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

Daniel Iong[1,1,1], Yang Chen[2,2,2], Gabor Toth[3,3,3], Shasha Zou[3,3,3], Tuija I. Pulkkinen[2,2,2], Jiaen Ren[4,4,4], Enrico Camporeale[5,5,5], and Tamas I. I. Gombosi[3,3,3]

[1]University of Michigan, Ann Arbor
[2]University of Michigan
[3]University of Michigan-Ann Arbor
[4]University of Michigan - Ann Arbor
[5]University of Colorado

November 30, 2022

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.

New Findings from Explainable SYM-H Forecasting using Gradient Boosting Machines

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:

- [illegible]
- [illegible]
- [illegible]

Corresponding author: Daniel Iong, [illegible]

Abstract

[illegible]

Plain Language Summary

[illegible]

1 Introduction

[illegible]

[illegible]

2 Data

[illegible] 16- [illegible] ([illegible] , 1998) , [illegible] 64- [illegible] IMF [illegible] field, [illegible] ([illegible] , 2007) .

[illegible] SYM-H [illegible] IMF [illegible] 5- [illegible] (2021) ; [illegible] (2021) .

[illegible] GBM [illegible] 3. 1, [illegible] 1998 [illegible] 2018 [illegible] SYM-H [illegible] -100 nT [illegible] 1 [illegible] 2 [illegible] (section 3. 1) [illegible] A1 [illegible] A2.

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).

St#	Start	End	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

[illegible] SYM-H Δt [illegible] t, [illegible] $y(t+\Delta t)$, [illegible] 3. [illegible] Δt [illegible] SYM-H [illegible]

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

St#	Start Date	End Date	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

[illegible] SYM-H [illegible] 30 [illegible] SYM-H [illegible] (2021), [illegible] R^2 [illegible] SYM-H [illegible] R^2 [illegible] 30 [illegible] SYM-H [illegible] SYM-H [illegible] R^2 [illegible] 90 to 180 [illegible] R^2 [illegible] 180 [illegible]

Table 3. Features used as input into our models.

Feature	History (min)
SYM-H (nT)	1
IMF: B_x, B_y, B_z (nT)	2 to 30
Solar wind: V_x (km/s), ρ (p/cm^3), T (K)	2 to 30
Derived quantities: ρV_x^2 (nPa), $E_s = \max(0, -\lvert V_x \rvert B_z)$ (mV/m)	2 to 30

[illegible] IMF [illegible] SYM-H [illegible] I_1, [illegible] IMF [illegible] 2% [illegible] 9% [illegible] 3.4 in [illegible] (2016) [illegible].

[illegible] I_3 [illegible] I_4 [illegible] ρV_x^2 [illegible] SYM-H, [illegible] y [illegible] $E_y = V_x B_z$ [illegible] $V_x < 0$ [illegible] (GSM) [illegible] $E_s = (0, E_y)$ [illegible] E_y [illegible] IMF [illegible] $B_z < 0$), [illegible] IMF [illegible] (B[illegible], 1975). [illegible] E_s [illegible] (T. P. O'B[illegible] & M[illegible], 2000; T. P. O'B[illegible], 2002).

[illegible] IMF [illegible] SYM-H [illegible] I_2 [illegible] I_4 [illegible] SYM-H [illegible] (see Table 4).

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

Inputs	Features
I_1	IMF, [illegible] SYM-H
I_2	IMF
I_3	IMF/[illegible], [illegible] SYM-H
I_4	IMF/[illegible]

3 Methods

3.1 Gradient Boosting Machines

G[illegible] (GBMs), [illegible] (N[illegible] & K[illegible], 2013). [illegible] (2021) [illegible] GBMs [illegible] GBMs [illegible] K[illegible] (C[illegible] & G[illegible], 2016). In [illegible] GBMs [illegible] (B[illegible], 2021) [illegible] (X[illegible], 2020).

In [illegible] (F[illegible], 2001). In [illegible] GBMs [illegible] GBMs [illegible] In [illegible] (B[illegible], 1996).

[illegible] SYM-H [illegible]

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

[illegible] $I(t)$ [illegible] t; $\epsilon(t)$ [illegible] t; T_m' [illegible] M [illegible] N [illegible] α [illegible] $I(t)$ [illegible] I_2 [illegible] $I(t) = (B_x(t), \ldots, B_x(t-115), B_y(t), \ldots, B_y(t-115), B_z(t), \ldots, B_z(t-115))$, [illegible] $B_z(t-60)$ [illegible] B_z 60 [illegible]

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

[illegible] w [illegible] q [illegible] Fig. 1 [illegible] GBM [illegible]

