Ann. Geophys., 44, 1259–1279, 2026
https://doi.org/10.5194/angeo-44-1259-2026
© Author(s) 2026. This work is distributed under
the Creative Commons Attribution 4.0 License.

Annales Geophysicae Open Access EGU

How well can we forecast local magnetic ground perturbations with existing space weather monitoring resources?

Stephen Omondi[1], Spencer Mark Hatch[2], Andreas Kvammen[3], Magnar Gullikstad Johnsen[3], Mathew J. Owens[4], Kristian Solheim Thinn[5], and Rodrigo López[6]

[1]Geophysical Institute, University of Alaska Fairbanks, Fairbanks, AK, USA
[2]Department of Physics and Technology, University of Bergen, Bergen, Norway
[3]Tromsø Geophysical Observatory (TGO), University of Tromsø, Tromsø, Norway
[4]Department of Meteorology, University of Reading, Reading, UK
[5]Electric Power Engineering, SINTEF Energi AS, Trondheim, Norway
[6]Statnett, Oslo, Norway

Correspondence: Spencer Mark Hatch (spencer.hatch@uib.no)

Received: 17 December 2025 – Discussion started: 12 January 2026
Revised: 10 September 2026 – Accepted: 26 October 2026 – Published: 9 November 2026

Abstract. In this study we examine how a deep-learning based forecast of local, ground-based geomagnetic field variations trained on solar wind parameters available in real time might be improved by including information contained in an accurate forecast of solar wind conditions. This is accomplished using a long short-term memory (LSTM) model together with magnetic field measurements made at the Rørvik magnetometer station in Mid-Norway. We use Advanced Composition Explorer (ACE) satellite measurements of solar wind and interplanetary magnetic field (IMF) conditions at the first Sun-Earth Lagrange point, and historical lists of coronal mass ejection (CME) impacts at Earth to train and validate the LSTM model. We find that accurate information about the IMF B_z component and solar wind speed are important for obtaining a reasonably accurate ($r^2 \geq 0.5$) forecast of local geomagnetic activity over forecasting horizons beyond ~ 3 h. Information about CME arrival time is only important when accompanied by accurate, relatively high-resolution information about IMF B_z. In the absence of the latter, CME arrival time information does not contribute to model performance. This empirical result amounts to a quantitative demonstration of the widely recognized impact of IMF orientation on CME geoeffectiveness. This result also highlights that new innovations, probably in the form of new prediction capabilities of conditions in interplanetary space, will be required to produce accurate forecasts of local geomagnetic disturbances beyond a forecast horizon of 1 h.

1 Introduction

Geomagnetic field disturbances at Earth's surface are driven by interactions between the solar wind and Earth's coupled magnetosphere-ionosphere system. The intensity of these disturbances is strongly tied to variations in the solar wind speed v and the interplanetary magnetic field (IMF) that is borne by the solar wind. The most intense disturbances are associated with transient solar phenomena such as coronal mass ejections (CMEs), with the degree of influence that a particular CME has on space weather conditions in near-Earth geospace – its so-called "geoeffectiveness" – critically dependent on IMF orientation upon arrival at Earth. Periods of strongly negative (southward) IMF B_z are invariably connected with the most intense geomagnetic storm activity (e.g., Gonzalez et al., 2011; Lakhina and Tsurutani, 2016; Sierra-Porta et al., 2024, and references therein).

At the surface of Earth and in its interior, geomagnetic disturbances give rise to geomagnetically induced electric fields (GEFs) that themselves give rise to geomagnetically induced currents (GICs). GICs can have deleterious effects on critical ground infrastructure including power grids, gas pipelines, and some navigation and communication systems (Ngwira and Pulkkinen, 2019; Press, 2021; CIGRE, 2019; EPRI, 2020; Patterson et al., 2023). Operational forecasts of geomagnetic disturbances, GEFs, and GICs are therefore naturally a central objective in space weather research

(see Hapgood, 2011; Fry, 2012; Pulkkinen et al., 2017; Abda et al., 2020; Press, 2021 and references therein).

Forecasting of GEFs and GICs, as well as proxies for them, has seen an explosion of interest during the past decade or so (Hapgood, 2011; Tóth et al., 2014; Pulkkinen, 2015; Keesee et al., 2020; Siddique and Mahmud, 2022; Conde et al., 2023; Wang et al., 2025), likely as a result of increasing awareness of the threat posed by GICs and space weather generally (Fry, 2012; Gannon et al., 2019; Abda et al., 2020) as well as increasing awareness and acceptance of machine learning-based solutions (Camporeale et al., 2018; Camporeale, 2019).

With few exceptions, forecasting methods rely on measurements of solar wind speed and density as well as the strength and orientation of the IMF made by solar wind monitors such as the Advanced Composition Explorer (ACE) and Deep Space Climate Observatory (DSCOVR) satellites located at L1, the first Sun-Earth Lagrange point. As L1 is $\sim 1.5 \times 10^6$ km from Earth, such measurements provide lead times of $\sim$ 30–60 min depending on v, which ranges from approximately 300–800 km s^{-1}. The accuracy of forecasts that rely on measurements at L1 therefore degrades rapidly with increasing forecasting horizon beyond 1 h.

One source of information that, at present, is little used in quantitative geomagnetic forecasting models described in the literature are outputs from solar wind models such as the Wang–Sheeley–Arge (WSA)–Enlil model (Arge and Pizzo, 2000; Arge et al., 2004; Odstrcil et al., 2004) and the heliospheric upwind extrapolation with time-dependence (HUXt) model (Barnard and Owens, 2022). The former provides a forecast of solar wind density and radial velocity. The latter produces a probabilistic forecast of solar wind speed v, and a probability distribution of the time of CME impact at Earth for any currently active CMEs.

In this study we seek to answer the following question: How accurately could one predict local geomagnetic activity if high-quality forecasts of solar wind properties such as speed, density, and IMF strength and orientation, as well as precise information about CME arrival time, were available 24 h in advance?

We answer this question via a deep learning-based forecasting model of the spectral power of dH/dt (the time rate of change of horizontal magnetic field disturbances, a well known proxy for ground induced currents) that is provided with four main sources of data: (a) ambient parameters such as time of day and the tilt of Earth's magnetic dipole; (b) real-time measurements of local magnetic disturbances measured by a ground-based magnetometer station in Rørvik, Norway; (c) real-time solar wind and IMF properties measured by the ACE satellite at L1; (d) "perfect" forecasts of solar wind and IMF properties as well as CME arrival times. The latter are "perfect" in the sense that they are based on actual solar wind and IMF observations and CME arrival times at Earth, ignoring any uncertainties in those observations.

This paper is organized as follows: In Sect. 2 we present the datasets used to train the model as well as the methodologies used for producing a training dataset. In Sect. 3 we present the model architecture. In Sect. 4 we present a comparison of model predictions with observations for three storms in 2021, 2023, and 2024. In Sect. 5 we discuss our results and present an evaluation of the model performance against simpler statistical models. In Sect. 6 we conclude our study.

2 Data

There are three primary sets of measurements used in this study: magnetometer measurements made at the Rørvik magnetometer station (64.95° N, 10.99° E geographic coordinates; 62.46°, 91.21° corrected geomagnetic coordinates) maintained by the Tromsø Geophysical Observatory (Johnsen, 2025); solar wind and IMF measurements made by the ACE satellite; and the Richardson and Cane (2024) CME list, which is extended from the original work of Cane and Richardson (2003). All of these datasets extend over the 15 year time period (2010–2024) that is considered in this study.

Our choice to focus on Rørvik magnetometer measurements is based on information from Norway's transmission grid operator, Statnett, that ground-based conducting infrastructure close to the area was responsive to elevated levels of geomagnetic activity in 2021 (Statnett, private communication, 2025). We therefore specifically examine dH/dt, the time rate of change of the horizontal magnetic field and a commonly used proxy for GICs (Viljanen et al., 2001), as estimated from Rørvik magnetometer measurements.

2.1 GIC proxy from Rørvik magnetometer data

The starting point for producing model training data is the set of Rørvik magnetometer measurements with a temporal resolution of 10 s. An example time series of the horizontal component H and the time rate of change dH/dt estimated using first-order backward differencing ($f'(x) \approx \frac{f(x)-f(x-\Delta x)}{\Delta x}$) during 3–4 November 2021, are shown in Fig. 1a and b, respectively. An interplanetary CME impacted Earth at 19:42 UT on 3 November 2021, as indicated by the red dash-dotted vertical line in all panels.

Figure 1c shows the power spectral density of dH/dt up to the Nyquist frequency 0.05 Hz calculated using a Python implementation (Prieto, 2022) of the multitaper technique (Thomson, 1982; Hatch and LaBelle, 2018). We use the multitaper technique because it yields a less biased estimate of the power spectral density than is obtained, for example, via Fourier transformation of the underlying time series. From Fig. 1c it is clear that the power is concentrated at the lowest frequencies both before and after CME impact.

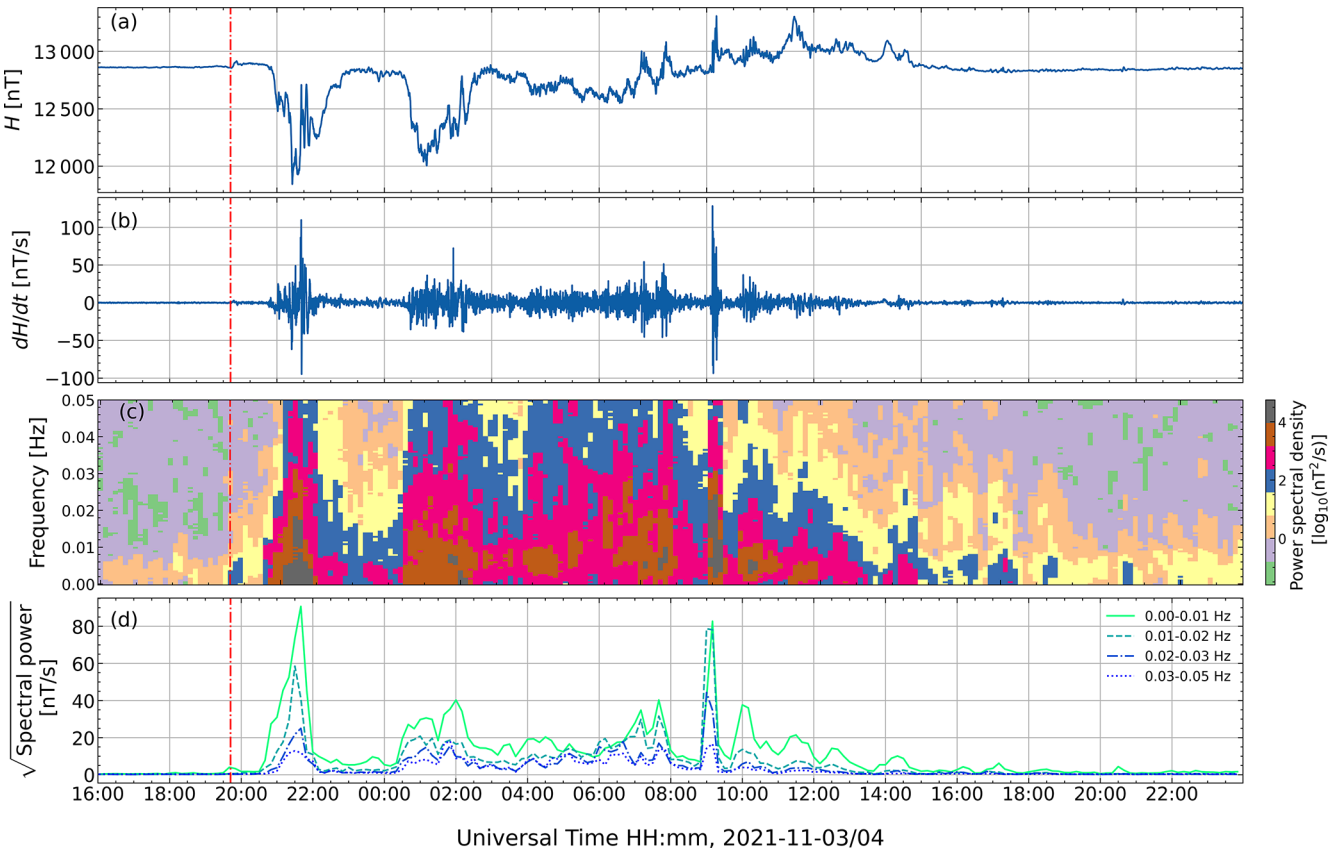


Figure 1. Illustration of procedure for preprocessing Rørvik magnetometer data for measurements made on 3–4 November 2021. **(a)** Horizontal magnetic field component H. **(b)** Time rate of change of H ($\mathrm{d}H/\mathrm{d}t$) estimated using backward differencing. **(c)** Spectrogram of logarithmic $\mathrm{d}H/\mathrm{d}t$ power spectral density. **(d)** Square root of integrated $\mathrm{d}H/\mathrm{d}t$ power spectral density over different frequency bands. The focus of this study is the power in the lowest frequency band (0–0.01 Hz, green solid line). The red dash-dotted vertical line at 19:42 UT on 3 November 2021 indicates the time of impact of an interplanetary CME as given in the Richardson and Cane (2024) CME list.

Figure 1d shows the integrated power spectral density, or spectral power, in various frequency bands. The majority of the power is generally concentrated at the lowest frequencies (here represented by 0–0.01 Hz, the green solid line), denoted by P_0 in Table 1, although during periods of intense geomagnetic activity the contributions from the other frequency bands can reach or exceed the spectral power in the lowest frequency band. This occurs, for example, near 09:00 UT on 4 November 2021.

