Extended lead-time geomagnetic storm forecasting with solar wind ensembles and machine learning

Billcliff, M. ORCID: https://orcid.org/0009-0007-2960-9388, Smith, A. W. ORCID: https://orcid.org/0000-0001-7321-4331, Owens, M. ORCID: https://orcid.org/0000-0003-2061-2453, Woo, W. L., Barnard, L. ORCID: https://orcid.org/0000-0001-9876-4612, Edward-Inatimi, N. ORCID: https://orcid.org/0009-0001-6211-5781 and Rae, I. J. ORCID: https://orcid.org/0000-0002-2637-4786 (2026) Extended lead-time geomagnetic storm forecasting with solar wind ensembles and machine learning. Space Weather, 24 (3). e2025SW004823. ISSN 1542-7390 doi: 10.1029/2025SW004823 Available at https://centaur.reading.ac.uk/128858/

AGU ADVANCING EARTH AND SPACE SCIENCES

Space Weather®

RESEARCH ARTICLE
10.1029/2025SW004823

Extended Lead-Time Geomagnetic Storm Forecasting With Solar Wind Ensembles and Machine Learning

M. Billcliff[1], A. W. Smith[1], M. Owens[2], W. L. Woo[1], L. Barnard[2], N. Edward-Inatimi[2], and I. J. Rae[1]

[1]Department of Mathematics, Physics and Electrical Engineering, Northumbria University, Newcastle upon Tyne, UK, [2]Department of Meteorology, University of Reading, Reading, UK

Key Points:

- Our model outperforms standard baseline approaches up to 24 hr with improved discriminative skill and reliable probabilistic forecasts
- Machine learning classifiers process an ensemble of future realizations of ambient solar wind conditions, weighted by historical OMNI
- The framework could be operationalized with minimal modification by switching to near-real-time data inputs and retraining the model

Correspondence to:
M. Billcliff,
matthew.billcliff@northumbria.ac.uk

Citation:
Billcliff, M., Smith, A. W., Owens, M., Woo, W. L., Barnard, L., Edward-Inatimi, N., & Rae, I. J. (2026). Extended lead-time geomagnetic storm forecasting with solar wind ensembles and machine learning. *Space Weather*, *24*, e2025SW004823. https://doi.org/10.1029/2025SW004823

Received 12 NOV 2025
Accepted 23 FEB 2026

Author Contributions:
Investigation: M. Billcliff
Methodology: M. Billcliff, A. W. Smith, W. L. Woo
Software: M. Owens, L. Barnard
Supervision: A. W. Smith, W. L. Woo, I. J. Rae
Writing – original draft: M. Billcliff
Writing – review & editing: M. Billcliff, A. W. Smith, M. Owens, L. Barnard, N. Edward-Inatimi

Abstract Geomagnetic storms are large disruptions of the magnetosphere, which can impact satellites, communications systems, and power grids, causing significant technological and economic impacts. Current forecasting models utilize L1 satellite data, constraining lead time to a few hours, often insufficient for effective mitigation. We investigate how to extend the lead times of these forecasts with solar data. Associated spatial and propagation uncertainties of solar data are captured with a solar-wind ensemble, of the computationally efficient one-dimensional HUXt numerical model. The solar-wind ensemble once propagated to Earth is processed through logistic regressions, weighting ensemble members by comparison with historical observed velocities, effectively filtering out high error ensemble members. Performance was evaluated across different storm intensities and lead times, demonstrating the models predictive capabilities in a variety of circumstances. Although not including transient phenomena such as Coronal Mass Ejections, our approach demonstrates strong predictive capability, achieving a Brier Skill Score relative to climatology (BSS_{clim}) of 0.595 and a Receiver Operating Characteristic Area Under the Curve (ROC AUC) of 0.751 at 6-hr lead time for storms defined as $Hp30_{MAX} \geq 5$ within a 24-hr forecast window. Overall, these results highlight the strong potential of the coupled numerical model and machine learning framework to extend the forecast lead time for geomagnetic storms.

Plain Language Summary Geomagnetic storms can impact satellites, communication networks, and power grids, but current forecasts rely on near-Earth solar wind data, providing only 30–90 min of warning. Since the solar wind takes 1–3 days to travel from the Sun to Earth, we demonstrate a method to extend forecast lead times using solar data instead. By processing this data as an ensemble, we model a range of possible future solar wind conditions and combine it with machine learning to predict the probability of a geomagnetic storm occurring within a given 24-hr window, and offering significantly more warning than existing methods.

1. Introduction

Geomagnetic storms are frequently occurring disturbances in the Earth's magnetosphere, driven by solar wind interactions with the Earth's magnetic field (Gonzalez et al., 1994). Due to their utility in quantifying geomagnetic activity and geomagnetic storms, many studies over the last decade have developed methods to forecast indices such as Kp (Shprits et al., 2019; Wing et al., 2005; Wintoft et al., 2017), Dst (Gruet et al., 2018), and SymH (Collado-Villaverde et al., 2023; Conde et al., 2023; Siciliano et al., 2021) using data from satellites located near the Lagrangian L1 gravitational equilibrium. These satellites remain in approximately the same position relative to Earth and the Sun, allowing us to use instruments that continuously monitor the solar wind between the Sun and Earth. Geomagnetic storms are triggered by high-speed solar wind streams from coronal holes on the Sun (Cranmer, 2009; Nitti et al., 2023; Richardson, 2018), as well as large eruptions of high energy plasma from the solar surface called Coronal Mass Ejections (CMEs) (Dumbović et al., 2015; Kilpua et al., 2019). Sufficiently intense geomagnetic storms can have impacts on infrastructure, including electrical grid disruptions caused by Geomagnetically Induced Currents (Bolduc, 2002; Boteler, 2003; Marshall et al., 2012), radiation damage to near-Earth satellites (Baker, 2001; Hands et al., 2018), and railway signaling misoperations (Patterson et al., 2024). Aurora are also a result of geomagnetic storms, producing magnificent bright light displays at Earth's magnetic poles.

Due to data availability, current forecasting models (Chakraborty & Morley, 2020; Tan et al., 2018) use data from satellites stationed near L1 to make forecasts. The time between solar wind arriving at L1 satellites and arriving at Earth is 30–90 min, based on solar wind speeds of 800 km/s to 300 km/s respectively with the travel time of solar wind from the Sun to Earth ranging from 1 to 3 days. As a result, the lead time potential of forecasting models

using only data collected at L1 is significantly less than that of models utilizing solar data. Longer lead times can be achieved from model-derived Carrington maps of the solar wind at 21.5 solar radii ($R_\odot$), produced from magnetograms using the Magnetohydrodynamic Algorithm outside a Sphere (MAS) model (Riley et al., 2001). At 21.5$R_\odot$ (0.1AU) the solar wind is both super-Alfvénic and supersonic so information flows only away from the Sun which simplifies the corona-solar wind coupling (M. Owens et al., 2020). Bernoux et al. (2022) moves in the direction of using solar data, forecasting daily maximum of geomagnetic activity between 2 and 7 days in advance using data from solar EUV imagers.

Planetary geomagnetic indices quantify the strength of storms, given as single value which describes the geomagnetic disturbances across the Earth. The Kp (Bartels, 1949) index is calculated using 13 sub-auroral magnetometers across Earth, by measuring the maximum horizontal deviation of the magnetic field with a 3 hr period, scaling these deviations to a quasi-logarithmic scale specific to each station, normalizing for geographic differences, and averaging the results to produce a global measure of geomagnetic activity. Kp takes values between 0 and 9 with 0 representing no activity, and 9 representing an extreme event which only occurs a few times every solar cycle. In this study, storm periods are defined using thresholds of the Hp30 index (Yamazaki et al., 2022), an open-ended (i.e., not capped at 9), higher resolution (30-min cadence) adaptation of the Kp index. The target variable is defined as $\mathrm{Hp30}_{\mathrm{MAX}} \geq 4.66$, that is, exceeding an Hp30 index threshold of 4.66 within a 24-hr forecast window.

The aims of this work on operational forecasting are to provide a longer lead time forecast, providing more time to perform mitigating actions for damage to infrastructure (Oughton et al., 2019; Schrijver, 2015). Forecasting the Hp30 index (Yamazaki et al., 2022) allows us to better distinguish between storms of very high intensity compared to the Kp index (Bartels, 1949). Instances of damage in the past have been observed and analyzed in a variety of high risk countries: New Zealand (Marshall et al., 2012); South Africa (Gaunt & Coetzee, 2007); Sweden (Pulkkinen et al., 2005); Canada (Bolduc, 2002; Boteler, 2003). These countries have been vulnerable in the past, likely as a combination of latitude (Thomson et al., 2011), geology and power network configuration. Oughton et al. (2019) calculates that for the United Kingdom, total loss in gross domestic product caused by a large Carrington size storm without mitigation would be around £15.9 Billion, dropping to £2.9 Billion under current forecasting capabilities. In optimal forecasting scenarios, we could bring this figure to as low as £900 Million.

Extending lead time using solar wind forecasts introduces both spatial uncertainties in the solar wind mapping and propagation uncertainties in the numerical modeling. When data with large uncertainties are used as input to a deterministic model, it could produce inaccurate results, since a model can only be as good as the quality of its inputs. An effective way to capture spatial uncertainties in the solar wind is through the use of an ensemble, to provide an array of solar wind inputs and initialize one numerical model for each input. The advantages of an ensemble are that given a single model realization of solar wind conditions near the Sun, we can generate many possible realizations of the evolution of solar wind at Earth, which in principle could capture all possible variability in the ambient solar wind for a better understanding of the dynamic solar wind environment and enhancing forecasts of geomagnetic storms. Computationally intensive 3D magnetohydrodynamic (MHD) models are too slow to run large ensembles, so we turn to one-dimensional models, which offer a more efficient approach by reducing the complex physics of solar wind propagation, enabling faster simulations and larger ensemble sizes compared to 3D MHD models. Examples of such one-dimensional models include the Heliospheric Upwind eXtrapolation with time dependency (HUXt) model (M. Owens et al., 2020; Barnard & Owens, 2022), or a ballistic solar wind propagation model (Issan & Kramer, 2023). The limitations of the these types of model are that they often don't output magnetic field components of the interplanetary magnetic field (IMF), density of the solar wind, nor solar wind dynamic pressure, all of which have been shown to be important parameters for forecasting geomagnetic storms (A. W. Smith et al., 2020) and the consequences of geomagnetic storms (Ma et al., 2024; A. Smith et al., 2024).

This study aims to extend the lead time of geomagnetic storm forecasting. We develop ensembles of computationally fast reduced-physics numerical modeling with machine learning to produce probabilistic forecasts for storm occurrences within a 24 hr window. In Section 2 we discuss the types of data used, with Section 3 describing the two numerical models we leverage, and how the outputs of these are processed with machine learning techniques. This includes how we create ensembles, how the data is prepared for input to numerical models, and the setup for making a forecast. Section 4 explains the performance metrics we use to evaluate our models predictive skill. In Section 5, we present an analysis of model performance and reliability, and the effect of

lead time and storm strength on performance metrics. Section 6 presents a detailed discussion on the implications of the model, where this study fits in the bigger picture of storm forecasting, and future work.

2. Data

This section presents the open source data used in this study. Section 2.1 discusses the choice of index, and the subsequent storm definition used throughout, with Section 2.2 giving an overview of the OMNI data set, and its use case for this study.

2.1. Planetary Geomagnetic Indices

We considered three different planetary geomagnetic indices to characterize geomagnetic storm times: Kp (Bartels, 1949), and Hpo (both Hp60 and Hp30) (Yamazaki et al., 2022). These indices provide an overall measure of the intensity of geomagnetic activity at Earth, given as a single value to represent intensity over a period of time determined by the index. The two main considerations for choosing an index to forecast are cadence, and upper limit. The cadences for Kp, Hp60, Hp30 are 3 hr, 1 hr, and 30 min respectively. The Kp index is also artificially capped at a value of 9. The distributions of the Hpo indices are consistent with the distribution of the Kp index, but are open-ended, that is, they don't have an artificial upper limit. The Kp index, limited at 9, does not differentiate between geomagnetic storms where Kp = 9, such as the 2003 Halloween Storms (Gopalswamy et al., 2005) and the 1859 Carrington Storm (Cliver & Svalgaard, 2004), which differ significantly in intensity. We therefore consider that the open-ended Hpo indices offer a more useful representation of geomagnetic activity.

