An interpretable machine learning method for forecasting the SYM-H Index

Daniel Iong[1], Yang Chen[1], Gabor Toth[2], Shasha Zou[2], Tuija Pulkkinen[2], Jiaen Ren[2], Enrico Camporeale[3,4], Tamas Gombosi[2]

[1]Dept. of Statistics, University of Michigan, Ann Arbor, MI, USA
[2]Dept. of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI, USA
[3]CIRES, University of Colorado, Boulder, CO, USA
[4]NOAA Space Weather Prediction Center, CO, USA

Key Points:

- We adapt gradient boosting machines (GBMs) for forecasting the SYM-H index multiple hours ahead.
- We quantify feature contributions using SHAP values to interpret model predictions.
- Our proposed method has similiar accuracy to existing methods, while being more interpretable.

Corresponding author: Daniel Iong, daniong@umich.edu

Abstract

In this work, we develop gradient boosting machines (GBMs) for forecasting the SYM-H index multiple hours ahead using different combinations of solar wind and interplanetary magnetic field (IMF) parameters, derived parameters, and past SYM-H values. Using Shapley Additive Explanation (SHAP) values to quantify the contributions from each input to predictions of the SYM-H index from GBMs, we show that our predictions are consistent with physical understanding while also providing insight into the complex relationship between the solar wind and Earth's ring current. We also perform a direct comparison between GBMs and neural networks presented in prior publications for forecasting the SYM-H index by training, validating, and testing them on the same data. We find that the GBMs have a comparable root mean squared error as the best published black-box neural network schemes and GBMs have better Heidke Skill Scores at predicting strong storms.

Plain Language Summary

Forecasting geomagnetic indices is crucial for mitigating potential effects of severe geomagnetic storms on critical infrastructures such as power grids. In this work, we adopt gradient boosting machines (GBMs) for SYM-H prediction hours ahead with various combinations of solar wind & interplanetary magnetic field parameters, past SYM-H values, and other derived parameters. The feature importance quantification that we derive provides important, new insight into the complex relationship between the solar wind and the Earth's ring current.

1 Introduction

Geomagnetic storms are the largest geomagnetic disturbances, during which severe space weather threats can occur and disrupt our technological society. During geomagnetic storms, petajoules of energy enter the Earth's magnetosphere from the solar wind, of which vast majority is stored in the ring current in the inner magnetosphere. The ring current indices such as Dst and SYM-H provide essential information about the current strength and evolution as well as the energy budget, and thus are of crucial practical importance. The ring current indices have been used in numerous space weather applications, such as in classification of the storms, as key input for empirical models of the magnetospheric magnetic topology (N. Tsyganenko, 1989; N. A. Tsyganenko, 1995, 2002a, 2002b), as features representing the geomagnetic activity level for machine learning forecasting the ionospheric total electron content (Liu et al., 2020), as parameters used for forecasting of the radiation belt energetic particle fluxes (Sakaguchi et al., 2015) and other magnetospheric quantities (Bortnik et al., 2018). Therefore, the ability to predict the ring current indices is crucial for space weather forecasts and end-users.

Several attempts have been made to use machine learning methods to forecast the SYM-H index. Cai et al. (2010) and Bhaskar and Vichare (2019) used a Nonlinear AutoRegressive with eXogeneous inputs (NARX) neural network to predict 5-minute averages of the SYM-H index one hour ahead using past SYM-H values, solar wind and IMF parameters as input. Cai et al. (2010) trained their neural networks with data from 67 geomagnetic storms from 1998 to 2006, while Bhaskar and Vichare (2019) used data from 25 additional geomagnetic storms from 2006 to 2013. With the goal of developing operationally feasible models, Siciliano et al. (2021) trained long short-term memory (LSTM) and convolutional (CNN) neural networks to predict the SYM-H index one hour ahead using only IMF parameters and past SYM-H values as input. Collado-Villaverde et al. (2021) took a similar approach to predict the SYM-H index several hours ahead, while also considering the effects of omitting past SYM-H values as input on predictive performance. Both Siciliano et al. (2021) and Collado-Villaverde et al. (2021) train and validate their networks on 25 strong geomagnetic storms (Dst $<$ -100 nT) from 1998 to 2017

and evaluate their performance using 17 strong test storms. To conduct a direct comparison of predictive performance, we use the same storms and features to train and test our proposed model. Comparison results are discussed in section 4.2.

A myriad of machine learning approaches have been taken to forecast the closely related Dst index and other geomagnetic indices such as the Kp index. Attempts to apply machine learning methods to forecast the Dst index date back to the 1990s starting with the works of Lundstedt and Wintoft (1994), Gleisner et al. (1996), and Wu and Lundstedt (1997). These authors generally observed that the initial and main phases were more accurately predicted than the recovery phase when the Dst index is not used as an input due to the fact that the initial and main phases are more strongly correlated with solar wind properties. Pallocchia et al. (2006) advocated for using only IMF parameters as inputs for operational forecasting of the Dst index because in situ solar wind plasma instruments tend to fail more often than space-based magnetometers. This was also the motivation for using only IMF parameters and past values to forecast the SYM-H index in Siciliano et al. (2021) and Collado-Villaverde et al. (2021).

Although a vast majority of machine learning approaches to forecasting geomagnetic indices use neural networks, other techniques have also been proposed: Chandorkar et al. (2017) investigated the use of Gaussian Processes for forecasting the Dst index; Lu et al. (2016) compared the use of support vector machines (SVM) with neural networks; Boynton et al. (2011) employed the Nonlinear AutoRegressive Moving Average with eXogeneous inputs (NARMAX) model to derive an analytic expression to forecast 1-hour-ahead Dst as function of its past values and of the history of a solar wind-magnetoshpere coupling function. Xu et al. (2020) combined neural networks with SVM to construct an ensemble model using bagging to predict the Dst index up to six hours ahead. We also construct an ensemble model but use gradient boosting instead of bagging (see Bauer and Kohavi (1999) for a detailed comparison between boosting and bagging). Another difference is that we create an ensemble of many simple tree-based models as opposed to a few complex models. A comprehensive review of machine learning models for geomagnetic indexes can be found in Camporeale (2019).

Despite the exciting fact that machine learning methods have made a lot of progress in many scientific fields and have become popular tools, the lack of interpretability has been a major drawback in these data-driven methods. Even if machine learning methods have typically focused on predictive performance, there has been a recent surge in interest in making these methods more interpretable (Molnar et al., 2020). The development of interpretable machine learning algorithms is of key importance especially in scientific fields such as space weather. Inspite of the fact that machine learning methods have repeatedly been shown to outperform operational models empirically, these methods have not been widely adopted in an operational setting due to a lack of trust and skepticism from the space weather community. Interpretability gives confidence to operational forecasters that relevant physical processes are captured to some degree and encoded in a black-box model, hence reassuring of its generalizability and robustness versus rare events, which are the main focus of space weather forecasting. Gray-box approaches, which combine physics-based models with black-box models, can also be used to make machine learning methods for space weather forecasting more reliable (Camporeale et al., 2020).

Interpretability, or explainability, can be achieved with either post-hoc explanation methods or intrinsically interpretable models. Examples of intrinsically interpretable models include linear regression, decision trees, and generalized additive models. Unfortunately, there is often a tradeoff between intrinsic model interpretability and predictive performance because interpretable models tend to make strong simplifying assumptions such as linearity or additivity. Recent efforts have been made to close this gap, starting with additive models that incorporate two-way feature interactions (Lou et al., 2013). Post-hoc explanation methods, to some extent, can be used to explain the predictions

made by more complex models, usually by constructing an approximate interpretable model after training the original model. For an overview of interpretable machine learning methods, see Molnar (2019).

In this work, we not only aim to obtain accurate predictions of the SYM-H index, but more importantly, to learn if the data-driven approach can reveal insights on the physical mechanisms. In turn, these insights could then be used to inform future physics-based or grey-box models. We achieve this by using a post-hoc explanation method known as Shapley Additive Explanations (SHAP) to quantify the contributions from each input on the predictions made by gradient boosting machines. SHAP has been successfully used to interpret predictions from tree-based models in other scientific fields such as medicine (Lundberg et al., 2018), solar power forecasting (Kuzlu et al., 2020; Mitrentsis & Lens, 2021), finance (Bluwstein et al., 2020; Mokhtari et al., 2019), and atmospheric science (Stirnberg et al., 2020). Section 3.2 continues this discussion on interpretability and describes the SHAP method in detail.

The remainder of the paper is organized as follows. In Section 2, we introduce the data sources and our data processing procedures. In Section 3, we describe the gradient boosting machine, hyperparameter tuning, and quantification of feature importance. In Section 4, we provide results of our predictions, compare them with those published in the existing literature, and more importantly, the new insights that we learn from the prediction model results. We conclude in Section 5 with a summary on key findings and some discussions on future work.

2 Data

The Disturbance Storm Time (Dst) index is computed as the H (magnetic north) component perturbation on equatorial magnetometers (Mayaud, 1980) on an hourly basis, and is a historical characterization of a magnetic storm. The Dst index represents the longitudinally averaged part of the external geomagnetic field measured at the equator (Sugiura, 1964). As the index includes only the field variation, during geomagnetically quiet times, it hovers around zero. The typical definition of a geomagnetic storm is that the Dst index reaches values below -50 nT.

The SYM-H index is a high-time-resolution version of the original Dst index, and is given at 1-minute cadence (Wanliss & Showalter, 2006). The SYM-H index is compiled from 11 low- and mid-latitude magnetometer stations. Quiet time fields, including local time and seasonal quiet time Sq current effects, are removed, and the residuals are averaged together, divided by the cosine of the co-latitude of the station to yield the component parallel with the magnetic dipole. Geomagnetic storms can be classified based on the SYM-H values: moderate ($-100\,\text{nT} < \text{SYM-H} < -50\,\text{nT}$), intense ($-250\,\text{nT} < \text{SYM-H} < -100\,\text{nT}$), and superstorms ($\text{SYM-H} < -250\,\text{nT}$).

We extract the SYM-H index data from the OMNI dataset compiled at NSSDC (`https://spdf.gsfc.nasa.gov`) using the open-source Python library *swmfpy* (King, 2005; Al Shidi, Qusai, 2020). We use the level-2 solar wind plasma and interplanetary magnetic field (IMF) parameters from the Advanced Composition Explorer (ACE) spacecraft provided by the NASA Space Physics Data Facility (`https://cdaweb.gsfc.nasa.gov/index.html/`) as inputs in our models. The original dataset contains the IMF components from the ACE Magnetic Field Experiment (MAG) instrument (Smith et al., 1998) at a 16-second cadence, as well as proton density, bulk speed, and ion temperature from the SWEPAM suite (McComas et al., 1998), at a 64-second cadence. In addition to solar wind plasma and IMF parameters, we also include derived quantities, in particular the solar wind dynamic pressure and electric field, as these have been shown to be relevant input parameters for predicting geomagnetic storms (Newell et al., 2007). Explanation methods, such as SHAP, allow us to confirm or disprove these expectations. To

remove some of the high frequency variation inherent in high time resolution data and to eliminate minor data gaps, we average the SYM-H index, solar wind and IMF parameters to a 5-min time resolution.

