

# Space Weather

## RESEARCH ARTICLE

10.1029/2021SW002788

### Special Section:

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

### Key Points:

- Three neural network variants can use solar wind inputs to provide skillful and reliable probabilistic forecasts of large dB/dt in the UK
- The forecast skill/reliability increases with forecast horizon, maximizing at a horizon of 180 min
- Increasing the volume of input solar wind input data without increasing the model complexity does not boost performance

### Correspondence to:

A. W. Smith,  
[andy.w.smith@ucl.ac.uk](mailto:andy.w.smith@ucl.ac.uk)

### Citation:

Smith, A. W., Forsyth, C., Rae, I. J., Garton, T. M., Bloch, T., Jackman, C. M., & 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, e2021SW002788. <https://doi.org/10.1029/2021SW002788>








Received 22 APR 2021

Accepted 1 AUG 2021

© 2021. The Authors.

This is an open access article under the terms of the [Creative Commons Attribution License](#), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

## Forecasting the Probability of Large Rates of Change of the Geomagnetic Field in the UK: Timescales, Horizons, and Thresholds

A. W. Smith<sup>1</sup> , C. Forsyth<sup>1</sup> , I. J. Rae<sup>2</sup> , T. M. Garton<sup>3</sup> , T. Bloch<sup>4</sup> , C. M. Jackman<sup>5</sup> , and M. Bakrania<sup>1</sup> 

<sup>1</sup>Mullard Space Science Laboratory, UCL, Dorking, UK, <sup>2</sup>Department of Mathematics, Physics and Electrical Engineering, Northumbria University, Newcastle upon Tyne, UK, <sup>3</sup>Space Environment Physics Group, Department of Physics and Astronomy, University of Southampton, Southampton, UK, <sup>4</sup>Department of Meteorology, University of Reading, Reading, UK, <sup>5</sup>School of Cosmic Physics, DIAS Dunsink Observatory, Dublin Institute for Advanced Studies, Dublin, Ireland

**Abstract** Large geomagnetically induced currents (GICs) pose a risk to ground based infrastructure such as power networks. Large GICs may be induced when the rate of change of the ground magnetic field is significantly elevated. We assess the ability of three different machine learning model architectures to process the time history of the incoming solar wind and provide a probabilistic forecast as to whether the rate of change of the ground magnetic field will exceed specific high thresholds at a location in the UK. The three models tested represent feed forward, convolutional and recurrent neural networks. We find all three models are reliable and skillful, with Brier skill scores, receiver-operating characteristic scores and precision-recall scores of approximately 0.25, 0.95 and 0.45, respectively. When evaluated during two example magnetospheric storms we find that all scores increase significantly, indicating that the models work better during active intervals. The models perform excellently through the majority of the storms, however they do not fully capture the ground response around the initial sudden commencements. We attribute this to the use of propagated solar wind data not allowing the models notice to forecast impulsive phenomenon. Increasing the volume of solar wind data provided to the models does not produce appreciable increases in model performance, possibly due to the fixed model structures and limited training data. However, increasing the horizon of the forecast from 30 min to 3 h increases the performance of the models, presumably as the models need not be as precise about timing.

**Plain Language Summary** Geomagnetically induced currents (GICs) are external electrical currents that can be created in power lines and pipe networks as a result of variability in near-Earth space. GICs are likely to be created when the ground magnetic field is changing rapidly. We test three different machine learning models to see whether they can predict if the ground magnetic field at a site in the UK is going to be highly variable in the future. We find that the models all perform excellently, providing useful forecasts as to when the rate of change of the magnetic field will be elevated.

## 1. Introduction

Geomagnetically induced currents (GICs) are a primary space weather hazard, caused by intense dynamical processes in near-Earth space. The generation of GICs in grounded infrastructure, such as pipelines or power networks, can damage components during intervals of exceptionally large GICs (e.g., Bolduc, 2002; Kappenman, 2005; Rajput et al., 2020). Some of the risks associated with large GICs may be mitigated with sufficient warning, and so forecasting when such intervals are likely to occur is a critical endeavor.

Space weather events that generate extremely large GICs are thankfully rare. However, the rarity of these events coupled with the relative sparsity of direct GIC measurements means that a proxy measurement is often necessary to provide a sufficient historical data set with which to train advanced forecasting models. The magnitude of GICs is predominantly dependent upon three factors: (a) the rate of change of the magnetic field, (b) the local subsurface conductivity and (c) the relative geometry and properties of the infrastructure (Beggan, 2015; Boteler, 2014; Divett et al., 2018; Mac Manus et al., 2017; Thomson et al., 2005; Viljanen et al., 2013). Because of the dependence upon the geology and the local field variations, significantly

different driving electric fields or GICs have been observed on geographical scales below  $\sim 100$  km (e.g., Bedrosian & Love, 2015; Dimmock et al., 2020; Ngwira et al., 2015). Overall however, larger rates of change of the magnetic field tend to be linked to larger GICs (Bolduc et al., 1998; Mac Manus et al., 2017; Rodger et al., 2017). For this reason, and the relative abundance of magnetometer observations, the rate of the change of the magnetic field has been used extensively as a proxy measurement to allow statistical investigation of how and when GICs may occur (e.g., Carter et al., 2015; Freeman et al., 2019; Kelly et al., 2017; Oliveira et al., 2018; Rogers et al., 2020; Smith et al., 2019; Turnbull et al., 2009; Thomson et al., 2011; Viljanen et al., 2001).

Physically, there are a multitude of different processes in near-Earth space that can result in large variations in the magnetic field (as measured on the ground), and consequently GICs (e.g., Rogers et al., 2020; Tsurutani & Hajra, 2021). These processes are ultimately driven by the interaction between the incident solar wind and its coupling to the magnetosphere. Some processes are an almost instantaneous response to structures in the solar wind, while others require specific conditions, such as the interplanetary magnetic field (IMF) to be directed southward, potentially for an extended period of time.

The fastest physical response to the solar wind is found during sudden commencements (SCs) (Curto et al., 2007). These sharp, predominantly northward deflections of the ground magnetic field are observed globally, and are related to rapid increases in solar wind dynamic pressure at the nose of the magnetosphere, the impact of a solar wind interplanetary shock for example, (Lühr et al., 2009; Takeuchi et al., 2002). The morphological signature of an SC is complex, with particularly strong latitudinal variations (Araki, 1994; Shinbori et al., 2012; Smith et al., 2021). At low latitudes, ground observations are dominated by a compressional signature that maximizes at noon and decreases toward midnight (Russell et al., 1992). Meanwhile at higher latitudes the compressional component couples to shear Alfvén waves and field aligned resonances (Southwood & Kivelson, 1990). While the magnitude of the SC deflection has been found to increase with latitude (Fiori et al., 2014; Smith et al., 2021), significant associated GICs have been recorded at mid and low latitudes (Beland & Small, 2004; Carter et al., 2015; Kappenman, 2003; Marshall et al., 2012; Rodger et al., 2017; Zhang et al., 2015).

SCs are often driven by the interplanetary shock that precedes a coronal mass ejection (CME) or corotating interaction region (CIR) (Kilpua et al., 2015). CMEs are large eruptions of solar material that explode through the solar system (e.g., Chen, 2011; Kilpua et al., 2017; Webb & Howard, 2012). In contrast, CIRs are found at the interface between fast and slow solar wind streams (see reviews by Gosling & Pizzo, 1999; Richardson, 2018). CMEs and their associated structure may create an interval of particularly strong coupling between the solar wind and magnetosphere known as a geomagnetic storm, and driving dynamics such as substorms (Akasofu & Chao, 1980; Brueckner et al., 1998; Gonzalez et al., 1994; Yue et al., 2010). If an SC is followed by a geomagnetic storm then it may be termed a storm sudden commencement (SSC), while if it is not then it may be termed a sudden impulse (SI) (Curto et al., 2007). In contrast to impulsive phenomena such as SCs, which are driven by rapid changes in the solar wind, magnetospheric storms and substorms depend on the hysteresis of the coupled solar wind and magnetospheric system.

Geomagnetic substorms are cycles of energy storage and release in the magnetosphere. Typically there are considered to be three substorm phases: (a) the growth phase during which time energy is stored in the magnetotail magnetic field and plasma (McPherron, 1970); (b) the expansion phase in which the energy is explosively released; (c) the recovery phase where the system returns to its quiescent configuration (Akasofu, 1964). The growth phase lasts of the order of an hour, during which time the IMF configuration must be conducive to reconnection at the dayside magnetopause (Li et al., 2013). The explosive expansion phase lasts for around 15–30 min (Forsyth et al., 2015), starting with rapid increases in auroral brightness (Voronkov et al., 2003) and enhanced ultra-low frequency (ULF) wave activity (Rae et al., 2012; Smith, Rae, Forsyth, Watt, & Murphy, 2020; Smith, Rae, Forsyth, Watt, Murphy, & Mann, 2020). In the magnetosphere, magnetotail currents are diverted into the ionosphere through field aligned currents (see reviews by Kepko et al., 2015; Milan et al., 2017), enhancing the high latitude auroral electrojets and resulting in sharp deflections of the ground magnetic field (Akasofu & Chapman, 1961; Davis & Sugiura, 1966; Forsyth et al., 2018; McPherron et al., 1973; Mann et al., 2008). Timescales for the response of the field aligned currents and ground magnetic field perturbations to the solar wind have been found to depend strongly on geographical location, and range between 10 min and several hours (Coxon et al., 2019; Shore et al., 2018). Finally, the

recovery phase can last for several hours, and is characterized by fading auroral brightness, omega bands, reductions in ULF wave activity and weakening auroral currents. Generally, the dynamic auroral currents in the expansion phase are associated with some of the strongest ground magnetic field deflections (Freeman et al., 2019; Thomson et al., 2011; Turnbull et al., 2009; Viljanen et al., 2006). At high geomagnetic latitudes in Canada (above  $\sim 65^\circ$ ), Engebretson et al. (2021) recently found that most periods of rapid magnetic variability occurred within 30 min of substorm onset, though a substorm in itself was not a necessary nor sufficient condition to predict elevated magnetic variability.

Like substorms, geomagnetic storms have a series of three phases: the initial phase when the CME shock impacts the magnetosphere (the SSC), the main phase and recovery phase (Gonzalez et al., 1994). However in contrast to substorms, the main and recovery phases generally last for hours to days. Storms are characterized by a global disturbance in the geomagnetic field: An enhancement in the ring current, located about  $\sim 4 - 7 R_E$  ( $1 R_E = 6371$  km) from the Earth (Daglis et al., 1999), generates a magnetic field that opposes the background geomagnetic field. Though geomagnetic storms are often broadly cited as a cause of severe space weather, there are numerous physical processes that occur during the extended intervals of strong coupling between the solar wind and magnetosphere. Examples of these physical process include the SSC at the start of the storm (e.g., Clilverd et al., 2018; Rodger et al., 2017), substorms that occur during the main and recovery phases (e.g., Dimmock et al., 2019; Ngwira et al., 2015, 2018; Pulkkinen et al., 2003; Pulkkinen et al., 2015; Viljanen et al., 2006), and ULF pulsations (e.g., Heyns et al., 2020). Such magnetic ULF waves with periods of between 2.5 and 10 min, termed Pc5 waves (Jacobs et al., 1964), may be observed in the post-noon sector as a result of the Kelvin-Helmholtz instability operating on the dusk flank (Mann et al., 1999; Rae et al., 2005), or related to the impact of an interplanetary shock (Zhang et al., 2010).

The UK is located at mid magnetic latitudes ( $\sim 48 - 58^\circ$ ), which corresponds to an interesting location regarding the phenomena which can generate large geomagnetic fluctuations. Studies utilizing extreme value statistics have shown that these approximate latitudes may be susceptible to some of the largest rates of change of the field at 100-year return levels (Thomson et al., 2011; Rogers et al., 2020; Wintoft et al., 2016). This is likely related to unusually large extensions of the auroral oval and associated currents during extreme events. More generally, Freeman et al. (2019) found that just over 50% of extreme geomagnetic perturbations in the UK (above the 99.97th percentile) were associated with the substorm expansion and recovery phases. This suggests that models of the geomagnetic field fluctuations in the UK must be able to account for substorm activity. Meanwhile, Smith et al. (2019) found that  $\leq 10\%$  of large field perturbations (above  $\sim 50$  nTmin<sup>-1</sup>) in the UK were associated with SCs, and that this fraction decreased dramatically toward the more northerly parts of the UK. However there remained a strong link between SCs and strong geomagnetic activity in the days that follow. Therefore we might expect that models forecasting UK ground variability would pick up SC-like activity as an indicator that future activity may be likely.

Given the variety of mechanisms that can generate significant ground magnetic field variability, in order to forecast large rates of change of the magnetic field any model must be able to skillfully incorporate information about the time history of the incident solar wind. These processes also vary depending on the latitude and local time. In the last 5–10 years machine learning methods have been increasingly used to study and forecast space weather phenomena (e.g., see review by Camporeale, 2019). Often this has taken the form of forecasting a geomagnetic index (Liemohn et al., 2018), for example, the Sym-H (Bhaskar & Vichare, 2019; Siciliano et al., 2020), Dst/Est (Chandorkar et al., 2017; Gruet et al., 2018; Kugblenu et al., 1999; Lethy et al., 2018; Lundstedt et al., 2002; Tasistro-Hart et al., 2021; Wintoft & Wik, 2018; Wu & Lundstedt, 1996) or Kp indices (Ji et al., 2013; Tan et al., 2018; Wing et al., 2005; Wintoft et al., 2017; Zhelavskaya et al., 2019). Models have also been produced that aim to predict phenomena such as ionospheric current systems (Kunduri et al., 2020), geomagnetic storms (Chakraborty & Morley, 2020), substorms (Maimaiti et al., 2019) or SSCs (Smith, Rae, Forsyth, Oliveira, et al., 2020). On a local level, studies have also looked at forecasting the geomagnetic perturbations at magnetometer stations (Camporeale et al., 2020; Keesee et al., 2020; Wintoft et al., 2015). These geomagnetic perturbations have been shown to be difficult to forecast directly, using either physics based or empirical models (Pulkkinen et al., 2013) or machine learning techniques (Keesee et al., 2020), however it has been shown that models can skillfully forecast when the perturbations will exceed pre-determined levels (Camporeale et al., 2020; Pulkkinen et al., 2013), or the maximum perturbation (Tóth et al., 2014). For operational purposes, predicting the exact perturbation amplitudes may not be

necessary since there is generally a threshold level at which the risk is deemed sufficiently high to warrant remedial action.