In this study we choose to focus on the spectral power contained in the lowest frequency band based on the findings of Oyedokun et al. (2020) that the most dominant frequencies in actual GIC measurements tend to be below 50 mHz.

We view the spectral power of $\mathrm{d}H/\mathrm{d}t$ as a natural choice for modeling power dissipation due to GICs in power lines and transformer windings, since the spectral power of $\mathrm{d}H/\mathrm{d}t$ as a function of frequency f, $|\mathrm{d}H/\mathrm{d}t(f)|^2$ is proportional to the spectral power of the GIC current $|I(f)|^2$ through a transformer, which is in turn directly proportional to the power dissipation $P = I^2R$. Here R represents the winding resistance of copper windings in a transformer. (The other form of impedance in a transformer, steel core resistance, is the main heat source. It unfortunately cannot be approximated via $P = I^2R$ as it is highly frequency dependent.)

2.2 Exogenous data inputs

Table 1 lists all inputs used for training the deep learning models we present in the next section. A selection of these inputs are shown in Fig. 2 for a four-day period during 2–5 November 2021. These are, in display order from top to bottom, dipole tilt angle ψ, IMF B_z and B_y, solar wind density n, solar wind speed v, and the "predicted" CME arrival time distribution. Quantities in panels b–e are generated from ACE measurements, while the arrival time distribution in panel f is produced by generating a Gaussian distribution with a standard deviation between 5 and 6 h around the time of CME impact as recorded in the Richardson and Cane (2024) CME list. The exact value of the standard deviation for a particular arrival time distribution is randomly sampled from a uniform distribution $U(5\,\mathrm{h}, 6\,\mathrm{h})$, where this distribution was selected based on typical arrival time dis-

Table 1. List of input variables used in training of deep learning models.

Variable	Symbol
log dH/dt spectral power over 0–0.01 Hz*	$\log_{10}(P_0)$
Cosine UT	cos(UT)
Sine UT	sin(UT)
Dipole tilt angle	ψ
IMF B_z	B_z
IMF B_y	B_y
Solar wind density	n
Solar wind velocity	v
CME arrival time distribution	pred_t
Predicted IMF B_z†	B_z^{P}
Predicted IMF B_y†	B_y^{P}
Predicted solar wind density†	n^{P}
Predicted solar wind velocity†	v^{P}
Standard normal random number	$\mathcal{N}(0, 1)$

* Only the time history of $\log_{10}(P_0)$ is used for training, as illustrated in Fig. 4.
†For Model A, predicted quantities are generated by shifting the original time series 24 h backward. For Model B, predicted quantities are first smoothed using a second-order Savitzky–Golay filter with a 36 h window and then shifted 24 h backward.

tribution widths observed in the HUXt model (Barnard and Owens, 2022). The total area under the curve is 1, and the arrival time distribution is shifted 24 h backward in time so that the model has information about incoming CMEs 24 h ahead of time.

In addition to observed solar wind and IMF properties, panels b–e of Fig. 2 also show two sets of artificially generated space weather forecast quantities labeled as "Set A" and "Set B", respectively denoted by blue and orange lines. Both "Set A" and "Set B" quantities consist of ACE measurements shifted 24 h backwards in time, but "Set B" quantities are additionally smoothed using a second-order Savitzky–Golay filter with a 36 h window. Each set of artificial space weather forecast products is used to train a separate model, as described in Sect. 3. Quantities in panels d–f are among the outputs of the WSA–Enlil and HUXt space weather models. (Note that although the heliospheric magnetic field polarity is routinely forecasted, no currently existing space weather forecasting model is capable of predicting IMF B_z and B_y.)

3 Modeling

In this section we briefly describe the design of the long short-term memory (LSTM) networks we employ, deep learning, and a persistence-based benchmark model for evaluating the performance of the LSTM networks. A summary of the four models used in this study are shown in Table 2.

Table 2. Summary of models used in this study.

Model	Description	Section
A	Based on conditions at L1 up to 24 h in advance	3.2
B	Based on smoothed conditions at L1 up to 24 h in advance	3.2
C	Same as Model A but without CME arrival time information	3.2
Benchmark	Weighted combination of persistence and climatology models	3.3

3.1 Long Short Term Memory

LSTM networks are a variant of recurrent neural networks that handles short-term memory problems by introducing a *cell state* (Hochreiter, 1997). The cell state keeps past information for a long time, solving the vanishing gradient problem and allowing the network to "remember" past information in the sequence (Wei et al., 2018).

As illustrated in Fig. 3a, the recurrent structure of the LSTM unit is made up of three gates: forget, input, and output (Siciliano et al., 2021). The forget gate keeps the relevant information in the network by filtering both the current and past values in the sequence. The input gate enables the network to determine which information to retain from the current values of the sequence. Thus, the input gate is where the current information is added to the network's long-term memory. The previous cell state, C_{k-1}, is then updated to the final cell state, C_k, which is finally uploaded to the output gate. The output gate is the final destination of the information after transitioning from the forget and input gates. The output gate then determines what information is allowed to the next element. In each LSTM unit, there is an element of a sequence being processed, X_k, and the hidden states are also calculated as h_k. In this case, both the hidden state h_k and the cell state C_k are passed on to the preceding element of the sequence, to ensure the past information is used as input in the subsequent element in the sequence being processed. This unique architecture makes the LSTM models better at handling time series problems. In a skillful model, several LSTM cell units per layer are looped together, forming a block of layers such as the one illustrated in Fig. 3b.

3.2 Deep Learning

In big data analytics, deep learning models attempt to learn a functional consistency between input features and future values of the label feature y (in this study $y = \log_{10}(P_0)$). The resulting model provides forecasts for the target feature at future time steps. Given a time series vector $\mathbf{x}(t)$ of length n (corresponding to n input features) measured at time t, developing a model to forecast a target variable $y(t+h)$ at a future time $t+h$ requires historical data $x(t-1)$ at time $t-1$.

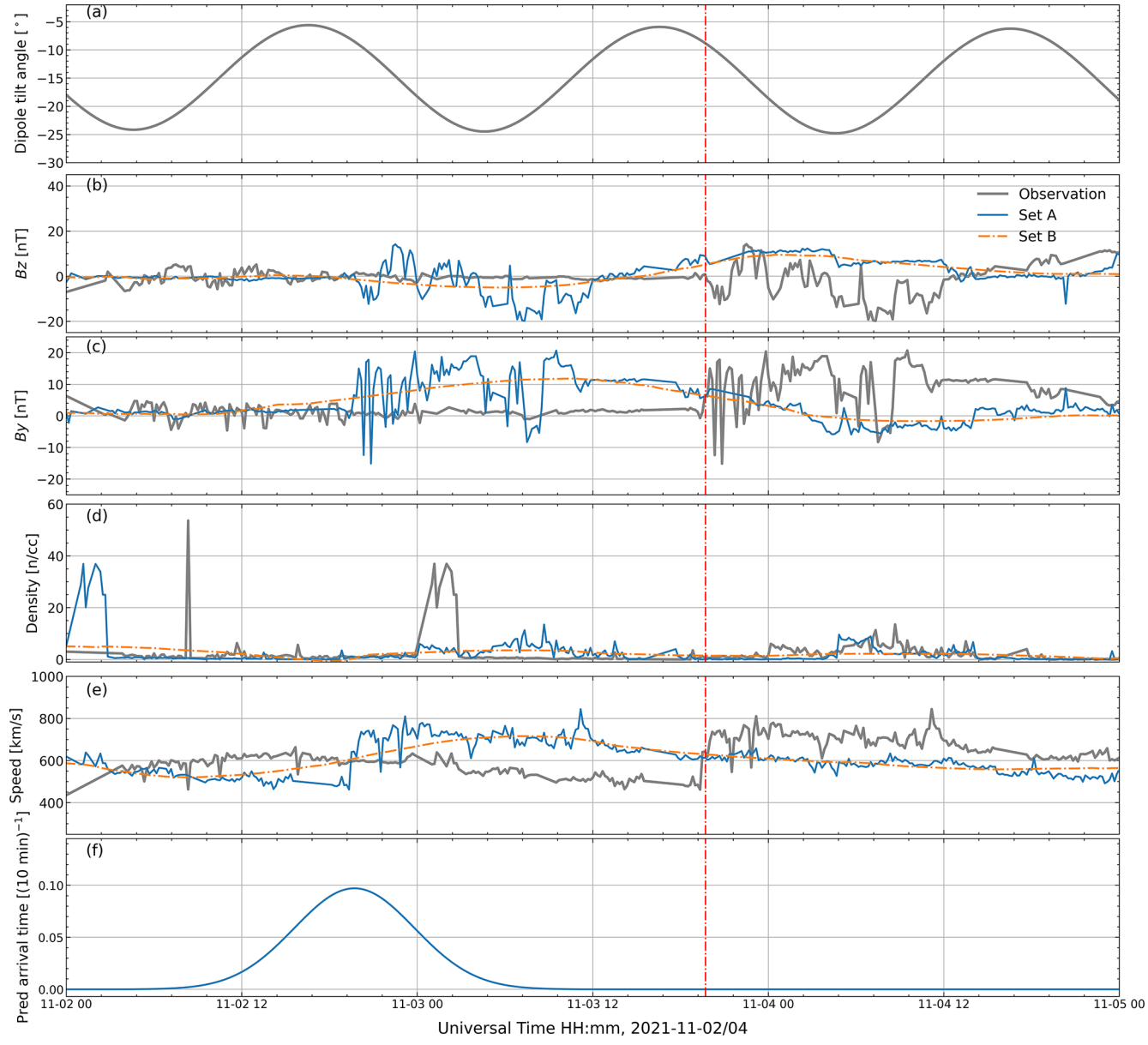


Figure 2. Illustration of exogenous model inputs during 2–5 November 2021. The quantities labeled "Set A" and "Set B" in panels **b–e** (indicated by blue solid lines and orange dash-dotted lines, respectively) are the artificial solar wind forecasts generated as described in Sect. 2.2. Both "Set A" and "Set B" are ACE measurements at L1 shifted 24 h backwards in time, but "Set B" is additionally smoothed using a second-order Savitzky–Golay filter with a 36 h window. The red dash-dotted vertical line at 19:42 UT on 3 November 2021 indicates the time of impact of an interplanetary CME according to the Richardson and Cane (2024) CME list. **(a)** Dipole tilt angle Ψ. **(b)** Observed and forecasted IMF B_z. **(c)** Observed and forecasted IMF B_y. **(d)** Observed and forecasted solar wind density n. **(e)** Observed and forecasted solar wind speed v. **(f)** CME arrival time distribution "forecasted" 24 h in advance. Note that the peak of the CME arrival time distribution is deliberately shifted 24 h prior to the actual arrival of the CME indicated by the red dash-dotted line.

The model thus requires uniform input sequences of a fixed length using a sliding lookback window $\Delta t_b = N_b \Delta t$, where N_b is the number of past samples that comprise a sequence and Δt is the cadence at which input features are measured or sampled. In our study these are $\Delta t = 10\,\text{min}$ and $N_b = 204$ so that $\Delta t_b = 34\,\text{h}$.

Mathematically, we may express the functional relationship between the input features and the target feature that the model has learned as

$$\hat{y}_t^h = g_h(\mathbf{x}_{t-\Delta t_b}, \ldots, \mathbf{x}_{t-\Delta t}, y_{t-\Delta t_b}, \ldots, y_{t-\Delta t}) \,, \quad (1)$$

where $\hat{y}_t^h$ is the expected forecast for the time t and h indicates the forecast horizon, while $y_{t-\Delta t_b}$ and $x_{t-\Delta t_b}$ are the

