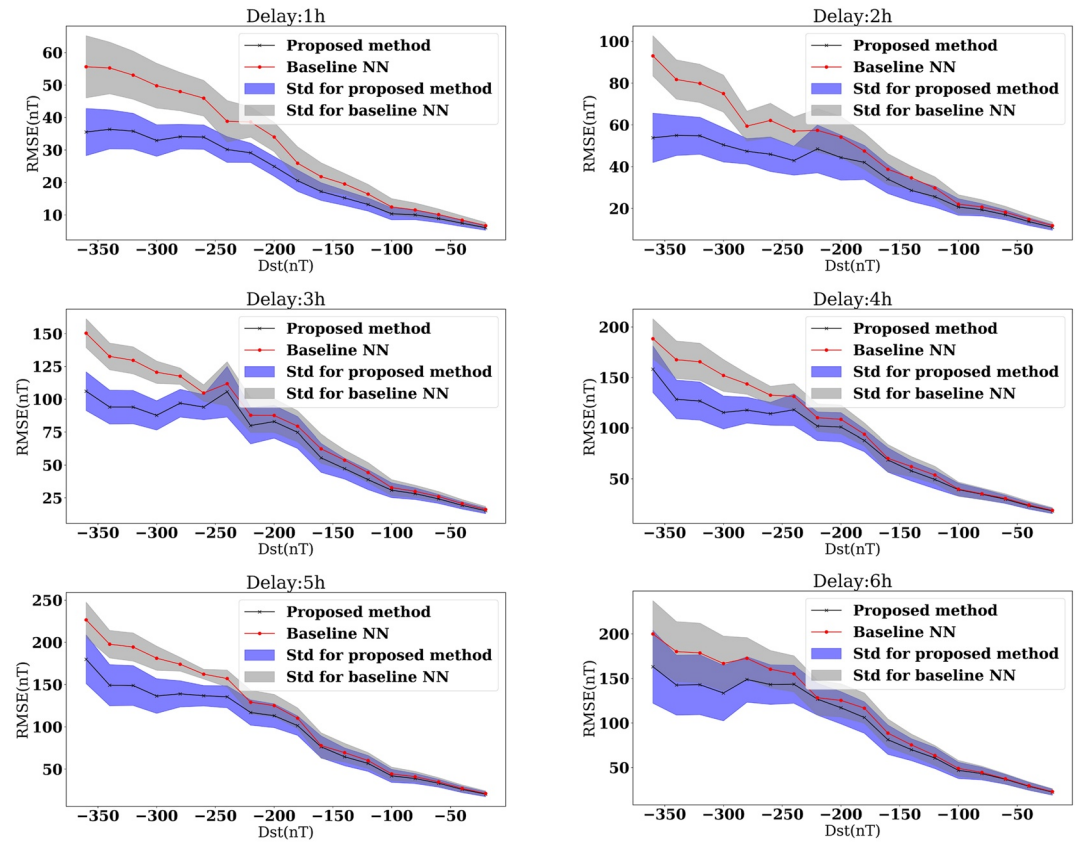
The accuracy of the models developed with different criteria are shown in Table 3, where different columns report the RMSE value calculated on different portions of the validation set, conditioned on the observed value of Dst (we have used four intervals:  $(-\infty, -100]$ ,  $(-100, -50]$ ,  $(-50, \infty)$ , and all). 20% of all samples are selected for this validation. Those samples will not be used for training. It is clear that the model developed with “z-score” outperforms the other two significantly within all Dst ranges. Hence, “z-score” is selected for the following study. All training samples will be separated into three subsets for training, early stop, and the aforementioned criteria validation. The training set needs to be shuffled before training in order to make sure it is homogeneously distributed.

