Forecasting occurrence and intensity of geomagnetic activity with pattern-matching approaches

Haines, C. ORCID: https://orcid.org/0000-0002-9010-0720, Owens, M. J. ORCID: https://orcid.org/0000-0003-2061-2453, Barnard, L. ORCID: https://orcid.org/0000-0001-9876-4612, Lockwood, M. ORCID: https://orcid.org/0000-0002-7397-2172, Ruffenach, A., Boykin, K. and McGranaghan, R. ORCID: https://orcid.org/0000-0002-9605-0007 (2021) Forecasting occurrence and intensity of geomagnetic activity with pattern-matching approaches. Space Weather, 19 (6). ISSN 1542-7390 doi: 10.1029/2020SW002624 Available at https://centaur.reading.ac.uk/98356/

Space Weather

RESEARCH ARTICLE
10.1029/2020SW002624

Special Section:
Space Weather Impacts on Electrically Grounded Systems at Earth's Surface

Forecasting Occurrence and Intensity of Geomagnetic Activity With Pattern-Matching Approaches

C. Haines[1], M. J. Owens[1], L. Barnard[1], M. Lockwood[1], A. Ruffenach[2], K. Boykin[1], and R. McGranaghan[3]

[1]Department of Meteorology, University of Reading, Reading, Berkshire, UK, [2]EDF Energy R&D UK Centre, Interchange, Croydon, UK, [3]Atmospheric and Space Technology Research Associates (ASTRA), Louisville, CO, USA

Key Points:

- Pattern-matching techniques are an effective way to forecast geomagnetic activity
- The analogue ensemble and support vector machine outperform 27 days recurrence and climatology
- The best forecast approach for the end user will depend on their need for probabilistic forecast information

Correspondence to:
C. Haines,
carl.haines@pgr.reading.ac.uk

Citation:
Haines, C., Owens, M. J., Barnard, L., Lockwood, M., Ruffenach, A., Boykin, K., & McGranaghan, R. (2021). Forecasting occurrence and intensity of geomagnetic activity with pattern-matching approaches. *Space Weather*, *19*, e2020SW002624. https://doi.org/10.1029/2020SW002624

Received 2 SEP 2020
Accepted 7 MAY 2021

Abstract Variability in near-Earth solar wind conditions gives rise to space weather, which can have adverse effects on space- and ground-based technologies. Enhanced and sustained solar wind coupling with the Earth's magnetosphere can lead to a geomagnetic storm. The resulting effects can interfere with power transmission grids, potentially affecting today's technology-centered society to great cost. It is therefore important to forecast the intensity and duration of geomagnetic storms to improve decision making capabilities of infrastructure operators. The 150 years aa_H geomagnetic index gives a substantial history of observations from which empirical predictive schemes can be built. Here we investigate the forecasting of geomagnetic activity with two pattern-matching forecast techniques, using the long aa_H record. The techniques we investigate are an Analogue Ensemble (AnEn) Forecast, and a Support Vector Machine (SVM). AnEn produces a probabilistic forecast by explicitly identifying analogs for recent conditions in the historical data. The SVM produces a deterministic forecast through dependencies identified by an interpretable machine learning approach. As a third comparative forecast, we use the 27 days recurrence model, based on the synodic solar rotation period. The methods are analyzed using several forecast metrics and compared. All forecasts outperform climatology on the considered metrics and AnEn and SVM outperform 27 days recurrence. A Cost/Loss analysis reveals the potential economic value is maximized using the AnEn, but the SVM is shown as superior by the true skill score. It is likely that the best method for a user will depend on their need for probabilistic information and tolerance of false alarms.

Plain Language Summary Space weather has the potential to disrupt society and the economy on a large scale. One such major impact is on power grids, which can be damaged by disturbances in Earth's magnetic field caused by space weather events. As a result, it would be useful to have an accurate forecast of space weather that can help power grid operators make decisions about taking mitigating action. In this work, we test three forecasting techniques which utilize long historical records to exploit patterns in the data and hence predict future disturbances in Earth's magnetic field. We find that all three of the techniques provide valuable information and the best method depends on the individual needs of the forecast user.

1. Introduction

Geomagnetic storms present a significant threat to critical infrastructure both in space and on the ground (Cannon et al., 2013; Oughton et al., 2017). Through solar wind energy input to the magnetosphere and the associated substorm process (e.g., Lockwood, 2019; Pulkkinen, 2007), Earth's ionospheric current systems can be dramatically enhanced (Buonsanto, 1999). Rapid fluctuations in these enhanced ionospheric currents can generate geomagnetically induced currents (GICs. e.g., Boteler, 1994; Pirjola, 2000) in ground-based conductors, posing a particular risk to power grids and pipelines. To ensure that we minimize service disruption and mitigate economic cost (e.g., Eastwood, Biffis, et al., 2017), there is a need for forecasting of both the intensity and duration of geomagnetic storms. Reliable forecasts improve the decision-making capabilities of operators of affected systems when taking mitigating action. However, current forecast capabilities are limited (Cannon et al., 2013; Koskinen et al., 2017).

Geomagnetic indices, which combine measurements from multiple ground-based magnetometers, are often used as a convenient measure of global geomagnetic activity because of their ability to reduce large-scale

physical processes into a single time-series of observations. Commonly used measures include low latitude indices *Dst* and *SYM-H*, the mid-latitude range index *Kp* and high latitude auroral index, *AE* (e.g., Lockwood, 2013).

Current approaches to forecasting geomagnetic indices cover a spectrum of techniques from first principal, physics-based attempts (Pulkkinen et al., 2013) (although even these often incorporate some empirical aspects in practice), through more empirical approaches, which range from those that still rely on domain-specific knowledge for their construction (Burton et al., 1975), to those that are almost entirely data-driven (e.g., Gu et al., 2019).

Global Magnetohydrodynamic (GMHD) models simulate the magnetosphere using solar wind data as the input. These provide a physics-based representation of the magnetosphere which can be run in real time (Eastwood, Nakamura, et al., 2017) and thus can be used operationally. Three of the main GMHD models were tested by Pulkkinen et al. (2013), with SWMF (Tóth et al., 2005, 2012) found to provide the most accurate reconstruction of $\frac{d\mathbf{B}}{dt}$, which is related to GIC intensity.

Owens et al. (2017b) argued for the use of empirical models for solar wind forecasting in conjunction with numerical physics-based models. Empirical models can add value because they have the advantage of being computationally cheap, meaning that they can be readily run in large ensembles to provide an estimate of uncertainty (Knipp, 2016). Empirical models can also parameterize unknown physics that a first-principle based model does not capture and act as a useful baseline with which to evaluate physics based models.

A range of empirical approaches has been attempted for geomagnetic index forecasting. Recently, Chandorkar et al. (2017) developed a "one step ahead" forecast of the *Dst* index using an auto-regressive Gaussian process approach. It was tested on a set of 68 storms and concluded that for a 1 h lead time, it out-preformed persistence on the metrics considered (mean absolute error (MAE), root-mean-square error, and correlation coefficient). Zhang and Moldwin (2015) produced a probabilistic forecast of *SYM-H* and *AE* using solar wind parameters to construct a cumulative probability distribution that the index would exceed the given intensity thresholds. A non-linear autoregressive with exogenous inputs (NARX) approach was employed by Ayala Solares et al. (2016) for forecasting the *Kp* index. They found that, in general, the NARX approach gave good results for short and long lead times, however it failed to surpass the neural network models of Wing et al. (2005), to which they were comparing. Other empirical approaches include: that of O'Brien and McPherron (2000) who employed a differential equation from Burton et al. (1975) which maps the evolution of the corrected *Dst* index, *Dst**; that by Vassiliadis and Klimas (1995) who used a driven harmonic oscillator circuit analogy; Vassiliadis et al. (1995) who used linear and nonlinear filters to predict the *AL*, *AU*, and *AE* indices. Camporeale (2019) summarized efforts using machine learning techniques to forecast geomagnetic indices. The majority of approaches use neural networks, however other machine learning techniques have been proposed. Lu et al. (2016) compared the use of Support Vector Machines (SVMs, Burges, 1998; Cortes & Vapnik, 1995), a machine learning approach that seeks to define a hyper-plane separating two classes, with neural networks for predicting intense storms in the *Dst* index using solar wind data as input parameters. Lu et al. (2016) concluded that SVMs out-perform neural networks for that application and can be improved further through the use of distance correlation learning (Székely et al., 2007). Liemohn et al. (2018) present an extensive list of works that forecast the behavior of *Dst*, SYM-H, *Kp*, *AE*, *AL*, and *AU* along with the metrics used to evaluate them. They showed no single metric is capable of metering a model for all applications. Therefore, we must be rigorous in our application of evaluative metrics. In this work, we explore numerous metrics, taking guide from Liemohn et al. (2018).