Between Hp30 and Hp60 the choice is largely arbitrary given that we are classifying 24-hr windows which are long relative to the cadence of the indices. After evaluation, we chose to forecast Hp30 due to its higher cadence. Hp30 should capture the necessary structures and aligns closely with the objectives of our study. Furthermore, Hp30 provides more data points within the same period of time compared to Kp, allowing observation of large, short-term disturbances not evident in Kp data. We provide a comparison of Hp30 and Kp in Figure 1, which shows one of the 2003 Halloween storms (Pulkkinen et al., 2005): an extreme CME driven geomagnetic storm, in which the Kp value reached the maximum value of 9.

Observing the differences between Hp30 and Kp, it is evident that the Hp30 index is capable of recognizing severe storms beyond that of the Kp indices, indicated by the large differences between 2003 and 10–30 19:00 and 22:00 where the Hp30 index is more reflective of the extremely high geomagnetic activity, reaching a peak value of 11.66, as compared to the capped Kp of 9. Hpo data is available in near real-time (from https://www.gfz-potsdam.de/en/hpo-index). The Hpo index is the limiting factor for data collection, available from 1995 to 01-01 to present, providing us with approximately 30 years of data. With the Hp30 index, we define a geomagnetic storm using the widely regarded threshold of $Hp30_{MAX} \geq 4.66$ (or 5^{-}), where $Hp30_{MAX}$ represents the maximum value of Hp30 observed within a given period. This is in alignment with the National Oceanic and Atmospheric Administration (NOAA) Geomagnetic Storm (G Scale) which relates the Kp index to the severity of the storm. The G Scale ranges from G1 (minor: Kp = 5) to G5 (extreme: Kp = 9).

2.2. OMNI

The OMNI data set is a processed collection of near-Earth solar wind properties, IMF data, and geomagnetic indices (Papitashvili & King, 2020). In this study, we used the low-resolution (hourly) solar wind velocity data from the OMNI data set, which is available from 1963 to the present (https://omniweb.gsfc.nasa.gov/html/ow_data.html); however, we specifically use data starting from 1995 onwards due to availability of Hpo data. This hourly data represents averaged values derived from higher resolution (1-min) data, rather than hourly measurements.

The data set is compiled from measurements by satellites in geocentric and L1 orbits, including missions such as ISEE 3, Wind, and ACE. To ensure consistency across different sources, the data has been extensively cross-compared and cross-normalized. Further processing includes time-shifting of higher resolution data for spacecraft at L1 to accurately reflect arrival times of solar wind streams.

In our study, OMNI data is not used directly as model input. Instead, we use it for comparison with a solar wind velocity ensemble from the HUXt model (M. Owens et al., 2020; Barnard & Owens, 2022). This comparison is incorporated both as a weighting method for the output of our machine learning models, and as a derived feature

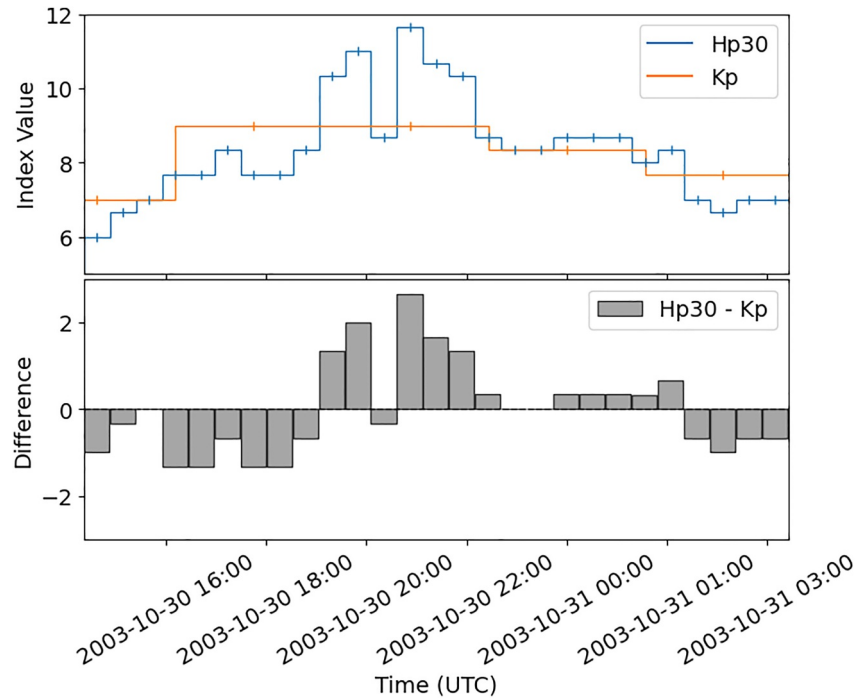

Figure 1. An example of the differences between Hp30 and Kp indices over a 13 hr period during the 2003 Halloween Solar Storms. Top panel: Observed values for Hp30 (blue) and Kp (orange) indices over a 13 hr period. Bottom panel: Difference between Hp30 and Kp indices.

defined as the difference between a solar wind ensemble (v) and OMNI, which we denote v − OMNI. Where data gaps are present in OMNI, we apply linear interpolation to ensure consistency in v − OMNI values.

We note that solar wind speed from the OMNI data set is a measurement of both ambient solar wind and transients (i.e., CMEs). For the purposes of this study, we do not filter out transient structures, leading to discrepancies between the forecasted ambient solar wind, and OMNI when transients are present. As CMEs drive many of the strongest geomagnetic disturbances, their exclusion reduces forecast performance during such events. Incorporating CME ensemble forecasts will be explored in future work.

3. Models

In order to make forecasts from Sun to Earth, we utilize a pipeline of models, with the aim of combining solar wind forecasting with geomagnetic storm forecasting. For this study, we combine 3 models, described throughout this section. A schematic of these models with their order in the model pipeline is illustrated in Figure 2. This schematic shows the flow of information from model A–MAS (Section 3.1), to model B - HUXt (Section 3.2) through the use of extracting an ensemble from the output of MAS. It then shows that separate classifiers are trained for each member of the HUXt ensemble, as information is passed from HUXt to model C(i)–an ensemble of logistic regression classifiers (Section 3.3.1.) Then model C(i) is aggregated into a final probabilistic forecast through model C(ii)–weighted mean (Section 3.3.2).

3.1. Model A–MAS

Near-Sun data can be used to extend lead times due to the travel time of solar wind. Typically, we observe travel times of the solar wind from 21.5 $R_\odot$ to Earth to be between 1 and 3 days. This study uses Carrington maps of solar wind evolution at 21.5 $R_\odot$ as boundary conditions for a numerical model which will model the propagation of solar wind properties from Sun to Earth. Typically, a Carrington map is constructed by solar images during one Carrington Rotation, which are stitched together to form a map of the entire solar surface. This concept allows us to represent the evolution of the Sun over the period of one Carrington rotation in a single two-dimensional plot. The Carrington map we used in this study is produced from magnetograms which cover the full Carrington

10.1029/2025SW004823

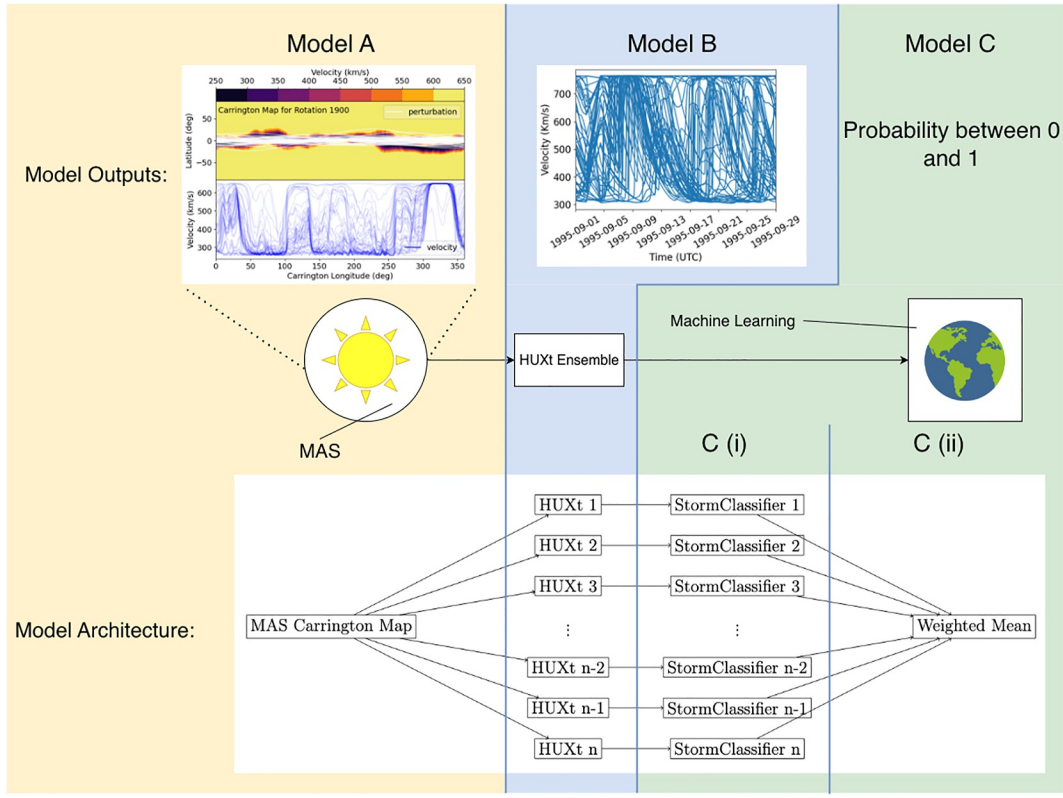

Figure 2. Schematic of the geomagnetic storm forecasting process, illustrating the propagation from the Sun to Earth through a numerical model. Model (a) MAS. The plot in this segment is the output of MAS at 21.5 $R_\odot$ for Carrington rotation 1900. Model (b) Heliospheric Upwind eXtrapolation with Time dependency (HUXt), initialized with boundary conditions from Model (a) The plot in this segment shows the output of HUXt at 1AU during the period of Carrington rotation 1,900. Model (c) Machine Learning models, split into two parts–C(i): Storm classifiers trained on Model B output and a target variable of $Hp30_{MAX} \geq$ storm threshold. C(ii): Returns a weighted average of the outputs from C(i) giving our final probabilistic forecast. The flow chart shows the architecture of the full forecasting model for n ensemble members.

rotation, and outputs a map of the nascent solar wind at 21.5$R_\odot$ for that same Carrington rotation, calculated as the solution to the 3D MHD equations.

For this work, we use numerical solutions to the MHD equations from the MAS model (Riley et al., 2001). MAS models the Solar Corona (1$R_\odot$ to 21.5$R_\odot$) and the inner Heliosphere (21.5$R_\odot$ to 5 AU). In this study, we use only the simulated Solar Corona to obtain the output at 21.5$R_\odot$. The Solar Coronal modeling of MAS corresponds to Model A in Figure 2 and represents a 3D MHD solution using observed photospheric magnetic fields as the inner boundary condition. The magnetic field data are sourced from various observatories to ensure full temporal coverage, consistent with the data set used by M. J. Owens et al. (2022). This setup allows MAS to simulate the corona and inner heliosphere up to 21.5$R_\odot$, producing a Carrington map of solar wind velocity at this radius.