For training and testing the GBMs discussed in section 3.1, we use 42 strong geomagnetic storms occurring between 1998 to 2018 which reached a minimum SYM-H index value of less than -100 nT. Information about these storms are given in tables 1 and 2. We use 5-fold cross validation to optimize hyperparameters (see section 3.1) instead of using a separate set of storms for validation, which allows us to use more data for training models.

Table 1. Storms used to train GBMs. These storms are identical to the ones used to train and validate models in Collado-Villaverde et al. (2021).

Storm #	Start time	End time	Min. SYM-H (nT)
1	1998-02-14	1998-02-22	-119
2	1998-08-02	1998-08-08	-168
3	1998-09-19	1998-09-29	-213
4	1999-02-16	1999-02-24	-127
5	1999-10-15	1999-10-25	-218
6	2000-07-09	2000-07-19	-335
7	2000-08-06	2000-08-16	-235
8	2000-09-15	2000-09-25	-196
9	2000-11-01	2000-11-15	-174
10	2001-03-14	2001-03-24	-165
11	2001-04-06	2001-04-16	-275
12	2001-10-17	2001-10-22	-210
13	2001-10-31	2001-11-10	-313
14	2002-05-17	2002-05-27	-113
15	2003-11-15	2003-11-25	-488
16	2004-07-20	2004-07-30	-208
17	2005-05-10	2005-05-20	-302
18	2006-04-09	2006-04-19	-110
19	1998-12-09	1998-12-19	-206
20	2012-03-01	2012-03-11	-149
21	1998-04-28	1998-08-05	-268
22	1999-09-19	1999-09-26	-160
23	2003-10-25	2003-11-03	-427
24	2015-06-18	2015-06-28	-207
25	2017-09-01	2017-09-11	-144

In order to predict SYM-H Δt hours ahead of time t, henceforth denoted as $y(t+\Delta t)$, we will consider different combinations of the features listed in table 3. We also consider lead times Δt of one and two hours. When the SYM-H index is included, the observations from the previous one hour are used as input. We set the history length for all other features to be either two hours, if the SYM-H index is included, or 30 hours, if the SYM-H index is excluded. The history length selections were motivated by Siciliano et al. (2021), who examined the coefficient of determination R^2 that quantifies the amount of observed variance that is explained by the predictions as a function of the history length, when the SYM-H index was either included or excluded as an input. They found that R^2 started to decrease when the history length was around 30 hours, if the SYM-H index was not included as input. When the SYM-H index was included as input, the R^2

Table 2. Storms used to test GBMs. These storms are identical to the ones used to test models in Collado-Villaverde et al. (2021).

Storm #	Start time	End time	Min. SYM-H (nT)
26	1998-06-22	1998-06-30	-120
27	1998-11-02	1998-11-12	-179
28	1999-01-09	1999-01-18	-111
29	1999-04-13	1999-04-19	-122
30	2000-01-16	2000-01-26	-101
31	2000-04-02	2000-04-12	-315
32	2000-05-19	2000-05-28	-159
33	2001-03-26	2001-04-04	-434
34	2003-05-26	2003-06-06	-162
35	2003-07-08	2003-07-18	-125
36	2004-01-18	2004-01-27	-137
37	2004-11-04	2004-11-14	-393
38	2012-09-10	2012-10-05	-138
39	2013-05-28	2013-06-04	-134
40	2013-06-26	2013-07-04	-110
41	2015-03-11	2015-03-21	-233
42	2018-08-22	2018-09-03	-205

results for history lengths of 90 to 180 minutes were similar, while R^2 started to decrease for time intervals longer than 180 minutes.

Table 3. Features used as input into our models.

Features	History length (in hours)
Past SYM-H index (nT)	1
IMF: B_x, B_y, B_z (nT)	2 or 30
Solar wind: V_x (km/s), ρ (amu/cm^3), T (K)	2 or 30
Derived quantities: ρV_x^2 (nPa), $E_s = \max(0, -\lvert V_x \rvert B_z)(mV/m)$	2 or 30

The different sets of features used as inputs are listed in table 4. Using different sets of features to train our models allows us to investigate how the inclusion of certain features affects predictive performance and feature contributions. The choice to train our models using only IMF parameters and past SYM-H (input set I_1, Table 4) was motivated by the high percentage of missing observations for solar wind plasma parameters. For IMF parameters and solar wind velocity, there is less than 2% of observations missing within our sample. However, this percentage is substantially higher for solar wind density and temperature. Although our proposed model handles missing data internally, we choose to impute missing observations using linear interpolation (see section 3.4 in Chen and Guestrin (2016) for details).

Including solar wind plasma and derived parameters in input sets I_3 and I_4 allows us to investigate how these contribute to predictions. In particular, a sudden increase of dynamic pressure ρV_x^2 can compress the magnetosphere that causes a positive jump in SYM-H, which typically happens at the beginning of the geomagnetic storms (sudden storm commencement). Another physically important parameter is the y component of the interplanetary electric field $E_y = V_x B_z$ that characterizes the amount of north-south magnetic flux carried by the solar wind. Note that $V_x < 0$ in the geocentric-

solar-magnetic (GSM) coordinate system used here. The rectified electric field $E_s = \max(0, E_y)$ is the same as E_y when the IMF has a southward component ($B_z < 0$), which facilitates the onset of dayside reconnection, and zero for northward IMF when dayside reconnection is limited to high latitudes beyond the polar cusps (Burton et al., 1975). To examine how solar wind and IMF parameters influence predictions without knowledge of past SYM-H values, we train models with input sets I_2 and I_4 which exclude past SYM-H values (see Table 4).

Table 4. Various sets of features used as inputs to train our models.

Input set	Features included
I_1	IMF, past SYM-H
I_2	IMF
I_3	IMF/solar wind/derived quantities, past SYM-H
I_4	IMF/solar wind/derived quantities

3 Methods

3.1 Gradient Boosting Machines

Gradient boosting machines (GBMs), also known as gradient boosted trees, have had considerable success in prediction tasks across a wide range of domains (Natekin & Knoll, 2013). Shwartz-Ziv and Armon (2021) recently performed a rigorous study showing GBMs outperformed several neural network models in terms of accuracy in classification and regresssion problems with tabular data. GBMs are consistently used in the winning solutions of various machine learning prediction competitions like Kaggle, showing its effectiveness in a wide range of problems. In the space sciences, GBMs and other ensemble methods have recently been used to predict ambient solar wind flow (Bailey et al., 2021) and the Dst index (Xu et al., 2020).

In contrast to algorithms that construct one complex model, gradient boosting sequentially constructs simple prediction models called base learners that improve upon previously constructed base learners and sums them together to obtain an ensemble model. This process is analogous to how gradient descent optimizes weights in a neural network. Seen as a form of "functional gradient descent", gradient boosting minimizes an objective function by iteratively adding a new base learner, usually a decision tree, that leads to the largest decrease in the loss function (Friedman, 2001). In the case of GBMs, the base learners are regression trees, which are a highly interpretable class of machine learning models that mimic human decision-making but are often too simplistic for most prediction problems when used alone. Fortunately, ensembles of regression trees, like GBMs, are capable of producing highly accurate predictions while still taking advantage of the interpretability of regression trees. In addition to gradient boosting, bagging is another widely used ensemble method that constructs multiple base learners in parallel and aggregates them by averaging (Breiman, 1996).

The gradient boosting machines that we use to forecast SYM-H have the form

$$y(t + \Delta t) = \alpha + \sum_{m=1}^{M} T_m(I(t)) + \epsilon(t), \ t = 1, \ldots, N, \tag{1}$$

where $I(t)$ is a vector of inputs used at time t; $\epsilon(t)$ is an error term at time t; T_m's are regression trees; M is the number of iterations (trees) in the training algorithm; N is the number of timepoints; and α is a constant intercept term. $I(t)$ depends on which input set from table 4 is used. For instance, if input set I_2 is used, $I(t) = \big(B_x(t), \ldots, B_x(t-$

115), $B_y(t), \ldots, B_y(t-115), B_z(t), \ldots, B_z(t-115)\big)$. The regression trees can be written mathematically as

$$T(x) = w_{q(x)}, \tag{2}$$

where w are the leaf weights of the tree; and q represents the tree structure by mapping an input to its corresponding leaf node index. Figure 1 shows the tree structure of one of the trees in a GBM that we trained.

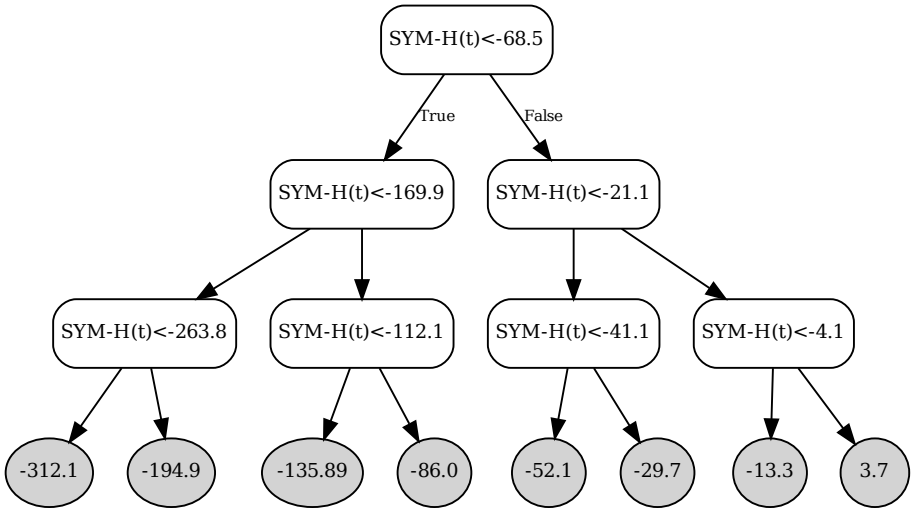

Figure 1. Structure of the first tree T_1 learned in a GBM trained with input set I_3 to predict the SYM-H index one hour ahead. The leaf nodes of the tree are shaded gray. The value in each leaf node is its corresponding leaf weight. Left splits correspond to the inequality in the previous node being true, and vice versa.

To train our GBMs, we use the open-source framework XGBoost that constructs the regression trees using gradient boosting and penalizes trees that are overly complex to avoid overfitting (Chen & Guestrin, 2016). More specifically, at each iteration m, we will construct a new regression tree T_m by minimizing the following objective function.

$$\mathcal{L}^{(m)}(T_m) = \sum_{t=1}^{N} \Big\{ y(t+\Delta t) - \big[\hat{y}^{(m-1)}(t+\Delta t) + T_m(I(t))\big] \Big\}^2 + \sum_{j=1}^{m} \Omega(T_j), \tag{3}$$

$$\text{where } \hat{y}^{(m-1)}(t+\Delta t) = \sum_{k=1}^{m-1} T_k(I(t)) \text{ and } \Omega(T_j) = \gamma K_j + \frac{1}{2}\lambda \sum_{k=1}^{K_j} w_{j,k}^2. \tag{4}$$

In eq. (4), K_j is the number of leaf nodes in T_j; $w_{j,k}$'s are the leaf node weights in T_j; and γ and λ are regularization hyperparameters. Ω is a regularization term that penalizes the complexity of the regression trees by limiting the number of leaf nodes and shrinking the leaf weights. Increasing γ results in shallower trees while increasing λ leads to smaller leaf weights. An alternative method for controlling tree size is to explicitly set the maximum tree depth. Besides increasing λ, we can also reduce the influence of individual trees by scaling their leaf weights by a learning rate.