We here implement two-pattern matching forecasts, both requiring a large data set for training, and an additional recurrence forecast. The first method is an analogue ensemble (AnEn) forecast, a purely empirical approach. This method assumes that previous observations provide a good analogue for likely future variations (Delle Monache et al., 2013). Thus a historical record that is sufficiently long and covers a large enough range of behavior of the system can be used to identify previous periods when conditions are similar to the present. A forecast is constructed on the basis of the trajectories of the analogs forward in time. The "best" forecast in a deterministic sense is typically taken to be the median of the chosen analogs, but a large ensemble of analogs can provide probabilistic information. Probabilistic forecasting helps quantify

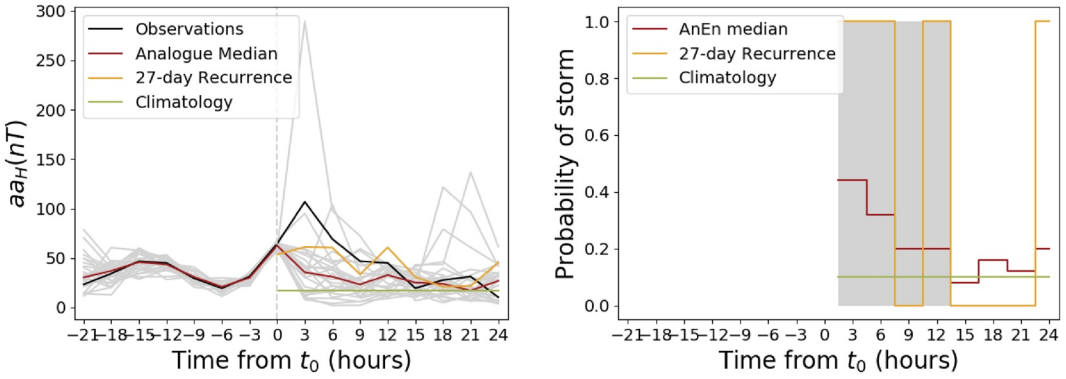


Figure 1. Left: An analogue ensemble (AnEn) forecast from 2017-12-05. The median AnEn forecast is shown in red with individual analogs in gray. The observed time-series of aa_H is shown in black. The benchmark forecasts are shown in yellow (27 days recurrence) and green (climatology). Right: The probability of a storm occurrence from each forecast method.

the forecast uncertainty, which benefits decision making, and can also be useful for evaluating a forecasting method (Knipp et al., 2018). An implementation of AnEn has been developed for this project in python.

AnEn has been used for a variety of parameters in terrestrial weather forecasting (e.g., Delle Monache et al., 2013; Van den Dool, 1989), but has been surpassed by physics-based models. This is largely due to the inherently chaotic nature of the system, which means states can rapidly diverge with a small perturbation to the initial conditions. Recently Owens et al. (2017a) and Riley et al. (2017) used an analogue forecast for solar wind parameters and the *Dst* index with some success, finding that it outperformed benchmarks of climatology and 27 days recurrence.

The second method investigated here is the SVM (Burges, 1998; Cortes & Vapnik, 1995), a supervised machine learning method for two-group classification. The volume and quality of data available, particularly in the aa_H index (see Section 2), and capabilities of modern computing means machine learning approaches are ripe for forecasting geomagnetic activity. SVMs seek to define a hyperplane which divides two classes (in this case "storm" and "no storm," for a given definition) and optimize it by maximizing the distance between it and the closest data-points, called support vectors. To aid linear separability of the classes, the vector space of input parameters is mapped into a higher dimension space using implicit mapping functions with a defined kernel function (Burges, 1998). A brief overview of the SVM and application to space-weather is given by McGranaghan et al. (2018). An implementation of the SVM has been developed for this project in python.

Here we will implement an SVM but, unlike many previous works, without the use of solar wind (exogenous) parameters as input and use solely the time history of observations before the time of forecast. This gives a more direct comparison with AnEn and, importantly, allows the best use of the 150 years aa_H data set, for most of which we do not have simultaneous solar wind observations. Given that both the solar wind transit time (between the usual point of observation the L1 point, and the magnetosphere), and the magnetospheric response time are small compared the time resolution of the aa_H data set (3 h), this is not expected to reduce forecast capability.

The third forecast type considered is 27 days recurrence, which implicitly assumes the structure of the coronal magnetic field, varies slowly compared to the solar rotation period. Thus, the same region of the Sun is directed toward the Earth every 27 days. This assumption is generally more valid during solar minimum and the late declining phase of the solar cycle than during solar maximum periods. Near-Earth solar wind conditions and the resulting geomagnetic activity have long been known to exhibit 27 days recurrence (Chree & Stagg, 1928; Bartels, 1932, 1934; Owens et al., 2013). The recurrence pattern is also present in the occurrence of moderate storms in the aa_H index (Haines et al., 2019, see also Section 2). Watari (2011) used a 27 days recurrence forecast for the *Kp* index, concluding that it was a viable forecast method during the declining phase of the solar cycle but not for other parts of the cycle.

Section 2 describes the aa_H data set. Section 3 describes our storm definition, the forecast methods considered and benchmarks used in this study. Section 4 compares the forecasts using metrics and techniques adopted from terrestrial weather forecasting (Henley & Pope, 2017). Many of the verification techniques are recommended by Liemohn et al. (2018) with the addition of Taylor diagrams and reliability diagrams (described below).

2. Data

We use the *aa* index (Mayaud, 1971), with recent recalibrations (aa_H Lockwood, Chambodut, et al., 2018; Lockwood, Finch, et al., 2018). Using a single magnetometer station in each of the UK and Australia, *aa* provides a quasi-global measure of geomagnetic activity with particular sensitivity to the substorm current wedge (Lockwood, 2013; Ganushkina et al., 2015). In order to span 150 years back to 1868, *aa* must be constructed from three different stations in each hemisphere, which introduces issues of calibration. However, this results in the longest available record of geomagnetic activity. The recent recalibrations account for the variation of mean geographic location of the midnight sector auroral oval, due to drifts of the Earth's geomagnetic poles. They also allow for time-of-day/time-of-year response pattern of the stations, thereby reducing uncertainties related to using just two stations. Lockwood et al. (2019) showed aa_H agrees well with *am* (Mayaud, 1981), a similar index but with much greater suppression of longitudinal sampling effects achieved by using multiple stations in each hemisphere. The disadvantage of *am* for the present study, of course, is that the data sequence is much shorter as the greater data requirement means it can only be constructed back to 1959.

Unfortunately, aa_H is limited by its temporal resolution of 3 h. Space weather impacts such as GICs occur due to magnetic fluctuation on a timescale of seconds and minutes. This means that a 3 h range index cannot give direct information on potential GICs but it can give an idea of the low frequency variation in the magnetosphere. The 3 h resolution of aa_H is more coarser than that of the 1 h resolution of *Dst*. Because of this resolution difference, *Dst* actually has more data points, despite the record only being available for approximately 60 years. Although more data points is useful for training models, aa_H spans around eight more solar cycles than *Dst* and so captures a more complete picture of the space climate.

A further limitation of *aa*, and hence aa_H, is that it is derived from K indices (Bartels et al., 1939) which are based on a quasi-logarithmic scale leaving quantization in the data set (Bubenik & Fraser-Smith, 1977). The uncertainty created from quantization is largest in the larger values of aa_H, but still present in the small.

Chapman et al. (2020) investigated the use of aa_H in characterizing extreme geomagnetic activity. They made a comparison of extreme aa_H with *Dst*, finding that there is good correspondence between the two indices and that it is possible to "read across" from extreme aa_H to extreme *Dst*.

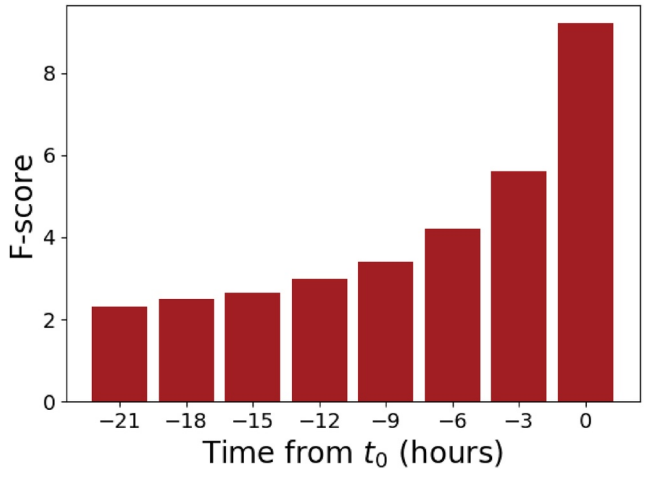


Figure 2. *F*-Scores of each 3 h aa_H data point in the 24 h leading up to the forecast time t_0. The *F*-scores show the relevance of each parameter to the data point immediately after t_0.

3. Methodology

3.1. Storm Definition

Various definitions of geomagnetic storms have been used, often dependent on the purpose of the study (Riley et al., 2018).The most common method is to set a threshold in a particular geomagnetic index (e.g., Vennerstrom et al., 2016) with values exceeding the threshold being defined as part of a storm, and a storm ends once the value of the index falls below the threshold. Kilpua et al. (2015) used a slight variation of this in which the last point of a storm is the first point below the threshold. Other approaches, such as that of Hutchinson et al. (2011), involve a manual inspection of the data, looking for characteristic storm traces for each event more intense than a chosen threshold. While a more nuanced approach than blindly applying a geomagnetic threshold, it is labor-intensive to apply to a large data set and is difficult to make truly repeatable.