It should be noted that using MAS here is chosen due to availability of data for the required periods, but this choice contributes to the uncertainties in the solar wind ensemble. Different results could be achieved if we were to use another method for estimating solar wind conditions at 21.5 $R_\odot$, such as the Wang-Sheeley-Arge (WSA) model (Arge & Pizzo, 2000).

3.1.1. Ensemble Extraction From a Carrington Map

An illustration of the MAS output and the solar wind ensemble extraction is shown in Figure 3. For visual clarity, we show a representative subset of 50 ensemble members in this figure. The full analysis throughout the paper will use 100 ensemble members. The top panel shows the Carrington map for Carrington rotation 1,900 (1995-09-02 to 1995-09-30), while the bottom panel shows the extracted velocities, which should be read from east to west (right to left on the map) following the Sun's rotation relative to Earth.

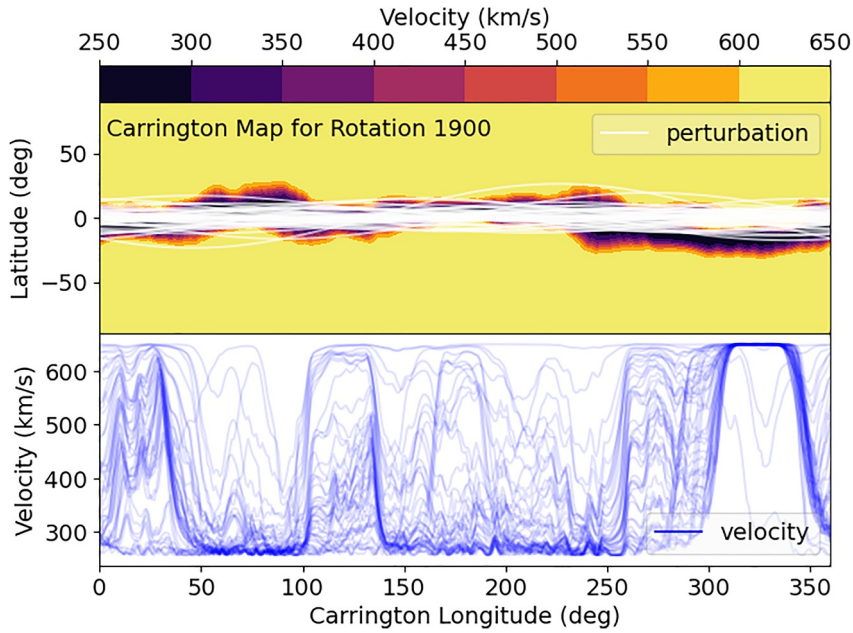

Figure 3. MAS output for Carrington rotation 1,900 (1995-09-02 to 1995-09-29). Top panel: Carrington map of solar wind velocity (km/s) with latitude (deg) on the *y*-axis and Carrington longitude (deg) on the *x*-axis. White lines indicate 50 sinusoidal perturbations on Earth's latitude relative to the Sun's equator. The color bar indicates the velocities of the color map. Bottom panel: presents the corresponding velocity profiles (km/s) for the extracted ensemble members.

As the Sun rotates, the Earth's orbital path traces an approximate straight line on the Carrington map. To account for the uncertainty in the location of solar wind speed features on the map, and their later propagation, we perturb this background map, which when rotated in latitude makes Earth's path appear as a sinusoid on the map. This allows us to sample different profiles from the map and better estimate the range of velocity profiles. To ensure our model captures the essential structures, we extract velocities from a large number ($\geq$50) of perturbations of the Carrington map, an approach used by M. J. Owens and Riley (2017). The shapes of the sinusoidal perturbations are determined by two constants, which are used to specify the distributions from which we generate the parameters for each perturbation. We create these perturbations in the same manner as Edward-Inatimi et al. (2024). In short, we create perturbations according to:

$$\theta(\phi) = \theta_E + \theta_{MAX} \sin\left(\phi + \phi_0\right) \tag{1}$$

where θ is heliolatitude of the perturbed path, ϕ is Carrington longitude, θ_E is the unperturbed Earth heliolatitude, θ_{MAX} is the maximum perturbation, and ϕ_0 defines the phase. Values of θ_{MAX} are drawn from a normal distribution with mean of Earth's path over this Carrington rotation and $\sigma = 7.5°$, sufficient to cover the spatial uncertainties on the map. ϕ_0 are selected from a uniform distribution in $[0, 2\pi]$.

In general, we observe slower solar wind velocities around the Sun's equator compared to higher or lower latitudes, however, high solar wind speeds can still be observed due to coronal holes, leading to the presence of high-speed solar wind streams. From examining Figure 3, it is clear that variations in the sinusoidal path of the Earth across the Sun can significantly impact the solar wind velocity.

In this study, we use many such profiles to create an ensemble of inner boundary conditions for the Heliospheric Upwind Extrapolation with time dependency (HUXt) model (M. Owens et al., 2020; Barnard & Owens, 2022) described in Section 3.2. Variations in the velocity profiles will affect the models input, and hence it's output. The variability we observed in Figure 3 is one reason we employ a large number of ensembles in our model, to capture the possible structures present in the solar wind accurately. In this way, each velocity profile provides an estimate of conditions near Earth.

3.2. Model B–HUXt

The Heliospheric Upwind Extrapolation with time dependency (HUXt) model (M. Owens et al., 2020; Barnard & Owens, 2022) corresponds to Model B in Figure 2. HUXt is an open-source Python package that solves the time-dependent 1D incompressible hydrodynamic equations to model the propagation of the solar wind. It is typically initialized using boundary conditions at a fixed heliocentric distance; for this study, we use $21.5R_{\odot}$. From these boundary conditions, the model extrapolates the solar wind flow speed to near-Earth distances. This method enables us to quantify the uncertainty in the forecasted ambient solar wind which results from positional uncertainties in the MAS estimation of near-Sun solar wind. This approach allows us to provide much more data to a machine learning model than is usually available. Specifically, we utilize this method to "see the future," where we can use estimations of the future ambient solar wind conditions to forecast storms.

HUXt also has functionality for simulating the interaction between CMEs and the ambient solar wind, but we don't use this in this study. Unlike full 3D MHD models (Mayank et al., 2025; Odstrčil, 1999; Odstrčil & Pizzo, 1999; Pizzo et al., 2011; Pomoell & Poedts, 2018), HUXt uses a reduced-physics approach. The main advantage of HUXt is the rapid computation times allowing us to run simulations for large ensembles - for example, running 100 ensemble simulations over a full solar cycle (excluding CMEs) took only 9 min on a laptop with an Apple M2 chip (8 cores), and 16 GB RAM. Disadvantages include the lack of magnetic field modeling and the assumption of radial flow, which limits the model's ability to capture more complex heliospheric dynamics. HUXt is publicly available as an open-source Python package (M. J. Owens & Barnard, 2024), making it accessible and easily integrable. In this study we do not initialize the model with CMEs, and simulate only ensembles of the ambient solar wind.

3.3. Model C–Storm Classifiers

The storm forecasting component of our framework is implemented through a series of machine learning classifiers that process the outputs of the numerical ensemble independently. These classifiers, denoted as C(i) in our schematic (Figure 2), function in parallel, each corresponding to a unique member of the HUXt ensemble, resulting in a set of probabilistic predictions. The individual outputs of these classifiers are then aggregated as a weighted mean (model C(ii)) using output of C(i) producing a final probabilistic forecast. This approach allows us to use the unique insights of each ensemble member while combining their strengths to enhance the overall predictions.

3.3.1. Model C(i)–Logistic Regression

Logistic regression was chosen for its simplicity and its ability to perform well on small data sets, like the one used in this study, which includes 2345 storm times and 2345 non-storm times (after dropout–discussed in Section 3.4). Logistic regression is a statistical model used for binary classification problems, where the goal is to estimate the probability of an event (e.g., a geomagnetic storm) based on input variables. The logistic function used in regression is given by:

$$P(Y=1|X)=\frac{1}{1+e^{-\left(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n\right)}} \tag{2}$$

Here, $P(Y=1|X)$ represents the probability of the event occurring, β_0 is the intercept, $\beta_1,\beta_2,\ldots,\beta_n$ are the coefficients for the predictor variables $X_1,X_2,\ldots,X_n$. The coefficients for the logistic regression are calculated using the default solver from the sci-kit learn Python package: Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS). This probabilistic output helps in classifying storm events with a clear decision boundary, making logistic regression a strong choice of model.

Model C(i) consists of an ensemble of logistic regression classifiers, each corresponding to a member of the HUXt numerical model ensemble. Each classifier uses the following time series input features: (a) predicted ambient solar wind velocity (v), given from the start of the input window to 48 hr after the time the forecast is made, (b) the gradient of v (calculated using finite differences) over the same interval as v, (c) the historical difference between predicted v and OMNI data (v − OMNI), at all time steps within the input window, and (d) Hp30 data over the input window. The output of each classifier is a probabilistic value between 0 and 1, representing the likelihood of a geomagnetic storm occurring based on these inputs.

Table 1
Availability of Input Data Relative to a Forecast Time, T_0. $T_0 - 24$ h Defines the Input Window, While $T_0 + 48$ h Represents the Period Following the Forecast Time

Variable	Input window ($T_0 - 24$ h; T_0)	Post input window (T_0; $T_0 + 48$ h)
HUXt (v)	×	×
HUXt (Δv)	×	×
Data (v − OMNI)	×	
Data (Hp30)	×	

3.3.2. Model C(ii)–Weighted Mean

The role of Model C(ii) is to process the forecasts made by individual ensemble members into a single probabilistic forecast by aggregating their outputs. To enhance model performance, we investigated how to use historical hourly OMNI observations to provide insight for potential ensemble member performance. Data gaps in OMNI were ignored for calculating Mean Absolute Error (MAE). We evaluated several aggregation strategies, including logistic regression on Model C(i) outputs and filtering weak ensemble members, and found that a weighted average provided the most consistent performance across metrics. The weights for each output from Model C(i) are proportional to the MAE of historical OMNI and the corresponding HUXt solar wind velocity profile. The equation for calculating the weighted average from the probabilities and MAEs is as follows:

$$w_i = \frac{1}{MAE_i^2} \tag{3}$$

$$w_i^{\text{norm}} = \frac{w_i}{\sum_j w_j} \tag{4}$$

$$\hat{y} = \sum_i w_i^{\text{norm}} p_i \tag{5}$$

where p is the array of probabilities from Model $C(i)$, w_i is the weight for each model based on its MAE (MAE_i), and $\hat{y}$ is the final weighted probabilistic prediction.

3.4. Forecast Setup

The goal of this project is to forecast, as a probability, whether the Hp30 index will exceed a specified storm threshold within a forecast window. We define the input window as the period preceding the time the forecast is made (T_0), fixed at 24 hr in length. All input parameters are available during the input window. The forecast window starts at T_0 + lead time (Lt) and is also fixed at 24 hr in length. Importantly, predicted solar wind velocities from HUXt are used both during the input window and 48 hr after the input window, which covers the forecast window and slightly beyond (for $Lt \leq 24$ hours. When $Lt \leq 24$ hours, we increase the length of the solar wind ensemble to fully cover the forecast window. This setup ensures a clear separation between the input data and the forecast period. A stride (gap between successive forecast windows) of 24 hr is chosen to maximize the number of forecast windows seen by the model, without overlap. This information is displayed in Table 1, to help with understanding of data availability, and to highlight the nature of having a forecasted solar wind during the forecast window.

Each period is labeled as either "Storm" if $\text{Hp30}_{\text{MAX}} \geq 4.66$ during the 24 hr forecast window or "Non-Storm," if $\text{Hp30}_{\text{MAX}} < 4.66$. 4.66 is chosen in alignment with a G1 geomagnetic storm on the NOAA G-scale (https://www.swpc.noaa.gov/noaa-scales-explanation). Additionally, we ensure that the storm to non-storm ratio remains consistent across training and testing sets. The data set used in this study exhibits a significant class imbalance, with a much higher prevalence of non-storm periods compared to storm periods. Specifically, of the 8739 windows between 1995-01-01 and 2024-01-18, only 2345 contain storms (27%). If we train models on the full imbalanced data set, many metrics such as accuracy, and true negative rate (specificity), could be artificially high. To counteract this, we dropout data such that we balance the number of storm and non-storm windows giving a better representation of the models ability to identify storms.