**Table 3**  
RMSE of the Developed Model With Different Boost Criteria

Predict time	Boost criteria	RMSE(nT)			
		All	$(-\infty, -100]$	$(-100, -50]$	$(-50, \infty)$
1h	Error	2.72	5.02	3.80	2.07
	$\sigma$	2.74	5.78	3.27	2.06
	z-score	<b>2.59</b>	<b>4.84</b>	<b>3.31</b>	<b>2.00</b>
2h	Error	5.46	10.63	7.80	4.12
	$\sigma$	5.98	12.68	7.09	4.69
	z-score	<b>3.81</b>	<b>7.57</b>	<b>3.89</b>	<b>3.13</b>
3h	Error	10.19	<b>20.83</b>	11.34	8.14
	$\sigma$	10.39	22.10	10.88	8.24
	z-score	<b>8.22</b>	<b>16.99</b>	<b>10.37</b>	<b>6.00</b>
4h	Error	12.04	24.63	13.78	9.43
	$\sigma$	12.68	26.37	13.28	10.01
	z-score	<b>10.44</b>	<b>22.19</b>	<b>11.85</b>	<b>7.99</b>

Note. Most accurate method is marked in bold.

In order to precisely assess the accuracy of a model, it is important that the performance metrics are computed on a test set independent from the training set, so-called unseen data. Hence, we ensure that the machine learning algorithm does actually learn meaningful patterns and does not merely memorize the training data. A “Leave one out” technique is adopted here. That is a K-fold cross validation taken to its logical extreme, with K equal to N, the number of selected storm cases. That means that the proposed model is trained on all the data except for one storm window and a prediction is made for that left-out storm. The procedure is repeated N times. Finally, the metrics are computed as averages over the N models. In this study, each of the 66 storm windows reported in Table 2 constitutes a fold. Root-mean-square error (RMSE) is used as the main metric to assess the accuracy of the proposed model. The continuous Dst prediction can also be transformed to a binary label upon defining a threshold, that is,  $-100$  nT in this study. In this way we can use standard metrics for binary classification such as the True Skill Statistic and Matthews correlation coefficient (MCC) (Camporeale et al., 2020). The MCC score is a reliable statistical rate that produces a high score only if the prediction obtained good results in all of the four confusion matrix categories—True positive (TP), False positive (FP), True Negative (TN) and False Negative (FN)—proportionally both to the size of positive elements



**Figure 6.** RMSE of the proposed model with 1–6 hr time delay. For each panel, the horizontal axis is the Dst threshold, and the vertical axis is the RMSE of all samples below that threshold, evaluated in the test set. In each panel, the black line denotes the average boost results, the red line shows the average Gated Recurrent Unit (GRU) results, and the purple and gray regions are the uncertainty for the boosting and the GRU results, respectively.

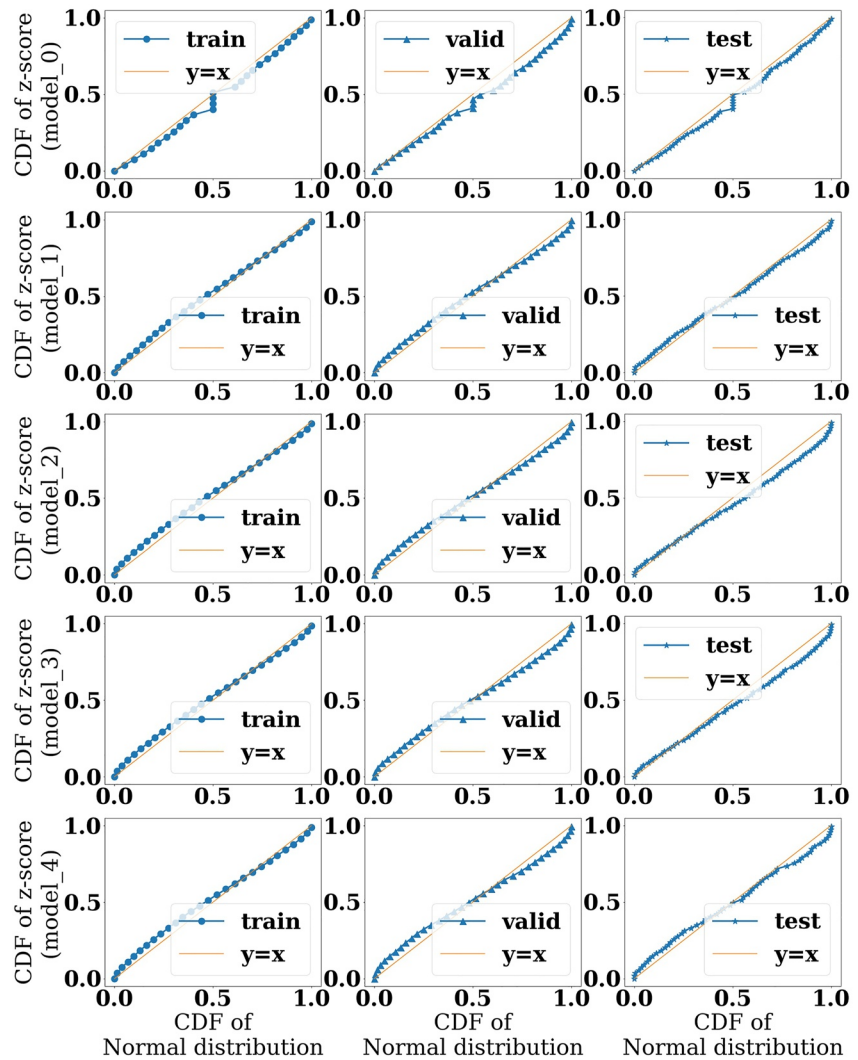
and the size of negative elements in the data set (Baldi et al., 2000). TSS is a useful metric that combines both types of information and should be as close as possible to 1. Those metrics have shown some advantages over the F1 score and accuracy in binary classification evaluation (Chicco & Jurman, 2020).