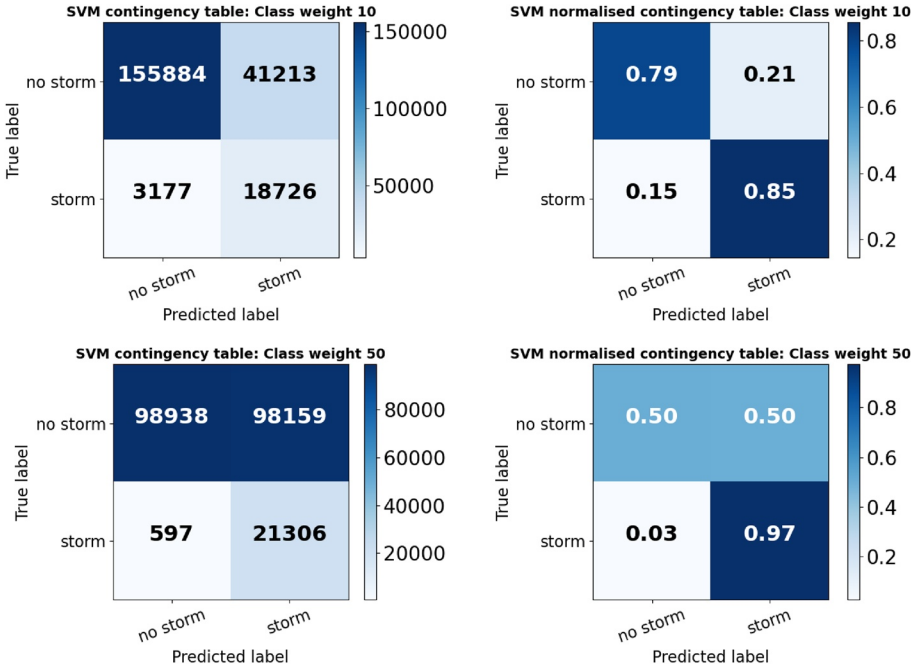


Figure 3. Examples of contingency tables, sometimes called a confusion matrix, for the Support Vector Machine (SVM) showing the number of occurrences (left) and normalized frequencies (right) of true and false positives and negatives. Top: SVM trained with a class weight of 10. Bottom: SVM trained with a class weight of 50. Changing the class weight affects the ratios of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

In this work we are concerned with the full spectrum of geomagnetic activity which can lead to adverse impacts on infrastructure. We point to the work of Schrijver (2015) and Schrijver et al. (2014) which examine the impact of moderate space weather using insurance claims data. Schrijver (2015) highlights that high frequency, low impact events may cumulatively be comparable in economic cost to low frequency, high impact events. Congruently, Schrijver et al (2014) examines insurance claims on electrical equipment identifying significant rises on both the top 5% and top third of geomagnetically active days. We therefore seek to choose a storm definition that captures moderate geomagnetic activity alongside the more rare, extreme events.

With the work of Schrijver (2015) and Schrijver et al. (2014) in mind, we use the same storm definition as Haines et al. (2019). This approach uses a simple threshold, similar to Vennerstrom et al. (2016) and Kilpua et al. (2015), but, as in Gonzalez et al. (1994), with a data-informed threshold set at the 90th percentile of the data set. For aa_H, this is a value of 40.1 nT. The start of the storm is the first point above the threshold and the end of the storm is the last point over the threshold. The effect of threshold on number of events is shown in Figure 2 of Haines et al. (2019).

3.2. Analogue Ensemble

To illustrate the methodology for the AnEn forecast, Figure 1 shows an event in the aa_H index from 2017-12-05. The observed time series (black) shows a storm, defined by exceeding a threshold of 40.1 nT, with storm onset at t_0. aa_H continues to rise until to a peak value of around 100 nT at the next data point (3 h later), then gradually falls back to non-storm conditions (i.e., below 40.1 nT). We identify the N previous events in the aa_H data set which most closely match the pattern of the observed time series in the 24 h time period before t_0, as described in more detail below. The time-series of these analogous periods are then projected forward to provide a probabilistic forecast after t_0. Also shown in Figure 1 is a 27 days recurrence forecast which can be used as a deterministic forecast of storm intensity, or, using the storm-definition threshold, to give a dichotomous storm forecast that is, that there will be a storm or that there will be no storm. The climatological mean value of aa_H is 17.5 nT, shown in green in the left panel, while the climatological prob-

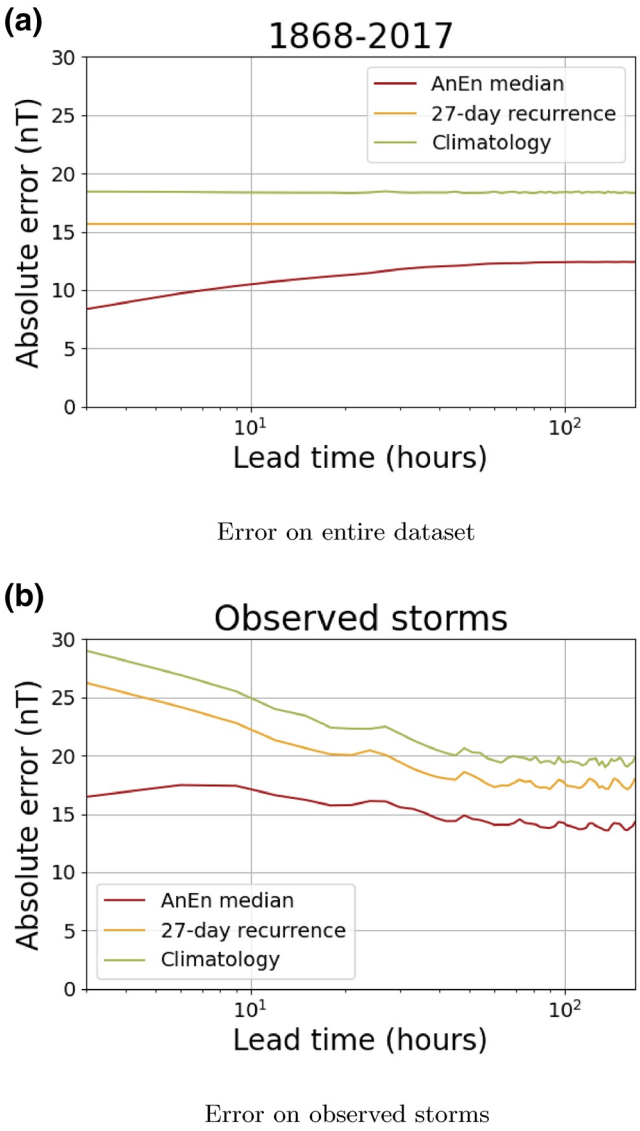


Figure 4. A comparison of the mean absolute error (MAE) of Analogue Ensemble (AnEn) median, 27 days recurrence and the intensity climatology for a range of lead times up to 300 h (a) MAE when the hindcast is run for t_0 at every point in the aa_H data set. (b) MAE for t_0 as the start of known storm events only.

ability of exceeding 40.1 nT is, by definition, 10% shown in green in the right-hand panel.

There are a number of aspects of the AnEn that must be tuned for the chosen application. The ensemble size, N, should be large enough to give sufficient resolution as a probabilistic forecast but small enough to ensure that the analogous periods are indeed analogous, particularly for rarer events such as larger storms. Values of N have been varied in the interval of (10, 50) without significant difference in results. Therefore, for clarity, we have selected a single ensemble size of 25 for the presentation of results in the remainder of this study.

While the input data to the AnEn is simply the recent time history of observations (the previous 24 h was shown in Figure 1), it is to be expected that some of these observations will be more relevant than others for predicting future behavior. For a highly driven system like the magnetosphere, the most recent observations are more likely to contain useful information about future evolution than observations from 24 h ago. We use the univariate F-score to determine the relevance of each input parameter, that is, each 3 hourly aa_H data point in the previous 24 h, to the subsequent data point of aa_H when forecasting with a 3 h lead time (Pedregosa et al., 2011). These F-scores are shown in Figure 2 and we see that the most recent observation is the most relevant as expected. The F-score is used as a weighting factor when selecting the best analogs. The total level of agreement is then the inverse of the mean of the weighted squared errors over the 24 h training window. The N analogs are then those with the lowest mean weighted squared error. These analogs are shown in Figure 1 by the gray lines converging as they approach t_0, and diverging significantly after t_0. Thus there is a wide range of possible future behavior on the basis of previous analogs to recent conditions. In Figure 1 the median of these analogs is shown in red. It matches the observations in the "training period," that is, −21–0 h, but in this particular example, under-predicts the observed intensity in the forecast window, that is, 0–24 h.

The right-hand panel of Figure 1 shows the probabilistic nature of AnEn, 27 days recurrence and climatology. The gray shaded region shows when a storm was actually observed to occur and the colored lines show the probability of storm conditions from each forecast considered. In this event the AnEn begins by predicting that a storm is likely with a probability of approximately 50%, which then drops over the next 12 h to around 25%. (i.e., 25% of the ensemble members are predicting $aa_H > 40.1$ nT at that time). The deterministic 27 days recurrence forecast does reasonably well in this particular example.