For machine learning purposes, we split the data using the train:test ratio of 80:20. To split the data, we use a systematic sampling approach based on Carrington rotations, rather than a random split. This was done to ensure that both train and test sets for machine learning are representative of different phases of the solar cycle, and to avoid the mixing of data within a given Carrington rotation. Additionally, this method ensures consistency in testing sets when performing cross validation. Chronologically, we repeat the pattern of 1 Carrington rotation worth of data in the test set, then 4 Carrington rotations worth of data in the training set, resulting in approximately

20% of the data in the test set and 80% of data in the training set. Given the length of the forecast windows, a slight gap of around 6 hr is introduced between successive Carrington rotations, allowing for extra separation between the test and training data sets.

We also assess the variability in model performance by performing cross validation by shifting the test fold, providing 5 folds for calculating uncertainty estimates for the evaluation metrics across different train–test partitions. Additionally, we test over 5 random seeds (1, 42, 100, 12345, 151201) which determine the random dropout of non-storms, as well as the perturbations which generate ensembles on the Carrington map. Combined, we run the model 25 times (5×5) corresponding to each combination of test fold and random seed in order to assess the variability of performance metrics and verify performance.

4. Metrics

The metrics used in this study include both the probabilistic metric of Area Under the Receiver Operating Characteristic Curve (ROC AUC Score) and Brier Skill Score (BSS), to compare our model to climatology. These metrics are frequently paired together in Space Weather related studies (A. Smith et al., 2021; Forsyth et al., 2020; Leka et al., 2019) since by combining these metrics, we provide a comprehensive evaluation of the models, capturing both their discriminative skill (ROC AUC Score) and their probabilistic accuracy and reliability ($\mathrm{BSS}_{\mathrm{clim}}$). Given our balanced data set, the climatology forecast is simply the probability 0.5 for all forecast windows. A contingency table categorizes the predictions into four groups: True Positives (TP), where the model correctly identifies a positive event; False Positives (FP), where the model incorrectly predicts a positive event; True Negatives (TN), where the model correctly identifies a negative event; and False Negatives (FN), where the model fails to identify a positive event.

ROC AUC measures the model's ability to separate classes (discriminative skill) and is calculated by varying the decision threshold, plotting the True Positive Rate (TPR) against the False Positive Rate (FPR), and computing the area under the resulting curve. TPR and FPR are given by:

$$\mathrm{TPR} = \frac{TP}{TP + FN}, \quad \mathrm{FPR} = \frac{FP}{FP + TN} \tag{6}$$

A ROC AUC Score >0.5 indicates performance better than random guessing, and a score of 1 indicating perfect discriminative skill.

Brier Skill Score (BSS) measures the comparative skill of a model against a reference. In this work, the reference is a climatology forecast, which on our balanced test set corresponds to a constant probability of 0.5. We therefore refer to this as the BSS (Climatology), denoted as $\mathrm{BSS}_{\mathrm{clim}}$. To calculate $\mathrm{BSS}_{\mathrm{clim}}$, we first compute the Brier Score (BS):

$$BS = \frac{1}{N}\sum_{i=1}^{N}\left(p_i - o_i\right)^2 \tag{7}$$

where N is the number of forecasts, p_i is the forecast probability, and $o_i \in 0, 1$ is the observed outcome. The BSS relative to climatology is then given by:

$$\mathrm{BSS}_{\mathrm{clim}} = 1 - \frac{BS}{BS_{\mathrm{clim}}} \tag{8}$$

where BS is the BS of the model under evaluation, and BS_{clim} is the BS of the climatology forecast. Positive values of $\mathrm{BSS}_{\mathrm{clim}}$ indicate improvement over climatology, and for this work we consider scores of 0.2 or above to indicate clear added value over the comparative model. This value is often considered to show meaningful reliability, as seen in Wilks (2011) and (A. Smith et al., 2021).

5. Results

The following section contains model performances over a variation of forecasting conditions, and events. We present a comparison of the model proposed in this paper (listed as "weighted mean") against baseline models (Section 5.1), and investigate how the model performs, as a look into storm strengths and lead times (Section 5.2).

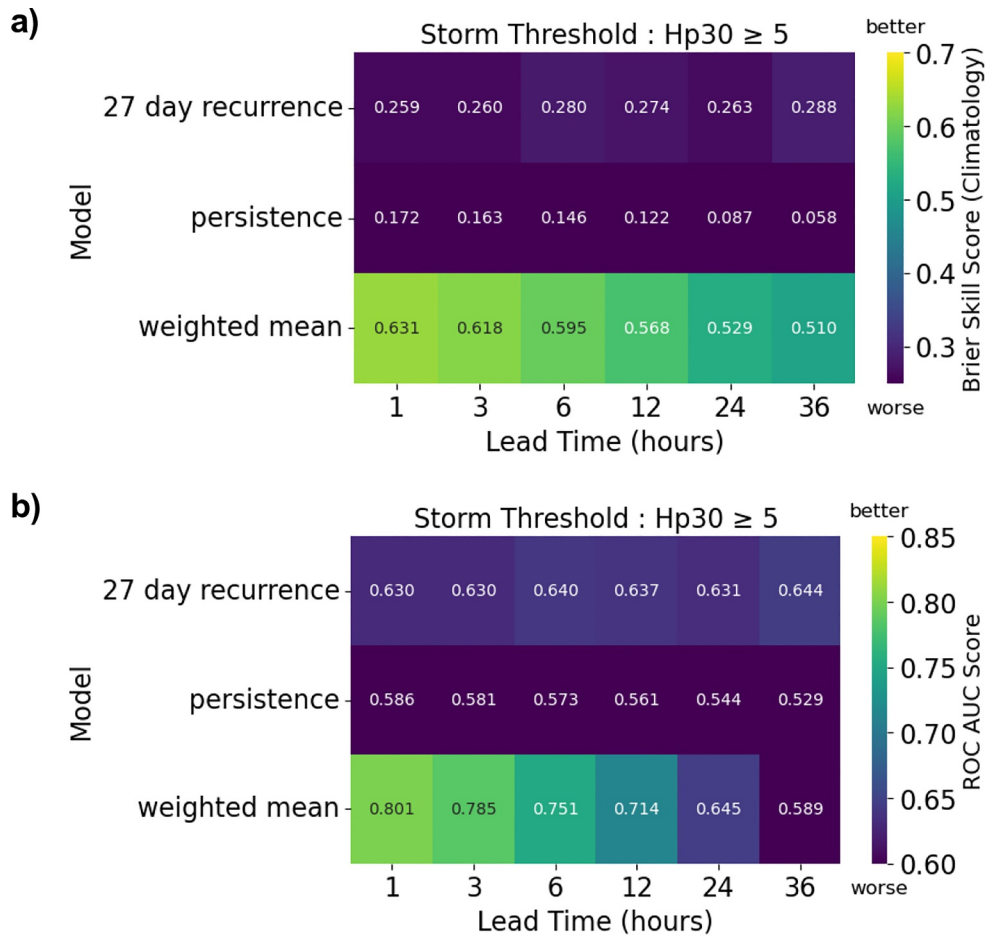

Figure 4. Heat maps of median model performance metrics at different forecast lead times, with Hp30 storm threshold fixed at $Hp30_{MAX} \geq 5$. Panel (a) shows Brier Skill Score (Climatology), and panel (b) shows ROC AUC Score. The *x*-axis represents lead times ranging from 1 to 36 hr, while the *y*-axis represents the model used. The color map highlights better scores in light colors, and worse scores in dark colors, indicated by the color bar to the right of each plot.

Beyond BSS_{clim}, we assess the models reliability, and look at the calibration of our forecasts in Section 5.3. All models are trained with 100 ensemble members, which proved to be a sufficient number to capture the variability in the ambient solar wind with more ensemble members showing little to no improvement.

5.1. Comparison to Baseline Models

The two baseline models we compare against are persistence, and 27-day recurrence. Persistence assumes current conditions will continue unchanged. The prediction for persistence is whether Hp30 exceeded the storm threshold in the time step immediately preceding the time the forecast is made (T_0). 27-day recurrence utilizes the repeating patterns in the solar wind every roughly 27.28 days (1 solar rotation period). Calculating autocorrelation of the Hp30 index at lags close to 1 solar rotation period, we observe a local maximum at a lag of 27.00 days, with Pearson correlation coefficient of 0.25, shown in Appendix Figure A1. Therefore, the forecast for 27-day recurrence is whether Hp30 exceeded the storm threshold in the equivalent period 27.00 days prior. This is similar to M. J. Owens et al. (2013) which uses the same principle to forecast the solar wind.

Figure 4 shows heat maps of median performance for BSS_{clim} (Figure 4a) and ROC AUC (Figure 4b) for our weighted mean model and the baseline approaches. The values shown are median metrics from the 25 runs described in Section 3.4. The color of each square indicates the value of the metric, with light colors being a better score and dark being worse, indicated by color bars on the right-hand side of each plot.

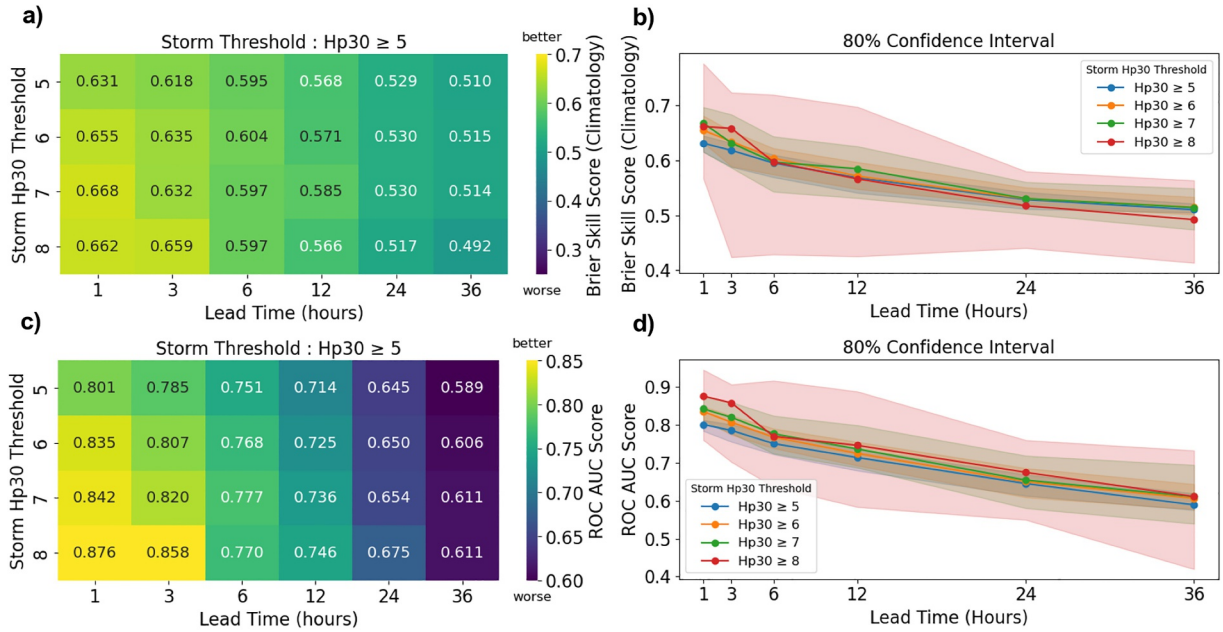

Figure 5. Metrics for our model at a variety of lead times and storm strengths. Panels (a), (b) show Brier Skill Score (Climatology), and (c), (d) show ROC AUC Score. Panels (a), (c) show heat maps of median model performance metrics as a function of forecast lead time and Hp30 storm threshold. The *x*-axis for all panels shows forecast lead time in hours. The *y*-axis represents various storm test thresholds. Color intensity indicates the score for each metric, with brighter representing better values, and darker representing worse values, indicated by the color bar to the right of each plot. Panels (b), (d) show the corresponding line plots of panels (a), (c), with shaded area representing the 80th percentile spread in performance across test folds and random seeds. The *y*-axis shows the corresponding metric values. Each colored line and associated shaded region corresponds to a different storm threshold $\left(\mathrm{Hp30}_{\mathrm{MAX}} \geq 5, 6, 7, 8\right)$.