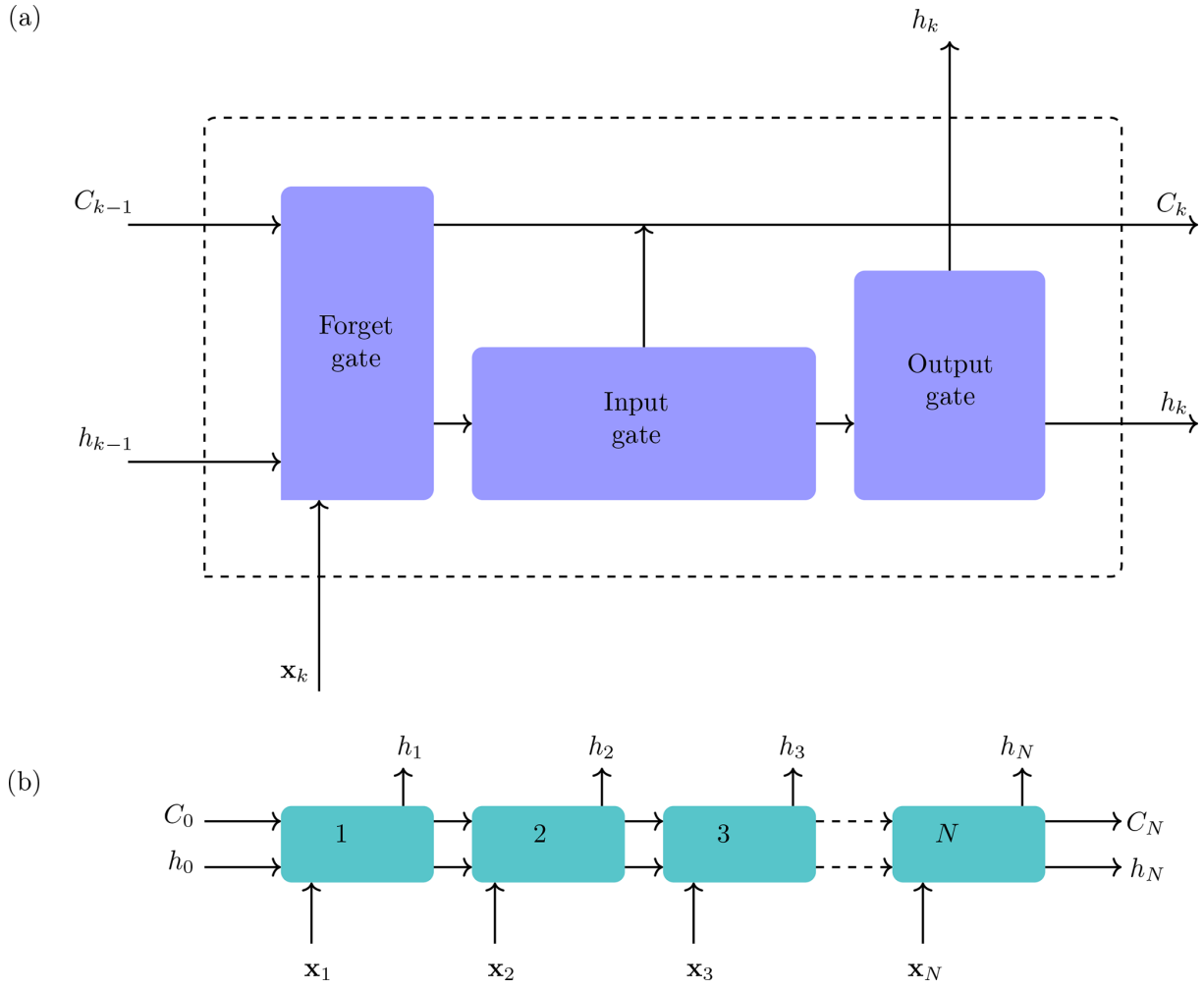


Figure 3. Simplified schematic of an LSTM network at two levels of detail: **(a)** An LSTM cell; **(b)** block stacking of LSTM cell units. The symbols X_t, h_t, and C_t are respectively input sequence data, the hidden state, and the cell state at time t.

input target and covariate input variables observed from the time $t - \Delta t_b$ to $t - \Delta t$. The forecast horizon h is the length of time to be forecasted into the future . For the models we present in the following section, the forecast horizon h takes on the values 10 min, 1 h, 3 h, 6 h, and 12 h.

The deep learning model we use consists of an input layer, LSTM units in the hidden layers, and a fully connected or dense layer in the output layer. The hidden layers comprised three layers, each with sequential LSTM units of 327, 56, and 249, and a tanh activation function. The model was optimised using the Optuna framework to obtain the best-performing model hyperparameters (Akiba et al., 2019). The tuned parameters had a batch size of 70, a look-back window $\Delta t_b = 34$ h (or $N_b = 204$ as previously mentioned), an optimizer with a learning rate of 5.4×10^{-5}, early stopping at a patience value of 8, and a dropout probability of zero. (The patience value is a hyperparameter used with the early stopping technique; it is the number of epochs the training process must wait for an improvement in the model's performance on the validation set before stopping the training. The dropout probability is the probability of ignoring a random neuron and its connections; it is part of a commonly used strategy for avoiding overfitting that is known as the dropout method (Srivastava et al., 2014).)

This multivariate modeling framework, with a single output for each of the five forecast horizons, results in five distinct models. The choice of this architecture over the multivariate input multi-output allows for the significant contribution of every input variable to the model output in each forecast horizon.

Two LSTM networks were trained with the 14 input variables listed in Table 1 together with a gradient descent approach with the ADAM optimizer (Kingma and Ba, 2015). The first network, hereafter "Model A", was trained using "Set A" quantities (see Table 1) and thus effectively had access to a perfect forecast of solar wind conditions at L1, as measured by ACE, up to 24 h in advance. The second network, "Model B", was trained using lower-resolution "Set B"

quantities. We therefore expected the performance of Model A to be much higher than that of Model B. Last, to quantify the value of CME arrival time information, we trained a third neural network ("Model C") using all Set A quantities except for the predicted arrival time "pred_t". We hypothesized that Model A would perform better than Model C.

Each network was trained using an early stopping mechanism and a maximum of 100 epochs.

We used the Mean Squared Error

$$\mathrm{MSE} = \frac{1}{m}\sum_{t}\left(y_t - \hat{y}_t^h\right)^2 \tag{2}$$

as the loss function in training each LSTM network, where y and $\hat{y}^h$ are as defined in Eq. (1) and m is the number of observations. This loss term is used because it increases the sensitivity of the resulting model to outliers relative to the root mean square error (RMSE) (Girosi et al., 1995) – or in the case of this study, to the relatively infrequent but extreme values of P_0 that occur during periods of elevated geomagnetic activity. The model that generates the minimum MSE on the validation set is saved, and its performance is evaluated using a test data set with the root mean square error (the square root of Eq. 2).

To calculate the prediction accuracy with the LSTM for forecasting the power of the rate of change of magnetic field, we used the coefficient of determination

$$R^2 = 1 - \frac{\sum_t \left(y_t - \hat{y}_t^h\right)^2}{\sum_t (y_t - \overline{y})^2}\ , \tag{3}$$

where $\overline{y}$ is the mean of the observed values. This metric indicates how much of the variance of the actual measurement is explained by the model (e.g., Liemohn et al., 2018).

Regarding the choice to train the models using the logarithm of the spectral power with an MSE loss function, we also trained sets of models directly using the spectral power (not the logarithm) as well as the RMSE. We found that the specific combination of the logarithm of the spectral power and an MSE loss function yielded the most predictive power. Chu et al. (2025) also found that training using the logarithm of their heavily tailed training data set yielded the most predictive power.

3.3 Benchmark model

We use a simple combination of a persistence model and a climatology model as a performance benchmark for the deep learning model.

The persistence model we employ is a naive forecasting model that assumes the target time series repeats itself (Bailey et al., 2022; Hu et al., 2024), such that

$$f(t+h) = f(t) \tag{4}$$

for forecast horizon h, with $f(t+h)$ and $f(t)$ respectively the forecast value and the last observed value. Persistence models work well in near-term forecasting and degrade as the forecast horizon increases.

For the climatology model, we first group the entire $\sim$ 15-year time series of dH/dt spectral power over 0–0.01 Hz (denoted by P_0) by month, and then for each month subgroup calculate the average value of P_0 within a one-hour window for each hour in UT. This procedure yields, for example, the average P_0 during 00:00–01:00 UT for the month of June. With 12 months per year and 24 h d^{-1}, we end up with $12 \times 24 = 288$ average values of P_0. To obtain a climatology model prediction for a specific timestamp, we use the month of the timestamp in question to first select a group of 24 coefficients. We then convert the timestamp of the observation to fractional hours and perform a linear interpolation between the two nearest hourly averages.

Finally, we weight the persistence and climatology model outputs equally:

$$P_0^{\mathrm{hybrid}}(t+h) = \frac{1}{2}\left(P_0^{\mathrm{clim}}(t+h) + P_0^{\mathrm{pers}}(t)\right)\ . \tag{5}$$

As we show in the following sections, the resulting benchmark (combined persistence and climatology) model performs well for near-term forecasting horizons and degrades with increasing forecast horizon.

3.4 Data preprocessing

We normalized the data using robust scaling and windowing techniques to create suitable sequences for time series forecasting. First, we scaled the target P_0 by calculating the logarithm of the square root of P_0. The logarithm reduces data skewness and effectively leads to faster convergence and more predictive power for events of interest in space weather forecasting, such as CMEs, which occur sporadically but are generally accompanied by extreme target values (see, e.g., Chu et al., 2025). We then scaled all input and output data using the robust scaler method. This method scales data by subtracting the median from each point and scaling it according to the quantile range. Data scaled in this way are robust to outliers (Pedregosa et al., 2011).

For time series forecasting, data are conventionally split into training, validation, and testing sets, based on statistical ratios; for example 70 : 20 : 10, respectively. This is desirable for static learning. Here, however, we are interested in dynamic learning, whereby only the best-performing data in the entire training dataset are used to train the model. In specific, we employ the time series cross-validation method using the TimeSeriesSplit software module (Pedregosa et al., 2011). In this scenario, the training data extends from January 2010 up to and including September 2021, and the test set extends from October 2021 up to and including September 2024. Data are split chronologically rather than randomly to avoid data leakage (Lones, 2024; Apicella et al., 2025). The training set was subjected to time series cross-validation, which splits the training set further into 80 : 20 proportions

per fold number for training and validation sets, respectively. In this work, we optimized fold number with Optuna, and out of 4 (0, 1, 2, 3), the optimal fold number used for training was 2. The training and validation sets were used to train and find the best model, and the test set was used to evaluate the model's performance.

An important difference between the training and test data set is that the training data set covers the beginning of the ascending phase of Solar Cycle 24 through the beginning of the ascending phase for Solar Cycle 25, while the test data set only represents the ascending phase of Solar Cycle 25. The occurrence rate of CMEs is higher in the test data set than in most of the preceding solar cycle, and the storms in the test data set are also generally more extreme. This difference is manifest in statistics of P_0, where for example the mean value of P_0 for the training and test data set is respectively 0.082 $(\mathrm{nT\,s^{-1}})^2$ and 0.14 $(\mathrm{nT\,s^{-2}})$. One seemingly intuitive way to make the training and test data sets more similar statistically would be to split the data randomly. Random splitting of time series data, however, introduces data leakage issues (Lones, 2024; Apicella et al., 2025). We therefore chose to split the data chronologically. (We also tested an alternative splitting procedure which gives similar results, as described in Appendix A.)

Finally, all the transformed data were sequenced using the sliding window method. With this method, a sliding window is created by taking a time series data sample of length T and creating individual subsamples of length n. The last m samples are to be forecasted, and $m < n < T$. The next subsample is generated by sliding the window forward to the next subsample of length n, such that the total number of subsamples is $T - n + 1$. This process is illustrated in Fig. 4.

3.5 Model Hyperparameter Tuning

In deep learning problems, hyperparameter search is a prerequisite step for successful optimal modeling. The growing popularity of deep learning and its complexity require an efficient automatic hyperparameter tuning method. There are many well developed hyperparameter optimization software packages such as Spearmint (Snoek et al., 2012), SMAC (Hutter et al., 2011), Hyperopt (Bergstra et al., 2015), Google Vizier (Golovin et al., 2017), Autotune (Koch et al., 2018), and Optuna (Akiba et al., 2019) that address this need. We chose to use Optuna, as it dynamically constructs the search space, is easy to set up, provides efficient sampling and pruning algorithms, and recently has been reported to be robust in optimizing model hyperparameters (Conde et al., 2023). For this work, we used Optuna to examine sequence window size, number of layers, LSTM hidden units, batch size, patience, learning rate, fold number, and dropout probability. The ADAM optimizer used in the model was preselected based on its successful performance in other modeling works (Kingma and Ba, 2015). To achieve the best optimizations with Optuna, various approaches are employed for each tuning. Traditional grid and random-search strategies have proven to be inefficient when the hyperparameter search space is large. The Hyperband approach, which is a bandit-based algorithm (Li et al., 2018), combined with the tree-structured Parzen estimator (TPE) sampler, is used to approximate the objective function (Bergstra et al., 2011). Given n trials in the optimizations using the Optuna tuner, many training cases may yield the worst outcome, which eventually will be ignored in selecting the optimal model. To save on computational cost, we pruned the training cases that would not yield good models using TFpruningCallback and HyperbandPruner with the minimum resource of three epochs and reduction factor of 3 to the maximum resource of a fixed training epoch number. The final useful trained models listed are again stored and ranked to produce the best optimal model using their loss ranking basis.

3.6 Feature importance

Neural networks are often regarded as black-box algorithms, though some useful external inference methods are used to explain deep learning models. Computing the feature importance of deep learning models, such as LSTM, provides a vital aspect of model understanding and interpretation. The degree of feature or input variable contribution to the output of a model is measured by feature importance. Feature importance is thus calculated based on Gini importance (Nembrini et al., 2018), such that more important features have higher values and less important features have lower values. Understanding the feature importance of a model can help improve model performance by training the model with the most relevant features. It also provides useful insights into the underlying dynamics, relationships, and patterns available in the examined time-series data.

In this study, we calculate the feature importance via the gradient tape technique as implemented in TensorFlow (Abadi et al., 2015). Gradient tape is a gradient-based, post-hoc method that works on an already trained model, and that keeps track of automatic differentiation operations of tensor variables. The capability of this method to monitor variables during computation makes it a useful technical tool in evaluating the contributions of trainable input variables in a model. We use the gradient tape technique instead of other widely used feature importance techniques, such as Shapley Additive exPlanations (Lundberg and Lee, 2017; Iong et al., 2022) and input permutation (Breiman, 2001; Fisher et al., 2019), that are not necessarily optimized for evaluating deep learning-based models. We present feature importance analysis of Models A–C in Sect. 5.