For analyzing and testing the performance of the forecast methods, we implement them here as hindcasts, predicting past events for which we already have the observations of the predicted period. The whole aa_H data set (excluding the 1 day prior, and 12.5 days subsequent, to t_0, which excludes the maximum extent of the training and evaluation windows) is made available for computing analogs. We use the full timeseries so that the results give an estimate of the predictive power of the model that would be deployed. The hindcasts have been run with the hindcast start time, t_0, at every point in the aa_H data set.

Due to the class imbalance between quiet and storm times, it is possible for a forecast to be valuable on average but perform poorly during storm times. For this reason, we additionally select and validate hindcasts during only the storm subset.

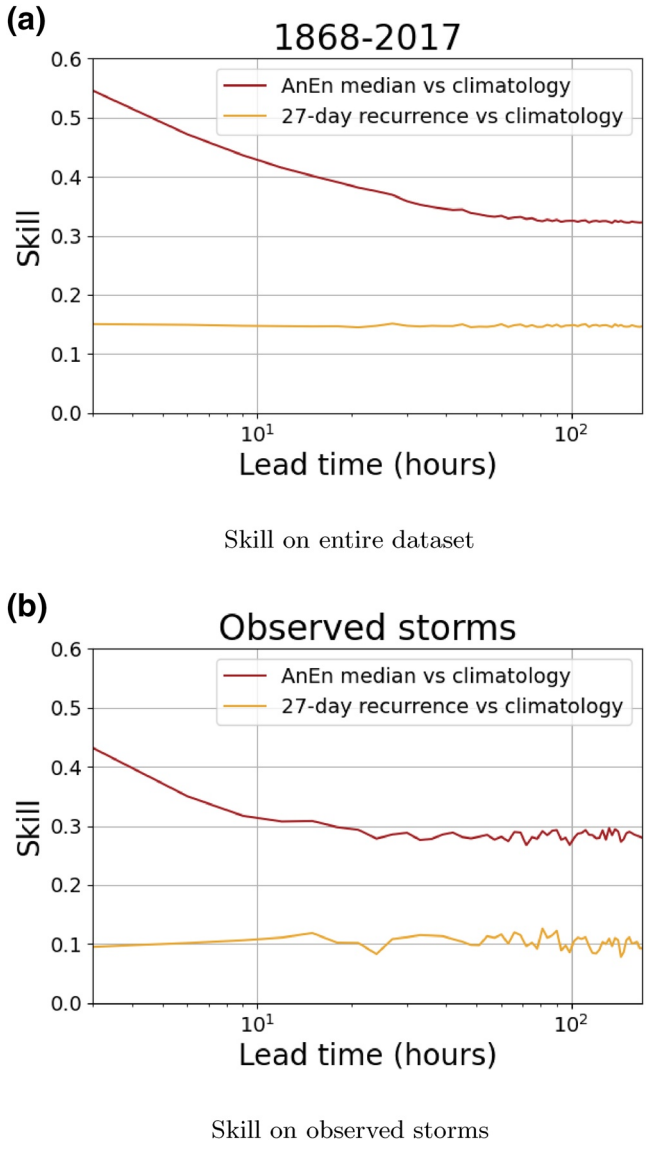


Figure 5. A comparison of skill for Analogue Ensemble (AnEn) median and 27 days recurrence with respect to climatology. (a) The entire aa_H data set. (b) Restricted to the period following observed storms.

3.3. SVM

The SVM is a commonly used classification algorithm, which we implement here to classify whether or not a storm will occur. Given a sample of the input and the associated classification labels, the SVM will find a function that separates these input features by their class label. This is simple if the classes are linearly separable, as the function is a hyperplane. The samples lying closest to the hyperplane are called support vectors and the distance between these samples and the hyperplane is maximized.

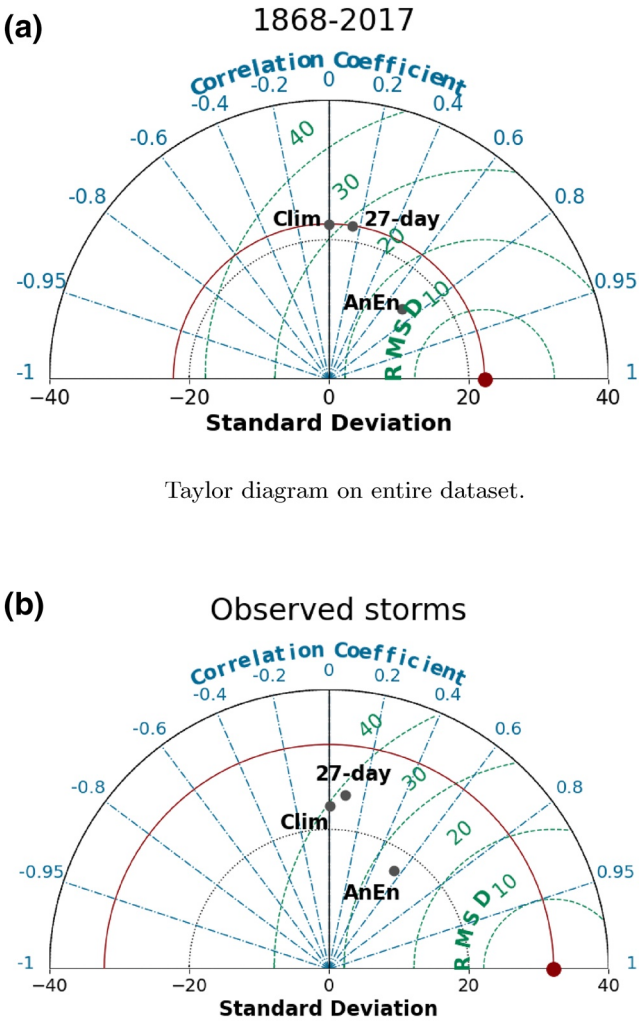
Typically, the samples are not linearly separable so we employ Cover's theorem (Cover, 1965) which states that linearly inseparable classification problems are more likely to be linearly separable when cast non-linearly into a higher dimensional spaces. Therefore, a kernel function is used to increase the dimensionality of the space. The Gaussian kernel is used for this purpose. It has a single hyperparameter, γ, which serves as a width parameter, determining the influence of a single data point on training. A sensitivity analysis has shown that an appropriate value for γ is 0.01.

On the basis of the aa_H values in the 24 h training window, the SVM predicts whether the next 3 h will be either a storm or not. By comparing this dichotomous hindcast with the observed aa_H, the outcome will be one of True Positive (TP, where a storm is correctly predicted), True Negative (TN, where no storm is correctly predicted), False Positive (FP, where a storm is predicted but not observed), or False Negative (FN, where a storm is not predicted but is observed). This is shown in the form of a contingency table in Figure 3 (top left).

For development of the SVM, the aa_H data has been separated into independent training and test intervals. These intervals are chosen to be alternate years. This is longer than the auto-correlation in the data (choosing, e.g., alternate 3 hourly data points, would not generate independent training and test data sets) but short enough that we assume there will not be significant aliasing with solar cycle variations.

Training is an iterative process, whereby a cost function is minimized. The cost function is a combination of the relative proportion of TP, TN, FP, and FN. Thus while training itself, an SVM attempts to classify labeled data that is, data belonging to a known category, in this case "storm" and "no storm" on the basis of the previous 24 h of aa_H. If the SVM makes an incorrect prediction it is penalized through a cost function which the SVM minimizes. The cost parameter determines the degree to which the SVM is penalized for a mis-classification in training which allows for noise in the data. A sensitivity analysis showed that an appropriate value for cost parameter is 0.1. See also Section 4.3.

It is common that data with a class imbalance, that is containing many more samples from one class than the other, causes the classifier to be biased toward the majority class (Longadge et al., 2013). In this case, there are far more non-storm intervals than storm intervals. Following McGranaghan et al. (2018), we define the cost of mis-classifying each class separately. This is done through the weight ratio (W_{storm}: $W_{no\ storm}$). Increasing the W_{storm} increases the frequency at which the SVM predicts a storm and it follows that it predicts "no storm" at a reduced frequency, as seen in the left column of Figure 3. The same results have been normalized to reveal more clearly how varying the class weights effects the predictions, shown in the right column of Figure 3. In this work we have varied W_{storm} and kept $W_{no\ storm}$ constant at 1. A user of the SVM method for forecasting may wish to tune the class weight ratio to give an appropriate ratio of false alarms and hit rate dependent on their needs.



Taylor diagram on entire dataset.

Taylor diagram on observed storms

Figure 6. Taylor diagrams comparing the Analogue Ensemble (AnEn) median hindcast to climatology and 27 days recurrence for a 3 h lead-time. The diagrams summarize the root mean square deviation (RMSD) (nT), correlation coefficient and standard deviation (nT) of the hindcast s. (a) The entire aa_H data set. (b) Restricted to the period following observed storms.

3.4. Bench Marking

Similar to Owens et al. (2017a) and following the recommendation of Liemohn et al. (2018), we use a benchmark hindcast method to distinguish between times when the studied hindcasts perform poorly and times when conditions are intrinsically more difficult to predict. For this purpose, we use climatology defined by the mean intensity of the entire data set or, for a probabilistic hindcast of storms, the fraction of measurements in the entire data set which qualify as storm events. These values are 17.5 nT and 10% respectively.