To reduce the risk of overfitting, we control model complexity by optimizing several hyperparameters: learning rate, maximum tree depth, feature subsampling percentage, minimum child weight, and number of boosting iterations (trees). We optimize these hyperparameters, except the number of iterations, using cross validation and a gradient-free optimization platform called Nevergrad (Rapin & Teytaud, 2018). To set the number of iterations (trees), we monitor performance using cross validation at each iteration and terminate the algorithm when the performance stops improving. This technique is commonly referred to as early stopping in the machine learning literature (Zhang & Yu, 2005). Cross validation is performed by splitting the training storms in table 1 into 5 sets and using each set for evaluation while training the model on the rest of the training storms. Using cross validation, as opposed to a separate validation set, allows us to use more data when training the final model. The specific hyperparameter values we set are given in table 5.

Table 5. Hyperparameter values for training GBMs using the different input sets in table 4.

Input set	Hyperparameter	Value
I_1, I_2	Learning rate	0.072
	Max. tree depth	4
	Min. child weight	4
	Column subsampling %	0.78
	# of trees	84
I_3, I_4	Learning rate	0.147
	Max. tree depth	3
	Min. child weight	2
	Column subsampling %	0.894
	# of trees	291

GBMs have several advantages over competing machine learning methods. GBMs, and tree-based methods in general, are invariant to monotonic transformations of the features so it is better equipped to handle inputs on different scales. A practical consequence of this property is that the features don't have to be standardized before training. GBMs are robust against issues arising from correlated features due to the greedy nature of gradient boosting and how regression trees are constructed. A downside of tree-based models is that they produce predictions that are not smooth due to the tree structure of the model (Hastie et al., 2001). This can be seen in fig. 2, where the predictions from our GBM looks noisier than the ones from LSTM. Despite this property, GBMs are still able to produce highly accurate predictions. Another problem is that trees have difficulty modeling additive structures (see 9.2.4 in Hastie et al. (2001) for an example).

GBMs can also suffer from over-specialization, wherein trees added in later iterations tend to only impact the predictions of a few instances (Korlakai Vinayak & Gilad-Bachrach, 2015). This may make the model highly sensitive to the contributions of the initially added trees. This issue is combated, to some extent, by selecting a small learning rate. To further alleviate this issue, we use a technique for employing dropouts in GBMs introduced by Korlakai Vinayak and Gilad-Bachrach (2015). Dropouts have been used successfully in neural networks, where a random subset of connections in the network is dropped during training (Srivastava et al., 2014). In the context of GBMs, at each training iteration, we replace $\hat{y}^{(m-1)}$ in eq. (3) with the sum of a random subset, instead of all, of the previously constructed trees and then normalize the newly constructed tree and dropped trees. Further details of this procedure can be found at (Korlakai Vinayak & Gilad-Bachrach, 2015).

3.2 Feature Importance

Methods for computing feature contribution, or feature importance, can be categorized as global versus local and model-specific versus model-agnostic. Global feature importance scores are used to explain a model's overall behavior across the entire training dataset, while local feature importance scores tells you how individual features contributed to a single prediction. Model-specific feature importance is provided directly by the model, while model-agnostic methods, such as SHAP, typically construct an approximate interpretable model to explain predictions from the original model. For tree-based models, global feature importance can be calculated using information gain (Breiman et al., 1984), permutation (Breiman, 2001), or split count (Chen & Guestrin, 2016). In this paper, we will focus primarily on local feature importance as the contribution from each feature is likely to vary over time depending on the storm phase.

While there are several methods for computing local feature contribution in tree-based models (Molnar, 2019), we chose to use Shapley additive explanation (SHAP) because of its desirable theoretical properties (Lundberg & Lee, 2017). SHAP is based on Shapley values in cooperative game theory (Shapley, 1953), where they are used to fairly distribute payoffs in a game among a coalition of players with unequal contributions. In the case of SHAP, the payoff is the prediction and the players are the features. SHAP belongs to the class of additive feature attribution methods which assumes the following linear explanation model for an individual prediction.

$$g(\mathbf{z}) = \phi_0 + \sum_{i=1}^{p} \phi_i z_i, \tag{5}$$

where ϕ_0 is a reference value (e.g. mean); p is the number of input features; $\mathbf{z} = \left(z_1 \dots z_p\right)'$, where z_i is a binary variable indicating whether feature i is present; and ϕ_i is the contribution from feature i. SHAP yields the unique solution to eq. (5) which satisfy three desirable theoretical properties: local accuracy, missingness, consistency. The local accuracy property ensures that the sum of feature contributions for given inputs sum up to the prediction. The consistency property ensures that the SHAP value for a feature increases if the marginal contribution from that feature increases. Missingness is mainly a theoretical property that says a missing feature has zero contribution. The only alternative tree-specific local explanation method that we are aware of is Saabas (2014), which doesn't have the consistency property. Note that SHAP values describe a particular model's decision-making process based on the data, as opposed to the physical mechanisms characterized by the data. Therefore, SHAP values cannot be used alone to make general statements about how the different inputs affect the SYM-H index.

Although SHAP values can, in theory, be computed for any black box model, they are more computationally efficient for tree-based models like GBMs due to a model-specific algorithm for computing exact SHAP values known as TreeSHAP (Lundberg et al., 2019), which reduces the computational complexity from exponential to polynomial. For other complex models like neural networks, computing SHAP values would require refitting the model with many subsets of features, which is impractical if training is expensive and more than a few features are used. Unfortunately, a downside of using TreeSHAP is that non-contributing features can potentially have a non-zero contribution if they are correlated with a contributing feature (Molnar, 2019).

4 Results

In this section, we will compare the predictive performance of GBMs with neural networks developed by Siciliano et al. (2021) and Collado-Villaverde et al. (2021), interpret model predictions using the methods discussed in section 3.2, and discuss how predictions vary when the different set of features listed in table 4 are used as inputs.

4.1 Metrics

To evaluate the predictive accuracy of GBMs for forecasting the SYM-H index, we use metrics that quantify fit and event detection performance recommended by Liemohn et al. (2018). Using both types of performance metrics allows us to pinpoint specific strengths and weaknesses of our models. The metrics used are the root mean squared error (RMSE) and Heidke skill score (HSS).

The root mean squared error (RMSE), defined in eq. (6), is arguably the most widely used metric for evaluating accuracy in machine learning. It provides insight into how well predictions match observations on average so a lower value is better. However, RMSE gives equal weight to all indices which isn't optimal for evaluating SYM-H index predictions because accurate predictions during storm events is more desirable than during quiet times. Therefore, additional metrics should be used in conjunction with RMSE to assess accuracy during storm events.

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

To evaluate how well our method is able to predict the timing of geomagnetic storms, we use the Heidke skill score (HSS), which measures how well a model predicts a predefined event. In our case, an event occurs when the SYM-H index drops below -100 nT, the commonly used threshold separating moderate and strong geomagnetic storms. The HSS is defined as

$$\mathrm{HSS} = \frac{2[(H \times N) - (M \times F)]}{(H + M)(M + N) + (H + F)(F + N)}, \tag{7}$$

where H, M, F, N denote the number of true positives (hits), false negatives (misses), false positives, and true negatives, respectively. HSS takes on values between $-\infty$ and 1, with higher values being more desirable. The value HSS = 1 means perfect prediction, while HSS = 0 indicates that the model is as good as random guess with the correct positive/negative ratio.

4.2 Comparison to existing methods

In this section, we compare the predictions obtained using our model with the neural networks developed in Siciliano et al. (2021) (**LSTM1/CNN1**) and Collado-Villaverde et al. (2021) (**LSTM2**) on the 17 test storms in table 2 using the metrics discussed in section 4.1. Collado-Villaverde et al. (2021) considers 1-2 hours ahead prediction, whereas Siciliano et al. (2021) only considers 1-hour. On the other hand, Siciliano et al. (2021) trains models with and without the SYM-H index as an input, whereas Collado-Villaverde et al. (2021) only trains models with SYM-H. We train GBM models to predict 1-2 hours ahead with and without the SYM-H index as an input and compare them to the corresponding neural network models. All models were trained using data from the same storms in table 1.

Both metrics in section 4.1 are computed for all considered models and shown in tables 6 to 8. Since Collado-Villaverde et al. (2021) made their predictions publicly available (Redxgit, 2021), we were able to compute the HSS for their predictions. However, we were not able to compute the HSS for predictions from Siciliano et al. (2021). Following the convention taken in Collado-Villaverde et al. (2021), we display our metrics using three decimal places.

4.2.1 1-hour ahead predictions

For predicting 1 hour ahead with SYM-H included as an input, our GBM achieves the mean and median RMSE and highest mean and median HSS among the considered

models. Furthermore, our GBM has the lowest RMSE and highest HSS for 14 out of 17 test storms. Table 6 and the first column of table 7 shows the RMSE and HSS, respectively, for 1 hr ahead predictions using our GBM, LSTM1, and LSTM2. Compared to LSTM2, our model has a slightly higher RMSE only for test storms 33, 34, and 40 and a lower HSS only for test storms 29, 34, and 41. LSTM1 has a slightly lower RMSE only for test storm 39. Our proposed model has a higher HSS for 1 hour ahead prediction than LSTM2 for the three strongest test storms with SYM-H < -300 nT (31, 33, 37), indicating that it has more skill in predicting the timing of intense geomagnetic storms. Figure 2 shows the 1 hour ahead predictions from our GBM and LSTM2 during the main and recovery phases of these test storms along with the corresponding prediction errors and HSS. The distribution of the prediction errors are roughly similiar for these three test storms. For the March 2001 storm (second row; fig. 2), our GBM was able to accurately predict the minimum SYM-H of around -400 nT that was reached around 06:00 to 12:00 UT Mar 31 even though the timing is slightly off. According to the metrics used, our proposed model outperforms LSTM1 by a significant margin and LSTM2 by a slight margin for predicting the SYM-H index 1 hour ahead.

Table 6. RMSEs for 1-hour ahead prediction over the test storm set with our GBM model, LSTM1 (Siciliano et al., 2021) and LSTM2 (Collado-Villaverde et al., 2021) neural networks and simple persistence. The lowest RMSE for each row is shown in **bold**.

Storm #	GBM	LSTM1	LSTM2	Persistence
26	**5.863**	6.630	6.700	7.631
27	**7.729**	8.913	8.900	9.623
28	**4.281**	5.858	5.400	5.814
29	**5.833**	6.683	7.200	7.174
30	**4.927**	5.200	5.600	4.810
31	**8.277**	8.584	10.700	10.429
32	**6.841**	7.259	8.300	10.528
33	14.492	**13.340**	16.300	21.167
34	10.190	**10.034**	11.300	10.913
35	**7.154**	7.693	8.500	8.011
36	**8.512**	9.525	8.700	9.708
37	**14.548**	15.184	17.500	19.698
38	**3.886**	4.080	4.200	4.842
39	5.901	6.431	**5.600**	7.597
40	4.976	**4.673**	5.500	5.057
41	**7.558**	7.882	9.000	9.984
42	**5.030**	5.669	5.900	6.036
Mean	**7.412**	7.860	8.550	9.354
Median	**6.841**	7.259	8.300	8.011
Min.	**3.886**	4.080	4.200	4.810
Max.	**14.548**	15.184	17.500	21.167

4.2.2 2-hour ahead predictions

Tables 7 and 8 show the HSS and RMSE values for 2-hour ahead predictions from GBM and LSTM2 with past SYM-H included as an input, respectively. Our GBM has a lower RMSE for 9 out of the 17 test storms and a slightly higher mean HSS than LSTM2. However, LSTM2 has a slightly lower mean and median RMSE due to the high RMSE for storm #33 from our GBM. Both models have a significantly lower RMSE than the persistence model.