4 Results

In this section, we present case study results for three space weather events. The first occurred at 19:42 UT on 3 Novem-

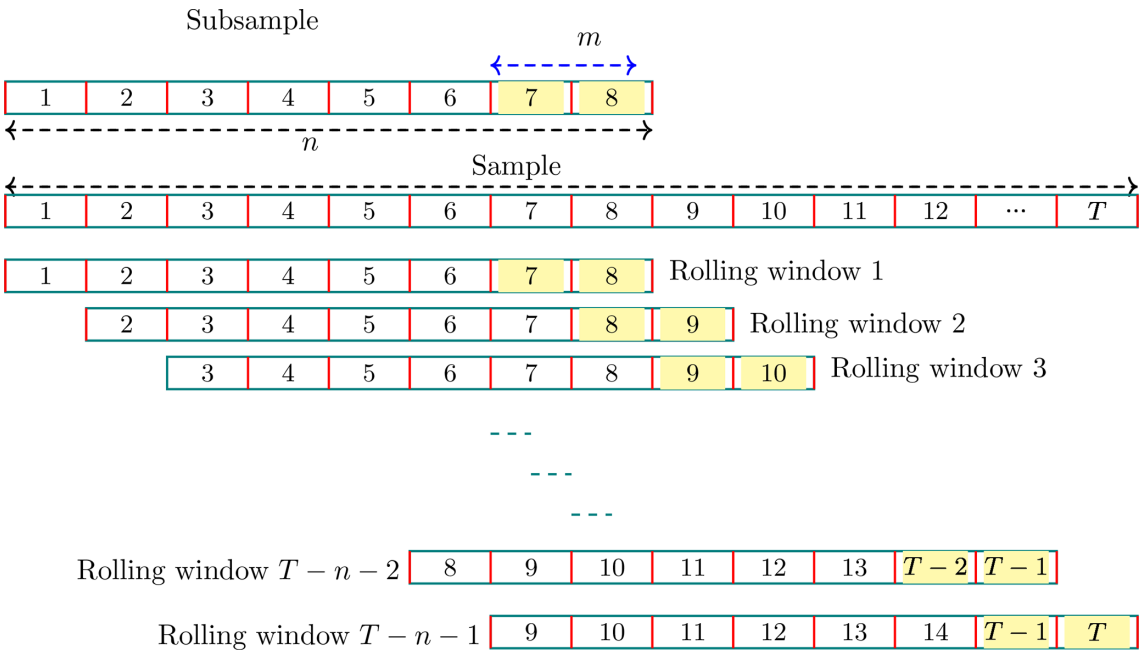


Figure 4. Data windowing using rolling strategies.

ber 2021, and involved a CME impact at Earth. The second involved consecutive CME impacts at 17:05 UT on 10 May and 09:17 UT on 12 May in 2024. We also present results from a high-speed stream/stream interaction region (HSS/SIR) storm that occurred during 5–7 May 2023. The CME impact events in 2021 and 2024 are identified in the Richardson and Cane (2024) list. The HSS/SIR storm is identified in the Space Weather Database Of Notifications, Knowledge, Information (DONKI) (Community Coordinated Modeling Center, 2026).

Figure 5 shows the minor geomagnetic storm of 3–4 November 2021. It includes values of P_0 (gray line) calculated from Rørvik magnetometer measurements as well as the forecasted power at forecasting horizons of 10 min, 1 h, 3 h, 6 h, and 12 h (panels a–e, respectively). Results from Models A–C (blue, orange, and green lines, respectively) and the benchmark model (dashed black line) for each forecasting horizon are also shown. In 10 min and 1 h forecasts, Models A–C and the benchmark model yielded similar forecast output as demonstrated in Fig. 5a and b.

For Models A–C, the trend in the forecasted time series is mostly consistent for forecasting horizons extending from $h = 10$ min up to $h = 3$ h. The forecast output amplitudes decrease with increasing forecast horizon. The deep learning models perform better than the benchmark model, which exhibits worsening performance with increasing forecast horizon. In panels d and e (6 and 12 h forecast, respectively) all of the models demonstrate relatively low predictive power, yielding output with diminished amplitudes. During this period v reached a maximum value of 850 km s^{-1}, and IMF B_z reached a minimum value of -20 nT, as shown in Fig. 2.

Figure 6 illustrates the forecast output of the 9–12 May 2024 geomagnetic storm. Models A–C yielded a generally good correlation with calculated P_0 values (thick gray line) up to 1 h in advance, as shown in panels a and b. However, the enhanced P_0 values that accompanied the last phase of this storm were only captured by the 10 min forecast of each model, while the enhancements associated with the last storm phase are mostly absent in the models with a 1 h or greater forecasting horizon. We return to this point in the discussion.

Panels c and d show results for forecasting horizons of 3–6 h; the storm initial and main phases were well captured in the forecast, while the last phase was completely missed. The model forecast results for $h = 12$ h, shown in panel e, shows some resemblance to the observed time series, but does not evince any of the rapid, larger amplitude variations that appear in the observed time series. In particular the forecasted values of P_0 are everywhere less than 30 nT s^{-1} for all models, whereas the observed time series frequently exceeds 60 nT s^{-1}.

Figure 7 illustrates observations and model results during 5–7 May 2023, for an HSS/SIR-driven storm. During this period v reached a maximum value of 580 km s^{-1}, and IMF B_z reached a minimum value of -18 nT (not shown). For forecasting horizons beyond 10 min, none of the deep learning models manage to predict the enhanced levels of P_0 that were observed at Rørvik.

Figure 8 summarizes performance metrics of each model for each of the five forecasting horizons, calculated over the entire test data set. Results for Models A and B are respectively indicated by solid blue and orange dash-dotted lines, while metrics for the benchmark model are indicated by

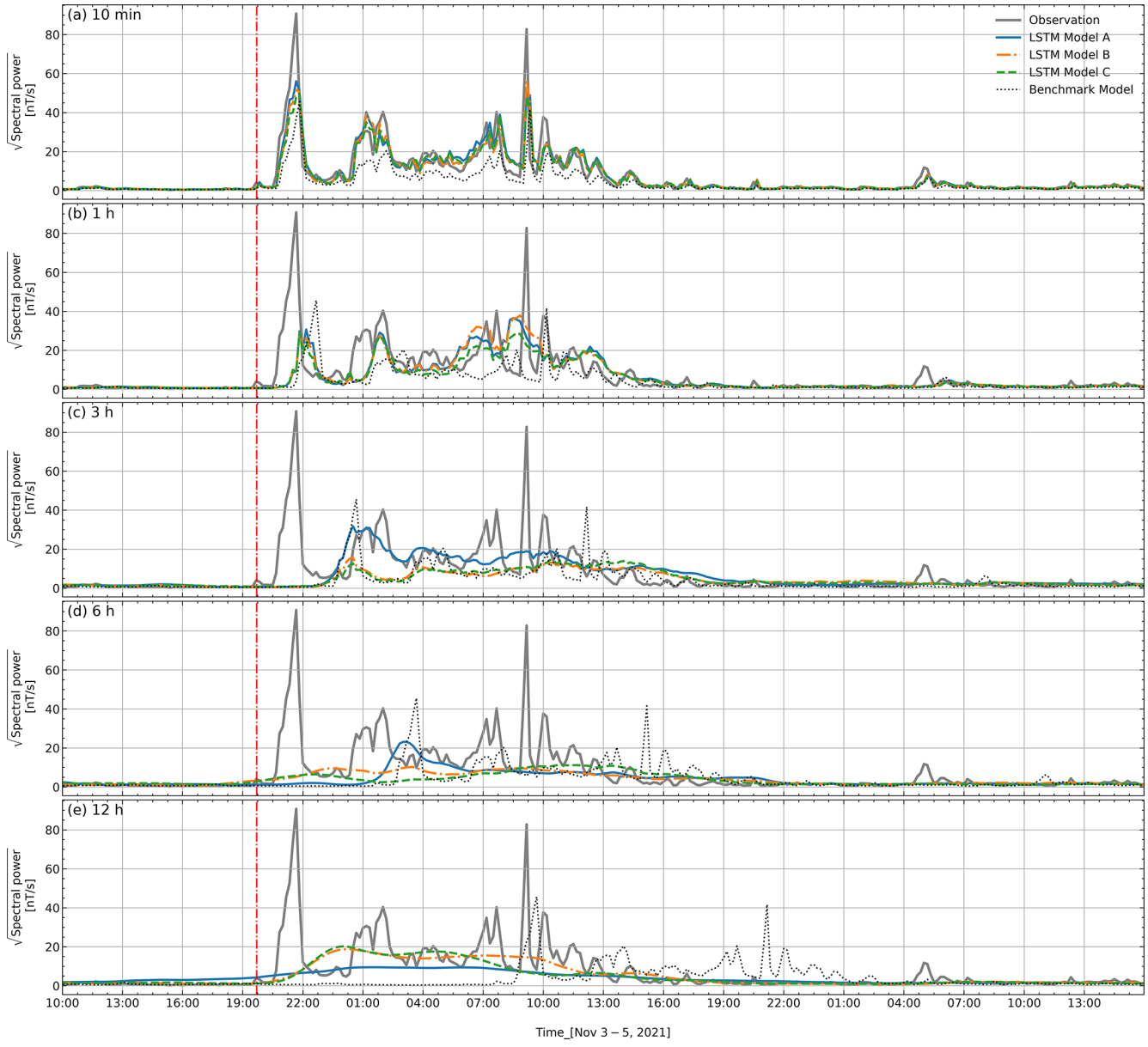


Figure 5. Comparison of observed and forecasted values of the GIC proxy P_0 (estimated dH/dt power over 0–0.01 Hz) derived from Rørvik magnetometer measurements during 3–5 November 2021. Forecasts with time horizons of 10 min, 1 h, 3 h, 6 h, and 12 h are shown in panels **a–e** as indicated in the panel caption. Forecast outputs from Model A ("high-resolution" artificial space weather forecast inputs), Model B (low-resolution inputs), and Model C (high-resolution inputs except for predicted arrival time) are respectively indicated by solid blue lines, orange dash-dotted lines, and green dashed lines. In each panel the observed P_0 time series and the benchmark model predictions are respectively indicated by a thick, gray line and a black dotted line. CME arrival at 19:42 UT on 3 November 2021, is indicated by the vertical dash-dotted red line.

black dotted lines. (Results for Model C are essentially identical to those of Model A and are omitted.) The coefficients of determination R^2 for Models A and B decrease uniformly from 88 % for forecast horizon $h = 10$ min to ~ 50 % for $h = 3$ h. The performance of Model A recorded performance metrics with a slight increasing trend over $h = 3 - 12$ h with values of 50 %, 52 %, and 56 %. On the other hand, Model B for $h = 3 - 12$ h continues to decrease gradually with values 52 %, 42 %, and 42 % respectively. The R^2 value for the benchmark model monotonically decreases with increasing forecast horizon h and is everywhere lower than R^2 values for Models A and B.

In addition to the benchmark model, we also examined the performance of a simple persistence model (not shown) in which the current value was predicted to be the same as the value measured at an earlier time (e.g., $h =$

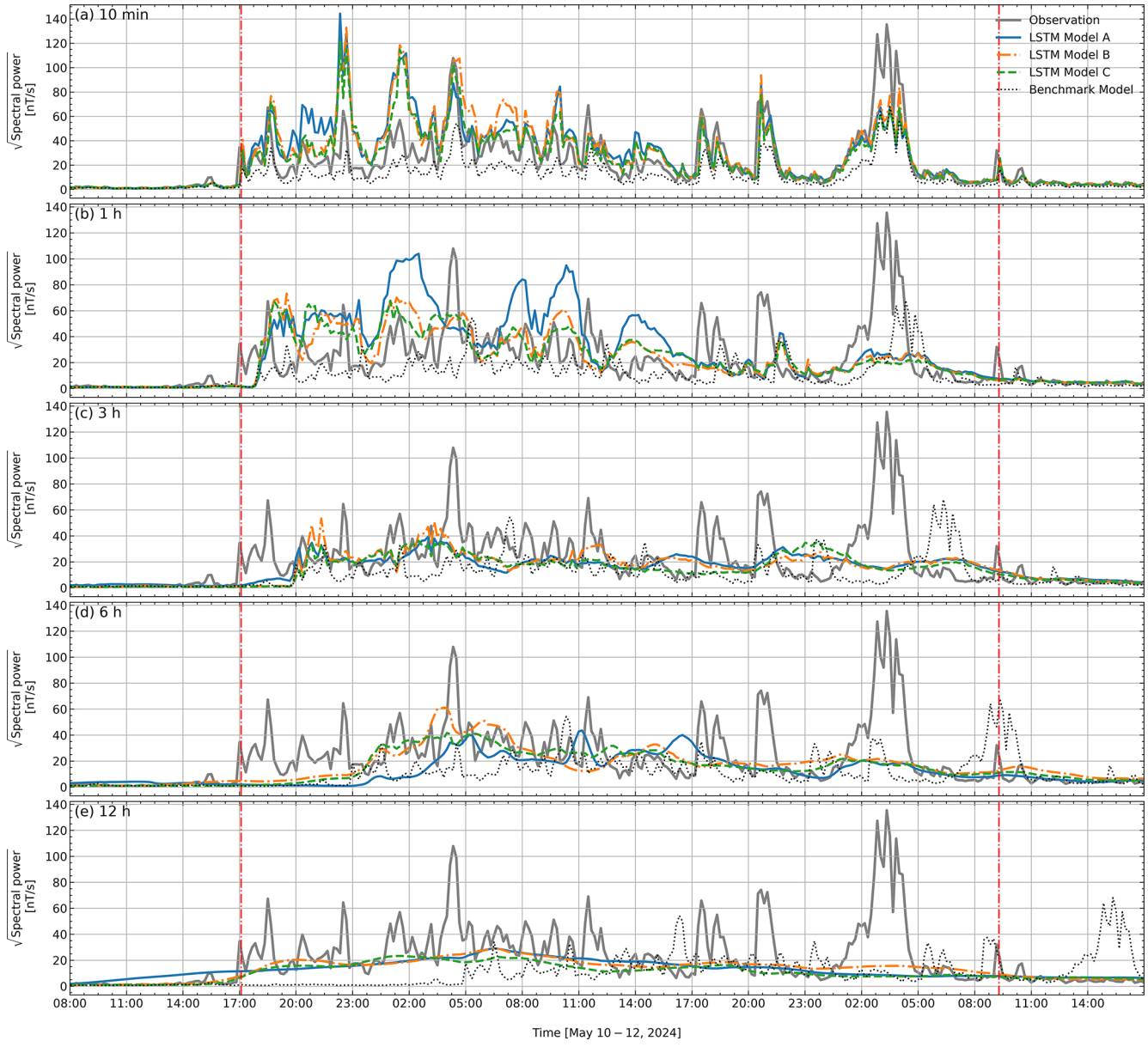


Figure 6. Estimated dH/dt power over 0–0.01 Hz (thick, gray line) derived from Rørvik magnetometer measurements as well as model forecast and benchmark model forecast for 10–12 May 2024, in the same format as Fig. 5. CME arrivals at 17:05 UT on 10 May and 09:17 UT on 12 May are indicated by vertical dash-dotted red lines.