4. Results

4.1. AnEn Deterministic Intensity Hindcast

In the present section we consider the deterministic performance of the AnEn by reducing the ensemble to the median value. In this and Section 4.2 we present results for three subsets of the aa_H data set. The subsets are: the entire aa_H data set; the occasions on which a storm was observed at a 3 h lead time; the occasions on which a storm was predicted by the AnEn median at a 3 h lead time. The second subsets includes only true positives, whereas the third subset includes false negatives and true positives. There are many more non-storm events than storm events in the data set so a hindcast always predicting no storm would fare well. These subsets of aa_H help to distinguish whether a hindcast method has any predictive power of storm events.

Figure 4 shows the MAE of the deterministic hindcasts of aa_H intensity compared to observations for lead times up to 300 h (12.5 days). Figure 4a shows the MAE when the hindcasts are initiated from every time step in the aa_H data set. The general pattern is for the AnEn median to produce the lowest MAE, followed by 27 days recurrence and climatology with the highest. While the MAE in the 27 days recurrence and climatology are relatively constant with lead-time, AnEn clearly displays higher accuracy for shorter lead times until it plateaus at approximately 50 h lead time. This suggests that the usefulness of information in the preceding 24 h to t_0 is greatest for short lead times. Figure 4b shows the error of hindcasts on which the point immediately after t_0 is classed as a storm, as defined by a threshold of 40.1 nT. All hindcast methods have a high MAE for short lead times which drops off for longer lead times. At long lead times, approximately the same order of accuracy of the hindcasts exists for this storm data set as for the whole data set. At shorter lead times, the storms are in progress, and thus the observed aa_H is high, and the same percentage error leads to a higher MAE than non-storm times. This issue, as well as that of storm conditions generally being more difficult to predict than quieter times, can be addressed by computing the skill of the hindcasts relative to a reference forecast. In essence, it allows discrimination between poor forecasts and periods, which are inherently difficult to forecast.

Skill is computed as:

$$\text{skill} = 1 - \frac{\textit{forecast error}}{\textit{reference error}}, \tag{1}$$

Thus, skill can vary between $-\infty$ and 1, where a more positive value is more skillfull, and zero is identical performance to the reference hindcast. Figure 5 shows the skill of the AnEn median and 27 days recurrence relative to climatology. This is done for the whole data set in Figure 5a. Both AnEn median and 27 days recurrence have

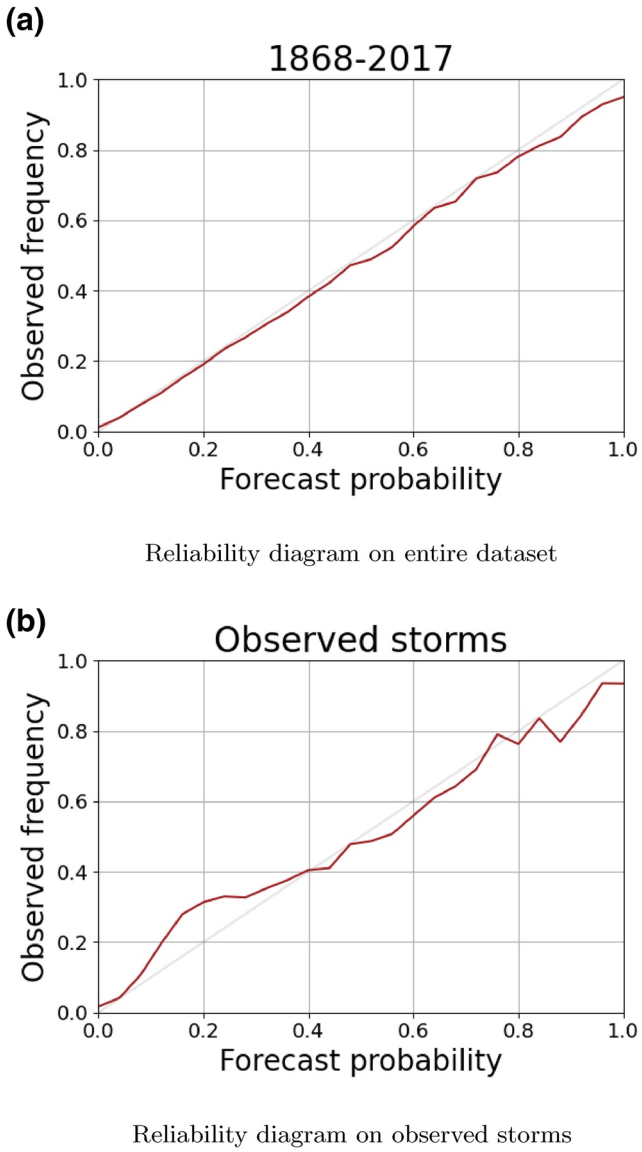


Figure 7. A reliability diagram of the analogue ensemble dichotomous storm hindcast for a 3 h lead-ime. The gray line represents the path of a perfectly reliable hindcast, wherein events are forecast with a probability equal to the observed occurrence rate. (a) The entire aa_H data set. (b) Restricted to the period following observed storms.

positive skill for all lead times. AnEn median achieves substantially higher skill, especially for shorter lead times. Figure 5b makes the same comparisons considering only the time periods immediately following observed storm onsets. We again see that AnEn median has positive skill, however skill is reduced by approximately 10% compared to the whole data set.

Figure 6 shows Taylor diagrams (Taylor, 2001; Owens, 2018) that summarize the performance of a hindcast in terms of three metrics, visualized on a single plot. These metrics are the standard deviation of the hindcast, linear correlation coefficient between hindcast and observed intensities, and the centered root-mean-squared distance (RMSD) between hindcast and observed intensities. These three metrics provide measures of agreement in both statistical terms (standard deviations) and the correspondence on a point-by-point basis (correlation and RMSD). A perfect hindcast would lie on the red dot, having a hindcast standard deviation matching that observed, correlation coefficient of 1, and centered RMSD of 0. Put simply, the further a hindcast lies from the red dot, the worse the forecast is, and the direction of displacement can help diagnose the problem with the forecast.

Figure 6a shows the three hindcast types for the whole 150 year period of aa_H data for a 3 h lead time. AnEn median provides the best hindcast by two of the three metrics considered but has a smaller and reduced standard deviation than both climatology and 27 days recurrence compared to the observations (by the construction, this is expected: both benchmarks are direct, unaveraged samples of the observations against which they are tested. Conversely, by taking the median of the AnEn, the variability will be reduced). Figure 6b shows the hindcasts run only for observed storm onsets. The general pattern is similar to that of when the hindcast is run on the whole data set.

4.2. AnEn Probabilistic Dichotomous Hindcast

In Section 4.1, the AnEn was reduced to a deterministic hindcast of intensity by considering only the ensemble median. But the AnEn can be used as a probabilistic hindcast. We here consider the probabilistic hindcast of (dichotomous) event occurrence, in this case the occurrence/non-occurrence of storms, by considering all the ensemble members together to form a probability distribution of future evolution. While a deterministic intensity hindcast looks to minimize the error of the prediction, a probabilistic dichotomous hindcast aims to predict event occurrence at the observed frequency. That is to say if a hindcast makes a prediction with $x\%$ certainty it is said to be reliable if, on average, an event is subsequently observed $x\%$ of the time. Systematic bias in hindcast probability can be quantified with a reliability diagram (Jolliffe & Stephenson, 2003; Sharpe & Murray, 2017) which compares predicted and observed probabilities. A perfectly reliable hindcast would follow the $y = x$ line, as shown in Figure 7 by the light gray line. A forecast giving a reliability curve below this line shows overestimation of event likelihood and a reliability curve over the line shows underestimation of the likelihood of an event. Figure 7a shows AnEn hindcast of storms from all data points in the aa_H data set for a hindcast lead-time of 3 h. On the whole, the curve fits well to $y = x$ with a slight overestimate of storm probability for larger values of hindcast probability. When considering only known storm events, the AnEn is less reliable, as shown Figure 7b. While the curve largely follows the $y = x$ line, there is an underestimate of storms for low hindcast probability and an overestimation for high.

This underestimate may be an indicator that there is insufficient information in the observed time-series leading up to t_0 in order to differentiate between storms and not storms, that is, many of the analogs found for the build up to a storm may be associated with only a small increase in aa_H because there are simply

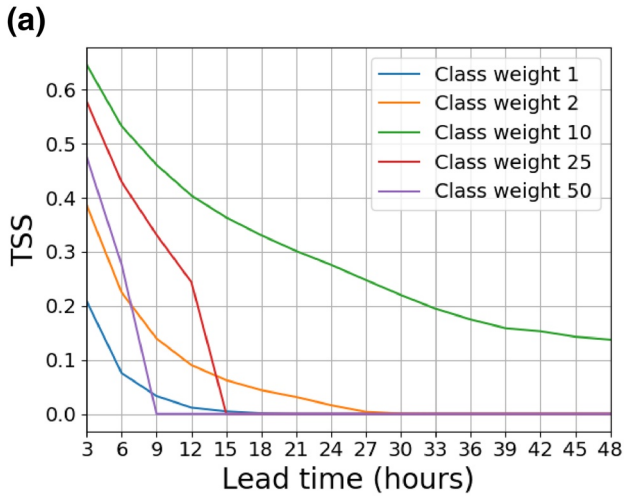


Figure 8. True Skill Score (TSS) for lead times of 3–48 h (a) shows the TSS for the Support Vector Machine (SVM) with a range of class weights. We see that using class weight 10 gives the best skill. (b) TSS for Analogue Ensemble (AnEn) median, 27 days recurrence and SVM with class weight 10. We see that the SVM skill exceeds that of the other hindcasts.

many more instances of smaller variations than larger. This would bias the hindcast toward predicting smaller storms and thus underestimating the probability of an event.