Table 7. Heidke skill scores (HSS) for GBM vs. LSTM2 (Collado-Villaverde et al., 2021) using a threshold of -100 nT. The highest HSS for each row and lead time is shown in **bold**.

	1-hour ahead		2-hour ahead	
Storm #	GBM	LSTM2	GBM	LSTM2
26	**0.417**	-0.004	**0.116**	0.000
27	**0.774**	0.731	**0.708**	0.635
28	**0.239**	0.112	**0.399**	0.310
29	0.337	**0.416**	**0.249**	0.089
30	**0.132**	0.095	**0.092**	0.040
31	**0.984**	0.975	**0.970**	0.956
32	**0.769**	**0.769**	0.677	**0.719**
33	**0.987**	0.986	**0.971**	0.960
34	0.757	**0.825**	0.592	**0.671**
35	**0.530**	0.497	**0.194**	-0.005
36	**0.435**	0.163	0.076	**0.556**
37	**0.934**	0.922	0.890	**0.904**
38	**0.950**	0.922	**0.890**	0.824
39	**0.844**	0.727	0.668	**0.674**
40	**0.380**	0.273	**0.417**	-0.001
41	0.925	**0.940**	**0.941**	0.892
42	**0.954**	**0.954**	0.864	**0.896**
Mean	**0.668**	0.606	**0.571**	0.536
Median	**0.763**	0.729	0.630	**0.653**
Min.	**0.132**	-0.004	**0.076**	-0.005
Max.	**0.987**	0.986	**0.971**	0.960

4.2.3 Predictions without past SYM-H

When we omit the SYM-H index as an input to predict 1-hour ahead, our GBM outperforms LSTM1 and has simliar performance as CNN1. Table 9 shows the RMSE for 1-hour ahead predictions from GBM, LSTM1, and CNN1 and 2-hour ahead predictions from GBM. Our GBM has the lowest overall mean and median RMSE and lowest RMSE for 11 out of the 17 test storms. However, CNN1 achieves a lower minimum and maximum RMSE and lower RMSEs for the 3 strongest test storms (33, 37, 40). Interestingly, the 2-hour ahead predictions from GBM have only a slightly higher overall RMSE than the 1-hour ahead prediction.

4.3 Interpreting predictions

Although existing methods for forecasting the SYM-H index have similar accuracy as our proposed model, they do not provide objective explanations for how their predictions were made. In this section, we explain how the input features we use contributed to our model's predictions using the methods discussed in section 3.2. To obtain the contributions from each feature in table 3, we sum up the contributions from the history of that feature. To reiterate from section 3.2, quantified feature contributions describe a particular model and not the data itself so they cannot be used to make statements about how the different inputs affect the SYM-H index in general.

Figure 3 shows the contributions to the 1-hour prediction from various features as a function of the SYM-H. Overall, the past SYM-H value dominates, which means that SYM-H varies relatively smoothly at a 1-hour time scale. This also means that beating the persistence model is not easy. The second most important contribution comes from

Table 8. RMSEs for 2-hour ahead prediction over the test storm set with our GBM model, the LSTM2 neural network (Collado-Villaverde et al., 2021) and persistence. The lowest RMSE for each row is shown in **bold**.

Storm #	GBM	LSTM2	Persistence
26	**8.285**	8.989	12.374
27	**11.585**	13.418	15.387
28	**5.650**	5.877	9.331
29	**8.826**	9.314	11.415
30	**7.280**	7.288	7.416
31	12.613	**12.436**	17.193
32	9.927	**8.937**	15.282
33	24.519	**18.481**	33.927
34	**13.736**	13.941	15.109
35	**9.504**	9.932	11.211
36	12.068	**12.058**	14.687
37	22.327	**21.084**	30.582
38	**5.153**	5.213	7.353
39	7.391	**6.798**	12.322
40	5.633	**5.281**	6.373
41	12.121	**11.707**	15.437
42	**7.976**	8.273	10.130
Mean	10.858	**10.530**	14.443
Median	9.504	**9.314**	12.374
Min.	**5.153**	5.213	6.373
Max.	24.519	**21.0840**	33.927

B_z, which is expected based on its importance in driving magnetic reconnection that allows energy entry into the magnetosphere. What is less expected is that the velocity V_x and the rectified electric field E_s are much less important for the storm peak values (SYM-H below -100 nT). In fact, the third most important feature is the dynamic pressure ρV_x^2. One would expect the dynamic pressure to be most important during the sudden storm commencement that produces a positive jump in SYM-H. Interestingly, the contributions of ρV_x^2 and B_z are comparable even for predicting positive SYM-H, except for the most positive values. Overall, we find that past SYM-H and B_z are the most important features. Density, velocity, the derived dynamic pressure and rectified electric field are comparable. The rest of the features, such as B_x, B_y and temperature provide quite small contributions. It is interesting to note that the rectified E_s is a less important contributor than B_z and the dynamic pressure, despite its physical significance of carrying the magnetic flux that induces dayside reconnection.

Figure 4 shows the contribution of various features of the model that is not using past SYM-H. As expected, B_z becomes the most important feature. Now velocity and density are the next most important features, especially for moderate values of SYM-H, and the dynamic pressure by itself does not have enough information (unlike in the previous case that used past SYM-H). Surprisingly, the rectified E_s is still a rather small contributor compared to B_z. This can be explained by examining the contributions of B_z and V_x: B_z completely dominates for large SYM-H values. On the other hand, the solar wind speed peaks at moderate storm values above -100 nT, and the contribution tapers off for the very strong storm values. While the electric field combines these two terms, one can see that they are most effective in different storms or different phases of the storm, suggesting that they need to be considered as independent variables rather than as a single parameter. The strong contribution of density for small and positive SYM-

Table 9. RMSEs for 1- and 2-hour ahead predictions using only the IMF as input (I_2) with our GBM model and the LSTM1 and CNN1 models of Siciliano et al. (2021). For 1-hour ahead predictions, the lowest RMSE in each row is shown in **bold**.

	1-hour ahead			2-hour ahead
Storm #	GBM	LSTM1	CNN1	GBM
26	**12.6**	18.0	19.8	12.9
27	20.1	**16.8**	23.4	20.9
28	**12.7**	18.6	14.4	12.4
29	**15.4**	21.1	20.0	16.7
30	**17.0**	24.2	25.8	17.1
31	**28.5**	32.5	32.1	29.6
32	21.8	23.4	**18.9**	21.9
33	35.7	33.8	**26.7**	38.1
34	**15.3**	17.9	16.6	15.5
35	**16.9**	21.3	18.6	17.3
36	**16.2**	20.4	21.4	16.8
37	41.6	42.6	**36.9**	42.7
38	**10.5**	18.6	13.0	10.6
39	**13.0**	20.3	16.5	12.8
40	10.9	13.6	**9.2**	10.6
41	**23.2**	27.3	25.4	23.7
42	16.9	17.8	**16.7**	17.1
Mean	**19.3**	22.8	20.9	19.8
Median	**16.9**	20.8	19.9	17.1
Min.	10.5	13.6	**9.2**	10.6
Max.	41.6	42.6	**36.9**	42.7

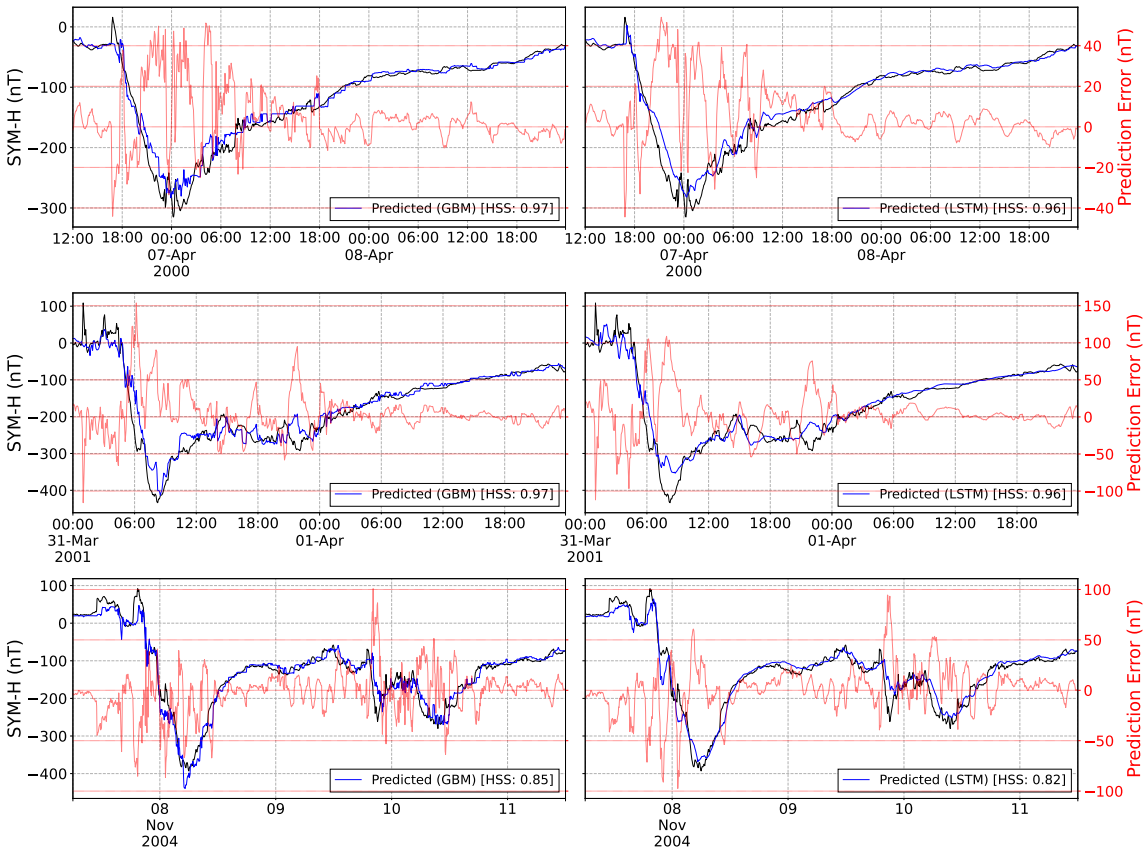

Figure 2. 1-hour ahead predictions for the 3 strongest geomagnetic storms in the test set during the main and recovery phases from our GBM (left column) and the LSTM2 developed by Collado-Villaverde et al. (2021) (right column). The observed SYM-H (black), the predicted SYM-H (blue) and the error (red) are shown for storms 31, 33, and 37 in the 3 rows, respectively. The HSS for each prediction is calculated using a threshold of -100 nT.

H values speaks to the importance of density pulses that often are found at the leading edges of solar wind structures impacting the Earth (Kilpua et al., 2017).