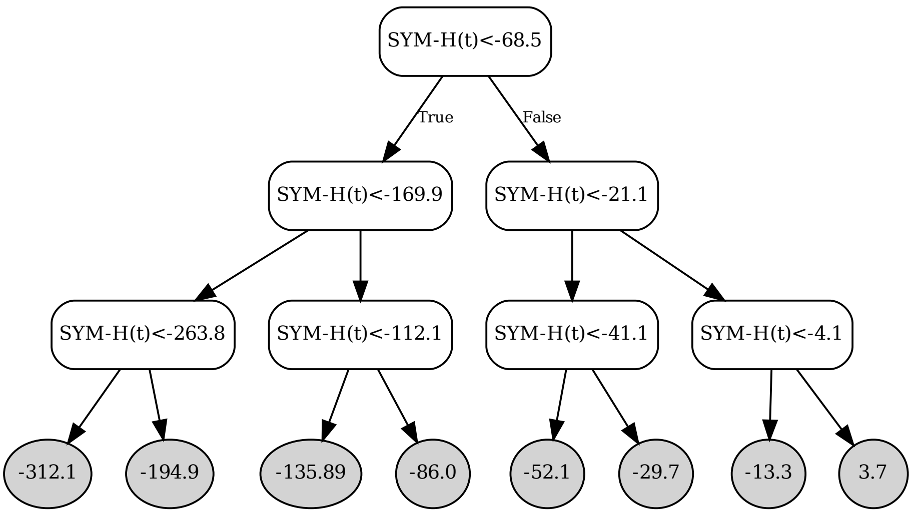

Figure 1. Structure of the first tree T_1 learned in a GBM trained with input set I_3 to predict the SYM-H index one hour ahead. The leaf nodes of the tree are shaded gray. The value in each leaf node is its corresponding leaf weight. Left splits correspond to the inequality in the previous node being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

[illegible] GBM [illegible] XGB [illegible] (Chen & Guestrin, 2016). [illegible] m, [illegible] T_m [illegible]

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

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

In eq. (4), K_j [illegible] T_j; $w_{j,k}$' [illegible] T_j; [illegible] γ [illegible] λ [illegible] Ω [illegible] I[illegible] γ [illegible] λ [illegible] A[illegible] B[illegible] λ, [illegible] It [illegible] GB [illegible] A[illegible] 3 of Chen and Guestrin (2016).

[illegible] (2018). [illegible] (2005). [illegible] 5 [illegible] 4 [illegible] 5.

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

Inputs	H[illegible]	V[illegible]
I_1, I_2	L[illegible]	0.072
	M[illegible]	4
	M[illegible]	4
	C[illegible]%	0.78
	# [illegible]	84
I_3, I_4	L[illegible]	0.147
	M[illegible]	3
	M[illegible]	2
	C[illegible]%	0.894
	# [illegible]	291

GBMs [illegible] GBMs [illegible] A[illegible] GBMs [illegible] A[illegible] (H[illegible], 2001). [illegible] fig 2, [illegible] GBM [illegible] LSTM [illegible] GBMs [illegible]

[illegible]

CGBMs [illegible] (2015). [illegible] CGBMs [illegible] (2015). [illegible] (2014). [illegible] CGBMs [illegible] $\hat{y}^{(m-1)}$ in eq (3) [illegible] (2015).

3.2 Feature Importance

[illegible] (1984), [illegible] (2001), [illegible] (2016). In [illegible]

[illegible] (2019), [illegible] (SHAP [illegible] (2017). SHAP [illegible] (1953), [illegible] In [illegible] SHAP [illegible] SHAP [illegible]

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

[illegible] ϕ_0 [illegible]; p [illegible] $\mathbf{z} = (z_1 \ldots z_p)'$, [illegible] z_i [illegible] i [illegible] ϕ_i [illegible] i. SHAP [illegible] eq (5) [illegible] SHAP [illegible] (2014), [illegible] SHAP [illegible] SHAP [illegible] CGBMs.

[illegible]

4 Results

[illegible]

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

[illegible]

[illegible]

$$\mathrm{FS}(y, \hat{y}, y_{\mathrm{burton}}) = 1 - \frac{\mathrm{MSE}(y, \hat{y})}{\mathrm{MSE}(y, y_{\mathrm{burton}})}, \tag{7}$$

[illegible] y_{burton} [illegible] $\mathrm{MSE}(y, \hat{y}) = (1/n)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. [illegible]

4.1 Comparison to existing methods

[illegible] (2021) (**LSTM1/CNN1**) [illegible] (2021) (**LSTM2**) [illegible]

[illegible] RMSEs [illegible] 5% [illegible] [illegible] I [illegible] [illegible] I [illegible] RMSEs [illegible]

4.1.1 1-[illegible]

[illegible] 6 [illegible] 7 [illegible] RMSE [illegible] 1 [illegible] SYM-H [illegible] GBM, LSTM1, LSTM2, [illegible] I [illegible] GBM [illegible] RMSE [illegible] GBM [illegible] 0.448 [illegible] (5.7%) [illegible] RMSE [illegible] LSTM2, [illegible] 1.138 [illegible] (13.3%) [illegible] RMSE [illegible] LSTM1, [illegible] 1.942 [illegible] (20.8%) [illegible] RMSE [illegible] GBM [illegible] RMSE [illegible] 14 [illegible] 17 [illegible] (26-32, 35-38, 41, 42). Fig 2 [illegible] 1 [illegible] GBM [illegible] LSTM2 [illegible] SYM-H < -300 [illegible] (31, 33, 37) [illegible] 2001 [illegible] fig 2), [illegible] GBM [illegible] SYM-H [illegible] -400 [illegible] 06:00 [illegible] 12:00 [illegible] 31 [illegible] A [illegible] A.