144 in Eq. (4) for a persistence model based on measurements made 24 h earlier with data at 10 min resolution). These models yielded coefficients of determination $R^2 =$ (85.7, 54.0, 31.2, 16.1, 2.0, −44.0) for forecasting horizons of $h =$ (10 min, 1 h, 3 h, 6 h, 12 h, 24 h). We also performed this calculation for the climatology model described in Sect. 3.3 and obtained $R^2 = -2.9$. In summary, the persistence and climatology generally performed much worse than Models A–C. The exception was the persistence model with a forecasting horizon of 10 min, which performed approximately as well as Models A–C.

Examining the RMSE of every forecast horizon, there are similar trends as a function of forecasting horizon h for all models, as shown in Fig. 8b. Models A and B had the same in lower forecast horizons up to 3 h in advance, with RMSE values of 1.4–2 nT s^{-1}, respectively. For longer forecasting horizons ($h = 3-12$ h), for Model A there is a slight decrease from 2.06–1.97 nT s^{-1}, while for Model B, RMSE increases from 2.0–2.2 nT s^{-1}. The RMSE values are calculated from the target and forecast variables $y(= \log_{10} P_0)$ and $\hat{y}$, and are then quoted here after raising them to the power of 10.

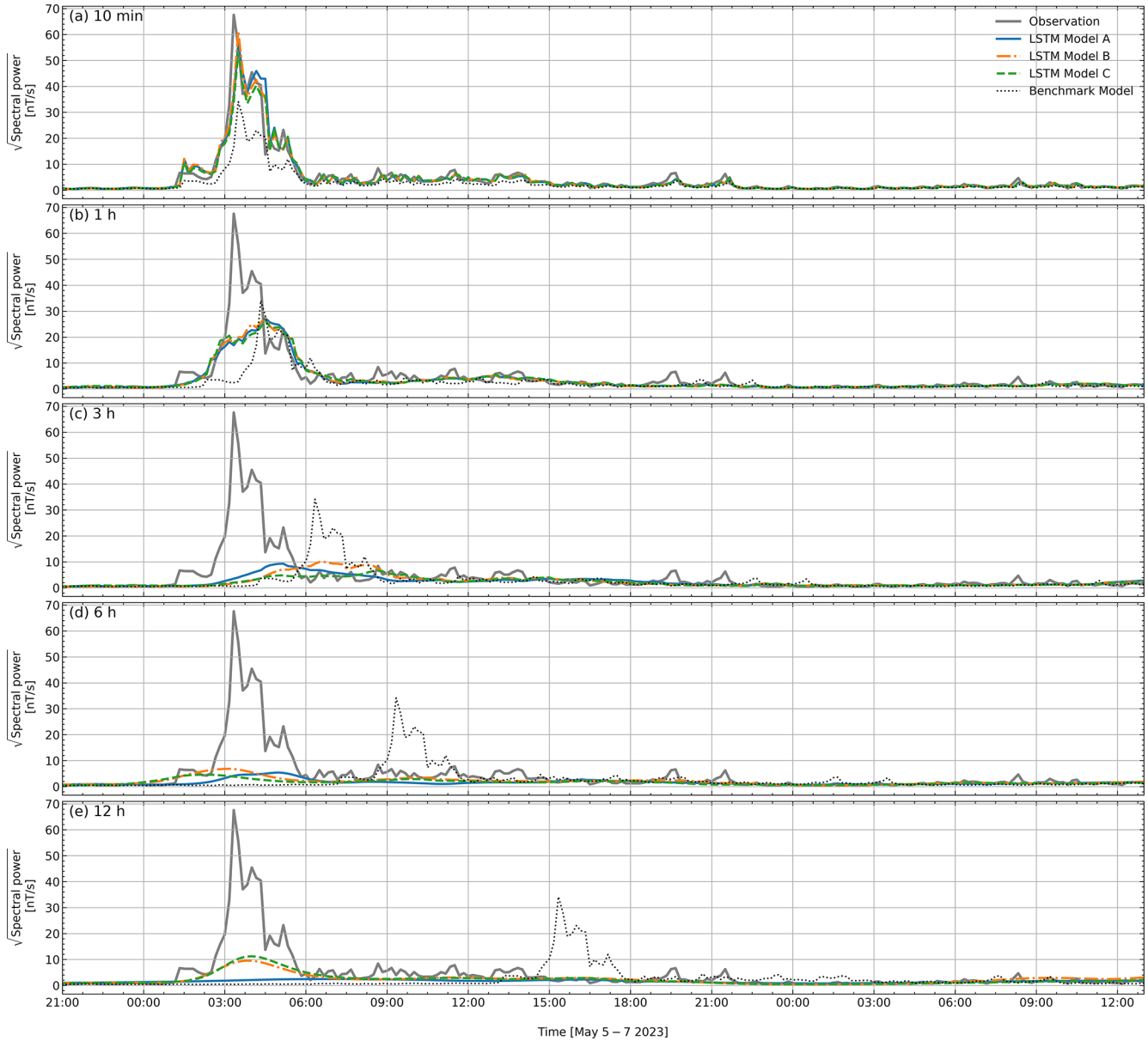


Figure 7. Calculated dH/dt power over 0–0.01 Hz (thick, gray line) derived from Rørvik magnetometer measurements as well as model forecast and benchmark model forecast for 5–7 May 2023, in the same format as Fig. 5. The DSCOVR satellite observed HSS/SIR activity beginning at approximately 05:30 UT on 6 May 2023 (Community Coordinated Modeling Center, 2026).

In the benchmark model, there is a steady decrease in model performance with increasing forecast horizon h, with $h = 12$ h showing the worst RMSE value of 2.75 nT s^{-1}, and $h = 10$ min slightly higher than those of intelligent models, with a value of 1.56 nT s^{-1}.

In this study feature importance is calculated using the gradient-based method described in Sect. 3.6. Figure 9 shows contributions of every input feature to the model output for all forecast horizons, demonstrated in panels a–e. For $h =$ 10 min (Fig. 9a) the input feature P_0 contributes the most, with a gradient slightly higher than 0.005. The other significant input features (gradients above 0.005) are IMF B_z, v, sin(UT), and cos(UT). All other features have little influence on the performance of the models, as indicated by their having gradients similar to that of the normally distributed random input denoted by $\mathcal{N}(0, 1)$ in Table 1.

For $h = 1$ h (Fig. 9b) the importance of IMF B_z is greatest, followed by P_0 and v with gradient values above 0.0015. Forecast input features, such as predicted IMF B_z and v, also make a non-negligible contribution to the 1 h forecast.

For $h = 3$ h (Fig. 9c) Models A and B respond differently to the various input features. The importance of B_z, v,

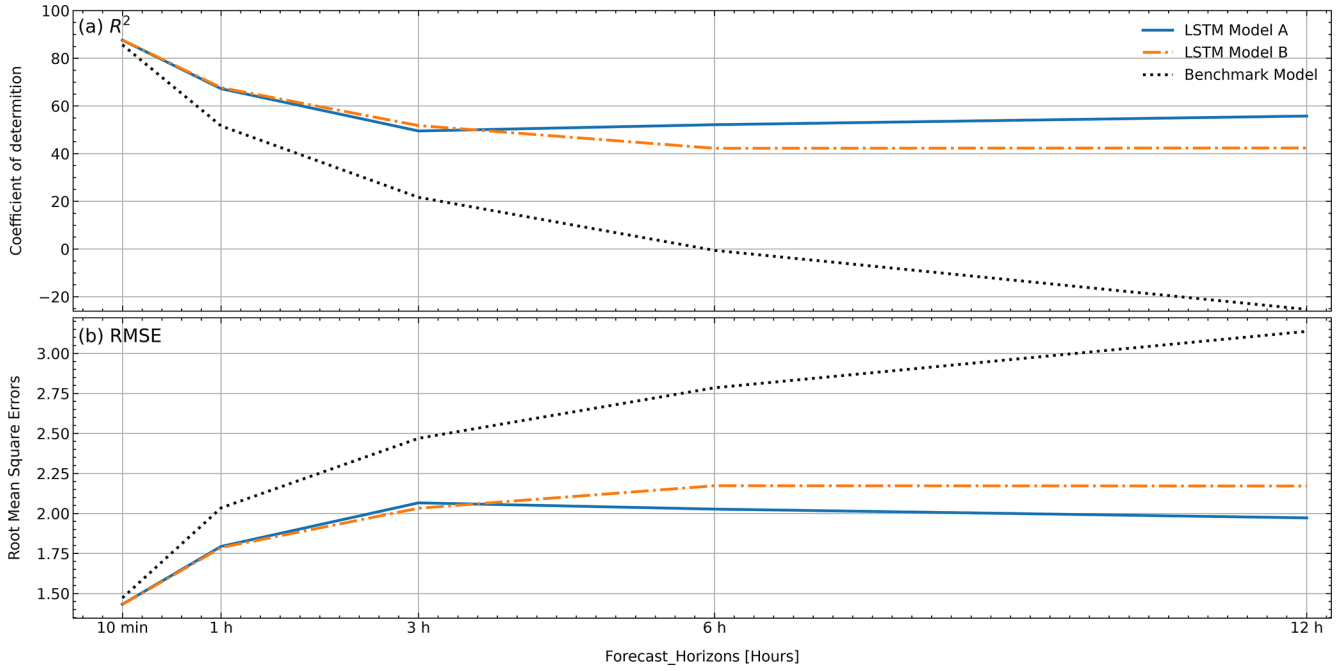


Figure 8. Statistical evaluation of model performance for each forecasting horizon h. **(a)** Coefficient of determination R^2. **(b)** RMSE metrics for each model. Results for Model C are essentially indistinguishable from those of Model A on the scales shown in this figure and are therefore omitted.

sin(UT), and pred_t dominates for Model A. (Predicted arrival time information is not provided to Model C and therefore has zero importance.) On the other hand, predicted arrival time and predicted IMF B_z are the two most important features for Model B, with multiple other features making smaller, yet still non-negligible contributions.

For $h = 6$ h (Fig. 9d), P_0 has little to no importance. Predicted IMF B_z and v are important for all three models, and predicted arrival time is also relatively important for Models A and B.

The trends for $h = 12$ h (Fig. 9e) are generally the same as those for $h = 6$ h. The most important features for all models are externally driven, with the gradients having higher values for Models A and C than those for Model B.

In addition to the foregoing case study and feature importance results, we have also carried out a binary classification analysis of events using the True Skill Statistic

$$\mathrm{TSS} = \mathrm{Sensitivity} + \mathrm{Specificity} - 1 \ , \tag{6}$$

where

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

and

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \ . \tag{8}$$

Defined in this way the TSS is a value between -1 and 1, where the former value corresponds to "no forecasting capability" and the latter corresponds to a perfect forecast. In these expressions "true positives" (TP) are those for which the predicted values $\hat{y}_{t+h}$ of the target variable $y = \log_{10} P_0$ for forecast horizon h both reach or exceed a chosen threshold value. "True negatives" (TN) are those for which both forecasted and observed values are less than the threshold. For false negatives (FN) the forecasted value does not reach the threshold percentile while the observed value exceeds it, and vice versa for false positives (FP) (Welling et al., 2018). We choose to use 0.985 as a percentile threshold for calculating the confusion matrix. This threshold corresponds to $\sim 10\,\mathrm{nT\,s^{-1}}$ in the target variable P_0, which serves as a practical threshold for assessing the potential impact of geomagnetic activity on nearby deployed infrastructure.

We also use the Matthew Correlation Coefficient

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \ . \tag{9}$$

as measures of model performance, accuracy, and unbiased balance between predicted classes. The metrics represented by Eqs. (6)–(9) are used in a number of previous investigations of space weather forecasting (e.g., Hu et al., 2024; Bailey et al., 2022).

Table 3 presents the confusion matrix and several other statistical classification metrics of model performance for Models A and B as well as the benchmark model. (Metrics for Model C are almost identical to those of Model A and are not shown.) All of the models perform well for $h = 10$ min, with Models A–C having the highest values of TP, MCC, and TSS.