Our model ("weighted mean") outperforms persistence at all lead times for both metrics. Looking at Figure 4b and comparing with the 27-day recurrence model, the ROC AUC Score is higher for our model up to a lead time of 24 hr where we only slightly outperform 27-day recurrence. Based on ROC AUC Score alone, we would conclude that 27-day recurrence is a better choice at long lead times (i.e., 36 hr), but $\mathrm{BSS}_{\mathrm{clim}}$ shows the benefit of providing a probability for more reliable risk estimates. The specific reliability of the probabilistic forecasts is discussed in Section 5.3. We note that 27-day recurrence is deterministic, and though it has decent skill at longer lead times, the reliability suffers since we have no indication of model confidence or uncertainty on the forecast.

5.2. Impact of Storm Strength and Lead Time on Model Performance

All models are trained on a storm threshold of $\mathrm{Hp30}_{\mathrm{MAX}} \geq 4.66$. Models are tested on incrementally increasing thresholds to assess performance on more intense geomagnetic storms. For evaluating performance, we limit the test set to only include geomagnetic storms of a specified intensity or greater. When we filter the test set, we only take storms where $\mathrm{Hp30}_{\mathrm{MAX}}$ in the forecast window exceeds the specified threshold, and randomly drop out non-storms to balance the storms and non-storms in the test set. We also vary the lead times between 1 and 36 hr Figure 5 shows median results and confidence intervals for all combinations of test folds and random seeds (discussed above), where the *x*-axis shows the varying lead times. For Figures 5a and 5c, the *y*-axis shows the varying storm thresholds. The colors of each box represent the metric, indicated by color bars on the right hand side, similar to Figure 4. For Figures 5b and 5d, the *y*-axis shows the metric for the corresponding lead time and geomagnetic storm threshold, with the shaded area around each line showing the 80% confidence interval (i.e., 10th and 90th percentiles).

In Figure 5, we observe strong performance, with $\mathrm{BSS}_{\mathrm{clim}}$ consistently high, well above the threshold of 0.2 at which we consider our model to have clear benefit beyond the comparison model - even at lead times of 36 hr. Both metrics show a clear decrease in performance as the lead time increases. The biggest jump in skill is between 3 and 6 hr, and flattening out at the longer lead times of 24 and 36 hr. We expect that as the lead time increases, the model becomes more reliant on the solar wind ensembles, rather than the current conditions of the magnetosphere provided by observed Hp30. For very short lead times (1–3 hr), both metrics are comparatively high, which is

10.1029/2025SW004823

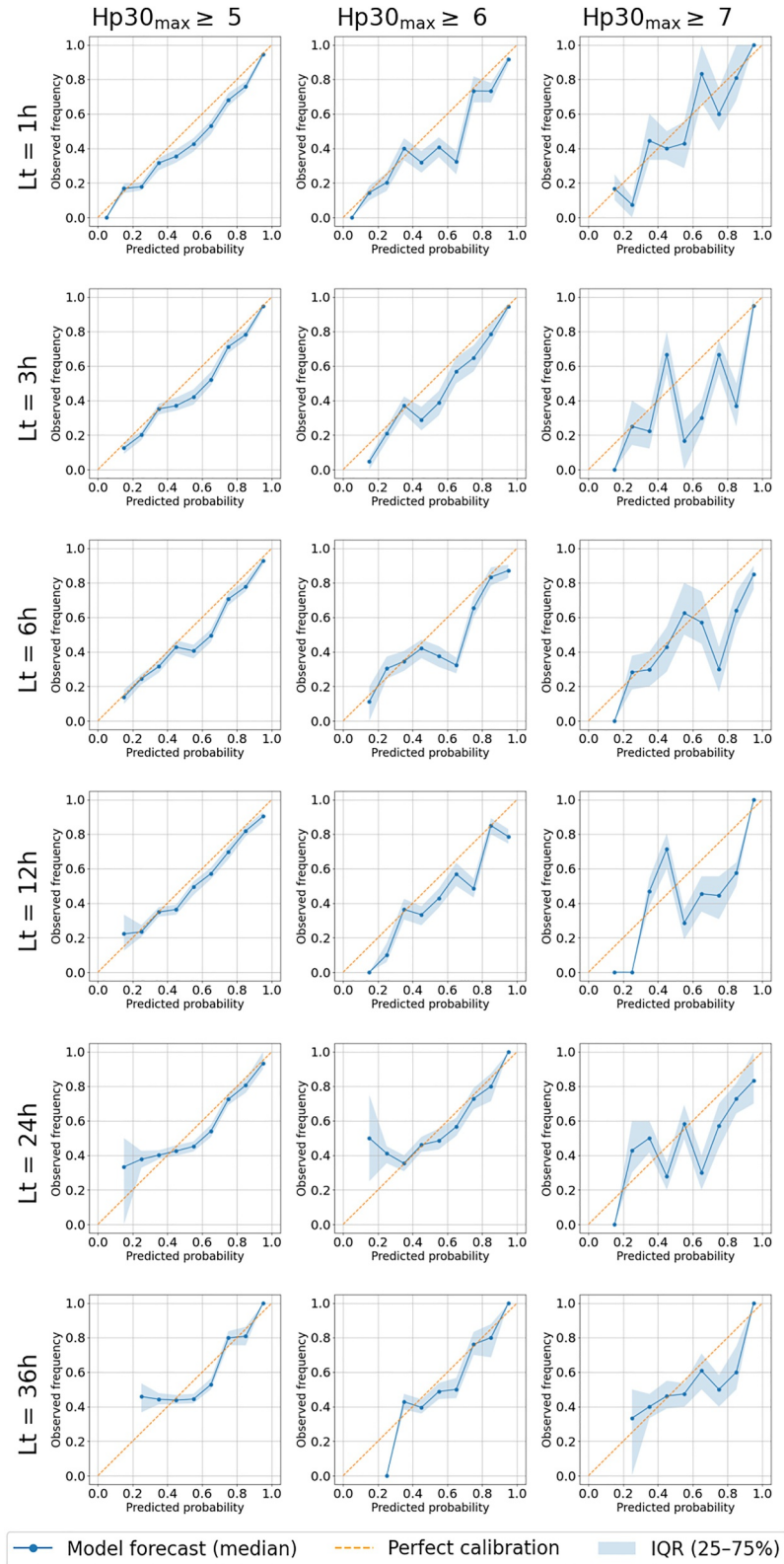

Figure 6. Calibration plots for varying lead times (Lt) and Hp30 storm thresholds. Each panel plots the models forecasted probabilities against the observed frequency (blue), along with the shaded region showing inter quartile range for the associated lead time and geomagnetic storm threshold. The orange line $y = x$ indicates the perfect calibration line.

10.1029/2025SW004823

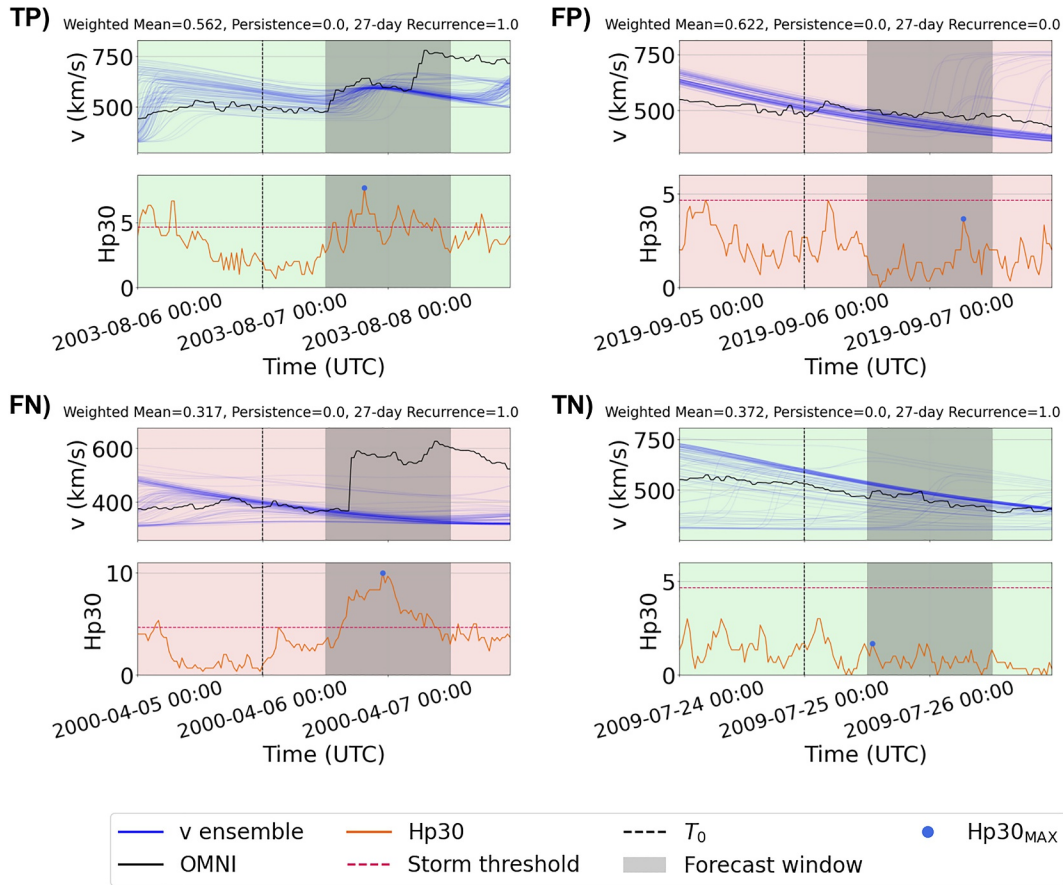

Figure 7. Four case studies of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) are shown. Each quadrant shows the ensemble of velocities (blue), observed velocities from OMNI (black), observed Hp30 (orange). The storm threshold of 4.66 is marked as a dashed red line, with $Hp30_{MAX}$ indicated by a blue dot. T_0 (dashed black), and the forecast window (gray) are also shown. A single shared legend is displayed at the bottom of the plot. Each quadrant is titled with the weighted mean, persistence and 27-day recurrence forecasts for the given forecast window.

likely because the model is able to use the current state of the solar wind and magnetosphere effectively. Figure A2 shows that the PACF of Hp30 drops below the statistically significant level at 21 hr indicating that Hp30 is a less useful input for longer lead time forecasting.

From the metrics in Figures 5a and 5c, alone, we could conclude that the performance at longer lead times is equal across all storm intensities, but we will discuss in Section 5.3 why this is not necessarily the case.

5.3. Forecast Uncertainty and Reliability

As discussed above, the model performance decreases according to both metrics as the lead time increases. In Figures 5b and 5d, the uncertainty of model performance for geomagnetic storm threshold of Hp30 $\geq$ 5 is particularly low, indicating that our model is very consistently performing well. The variability is notably larger for Hp30 $\geq$ 7 and Hp30 $\geq$ 8 thresholds. This is likely a reflection of the sample size for such large events, many of which are CMEs. Taking the median values across test folds, there are 43 periods where Hp30 $\geq$ 7, and 12 periods where Hp30 $\geq$ 8 in the test set. Given only 24 periods (12 storm and 12 non-storm) to test on, we expect the confident intervals to widen substantially for higher Hp30 thresholds.