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, Burton equation (T. O'Brien & McPherron, 2000) and simple persistence. Here, the GBM, LSTM1, and LSTM2 were trained with past SYM-H and IMF parameters as inputs. The lowest RMSE for each row is shown in **bold**.

St#	GBM	LSTM2	LSTM1	Burton	Persistence
26	**5.863**	6.630	6.700	6.839	7.631
27	**7.729**	8.913	8.900	7.954	9.623
28	**4.281**	5.858	5.400	5.697	5.814
29	**5.833**	6.683	7.200	6.511	7.174
30	4.927	5.200	5.600	**4.614**	4.810
31	**8.277**	8.584	10.700	8.838	10.429
32	**6.841**	7.259	8.300	9.487	10.528
33	14.492	**13.340**	16.300	16.630	21.167
34	10.190	**10.034**	11.300	10.888	10.913
35	**7.154**	7.693	8.500	7.918	8.011
36	**8.512**	9.525	8.700	9.082	9.708
37	**14.548**	15.184	17.500	15.713	19.698
38	**3.886**	4.080	4.200	4.572	4.842
39	5.901	6.431	**5.600**	6.663	7.597
40	4.976	**4.673**	5.500	5.371	5.057
41	**7.558**	7.882	9.000	8.358	9.984
42	**5.030**	5.669	5.900	5.549	6.036
Mean	**7.412**	7.860	8.550	8.276	9.354
Median	**6.841**	7.259	8.300	7.918	8.011
Min	**3.886**	4.080	4.200	4.572	4.810
Max	**14.548**	15.184	17.500	16.630	21.167
[illegible]	0.763	0.713	0.901	0.840	1.131

Table 7. Forecast skill scores (using the Burton equation (T. O'Brien & McPherron, 2000) as the baseline) 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. Here, the GBM, LSTM1, and LSTM2 were trained with past SYM-H and IMF parameters as inputs. The highest skill score for each row is shown in **bold**.

St#	GBM	LSTM2	LSTM1
26	**0.143**	0.031	0.020
27	**0.028**	-0.120	-0.119
28	**0.249**	-0.028	0.052
29	**0.104**	-0.026	-0.106
30	-0.068	-0.127	-0.214
31	**0.063**	0.029	-0.211
32	**0.279**	0.235	0.125
33	0.129	**0.198**	0.020
34	0.064	**0.078**	-0.038
35	**0.096**	0.028	-0.074
36	**0.063**	-0.049	0.042
37	**0.074**	0.034	-0.114
38	**0.150**	0.108	0.081
39	0.114	0.035	**0.160**
40	0.074	**0.130**	-0.024
41	**0.096**	0.057	-0.077
42	**0.094**	-0.022	-0.063

4.1.2 2-[illegible]

[illegible] RMSE [illegible] 2-[illegible] GBM [illegible] LSTM2 [illegible] SYM-H [illegible] GBM [illegible] RMSE [illegible] 3.585 nT (24.8%) [illegible] RMSE [illegible] RMSE [illegible] GBM [illegible] 3.28 nT (3.1%) [illegible] LSTM2 [illegible] LSTM2 [illegible] RMSE [illegible] 8 of the 17 [illegible] (31-33, 36, 37, 39-41).

4.1.3 [illegible] SYM-H

[illegible] SYM-H [illegible] 1-[illegible] GBM [illegible] LSTM1 [illegible] CNN1. [illegible] 10 [illegible] RMSE [illegible] 1-[illegible] GBM, LSTM1, [illegible] CNN1 [illegible] 2-[illegible] GBM [illegible] GBM [illegible] 3.51 nT (15.4%) [illegible] RMSE [illegible] LSTM1 [illegible] 1.6 nT (7.7%) [illegible] RMSE [illegible] CNN1. [illegible] GBM [illegible] RMSE [illegible] 11 of 17 [illegible] CNN1 [illegible] RMSE [illegible] 3 [illegible] (33, 37, 40).

4.1.4 [illegible]

[illegible] 1 [illegible] 4.1. [illegible] RMSE [illegible] GBM [illegible] 5% [illegible] (p [illegible] ≤ 0.05) [illegible] 2-[illegible] LSTM2.

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), Burton equation (T. O'Brien & McPherron, 2000) and persistence. Here, the GBM and LSTM2 model were trained with past SYM-H and IMF parameters as inputs. The lowest RMSE for each row is shown in **bold**.

Storm #	GBM	LSTM2	Burton	Persistence
26	**8.285**	8.989	10.690	12.374
27	**11.585**	13.418	12.465	15.387
28	**5.650**	5.877	8.858	9.331
29	**8.826**	9.314	9.776	11.415
30	7.280	7.288	**6.266**	7.416
31	12.613	**12.436**	13.604	17.193
32	9.927	**8.937**	13.766	15.282
33	24.519	**18.481**	25.729	33.927
34	**13.736**	13.941	14.695	15.109
35	**9.504**	9.932	10.586	11.211
36	12.068	**12.058**	13.117	14.687
37	22.327	**21.084**	24.446	30.582
38	**5.153**	5.213	6.546	7.353
39	7.391	**6.798**	10.159	12.322
40	5.633	**5.281**	6.032	6.373
41	12.121	**11.707**	12.622	15.437
42	**7.976**	8.273	8.877	10.130
Mean	10.858	**10.530**	12.249	14.443
Median	9.504	**9.314**	10.690	12.374
Min	**5.153**	5.213	6.032	6.373
Max	24.519	**21.0840**	25.729	33.927
Std error	1.310	1.077	1.338	1.808