### 3. Results

In this section, the proposed boosting model is first compared against a single GRU model. The 2003 and 2022 Halloween storm events are selected to exhibit the advantages of this method. It should be noted that there are two storm events around 2003 Halloween (October 30 and November 20). The latter one is chosen because it has a larger negative peak (hereinafter we call it “the 2003-Halloween storm”). Note that this baseline model is trained with a random initialization and without performing a hyper-parameter optimization.

Figure 6 shows a statistical analysis of the proposed model with 1–6 hr time delay. In each panel, each dot represents the mean RMSE computed over all test samples with Dst smaller than a given Dst threshold (on the horizontal axis). The red line shows the average RMSE of the GRU results, and the gray bar is the corresponding RMSE standard deviation. The red line and the purple bar show the corresponding average RMSE and standard deviation for the developed boosting method, respectively.

It is clear that the boosting method significantly outperforms a single GRU method for all time delays. The corresponding Q-Q plots are shown in Figure 7. If z-scores estimated by the ACCRUE method are distributed as a perfect Gaussian, then the Q-Q lines (in blue) would perfectly overlap the diagonal line (in orange). In Figure 7, the Q-Q lines agree well with the diagonal lines in the training and validation sets. This correlation is slightly worse in the test set because the distribution of the test set might differ from that of the training and validation sets.