We now take a look at how well calibrated our model is. A calibration plot shows how well a model's predicted probabilities match the observed outcomes. A probabilistic model is well-calibrated if, among all instances where it predicts probability p, the empirical frequency of the event is equal to p. We can plot the observed frequency of a geomagnetic storm for varied model forecasts. In Figure 6, the x-axis represents the models predicted probability split into 10 equal bins, while the y-axis shows the actual observed frequency of events. A perfectly calibrated

model will lie along the diagonal line $y = x$, where predicted probabilities match observed frequencies exactly. Points above the diagonal indicate the model is under confident: events occur more frequently than predicted. Points below the diagonal indicate the model is overconfident: events occur less frequently than predicted. Note that the Hp30 ≥ 8 threshold is omitted because the sample size is too small to generate any meaningful calibration plot.

On our whole test set (i.e., Hp30 threshold $\geq$5), our model appears to be well calibrated across all lead times, with the lines showing a smooth general trend and fair proximity to the perfect calibration line. The inter quartile range of the calibration lines also stay consistently low, highlighting consistency in the probabilities across the test set. Although the general trend follows the $y = x$ reference line (of perfect calibration), the model is slightly over-predicting. This over-prediction will be reflected in both the ROC AUC and BSS_{clim} metrics, and we would expect a better calibrated model to perform better in terms of these metrics. We expect our model to provide a useful probability of geomagnetic storms on a generic test set even at lead time of 24 hr. At a lead time of 36 hr, predicted probabilities below 0.5 exhibit limited skill, as shown by the flat calibration curve at low predicted probabilities in the bottom-left panel of Figure 6. We also see that at large lead times (e.g., 24 and 36 hr) the lowest probabilities are the least well calibrated. This is likely due to CMEs driving geomagnetic storms, which are not included in our current modeling framework. As we increase the storm threshold to $Hp30_{MAX} \geq 6$ and $Hp30_{MAX} \geq 7$, the calibration is worse. This is likely a combination of decreasing sample size, and that the model is less effective on larger scale events due to the lack of transients in our numerical modeling. On these larger events, our distribution of forecasts is more heavily peaked around 0.5, showing a lack of decisiveness of our model on these types of geomagnetic storms.

6. Discussion

6.1. Ambient Solar Wind Forecasting

We pick a case study corresponding to each element in a contingency table: TP, FP, FN, and TN as described in Section 4. This is useful to pick out what is likely contributing to both correct and incorrect forecasts for our model, and we show each of the four cases in Figure 7. The chosen case studies are all for 12-hr lead time, and were selected because the properties they exhibit are consistently present in our test set.

The true positive (TP) in Figure 7 is a standard example of a correctly forecasted storm, where our solar wind profiles match the same general trend as OMNI up to the point of the storm. Every one of our solar wind ensemble exhibits an increase in solar wind velocity occurring at the start of the forecast window, triggering the geomagnetic storm, with the increase only differing in time and magnitude. This rise in solar wind velocity is likely a corotating interaction region given that it's present in the HUXt ensemble, and that the storm is correctly forecasted by the 27-day recurrence model. The second rise in OMNI occurs after $Hp30_{MAX}$ and is misaligned with the solar wind ensemble where the majority of the ensemble members expect the increase approximately 12 hr later.

The false positive (FP) in Figure 7 is an incorrect forecast of a storm. Within the input window, there is low discrepancy between ensemble members, leading to similar weighting for the final forecast. The solar wind ensemble exhibits a similar downwards trend as seen in OMNI, but with 19 of the 100 ensemble members incorrectly expecting a sharp rise in solar wind velocities. Infact, the three ensemble members with the lowest MAE with OMNI in the input window (and hence highest weighting) all contain this sharp rise which is likely contributing to the false positive.

The false negative (FN) in Figure 7 is a common situation. There is a CME arrival near the beginning of the forecast window, and because of the exclusion of CMEs in the solar wind ensemble, we never expect to be able to forecast this type of event. Notably here, the velocity of the CME is not particularly fast, peaking at just over 600 km/s, indicating that there is likely some other feature of the solar wind triggering the storm. The north-south magnetic field component of the solar wind (B_Z) when southward (negative) is linked to strong geomagnetic activity. At the time of the CME, B_z flipped from northward to southward (not shown), which contributed to the particularly high Hp30 activity observed in this case, even though the solar wind velocity remains relatively low. IMF orientation can be a useful parameter for storm forecasting, especially in cases like this, but providing a model with a reliable forecast of B_z is very challenging. B_z can be forecasted through the use of a full 3D MHD model, but this is very computationally expensive and it would likely require a large ensemble to capture B_z effectively.

The true negative (TN) in Figure 7 we observe that the majority of the solar wind ensemble shows the same general decreasing solar wind trend as OMNI. Some of our ensemble members are notably low at 300 km/s. 27-day recurrence forecasts a storm here, but since the feature contributing to the fast solar wind is dissipating, only some of the ensemble members expect a rise in the solar wind speed during the forecast window.

The tools used in this study give us an idea what the ambient solar wind velocity is going to look like in the near future, to enable longer lead time forecasts. Shorter lead time Kp forecasting models for example, Tan et al. (2018); Chakraborty and Morley (2020), which predict 3 hr ahead, don't include forecasted solar wind in their input, and consequently don't have an indication of possible future solar wind conditions. Our approach is fundamentally different to forecasting a single velocity profile for the solar wind, and is instead a model trained on a range of potential solar wind conditions. We note that our "future" solar wind profiles are ambient solar wind forecasts, and do not yet explicitly include CMEs. Additionally, 27-day recurrence correctly forecasts 3 out of 4 cases in Figure 7. 27-day recurrence is a good deterministic model, but the metrics shown in Figures 4 and 5 highlight our models probabilistic insight.

6.2. Ensemble Calibration

Significant work has been done to improve the quality of our solar wind ensemble using a data assimilative approach (Barnard et al., 2023; Lang et al., 2021; Turner et al., 2023). Lang et al. (2021) uses comparisons with in situ data from STEREO-A, STEREO-B, and ACE satellites to update the inner boundary conditions for HUXt, and shows a 31.4% reduction in root mean squared error of solar wind forecasts compared to non data assimilative methods. Data assimilations lies beyond the scope of this study, but we aim to incorporate this modeling technique in future work.

Finding the optimal number of ensemble members is complex and will be investigated in the future, since we have sufficient ensembles to capture the range of variability that we expect, and a simple problem framework. Milinski et al. (2020) proposes a standard method for calculating the required number of ensembles to reduce the uncertainties to the required error, specific to the problem. The acceptable error needs to be uniquely determined for each problem. The method for finding sufficient ensembles involves subsampling a large ensemble and estimating error on a specified metric. This can be used to find the minimum ensemble size that satisfies the required error condition.

6.3. Forecasting Large Storms With Long Lead Time

We expect performance to decrease across lead times, consistent with the influence of Hp30 over time (Figure A2) and the diminishing impact of solar wind velocity with increasing forecast horizon. For our whole test set ($Hp30_{MAX} \geq 5$, Figure 5 shows that model skill is highest at short lead times: at 1 hr (ROC AUC = 0.801, BSS = 0.631) and 3 hr (ROC AUC = 0.785, BSS = 0.618), performance remains strong. At intermediate lead times of 6 hr (ROC AUC = 0.751, BSS = 0.595) and 12 hr (ROC AUC = 0.714, BSS = 0.568), the model maintains good skill, though a slight decrease is evident. From Figure 5, we note that for longer lead times of 24 and 36 hr, the BSS indicates that performance on very strong storms weakens somewhat. This behavior is expected, as high Hp30 storms are more likely to be CME-driven, and since CMEs are not included in the model inputs, the model's ability to forecast these events is naturally limited.

Limiting the storm subset to those where Hp30 exceeds 8 introduces a degree of randomness to the results, limiting confidence in extrapolating this performance to larger data sets. This limitation stems from the rarity of such extreme storms in the available Hp30 data set (1995–present), which is the limiting factor for out data. This is reflected in the large uncertainty for $Hp30 \geq 7$ and $Hp30 \geq 8$ seen in Figure 5.

There is a gradual drop in both metrics as lead time increases (i.e., Figures 4 and 5), but for most storm sizes the smallest change in metrics is between 24- and 36-hr lead times. This suggests that initial conditions in the input window (i.e., Hp30 values) become less influential over longer lead times due to the highly dynamic nature of the magnetosphere and relative autocorrelation timescales. Instead of relying on Hp30 values, the model relies more heavily on ensemble data for these extended forecasts at these long lead times. This highlights the importance of accurate ensemble initialization and selection for extended lead time predictions.

We note that the 27-day recurrence model demonstrates strong predictive skill at long lead times (≥24 hours) compared with our model. The 27-day recurrence model will perform better during solar minima, when

storms are more often driven by stable solar structures rather than transient CMEs. We tested the performance of the baseline models as we increase the storm threshold. When taking median values of metrics for 27-day recurrence across all lead times, ROC AUC Score decreases from 0.634 for.

$Hp30_{MAX} \geq 5$ to 0.619 for $Hp30_{MAX} \geq 8$, and BSS_{clim} decreases from 0.269 to 0.239, as expected. In contrast, persistence shows the opposite trend: ROC AUC Score increases from 0.567 for $Hp30_{MAX} \geq 5$ to 0.625 for $Hp30_{MAX} \geq 8$, and BSS_{clim} increases from 0.134 to 0.250. Given the definition of the problem, we expect persistence to perform better on geomagnetic storms with higher $Hp30_{MAX}$ since these storms tend to persist longer, as it takes time for total geomagnetic disturbance in the magnetosphere to decay below $Hp30_{MAX} < 4.66$.

6.4. Limitations of Current Ensembles

We have extensively explored the limitations of using forecasts of the ambient solar wind, and so notably we do not (as yet) include solar wind transients (i.e., CMEs) in our forecasts. The 1D incompressible hydrodynamic approach of HUXt has the benefit of computational efficiency; however, it does limit the solar wind forecast to velocity only. This is in contrast to 3D MHD approaches, which would allow the estimation of properties such as the magnetic field. Such forecasts would undoubtedly be beneficial, as many previous studies have noted that parameters such as the dawn-dusk electric field or magnetic field orientation are powerful predictors of magnetospheric activity (A. W. Smith et al., 2020; Ma et al., 2024). Future methods could explore the use of more computationally intensive solar wind propagation methods, though this would likely limit the practical ensemble size.

Another potential limitation is that we have chosen MAS due to data availability, but this inevitably contributes to the quality and uncertainty of the solar wind ensemble. A comparison of solar wind models is shown and discussed in (Barnard & Owens, 2022), and in the future we aim to investigate the direct impact of HUXt boundary conditions on the quality of forecast we provide. For example, we would expect to extract a different solar wind ensemble from the WSA model (Arge & Pizzo, 2000). This also ties in with future work on how we would operationalize a similar model, since MAS is not available in near real time.

6.5. Operational Forecasting

From an operational perspective, the consistently high BSS at short lead times suggests that forecasts are particularly useful for shorter warning windows, where actionable confidence is highest. At longer lead times, users may rely more on the discriminative capacity of the model (ROC AUC) rather than absolute probabilistic calibration.

Operationalizing HUXt is highly feasible with minimal modifications and HUXt solar wind ensembles are run live at the University of Reading: https://research.reading.ac.uk/met-spate/huxt-forecast/. The primary challenge is that MAS solutions (https://www.predsci.com/mhdweb/home.php) are not available in real time; however, we could substitute them with WSA solutions, which serve as a like-for-like replacement. This substitution would require retraining our model due to significant differences between WSA and MAS.