4.3.1 November 2004 Storm

We now look into how the prediction is obtained during the strongest test storm. Figure 5 shows the absolute and relative contributions of various features to the 1-hour and 2-hour ahead predictions of SYM-H during the November 2004 geomagnetic storm. The minimum SYM-H is close to -400 nT for this extreme event, so the RMSE of about 30 nT for 1-hour and 39 nT for 2-hour forecast are quite accurate (top row). The absolute and relative contributions shown in the subsequent rows vary substantially during the storm. From 18:00 to 20:45 UT (following the Storm Sudden Commencement, SSC), the observed SYM-H is positive, and this is roughly captured by the model for 1-hour prediction, but is completely missed by the 2-hour forecast. This is not very surprising, since there is no information in the solar wind that would predict the sudden commencement prior to the arrival of the shock. The only reason the 1-hour prediction can get the SSC about half an hour rather than 1 hour late is the lead time provided by the time it takes the high speed solar wind to propagate from L1 to the Earth. The main con-

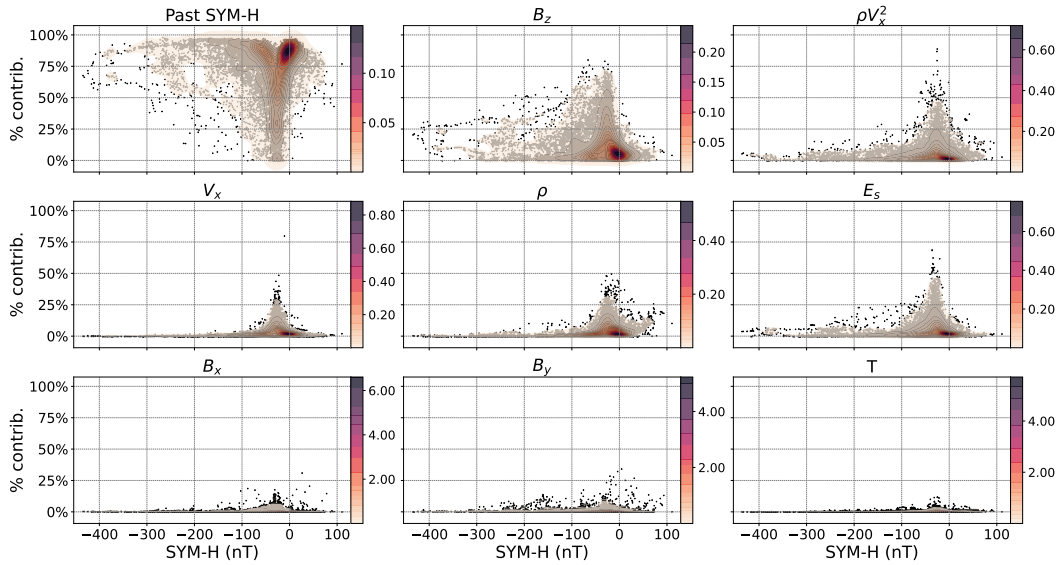

Figure 3. Scatter plot of percentage contributions (y-axis) against SYM-H (x-axis) for all the geomagnetic storms. The panels show the contributions of all considered features to the 1-hour ahead GBM prediction. Kernel density estimates using a Gaussian kernel are shown in color with the corresponding color legend on the right of each scatter plot.

tributors to the 1-hour prediction during this period are the density and dynamic pressure, and to some extent the IMF B_z. Based on our physical understanding, we would expect the dynamic pressure to be a more important predictor than the density, but that is clearly not the case, perhaps associated with the relatively constant value of the solar wind speed over that period.

During the main phase (22:00 Nov 7 to 06:00 Nov 8) of the storm, the SYM-H gradually drops to its minimum value near -400 nT. Focusing on the two-hour prediction, the relative contribution of B_z peaks around 22:00 on November 7, and 01:00 and after 04:00 UT. The first peak corresponds to the time when B_z decreases rapidly to nearly -50 nT value. The following period of very intense southward IMF shows initially low contribution from B_z, but then consistently high values with a peak at 04:00 close to the SYM-H minimum demarking the end of the storm main phase. Interestingly, the contribution from B_y, while generally low, has a broad peak between 20:00 and 00 UT on November 7. During that period, B_y is first positive and then turns strongly negative. As the B_z is negative during that time, the strong B_y component adds to the efficiency of the dayside reconnection process, which may account for its independent role as a predictor. Finally, during the recovery phase the prior SYM-H dominates (SYM-H evolution dominated by internal ring current loss processes), with B_z playing a secondary role.

Figure 6 shows the contribution of features as a function of time when the prior SYM-H is not used. The RMSE values become 33 nT and 37 nT for the 1 and 2-hour predictions, respectively. For the 1-hour prediction, RMSE slightly increases by about 3 nT, but for the 2-hour prediction, RMSE surprisingly decreases by roughly 2 nT. This suggests that there is no additional information from the 2-hour old SYM-H compared to what the model can infer from a longer history of L1 observations, at least for this event. If this held in general, it would put a prediction window limit on using past SYM-H for data assimilation purposes. Another unexpected result is that the 1-hour prediction misses the positive SYM-H period despite using the dynamic pressure. This is in contrast with

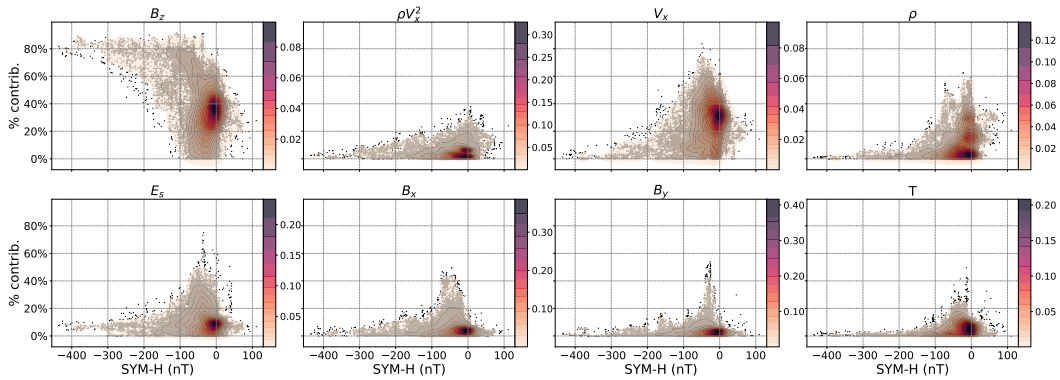

Figure 4. Scatter plot of percentage contributions (y-axis) against SYM-H (x-axis) from solar wind and IMF parameters for 1-hour ahead prediction from GBM using only solar wind and IMF parameters as input. Kernel density estimates using a Gaussian kernel are shown in color with the corresponding color legend on the right of each scatter plot.

the 1-hour prediction that includes past SYM-H, which produced a larger positive SYM-H, although still lower than observed.

The relative contributions (bottom row) show a rather complicated and interesting pattern. In the initial storm period 18:00 to 21:00 UT, when the observed SYM-H is positive, the main contributors are density and velocity. Once SYM-H goes negative, B_z gradually becomes the main contributing feature with E_s and, surprisingly, B_x (for 1-hour prediction) and B_y (for 2-hour prediction) being the second and third most important. Once SYM-H drops below -100 nT, the contribution from B_z becomes dominant and this remains true during the whole recovery phase. The other features start to contribute more after 12:00 UT Nov 8 when B_z turns positive. Even with positive B_z, however, the main contributor remains B_z. This shows that the rectified E_s, which simply zeroes out the electric field for positive B_z, is throwing away potentially important information.

4.3.2 January 2004 Storm

Next, we study the storm of January 2004 that has a minimum SYM-H of about -140 nT, so it is an intense storm, but not as extreme as the November 2004 super storm. As shown in figure 7, this is a very complicated storm due to the highly variable B_z field in the CME sheath (00:00 UT to 11:00 UT Jan 22) preceding the magnetic cloud with consistently negative B_z. The model prediction has 14.22 nT and 19.96 nT RMSE for the 1- and 2-hour predictions, respectively, which is quite good for such a complicated event. In the ICME sheath, the main contributor is the previous SYM-H followed by the dynamic pressure.

The 1-hour ahead model predicts the jump of SYM-H from 0 to about +30 nT at 2:00UT, which is about half an hour late compared to observations. This cannot be based on prior SYM-H that is observed 1 hour earlier, and it is clearly obtained from the dynamic pressure as expected from physical understanding. The 2-hour prediction, however, completely misses predicting positive SYM-H values (except for following the increase of the observed SYM-H with a 2-hour delay), similarly to the extreme event case.

Between 01:00 and 11:00 UT the main contributors are the prior SYM-H and the dynamic pressure, with B_z playing a minor role only. After 11:00 UT, however, B_z turns consistently negative and it becomes the main contributor of predicting the main phase

of the storm 1 hour or 2 hours later for the two models, respectively. The 2-hour prediction also relies heavily on B_y between 10 and 12:00 UT. A possible explanation is that the strong magnetic field in the magnetic cloud rotates, so a strong signal in B_x or B_y may be a predictor for a strong, possibly negative, B_z value that has strong geomagnetic impact.

The model correctly predicts the minimum value of SYM-H, but it is late by an hour and two hours for the 1- and 2-hour predictions, respectively. This means that the prior SYM-H was the primary contributor to the prediction of the minimum SYM-H. We note that the last available B_z is negative, but has a small amplitude at this point (about $-5\,\mathrm{nT}$). Clearly the model is not capable of predicting the behavior of the storm very well during this time period for this particular event. The recovery phase is correctly captured with the prior SYM-H dominating, as expected. B_z becomes slightly more negative from 19:00 to 23:00, and the importance of B_z and E_s becomes significant during this time correctly predicting the slow down of the recovery, although with considerable delay.

Figure 8 shows the model predictions for the January 2004 storm without relying on the prior SYM-H values. The RMSE is around $33\,\mathrm{nT}$ for both the 1-hour and 2-hour ahead forecast. The positive SYM-H values are missed by the model and in fact there is a considerable underprediction of SYM-H until 11:00UT. The main phase of the storm corresponding the rapid decrease of SYM-H is quite well captured. It is slightly too early for the 1-hour prediction, and quite spot on for the 2-hour prediction. The minimum SYM-H is correctly predicted by both models with an hour delay, and it is actually somewhat better predicted by the 2-hour ahead model. The recovery phase is reasonably predicted, although the predicted recovery rate is somewhat slower than what is observed.

The main contributors to the prediction before 11:00UT are velocity,the rectified electric field and density. During the main phase and the recovery, B_z becomes an important contributor, but the velocity and E_s still play considerable roles. Quite surprisingly, B_x becomes the most important contributor during the recovery phase. Figure 4 confirms that B_x and B_y become significant contributors when prior SYM-H is not used.

One of the surprises mentioned above was that B_z is a better predictor than E_s. However, these features are highly correlated so it is not clear if the GBM prefers B_z over E_s by chance only. To investigate this question, we have performed experiments to see whether B_z or E_y, or the rectified E_s is the best predictor out of the three for future SYM-H. To make E_y (or E_s) and B_z fully independent of each other, we have removed the V_x and ρV_x^2 features and used only one the three quantities (B_z, E_y, and rectified E_s) together with density and temperature while training the GBM. The RMSE values are shown in Table 10 including both cases with and without prior SYM-H.

Table 10. RMSE from models with only one of B_z, E_y, and E_s included as input calculated using all test storms. The RMSE from a model trained with B_z, E_s, and ρV_x^2 is shown in the last column as reference. For these experiments, density and temperature were also used as features.