In this work we will examine the ability of several different machine learning architectures to produce skillful probabilistic forecasts of the ground magnetic perturbations in the UK exceeding set values. Section 2 will describe the data, models and model evaluation methods. Section 3 will first discuss the overall performance of the models when applied to an unseen 2-year interval, before qualitatively demonstrating the models during an example storm for a single combination of input data window, magnetic field variability threshold and forecast horizon. The results of the three models for all ground magnetic field variability thresholds will then be qualitatively discussed for two example magnetospheric storms. The corresponding metrics achieved by the models during these storm periods will then be quantitatively assessed. Next, the impact of changing the volume of solar wind input provided to the models, and the horizon with which the forecast is made will be evaluated. Finally, Section 4 will further discuss the performance of the models, as well as the implications for forecasting ground magnetic field variability and future development.

## 2. Data, Method and Models

In this section we will outline the input data used to train our machine learning models, the pre-processing and preparation applied to these data sets, the metrics whereby model performance is measured and validated, and the models to be tested.

### 2.1. Input Data

For this work we drive the models using data obtained upstream of the Earth at L1. To maximize the homogeneity and continuity of the data, we use the OMNI database (<http://nssdc.gsfc.nasa.gov/omniweb/>). To produce the OMNI data, the spacecraft data obtained at L1 is post-processed and propagated to the bow shock, negating most of the requirement to consider applying time-lags to the data (c.f. Wintoft et al., 2015). The OMNI data has been used with success in other similar forecasting studies (e.g., Keese et al., 2020). We use 1-minute resolution data for this work.

We manually select variables (or features) from the OMNI data set to describe the solar wind at L1. We do not select multi-variable quantities such as the dynamic pressure or convection electric field as the models will already have the composite information in some form. We select those features that are often represented in solar wind-magnetosphere coupling functions (e.g., Milan et al., 2012; Perreault & Akasofu, 1978), specifically we use: the solar wind velocity, density, magnetic field components in the geocentric solar magnetospheric (GSM) system and magnetic field magnitude. The inclusion of other available variables less commonly included in coupling functions (e.g., temperature) were tested and not found to increase the skill of the models, and so they were discarded. Importantly, our variables could all be provided in near-real time from spacecraft at L1, although we note that there are differences between the OMNI and real-time data, not least the fact that the OMNI data is propagated to the bow shock.

In addition to the OMNI data, we also need to provide as an input the location at which the “ground truth” magnetic field variability has been collected. The latitude considered by the model is set by the magnetometer station used for ground comparison, any local effects are also included by the choice of ground station. If forecasts were needed for a different latitude, then a new station could be selected. However, we do need to include the magnetic local time (MLT) of the station as an input as this will cyclically vary over time. We transform the MLT of the station from a linear variable with a jump at midnight (i.e.,  $m = 0-24$ ) to a pair of continuous cyclical variables:

$$M_1 = \sin\left(\frac{2\pi m}{24}\right); M_2 = \cos\left(\frac{2\pi m}{24}\right) \quad (1)$$

Similar such transformations have been used in previous studies of phenomena that depend on MLT (e.g., Bentley et al., 2020). In total therefore there are eight features provided to the models, six describing the solar wind properties and two for the MLT of the station. We also wish to provide the models with the time history of the solar wind. In principle, this can be either done by providing the time series data to the models

explicitly and letting the model learn the most important information (e.g., Kunduri et al., 2020), or by using features that describe the variability in a time window (e.g., Camporeale et al., 2020). In this work we provide the time history explicitly, testing the ability of several different neural network architectures to extract the important and necessary information from the rich input data. Effectively, we provide an input matrix of shape  $(\Delta T, 8)$ , where  $\Delta T$  is the length of the input window in minutes, and eight is the number of features. In this work we test the performance of the models using different lengths of input window, selecting either 30 min, 1 or 2 h of historical data.

We note here that we do not include the prior ground magnetic field variability as an input, which would allow the models to develop some form of persistence forecast. We do this as we wish to assess how the models are processing the solar wind data and inferring the strength of the myriad of magnetospheric processes described above. In the future, better model skill could likely be achieved through the inclusion of local data, such as the ground field variability, or through more global parameters such as indices which would indicate the current state of the magnetospheric system. Additionally, whilst we have selected the OMNI data for the basis of this investigation, were similar models to be used in an operational space weather capacity then care would need to be taken transitioning to real-time solar wind data obtained at L1.

## 2.2. Output Data Processing

For this work we create and evaluate probabilistic forecasts of whether the rate of change of the ground magnetic field in the UK will exceed a given threshold. We have chosen to create this forecast for the Lerwick (LER) magnetometer station, as it is the highest latitude magnetometer station in the UK and generally sees the greatest rates of change of the magnetic field of the three UK INTERMAGNET stations (e.g., Freeman et al., 2019). The LER station is located at a geomagnetic latitude of  $57.85^\circ$  and a longitude of  $81.15$ , and so is most comparable to the mid latitude stations considered by Pulkkinen et al. (2013) and Tóth et al. (2014), and also to the Ottawa station considered by Keesee et al. (2020).

We define  $R$ , the rate of change of the magnetic field, as the rate of change of the horizontal magnetic field vector as follows (where  $X$  and  $Y$  are the northward and eastwards components of the magnetic field respectively):

$$R = \frac{\delta \mathbf{H}}{\delta t} = \frac{\sqrt{[X(t + \delta t) - X(t)]^2 + [Y(t + \delta t) - Y(t)]^2}}{\delta t} \quad (2)$$

This definition has been used in the past as a proxy for GICs as it captures directional changes in the field which could be significant (Freeman et al., 2019; Smith et al., 2019; Viljanen et al., 2001). In this work we are interested in forecasting whether  $R$  will exceed predefined thresholds in the future. We follow Pulkkinen et al. (2013), who evaluated a series of models using similar thresholds of ground magnetic field variability. These levels are placed at: 18, 42, 66 and  $90 \text{ nTmin}^{-1}$ , accounting for the different cadence of the observations (c.f. Pulkkinen et al., 2013). We then can structure the forecast as a binary classification problem, positive if the threshold is exceeded, and negative if not. Specifically we train the models to produce a probabilistic output to the classification problem.

We consider the time horizon with which we train the models to predict elevated  $R$ . This involves looking ahead at each time step to determine if the threshold of  $R$  is exceeded within  $T_H$  minutes, where  $T_H$  is our forecast horizon. If the threshold is exceeded, then the ground truth is classified as a “positive” event in the data set. As discussed above, the maximum possible length of this horizon will depend on the processes that drive the elevated rates of change of the ground magnetic field. Previous works have considered a fixed forecast horizon of 20 min, finding this to be a good balance between phenomena (e.g., Camporeale et al., 2020; Tóth et al., 2014). Others have considered different horizons, for example, Wintoft et al. (2015) considered a horizon of 30 min. However, we note that this was using data directly from L1, and so also included the travel time between the spacecraft and the magnetopause, meaning that in practice the time horizon was very short. In this work we investigate a series of horizons, choosing 30 min, 3 and 12 h. These correspond approximately to the substorm expansion phase, an approximate substorm length and the main phase of a geomagnetic storm, respectively (e.g., Forsyth et al., 2015; Walach & Grocott, 2019). In this way we are ask-

ing whether the models can predict these large scale coupling processes in advance, and whether they will cause large magnetic perturbations at the latitude and MLT of the magnetometer station. These horizons are longer than those previously considered in the literature, because from an operational perspective as large a warning period as possible would be preferable.

### 2.3. Data Preparation

To create an effective ML model it must be trained using a well structured data set. For training to identify temporal features, samples should be continuous, and of same shape and resolution to create an accurate model. The OMNI data set is noted to have dropouts, particularly in the plasma data (e.g., Keese et al., 2020). Some of these dropouts are due to saturation of the instruments during extreme conditions, an unfortunate circumstance that should be mitigated in future space weather missions (Nicolaou et al., 2020). Many of these data gaps are very short, of the order of minutes, and so to maximize the volume of training data we employ linear interpolation to fill small data gaps smaller than 15 min (e.g., Wintoft et al., 2015). This is particularly significant if two hours of continuous data are required to produce an output, as in our longest input window. In this regard, providing single numbers to describe the conditions or variability (e.g., RMS, Range) could be advantageous (e.g., Camporeale et al., 2020), however potentially important information could be lost unless the most significant measures are used. Most of the data drop outs at L1 are in the plasma data, and so some models have been trained that use only the magnetic field data. This allows a model, albeit with more limited performance, to perform predictions when the requisite plasma data is not available (e.g., Wintoft et al., 2015). These show reduced performance, but allow forecasting when it would otherwise not be possible. Though this is an excellent method, we do not employ it in this study as we wish to evaluate how the models can process the data and determine the coupling and driving of the magnetospheric system. We limit our data set to those intervals with both magnetic field and plasma data, noting that data gaps smaller than 15 min have been filled through linear interpolation.

It is also important to consider the scale and range of data with which we are presenting the models. Each feature has its own distinct mean and range; for example, the velocity will vary between values of the order of hundreds of  $\text{kms}^{-1}$ , while the density will vary from several  $\text{cm}^{-3}$  to tens of  $\text{cm}^{-3}$ . To prevent any single variable from dominating the numerical models, as a result of the nature of the units and order of magnitude of the measurement, we scale the values of each feature independently using the mean and standard deviation of that feature. This effectively normalizes the values such that each feature now has a mean of zero and a standard deviation of one.

### 2.4. Metrics

In this work we are producing probabilistic forecasts, and we can use a series of metrics to evaluate the reliability and skill of these predictions. Previous space physics and space weather studies have used metrics such as the receiver-operator characteristic (ROC) and Brier skill scores (BSSs) (Azari et al., 2018; Crown, 2012; Forsyth et al., 2020; Leka et al., 2019; Murray et al., 2017; Smith, Rae, Forsyth, Oliveira, et al., 2020), which have a strong heritage in terrestrial weather forecasting.

The BSS is a measure of the calibration of the probabilities, compared to a reference prediction. The Brier score (BS) represents the mean square of the probability error in the predictions (Brier, 1950):

$$BS = \frac{1}{N} \sum_{i=1}^N (\rho_i - a_i)^2 \quad (3)$$

where  $N$  is the number of events,  $\rho_i$  is the probabilistic forecast for event  $i$  and  $a_i$  is the corresponding observation (zero or one). The BS can vary between zero for a perfect forecast (i.e., the model predicts zero or one as required) and one for a completely incorrect forecast (i.e., the model always predicts the opposite outcome). The BSS compares the BS of the model to a reference forecast ( $BS_{Ref}$ ):

$$BSS = \frac{BS_{Ref} - BS_{Model}}{BS_{Ref}} \quad (4)$$

The BSS will be one for a perfect forecast, zero if the predictions are comparable to the reference forecast, and negative if the model performs worse than the reference. As is often the case, here we compare to climatology: The forecast probability is simply the fraction of data represented by the positive class.

We note here that a perfectly reliable model need not forecast 100% probabilities for all positive events, but over the full data set when it estimates a 0.5 probability it should identify the positive class 50% of the time. Because of this, while the BSS is a good measure of the reliability of a probabilistic forecast, it is most useful when coupled with a metric that describes the skill of the model; one that describes how well the model can separate the positive and negative classes. Classification problems often use a contingency table to evaluate the skill of their results: a table of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Metrics can then use combinations of these categories to create a single statistic. When dealing with probabilistic predictions such contingency tables can be evaluated as a function of probability for event acceptance.

A common and useful metric for evaluating the skill of a probabilistic model is the ROC (Swets, 1988). It is a comparison between the FP rate ( $1 - TN/(TN + FP)$ ) against the TP rate ( $TP/(TP + FN)$ ) as the threshold for event acceptance is increased. The ROC score is a method of quantifying the idea that a skillful model should maximize its hit rate before encountering significant false alarms. The ROC score is measured between zero and one, with one representing perfect skill and 0.5 the equivalent of randomly forecasting a coin toss (Zweig & Campbell, 1993). However, we note that our data set is imbalanced, with many fewer positive events (for which the threshold of variability is exceeded within the forecast horizon) than negative events. In fact, the thresholds we test in this work are all above the ninety ninth percentile of variability at LER. This imbalance is somewhat mitigated by the use of the forecast horizon in this work, effectively extending the positive class for greater intervals. However, even in the most favorable combination of parameters (threshold =  $18 \text{ nTmin}^{-1}$ , forecast horizon =  $720 \text{ min}$ ) the positive class only accounts for ~6% of the data. When dealing with imbalanced data sets high ROC scores can be achieved whilst the minority class is not well identified. For this scenario, and particularly where the positive (and minority) class is important, a metric such as the precision-recall (PR) score which focuses on evaluating the positive class can be an excellent choice. The PR score is calculated in a similar manner to the ROC score, where the threshold for event acceptance is adjusted, but now the precision ( $TP/(TP + FP)$ ) and recall ( $TP/(TP + FN)$ ) of the model are tested (e.g., Jonas et al., 2018). We note that a model which maximizes the ROC score may not maximize the PR score (Davis & Goadrich, 2006).

We derive uncertainties in the BSS, ROC and PR scores by bootstrapping the test data set (e.g., Yousef et al., 2005). The provided uncertainties are the 95% confidence intervals returned by performing randomized bootstrapping with replacement 100 times. This enables a robust comparison between the different models and input choices. We note that though we do not present the results of multiple independently trained models, the chosen bootstrapping method well represented the variation observed by different training runs.

## 2.5. Cross Validation

We use the data between 1996 and 2016, covering almost two solar cycles, to train and validate the models. We use the years 2003–2014 to train the models, with the years 1996–2002 used as a validation set. This leaves 2015 and 2016 as an unseen test data set from which we can report our performance metrics. The division between the train, validation and test sets has been done with the aim of creating continuous sets with an approximate 70/20/10 fractional data split, while also maintaining as even a proportion of large  $R$  between the sets as possible. The distribution of  $R$  in the three subsets is shown in Figure A1. We also shuffle the data during training to ensure the models see a variety of different conditions between model updates. The shuffling is performed when generating the batches for model training, that is, in such a way that preserves the temporal sequence for each prediction. Unfortunately, the splitting does mean that only one storm from the Welling et al. (2018) evaluation set is found in the test data set: the storm of March 17th, 2015. However, we will show results from the application of our models to this storm.

We note that when we perform the scaling of each feature based on their mean and standard deviation we only use data from the Training set in order to determine the scaling parameters. This ensures that the Test set is completely unseen by the models before being used to assess their performance.

## 2.6. Models

Neural network models have been shown to perform admirably at classification, regression and forecasting tasks in the fields of space plasma physics and space weather (e.g., Bakrania et al., 2020; Bloch et al., 2021; Bortnik et al., 2016; Clausen & Nickisch, 2018; Garton et al., 2021; James et al., 2020; Lethy et al., 2018; McGranaghan et al., 2020; Wintoft et al., 2017; Zhelavskaya et al., 2021). For space weather forecasting in particular, models can be structured to have a “memory” of the preceding solar wind conditions (Bhaskar & Vichare, 2019; Kugblenu et al., 1999). This has typically been done through the use of recurrent layers (Gruet et al., 2018; Keese et al., 2020; Liu et al., 2020; Tan et al., 2018; Wu & Lundstedt, 1996), or by using filters on the historical data to extract important information, convolution layers for example (Kunduri et al., 2020; Siciliano et al., 2020).