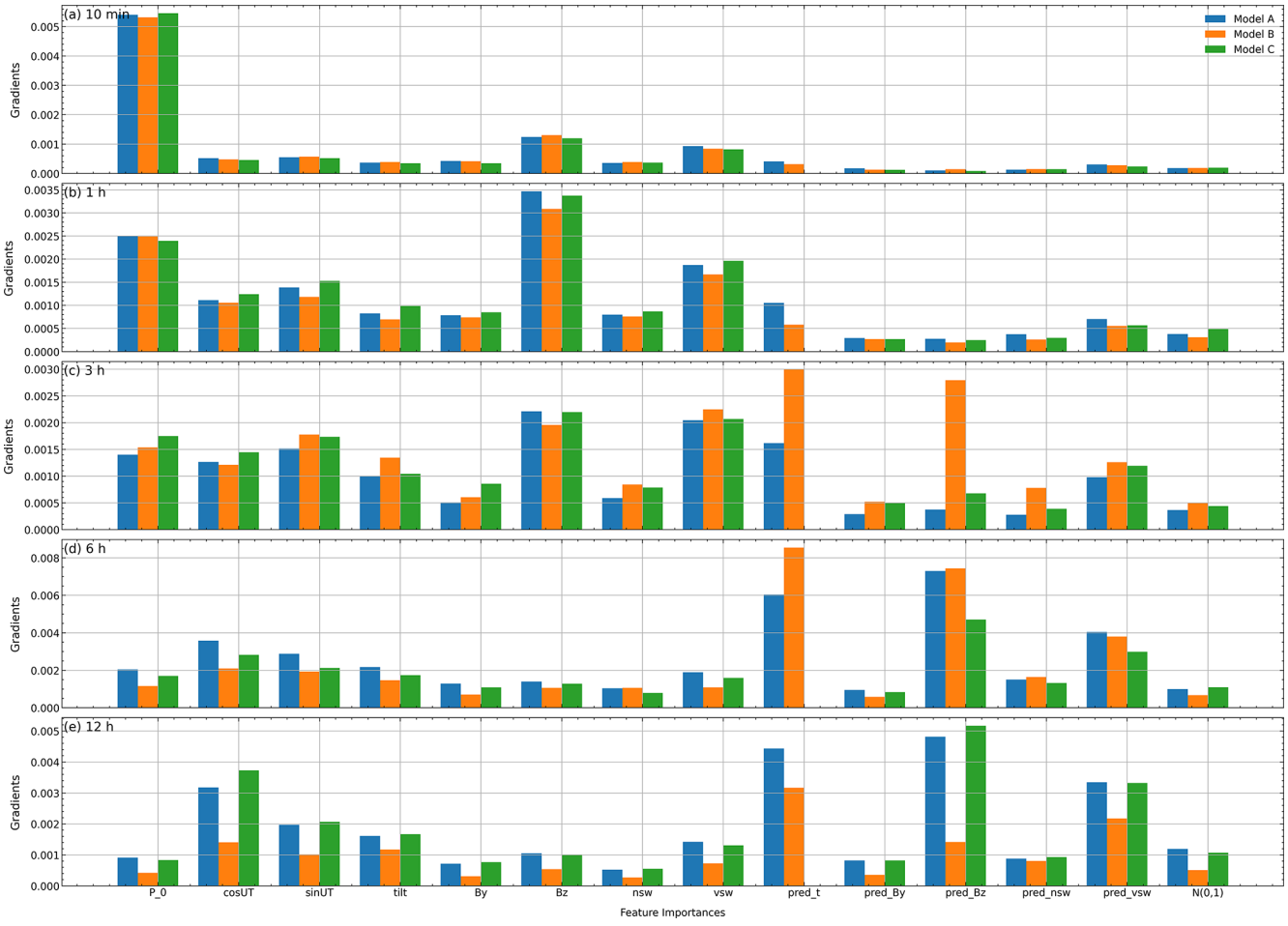


Figure 9. Feature importance for each model forecasting horizon defined using the gradient tape method described in Sect. 3.6. The y axis indicates the relative importance of each feature.

Comparing our Table 3 with Table 3 in Hu et al. (2024), only TSS for both models in a 10 min forecasting horizon matches the value TSS = 0.66 that they obtained, while for a forecasting horizon $h = 1$ h we find our models have slightly higher TSS values. Here it is important to observe that their study examined the geoelectric field at mid-latitudes, whereas this study is focused on a GIC proxy at high latitudes, so this indirect comparison is only suggestive. Regarding our choice of 0.985 as the threshold percentile, we also tested increasing the percentile threshold to 0.998 (corresponding to $P_0 \approx 20\,\mathrm{nT\,s^{-1}}$) and found that this only slightly decreased the TSS and MCC. We therefore deem the results with a threshold of 0.985 to be representative.

5 Discussion

Results presented in the previous section illustrate that the deep learning models we employ exhibit modest learning levels and predictive power. The forecasting power of the three deep learning models initially decreases for forecasting horizons up to about 3 h, and then show a trend in Fig. 8 toward constant or slightly increasing forecasting power for horizons above 3 h, with a similar degree of performance ($R^2 = 50\,\%$). A similar finding was reported in the study of Zewdie et al. (2021).

Despite the rather modest levels of predictive power exhibited by the deep learning models, they all outperform the benchmark model for all horizons. These findings agree with the general conclusion from Wang et al. (2026, see their Table 6) that deep learning-based models tend to perform better than other forecasting strategies when the goal is long-term forecasting. It is however also important to observe that alternative forecasting strategies based on, e.g., persistence modeling (Zewdie et al., 2021; Bhandari and Shakya, 2026), gradient boost machines (Iong et al., 2022; Bhandari and Shakya, 2026), or hybrids of persistence and deep learning (Hu et al., 2023) can perform as well or better than deep learning-based models depending on the forecasting horizon and the model design.

Table 3. Comparison of model performance parameters defined in Eqs. (6)–(9) between LSTM and benchmark models with forecast horizons between 10 min and 12 h. Values for TP, TN, FP, and FN are given as a percentage of the total number of measurements in the test data set ($N = 153\,992$). The total numbers of positives and negatives are respectively 2310 and 151 682.

		Forecast Horizons				
Metrics	Model	10 min	1 h	3 h	6 h	12 h
TP	Model A*	1.05	0.56	0.21	0.34	0.49
	Model B	1.04	0.55	0.30	0.22	0.19
	Benchmark	0.39	0.17	0.09	0.07	0.04
TN	Model A	98.03	97.89	98.21	97.93	97.78
	Model B	98.04	97.89	98.12	98.24	98.30
	Benchmark	98.42	98.36	98.36	98.37	98.40
FP	Model A	0.41	0.55	0.23	0.50	0.66
	Model B	0.39	0.54	0.32	0.19	0.13
	Benchmark	0.02	0.08	0.07	0.06	0.03
FN	Model A	0.52	1.00	1.35	1.23	1.08
	Model B	0.53	1.01	1.27	1.35	1.37
	Benchmark	1.17	1.39	1.48	1.50	1.52
MCC	Model A	0.69	0.42	0.25	0.29	0.36
	Model B	0.69	0.42	0.30	0.27	0.26
	Benchmark	0.49	0.27	0.17	0.15	0.12
TSS	Model A	0.66	0.35	0.13	0.21	0.31
	Model B	0.66	0.35	0.19	0.14	0.12
	Benchmark	0.25	0.11	0.06	0.04	0.03

* Model C values are very similar to those for Model A and are therefore not shown.

The feature importance shown in Fig. 9 was calculated for each model and forecasting horizon h. This was done using the post hoc gradient based method (e.g., Jin et al., 2023, and references therein), which we find to be most appropriate for evaluation of deep learning models of the sort we have presented in this study. Results in Fig. 9 show that the importance of the various input features varies significantly with forecasting horizon h. For instance, for $h = 10$ min and $h = 1$ h, only IMF B_z and solar wind velocity v significantly contribute beyond that of the target variable P_0 itself. We also see that as forecasting horizon h increases, the models rely increasingly on accurate information about predicted IMF B_z and predicted v.

While Fig. 9c–e indicate that CME arrival time is an important input feature for time horizons $h \geq 3$ h, Fig. 8 clearly demonstrates that Model C, which is constructed *without* CME arrival time information, performs just as well as Model A. We therefore conclude that high-resolution solar wind and IMF forecasts are necessary to achieve the model performances we have reported for Models A and C. CME arrival time information is apparently unimportant when high-resolution forecasts of IMF and solar wind conditions are available.

In contrast, it is somewhat unclear what role CME arrival time information plays when only low-resolution (but nevertheless accurate) solar wind and IMF forecasts are provided to the model: According to Fig. 9c–e, CME arrival time is the single most important piece of information for Model B's predictions. To test this we trained another version of Model B (low-resolution space weather forecast inputs) in which information about IMF B_z and IMF B_y was excluded but CME arrival time was retained. In this model (not shown), CME arrival time had no importance for any of the five forecast horizons; this model also performed generally much worse than Model B for forecasting horizons $h = 6$ h and $h = 12$ h. We therefore conclude that when only low-resolution solar wind and IMF forecasts are available, information about CME arrival time could contribute to model performance. This result incidentally highlights that a study of model performance as a function of temporal resolution and accuracy of solar wind forecast parameters would likely provide clarity around this point.

To summarize, our results indicate that when only low-resolution forecasts of IMF and solar wind conditions are available, CME arrival time information may improve model performance. In the absence of any information about future IMF conditions, CME arrival time is apparently unimportant. This conclusion corresponds to the qualitative but well known idea that the geoeffectiveness of a CME is closely tied to the accompanying IMF conditions (e.g., Kang et al., 2006).

One unresolved aspect of the model performance is that the model inputs do not always seem to provide the information necessary for the model to predict enhanced geomagnetic activity. This is illustrated in Fig. 7, where we observe that the deep learning models performed poorly in longer forecast horizons beyond 1 h. It is interesting to note that this particular event is associated with HSS/SIR. By contrast, the event in Fig. 5 is a purely CME-driven storm. This leads us to speculate that perhaps the nature of the solar wind and IMF signatures of some types of disturbances, such as HSS/SIR events, are not clearly identifiable from the inputs provided to the model for longer forecasting horizons. This same seeming insensitivity to elevated solar and IMF conditions is apparent in model outputs for the latter half of the time window shown in Fig. 6 for the May 2024 storm: a large enhancement in geomagnetic activity was observed at Rørvik, several hours before the arrival of a second CME which was not geoeffective at 09:17 UT on 12 May 2024. The second enhancement was not forecasted by the models for forecasting horizons beyond 1 h. We reserve determining whether model predictive power in fact varies based on the type of disturbance carried by the solar wind as the topic of a potential future study.

Our finding about the limited utility of existing space weather forecast products likely applies to other domains concerned with space weather, such as predictions of total electron content (TEC), GPS scintillation, and satellite drag. A possible counterexample to this suggestion is the work of Adolfs et al. (2024), who find that the performance of their TEC forecasting model did not improve significantly with the addition of external inputs.

6 Conclusions

A central goal of this study was to determine whether we could produce an accurate forecast of local geomagnetic activity at the Rørvik magnetometer station up to 12 h in advance using artificial forecasts of space weather conditions and information about CME arrival. We find that high-resolution (time resolution of 10 min) space weather forecasts of IMF B_z and v 24 h in advance would enable a reasonably accurate ($R^2 = 55\,\%$) forecast of local geomagnetic activity 12 h in advance. In this case the inclusion of CME arrival time provides no useful information to the model. When only low-resolution (time resolution of 10 min but smoothed with a 36 h window) space weather forecasts are available, CME arrival time information may improve model performance. In the absence of any information about IMF B_z, CME arrival time is apparently unimportant.

Increasing the predictive power of forecasts with horizons beyond 1 h therefore seems to require either heliosphere/solar wind models that are able to accurately predict the IMF – a capability that does not exist at present – or in situ solar wind monitors that, for example, have a different viewing angle, such as the European Space Agency's upcoming Vigil mission which will be located at L5.

This study is, to our knowledge, the first to demonstrate via artificial intelligence-based models what is otherwise well established in the space weather community: That information about the IMF B_z component is crucial to predicting the geoeffectiveness of an inbound CME. (It is not, however, the first machine learning-based demonstration of the importance of IMF B_z; see, e.g., Lotz and Cilliers, 2015; Coughlan et al., 2025.) This study therefore suggests how one might incorporate other potentially relevant measures of space weather, such as solar energetic proton intensity or X-ray flux. We speculate that these quantities might allow a deep learning-based model to further modulate the intensity of predicted local geomagnetic activity.

Appendix A: Alternative splitting

The alternative splitting strategy that we employed divides the complete data set into subsets made up of different years and seeks to ensure that the mean and standard deviation of P_0 values is comparable between the training and test subsets, similar to the work of Wintoft and Wik (2021) (see their Table 1). Table A1 summarizes the chronological splitting strategy we use in the main text (first two rows) and the alternative splitting strategy (last two rows).

One small difficulty presented by this alternative splitting strategy is that it introduces large temporal gaps into the training and test subsets which, if not handled correctly, would result in many of the sample windows generated by the process illustrated in Fig. 4 to contain measurements from disparate years. These inaccurate sample windows would inevitably degrade model performance. To avoid this, we first generated sample windows from each individual contiguous period – five for training and three for testing, as shown in the last two rows of Table A1. We then combined the sample windows generated from each contiguous period into a final data set that was used for either training and validation or testing.

As discussed in Sect. 3.4 and shown in Fig. A1a, the values of $\log_{10} P_0$ used for training the models described throughout this manuscript (blue histogram) has a slightly lower mean than the distribution of $\log_{10} P_0$ values used for testing (orange histogram). The standard deviation of each distribution is nevertheless comparable. Figure A1b shows the distributions of $\log_{10} P_0$ produced by the alternative splitting strategy, which are clearly better aligned than the distributions shown in Fig. A1a.

Figure A2 shows that the correlation coefficients and measurement-model residuals produced by this alternative splitting strategy are very comparable to those shown in Fig. 8 that are produced by the chronological data splitting strategy. From this comparison we conclude that both of these splitting strategies are valid for this study.