4.3. SVM Classification

To evaluate the SVM and AnEn for storm classification we need a metric that is robust to class imbalance. This is because using a storm definition of the 90th percentile means we have nine non-storm events for every storm, so a prediction method that always predicts non-storm would do very well under many metrics. The True Skill Score (TSS) is a combination of TP, FP and FN only, meaning that it can handle imbalanced classes, though it neglects a model's ability to correctly predict non events, which can be valuable in its own right. TSS has been recommended and used in the space weather community (e.g., Bloomfield et al., 2012; McGranaghan et al., 2018).

TSS is defined as

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN} \tag{2}$$

which gives a score between −inf and 1 where 0 is a hindcast with no skill and 1 is a perfect hindcast.

TSS has been computed for the SVM, AnEn median and 27 days recurrence in Figure 8. Figure 8a shows TSS for SVMs with different class weights. A unique SVM has been trained for each value of lead-time. We see that using class weight 10 gives the best performance with positive, reducing skill for the full 48 h. The other SVMs perform considerably worse, particularly for lead times greater than 3 h. SVMs with class weight of 1 and 2 end up predicting no storm events will occur at longer lead times and SVMs with class weights 25 and 50 predict storms always occur at longer lead times. SVM with class weight 10 seems to strike a good balance, as it approaches the proportions of storm and non-storm events in the data set.

In Figure 8b we compare the TSS of SVM class weight 10 to TSS of AnEn median and 27 days recurrence. Both SVM and AnEn median have a similar shape of diminishing skill with lead-time, however SVM has a far superior TSS at all lead times considered. The TSS of 27 days recurrence is a flat line since its lead-time is, in essence, always 27 days 27 days recurrence exceeds AnEn median at 9 h and longer and is approximately equivalent to SVM at 48 h. It suggests that the AnEn median does not have predictive power for the storm class at longer lead times and quickly goes back to predicting quiet-time.

Different forecast applications will have different tolerances for false alarms and missed events. A limitation of TSS is that it treats FP and FN the same and does not give useful information for users with an unbalanced tolerance. To accommodate this, and as a further comparison of the hindcasts, a Cost/Loss analysis (Murphy, 1977; Richardson, 2000; Owens & Riley, 2017) is implemented in Figure 9. A space-weather example of how a Cost/Loss analysis is carried out is shown in Figure 7 of Owens and Riley (2017). In short, C is the economic cost associated with taking mitigating action when an event is predicted (whether or not it actually occurs) and L is the economic loss suffered due to damage if no mitigating action is taken when needed. For a deterministic method, such as the SVM, each time a storm is predicted will incur a cost C. Each time a storm is not predicted but a storm occurs a loss L is incurred. If no storm is predicted and no storm occurs then no expense in incurred. By considering some time interval, the total expense can be computed by summing C and L.

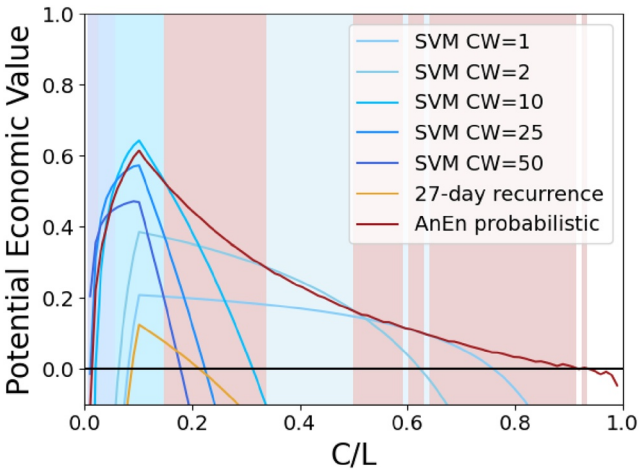


Figure 9. A Cost/Loss analysis showing the potential economic value of the hindcasts relative to value of a perfect hindcast (PEV = 1) and climatology (PEV = 0). The shaded areas indicate which hindcast type has the highest PEV for that C/L ratio. Negative values of potential economic value are shown only down to −0.1.

A particular forecast application will have a C/L ratio in the domain (0,1). This is because a C/L of 0 would mean it is most cost effective to take constant mitigating action and a C/L of 1 or more means that mitigating action is never cost effective. In either case, no forecast would be helpful. The power of a Cost/Loss analysis is that it allows us to evaluate our methods for the entire range of potential forecast end users without specific knowledge of the forecast application requirements. End users can then easily interpret whether our methods fit their situation.

For a probabilistic forecast, a similar process is applied with the difference that action is taken only when the forecast probability exceeds C/L. See Owens et al. (2020) for more information.

Once total costs have been calculated, the potential economic value (PEV) is given by

$$PEV = \frac{E_C - E}{E_C - E_0}, \tag{3}$$

where E_C is the total expense of using a probabilistic climatological forecast, E is the total cost of the forecast under consideration and E_0 is the total cost of a perfect forecast. The PEV of a forecast is therefore equivalent to climatology where PEV = 0 and to a perfect forecast where PEV = 1. Note that a user's Cost and Loss do not need to be computed in financial terms, only the ratio of the two values is necessary: high C/L suggests that false alarms should be avoided, whereas low C/L suggests missed events would be more problematic.

Figure 9 shows the PEV of the SVM with a range of class weights (CW), probabilistic AnEn and 27 days recurrence. Here a deterministic Cost/Loss analysis has been implemented for SVM and 27-days recurrence, and a probabilistic Cost/Loss for AnEn. The shaded regions indicate which hindcast has the highest PEV for that Cost/Loss ratio. The probabilistic AnEn has the highest PEV for the majority of the Cost/Loss domain although SVM has higher PEV for lower Cost/Loss ratios. It is possible that an increased resolution in the scan of class weights would bring the SVM out on top for a larger part of the domain. However certain users may appreciate that the hindcasts generally have a similar PEV for parts of the Cost/Loss domain and will find it more valuable to have the probabilistic hindcast of the AnEn. It also highlights that the "best" hindcast is dependent on the context in which it is to be employed.

5. Future Directions

There are a number of possible ways the forecast schemes presented here could be improved in the future. By taking the fraction of ensemble members which result in a storm to be the AnEn hindcast probability of a storm we are implicitly assuming that the analogs form a single distribution. This potentially throws away information about different modes of behavior. Clustering ensemble members together using K-means clustering is a way in which we could use the data to extract a number of possible future scenarios. An example is shown in Figure 10. The observed storm peaks at $t = 21$ h, however this behavior is not captured by the median of the ensemble members or easily visible amongst the gray ensemble member lines. However the scenario in which the storm has a late peak is picked out as a possible mode of behavior by the red cluster in the right of Figure 10, identified by K-means clustering. Here, the clustering algorithm aims to minimize the sum of the square error between the ensemble member and the cluster it is in. The number of clusters has been chosen by using an "elbow plot" which identifies appropriate K values by minimizing both the sum of square errors and the number of clusters.

6. Discussion and Conclusions

This study has considered the effectiveness of two pattern-matching methods in hindcasting the aa_H index. These are an AnEn and a SVM. We have additionally considered the 27 days recurrence hindcast for context. AnEn and 27 days recurrence can be used as intensity hindcasts and AnEn can also give a probability

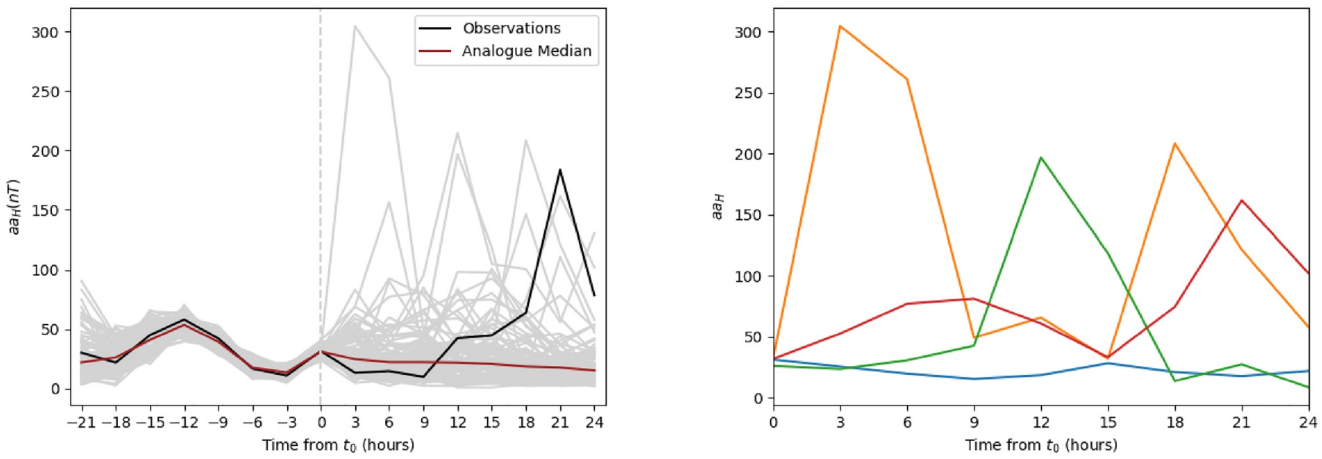


Figure 10. (Left) An event from 27/2/1997 with an analogue ensemble (AnEn) hindcast. (Right) Clusters from the ensemble members using K-means clustering.

distribution for dichotomous-event hindcast. SVM has only been implemented as a deterministic dichotomous-event hindcast.