	B_z	E_y	E_s	B_z, E_s, ρV_x^2
Including SYM-H	7.35	8.00	8.26	7.26
Excluding SYM-H	20.84	21.12	21.45	18.39

Based on the RMSE values in the table, we conclude that B_z is the best predictor followed by E_y and the rectified E_s. It is also interesting to see that past SYM-H and B_z together are pretty much all that the model needs. The velocity V_x, for example, plays no significant role in contributing to the quality of the prediction as it only improves the

RMSE from 7.35 to 7.26 nT. When past SYM-H is not used, the velocity plays a more important role by improving the RMSE from 20.84 to 18.39, but still much less important than B_z, E_y or E_s. A possible reason may be that V_x varies only about a factor of 2 between about -350 km/s and -700 km/s even during storm events (see fig. 9).

5 Discussion and conclusions

We apply an interpretable machine learning method to quantify the contribution of prior SYM-H values, solar wind, IMF, and derived parameters to predictions of the SYM-H index 1 to 2 hours ahead. In particular, gradient boosting machines (GBM) are used and the interpretation is based on the TreeSHAP method. We showed that gradient boosting machines have similar, if not better, performance compared to the less interpretable but highly effective LSTM method for forecasting the SYM-H index.

From the quantified feature contributions, we were able to show that our proposed model makes predictions in a physically consistent manner, while also challenging some of the commonly assumed relationships among the interplanetary magnetic field, the solar wind and the formation of Earth's ring current. In particular, we found that past SYM-H and B_z are the most important features overall but feature contributions vary depending on the storm phase and the storm itself. During the storm sudden commencement, past SYM-H, density, velocity, and to some extent, dynamic pressure and electric field, became the main contributors to predictions. As SYM-H decreases during the main phase, past SYM-H and B_z played an increasingly larger role.

SHAP values revealed several surprising ways that our models made predictions during the two storms we investigated in detail: density and velocity had a larger independent contribution than dynamic pressure during the storm sudden commencement; B_y had a non-negligible contribution during the storm sudden commencement and main phase; and B_z was a better predictor than the rectified E_s. Further efforts will be made to investigate the robustness of these findings and to perform a comparison of feature contributions for many different storms.

Along with gray-box approaches, this work takes the first steps in making machine learning methods more reliable and trustworthy for operational forecasting of geomagnetic activity. However, explanation methods like SHAP should be used with caution, especially in high-stakes decision making, as they do not always provide explanations that are faithful to the original model (Rudin, 2019). Thus, developing highly accurate but intrinsically interpretable models should be prioritized. In addition to interpretability, quantified uncertainty is also equally as important. Consequently, we will devote future efforts to developing interpretable methods for forecasting other types of geomagnetic indices and geomagnetic activity that also estimate predictive uncertainty.

Acknowledgments

We thank Austin Brenner, Qusai Al Shidi, and Professor Michael Liemohn from the Dept. of Climate and Space Sciences and Engineering (CLaSP) at the University of Michigan for helpful comments and fruitful discussions. This work was supported by NASA DRIVE Science Center grant 80NSSC20K0600, NASA MMS grant 80NSSC19K0564, NSF PRE-EVENTS grant 1663800 and NSF SWQU grant PHY-2027555. EC is partially funded by NASA under grants 80NSSC20K1580 and 80NSSC20K1275. The ACE level 2 data used as inputs into our models in this study is available through NASA/GSFC's Space Physics Data Facility's (SPDF) Coordinated Data Analysis Web (CDAWeb) at `https://cdaweb.gsfc.nasa.gov/`. The SYM-H index data is available through SPDF's OMNIWeb at `https://omniweb.gsfc.nasa.gov/`. All relevant digital materials used in this manuscript will be permanently archived at the University of Michigan (UM) Library Deep Blue Data Repository (`https://deepblue.lib.umich.edu/data`), which is specif-

ically designed for UM researchers to share their research data and to ensure its long-term viability. Processed data will be assigned Digital Object Identifiers (DOIs), which will serve as identifiers for the data, enabling them to be cited in publications.

References

Al Shidi, Qusai. (2020). *swmfpy.* Retrieved from `https://gitlab.umich.edu/swmf_software/swmfpy` (version 2020.5)

Bailey, R. L., Reiss, M. A., Arge, C. N., Möstl, C., Owens, M. J., Amerstorfer, U. V., ... Hinterreiter, J. (2021). Using gradient boosting regression to improve ambient solar wind model predictions. *arXiv:2006.12835 [astro-ph, physics:physics]*.

Bauer, E., & Kohavi, R. (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning*, *36*(1/2), 105–139. doi: 10.1023/A:1007515423169

Bhaskar, A., & Vichare, G. (2019). Forecasting of SYMH and ASYH indices for geomagnetic storms of solar cycle 24 including St. Patrick's day, 2015 storm using NARX neural network. *Journal of Space Weather and Space Climate*, *9*. doi: 10.1051/swsc/2019007

Bluwstein, K., Buckmann, M., Joseph, A., Kang, M., Kapadia, S., & Simsek, Ö. (2020). Credit Growth, the Yield Curve and Financial Crisis Prediction: Evidence from a Machine Learning Approach. *SSRN Journal*. doi: 10.2139/ssrn.3520659

Bortnik, J., Chu, X., Ma, Q., Li, W., Zhang, X., Thorne, R. M., ... others (2018). Artificial neural networks for determining magnetospheric conditions. In *Machine learning techniques for space weather* (pp. 279–300). Elsevier. doi: 10.1016/B978-0-12-811788-0.00011-1

Boynton, R., Balikhin, M., Billings, S., Sharma, A., & Amariutei, O. (2011). Data derived narmax dst model. In *Annales geophysicae* (Vol. 29, pp. 965–971).

Breiman, L. (1996). Bagging predictors. *Mach Learn*, *24*(2), 123–140. doi: 10.1007/BF00058655

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Wadsworth and Brooks.

Burton, R. K., McPherron, R. L., & Russell, C. T. (1975). An empirical relationship between interplanetary conditions and Dst. *Journal of Geophysical Research*, *80*, 4204. doi: 10.1029/JA080i031p04204

Cai, L., Ma, S. Y., & Zhou, Y. L. (2010). Prediction of SYM-H index during large storms by NARX neural network from IMF and solar wind data. *Annales Geophysicae*, *28*(2), 381–393. doi: 10.5194/angeo-28-381-2010

Camporeale, E. (2019). The challenge of machine learning in space weather: Nowcasting and forecasting. *Space Weather*, *17*(8), 1166–1207. doi: 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. *J. Geophys. Res. Space Physics*, *125*(11). doi: 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. doi: 10.1002/2017SW001627

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. doi: 10.1145/2939672.2939785

Collado-Villaverde, A., Muñoz, P., & Cid, C. (2021). Deep Neural Networks With Convolutional and LSTM Layers for SYM-H and ASY-H Forecasting. *Space Weather*. doi: 10.1029/2021SW002748

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, *29*(5), 1189–1232. doi: 10.1214/aos/1013203451

Gleisner, H., Lundstedt, H., & Wintoft, P. (1996). Predicting geomagnetic storms from solar-wind data using time-delay neural networks. *Ann. Geophys.*, *14*(7), 679–686. doi: 10.1007/s00585-996-0679-1

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York, NY, USA: Springer New York Inc.

Kilpua, E. K. J., Balogh, A., von Steiger, R., & Liu, Y. D. (2017). Geoeffective Properties of Solar Transients and Stream Interaction Regions. *Space Sci. Rev.*, *212*, 1271-1314. doi: 10.1007/s11214-017-0411-3

King, J. H. (2005). Solar wind spatial scales in and comparisons of hourly Wind and ACE plasma and magnetic field data. *Journal of Geophysical Research*, *110*(A2), 2104. doi: 10.1029/2004JA010649

Korlakai Vinayak, R., & Gilad-Bachrach, R. (2015). DART: Dropouts meet multiple additive regression trees. In G. Lebanon & S. V. N. Vishwanathan (Eds.), *Proceedings of the eighteenth international conference on artificial intelligence and statistics* (Vol. 38, pp. 489–497). San Diego, California, USA: PMLR.

Kuzlu, M., Cali, U., Sharma, V., & Guler, O. (2020). Gaining Insight Into Solar Photovoltaic Power Generation Forecasting Utilizing Explainable Artificial Intelligence Tools. *IEEE Access*, *8*, 187814–187823. doi: 10.1109/ACCESS.2020.3031477

Liemohn, M. W., Mccollough, J. P., Jordanova, V. K., Ngwira, C. M., Morley, S. K., & Cid, C. (2018). Model evaluation guidelines for geomagnetic index predictions. *Space Weather*, 16. doi: 10.1029/2018SW002067

Liu, L., Zou, S., Yao, Y., & Wang, Z. (2020). Forecasting Global Ionospheric TEC Using Deep Learning Approach. *Space Weather*, *18*(11). doi: 10.1029/2020SW002501

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 623–631). Chicago Illinois USA: ACM. doi: 10.1145/2487575.2487579

Lu, J., Peng, Y., Wang, M., Gu, S., & Zhao, M. (2016). Support vector machine combined with distance correlation learning for dst forecasting during intense geomagnetic storms. *Planetary and Space Science*, *120*, 48–55. doi: 10.1016/j.pss.2015.11.004

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2019). Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv:1802.03888 [cs, stat]*.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., ... Lee, S.-I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*, *2*(10), 749–760. doi: 10.1038/s41551-018-0304-0

Lundstedt, H., & Wintoft, P. (1994). Prediction of geomagnetic storms from solar wind data with the use of a neural network. *Ann. Geophys.*, *12*(1), 19–24. doi: 10.1007/s00585-994-0019-2

Mayaud, P. N. (1980). The dst index. In *Derivation, meaning, and use of geomagnetic indices* (p. 115-129). American Geophysical Union (AGU). doi: 10.1002/9781118663837.ch8

McComas, D. J., Bame, S. J., Barker, P., Feldman, W. C., Phillips, J. L., Riley, P., & Griffee, J. W. (1998). Solar Wind Electron Proton Alpha Monitor

(SWEPAM) for the Advanced Composition Explorer. *Space Science Reviews*, *86*, 563-612. doi: 10.1023/A:1005040232597

Mitrentsis, G., & Lens, H. (2021). An Interpretable Probabilistic Model for Short-Term Solar Power Forecasting Using Natural Gradient Boosting. *arXiv:2108.04058 [cs, stat]*.

Mokhtari, K. E., Higdon, B. P., & Başar, A. (2019). Interpreting financial time series with shap values. In *Proceedings of the 29th annual international conference on computer science and software engineering* (p. 166–172). USA: IBM Corp.

Molnar, C. (2019). *Interpretable machine learning.* (`https://christophm.github.io/interpretable-ml-book/`)

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. *arXiv:2010.09337 [cs, stat]*.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Front. Neurorobot.*, *7*. doi: 10.3389/fnbot.2013.00021

Newell, P. T., Sotirelis, T., Liou, K., Meng, C. I., & Rich, F. J. (2007). A nearly universal solar wind-magnetosphere coupling function inferred from 10 magnetospheric state variables. *Journal of Geophysical Research*, *112*, 01206. doi: 10.1029/2006JA012015

Pallocchia, G., Amata, E., Consolini, G., Marcucci, M. F., & Bertello, I. (2006). Geomagnetic Dst index forecast based on IMF data only. *Ann. Geophys.*, *24*(3), 989–999. doi: 10.5194/angeo-24-989-2006

Rapin, J., & Teytaud, O. (2018). *Nevergrad - A gradient-free optimization platform.* `https://GitHub.com/FacebookResearch/Nevergrad`. GitHub.