4.2 Explaining predictions

In [illegible] hs [illegible] 3.2 [illegible] m [illegible] 3, [illegible]

Fig. 3 [illegible] 1- [illegible] SYM-H [illegible] SYM-H [illegible] SYM-H [illegible] 1- [illegible] B_z, [illegible] V_x [illegible] E_s [illegible] (SYM-H [illegible] -100 nT. In [illegible] ρV_x^2. [illegible] SYM-H [illegible] ρV_x^2 [illegible] B_z [illegible] SYM-H, [illegible] SYM-H [illegible] B_z [illegible] B_x, B_y [illegible] E_s [illegible] B_z [illegible]

Table 9. Forecast skill scores (using the Burton equation (T. O'Brien & McPherron, 2000) as the baseline) for 2-hour ahead prediction over the test storm set with our GBM model and the LSTM2 neural network (Collado-Villaverde et al., 2021). Here, the GBM and LSTM2 model were trained with past SYM-H and IMF parameters as inputs. The highest skill score for each row is shown in **bold**.

Storm #	GBM	LSTM2
26	**0.225**	0.159
27	**0.071**	-0.076
28	**0.362**	0.337
29	**0.097**	0.047
30	**-0.162**	-0.163
31	0.073	**0.086**
32	0.279	**0.351**
33	0.047	**0.282**
34	**0.065**	0.051
35	**0.102**	0.062
36	0.080	**0.081**
37	0.087	**0.138**
38	**0.213**	0.204
39	0.272	**0.331**
40	0.066	**0.125**
41	0.040	**0.072**
42	**0.101**	0.068

Fig. 4 [illegible] SYM-H [illegible] B_z [illegible] SYM-H [illegible] SYM-H) [illegible] E_s [illegible] B_z. [illegible] B_z [illegible] V_x: B_z [illegible] SYM-H [illegible] V_x [illegible] SYM-H [illegible] -100 [illegible] E_s [illegible] SYM-H [illegible] (K[illegible] et al., 2017).

4.2.1 [illegible] 2004 [illegible]

[illegible] Fig. 5 [illegible] 1-h[illegible] 2-h[illegible] SYM-H [illegible] 2004 [illegible] SYM-H [illegible] -400 [illegible] RMSE [illegible] 30 [illegible] 1-h[illegible] 39 [illegible] 2-h[illegible] [illegible] 18:00 [illegible] 20:45 UT [illegible], [illegible] SYM-H [illegible] 1-h[illegible] 2-h[illegible]

Table 10. 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
Std. error	2.284	1.994	1.853	2.402