Table A1. Description of splitting used in main study and alternative data splitting for $\log_{10}(P_0)$.

Splitting strategy		Period	Count	Mean	Std
Chronological	Training[a]	2010–September 2021	603 438	−2.33	0.86
	Testing	October 2021–September 2024	154 124	−2.15	0.89
Wintoft and Wik (2021)	Training[a]	2010–2011, 2013–2014, 2016, 2019–2020, 2022–2023	616 683	−2.28	0.87
	Testing	2013, 2021, 2024[b]	140 933	−2.37	0.87

[a] Training data is automatically split into training and validation with the ratio 80 : 20 by TimeSeriesSplit; see Sect. 3.4.
[b] January to September 2024, inclusive.

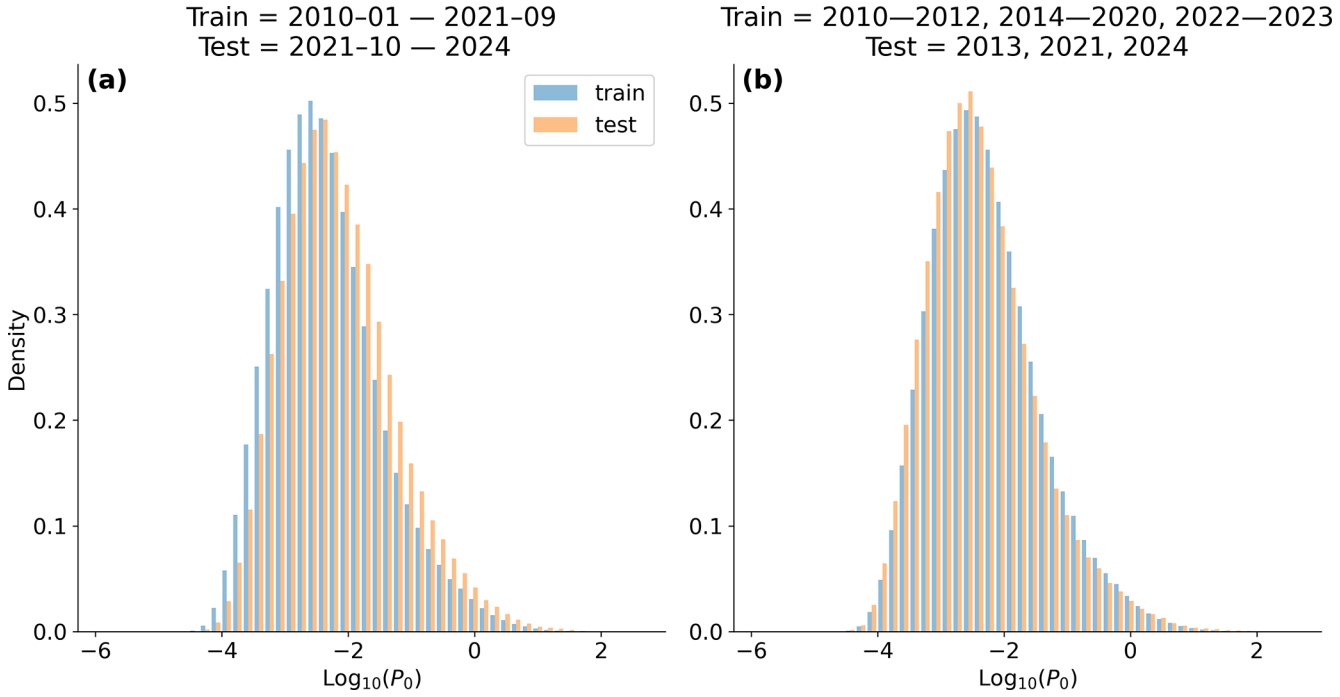


Figure A1. Histograms of $\log_{10}(P_0)$ values used for training (blue) and testing (orange). Histograms for the splitting used in the main study are shown in panel **a**, and those for the alternative splitting described in this appendix are shown in panel **b**.

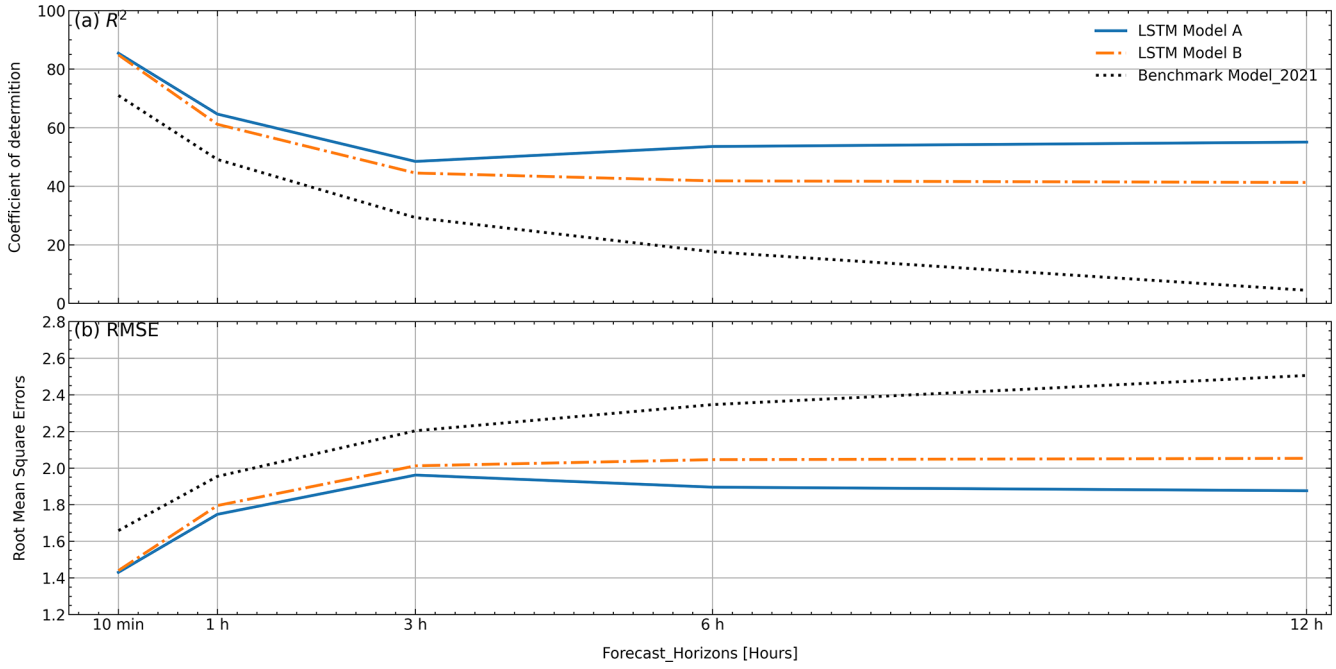


Figure A2. Statistical evaluation of model performance for each forecasting horizon h when trained using the alternative splitting strategy, in the same layout as Fig. 8.

Data availability. ACE satellite measurements of IMF and solar wind speed are available at https://sohoftp.nascom.nasa.gov/sdb/goes/ace/daily/ (last access: 15 August 2025). Rørvik magnetometer data are made available by the Tromsø Geophysical Observatory at https://flux.phys.uit.no/ascii/ (last access: 15 August 2025). The latest version of the Richardson and Cane near-Earth ICME list (https://doi.org/10.7910/DVN/C2MHTH, Richardson and Cane, 2024) is available at https://izw1.caltech.edu/ACE/ASC/DATA/level3/icmetable2.htm (last access: 21 March 2025). The Space Weather Database Of Notifications, Knowledge, Information (DONKI) (Community Coordinated Modeling Center, 2026) is located at https://kauai.ccmc.gsfc.nasa.gov/DONKI/ (last access: 10 September 2026).

Author contributions. SO: model design and creation (lead), formal analysis, writing (equal), and visualization (lead). SMH: project administration, funding acquisition, supervision (lead), training dataset production, study conceptualization (lead), writing (equal). AK: model design, conceptualization, validation, and writing. MGJ: supervision, conceptualization, validation, writing, and curation of the Rørvik magnetometer dataset. MJO: conceptualization including expertise on solar wind forecasting methods, writing – review and editing, and validation. KST: supervision, contribution of expertise on induction in power systems, writing – review and editing. RL: validation, contribution of expertise on GIC events, writing – review and editing.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Financial support. The work of Stephen Omondi, Spencer Mark Hatch, and Andreas Kvammen was supported by the University of Bergen via the UiB Idé program. Spencer Mark Hatch was additionally supported by Research Council of Norway project 355484. Andreas Kvammen acknowledges support from the Research Council of Norway through project 353378. Mathew J. Owens is partially funded by Natural Environment Research Council (NERC) grant no. NE/Y001052/1 and Science and Technology Facilities Council (STFC) grant no. UKRI1207.

Review statement. This paper was edited by Georgios Balasis and reviewed by Amy Keesee and one anonymous referee.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, https://www.tensorflow.org/ (last access: 10 August 2025), 2015.

Abda, Z. M. K., Aziz, N. F. A., Kadir, M. Z. A. A., and Rhazali, Z. A.: A Review of Geomagnetically Induced Current Effects on Electrical Power System: Principles and Theory, IEEE Access, 8, 200237–200258, https://doi.org/10.1109/ACCESS.2020.3034347, 2020.

Adolfs, M., Hoque, M. M., and Shprits, Y. Y.: Forecasting 24-Hr Total Electron Content With Long Short-Term Memory Neural Network, Journal of Geophysical Research: Machine Learning and Computation, 1, https://doi.org/10.1029/2024JH000123, 2024.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining, Anchorage, Alaska, USA, 4–8 August 2019, 2623–2631, https://doi.org/10.1145/3292500.3330701, 2019.

Apicella, A., Isgrò, F., and Prevete, R.: Don't push the button! Exploring data leakage risks in machine learning and transfer learning, Artif. Intell. Rev., 58, 339, https://doi.org/10.1007/s10462-025-11326-3, 2025.

Arge, C. N. and Pizzo, V. J.: Improvement in the prediction of solar wind conditions using near-real time solar magnetic field updates, J. Geophys. Res.-Space, 105, 10465–10479, https://doi.org/10.1029/1999JA000262, 2000.

Arge, C. N., Luhmann, J., Odstrcil, D., Schrijver, C., and Li, Y.: Stream structure and coronal sources of the solar wind during the May 12th, 1997 CME, J. Atmos. Sol-Terr. Phy., 66, 1295–1309, https://doi.org/10.1016/j.jastp.2004.03.018, 2004.

Bailey, R. L., Leonhardt, R., Möstl, C., Beggan, C., Reiss, M. A., Bhaskar, A., and Weiss, A. J.: Forecasting GICs and Geoelectric Fields From Solar Wind Data Using LSTMs: Application in Austria, Space Weather, 20, e2021SW002907, https://doi.org/10.1029/2021SW002907, 2022.

Barnard, L. and Owens, M.: HUXt – An open source, computationally efficient reduced-physics solar wind model, written in Python, AIP Conf. Proc. (Frontiers in Physics), 10, https://doi.org/10.3389/fphy.2022.1005621, 2022.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B.: Algorithms for hyper-parameter optimization, Adv. Neur. Inf., 24,https://papers.nips.cc/paper_files/paper/2011 (last access: 6 November 2026), 2011.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D. D.: Hyperopt: a python library for model selection and hyperparameter optimization, Computational Science and Discovery, 8, 014008, https://doi.org/10.1088/1749-4699/8/1/014008, 2015.

Bhandari, P. and Shakya, S.: Comparative Benchmarking Study of Deep Learning Models for Geomagnetic Storms Prediction, Journal of Trends in Computer Science and Smart Technology, 8, 884–902, https://doi.org/10.36548/jtcsst.2026.3.022, 2026.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, 2001.

Camporeale, E.: The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting, Space Weather, 17, 1166–1207, https://doi.org/10.1029/2018SW002061, 2019.

Camporeale, E., Wing, S., Johnson, J., Jackman, C. M., and McGranaghan, R.: Space Weather in the Machine Learning Era: A Multidisciplinary Approach, Space Weather, 16, 2–4, https://doi.org/10.1002/2017SW001775, 2018.

Cane, H. V. and Richardson, I. G.: Interplanetary coronal mass ejections in the near-Earth solar wind during 1996–2002, J. Geophys. Res., 108, 1156, https://doi.org/10.1029/2002JA009817, 2003.

Chu, X., Jia, L., McPherron, R. L., Li, X., and Bortnik, J.: Imbalanced Regression Artificial Neural Network Model for Auroral Electrojet Indices (IRANNA): Can We Predict Strong Events?, Space Weather, 23, https://doi.org/10.1029/2024SW004236, 2025.

CIGRE: TB 780 – Understanding of geomagnetic storm environment for high voltage power grids, Tech. rep., https://www.e-cigre.org/publications/detail/780-understanding-of-geomagnetic-storm-environment-for-high-voltage-power-grids.html (last access: 13 February 2024), 2019.

Community Coordinated Modeling Center: Space Weather Database Of Notifications, Knowledge, Information (DONKI), https://ccmc.gsfc.nasa.gov/DONKI/ (last access: 6 November 2026), 2026.

Conde, D., Castillo, F. L., Escobar, C., García, C., García, J. E., Sanz, V., Zaldívar, B., Curto, J. J., Marsal, S., and Torta, J. M.: Forecasting Geomagnetic Storm Disturbances and Their Uncertainties Using Deep Learning, Space Weather, 21, e2023SW003474, https://doi.org/10.1029/2023SW003474, 2023.