The Hpo indices are available in near real time, allowing us to make a forecasting with a 1-hr buffer region. Additionally, the solar wind velocities from L1, taken from the OMNI data set are not available in real-time due to processing time. Near real-time solar wind velocities can be inconsistent and could inhibit ensemble calibration. Near real-time solar wind is available from sources such as the DSCOVR satellite (Valero & Herman, 2006), or ACE satellite (Zwickl et al., 1998), which can be utilized by adjusting model input and retraining accordingly, as stated in A. Smith et al. (2022). Turner et al. (2023) presents an improvement on the reliability of real-time solar wind data with a data-assimilation approach, showing that solar wind forecasts using near real-time data are comparable to forecasts based on science-level data.

7. Conclusion

This study presents a novel approach to geomagnetic storm forecasting by integrating solar data with advanced modeling and machine learning techniques. Using the reduced-physics HUXt numerical model, an ensemble of ambient solar wind velocity profiles at Earth was generated based on output from the 3D MHD MAS algorithm. Machine learning models were trained on these profiles to classify geomagnetic storm events, leveraging an ensemble aggregation method, calibrating our ensemble members by their historical performance with observed solar wind data. Model performance was evaluated across different storm intensities and lead times, with particular focus on two metrics: ROC AUC Score and BSS against Climatology. The study highlights the

importance of solar wind ensembles, and the use of solar data for extending forecast lead times. The following is a summary of our methods and key findings:

Methods:

- Utilized a relatively large ensemble of boundary conditions to capture the variability in ambient solar wind.
- Used a multi-model ensemble approach to classify geomagnetic storms by coupling physics-based numerical models with machine learning models.
- Adopted the Hp30 geomagnetic activity index for its finer 30-min cadence and open-ended range compared to the 3-hourly, capped Kp index.

Key Findings:

- Demonstrated improvements in predictive performance over baseline models by selecting and weighting solar wind ensemble members based on error alignment with historical solar wind observations to provide a probability of exceeding a specified storm threshold.
- Forecast ability decreases with increasing lead times, though performance does not decrease to the same extent between 24 and 36 hr, as the quality of the underlying solar wind ensemble becomes dominant.
- We perform equally well across varying storm intensities for the ROC AUC Score and BSS_{clim} metrics, but expect the probabilities on larger events to be less well calibrated.
- Discussed the steps required to operationalize this model, which can be done with few adjustments: substitute MAS solutions for WSA solutions; substitute observed OMNI solar wind data for ACE or DSCOVR solar wind data; and retraining Model C(i) accordingly.

This study represents a significant advance in lead time forecasting of geomagnetic storms, crucial for mitigating the impacts of space weather on Earth-based infrastructure. By combining numerical models with machine learning, the approach begins to bridge the gap between physics-driven simulations and data-driven predictions, enabling new insights and forecasts. Future work will focus on incorporating CME-specific ensembles to address existing challenges, paving the way for robust operational forecasting.

Appendix A: Autocorrelation of Hp30

Autocorrelation is defined as the correlation between a time series, and that same time series shifted by a number of time steps (lag). Figure A1 presents the autocorrelation of the Hp30 index for lags between 22 and 32 days, chosen due to the 27-day rotation of the Sun from Earth's perspective, plus and minus 5 days.

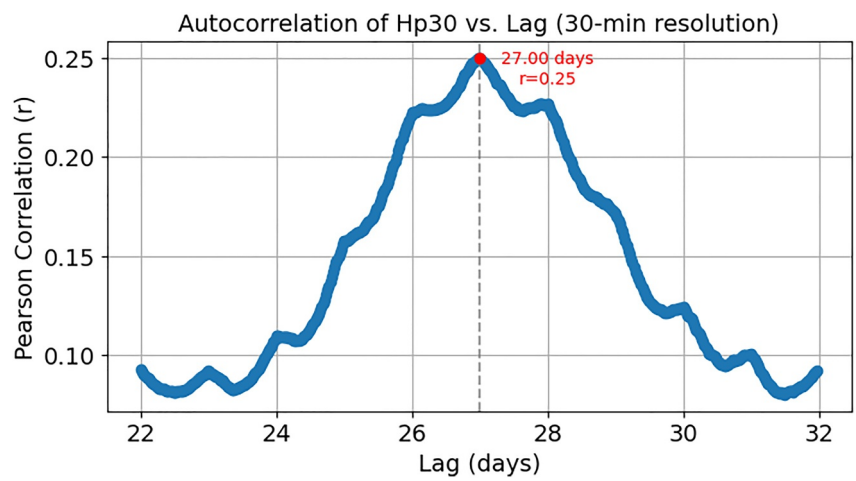

Figure A1. Autocorrelation of Hp30 index between 22 and 32 days. The *x*-axis shows lag for each calculation, with *y*-axis showing the corresponding Pearson correlation coefficient for each lag. The red dot indicates the highest autocorrelation for the lags shown.

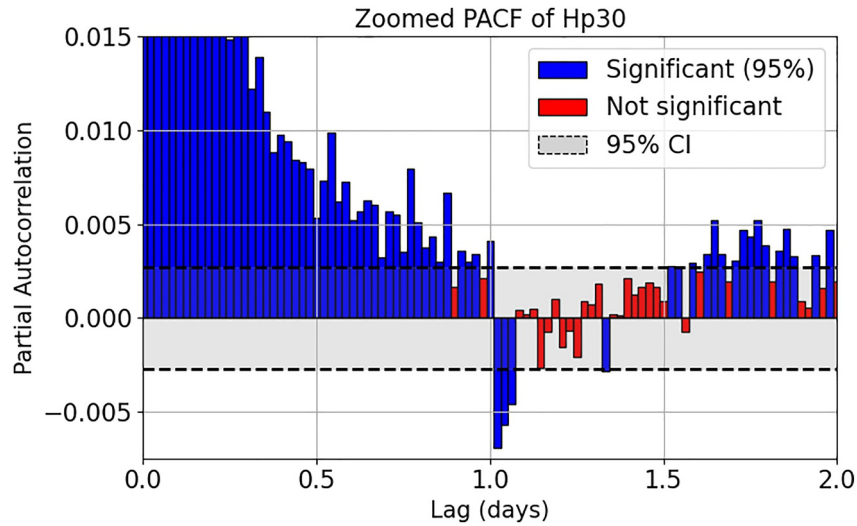

Figure A2. Partial Autocorrelation Function of the Hp30 index. The gray shaded area shows the 95% Confidence Interval (CI). Blue bars represent statistically significant autocorrelation (outside 95% CI), while red bars represent insignificant autocorrelation (inside 95% CI.).

Partial autocorrelation function (PACF) measures the autocorrelation at a specific lag when the effects of all shorter lags have been removed. Figure A2 shows the PACF of the Hp30 index, with 95% confidence intervals (CI) indicating at which point the autocorrelation is insignificant.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

HUXt is an open-source solar-wind model available at: https://doi.org/10.5281/zenodo.4889326 (M. J. Owens & Barnard, 2024). Ensemble calibration data was gathered from OMNI solar-wind observations (Papitashvili & King, 2020). MAS coronal model solutions used in this study: https://www.predsci.com/mhdweb/home.php. Hpo is sourced from https://www.gfz-potsdam.de/en/hpo-index. The code for training and testing the model used in this manuscript is available at: https://doi.org/10.5281/zenodo.17571893 (Billcliff, 2025).

Acknowledgments
We acknowledge use of NASA/GSFC's Space Physics Data Facility's OMNIWeb (or CDAWeb or ftp) service, and OMNI data. We would like to thank all those who provided data used in this study. MB was supported by a Northumbria University PhD studentship. AWS was supported by NERC Independent Research Grant NE/W009129/1. MO is part-funded by Science and Technology Facilities Council (STFC) Grant ST/V000497/1 and NERC Grant NE/Y001052/1. NE-I is funded through SCENARIO Grant NE/S007261/1. LB is supported by UKRI Future Leaders Fellowship MR/Y021207/1.

References

Arge, C., & Pizzo, V. (2000). Improvement in the prediction of solar wind conditions using near-real time solar magnetic field updates. *Journal of Geophysical Research*, *105*(A5), 10465–10479. https://doi.org/10.1029/1999ja000262

Baker, D. N. (2001). Satellite anomalies due to space storms: The effects of space weather on spacecraft systems and subsystems. *Space storms and space weather hazards*, 285–311. https://doi.org/10.1007/978-94-010-0983-6_11

Barnard, L., & Owens, M. (2022). Huxt—An open source, computationally efficient reduced-physics solar wind model, written in python. *Frontiers in Physics*, *10*, 1005621. https://doi.org/10.3389/fphy.2022.1005621

Barnard, L., Owens, M., Scott, C., Lang, M., & Lockwood, M. (2023). Sir-huxt—A particle filter data assimilation scheme for Cme time-elongation profiles. *Space Weather*, *21*(6), e2023SW003487. https://doi.org/10.1029/2023sw003487

Bartels, J. (1949). The standardized index, ks, and the planetary index, kp. *IATME Bulletin*, *12B*, 97–120. Retrieved from http://isgi.unistra.fr/IAGABulletins/IATME_Bulletin_12b_Herbert_Weisman_Bartels_1949.pdf

Bernoux, G., Brunet, A., Buchlin, É., Janvier, M., & Sicard, A. (2022). Forecasting the geomagnetic activity several days in advance using neural networks driven by solar euv imaging. *Journal of Geophysical Research: Space Physics*, *127*(10), e2022JA030868. https://doi.org/10.1029/2022ja030868

Billcliff, M. (2025). storm_forecasting_MB: Code for "Extended Lead-Time…" (v1.0.0). *Zenodo*. https://doi.org/10.5281/zenodo.17571893

Bolduc, L. (2002). Gic observations and studies in the hydro-québec power system. *Journal of Atmospheric and Solar-Terrestrial Physics*, *64*(16), 1793–1802. https://doi.org/10.1016/s1364-6826(02)00128-1

Boteler, D. (2003). Geomagnetic hazards to conducting networks. *Natural Hazards*, *28*(2–3), 537–561. https://doi.org/10.1023/a:1022902713136

Chakraborty, S., & Morley, S. K. (2020). Probabilistic prediction of geomagnetic storms and the kp index. *Journal of Space Weather and Space Climate*, *10*, 36. https://doi.org/10.1051/SWSC/2020037

Cliver, E. W., & Svalgaard, L. (2004). The 1859 solar–terrestrial disturbance and the current limits of extreme space weather activity. *Solar Physics*, *224*(1–2), 407–422. https://doi.org/10.1007/s11207-005-4980-z