In this study we compare and contrast three different model architectures using Keras and TensorFlow (Abadi et al., 2015). The properties of the networks were determined using an iterative testing regime, making incremental adjustments of the number of layers, neurons, dropout and regularization and comparing the subsequent performance. First, we assess a “Dense” model with a series of two hidden layers, the first with 32 neurons and the second with 16. The activation function for these layers is a rectified linear unit (“ReLU”). To prevent overfitting an L2 regularization (factor of 0.001) is used, as well as intermediate dropout layers (at a rate of 0.2). No dropout is applied on the input layer. There is a final layer of a single neuron that uses the sigmoid activation function to enable a probabilistic output. We note that the input for this model is not a matrix of shape  $(\Delta T, 8)$ , but instead is a flattened input of shape  $(\Delta T \times 8)$ , similar to the approach of Garton et al. (2021).

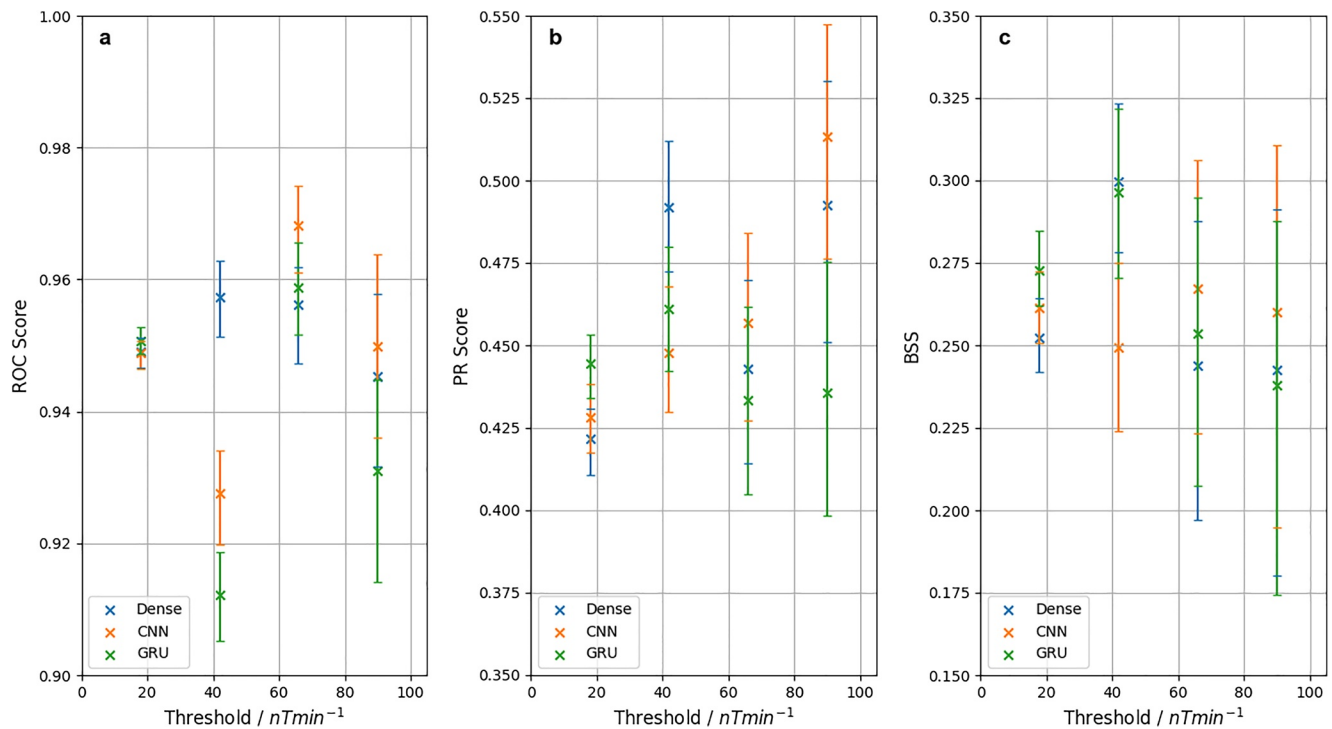
Second, we test a convolutional neural network (CNN) approach. The initial convolutional layer has 64 filters, and this is followed by two Dense layers with 32 and 16 neurons, with a final output layer consisting of a single neuron. Again, the activation for these layers is the ReLU function, apart from the final layer which uses a sigmoid function. L2 regularization (factor of 0.001) and intermediate dropout layers (rate of 0.3) are also used to prevent overfitting.

Third, we test a recurrent neural network. Specifically we use an initial gated recurrent unit (GRU) layer with 32 units. A GRU is similar to the popular long short-term memory unit (LSTM: Hochreiter and Schmidhuber (1997); e.g., Keese et al. (2020)), but has fewer trainable parameters (Cho et al., 2014) and comparable performance modeling sequences (Chung et al., 2014) and showed equal or better performance during our initial testing. The GRU layer is followed by a dense layer of 16 neurons. We use a recurrent dropout (rate of 0.3), intermediate dropout layers (at rates of 0.3) and L2 regularization of the Dense layer (factor of 0.001) to minimize overfitting. The final output layer again uses the sigmoid activation function.

The models were compiled with a binary cross-entropy loss and Adam optimizer, while maximizing the ROC and PR metrics (discussed above). A learning rate of 0.0005 and a batch size of 10,080 (representing 1 week of data) were utilized. The slower than default learning rate and large batch size were selected through an iterative testing regime and are likely a result of the class imbalance in the data set. The validation loss was monitored, and an early stopping procedure used to break the training when it increased for two epochs. The early stopping procedure was chosen as suitable by observation of the validation loss over a large number of epochs. With the combination of parameters above, the models train for around 30–60 epochs, achieving a smooth loss curve and reaching a plateau in model loss.

We note here that we have used the same model architectures to evaluate the models when using different input lengths, forecast horizons and thresholds, despite these changing the quantity of input data and the fraction of positives in the training data. Ideally the model architecture would be individually tuned to each setup, however this would make comparisons between the results difficult and very time intensive. We did however optimize the models using results from a series of different thresholds, windows and horizons to ensure we maximize performance while avoiding overfitting. We also note that each combination of





**Figure 1.** The (a) receiver-operating characteristic, (b) precision-recall and (c) Brier skill score metrics achieved by the models on the test data set (2015–2016) as a function of threshold for the Lerwick magnetometer station using 30 min of solar wind history and a horizon of 30 min. The uncertainty in the values are calculated using a bootstrap method and represent the 95% confidence intervals from 100 iterations.

threshold, input and horizon is represented by an independently trained model, an alternative approach would be to use a single model to evaluate the probability of multiple different thresholds being exceeded.

Though there are not firm rules to determine the architecture of a neural network, there are best practice guidelines (e.g., Heaton, 2008; Ranjan, 2020). First, we consider that the number of free parameters in the model should not exceed the number of samples used for training, if possible. Our three models have approximately 8,000, 18,000 and 4,500 trainable parameters respectively at the smallest input size (30 min). The number of trainable parameters in the Dense and CNN models increases with larger input windows, to approximately 31,000 and 64,000 parameters at an input of two hours, respectively. In comparison, for the shortest horizon and largest threshold considered we have 3,710 examples of the positive class in our training set, representing the smallest number of positive samples. At the largest time window this increases to 34,113 positive samples. Though not quantitative this comparison shows how the complexity of the models could be increased if only dealing with the larger forecast horizons due to the availability of the data. Second, it is generally accepted that the number of nodes in a hidden layer should be intermediate between the layers on either side, creating a tapered network. We have followed this convention.

### 3. Results

Figure 1 shows the metrics achieved by the models using 30 min of the time history of the solar wind and a forecast horizon of 30 min as the threshold of ground magnetic field variability is adjusted. These results are obtained by applying the models to the unseen test data set (obtained during the years 2015 and 2016). All three models reach ROC scores of 0.9–0.97 for the thresholds tested. The models report PR scores of between 0.4 and 0.5. As noted above, this likely reflects the imbalanced data set where the ROC score is inflated by considering the correct majority class where “nothing” occurs. The BSSs of the models are around 0.2 – 0.3, indicating good reliability compared to climatology.

While the metrics are relatively constant with increasing threshold, it is interesting to note that the size of the uncertainty increases, moving left to right in the panels in Figure 1. This likely reflects the relatively

small fraction of data for which a higher threshold of  $R$  is exceeded. When the bootstrapping process is used to estimate uncertainty, the subsets selected will have fractionally less in common if the positive class is fewer in number.

For the majority of thresholds tested (and metrics evaluated) the three models perform excellently, and to a similar level. An exception to this is the ROC score for a  $42 \text{ nTmin}^{-1}$  threshold, where the Dense model outperforms the more complex CNN and GRU models. If we compare the PR scores for these models we find that the scores are more similar, and mostly within uncertainties. This suggests that the Dense model is getting credit in its ROC score for correctly predicting when the  $42 \text{ nTmin}^{-1}$  threshold is not exceeded. This inference highlights the utility of evaluating both the ROC and PR scores, with their focus on general classifications and on the positive class, respectively.

### 3.1. Example Output: Storm March 17th, 2015

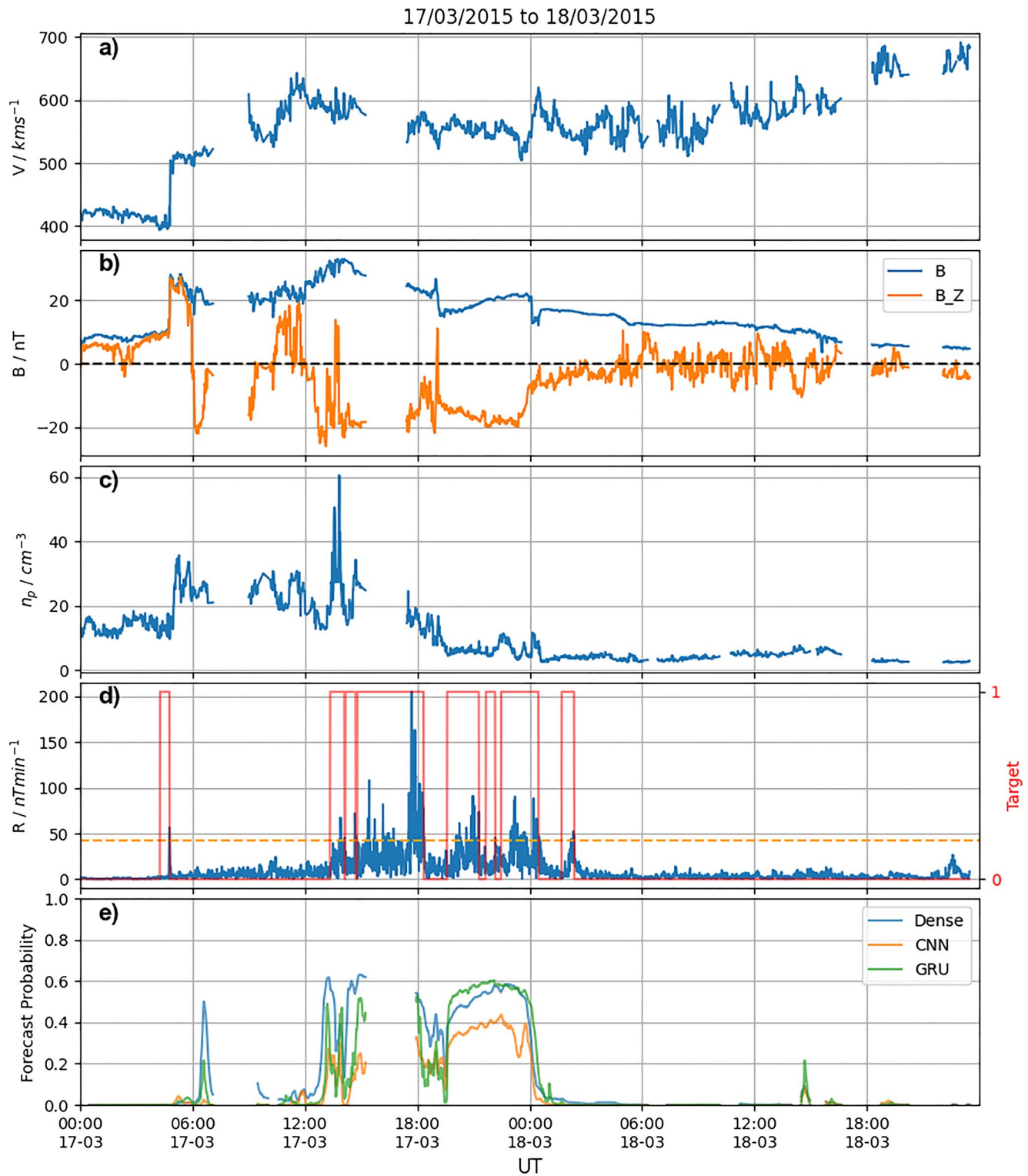
Figure 2 shows the results of testing the models on two days around a severe magnetospheric storm in March 2015. The space weather impacts of this storm, sometimes called the St Patrick's Day storm, have been analyzed in terms of the ionospheric response (e.g., Astafyeva et al., 2015), as well as the geomagnetic, geoelectric fields and GICs (e.g., Blake et al., 2016; Carter et al., 2016; Kozyreva et al., 2018; Marshalko et al., 2020). Additionally this storm has been selected as it was featured in the Welling et al. (2018) set of storms, and can be found within our unseen test data set: the models have not been trained on this data. We initially use a threshold of  $R = 42 \text{ nTmin}^{-1}$  and a forecast horizon of 30 min. Figures 2a–2c show the solar wind data from OMNI, Figure 2d shows  $R$  measured at the LER ground station, with the perfect model forecast (Target) in red. The Target in Figure 2d would change with the use of a different threshold of  $R$  or forecast horizon. The forecast probabilities from the three models are shown in Figure 2e.

We can make several links between the upstream solar wind data and the  $R$  measured on the ground, and consequently with the required “Target” during this storm. First, we see a shock in the solar wind at around 04:30 UT on the March 17th, clearly shown in the solar wind velocity, IMF and solar wind density (Figures 2a–2c). During the Sudden Commencement on the ground,  $R$  increases suddenly above the threshold indicated with the horizontal dashed line. As discussed in Section 2.2, this breaking of the threshold is reflected in the “Target”, shown in red, where a 30 min window prior is flagged as “1”. This represents the forecast horizon with which we would wish the models to indicate the increased likelihood of the threshold being broken. The CNN and GRU models do show very small increases in probability around this interval, but they are less than 5%. The increases in probability are also delayed from the SC, the use of the propagated OMNI data is unsuitable for forecasting such rapidly driven phenomena such as SCs: They do not see the shock far enough in advance to create a useful forecast.