Reducing the AnEn to a deterministic intensity hindcast by taking the median value, it outperformed the benchmark of climatology both when applied to the whole aa_H data set and limited only to observed storm onsets. AnEn clearly outperformed the benchmark for absolute error and skill for lead times up to a week. 27 days recurrence outperformed the benchmark but did not perform as well as AnEn.

When considering the AnEn as a probabilistic hindcast of storm occurrence, it was found to be highly reliable when hindcasting each data point in the aa_H data set, in that the predicted probability closely matches the observed frequency of events. Reliability was found to drop slightly when considering only storm events. In particular, the AnEn underestimates storms when it had a low certainty of a storm and overestimates the probability of a storm when it was reasonably certain. The underestimation may be an indicator that there is insufficient information in the observed time-series leading up to t_0 in order to differentiate between storms and not storms. That is, many of the analogs found for the build up to a storm may be associated with only a small increase in aa_H because there are simply many more instances of smaller variations than larger. This would bias the hindcast toward predicting smaller storms and thus underestimating the probability of an event.

Finally, an SVM was implemented for a range of class weights and compared to AnEn and 27 days recurrence using TSS and a Cost/Loss analysis. The SVM was more skillfull than AnEn by TSS, though neither hindcast had a conclusively higher potential economic value across the Cost/Loss domain. It is likely that the best method for a user will depend on their individual circumstances.

Data Availability Statement

The aa_H data is available at https://www.swsc-journal.org/articles/swsc/olm/2018/01/swsc180022/swsc180022-2-olm.txt. Code for AnEn is available at https://doi.org/10.5281/zenodo.4604487. Code for SVM is available at https://doi.org/10.5281/zenodo.4604485 which includes the code for splitting the data into train and test sets. A data file containing a list of storms in aa_H is also here.

Acknowledgments
The authors thank the National Environmental Research Council (NERC) for funding this work under grants NE/L002566/1 and NE/P016928/1.

References

Ayala Solares, J. R., Wei, H. L., Boynton, R. J., Walker, S. N., & Billings, S. A. (2016). Modeling and prediction of global magnetic disturbance in near-Earth space: A case study for Kp index using NARX models. *Space Weather*, *14*(10), 899–916. https://doi.org/10.1002/2016SW001463

Bartels, J. (1932). Terrestrial-magnetic activity and its relations to solar phenomena. *Journal of Geophysical Research*, *37*(1), 1–52. https://doi.org/10.1029/te037i001p00001

Bartels, J. (1934). Twenty-seven day recurrences in terrestrial-magnetic and solar activity, 1923-1933. *Journal of Geophysical Research*, *39*(3), 201–202. https://doi.org/10.1029/te039i003p00201

Bartels, J., Heck, N. H., & Johnston, H. F. (1939). The three-hour-range index measuring geomagnetic activity. *Journal of Geophysical Research*, *44*(4), 411–424. https://doi.org/10.1029/te044i004p00411

Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. (2012). Toward reliable benchmarking of solar flare forecasting methods. *Acta Pathologica Japonica*, *747*(2), L41. https://doi.org/10.1088/2041-8205/747/2/L41

Boteler, D. H. (1994). Geomagnetically induced currents: Present knowledge and future research. *IEEE Transactions on Power Delivery*, *9*(1), 50–58. https://doi.org/10.1109/61.277679

Bubenik, D. M., & Fraser-Smith, A. C. (1977). Evidence for strong artificial components in the equivalent linear amplitude geomagnetic indices. *Journal of Geophysical Research*, *82*(19), 2875–2878. https://doi.org/10.1029/ja082i019p02875

Buonsanto, M. J. (1999). Ionospheric storms - A review. *Space Science Reviews*, *88*(3–4), 563–601. https://doi.org/10.1023/a:1005107532631

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167. https://doi.org/10.1023/A:1009715923555

Burton, R. K., McPherron, R. L., & Russell, C. T. (1975). An empirical relationship between interplanetary conditions and Dst. *Journal of Geophysical Research*, *80*(31), 4204–4214. https://doi.org/10.1029/ja080i031p04204

Camporeale, E. (2019). The challenge of machine learning in space weather: Nowcasting and forecasting. *Space Weather*, *17*, 1166–1207. https://doi.org/10.1029/2018sw002061

Cannon, P., Angling, M., Barclay, L., Curry, C., Dyer, C., Edwards, R., & Underwood, C. (2013). Extreme space weather: Impacts on engineered systems and Infrastructures (Vol. 70). Royal Academy of Engineering. ISBN:1-903496-96-9.

Chandorkar, M., Camporeale, E., & Wing, S. (2017). Probabilistic forecasting of the disturbance storm time index: An autoregressive Gaussian process approach. *Space Weather*, *15*(8), 1004–1019. https://doi.org/10.1002/2017SW001627

Chapman, S. C., Horne, R. B., & Watkins, N. W. (2020). Using the index over the last 14 solar cycles to characterize extreme geomagnetic activity. *Geophysical Research Letters*, *47*(3), e2019GL086524. https://doi.org/10.1029/2019GL086524

Chree, C., & Stagg, M. (1928). Recurrence phenomena in terrestrial magnetism (Vol. 227, pp. 21–62). Royal Society. https://doi.org/10.1098/rsta.1928.0002

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297. https://doi.org/10.1007/BF00994018

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, *14*(3), 326–334. https://doi.org/10.1109/PGEC.1965.264137

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., & Searight, K. (2013). Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, *141*(10), 3498–3516. https://doi.org/10.1175/mwr-d-12-00281.1

Eastwood, J. P., Biffis, E., Hapgood, M. A., Green, L., Bisi, M. M., Bentley, R. D., et al. (2017). The economic impact of space weather: Where do we stand? *Risk Analysis*, *37*(2), 206–218. https://doi.org/10.1111/risa.12765

Eastwood, J. P., Nakamura, R., Turc, L., Mejnertsen, L., & Hesse, M. (2017). The Scientific Foundations of Forecasting Magnetospheric Space Weather. *Space Science Reviews*, *212*(3–4), 1221–1252. https://doi.org/10.1007/s11214-017-0399-8

Ganushkina, N. Y., Liemohn, M. W., Dubyagin, S., Daglis, I. A., Dandouras, I., De Zeeuw, D. L., et al. (2015). Defining and resolving current systems in geospace. *Annales Geophysicae*, *33*(11), 1369–1402. https://doi.org/10.5194/angeo-33-1369-2015

Gonzalez, W. D., Joselyn, J. A., Kamide, Y., Kroehl, H. W., Rostoker, G., Tsurutani, B. T., & Vasyliunas, V. M. (1994). What is a geomagnetic storm? *Journal of Geophysical Research*, *99*(A4), 5771. https://doi.org/10.1029/93ja02867

Gu, Y., Wei, H. L., Boynton, R. J., Walker, S. N., & Balikhin, M. A. (2019). System identification and data-driven forecasting of AE index and prediction uncertainty analysis using a new cloud-NARX model. *Journal of Geophysical Research: Space Physics*, *124*(1), 248–263. https://doi.org/10.1029/2018JA025957

Haines, C., Owens, M. J., Barnard, L., Lockwood, M., & Ruffenach, A. (2019). *The variation of geomagnetic storm duration with intensity* (p. 154). https://doi.org/10.1007/s11207-019-1546-zThe

Henley, E. M., & Pope, E. C. D. (2017). Cost-loss analysis of ensemble solar wind forecasting: Space weather use of terrestrial weather tools. *Space Weather*, *15*(12), 1562–1566. https://doi.org/10.1002/2017SW001758

Hutchinson, J. A., Wright, D. M., & Milan, S. E. (2011). Geomagnetic storms over the last solar cycle: A superposed epoch analysis. *Journal of Geophysical Research*, *116*(9), A09211. https://doi.org/10.1029/2011JA016463

Jolliffe, I., & Stephenson, D. (2003). Forecast verification: A practioners guide in atmospheric science (Vol. 4).