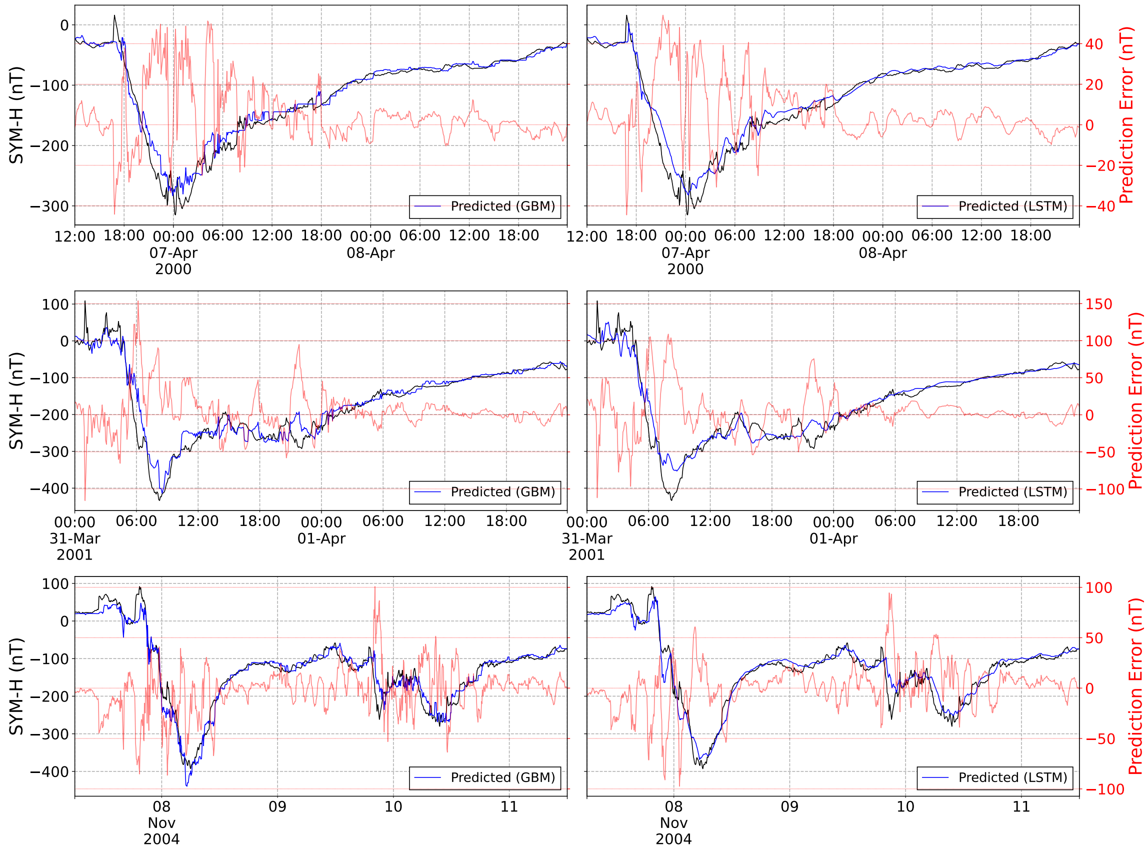

Figure 2. 1-hour ahead predictions for the 3 strongest geomagnetic storms in the test set during the main and recovery phases from our GBM (left column) and the LSTM2 developed by Collado-Villaverde et al. (2021) (right column). The observed SYM-H (black), the predicted SYM-H (blue) and the error (red)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

[illegible] IMF [illegible] B_z. [illegible]

[illegible] (22:00 Nov 7 to 06:00 Nov 8) [illegible] SYM-H [illegible] −400 nT [illegible] B_z [illegible] 22:00 on Nov 7, [illegible] 01:00 [illegible] 04:00 UT [illegible] B_z [illegible] -50 nT [illegible] IMF [illegible] B_z, [illegible] 04:00 [illegible] SYM-H [illegible] B_y, [illegible] 20:00 [illegible] 00 UT on Nov 7. [illegible] B_y [illegible] B_z [illegible] B_y [illegible]

Table 11. P-values from paired t-tests for null hypothesis that the mean difference in RMSE across storms for GBM vs. competing methods is zero.

	1 hr [illegible]	2 hr [illegible]
LSTM2 [illegible]	0.008	0.419
LSTM1 [illegible]	0.000	
[illegible]	0.000	0.000

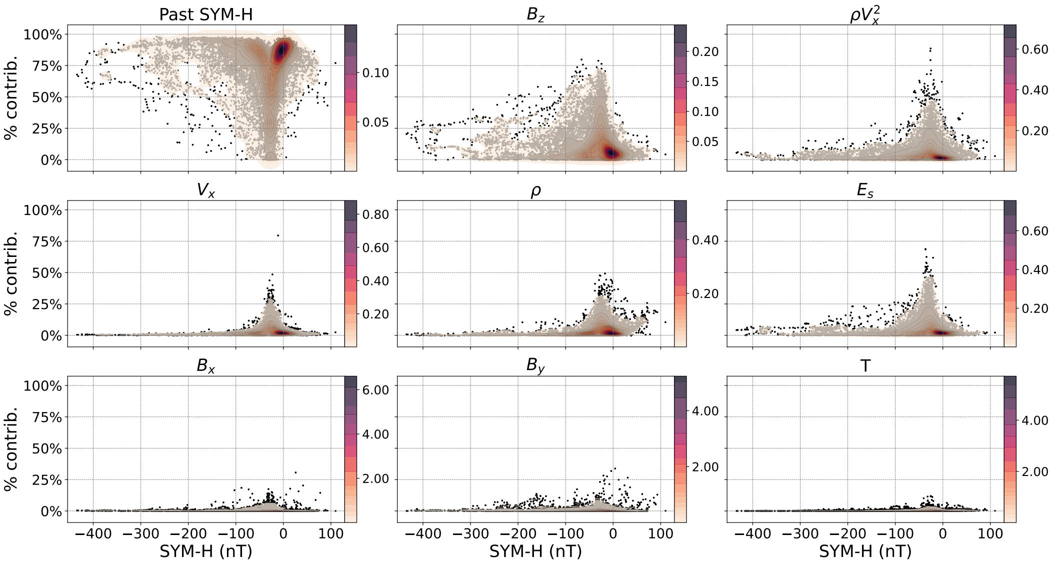

Figure 3. Scatter plot of percentage contributions (y-axis) against SYM-H (x-axis) for all the geomagnetic storms. The panels show the contributions of all considered features to the 1-hour ahead GBM prediction. Each prediction is represented as black dots. Kernel density estimates using a Gaussian kernel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

[illegible] SYM-H [illegible] B_z [illegible]

[illegible] SYM-H [illegible] 1- [illegible] 2- [illegible] SYM-H [illegible] L 1 [illegible] SYM-H [illegible] 1- [illegible] SYM-H [illegible] SYM-H [illegible]

[illegible] 18:00 [illegible] 21:00 [illegible] SYM-H [illegible] SYM-H [illegible] B_z [illegible] E_s [illegible] B_x ([illegible] 1- [illegible]

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. Each prediction is represented as black dots. Kernel density estimates using a Gaussian kernel are shown in color with the corresponding color legend on the right of each scatter plot.