A few hours later in the interval, there is an increase in the forecast probabilities from the Dense and GRU models just after 06:00 UT. This appears to be associated with a strong southward turn of the IMF (Figure 2b), but does not correspond to an increase in  $R$  (Figure 2), and so would represent a false alarm.

Next, just after 14:00 UT we see another southward turn of  $B_{z,GSM}$ , this is accompanied by elevated levels of  $R$  at LER. These breach the  $42 \text{ nTmin}^{-1}$  threshold, and we see that all three models show correspondingly elevated probabilities. The Dense model predicts the highest probability, followed by the GRU model, with the CNN model being the most conservative. Overall, observations of large  $R$  persist for at least six hours and for the first few hours this is reflected in the models' forecasts. However, starting at 15:30 UT and lasting for around 2 h there is a data gap. During this interval, and for 30 min after (until the model input window is once more fully populated by data) the models cannot make a prediction. For this storm the data gap coincides with the largest  $R$  observed at LER. A second period of  $R$  often exceeding  $42 \text{ nTmin}^{-1}$  is observed between roughly 19:30 and midnight UT. All three models forecast elevated probabilities in this interval.

Overall, during the bulk of the storm the models perform excellently, with elevated probabilities being shown during those periods before and during intervals at which the levels of  $R$  exceed the threshold. One interval where this is not the case is around the SC at the start. Here, either the models do not perform as would be hoped, or the use of propagated solar wind data inhibits the ability to produce a forecast. Additionally, input data gaps clearly negatively impact the forecasting ability of the models. Note that from a single



**Figure 2.** The performance of the models for a two day interval during the March 2015 storm. The data obtained in the solar wind from the OMNI data set: (a) the solar wind velocity, the magnetic field magnitude and (b)  $B_{Z,GSM}$  component and (c) the density. The rate of change of the magnetic field observed at Lerwick in blue, defined as in Equation 2, with the Target forecast defined by a threshold of  $42 \text{ nTmin}^{-1}$  and (d) a forecast horizon of 30 min in red. The horizontal dashed line indicates the  $42 \text{ nTmin}^{-1}$  threshold. The forecast probabilities returned by (e) the Dense, convolutional neural network and gated recurrent unit models.

epoch we cannot assess the accuracy of the forecast probability, that is, its reliability, these will be evaluated over the entire intervals.

### 3.2. Multiple Thresholds

Next, we compare and contrast the results of running the models with different thresholds of  $R$ . We will look at the results of the models during two example storms in 2015.

#### 3.2.1. Storm March 17th, 2015

Figure 3 shows the results for the same storm as in Figure 2, which occurred in March 2015. Figures 3a and 3b detail the solar wind velocity and magnetic field observations, while Figure 3c shows  $R$  measured at the LER station. Figures 3d–3f then show the outputs from the Dense, CNN and GRU models respectively. For context, the orange traces in Figures 3d–3f show the model results for the combination of parameters detailed in Figure 2e. The horizontal bars at the top of Figures 3d–3f represent the target forecast, with the color indicating the largest threshold of  $R$  that is exceeded within the forecast horizon (here 30 min). In effect, the target forecast would see an increase in the probabilities of the model of the corresponding color and those of lower thresholds, but not for the models with higher thresholds. For example, if the bar is orange then the 42 nTmin<sup>-1</sup> threshold is broken within 30 min, and an ideal model would see increased probabilities of the green and orange models (18 and 42 nTmin<sup>-1</sup>), but not the red or purple models (66 or 90 nTmin<sup>-1</sup>).

Analyzing the storm chronologically, we again see an SC in Figure 3c at 04:30 UT. This reaches above 42 nTmin<sup>-1</sup>, but not 66 nTmin<sup>-1</sup>, and so the target forecast in this interval is represented by an orange bar in Figures 3d–3f. So we would hope to see the 18 and 42 nTmin<sup>-1</sup> threshold models (green and orange) forecast increased probabilities in that time. However, as discussed above with regards to the 42 nTmin<sup>-1</sup> model in Figure 2, this is not seen. In the following hours the upper envelope of  $R$  does increase to nearly the 18 nTmin<sup>-1</sup> level, and this can be reflected in all three models' forecasts (Figures 3d–3f) where the green models are forecasting elevated probabilities.

Just after 06:00 UT there is a southward field deflection, and all three 18 nTmin<sup>-1</sup> models increase their probabilities. However this threshold is not broken for another few hours, though we note that  $R$  is elevated, and nearly at this level. The Dense and GRU models also begin to predict much higher levels of  $R$  to be broken, with significant probabilities returned for the models up to the 66 nTmin<sup>-1</sup> (red) level. The CNN model on the other hand, does not increase for any of the three higher thresholds, which turns out to more accurately reflect the target.

For the remainder of the storm all three models report elevated probabilities that reflect the “positive” target forecasts indicated with the horizontal bars, which, given the strongly negative IMF  $B_z$ , likely represent increased  $R$  due to magnetospheric substorms or convection. However, we note that all three models show less variability in their forecast probabilities than we see in the the horizontal target forecast bar, which often changes between levels rapidly (e.g., from orange to purple within tens of minutes). This likely indicates that the models are not able to predict the timing of such large  $R$  precisely, but recognize that the magnetosphere is experiencing a highly dynamic interval.

#### 3.2.2. Storm June 21st, 2015

A storm during June 2015 provides a second example as to how the models perform during an active interval. This storm has previously been studied in detail from the perspective of multiple spacecraft and data sets by Reiff et al. (2016), while its impact on a mid-latitude high speed rail network has also been documented (Liu et al., 2016). In the interval around this storm there were a series of three step-like increases in solar wind velocity. The first, at around 18:00 UT on the June 21st did not result in the 18 nTmin<sup>-1</sup> threshold at LER being exceeded, however there is a small signature in  $R$ . There are very small increases in the corresponding Dense and GRU models, but these are less than a few percent. The second increase in solar wind velocity at around 06:00 UT on the June 22nd does cause an  $R$  above the 18 nTmin<sup>-1</sup> threshold, but neither

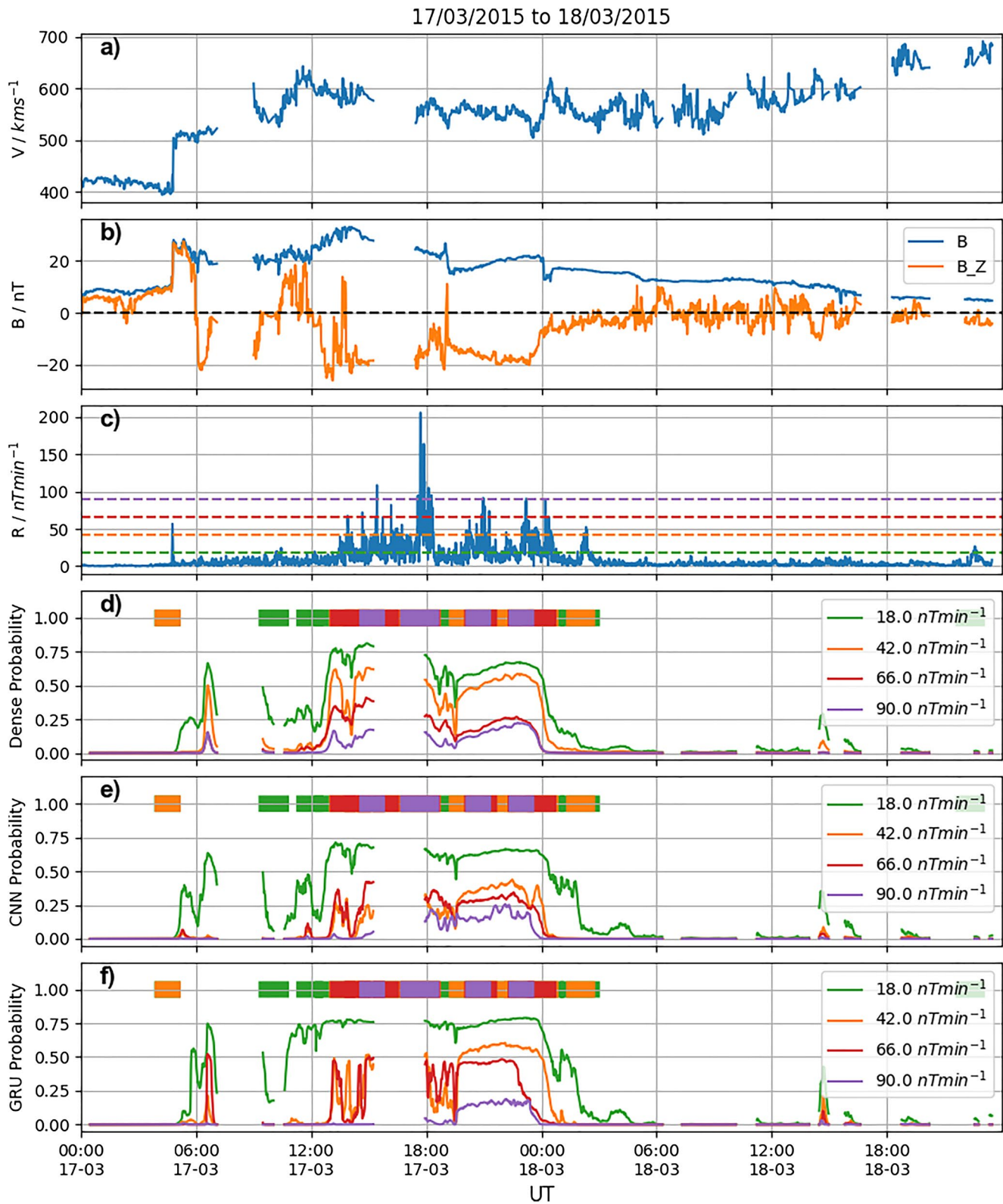


Figure 3.

the CNN or GRU models forecast increased probabilities. The Dense model on the other hand does forecast an increased probability, however this was slightly delayed likely due to the propagated solar wind data.

Later in the interval, at around 15:00 UT on the June 22nd there is a period of moderate southward IMF during which all three models forecast an increased likelihood of  $18 \text{ nTmin}^{-1}$  or greater  $R$ , of the order 10 – 20% chances. This does not occur, though  $R$  does appear elevated at this time, but more of the order of  $\sim 10 \text{ nTmin}^{-1}$ , and so this would count as a false alarm. The third and largest step like increase in solar wind velocity, at approximately 18:00 UT on the June 22nd, is accompanied by strong southward IMF and all three models forecast high probabilities of  $R$  in excess of  $90 \text{ nTmin}^{-1}$ . This excellently reflects the observations at LER, highlighting the utility of these models. The models also capture a second interval of extremely large  $R$  just after midnight on the June 22nd. This second interval of extreme  $R$  is not associated with a shock or increase in solar wind velocity, but instead related to a period of strong southward IMF, suggesting the models are not solely reliant upon changes in velocity but can skillfully use the IMF information.

### 3.2.3. Storm Metrics

The qualitative analysis above is useful to assess areas where the models work well, for example, periods of activity associated southward IMF, while also flagging phenomena that may not be suitably captured by the current models and data input (SCs for example). However, we can also use the metrics set out above to quantitatively compare and contrast the different models, as shown in Figure 1 for the full 2-year test data set. Figure 5 focuses on the results for the March 2015 (Figures 5a–5c) and June 2015 (Figures 5d–5f) storm periods in a similar format to Figure 1.

For both storms the models achieve ROC scores above 0.9, indicating excellent skill at discriminating intervals where  $R$  will be elevated from those times when it will not. These scores are also comparable to those obtained from the full (2-year) test data set (Figure 1).

When considering the PR scores, for the March 2015 storm (Figure 5b) the scores decrease with threshold from a very high score of 0.8, showing that the models are less successful at forecasting the very large  $R$  during the interval. This may be explained because of two main factors, first there is a large data gap around the interval with the largest  $R$ , when the models might hope to perform best. Second, there are several occasions where the  $90 \text{ nTmin}^{-1}$  threshold is only just broken, these instances may be harder to distinguish than times when the threshold is broken by a large margin. For the storm in June 2015 we see higher PR scores of between 0.7 and 0.8 for all thresholds. In contrast, during the June storm there were no data gaps, and in the intervals in which the  $90 \text{ nTmin}^{-1}$  threshold was exceeded  $R$  was significantly greater than the threshold. We note that these scores are greater than that achieved for the test data set as a whole, indicating that the models perform better when there is a greater period of solar wind driving, and there are no data gaps. From the other perspective, it may indicate that during more sporadic intervals of solar wind driving, which would be represented in the larger test data set, the performance of the models is not as good.

Finally, in terms of BSSs (Figures 5c and 5f) we see scores of between 0.4 and 0.6 for the June 2015 storm, indicating the excellent reliability of the forecasts, above that found for the full test data set. Meanwhile, for the March 2015 storm we again see a decrease in the reliability of the predictions with threshold, likely for similar reasons to the decrease in PR score discussed above.

For the majority of thresholds and metrics the three models achieve similar performance, with a few notable exceptions. During the March 2015 storm the Dense model outperforms the other two at the largest thresholds ( $66$  and  $99 \text{ nTmin}^{-1}$ ). This likely reflects the fact that the CNN and GRU models do not well forecast the first interval during which the  $90 \text{ nTmin}^{-1}$  threshold is exceeded at around 15:00 UT, just before the data gap (Figure 3). Evaluating the metrics over a single storm interval increases the importance of each

**Figure 3.** The performance of the models for a 2-day interval during the March 2015 storm for the Lerwick (LER) station. The data obtained in the solar wind from the OMNI data set: (a) the solar wind velocity, and (b) the magnetic field magnitude and  $B_z^{GSM}$  component. (c) The rate of change of the magnetic field observed at LER in blue, defined as in Equation 2. The horizontal dashed lines indicate the four thresholds chosen for this study: 18, 42, 66 and  $90 \text{ nTmin}^{-1}$  in green, orange, red and purple respectively. The forecasts produced by (d and f) the Dense, CNN and GRU models. The horizontal bars indicate the perfect forecast for the 30 min horizon, the color indicating the largest of the thresholds of  $R$  that is, exceeded.

“TP” or “FP” interval. Overall, the similarity of the scores indicates that all three methods of encoding the solar wind history allow skillful models to be created.