Kilpua, E. K. J., Olspert, N., Grigorievskiy, A., Käpylä, M. J., Tanskanen, E. I., Miyahara, H., et al. (2015). Statistical study of strong and extreme geomagnetic disturbances and solar cycle characteristics. *Acta Pathologica Japonica*, *806*(2), 272. https://doi.org/10.1088/0004-637X/806/2/272

Knipp, D. (2016). Advances in space weather ensemble forecasting. *Space Weather*, *14*(2), 113–135. https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2016SW001366

Knipp, D. J., Hapgood, M. A., & Welling, D. (2018). Communicating uncertainty and reliability in space weather data, models, and applications. *Space Weather*, *16*(10), 1453–1454. https://doi.org/10.1029/2018SW002083

Koskinen, H. E. J., Baker, D. N., Balogh, A., Gombosi, T., Veronig, A., & von Steiger, R. (2017). Achievements and challenges in the science of space weather. *Space Science Reviews*, *212*(3–4), 1137–1157. https://doi.org/10.1007/s11214-017-0390-4

Liemohn, M. W., McCollough, J. P., Jordanova, V. K., Ngwira, C. M., Morley, S. K., Cid, C., et al. (2018). Model evaluation guidelines for geomagnetic index predictions. *Space Weather*, *16*(12), 2079–2102. https://doi.org/10.1029/2018SW002067

Lockwood, M. (2013). Reconstruction and prediction of variations in the open solar magnetic flux and interplanetary conditions. *Living Reviews in Solar Physics*, *10*, 4. https://doi.org/10.12942/lrsp-2013-4

Lockwood, M. (2019). Does adding solar wind poynting flux improve the optimum solar wind-magnetosphere coupling function? *Journal of Geophysical Research: Space Physics*, *124*(7), 5498–5515. https://doi.org/10.1029/2019JA026639

Lockwood, M., Chambodut, A., Barnard, L. A., Owens, M. J., Clarke, E., & Mendel, V. (2018). *A homogeneous aa index: 1. Secular variation* (pp. A53–A57), Space Weather and Space Climate. https://doi.org/10.1051/swsc/2018038

Lockwood, M., Chambodut, A., Finch, I. D., Barnard, L. A., Owens, M. J., & Haines, C. (2019). Time-of-day/time-of-year response functions of planetary geomagnetic indices. *Journal of Geophysical Research: Space Climate*, *9*, A20. https://doi.org/10.1051/swsc/2019017

Lockwood, M., Finch, I. D., Chambodut, A., Barnard, L. A., Owens, M. J., & Clarke, E. (2018). A homogeneous aa index: 2. Hemispheric asymmetries and the equinoctial variation. *Journal of Space Weather and Space Climate*, *8*(A58), A58. https://doi.org/10.1051/swsc/2018044

Longadge, R., Dongre, S., & Malik, L. (2013). Class imbalance problem in data mining: Review. *Internation Journal of Computer Science and Network*, *2*(1).

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

Mayaud, P.-N. (1971). Une mesure planétaire d'activité magnetique, basée sur deux observatoires antipodaux. *Annales Geophysicae*, *27*, 67–70. https://doi.org/10.1002/9781118663837

Mayaud, P. N. (1981). Planetary indices derived from K indices (Kp, am, and aa). In P.N. Mayaud (Ed.), *Derivation, meaning, and use of geomagnetic indices*. (pp. 40–85). Geophysical Monograph Series. https://doi.org/10.1002/9781118663837.ch5

McGranaghan, R. M., Mannucci, A. J., Wilson, B., Mattmann, C. A., & Chadwick, R. (2018). New capabilities for prediction of high-latitude ionospheric scintillation: A novel approach with machine learning. *Space Weather*, *16*(11), 1817–1846. https://doi.org/10.1029/2018SW002018

Murphy, A. H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *American Meteorological Society*, *105*, 803–816. https://doi.org/10.1175/1520-0493(1977)105<0803:tvocca>2.0.co;2

O'Brien, T., & McPherron, R. L. (2000). Forecasting the ring current index Dst in real time. *Journal of Atmospheric and Solar-Terrestrial Physics*, *62*(14), 1295–1299. https://doi.org/10.1016/s1364-6826(00)00072-9

Oughton, E. J., Skelton, A., Horne, R. B., Thomson, A. W. P., & Gaunt, C. T. (2017). Quantifying the daily economic impact of extreme space weather due to failure in electricity transmission infrastructure. *Space Weather*, *15*(1), 65–83. https://doi.org/10.1002/2016SW001491

Owens, M. J. (2018). Time-window approaches to space-weather forecast metrics: A solar wind case study. *Space Weather*, *16*(11), 1847–1861. https://doi.org/10.1029/2018SW002059

Owens, M. J., Challen, R., Methven, J., Henley, E., & Jackson, D. R. (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., Lockwood, M., & Barnard, L. A. (2020). The value of CME arrival-time forecasts for space weather mitigation. *Space Weather*, *18*, e2020SW002507. https://doi.org/10.1029/2020sw002507

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

Owens, M. J., Riley, P., & Horbury, T. S. (2017a). Probabilistic solar wind and geomagnetic forecasting using an analogue ensemble or "similar day" approach. *Solar Physics*, *292*(5), 1–16. https://doi.org/10.1007/s11207-017-1090-7

Owens, M. J., Riley, P., & Horbury, T. S. (2017b). The role of empirical space-weather models (in a world of physics-based numerical simulations). *Proceedings of the International Astronomical Union*, *13*(S335), 254–257. https://doi.org/10.1017/S1743921317007128

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pirjola, R. (2000). Geomagnetically induced currents during magnetic storms. *IEEE Transactions on Plasma Science*, *28*(6), 1867–1873. https://doi.org/10.1109/27.902215

Pulkkinen, A., Rastätter, L., Kuznetsova, M., Singer, H., Balch, C., Weimer, D., et al. (2013). Community-wide validation of geospace model ground magnetic field perturbation predictions to support model transition to operations. *Space Weather*, *11*(6), 369–385. https://doi.org/10.1002/swe.20056

Pulkkinen, T. (2007). Space weather: Terrestrial perspective. *Living Reviews in Solar Physics*, *4*, 1. https://doi.org/10.12942/lrsp-2007-1

Richardson, D. S. (2000). Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *126*, 649–667. https://doi.org/10.1002/qj.49712656313

Riley, P., Baker, D., Liu, Y. D., Verronen, P., Singer, H., & Güdel, M. (2018). Extreme space weather events: From cradle to grave. *Space Science Reviews*, *214*, 21. https://doi.org/10.1007/s11214-017-0456-3

Riley, P., Ben-Nun, M., Linker, J. A., Owens, M. J., & Horbury, T. S. (2017). Forecasting the properties of the solar wind using simple pattern recognition. *Space Weather*, *15*(3), 526–540. https://doi.org/10.1002/2016SW001589

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

Schrijver, C. J., Dobbins, R., Murtagh, W., & Petrinec, S. M. (2014). Assessing the impact of space weather on the electric power grid based on insurance claims for industrial electrical equipment. *Space Weather*, *12*(7), 487–498. https://doi.org/10.1002/2014SW001066

Sharpe, M. A., & Murray, S. A. (2017). Verification of space weather forecasts issued by the met office space weather operations centre. *Space Weather*, *15*(10), 1383–1395. https://doi.org/10.1002/2017SW001683

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, *35*(6), 2769–2794. https://doi.org/10.1214/009053607000000505

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, *106*, 7183–7192. https://doi.org/10.1029/2000JD900719

Tóth, G., Sokolov, I. V., Gombosi, T. I., Chesney, D. R., Clauer, C. R., De Zeeuw, D. L., et al. (2005). Space weather modeling framework: A new tool for the space science community. *Journal of Geophysical Research*, *110*(A12), 1–21. https://doi.org/10.1029/2005JA011126

Tóth, G., van der Holst, B., Sokolov, I. V., De Zeeuw, D. L., Gombosi, T. I., Fang, F., et al. (2012). Adaptive numerical algorithms in space weather modeling. *Journal of Computational Physics*, *231*(3), 870–903. https://doi.org/10.1016/j.jcp.2011.02.006

Van den Dool, H. M. (1989). A new look at weather forecasting through analogs. *Monthly Weather Review*, *117*(10), 2230–2247. https://doi.org/10.1175/1520-0493(1989)117<2230:anlawf>2.0.co;2

Vassiliadis, D., & Klimas, A. J. (1995). On the uniqueness of linear moving-average filters for the solar wind-auroral geomagnetic activity coupling. *Journal of Geophysical Research*, *100*(A4), 5637–5641. https://doi.org/10.1029/94ja03303

Vassiliadis, D., Klimas, A. J., Baker, D. N., & Roberts, D. A. (1995). A description of the solar wind-magnetosphere coupling based on nonlinear filters. *Journal of Geophysical Research*, *100*(A3), 3495–3512. https://doi.org/10.1029/94ja02725

Vennerstrom, S., Lefevre, L., Dumbović, M., Crosby, N., Malandraki, O., Patsou, I., et al. (2016). Extreme geomagnetic storms – 1868 – 2010. *Solar physics*, *291*, 1447–1481. https://doi.org/10.1007/s11207-016-0897-y

Watari, S. (2011). Forecast of recurrent geomagnetic storms. *Advances in Space Research*, *47*(12), 2162–2171. https://doi.org/10.1016/j.asr.2010.07.029

Wing, S., Johnson, J. R., Jen, J., Meng, C.-I., Sibeck, D. G., Bechtold, K., et al. (2005). Kp forecast models. *Journal of Geophysical Research*, *110*(A4), 1–14. https://doi.org/10.1029/2004JA010500

Zhang, X.-Y., & Moldwin, M. B. (2015). Probabilistic forecasting analysis of geomagnetic indices for southward IMF events. *Space Weather*, *13*(3), 130–140. https://doi.org/10.1002/2014SW001113