Collado-Villaverde, A., Muñoz, P., & Cid, C. (2023). Neural networks for operational sym-h forecasting using attention and swics plasma features. *Space Weather*, *21*(8), e2023SW003485. https://doi.org/10.1029/2023sw003485
Conde, D., Castillo, F., Escobar, C., García, C., García, J., Sanz, V., et al. (2023). Forecasting geomagnetic storm disturbances and their uncertainties using deep learning. *Space Weather*, *21*(11), e2023SW003474. https://doi.org/10.1029/2023sw003474
Cranmer, S. R. (2009). Coronal holes. *Living Reviews in Solar Physics*, *6*(1), 3. https://doi.org/10.12942/lrsp-2009-3
Dumbović, M., Devos, A., Vršnak, B., Sudar, D., Rodriguez, L., Ruždjak, D., et al. (2015). Geoeffectiveness of coronal mass ejections in the soho era. *Solar Physics*, *290*(2), 579–612. https://doi.org/10.1007/s11207-014-0613-8
Edward-Inatimi, N. O., Owens, M. J., Barnard, L., Turner, H., Marsh, M., Gonzi, S., et al. (2024). Adapting ensemble-calibration techniques to probabilistic solar-wind forecasting. *Space Weather*, *22*(12), e2024SW004164. https://doi.org/10.1029/2024sw004164
Forsyth, C., Watt, C., Mooney, M., Rae, I., Walton, S., & Horne, R. (2020). Forecasting goes 15> 2 mev electron fluxes from solar wind data and geomagnetic indices. *Space Weather*, *18*(8), e2019SW002416. https://doi.org/10.1029/2019sw002416
Gaunt, C., & Coetzee, G. (2007). Transformer failures in regions incorrectly considered to have low gic-risk. In *2007 ieee lausanne power tech* (pp. 807–812). https://doi.org/10.1109/PCT.2007.4538419
Gonzalez, W., Joselyn, J.-A., Kamide, Y., Kroehl, H. W., Rostoker, G., Tsurutani, B., & Vasyliunas, V. (1994). What is a geomagnetic storm? *Journal of Geophysical Research*, *99*(A4), 5771–5792. https://doi.org/10.1029/93ja02867
Gopalswamy, N., Barbieri, L., Cliver, E., Lu, G., Plunkett, S., & Skoug, R. M. (2005). Introduction to violent sun-earth connection events of October–November 2003. *Journal of Geophysical Research*, *110*(A9). https://doi.org/10.1029/2005ja011268
Gruet, M. A., Chandorkar, M., Sicard, A., & Camporeale, E. (2018). Multiple-hour-ahead forecast of the dst index using a combination of long short-term memory neural network and gaussian process. *Space Weather*, *16*(11), 1882–1896. https://doi.org/10.1029/2018sw001898
Hands, A. D., Ryden, K. A., Meredith, N. P., Glauert, S. A., & Horne, R. B. (2018). Radiation effects on satellites during extreme space weather events. *Space Weather*, *16*(9), 1216–1226. https://doi.org/10.1029/2018sw001913
Issan, O., & Kramer, B. (2023). Predicting solar wind streams from the inner-heliosphere to Earth via shifted operator inference. *Journal of Computational Physics*, *473*, 111689. https://doi.org/10.1016/j.jcp.2022.111689
Kilpua, E., Fontaine, D., Moissard, C., Ala-Lahti, M., Palmerio, E., Yordanova, E., et al. (2019). Solar wind properties and geospace impact of coronal mass ejection-driven sheath regions: Variation and driver dependence. *Space Weather*, *17*(8), 1257–1280. https://doi.org/10.1029/2019sw002217
Lang, M., Witherington, J., Turner, H., Owens, M. J., & Riley, P. (2021). Improving solar wind forecasting using data assimilation. *Space Weather*, *19*(7), e2020SW002698. https://doi.org/10.1029/2020sw002698
Leka, K., Park, S.-H., Kusano, K., Andries, J., Barnes, G., Bingham, S., et al. (2019). A comparison of flare forecasting methods. Ii. Benchmarks, metrics, and performance results for operational solar flare forecasting systems. *The Astrophysical Journal Supplement Series*, *243*(2), 36. https://doi.org/10.3847/1538-4365/ab2e12
Ma, D., Bortnik, J., Ma, Q., Hua, M., & Chu, X. (2024). Machine learning interpretability of outer radiation belt enhancement and depletion events. *Geophysical Research Letters*, *51*(1), e2023GL106049. https://doi.org/10.1029/2023gl106049
Marshall, R., Dalzell, M., Waters, C., Goldthorpe, P., & Smith, E. (2012). Geomagnetically induced currents in the New Zealand power network. *Space Weather*, *10*(8). https://doi.org/10.1029/2012sw000806
Mayank, P., J. Athalathil, J., Nandy, S., Vaidya, B., Navanit, A., & Paul, A. (2025). Swasti: A physics-based modelling toolkit for space weather. *P. mayank et al. Journal of Astrophysics and Astronomy*, *46*(2), 80. https://doi.org/10.1007/s12036-025-10107-2
Milinski, S., Maher, N., & Olonscheck, D. (2020). How large does a large ensemble need to be? *Earth System Dynamics*, *11*(4), 885–901. https://doi.org/10.5194/esd-11-885-2020
Nitti, S., Podladchikova, T., Hofmeister, S. J., Veronig, A. M., Verbanac, G., & Bandić, M. (2023). Geomagnetic storm forecasting from solar coronal holes. *Monthly Notices of the Royal Astronomical Society*, *519*(2), 3182–3193. https://doi.org/10.1093/mnras/stac3533
Odstrčil, D., & Pizzo, V. J. (1999). Three-dimensional propagation of coronal mass ejections (CMES) in a structured solar wind flow: I. CME launched within the streamer belt. *Journal of Geophysical Research*, *104*, 483–492. https://doi.org/10.1029/1998JA900019
Odstrčil, D., & Pizzo, V. J. (1999). Three-dimensional propagation of coronal mass ejections (CMES) in a structured solar wind flow: II. CME launched adjacent to the streamer belt. *Journal of Geophysical Research*, *104*(A1), 493–503. https://doi.org/10.1029/1998JA900038
Oughton, E. J., Hapgood, M., Richardson, G. S., Beggan, C. D., Thomson, A. W., Gibbs, M., et al. (2019). A risk assessment framework for the socioeconomic impacts of electricity transmission infrastructure failure due to space weather: An application to the United Kingdom. *Risk Analysis*, *39*(5), 1022–1043. https://doi.org/10.1111/RISA.13229
Owens, M., Lang, M., Barnard, L., Riley, P., Ben-Nun, M., Scott, C. J., et al. (2020). A computationally efficient, time-dependent model of the solar wind for use as a surrogate to three-dimensional numerical magnetohydrodynamic simulations. *Solar Physics*, *295*(3), 43. https://doi.org/10.1007/s11207-020-01605-3
Owens, M. J., & Barnard, L. (2024). University-of-reading-space science/huxt: Huxt v4.2.0 (Software). https://doi.org/10.5281/zenodo.4889326
Owens, M. J., Chakraborty, N., Turner, H., Lang, M., Riley, P., Lockwood, M., et al. (2022). Rate of change of large-scale solar-wind structure. *Solar Physics*, *297*(7), 83. https://doi.org/10.1007/s11207-022-02006-4
Owens, M. J., Challen, R., Methven, J., Henley, E., & Jackson, D. (2013). A 27 day persistence model of near-earth solar wind conditions: A long lead-time forecast and a benchmark for dynamical models. *Space Weather*, *11*(5), 225–236. https://doi.org/10.1002/swe.20040
Owens, M. J., & Riley, P. (2017). Probabilistic solar wind forecasting using large ensembles of near-sun conditions with a simple one-dimensional "upwind" scheme. *Space Weather*, *15*(11), 1461–1474. https://doi.org/10.1002/2017SW001679
Papitashvili, N. E., & King, J. H. (2020). Omni hourly data [dataset]. *NASA Space Physics Data Facility*. https://doi.org/10.48322/1shr-ht18
Patterson, C. J., Wild, J. A., Beggan, C. D., Richardson, G. S., & Boteler, D. H. (2024). Modelling electrified railway signalling misoperations during extreme space weather events in the UK. *Scientific Reports*, *14*(1), 1583. https://doi.org/10.1038/s41598-024-51390-3
Pizzo, V., Millward, G., Parsons, A., Biesecker, D., Hill, S., & Odstrcil, D. (2011). Wang-sheeley-arge–enlil cone model transitions to operations. *Space Weather*, *9*(3). https://doi.org/10.1029/2011SW000663
Pomoell, J., & Poedts, S. (2018). Euhforia: European heliospheric forecasting information asset. *Journal of Space Weather and Space Climate*, *8*, A35. https://doi.org/10.1051/swsc/2018020
Pulkkinen, A., Lindahl, S., Viljanen, A., & Pirjola, R. (2005). Geomagnetic storm of 29–31 October 2003: Geomagnetically induced currents and their relation to problems in the Swedish high-voltage power transmission system. *Space Weather*, *3*(8). https://doi.org/10.1029/2004sw000123
Richardson, I. G. (2018). Solar wind stream interaction regions throughout the heliosphere. *Living Reviews in Solar Physics*, *15*(1), 1. https://doi.org/10.1007/s41116-017-0011-z
Riley, P., Linker, J. A., & Mikić, Z. (2001). An empirically-driven global mhd model of the solar Corona and inner heliosphere. *Journal of Geophysical Research*, *106*(A8), 15889–15901. https://doi.org/10.1029/2000JA000121

Schrijver, C. J. (2015). Socio-economic hazards and impacts of space weather: The important range between mild and extreme. *Space Weather*, *13*(9), 524–528. https://doi.org/10.1002/2015sw001252

Shprits, Y. Y., Vasile, R., & Zhelavskaya, I. S. (2019). Nowcasting and predicting the k p index using historical values and real-time observations. *Space Weather*, *17*(8), 1219–1229. https://doi.org/10.1029/2018sw002141

Siciliano, F., Consolini, G., Tozzi, R., Gentili, M., Giannattasio, F., & De Michelis, P. (2021). Forecasting sym-h index: A comparison between long short-term memory and convolutional neural networks. *Space Weather*, *19*(2), e2020SW002589. https://doi.org/10.1029/2020sw002589

Smith, A., Forsyth, C., Rae, I., Garton, T., Bloch, T., Jackman, C., & Bakrania, M. (2021). Forecasting the probability of large rates of change of the geomagnetic field in the uk: Timescales, Horizons, and thresholds. *Space Weather*, *19*(9), e2021SW002788. https://doi.org/10.1029/2021sw002788

Smith, A., Forsyth, C., Rae, I., Garton, T., Jackman, C., Bakrania, M., et al. (2022). On the considerations of using near real time data for space weather hazard forecasting. *Space Weather*, *20*(7), e2022SW003098. https://doi.org/10.1029/2022sw003098

Smith, A., Rae, I., Forsyth, C., Coxon, J., Walach, M.-T., Lao, C., et al. (2024). Space weather forecasts of ground level space weather in the uk: Evaluating performance and limitations. *Space Weather*, *22*(11), e2024SW003973. https://doi.org/10.1029/2024sw003973

Smith, A. W., Rae, I., Forsyth, C., Oliveira, D. M., Freeman, M. P., & Jackson, D. (2020). Probabilistic forecasts of storm sudden commencements from interplanetary shocks using machine learning. *Space Weather*, *18*(11), e2020SW002603. https://doi.org/10.1029/2020sw002603

Tan, Y., Hu, Q., Wang, Z., & Zhong, Q. (2018). Geomagnetic index kp forecasting with lstm. *Space Weather*, *16*(4), 406–416. https://doi.org/10.1002/2017SW001764

Thomson, A. W., Dawson, E. B., & Reay, S. J. (2011). Quantifying extreme behavior in geomagnetic activity. *Space Weather*, *9*(10). https://doi.org/10.1029/2011sw000696

Turner, H., Lang, M., Owens, M., Smith, A., Riley, P., Marsh, M., & Gonzi, S. (2023). Solar wind data assimilation in an operational context: Use of near-real-time data and the forecast value of an l5 monitor. *Space Weather*, *21*(5), e2023SW003457. https://doi.org/10.1029/2023sw003457

Valero, F. P., & Herman, J. (2006). Dscovr, the first deep space Earth and solar observatory. *Cell Biology and Instrumentation: UV Radiation, Nitric Oxide and Cell Death in Plants*, *371*, 44.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (Vol. 100). Academic Press.

Wing, S., Johnson, J., Jen, J., Meng, C.-I., Sibeck, D., Bechtold, K., et al. (2005). Kp forecast models. *Journal of Geophysical Research*, *110*(A4). https://doi.org/10.1029/2004ja010500

Wintoft, P., Wik, M., Matzka, J., & Shprits, Y. (2017). Forecasting kp from solar wind data: Input parameter study using 3-hour averages and 3-hour range values. *Journal of Space Weather and Space Climate*, *7*, A29. https://doi.org/10.1051/swsc/2017027

Yamazaki, Y., Matzka, J., Stolle, C., Kervalishvili, G., Rauberg, J., Bronkalla, O., et al. (2022). Geomagnetic activity index hpo. *Geophysical Research Letters*, *49*(10), e2022GL098860. https://doi.org/10.1029/2022GL098860

Zwickl, R., Doggett, K., Sahm, S., Barrett, W., Grubb, R., Detman, T., et al. (1998). The noaa real-time solar-wind (rtsw) system using ace data. *The advanced composition explorer mission*, *86*(1–4), 633–648. https://doi.org/10.1023/a:1005044300738