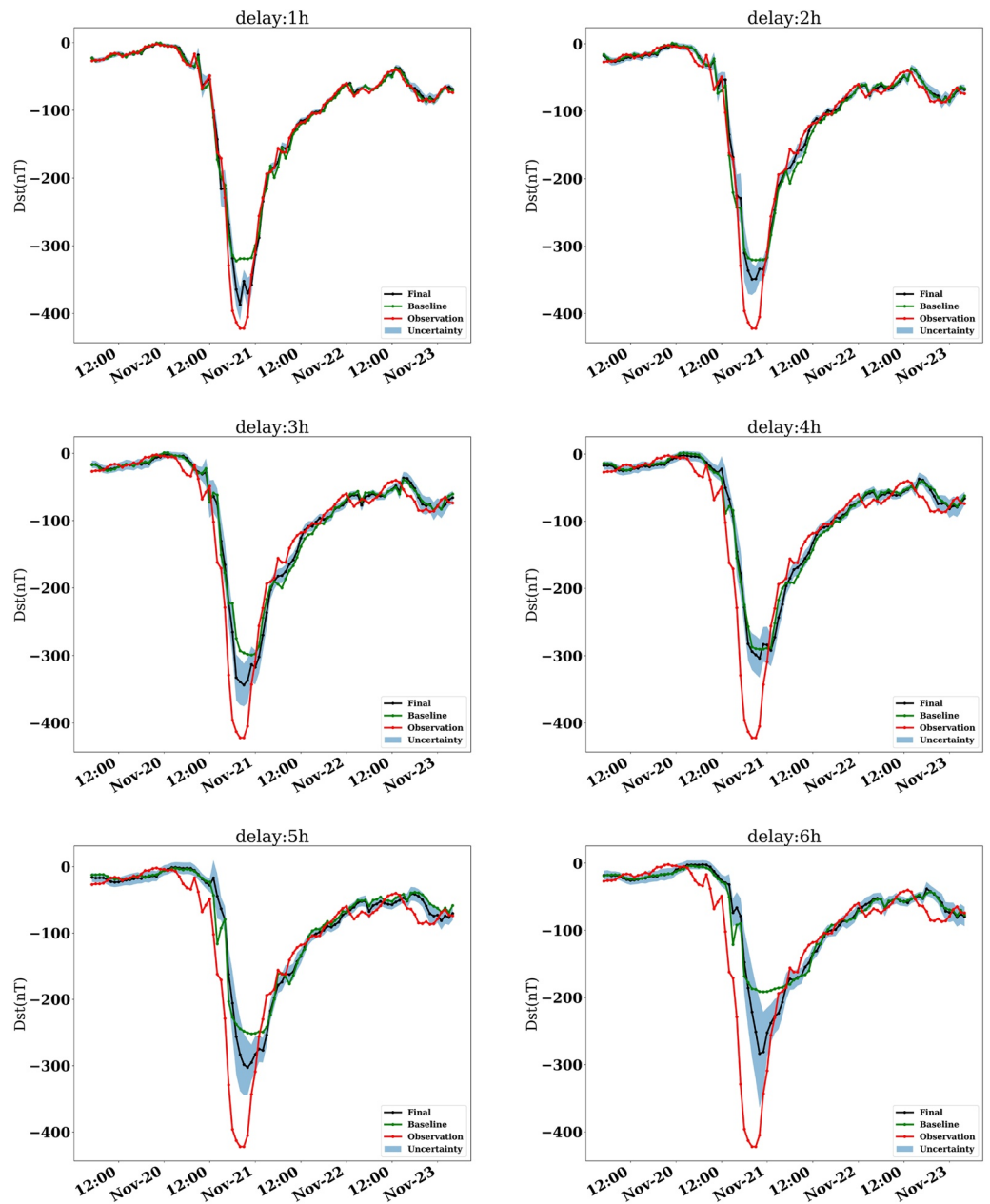
**Figure 7.** Quantile-quantile (Q-Q) plots (in blue) averaged over all storm events (6-hr ahead predictions for the final multi-fidelity results). From left to right, panels are QQ-plots for training, validation, and test sets. In each panel, the horizontal axis is a CDF of a normal distribution and the vertical axis is a CDF of the z-score for each submodel. The Q-Q line is blue, and the orange line is a 45-degree reference line (i.e.,  $y = x$ ).

In this section we analyze the model results for two case studies: (a) The 2003-Halloween storm, the biggest storm in the past 20 years, from 2003-11-19 to 2003-11-23; and, (b) The 2022-Halloween storm, a recent storm with Dst peak  $< -100$  nT, between 2022-11-06 and 2022-11-08. Results for all other storms are reported in the link <https://doi.org/10.5281/zenodo.7671247>. Figure 8 displays the Dst predictions of the proposed model (green for baseline GRU results and black for multi-fidelity boosting results) and the corresponding observed Dst (red) during the 2003-Halloween storm. The time delay of each panel is corresponding to the same panels in Figure 6.

There are two reasons for the under-prediction of the model in Figure 8 during the main phase of the storm. One has to do with the Dst residuals not being exactly normally distributed. Their distribution is skewed and that is not accounted for in the proposed multi-fidelity boosted method. A second reason has to do with the fact that the storm under consideration is very rare and not appropriately represented in the training set.

It is clear that the GRU model can better predict the time shift of the peak, especially between the main phase and the recovery phase. Meanwhile, the peak Dst is always better predicted by the persistence model. This proves that the multi-fidelity boosting model can take advantage of both models and further verifies our findings presented in Figure 6.