Redxgit. (2021). *Redxgit/supporting_data_paper_sw: Space Weather Submission.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.4562456` doi: 10.5281/zenodo.4562456

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. doi: 10.1038/s42256-019-0048-x

Saabas, A. (2014). *Interpreting random forests — Diving into data.* Retrieved from `https://blog.datadive.net/interpreting-random-forests/`

Sakaguchi, K., Nagatsuma, T., Reeves, G. D., & Spence, H. E. (2015). Prediction of MeV electron fluxes throughout the outer radiation belt using multivariate autoregressive models. *Space Weather*, *13*(12), 853–867. Retrieved 2021-08-05, from `https://onlinelibrary.wiley.com/doi/10.1002/2015SW001254` doi: 10.1002/2015SW001254

Shapley, L. S. (1953). A Value for n-Person Games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the Theory of Games (AM-28), Volume II* (pp. 307–318). Princeton University Press. doi: 10.1515/9781400881970-018

Shwartz-Ziv, R., & Armon, A. (2021). Tabular Data: Deep Learning is Not All You Need. *arXiv:2106.03253 [cs]*.

Siciliano, F., Consolini, G., Tozzi, R., Gentili, M., Giannattasio, F., & De Michelis, P. (2021). Forecasting SYM-H Index: A Comparison Between Long Short-Term Memory and Convolutional Neural Networks. *Space Weather*, *19*(2). doi: 10.1029/2020SW002589

Smith, C. W., L'Heureux, J., Ness, N. F., Acuña, M. H., Burlaga, L. F., & Scheifele, J. (1998). The ACE Magnetic Fields Experiment. *Space Science Reviews*, *86*, 613-632. doi: 10.1023/A:1005092216668

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*(56), 1929–1958.

Stirnberg, R., Cermak, J., Kotthaus, S., Haeffelin, M., Andersen, H., Fuchs, J., ... Favez, O. (2020). *Meteorology-driven variability of air pollution (PM1) revealed with explainable machine learning* (Preprint). Aerosols/Field Mea-

surements/Troposphere/Physics (physical properties and processes). doi: 10.5194/acp-2020-469

Sugiura, M. (1964). Oart 1. In *Hourly values of equatorial dst for the igy.* Pergamon Press.

Tsyganenko, N. (1989). A magnetospheric magnetic field model with a warped tail current sheet. *Planetary and Space Science*, *37*(1), 5–20. doi: 10.1016/0032-0633(89)90066-4

Tsyganenko, N. A. (1995). Modeling the Earth's magnetospheric magnetic field confined within a realistic magnetopause. *Journal of Geophysical Research*, *100*(A4), 5599. doi: 10.1029/94JA03193

Tsyganenko, N. A. (2002a). A model of the near magnetosphere with a dawn-dusk asymmetry 1. Mathematical structure: A NEW MAGNETOSPHERE MAGNETIC FIELD MODEL, 1. *Journal of Geophysical Research: Space Physics*, *107*(A8), SMP 12–1–SMP 12–15. doi: 10.1029/2001JA000219

Tsyganenko, N. A. (2002b). A model of the near magnetosphere with a dawn-dusk asymmetry 2. Parameterization and fitting to observations: A NEW MAGNETOSPHERE MAGNETIC FIELD MODEL, 2. *Journal of Geophysical Research: Space Physics*, *107*(A8), SMP 10–1–SMP 10–17. Retrieved 2021-08-05, from `http://doi.wiley.com/10.1029/2001JA000220` doi: 10.1029/2001JA000220

Wanliss, J. A., & Showalter, K. M. (2006). High-resolution global storm index: Dst versus sym-h. *Journal of Geophysical Research: Space Physics*, *111*(A2). Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JA011034` doi: https://doi.org/10.1029/2005JA011034

Wu, J.-G., & Lundstedt, H. (1997). Geomagnetic storm predictions from solar wind data with the use of dynamic neural networks. *J. Geophys. Res.*, *102*(A7), 14255–14268. doi: 10.1029/97JA00975

Xu, S. B., Huang, S. Y., Yuan, Z. G., Deng, X. H., & Jiang, K. (2020). Prediction of the Dst Index with Bagging Ensemble-learning Algorithm. *ApJS*, *248*(1), 14. doi: 10.3847/1538-4365/ab880e

Zhang, T., & Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *Ann. Statist.*, *33*(4). doi: 10.1214/009053605000000255

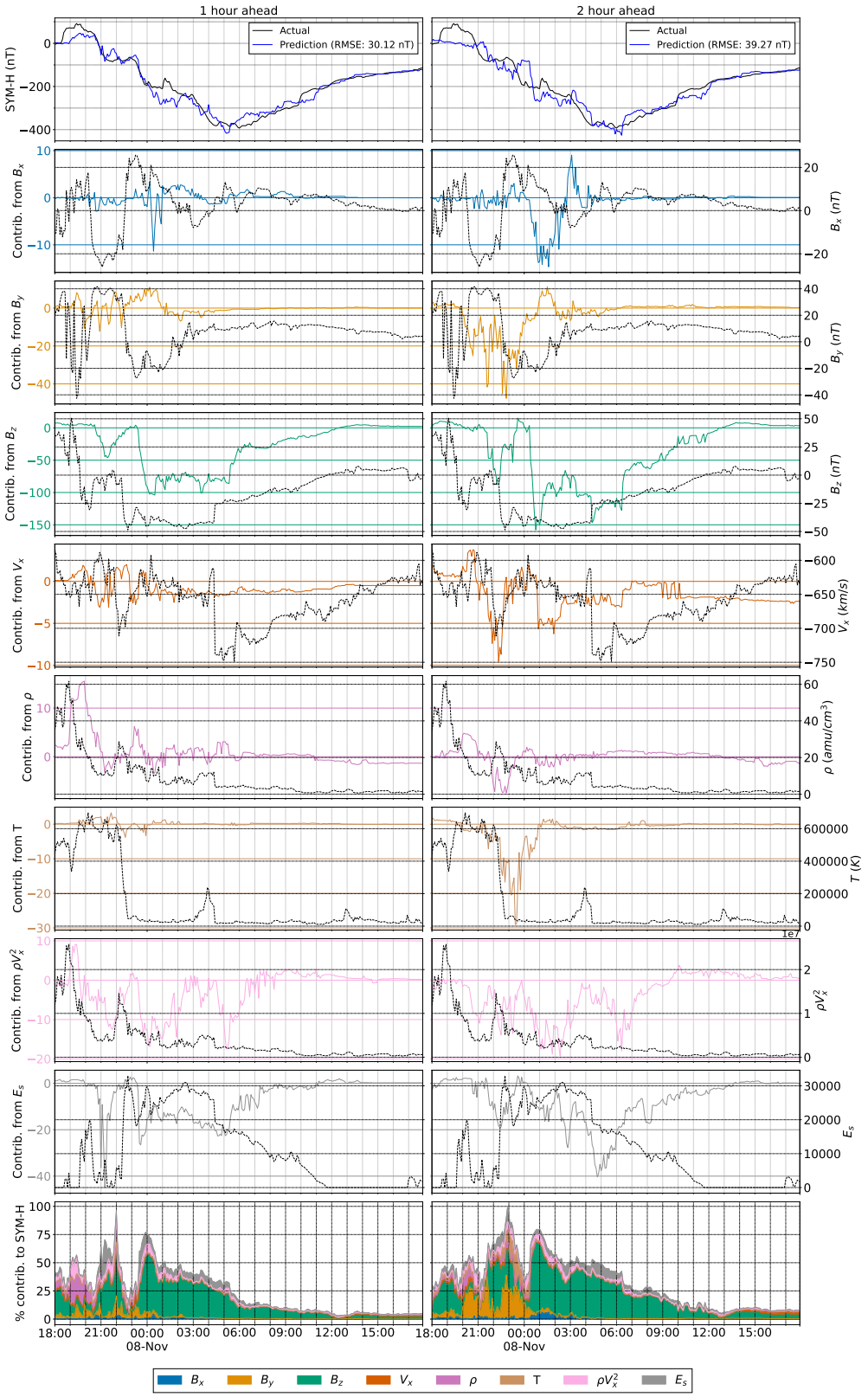

Figure 5. 1-hour (left) and 2-hour (right) ahead predictions for the Nov. 2004 storm using GBM trained on all considered features. The first row shows the observed (black) and predicted (blue) SYM-H values. Rows 2-9 show the contributions from each feature (left axis, colored) and its value (right axis, black). The percentage contributions are shown in the last row. The contribution from past SYM-H on predictions is omitted, but its percentage contribution is implicitly shown as the remaining white area in the last row.

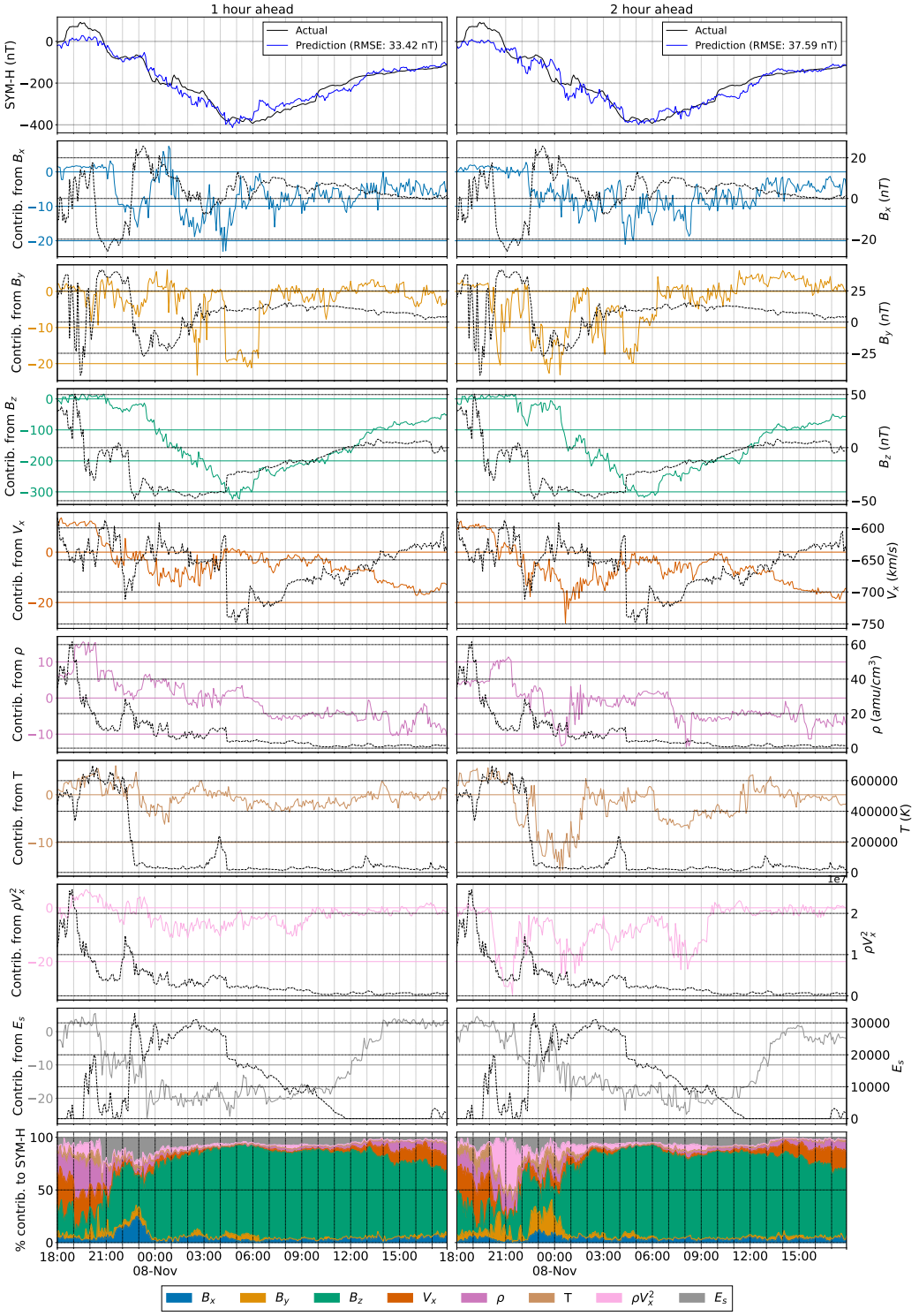

Figure 6. 1-hour (left) and 2-hour (right) ahead predictions for the Nov. 2004 storm using GBM trained on only solar wind and IMF parameters (first row), corresponding feature contributions and values (rows 2-9), and percentage contributions (last row).