[illegible]

4.2.2 January 2004 Storm

[illegible]

[illegible]

[illegible]

5 Discussion and conclusions

[illegible]

[illegible]

[illegible]

[illegible]

Appendix A

A1 Graphical comparison with persistence model & Burton equation

Fig A1 [illegible] GB M ([illegible] SYM-H and IMF [illegible] SYM-H [illegible] < 300 nT (31, 33, 37) [illegible] GB M [illegible] 37, [illegible] > 100 nT [illegible] SYM-H [illegible] GB M [illegible] -100 to 40 nT [illegible] SYM-H Fig A2 [illegible] GB M [illegible] B [illegible] I [illegible] GB M [illegible] B [illegible] RMSE [illegible] 6.

A2 Descriptive statistics of solar wind & IMF parameters

Table A1. Descriptive statistics for the solar wind and IMF parameters in the 25 storms used for training listed in table 1. The minimum temperature (MK) is most likely a measurement error.

Param	Min	25%	Med	75%	Max
B_x (nT)	-43.700	-3.131	0.340	3.378	34.681
B_y (nT)	-51.968	-2.901	0.221	3.289	46.862
B_z (nT)	-77.258	-2.296	-0.092	2.179	38.717
V_x (km/s)	-1233.693	-539.489	-445.287	-384.021	-264.722
Density (cm^{2})	0.041	2.912	5.027	8.477	76.239
Temp (MK)	0.0032	0.0385	0.0702	0.1262	1.0983

Table A2. Descriptive statistics for the solar wind and IMF parameters in the 25 test storms listed in table 2. The minimum temperature (MK) is most likely a measurement error.

Param	Min	25%	Med	75%	Max
B_x (nT)	-48.717	-2.868	0.221	3.444	33.827
B_y (nT)	-48.963	-2.816	-0.205	2.855	54.563
B_z (nT)	-48.585	-2.357	-0.084	1.933	53.002
V_x (km/s)	-887.784	-535.138	-424.304	-373.465	-251.481
Density (cm^{3})	0.295	2.760	4.424	7.643	113.982
Temp (MK)	0.0052	0.037	0.0658	0.122	0.9909

Acknowledgments

[illegible] B [illegible] (CL [illegible] NASA [illegible] 80NSSC20K0600, NASA [illegible] 80NSSC19K0564, NSF [illegible] 1663800 and NSF [illegible] 2027555. EC [illegible]

[illegible] NASA [illegible] 80NSSC20K1580 and 80NSSC20K1275. [illegible] ACE [illegible] [illegible] NASA [illegible] GSFC's [illegible] [illegible] Data Facility (SPDF) [illegible] CDAWeb [illegible] https://cdaweb.gsfc.nasa.gov/. [illegible] H [illegible] SPDF's OMNIWeb https://omniweb.gsfc.nasa.gov/. [illegible] (UM) L [illegible] Deep Blue Data [illegible] (https://deepblue.lib.umich.edu/data), [illegible] [illegible] (DOIs), [illegible]

References

[illegible] (2020). *swmfpy* [illegible] https://gitlab.umich.edu/swmf_software/swmfpy ([illegible] 2020 5)

[illegible] J. R., [illegible] W. H.-L., B[illegible] R. J., [illegible] V. S. N., & B[illegible] S. A. (2016). [illegible] *Kp* [illegible] NA[illegible] MODELING AND [illegible] *Kp* INDEX. *Space Weather*, *14*(10), 899–916. [illegible] 10.1002/2016SW001463

B[illegible] R. L., [illegible] M. A., [illegible] C. N., [illegible] M. C., [illegible] M. J., [illegible] U. V., ... H[illegible] J. (2021). [illegible] *arXiv:2006.12835 [astro-ph, physics:physics]*

B[illegible] E., & K[illegible] R. (1999). [illegible] *Machine Learning*, *36*(1/2), 105–139. [illegible] 10.1023/A:1007515423169

B[illegible] A., & [illegible] G. (2019). [illegible] 2015 [illegible] *Journal of Space Weather and Space Climate*, *9*. [illegible] 10.1051/swsc/2019007

B[illegible] K., B[illegible] M., J[illegible] A., K[illegible] M., K[illegible] S., & [illegible] Ö. (2020). [illegible] *SSRN Journal*. [illegible] 10.2139/ssrn.3520659

B[illegible] J. E. (2018). [illegible] *Journal of Geophysical Research: Space Physics*, *123*, 2029. [illegible] 10.1002/2017JA024650

B[illegible] J., C[illegible] X., M[illegible] Q., Li, W., [illegible] X., [illegible] R. M., ... [illegible] (2018). [illegible] In *Machine learning techniques for space weather* (pp. 279–300). E[illegible] [illegible] 10.1016/B978-0-12-811788-0.00011-1

B[illegible] R., B[illegible] M., B[illegible] S., [illegible] A., & A[illegible] O. (2011). [illegible] In *Annales geophysicae* ([illegible] 29, p. 965–971).