Coughlan, M., Keesee, A., Pinto, V., Mukundan, R., Marchezi, J. P., Adewuyi, M., Tibbetts, J., Johnson, J., Connor, H., and Hampton, D.: Using Machine Learning Explainability Techniques to Examine Drivers of Ground Magnetic Field Localization, Space Weather, 23, https://doi.org/10.1029/2025SW004391, 2025.

EPRI: Research Findings for Geomagnetic Disturbance Research Work Plan: Summary Report, Tech. rep., https://www.epri.com/research/products/000000003002019720 (last access: 6 November 2026), 2020.

Fisher, A., Rudin, C., and Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, J. Mach. Learn. Res., 20, 1–81, 2019.

Fry, E. K.: The risks and impacts of space weather: Policy recommendations and initiatives, Space Policy, 28, 180–184, https://doi.org/10.1016/j.spacepol.2012.06.005, 2012.

Gannon, J. L., Swidinsky, A., and Xu, Z. (Eds.): Geomagnetically Induced Currents from the Sun to the Power Grid, Wiley, https://doi.org/10.1002/9781119434412, 2019.

Girosi, F., Jones, M., and Poggio, T.: Regularization theory and neural networks architectures, Neural Comput., 7, 219–269, 1995.

Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., and Sculley, D.: Google vizier: A service for black-box optimization, in: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 13–17 Au-

gust 2017, Halifax, Nova Scotia, Canada, https://www.kdd.org/kdd2017 (last access: 6 November 2026), 1487–1495, 2017.

Gonzalez, W. D., Echer, E., Tsurutani, B. T., Gonzalez, A. L. C., and Lago, A. D.: Interplanetary Origin of Intense, Superintense and Extreme Geomagnetic Storms, Space Sci. Rev., 158, 69–89, https://doi.org/10.1007/s11214-010-9715-2, 2011.

Hapgood, M.: Towards a scientific understanding of the risk from extreme space weather, Adv. Space Res., 47, 2059–2072, https://doi.org/10.1016/j.asr.2010.02.007, 2011.

Hatch, S. M. and LaBelle, J.: Application of a new method for calculation of low-frequency wave vectors, in: Planetary Radio Emissions VIII, edited by: Fischer, G., Mann, G., Panchenko, M., and Zarka, P., 247–260, Austrian Academy of Sciences Press, https://doi.org/10.1553/PRE8s247, 2018.

Hochreiter, S.: Long Short-term Memory, Neural Computation, MIT-Press, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Hu, A., Camporeale, E., and Swiger, B.: Multi-Hour-Ahead Dst Index Prediction Using Multi-Fidelity Boosted Neural Networks, Space Weather, 21, https://doi.org/10.1029/2022SW003286, 2023.

Hu, A., Camporeale, E., Lucas, G., and Berger, T.: LiveWire: Horizontal Geoelectric Field Prediction With 1-hr Lead-Time Using Multi-Fidelity Boosted Neural Networks, Journal of Geophysical Research: Machine Learning and Computation, 1, https://doi.org/10.1029/2024JH000151, 2024.

Hutter, F., Hoos, H. H., and Leyton-Brown, K.: Sequential model-based optimization for general algorithm configuration, in: Learning and intelligent optimization: 5th International Conference, LION 5, Rome, Italy, January 17–21 2011, selected papers 5, 507–523, Springer, https://doi.org/10.1007/978-3-642-25566-3_40, 2011.

Iong, D., Chen, Y., Toth, G., Zou, S., Pulkkinen, T., Ren, J., Camporeale, E., and Gombosi, T.: New Findings From Explainable SYM-H Forecasting Using Gradient Boosting Machines, Space Weather, 20, e2021SW002928, https://doi.org/10.1029/2021SW002928, 2022.

Jin, W., Li, X., Fatehi, M., and Hamarneh, G.: Generating post-hoc explanation from deep neural networks for multi-modal medical image analysis tasks, MethodsX, 10, 102009, https://doi.org/10.1016/j.mex.2023.102009, 2023.

Johnsen, M. G.: Tromsø Geophysical Observatory geomagnetic data, https://flux.phys.uit.no/ascii/ (last access: 15 August 2025), 2025.

Kang, S.-M., Moon, Y.-J., Cho, K.-S., Kim, Y.-H., Park, Y. D., Baek, J.-H., and Chang, H.-Y.: Coronal mass ejection geoeffectiveness depending on field orientation and interplanetary coronal mass ejection classification, J. Geophys. Res.-Space, 111, https://doi.org/10.1029/2005JA011445, 2006.

Keesee, A. M., Pinto, V., Coughlan, M., Lennox, C., Mahmud, M. S., and Connor, H. K.: Comparison of Deep Learning Techniques to Model Connections Between Solar Wind and Ground Magnetic Perturbations, Frontiers in Astronomy and Space Sciences, 7, https://doi.org/10.3389/fspas.2020.550874, 2020.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May, 2015, Conference Track Proceedings, https://dblp.org/rec/journals/corr/KingmaB14 (last access: 6 November 2026), 2015.

Koch, P., Golovidov, O., Gardner, S., Wujek, B., Griffin, J., and Xu, Y.: Autotune: A derivative-free optimization framework for hyperparameter tuning, in: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining, London, England, 19–23 August 2018, 443–452, https://doi.org/10.1145/3219819.3219837, 2018.

Lakhina, G. S. and Tsurutani, B. T.: Geomagnetic storms: historical perspective to modern view, Geoscience Letters, 3, 5, https://doi.org/10.1186/s40562-016-0037-4, 2016.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A.: Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization, J. Mach. Learn. Res., 18, 1–52, http://jmlr.org/papers/v18/16-558.html (last access: 6 November 2026), 2018.

Liemohn, M. W., McCollough, J. P., Jordanova, V. K., Ngwira, C. M., Morley, S. K., Cid, C., Tobiska, W. K., Wintoft, P., Ganushkina, N. Y., Welling, D. T., Bingham, S., Balikhin, M. A., Opgenoorth, H. J., Engel, M. A., Weigel, R. S., Singer, H. J., Buresova, D., Bruinsma, S., Zhelavskaya, I. S., Shprits, Y. Y., and Vasile, R.: Model Evaluation Guidelines for Geomagnetic Index Predictions, Space Weather, 16, 2079–2102, https://doi.org/10.1029/2018SW002067, 2018.

Lones, M. A.: Avoiding common machine learning pitfalls, Patterns, 5, 101046, https://doi.org/10.1016/j.patter.2024.101046, 2024.

Lotz, S. and Cilliers, P.: A solar wind-based model of geomagnetic field fluctuations at a mid-latitude station, Adv. Space Res., 55, 220–230, https://doi.org/10.1016/j.asr.2014.09.014, 2015.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, Adv. Neur. In., 30, 2017.

Nembrini, S., König, I. R., and Wright, M. N.: The revival of the Gini importance?, Bioinformatics, 34, 3711–3718, https://doi.org/10.1093/bioinformatics/bty373, 2018.

Ngwira, C. M. and Pulkkinen, A. A.: An Introduction to Geomagnetically Induced Currents, in: Geomagnetically Induced Currents from the Sun to the Power Grid, edited by: Gannon, J. L., Swidinsky, A., and Xu, Z., https://doi.org/10.1002/9781119434412.ch1, 2019.

Odstrcil, D., Riley, P., and Zhao, X. P.: Numerical simulation of the 12 May 1997 interplanetary CME event, J. Geophys. Res.-Space, 109, https://doi.org/10.1029/2003JA010135, 2004.

Oyedokun, D., Heyns, M., Cilliers, P., and Gaunt, C.: Frequency Components of Geomagnetically Induced Currents for Power System Modelling, in: 2020 International SAUPEC/RobMech/PRASA Conference, Cape Town, South Africa, 29 to 31 January 2020, 1–6, IEEE, https://doi.org/10.1109/SAUPEC/RobMech/PRASA48453.2020.9041021, 2020.

Patterson, C. J., Wild, J. A., and Boteler, D. H.: Modeling "Wrong Side" Failures Caused by Geomagnetically Induced Currents in Electrified Railway Signaling Systems in the UK, Space Weather, 21, https://doi.org/10.1029/2023SW003625, 2023.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res., 12, 2825–2830, 2011.

Press, N. A.: Planning the Future Space Weather Operations and Research Infrastructure, National Academies Press, https://doi.org/10.17226/26128, 2021.

Prieto, G. A.: The Multitaper Spectrum Analysis Package in Python, Seismol. Res. Lett., 93, 1922–1929, https://doi.org/10.1785/0220210332, 2022.

Pulkkinen, A.: Geomagnetically Induced Currents Modeling and Forecasting, Space Weather, 13, 734–736, https://doi.org/10.1002/2015SW001316, 2015.

Pulkkinen, A., Bernabeu, E., Thomson, A., Viljanen, A., Pirjola, R., Boteler, D., Eichner, J., Cilliers, P. J., Welling, D., Savani, N. P., Weigel, R. S., Love, J. J., Balch, C., Ngwira, C. M., Crowley, G., Schultz, A., Kataoka, R., Anderson, B., Fugate, D., Simpson, J. J., and MacAlester, M.: Geomagnetically induced currents: Science, engineering, and applications readiness, Space Weather, 15, 828–856, https://doi.org/10.1002/2016SW001501, 2017.

Richardson, I. and Cane, H.: Near-Earth Interplanetary Coronal Mass Ejections Since January 1996, Harvard Dataverse [data set], https://doi.org/10.7910/DVN/C2MHTH, 2024.

Siciliano, F., Consolini, G., Tozzi, R., Gentili, M., Giannattasio, F., and De Michelis, P.: Forecasting SYM-H Index: A Comparison Between Long Short-Term Memory and Convolutional Neural Networks, Space Weather, 19, e2020SW002589, https://doi.org/10.1029/2020SW002589, 2021.

Siddique, T. and Mahmud, M. S.: Ensemble deep learning models for prediction and uncertainty quantification of ground magnetic perturbation, Frontiers in Astronomy and Space Sciences, 9, https://doi.org/10.3389/fspas.2022.1031407, 2022.

Sierra-Porta, D., Petro-Ramos, J., Ruiz-Morales, D., Herrera-Acevedo, D., García-Teheran, A., and Alvarado, M. T.: Machine learning models for predicting geomagnetic storms across five solar cycles using Dst index and heliospheric variables, Adv. Space Res., 74, 3483–3495, https://doi.org/10.1016/j.asr.2024.08.031, 2024.

Snoek, J., Larochelle, H., and Adams, R. P.: Practical bayesian optimization of machine learning algorithms, in: Proceedings of the 26th International Conference on Neural Information Processing Systems – Volume 2 (NIPS'12), Vol. 2, Curran Associates Inc., Red Hook, NY, USA, 2951–2959, https://papers.nips.cc/paper_files/paper/2012 (last access: 6 November 2026), 2012.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, J. Mach. Learn. Res., 15, 1929–1958, http://jmlr.org/papers/v15/srivastava14a.html (last access: 6 November 2026), 2014.

Thomson, D.: Spectrum estimation and harmonic analysis, P. IEEE, 70, 1055–1096, https://doi.org/10.1109/PROC.1982.12433, 1982.

Tóth, G., Meng, X., Gombosi, T. I., and Rastätter, L.: Predicting the time derivative of local magnetic perturbations, J. Geophys. Res.-Space, 119, 310–321, https://doi.org/10.1002/2013JA019456, 2014.

Viljanen, A., Nevanlinna, H., Pajunpää, K., and Pulkkinen, A.: Time derivative of the horizontal geomagnetic field as an activity indicator, Ann. Geophys., 19, 1107–1118, https://doi.org/10.5194/angeo-19-1107-2001, 2001.

Wang, T., Luo, B., Wang, J., Ao, X., Shi, L., Zhong, Q., and Liu, S.: Forecasting of the Geomagnetic Activity for the Next 3 Days Utilizing Neural Networks Based on Parameters Related to Large-Scale Structures of the Solar Corona, Space Weather, 23, https://doi.org/10.1029/2024SW004090, 2025.

Wang, Z., Wu, S., Huang, Y., Liu, R., and Liu, X.: A comprehensive review of deep learning for solar nowcasting: Enhancing accuracy, reliability, and interpretability, Appl. Energ., 407, 127378, https://doi.org/10.1016/j.apenergy.2026.127378, 2026.

Wei, L., Zhong, Q., Lin, R., Wang, J., Liu, S., and Cao, Y.: Quantitative Prediction of High-Energy Electron Integral Flux at Geostationary Orbit Based on Deep Learning, Space Weather, 16, 903–916, https://doi.org/10.1029/2018SW001829, 2018.

Welling, D. T., Ngwira, C. M., Opgenoorth, H., Haiducek, J. D., Savani, N. P., Morley, S. K., Cid, C., Weigel, R., Weygand, J. M., Woodroffe, J. R., Singer, H., Rosenqvist, L., and Liemohn, M.: Recommendations for Next-Generation Ground Magnetic Perturbation Validation, Space Weather, 16, 1912–1920, https://doi.org/10.1029/2018SW002064, 2018.

Wintoft, P. and Wik, M.: Exploring Three Recurrent Neural Network Architectures for Geomagnetic Predictions, Frontiers in Astronomy and Space Sciences, 8, https://doi.org/10.3389/fspas.2021.664483, 2021.

Zewdie, G. K., Valladares, C., Cohen, M. B., Lary, D. J., Ramani, D., and Tsidu, G. M.: Data-Driven Forecasting of Low-Latitude Ionospheric Total Electron Content Using the Random Forest and LSTM Machine Learning Methods, Space Weather, 19, e2020SW002639, https://doi.org/10.1029/2020SW002639, 2021.