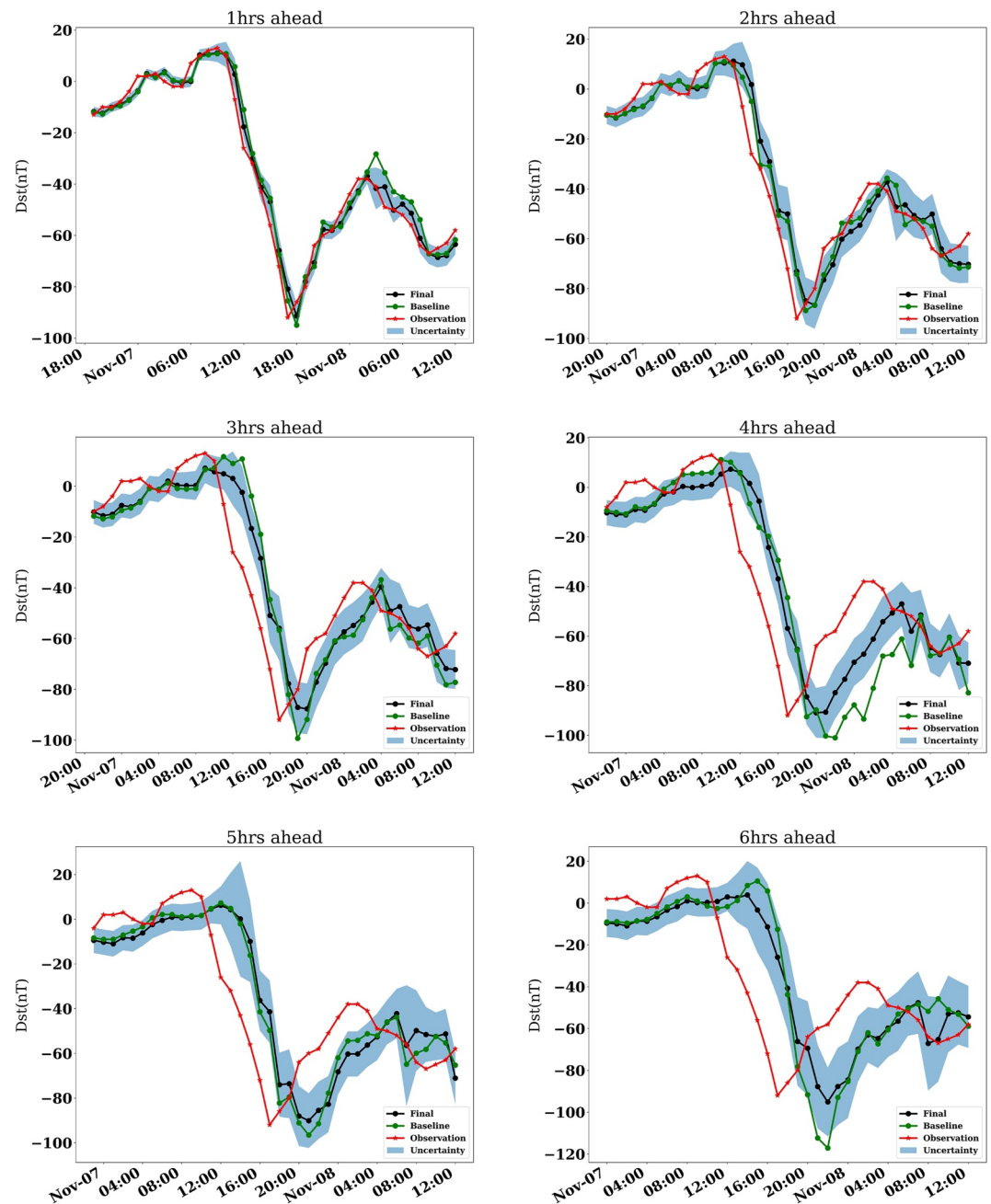


**Figure 8.** Dst predictions between 2003-11-19 and 2003-11-23 with 1-to-6 hr delays. In each panel, the red line is the observed Dst. The green and black lines are the baseline Gated Recurrent Unit and the multi-fidelity predictions. The blue regions are the multi-fidelity uncertainties.

Figure 9 is similar to Figure 8, except it shows the 2022-Halloween event. We can see that the proposed model, similar to the one for the 2003-Halloween storm, also better predicts Dst during the main phase of this 2022-Halloween storm than the baseline model. However, the time shift of the peak is much more significant here. This is because the model emphasizes storm events by putting more weights on larger negative Dst. A website (<https://swx-trec.com/dst/>) has been constructed to display real-time model predictions, such as those shown in Figure 9.