### 3.3. Length of Input and Forecast Horizons

We now examine how adjusting the length of the input window and increasing the forecast horizon change the skill and reliability of the models. Figure 6 shows the PR scores obtained by the models over the 2-year test data set as the forecast input and horizon are adjusted. The PR score is presented as we wish to focus on the skill of the models in identifying the positive class (when the thresholds are exceeded), as discussed above. Figure 6a represents the original PR metrics presented in Figure 1b. We use this as a benchmark, and note that the models achieve PR scores between 0.4 and 0.5 in this setup.

As we increase the quantity of input data to the models, moving from left to right (e.g., Figures 6a–6c) we might expect to see increasing scores as we are providing the models with more contextual information from which to make their forecast. However, while we see some moderate gains at low thresholds (of the order of 0.05 increases) we mostly see decreases in the scores at higher thresholds. There are competing considerations here, the model architectures (in terms of hidden layers) are the same between the different input windows, and it is likely that different or more complex models are required to utilize the additional input data effectively. We note that the number of trainable parameters does increase for the Dense and CNN models, in order to deal with the larger input volume (Section 2.6). However, it is likely that the Dense and CNN models are limited by the number of positive examples in the training data, where the larger number of trainable parameters are unsupportable with the available training data, particularly at high thresholds. Meanwhile, the number of trainable parameters does not change for the GRU model, perhaps emphasizing the limitations of the fixed architecture.

Increasing the forecast horizon, moving top to bottom (e.g., Figures 6a–6g) we see a general increase in the skills of the models when the horizon is increased to 180 min. There are several factors that could account for this. First, the models may be able to identify the consequences of coupling behavior for which the magnetospheric processing time (time delay between solar wind and subsequent ground impact) is greater than 30 min. Second, the increased forecast horizon allows the models to be less precise in their timing of when the threshold of  $R$  will be broken. Additionally, increasing the forecast horizon length will also lessen the class imbalance present in the data set, providing more “positive” examples. However, on the other hand increasing the forecast horizon also requires that the models can identify intervals when the thresholds of  $R$  will be broken further in advance. This is clearly not possible for some magnetospheric phenomena that cause large  $R$ , SCs for example. This consideration is likely why at the very large forecast horizon (720 min) we see strong decreases in the performance of the models. This is simply asking the models to forecast the higher thresholds too far in advance. We do note that the dropoff in performance is substantially less at the lowest threshold of  $R$ , indicating that the lower levels of activity that we are forecasting in this work are more common and predictable at a longer lead time.

Whilst the models provide comparable performance over most of the combinations of thresholds, inputs and horizons tested in Figure 6, there are several cases where the recurrent GRU model notably underperforms the other types of model, particularly at high thresholds. These cases would correspond to those with the most severe class imbalances (fewest positive cases where the threshold is exceeded) and therefore less training examples. Nonetheless, all three models generally provide skillful forecasts.

Now we assess how the BSSs achieved by the models change as the input data window and forecast horizon are adjusted. Figure 7 shows the BSS metrics, in the same format as Figure 6. We find that in the benchmark case (Figure 7a, first shown in Figure 1c) the BSSs returned by the models are between 0.2 and 0.3, indicating good reliability.

As we increase the input data from 30 to 60 min and then 120 min, moving from Figures 7a–7c we find that the models increase in reliability at the lowest thresholds, but decrease at the largest. This is a similar result to that found above when considering the model skill. It is again likely that the fixed model architectures are not able to fully utilize the increasing quantity and complexity of the data input.

The reliability of the models as we increase the forecast horizons are also similar to those found when assessing the skill of the models. Increasing the horizon to 180 min provides moderate increases in performance ( $\Delta BSS \sim 0.05 - 0.1$ ), but increasing it further to 720 min provides a similar magnitude of performance decrease.

As with the skill based performance evaluation above, the three models all achieve similar reliability in their forecasts for almost all of the combinations tested in Figure 7, with a few exceptions where one model provides inferior reliability. The under-performing model in these few cases is often either the GRU or the Dense model.

## 4. Discussion

We will now discuss the ability of the models to provide skillful and reliable forecasts, and evaluate their potential use in forecasting ground-based space weather impacts.

### 4.1. Thresholds of Ground Variability

In this work we trained three models to forecast when fixed thresholds of ground magnetic field variability would be exceeded (e.g., Camporeale et al., 2020). This approach is in contrast to a more direct forecast of the ground magnetic field variability (c.f. Keese et al., 2020), providing a simpler problem framework for a machine learning model to solve. We showed excellent correspondence between the target and model forecasts in the examples in Figures 3 and 4. In particular, we have shown that during two example storms in March and June 2015 the models for each threshold skillfully and reliably represent the observed ground magnetic field variability (e.g., Figures 3 and 4).

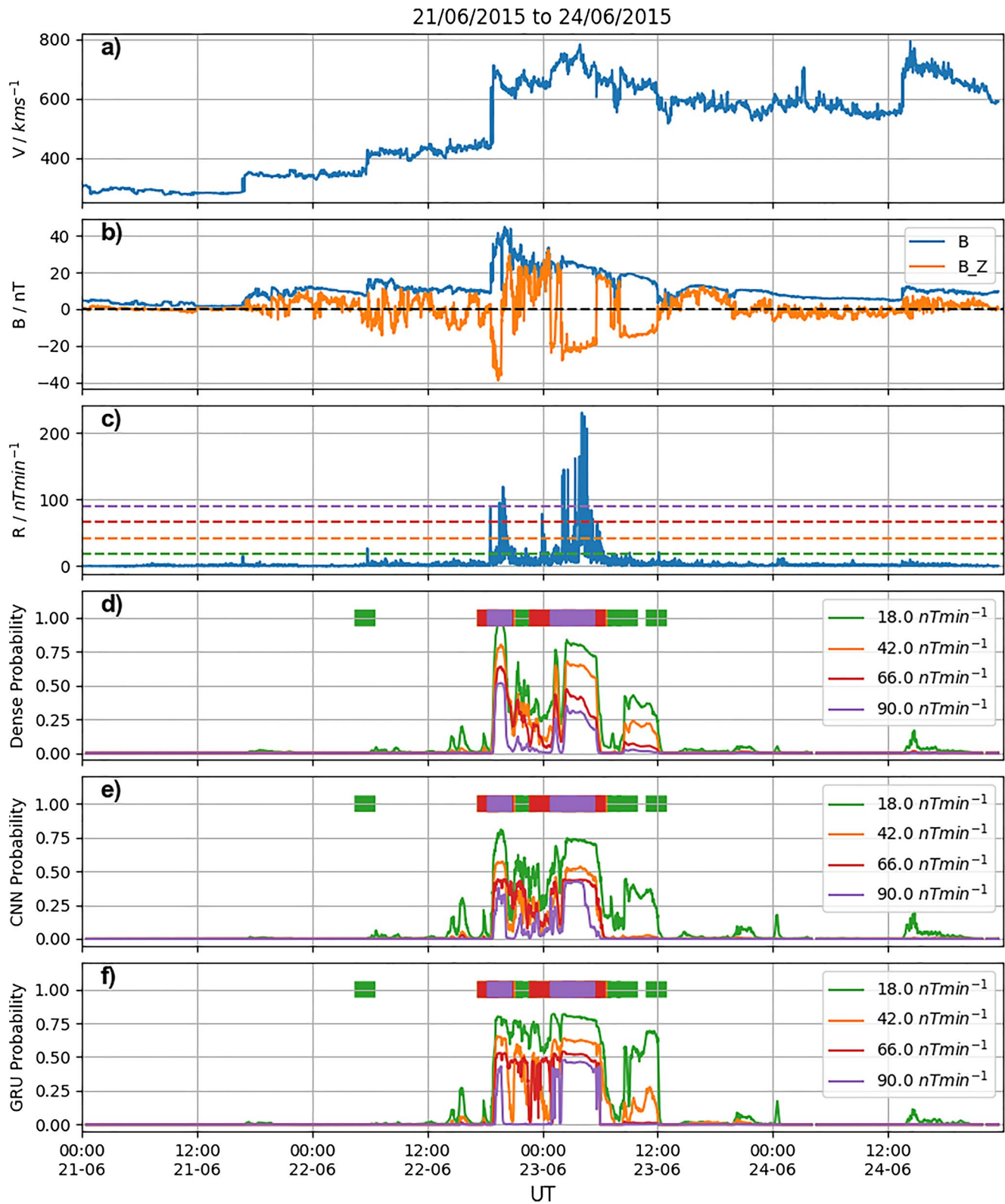
Our results further validate this approach, and suggest that it is a promising method to continue to explore in the future. However, a consideration that we note from this work is the different and increasing class imbalance present when using higher thresholds. These higher thresholds are exceeded less often, and so fewer “positive” examples are present during the training of the models. This can limit the complexity of the model architectures that can be employed and ultimately impact the performance. In future, distinct and tailored architectures would make the most of the available training data at each threshold.

### 4.2. Input Window

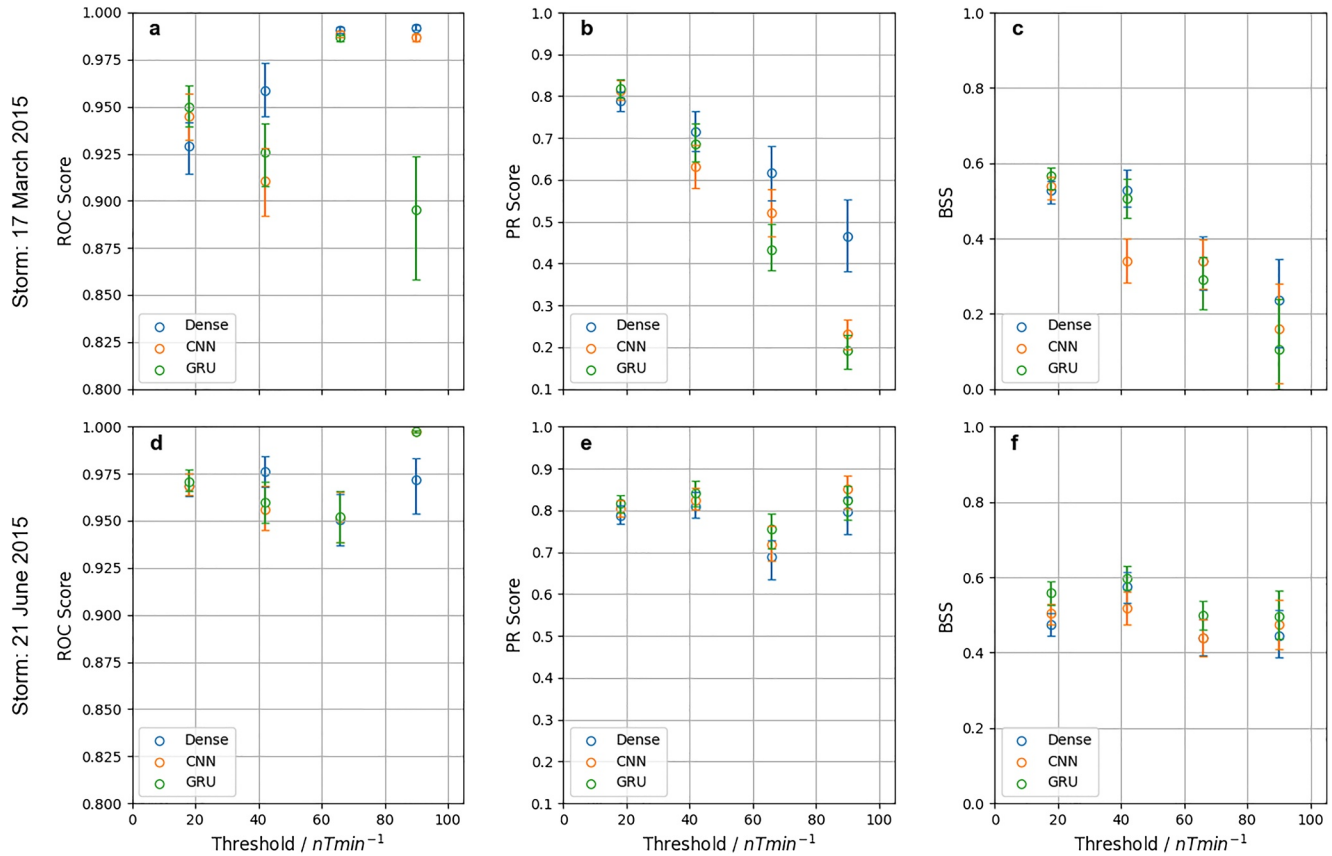
It might be expected that providing the models with additional input data, for example, increasing the window from 30 min to an hour, would have increased the skill and reliability of the models. It was thought that the additional information would provide the models with important context, given the variety of physical mechanisms that can cause elevated rates of change of the magnetic field on the ground. For example, this would allow the models to “know” that there was an historical southward turning of the IMF for a longer interval. However, we find little evidence that providing the extra data to the models increases their performance significantly. This is likely a result of the limited “positive” input data which we use to train the models, which also means less complex model architectures can be trained. As such, when forecasting rare events, associated with small positive sample sizes, there is a limit on the machine learning architectures that can be supported without overfitting, and this has an effect of limiting the historical information that can be effectively processed and usefully incorporated.

When providing a warning interval for space weather, it is desirable to provide as great a warning period as possible. We have assessed the ability of the models to provide forecasts at three different desired horizons: 30 min, 180 and 720 min. From the perspective of the major space weather phenomena responsible for elevated variability of the ground magnetic field we have selected these horizons to approximately correspond to the substorm expansion phase, substorm length and the duration of the main phase of a geomagnetic storm respectively. We note that in this work we have employed OMNI data, which is propagated to the bow shock, severely limiting the forecast that can be provided for some phenomena, SCs for example. Nonetheless, it was considered that longer forecasts horizons may be possible when considering large scale coupling of the solar wind and magnetosphere. There is also a balance when considering the performance of a given model. If a short horizon is used then the models are, in effect, being asked to narrowly and precisely





**Figure 4.** The performance of the models for a 4-day interval during the June 2015 storm. The format is the same as in Figure 3.