B[illegible] L. (1996). B[illegible] *Mach Learn*, *24*(2), 123–140. [illegible] 10.1007/BF00058655

B[illegible] L. (2001). [illegible] *Machine Learning*, *45*(1), 5–32. [illegible] 10.1023/A:1010933404324

B[illegible] L., F[illegible] J. H., [illegible] R. A., & [illegible] C. J. (1984). *Classification and regression trees*. [illegible]

B[illegible] R. K., M[illegible] R. L., & [illegible] C. T. (1975). [illegible] *Journal of Geophysical Research*, *80*, 4204. [illegible] 10.1029/JA080i031p04204

C[illegible] L., M[illegible] S. Y., & [illegible] Y. L. (2010). [illegible] H[illegible] NA[illegible] IMF [illegible] *Annales Geo-*

physicae, 28(2), 381-393. [illegible] 10.5194/[illegible]28-381-2010
C[illegible] E. (2019). [illegible] *Space Weather, 17*(8), 1166-1207. [illegible] 10.1029/2018SW002061
C[illegible] E., C[illegible] M. D., [illegible] H. J., B[illegible] C. C., H[illegible] Z., & [illegible] G. (2020). A G[illegible] B[illegible] [illegible] *J. Geophys. Res. Space Physics, 125*(11). [illegible] 10.1029/2019JA027684
C[illegible] M., C[illegible] E., & [illegible] S. (2017). [illegible] *Space Weather, 15*(8), 1004-1019. [illegible] 10.1002/2017SW001627
C[illegible] T., & G[illegible] C. (2016). [illegible] *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. [illegible] 10.1145/2939672.2939785
C[illegible] A., [illegible] P., & C[illegible] C. (2021). [illegible] *Space Weather*. [illegible] 10.1029/2021SW002748
F[illegible] J. H. (2001). [illegible] *Ann. Statist., 29*(5), 1189-1232. [illegible] 10.1214/[illegible]1013203451
G[illegible] N., J[illegible] A., & L[illegible] M. (2017). [illegible] *Space Science Reviews, 212*(3-4), 1315-1344. [illegible] 10.1007/s11214-017-0412-2
G[illegible] H., L[illegible] H., & [illegible] P. (1996). [illegible] *Ann. Geophys., 14*(7), 679-686. [illegible] 10.1007/s00585-996-0679-1
G[illegible] Y., [illegible] H.-L., B[illegible] R. J., [illegible] S. N., & B[illegible] M. A. (2019). [illegible] *Journal of Geophysical Research: Space Physics, 124*(1), 248-263. [illegible] 10.1029/2018JA025957
H[illegible] T., [illegible] R., & F[illegible] J. (2001). *The elements of statistical learning*. N[illegible] Y[illegible] NY [illegible]
I[illegible] T. (1990). [illegible] *Journal of geomagnetism and geoelectricity, 42*(11), 1249-1265.
K[illegible] E. K. J., B[illegible] A., [illegible] R., & Liu, Y. D. (2017). [illegible] *Space Sci. Rev., 212*, 1271-1314. [illegible] 10.1007/s11214-017-0411-3
K[illegible] J. H. (2005). [illegible] *Journal of Geophysical Research, 110*(A2), 2104. [illegible] 10.1029/2004JA010649
K[illegible] R., & G[illegible] B[illegible] R. (2015). [illegible] In G. L[illegible] & S. V. N. [illegible] (Eds.), *Proceedings of the eighteenth international conference on artificial intelligence and statistics* (Vol. 38, pp. 489-497). [illegible] PMLR.
K[illegible] M., C[illegible] U., [illegible] V., & G[illegible] O. (2020). [illegible] *IEEE Access, 8*, 187814-187823. [illegible] 10.1109/ACCESS.2020.3031477
Liu, L., [illegible] S., [illegible] Y., & [illegible] Z. (2020). [illegible] *Space Weather, 18*(11). [illegible] 10.1029/2020SW002501
Lou, Y., C[illegible] R., G[illegible] J., & H[illegible] G. (2013). [illegible] In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 623-631).

[illegible] ACM. [illegible] 10.1145/2487575.2487579

Lu, J., [illegible] Y., [illegible] M., [illegible] S., & [illegible] M. (2016). [illegible] *Planetary and Space Science*, 120, 48–55. [illegible] 10.1016/j.pss.2015.11.004

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2019). [illegible] *arXiv:1802.03888 [cs, stat]*.

Lundberg, S. M., & Lee, S.-I. (2017). [illegible] In *Advances in Neural Information Processing Systems* (Vol. 30). [illegible] Inc.

Lundberg, S. M., Nair, B., [illegible] M. S., [illegible] M., [illegible] M. J., [illegible] T., ... Lee, S.-I. (2018). [illegible] *Nat Biomed Eng*, 2(10), 749–760. [illegible] 10.1038/s41551-018-0304-0

Lundstedt, H., & [illegible] P. (1994). [illegible] *Ann. Geophys.*, 12(1), 19–24. [illegible] 10.1007/s00585-994-0019-2

Mayaud, P. N. (1980). [illegible] In *Derivation, meaning, and use of geomagnetic indices* (pp. 115–129). [illegible] (AGU). [illegible] 10.1002/9781118663837.ch8

McComas, D. J., Bame, S. J., Barker, P., Feldman, W. C., [illegible] J. L., [illegible] P., & Griffee, J. W. (1998). [illegible] (SWEPAM) [illegible] *Space Science Reviews*, 86, 563–612. [illegible] 10.1023/A:1005040232597

Mian, G., & Lam, H. (2021). [illegible] *arXiv:2108.04058 [cs, stat]*.