#### 4. Summary and Outlook

We have developed a multi-fidelity boosting model to predict Dst during geomagnetic storm periods, 1–6 hr ahead. Sixty-seven selected storm events were chosen during a long-span historical data set (20 years), between



**Figure 9.** Dst predictions between 2022-11-07 and 2022-11-09 with 1-to-6 hr delays. In each panel, the red line is the observed Dst. The green and black lines are the Gated Recurrent Unit and the multi-fidelity predictions. The blue regions are the multi-fidelity uncertainties.

2000-01-01 and 2020-01-01. One of the crucial points of this work is that an innovative multi-fidelity boosting method is developed to enhance a simpler prediction algorithm (here based on GRU networks), especially during those “super” storms as defined in Sec. 2. We have shown that this proposed model provides a good RMSE (8.22 nT) for predicting Dst up to 3 hr ahead during strong storm periods. The prediction becomes worse with a longer lead time (RMSE of 13.54 nT at 6 hr) because of the increasing errors of the baseline model. We have also discussed how this model performs during the 2003 and 2022 Halloween storms, as case studies. The Dst peak can be well captured by the developed model.

## Data Availability Statement

We thank OMNIWeb for providing the solar wind data (<https://omniweb.gsfc.nasa.gov/>), and WDC, Kytot for Dst data (<https://wdc.kugi.kyoto-u.ac.jp/index.html>). All the results have been made available in <https://doi.org/10.5281/zenodo.7671247>.

## Acknowledgments

This project has been developed in the framework of the National Aeronautics and Space Administration under Grants 80NSSC20K1580 and 80NSSC20K1275.