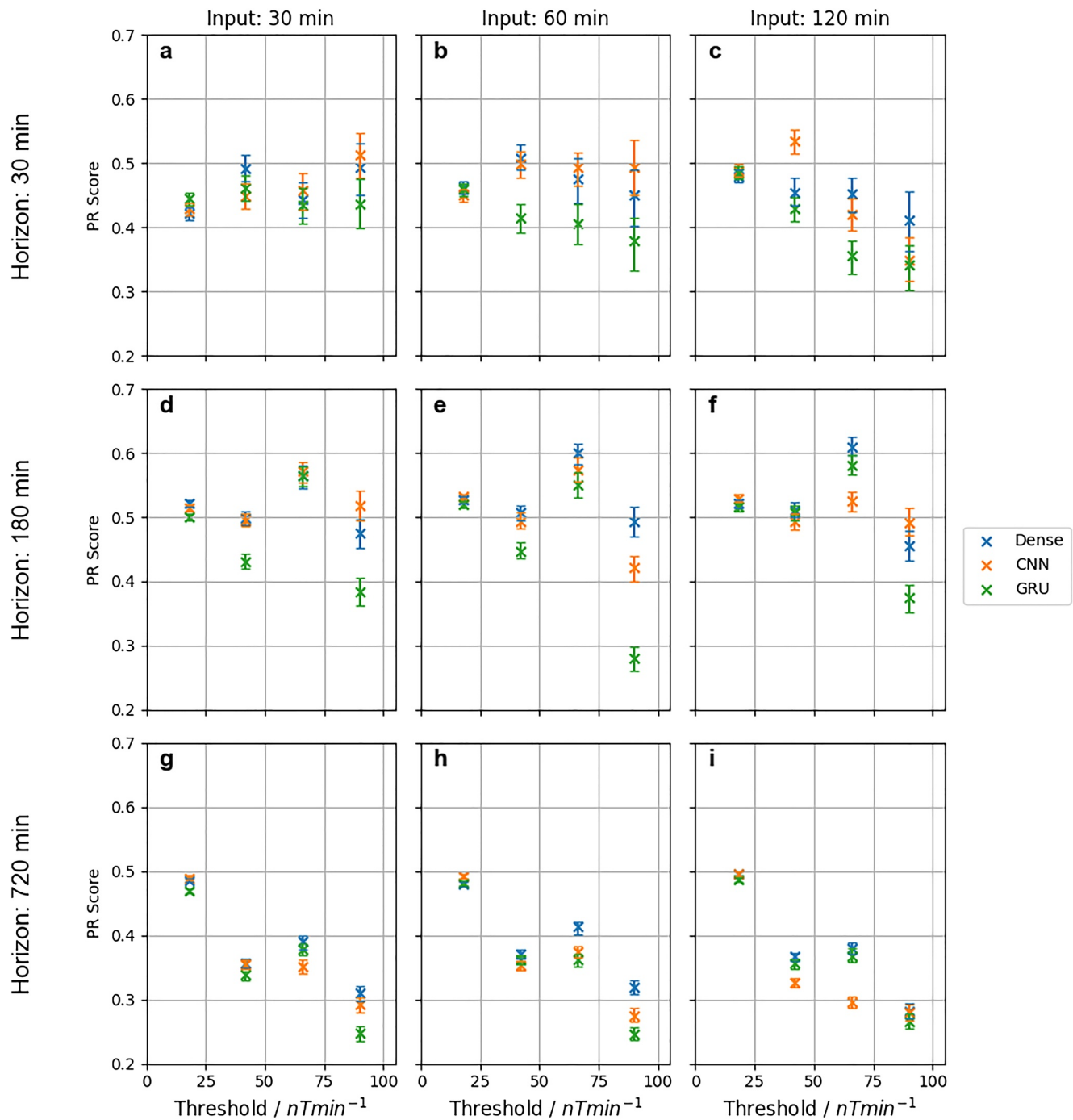


**Figure 5.** (a and d) The receiver-operating characteristic, (b and e) precision-recall and (c and f) Brier skill score metrics achieved by the models on two example storms in 2015 as a function of threshold for the Lerwick magnetometer station using 30 min of solar wind history and a horizon of 30 min. The time period evaluated is extended for five days before and after the storm periods in Figures 3 and 4. The results for the March 17th, 2015 storm are shown in panels a–c, while the results for the June 21st, 2015 storm are shown in panels d–f. The uncertainty in the values are calculated using a bootstrap method and represent the 95% confidence intervals from 100 iterations.

identify when phenomena of interest are going to occur, which is challenging for some phenomena (e.g., substorms) (c.f. Maimaiti et al., 2019). On the other hand if a long horizon is requested by stakeholders, then the models are asked to make predictions far into the future and gauge the impact of impinging solar wind that has not yet been observed upstream of the Earth.

We showed that increasing the horizon from 30 to 180 min provided increases in model performance, which we attribute to the models being able to forecast certain solar wind-magnetosphere coupling phenomena with less precision. Meanwhile, increasing this horizon to 720 min was associated with a strong decrease in all performance metrics. This suggests that either solely relying on data from upstream of the Earth or the limited model architectures are not capable of making as skillful forecasts at horizons of this length.

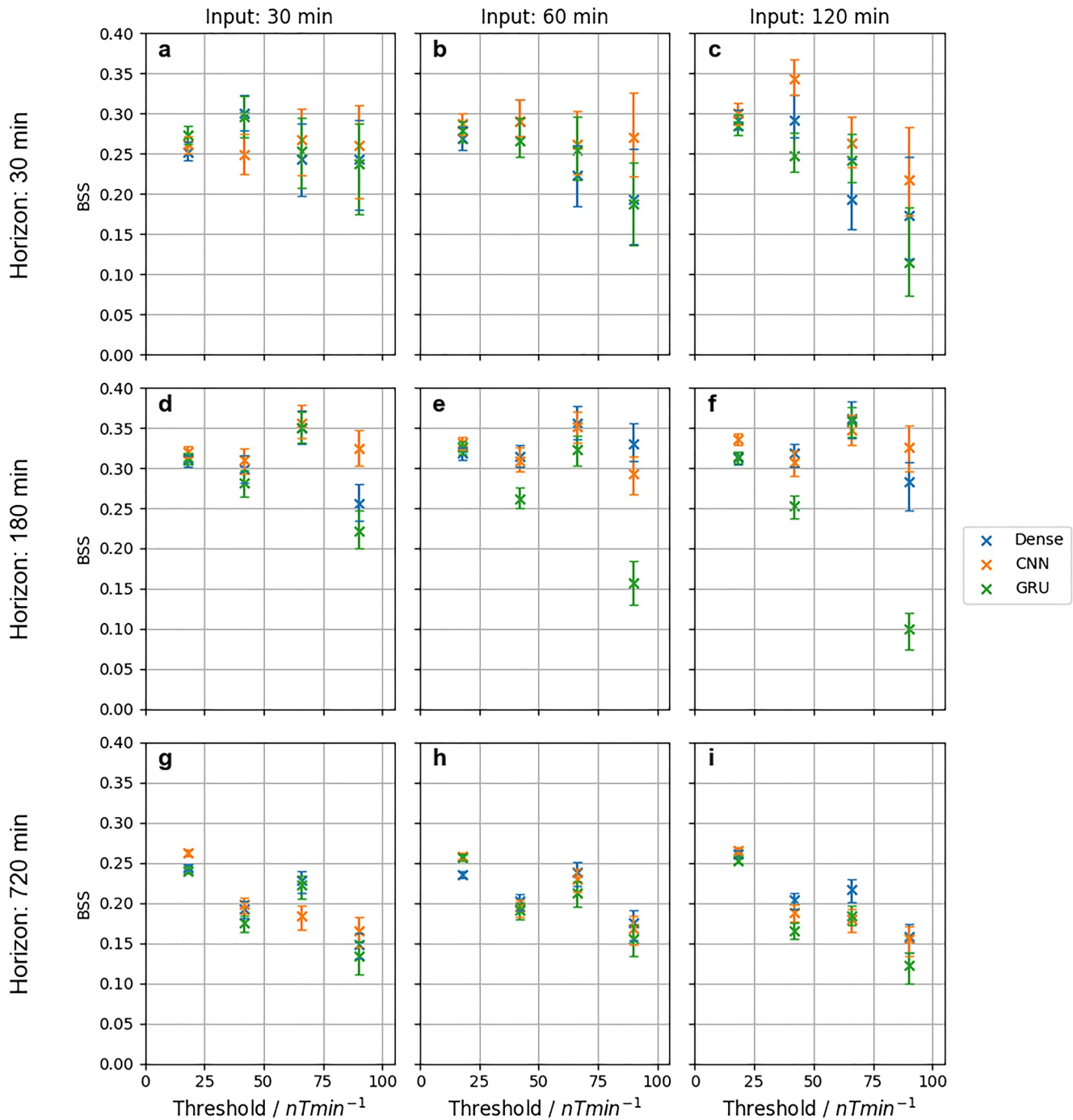
If a longer horizon is required, beyond that which the upstream in-situ data can provide, then it may be necessary to include input from other sources. Two suggested sources are the L5 point (e.g., Bailey et al., 2020; Owens et al., 2019; Thomas et al., 2018), or observations of the solar disk (e.g., Chakraborty & Morley, 2020; Tasistro-Hart et al., 2021). Future forecast models of the ground magnetic field variability could use these to extend the forecast horizon with which it is possible to obtain skillful results. The data from the solar disk in particular may be of use in forecasting impulsive solar wind phenomena, though a large lead time of these observations, that is, beyond a few hours, may be required.



**Figure 6.** The precision-recall (PR) Score achieved by the models as a function of threshold assessed for combinations of input window lengths and forecast horizons. The models have been assessed using the 2-year test data set (2015–2016). Columns are shown for input windows of (a, d, and g) 30, (b, e, and h) 60 and (c, f, and i) 120 min. Rows are shown for forecast horizons of (a, b, and c) 30, (d, e, and f) 180 and (g, h, and i) 720 min. The uncertainty in the PR Scores is calculated using a bootstrap method.

### 4.3. Neural Network Variant Performance Comparison

For the vast majority of input windows, thresholds and forecast horizons (parameters) the three types of neural network model tested (Dense, CNN and GRU) perform similarly, that is, within the bootstrap defined uncertainty. On the other hand, there are a few specific combinations of parameters where one model significantly outperforms or under-performs the other models. Some of these scenarios may be due to the



**Figure 7.** The Brier skill score achieved by the models as a function of threshold assessed for combinations of input window lengths and forecast horizons. The format is as in Figure 6.

random training process and initialization, however some could be due to the inherent training setup and model architectures. For example, the GRU models seem to under-perform at high thresholds, when the class imbalance between the number of “positive” and “negative” cases are most extreme. Additionally, when more input data were provided the Dense and CNN models required more trainable parameters to deal with the input, and perhaps became more limited by the class imbalance with more training required from the same data. In contrast, the GRU model did not require more trainable parameters, but the already limited architecture was not necessarily able to effectively process the additional information.

This highlights the need to refine the network architecture for each parameter combination, accounting for the input samples for each class. However, with the presented setup it appears that the Dense or CNN architectures would provide the greatest performance over most combinations of parameters.

#### 4.4. Implications for Forecasting and Future Development

Overall, these models show reliability and skill in forecasting intervals when the rate of change of the ground magnetic field exceeds high thresholds. In particular the presented models appear to excel at forecasting large scale phenomena driven by periods of enhanced coupling between the solar wind and magnetosphere, but may not satisfactorily forecast impulsive phenomena, such as SCs.

In order to better capture impulsive phenomena, other data sources could be explored, for example, the unpropagated data from ACE or DSCOVR (e.g., Wintoft et al., 2015). While this would provide a variable time delay from the solar wind data to the magnetopause, it would give the models the opportunity to forecast the future impact of phenomena such as solar wind shocks, and resulting SCs. Further, if such a system were intended to work in near-real time then it would be desirable to use the data that would be available on such timescales, instead of the fully calibrated science level data.

Figure 3 showed an example of the models output when applied to a geomagnetic storm in March 2015. During this storm the data from L1 was interrupted with several data gaps. The timing of these data gaps prevented the models from forecasting during the interval when the maximum  $R$  was observed in the storm, potentially due to saturation of the instrument during extreme solar wind conditions (e.g., Nicolaou et al., 2020). On a quantitative level, this was inferred to reduce the performance of the models on the derived metrics (e.g., Figure 5). However, more qualitatively it is distinctly undesirable for space weather forecasting models to be susceptible to such data drop outs. Future work should investigate other methods of providing a forecast during these intervals. For example, a model could be created for the specific circumstances when the data are unavailable, which may occur predominantly during extreme solar wind. Alternatively, if the data are missing then the last recorded data point could be repeated. In the case of the storm in March 2015 the last known data showed an elevated velocity ( $\sim 580 \text{ km s}^{-1}$ ) and strongly negative  $B_z$ , indicating that strong magnetospheric coupling is likely. With this method at least some forecasting capacity would remain. Finally, the models used for this study require continuous input data, with no gaps. If a large input window (e.g., 2 h) is used then the full period must have complete data to provide a forecast. However, models with shorter inputs, shown here to potentially provide at least comparable performance (e.g., Figures 6 and 7), would come back “online” faster and could be used in the interim period.

The models used here could be easily adapted to forecast the ground magnetic field at other ground magnetometer stations. This would provide estimates at other latitudes and local times, noting the dependence of the rate of change of the field on local features such as geology (e.g., Dimmock et al., 2020; Ngwira et al., 2015). At different latitudes other magnetospheric phenomena will be more important (e.g., Rogers et al., 2020; Smith et al., 2021), which may impact the performance of the models with different combinations of input window and forecast horizons.