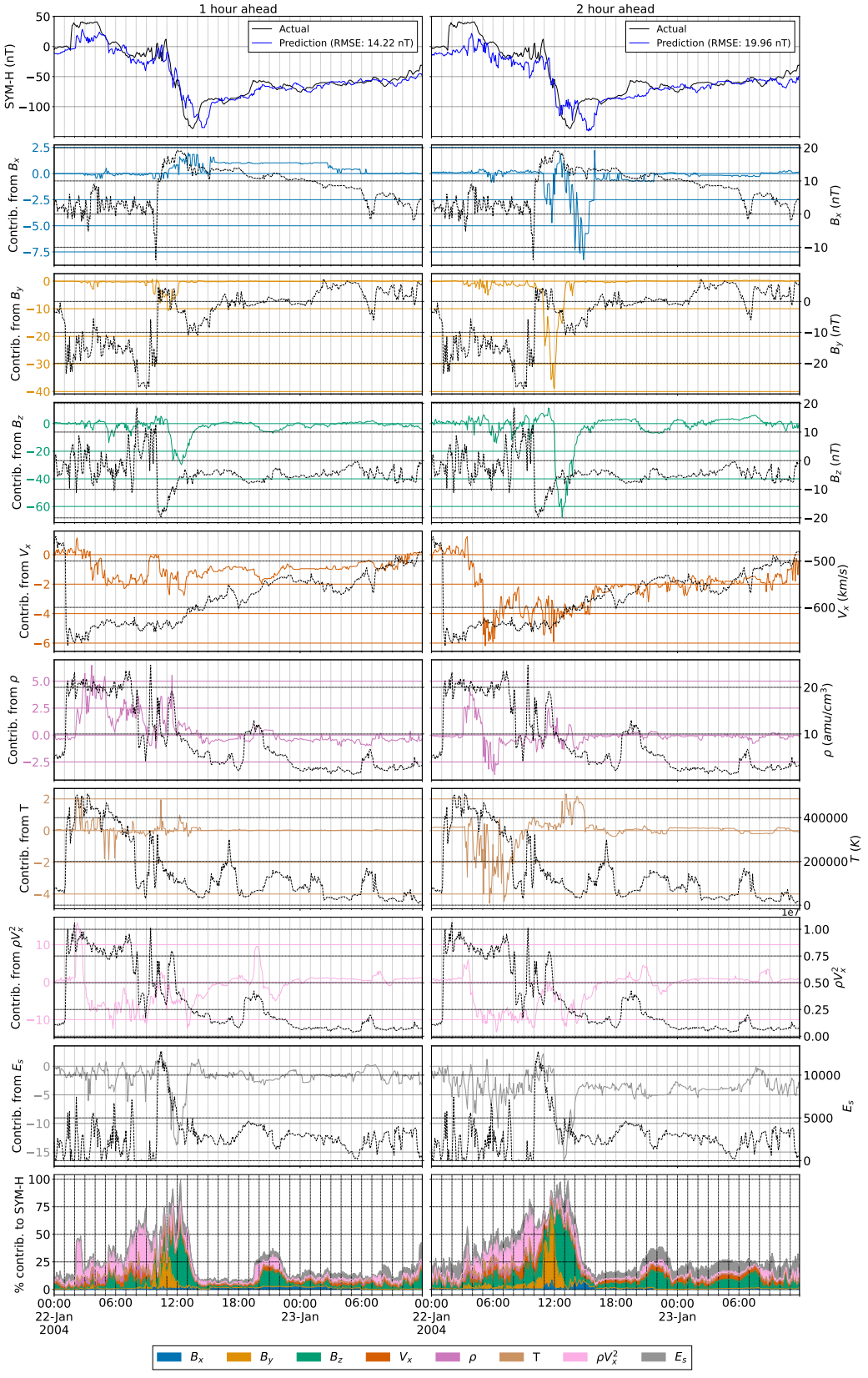

Figure 7. 1-hour (left) and 2-hour (right) ahead predictions for the Jan. 2004 storm using GBM trained on all considered features (first row), corresponding feature contributions and values (rows 2-9), and percentage contribution (last row). The contribution from past SYM-H on predictions is omitted but the percentage contribution is implicitly shown as the remaining white area in the last row.

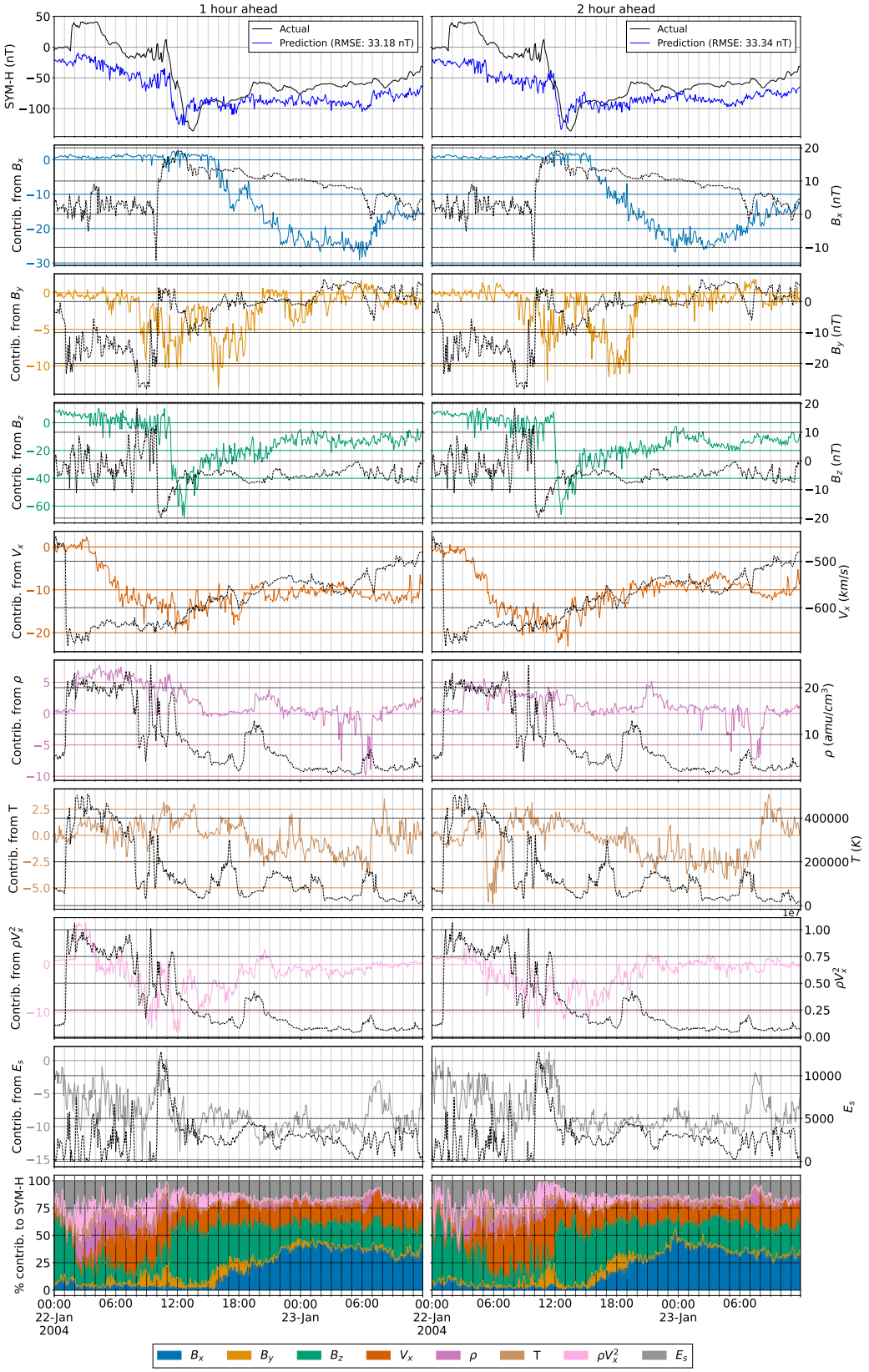

Figure 8. 1-hour (left) and 2-hour (right) ahead predictions for the Jan. 2004 storm using GBM trained on all considered features (first row), corresponding feature contributions (rows 2-9), and percentage contribution (last row).

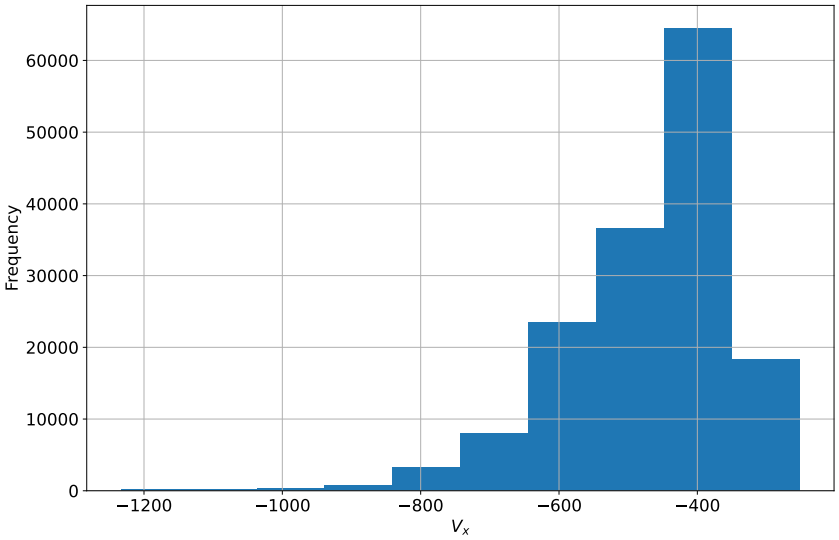

Figure 9. Histogram of V_x for all storms in tables 1 and 2.