## References

- Agapitov, O., Artemyev, A., Mourenas, D., Mozer, F., & Krasnoselskikh, V. (2015). Empirical model of lower band chorus wave distribution in the outer radiation belt. *Journal of Geophysical Research: Space Physics*, 120(12), 10–425. <https://doi.org/10.1002/2015ja021829>
- Bala, R., & Reiff, P. (2012). Improvements in short-term forecasting of geomagnetic activity. *Space Weather*, 10(6). <https://doi.org/10.1029/2012sw000779>
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, 16(5), 412–424. <https://doi.org/10.1093/bioinformatics/16.5.412>
- Boardsen, S., Eastman, J., Sotirelis, T., & Green, J. (2000). An empirical model of the high-latitude magnetopause. *Journal of Geophysical Research*, 105(A10), 23193–23219. <https://doi.org/10.1029/1998ja000143>
- Burton, R. K., McPherron, R., & Russell, C. (1975). An empirical relationship between interplanetary conditions and Dst. *Journal of Geophysical Research*, 80(31), 4204–4214. <https://doi.org/10.1029/ja080i031p04204>
- 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., & Carè, A. (2021). Accrue: Accurate and reliable uncertainty estimate in deterministic models. *International Journal for Uncertainty Quantification*, 11(4), 81–94. <https://doi.org/10.1615/int.j.uncertaintyquantification.2021034623>
- Camporeale, E., Cash, M. D., Singer, H. J., Balch, C. C., Huang, Z., & Toth, G. (2020). A gray-box model for a probabilistic estimate of regional ground magnetic perturbations: Enhancing the NOAA operational geospace model with machine learning. *Journal of Geophysical Research: Space Physics*, 125(11), e2019JA027684. <https://doi.org/10.1029/2019ja027684>
- Chandorkar, M., Camporeale, E., & Wing, S. (2017). Probabilistic forecasting of the disturbance storm time index: An autoregressive Gaussian process approach. *Space Weather*, 15(8), 1004–1019. <https://doi.org/10.1002/2017sw001627>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 1–13. <https://doi.org/10.1186/s12864-019-6413-7>
- Gonzalez, W., Joselyn, J.-A., Kamide, Y., Kroehl, H. W., Rostoker, G., Tsurutani, B., & Vasylunas, V. (1994). What is a geomagnetic storm? *Journal of Geophysical Research*, 99(A4), 5771–5792. <https://doi.org/10.1029/93ja02867>
- Gruet, M. A., Chandorkar, M., Sicard, A., & Camporeale, E. (2018). Multiple-hour-ahead forecast of the Dst index using a combination of long short-term memory neural network and Gaussian process. *Space Weather*, 16(11), 1882–1896. <https://doi.org/10.1029/2018sw001898>
- Hu, A., & Zhang, K. (2018). Using bidirectional long short-term memory method for the height of f2 peak forecasting from ionosonde measurements in the Australian region. *Remote Sensing*, 10(10), 1658. <https://doi.org/10.3390/rs10101658>
- Kaiser, L., & Sutskever, I. (2015). Neural GPU learn algorithms. arXiv preprint arXiv:1511.08228.
- Laperre, B., Amaya, J., & Lapenta, G. (2020). Dynamic time warping as a new evaluation for Dst forecast with machine learning. *Frontiers in Astronomy and Space Sciences*, 39. <https://doi.org/10.3389/fspas.2020.00039>
- Lazzús, J., Vega, P., Rojas, P., & Salfate, I. (2017). Forecasting the Dst index using a swarm-optimized neural network. *Space Weather*, 15(8), 1068–1089. <https://doi.org/10.1002/2017sw001608>
- Li, S., Galas, R., Ewert, D., & Peng, J. (2016). An empirical model for the ionospheric global electron content storm-time response. *Acta Geophysica*, 64(1), 253–269. <https://doi.org/10.1515/acgeo-2015-0067>
- Li, Z., Wang, N., Liu, A., Yuan, Y., Wang, L., Hernández-Pajares, M., et al. (2021). Status of CAS global ionospheric maps after the maximum of solar cycle 24. *Satellite Navigation*, 2(1), 1–15. <https://doi.org/10.1186/s43020-021-00050-2>
- Lundstedt, H., Gleisner, H., & Wintoft, P. (2002). Operational forecasts of the geomagnetic Dst index. *Geophysical Research Letters*, 29(24), 34–41. <https://doi.org/10.1029/2002gl016151>
- Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Interspeech* (Vol. 2, pp. 1045–1048).
- O'Brien, T., & Moldwin, M. (2003). Empirical plasmopause models from magnetic indices. *Geophysical Research Letters*, 30(4), 1152. <https://doi.org/10.1029/2002gl016007>
- Rostoker, G. (1972). Geomagnetic indices. *Reviews of Geophysics*, 10(4), 935–950. <https://doi.org/10.1029/r10i004p00935>
- Saiz, E., Cid, C., & Cerrato, Y. (2008). Forecasting intense geomagnetic activity using interplanetary magnetic field data. *Annales Geophysicae*, 26(12), 3989–3998. <https://doi.org/10.5194/angeo-26-3989-2008>
- Sugiura, M., & Kamei, T. (1991). Equatorial Dst index 1957–1986, IAGA bull., 40. In A. Berthelier & M. Menville (Eds.), *International Service Geomagnetic*. Indices Publication Office.
- Tobiska, W., Knipp, D., Burke, W., Bouwer, D., Bailey, J., Odstreil, D., et al. (2013). The anomomilos prediction methodology for Dst. *Space Weather*, 11(9), 490–508. <https://doi.org/10.1002/swe.20094>
- 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>
- Wan, X., Xiong, C., Gao, S., Huang, F., Liu, Y., Aa, E., et al. (2021). The nighttime ionospheric response and occurrence of equatorial plasma irregularities during geomagnetic storms: A case study. *Satellite Navigation*, 2(1), 1–14. <https://doi.org/10.1186/s43020-021-00055-x>
- 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>
- Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis for the analysis of data. *Biometrika*, 55(1), 1–17. <https://doi.org/10.2307/2334448>
- Xu, S., Huang, S., Yuan, Z., Deng, X., & Jiang, K. (2020). Prediction of the Dst index with bagging ensemble-learning algorithm. *The Astrophysical Journal - Supplement Series*, 248(1), 14. <https://doi.org/10.3847/1538-4365/ab880e>
- Zhao, H., Friedel, R., Chen, Y., Reeves, G., Baker, D., Li, X., et al. (2018). An empirical model of radiation belt electron pitch angle distributions based on van Allen probes measurements. *Journal of Geophysical Research: Space Physics*, 123(5), 3493–3511. <https://doi.org/10.1029/2018ja025277>