## 5. Summary

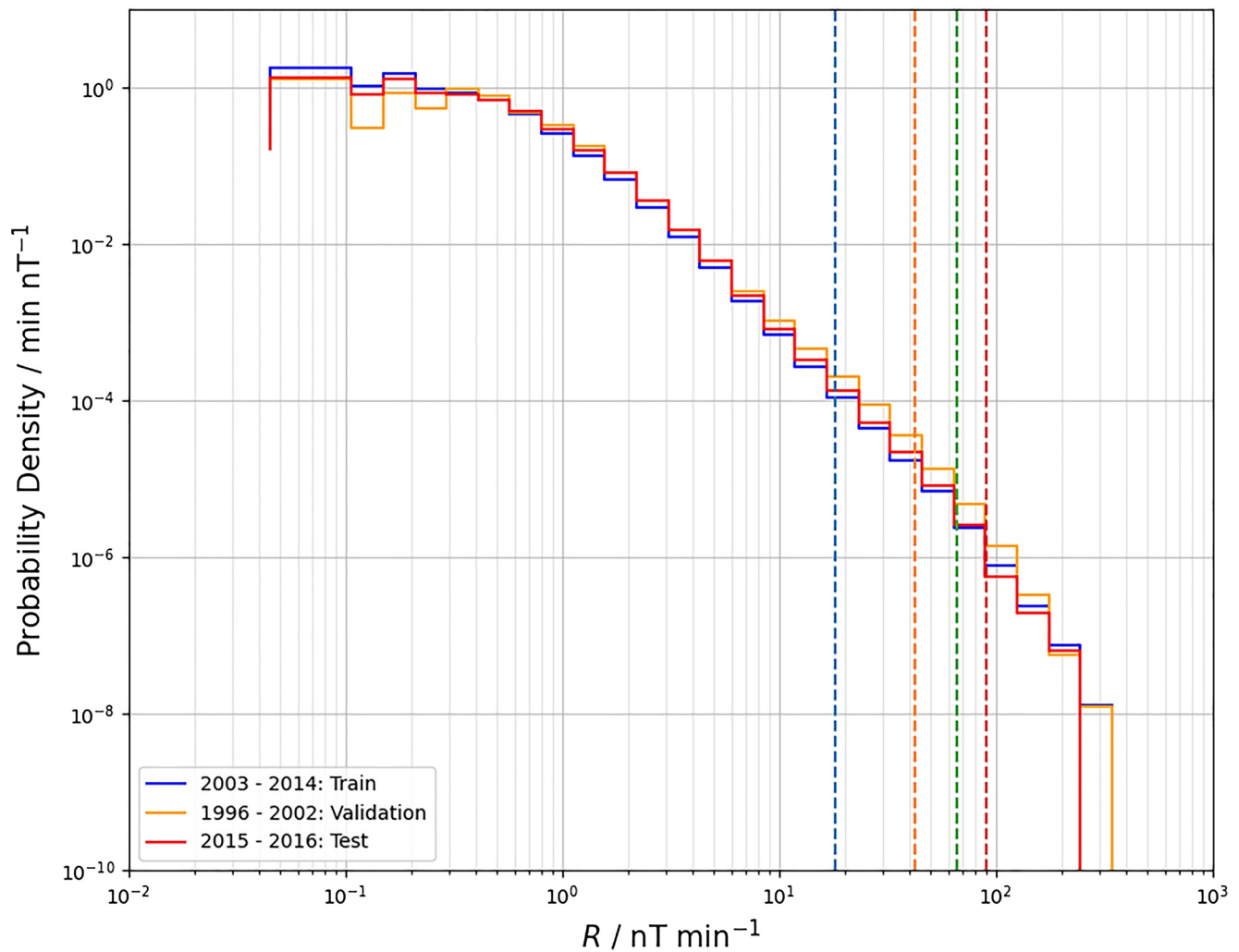
In this study we have created and analyzed a series of Machine Learning models to provide a probabilistic forecast as to whether ground magnetic perturbations ( $R$ ) at the LER ground station in the UK will exceed a given threshold. We tested three models: a “Dense” feed forward neural network with two hidden layers, a convolutional network (CNN) and a recurrent network (GRU). The models were trained using 12 years of data, with 6 years used to validate the training, and 2 years of unseen data was used to evaluate the model performance. Additionally, the models were evaluated over a range of input data interval lengths, thresholds of magnetic field variability and forecast horizons. We summarize our key results below:

1. With 30 min of input data, and forecasting 30 min into the future, the models are reliable and skillful with BSS of  $\sim 0.25$ , ROC scores of  $\sim 0.95$  and PR scores of  $\sim 0.45$ .

2. Limiting the evaluation to two example magnetospheric storms during 2015, we find that the models' performance was increased in these intervals. All three metrics increased, though the PR and BSSs increased most considerably. This perhaps indicates the models perform best during extended periods of strong coupling to the solar wind (i.e., storms), and less well during more sporadic activity that would be found throughout the test data set.
3. We note that during short, sharp increases in solar wind density and velocity, that is, solar wind shocks, the corresponding ground responses (SCs) are not well forecast. This can be attributed to the use of propagated solar wind data (i.e., OMNI) not providing sufficient lead time.
4. Increasing the data provided to the models to 60 or 120 min of history from the solar wind resulted in only slight increases, or most often decreases in performance. This was inferred to be a result of the fixed model architectures and limited training data.
5. Increasing the forecast horizon from 30 to 180 min did increase the performance of the models moderately, we infer that though this requires the models to predict further into the future it allows the models to be less precise in forecasting when a magnetospheric phenomenon may occur.
6. Over most combinations of threshold, horizons and input window the three models perform similarly, to within the bootstrapped uncertainties in the metrics. However, the GRU model is noted to occasionally under perform in the cases where there are fewest "positive" examples with which to train the models. Models specifically tailored to each combination of input window, threshold and forecast horizon would be able to ameliorate this effect, however in the current setup the Dense and CNN models are found to perform best over most combinations.

Overall we have shown that machine learning models can make skillful and reliable predictions of when the ground magnetic perturbations at the LER station will exceed several thresholds. Future work can further develop these models to better represent impulsive phenomena, deal with missing data and perform optimally for the desired forecast horizon and threshold of variability.

Appendix A: Cross Validation: Distributions of R



**Figure A1.** Distributions of  $R$  at the Lerwick magnetometer station in the train, test and validation data sets. The vertical dashed lines indicate the 18, 42, 66 and 90  $\text{nTmin}^{-1}$  thresholds employed in the study, and discussed in Section 2.2.

Data Availability Statement

The magnetometer data used in this study were collected at the Lerwick observatory. We thank the British Geological Survey for supporting their operation and INTERMAGNET for promoting high standards of magnetic observatory practice ([www.intermagnet.org](http://www.intermagnet.org)). The data were downloaded from <https://intermagnet.github.io> and are freely available there. We acknowledge and thank NASA GSFC's Space Physics Data Facility's OMNIWeb (or CDAWeb or ftp) service for the use of OMNI data (<https://omniweb.gsfc.nasa.gov>). The analysis in this paper was performed using python, including the TensorFlow (Abadi et al., 2015), pandas (McKinney, 2010), numpy (Van Der Walt et al., 2011), scikit-learn (Pedregosa et al., 2011), scipy (Virtanen et al., 2020), and matplotlib (Hunter, 2007) libraries.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., & Research, G. (2015). *TensorFlow: Large-Scale machine learning on heterogeneous distributed systems*. Retrieved from [www.tensorflow.org](http://www.tensorflow.org)  
 Akasofu, S.-I. (1964). The development of the auroral substorm. *Planetary and Space Science*, 12(4), 273–282. [https://doi.org/10.1016/0032-0633\(64\)90151-5](https://doi.org/10.1016/0032-0633(64)90151-5)

Acknowledgments

A. W. Smith and C. Forsyth were supported by STFC consolidated grant ST/S000240/1, and NERC grants NE/P017150/1 and NE/V002724/1. C. Forsyth was also supported by the NERC Independent Research Fellowship NE/N014480/1. I. J. Rae was supported by STFC consolidated grant ST/V006320/1. C. M. Jackman's work was supported by the Science Foundation Ireland grant 18/FRL/6199. M. R. Bakrania was supported by a UCL Impact Studentship, joint funded by the ESA NPI program.

- Akasofu, S.-I., & Chao, J. (1980). Interplanetary shock waves and magnetospheric substorms. *Planetary and Space Science*, 28(4), 381–385. [https://doi.org/10.1016/0032-0633\(80\)90042-2](https://doi.org/10.1016/0032-0633(80)90042-2)
- Akasofu, S.-I., & Chapman, S. (1961). The ring current, geomagnetic disturbance and the Van Allen radiation belts. *Journal of Geophysical Research*, 66(5), 1321–1350. <https://doi.org/10.1029/JZ066i005p01321>
- Araki, T. (1994). A physical model of the geomagnetic sudden commencement. In M. Engebretson, K. Takahashi, & M. Scholer (Eds.), *Solar wind sources of magnetospheric ultra-low-frequency waves* (p. 183).
- Astafeyeva, E., Zakharenkova, I., & Förster, M. (2015). Ionospheric response to the 2015 St. Patrick's Day storm: A global multi-instrumental overview. *Journal of Geophysical Research: Space Physics*, 120(10), 9023–9037. <https://doi.org/10.1002/2015JA021629>
- Azari, A. R., Liemohn, M. W., Jia, X., Thomsen, M. F., Mitchell, D. G., Sergis, N., et al. (2018). Interchange injections at Saturn: Statistical survey of energetic H<sup>+</sup> sudden flux intensifications. *Journal of Geophysical Research: Space Physics*, 123, 4692–4711. <https://doi.org/10.1029/2018JA025391>
- Bailey, R. L., Möstl, C., Reiss, M. A., Weiss, A. J., Amerstorfer, U. V., Amerstorfer, T., et al. (2020). Prediction of Dst during solar minimum using in situ measurements at L5. *Space Weather*, 18(5), e2019SW002424. <https://doi.org/10.1029/2019SW002424>
- Bakrania, M. R., Rae, I. J., Walsh, A. P., Verscharen, D., & Smith, A. W. (2020). Using dimensionality reduction and clustering techniques to classify space plasma regimes. *Frontiers in Astronomy and Space Sciences*, 7, 80. <https://doi.org/10.3389/fspas.2020.593516>
- Bedrosian, P. A., & Love, J. J. (2015). Mapping geoelectric fields during magnetic storms: Synthetic analysis of empirical United States impedances. *Geophysical Research Letters*, 42(23), 10160–10170. <https://doi.org/10.1002/2015GL066636>
- Beggan, C. D. (2015). Sensitivity of geomagnetically induced currents to varying auroral electrojet and conductivity models. *Earth, Planets and Space*, 67(1), 24. <https://doi.org/10.1186/s40623-014-0168-9>
- Beland, J., & Small, K. (2004). Space weather effects on power transmission systems: The cases of Hydro-Quebec and transpower New Zealand Ltd [Proceedings Paper]. In I. Daglis (Ed.), *Effects of space weather on technology infrastructure* (Vol. 176, pp. 287–299). Dordrecht, Netherlands: Springer.
- Bentley, S. N., Stout, J., Bloch, T. E., & Watt, C. E. J. (2020). Random forest model of ultralow frequency magnetospheric wave power. *Earth and Space Science*, 7, e2020EA001274. <https://doi.org/10.1029/2020EA001274>
- Bhaskar, A., & Vichare, G. (2019). Forecasting of SYMH and ASYH indices for geomagnetic storms of solar cycle 24 including St. Patrick's Day, 2015 storm using NARX neural network. *Journal of Space Weather and Space Climate*, 9, A12. <https://doi.org/10.1051/swsc/2019007>
- Blake, S. P., Gallagher, P. T., McCauley, J., Jones, A. G., Hogg, C., Campaña, J., et al. (2016). Geomagnetically induced currents in the Irish power network during geomagnetic storms. *Space Weather*, 14(12), 1136–1154. <https://doi.org/10.1002/2016SW001534>
- Bloch, T., Watt, C. E. J., Owens, M. J., Thompson, R. L., & Agiwal, O. (2021). Constraining the location of the outer boundary of Earth's outer radiation belt. *Earth and Space Science*, 8, e2020EA001610. <https://doi.org/10.1029/2020EA001610>
- 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](https://doi.org/10.1016/S1364-6826(02)00128-1)
- Bolduc, L., Langlois, P., Boteler, D., & Pirjola, R. (1998). A study of geoelectromagnetic disturbances in Quebec. I. General results. *IEEE Transactions on Power Delivery*, 13(4), 1251–1256. <https://doi.org/10.1109/61.714492>
- Bortnik, J., Li, W., Thorne, R. M., & Angelopoulos, V. (2016). A unified approach to inner magnetospheric state prediction. *Journal of Geophysical Research: Space Physics*, 121(3), 2423–2430. <https://doi.org/10.1002/2015JA021733>
- Boteler, D. (2014). Methodology for simulation of geomagnetically induced currents in power systems. *Journal of Space Weather and Space Climate*, 4, A21. <https://doi.org/10.1051/swsc/2014018>
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- Breckner, G. E., Delaboudiniere, J.-P., Howard, R. A., Paswaters, S. E., Cyr, O. S., Schwenn, R., et al. (1998). Geomagnetic storms caused by coronal mass ejections (CMEs): March 1996 through June 1997. *Geophysical Research Letters*, 25(15), 3019–3022. <https://doi.org/10.1029/98GL00704>
- Camporeale, E. (2019). The challenge of machine learning in space weather: Nowcasting and forecasting. *Space Weather*, 17(8), 1166–1207. <https://doi.org/10.1029/2018SW002061>
- Camporeale, E., Cash, M. D., Singer, H. J., Balch, C. C., Huang, Z., & Toth, G. (2020). A graybox model for a probabilistic estimate of regional ground magnetic perturbations: Enhancing the NOAA operational Geospace model with machine learning. *Journal of Geophysical Research: Space Physics*, 125, e2019JA027684. <https://doi.org/10.1029/2019JA027684>
- Carter, B. A., Pradipta, R., Zhang, K., Yizengaw, E., Halford, A. J., & Norman, R. (2015). Interplanetary shocks and the resulting geomagnetically induced currents at the equator. *Geophysical Research Letters*, 42(16), 6554–6559. <https://doi.org/10.1002/2015gl065060>
- Carter, B. A., Yizengaw, E., Pradipta, R., Weygand, J. M., Piersanti, M., Pulkkinen, A., et al. (2016). Geomagnetically induced currents around the world during the 17 March 2015 storm. *Journal of Geophysical Research: Space Physics*, 121, 10496–10507. <https://doi.org/10.1002/2016JA023344>
- Chakraborty, S., & Morley, S. K. (2020). Probabilistic prediction of geomagnetic storms and the  $K_p$  index. *Journal of Space Weather and Space Climate*, 10, 36. <https://doi.org/10.1051/swsc/2020037>
- 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>
- Chen, P. F. (2011). Coronal mass ejections: Models and their observational basis. *Living Reviews in Solar Physics*, 8(1), 1–92. <https://doi.org/10.12942/lrsp-2011-1>
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Emnlp 2014 - 2014 conference on empirical methods in natural language processing, proceedings of the conference* (pp. 1724–1734). <https://doi.org/10.3115/v1/d14-1179>
- Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. In *Nips 2014 deep learning and representation learning workshop*. Retrieved from <http://arxiv.org/abs/1412.3555>
- Clausen, L. B. N., & Nickisch, H. (2018). Automatic classification of auroral images from the Oslo Auroral THEMIS (OATH) data set using machine learning. *Journal of Geophysical Research: Space Physics*, 123, 5640–5647. <https://doi.org/10.1029/2018JA025274>
- Cilverd, M. A., Rodger, C. J., Brundell, J. B., Dalzell, M., Martin, I., Mac Manus, D. H., et al. (2018). Long-lasting geomagnetically induced currents and harmonic distortion observed in New Zealand during the 7–8 September 2017 disturbed period. *Space Weather*, 16(6), 704–717. <https://doi.org/10.1029/2018SW001822>
- Coxon, J. C., Shore, R. M., Freeman, M. P., Fear, R. C., Browett, S. D., Smith, A. W., et al. (2019). Timescales of Birkeland currents driven by the IMF. *Geophysical Research Letters*, 46(14), 7893–7901. <https://doi.org/10.1029/2018GL081658>