Moran, K. E., [illegible] B. P., & B[illegible], A. (2019). [illegible] In *Proceedings of the 29th annual international conference on computer science and software engineering* (pp. 166–172). [illegible] IBM Corp.

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

Molnar, C., [illegible] G., & B[illegible], B. (2020). [illegible] *arXiv:2010.09337 [cs, stat]*.

Murphy, A. H. (1988). [illegible] *Monthly Weather Review*, 116(12), 2417–2424. [illegible] https://journals.ametsoc.org/view/journals/mwre/116/12/1520-0493_1988_116_2417_ssbotm_2_0_co_2.xml [illegible] 10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

Natekin, A., & Knoll, A. (2013). [illegible] *Front. Neurorobot.*, 7. [illegible] 10.3389/fnbot.2013.00021

Newell, P. T., [illegible] T., Liou, K., Meng, C. I., & [illegible] F. J. (2007). [illegible] *Journal of Geophysical Research*, 112, A01206. [illegible] 10.1029/2006JA012015

O'Brien, T., & McPherron, R. L. (2000). [illegible] *Journal of Atmospheric and Solar-Terrestrial Physics*, 62(14), 1295–1299. [illegible] https://www.sciencedirect.com/science/article/pii/S1364682600000729 [illegible] 10.1016/S1364-6826(00)00072-9

O'Brien, T. P. (2002). [illegible] *Journal of Geophysical Research*, 107(A11), 1341. [illegible] 10.1029/2002JA009435

O'Brien, T. P., & McPherron, R. L. (2000). [illegible]

Journal of Geophysical Research: Space Physics, [illegible](A[illegible]), 7707–7719. [illegible] 10.1029/1998JA[illegible]437

[illegible] G., [illegible] E., [illegible] G., [illegible] M. F., & B[illegible] I. (2006). [illegible] IMF [illegible] *[illegible]n. Geophys.*, *24*(3), 989–999. [illegible] 10.5194/[illegible]-24-989-2006

[illegible] J., & [illegible] O. (2018). *Nevergrad - A gradient-free optimization platform*. https://GitHub.com/FacebookResearch/Nevergrad

[illegible] C. (2019). [illegible] *Nature Machine Intelligence*, *1*(5), 206–215. [illegible] 10.1038/s[illegible]256-019-0048-x

[illegible] A. (2014). *Interpreting random forests – Diving into data*. [illegible] https://blog.datadive.net/interpreting-random-forests/

[illegible] K., N[illegible] T., [illegible] G. D., & [illegible] H. E. (2015). [illegible] *Space Weather*, *1*[illegible](12), 853–867. [illegible]2021-08-05, [illegible] https://onlinelibrary.wiley.com/doi/10.1002/2015SW001254 10.1002/2015[illegible]254

[illegible] L. S. (1953). [illegible] In H. W. K[illegible] & A. W. [illegible] (Eds.), *Contributions to the Theory of Games (AM-28), Volume* [illegible] (pp. 307–318). [illegible] 10.1515/9781400881970-018

[illegible] R., & [illegible] A. (2021). [illegible] arXiv:2106.03253 [cs]

[illegible] F., [illegible] G., [illegible] R., [illegible] M., [illegible] F., & DeM[illegible] P. (2021). [illegible] *Space Weather*, *1*[illegible](2). [illegible] 10.1029/2020[illegible]2589

[illegible] C. W., L'H[illegible] J., N[illegible] N. F., [illegible] M. H., B[illegible] L. F., & [illegible] J. (1998). [illegible] ACE M[illegible] *Space Science Reviews*, *8*[illegible], 613–632. [illegible] 10.1023/A:1005092216668

[illegible] N., H[illegible] G., K[illegible] A., [illegible] I., & [illegible] R. (2014). [illegible] *Journal of Machine Learning Research*, *15*(56), 1929–1958.

[illegible] R., [illegible] J., K[illegible] S., [illegible] M., [illegible] H., F[illegible] J., ... F[illegible] O. (2020). *Meteorology-driven variability of air pollution (PM1) revealed with explainable machine learning*. [illegible] 10.5194/[illegible]-2020-469

[illegible] M. (1964). [illegible] In *Hourly values of equatorial dst for the [illegible]y* [illegible]

[illegible] M., & K[illegible] T. (1991). [illegible] 1957–1986, [illegible] 40 by *A. Berthelier and M. Menvielle (Int. Serv. Geomagn. Indices Publ. Off., Saint Maur, 1991)*

[illegible] N. (1989). [illegible] *Planetary and Space Science*, *3*[illegible](1), 5–20. [illegible] 10.1016/0032-0633(89)90066-4

[illegible] N. A. (1995). [illegible] *Journal of Geophysical Research*, *10*[illegible](A[illegible]), 5599. [illegible] 10.1029/94JA03193

[illegible] N. A. (2002a). [illegible] A NEW [illegible] MAGNETIC FIELD MODEL, 1. *Journal of Geophysical Research: Space Physics*, *10*[illegible](A8), SMP 12–1–SMP 12–15. [illegible] 10.1029/2001JA000219

[illegible] N. A. (2002b). [illegible] A NEW MAGNETOSPHERE MAGNETIC FIELD MODEL, 2 *Journal of Geophys-*

ical 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 rst 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.