- Crown, M. D. (2012). Validation of the NOAA Space Weather Prediction Center's solar flare forecasting look-up table and forecaster-issued probabilities. *Space Weather*, 10(6), S06006. <https://doi.org/10.1029/2011SW000760>
- Curto, J. J., Araki, T., & Alberca, L. F. (2007). Evolution of the concept of Sudden Storm Commencements and their operative identification. *Earth, Planets and Space*, 59(11), i–xii. <https://doi.org/10.1186/BF03352059>
- Daglis, I. A., Thorne, R. M., Baumjohann, W., & Orsini, S. (1999). The terrestrial ring current: Origin, formation, and decay. *Reviews of Geophysics*, 37(4), 407–438. <https://doi.org/10.1029/1999RG900009>
- Davis, J., & Goadrich, M. (2006). *The relationship between precision-recall and ROC curves* (Tech. Rep.). Retrieved from [https://www.biostat.wisc.edu/~sim\\$Page/rocpr.pdf](https://www.biostat.wisc.edu/~sim$Page/rocpr.pdf)
- Davis, T. N., & Sugiura, M. (1966). Auroral electrojet activity index *AE* and its universal time variations. *Journal of Geophysical Research*, 71(3), 785–801. <https://doi.org/10.1029/JZ071i003p00785>
- Dimmock, A., Rosenqvist, L., Hall, J., Viljanen, A., Yordanova, E., Honkonen, I., et al. (2019). The GIC and geomagnetic response over Fennoscandia to the 7–8 September 2017 geomagnetic storm. *Space Weather*, 17, 989–1010. <https://doi.org/10.1029/2018SW002132>
- Dimmock, A. P., Rosenqvist, L., Welling, D. T., Viljanen, A., Honkonen, I., Boynton, R. J., & Yordanova, E. (2020). On the regional variability of dB/dt and its significance to GIC. *Space Weather*, 18(8), e2020SW002497. <https://doi.org/10.1029/2020SW002497>
- Divett, T., Richardson, G. S., Beggan, C. D., Rodger, C. J., Boteler, D. H., Ingham, M., et al. (2018). Transformer-level modeling of geomagnetically induced currents in New Zealand's South Island. *Space Weather*, 16(6), 718–735. <https://doi.org/10.1029/2018SW001814>
- Engelbreton, M. J., Pilipenko, V. A., Steinmetz, E. S., Moldwin, M. B., Connors, M. G., Boteler, D. H., et al. (2021). Nighttime magnetic perturbation events observed in Arctic Canada: 3. Occurrence and amplitude as functions of magnetic latitude, local time, and magnetic disturbance indices. *Space Weather*, 19, e2020SW002526. <https://doi.org/10.1029/2020SW002526>
- Fiori, R. A. D., Boteler, D. H., & Gillies, D. M. (2014). Assessment of GIC risk due to geomagnetic sudden commencements and identification of the current systems responsible. *Space Weather*, 12(1), 76–91. <https://doi.org/10.1002/2013SW000967>
- Forsyth, C., Rae, I. J., Coxon, J. C., Freeman, M. P., Jackman, C. M., Gjerloev, J., & Fazakerley, A. N. (2015). A new technique for determining Substorm Onsets and Phases from Indices of the Electrojet (SOPHIE). *Journal of Geophysical Research: Space Physics*, 120, 10592–10606. <https://doi.org/10.1002/2015JA021343>
- Forsyth, C., Shortt, M., Coxon, J. C., Rae, I. J., Freeman, M. P., Kalmoni, N. M. E., et al. (2018). Seasonal and temporal variations of field-aligned currents and ground magnetic deflections during substorms. *Journal of Geophysical Research: Space Physics*, 123(4), 2696–2713. <https://doi.org/10.1002/2017JA025136>
- Forsyth, C., Watt, C. E. J., Mooney, M. K., Rae, I. J., Walton, S. D., & Horne, R. B. (2020). Forecasting GOES 15 > 2 MeV electron fluxes from solar wind data and geomagnetic indices. *Space Weather*, 18, e2019SW002416. <https://doi.org/10.1029/2019SW002416>
- Freeman, M. P., Forsyth, C., & Rae, I. J. (2019). The influence of substorms on extreme rates of change of the surface horizontal magnetic field in the United Kingdom. *Space Weather*, 17, 827–844. <https://doi.org/10.1029/2018SW002148>
- Garton, T. M., Jackman, C. M., Smith, A. W., Yeakel, K. L., Maloney, S. A., & Vandegriff, J. (2021). Machine learning applications to Kronian magnetospheric reconnection classification. *Frontiers in Astronomy and Space Sciences*, 7, 104. <https://doi.org/10.3389/fspas.2020.600031>
- Gonzalez, W. D., Joselyn, J. A., Kamide, Y., Kroehl, H. W., Ros, G., Tsuru, B. T., & Vasyliunas, V. M. (1994). *What is a geomagnetic storm?* (Vol. 99; Tech. 1156 Rep. No. A4). <https://doi.org/10.1029/93JA02867>
- Gosling, J., & Pizzo, V. (1999). Formation and evolution of corotating interaction regions and their three dimensional structure. *Space Science Reviews*, 89(1/2), 21–52. <https://doi.org/10.1023/A:1005291711900>
- 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>
- Heaton, J. (2008). *Introduction to neural networks with Java*. Heaton Research. Retrieved from <https://books.google.it/books?id=Swlcw7M4uD8C&pg=PA158&lpg=PA158&dq=IntroductiontoNeuralNetworksforJava>
- Heyns, M. J., Lotz, S. I., & Gaunt, C. T. (2020). Geomagnetic pulsations driving geomagnetically induced currents. *Space Weather*, 19, e2020SW002557. <https://doi.org/10.1029/2020SW002557>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Jacobs, J. A., Kato, Y., Matsushita, S., & Troitskaya, V. A. (1964). Classification of Geomagnetic Micropulsations. *Geophysical Journal of the Royal Astronomical Society*, 8(3), 341–342. <https://doi.org/10.1111/j.1365-246X.1964.tb06301.x>
- James, M. K., Imber, S. M., Raines, J. M., Yeoman, T. K., & Bunce, E. J. (2020). A machine learning approach to classifying MESSENGER FIPS proton spectra. *Journal of Geophysical Research: Space Physics*, 125(6), e2019JA027352. <https://doi.org/10.1029/2019JA027352>
- Ji, E.-Y., Moon, Y.-J., Park, J., Lee, J.-Y., & Lee, D.-H. (2013). Comparison of neural network and support vector machine methods for Kp forecasting. *Journal of Geophysical Research: Space Physics*, 118(8), 5109–5117. <https://doi.org/10.1002/jgra.50500>
- Jonas, E., Bobra, M., Shankar, V., Todd Hoeksema, J., & Recht, B. (2018). Flare prediction using photospheric and coronal image data. *Solar Physics*, 293(3), 48. <https://doi.org/10.1007/s11207-018-1258-9>
- Kappenman, J. G. (2003). Storm sudden commencement events and the associated geomagnetically induced current risks to ground-based systems at low-latitude and midlatitude locations. *Space Weather*, 1(3), 1016. <https://doi.org/10.1029/2003SW000009>
- Kappenman, J. G. (2005). An overview of the impulsive geomagnetic field disturbances and power grid impacts associated with the violent Sun-Earth connection events of 29–31 October 2003 and a comparative evaluation with other contemporary storms. *Space Weather*, 3(8), S08C01. <https://doi.org/10.1029/2004SW000128>
- Keesee, A. M., Pinto, V., Coughlan, M., Lennox, C., Mahmud, M. S., & Connor, H. K. (2020). Comparison of deep learning techniques to model connections between solar wind and ground magnetic perturbations. *Frontiers in Astronomy and Space Sciences*, 7, 72. <https://doi.org/10.3389/fspas.2020.550874>
- Kelly, G. S., Viljanen, A., Beggan, C. D., & Thomson, A. W. (2017). Understanding GIC in the UK and French high-voltage transmission systems during severe magnetic storms. *Space Weather*, 15(1), 99–114. <https://doi.org/10.1002/2016SW001469>
- Kepko, L., McPherron, R. L., Amm, O., Apatenkov, S., Baumjohann, W., Birn, J., & Sergeev, V. (2015). Substorm current wedge revisited. *Space Science Reviews*, 190(1–4), 1–46. <https://doi.org/10.1007/s11214-014-0124-9>
- Kilpua, E., Koskinen, H. E. J., & Pulkkinen, T. I. (2017). Coronal mass ejections and their sheath regions in interplanetary space. *Living Reviews in Solar Physics*, 14(1), 1–83. <https://doi.org/10.1007/s41116-017-0009-6>

- Kilpua, E. K. J., Lumme, E., Andreeva, K., Isavnin, A., & Koskinen, H. E. J. (2015). Properties and drivers of fast interplanetary shocks near the orbit of the Earth (1995–2013). *Journal of Geophysical Research: Space Physics*, *120*(6), 4112–4125. <https://doi.org/10.1002/2015JA021138>
- Kozyreva, O. V., Pilipenko, V. A., Belakhovsky, V. B., & Sakharov, Y. A. (2018). Ground geomagnetic field and GIC response to March 17, 2015, storm. *Earth, Planets and Space*, *70*(1), 157. <https://doi.org/10.1186/s40623-018-0933-2>
- Kugblenu, S., Taguchi, S., & Okuzawa, T. (1999). Prediction of the geomagnetic storm associated Dst index using an artificial neural network algorithm. *Earth, Planets and Space*, *51*(4), 307–313. <https://doi.org/10.1186/BF03352234>
- Kunduri, B. S., Maimaiti, M., Baker, J. B., Ruohoniemi, J. M., Anderson, B. J., & Vines, S. K. (2020). A deep learning-based approach for modeling the dynamics of AMPERE Birkeland currents. *Journal of Geophysical Research: Space Physics*, *125*(8), e2020JA027908. <https://doi.org/10.1029/2020JA027908>
- Leka, K. D., Park, S.-H., Kusano, K., Andries, J., Barnes, G., Bingham, S., & Terkildsen, M. (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>
- Lethy, A., El-Eraki, M. A., Samy, A., & Deebes, H. A. (2018). Prediction of the Dst index and analysis of its dependence on solar wind parameters using neural network. *Space Weather*, *16*(9), 1277–1290. <https://doi.org/10.1029/2018SW001863>
- Li, H., Wang, C., & Peng, Z. (2013). Solar wind impacts on growth phase duration and substorm intensity: A statistical approach. *Journal of Geophysical Research: Space Physics*, *118*(7), 4270–4278. <https://doi.org/10.1002/jgra.50399>
- Liemohn, M. W., McCollough, J. P., Jordanova, V. K., Ngwira, C. M., Morley, S. K., Cid, C., & Vasile, R. (2018). Model evaluation guidelines for geomagnetic index predictions. *Space Weather*, *16*(12), 2079–2102. <https://doi.org/10.1029/2018SW002067>
- Liu, L., Ge, X., Zong, W., Zhou, Y., & Liu, M. (2016). Analysis of the monitoring data of geomagnetic storm interference in the electrification system of a high-speed railway. *Space Weather*, *14*(10), 754–763. <https://doi.org/10.1002/2016SW001411>
- Liu, L., Zou, S., Yao, Y., & Wang, Z. (2020). Forecasting global ionospheric TEC using deep learning approach. *Space Weather*, *18*(11), e2020SW002501. <https://doi.org/10.1029/2020SW002501>
- Lühr, H., Schlegel, K., Araki, T., Rother, M., & Förster, M. (2009). Night-time sudden commencements observed by CHAMP and ground-based magnetometers and their relationship to solar wind parameters. *Annales Geophysicae*, *27*(5), 1897–1907. <https://doi.org/10.5194/angeo-27-1897-2009>
- Lundstedt, H., Gleisner, H., & Wintoft, P. (2002). Operational forecasts of the geomagnetic Dst index. *Geophysics Research Letters*, *29*(24), 1016. <https://doi.org/10.1029/2002GL016151>
- Mac Manus, D. H., Rodger, C. J., Dalzell, M., Thomson, A. W. P., Chilverd, M. A., Petersen, T., & Divett, T. (2017). Long-term geomagnetically induced current observations in New Zealand: Earth return corrections and geomagnetic field driver. *Space Weather*, *15*(8), 1020–1038. <https://doi.org/10.1002/2017SW001635>
- Maimaiti, M., Kunduri, B., Ruohoniemi, J. M., Baker, J. B. H., & House, L. L. (2019). A deep learning-based approach to forecast the onset of magnetic substorms. *Space Weather*, *17*, 1534–1552. <https://doi.org/10.1029/2019SW002251>
- Mann, I. R., Milling, D. K., Rae, I. J., Ozeke, L. G., Kale, A., Kale, Z. C., & Singer, H. J. (2008). The upgraded CARISMA magnetometer array in the THEMIS era. *Space Science Reviews*, *141*(1–4), 413–451. <https://doi.org/10.1007/s11214-008-9457-6>
- Mann, I. R., Wright, A. N., Mills, K. J., & Nakariakov, V. M. (1999). Excitation of magnetospheric waveguide modes by magnetosheath flows. *Journal of Geophysical Research*, *104*(A1), 333–353. <https://doi.org/10.1029/1998JA900026>
- Marshalko, E., Kruglyakov, M., Kuvshinov, A., Murphy, B. S., Rastätter, L., Ngwira, C., & Pulkkinen, A. (2020). Exploring the influence of lateral conductivity contrasts on the storm time behavior of the ground electric field in the Eastern United States. *Space Weather*, *18*(3), e2019SW002216. <https://doi.org/10.1029/2019SW002216>
- Marshall, R. A., Dalzell, M., Waters, C. L., Goldthorpe, P., & Smith, E. A. (2012). Geomagnetically induced currents in the New Zealand power network. *Space Weather*, *10*(8), S08003. <https://doi.org/10.1029/2012SW000806>
- McGranaghan, R. M., Ziegler, J., Bloch, T., Hatch, S., Camporeale, E., Lynch, K., et al. (2020). Toward a next generation particle precipitation model: Mesoscale prediction through machine learning (a case study and framework for progress). *Space Weather*, *19*, e2020SW002684. <https://doi.org/10.1029/2020SW002684>
- McKinney, W. (2010). *Data structures for statistical computing in Python*. Retrieved from <http://conference.scipy.org/proceedings/scipy2010/mckinney.html>
- McPherron, R. L. (1970). *Growth Phase of Magnetospheric Substorms* (Vol. 75; Tech. Rep. No. 28). Retrieved from <http://citeserx.ist.psu.edu/viewdoc/download?doi=10.1.1.452.9720&rep=rep1&type=pdf>
- McPherron, R. L., Aubry, M. P., Russell, C. T., & Coleman, P. J. (1973). *Satellite studies of magnetospheric substorms on August 15, 1968 4. Ogo 5 magnetic field observations* (Vol. 78; Tech. Rep. No. 16). Retrieved from [https://www.researchgate.net/publication/254933902\\_Satellite\\_Studies\\_of\\_Magnetospheric\\_Substorms\\_on\\_August\\_15\\_1968](https://www.researchgate.net/publication/254933902_Satellite_Studies_of_Magnetospheric_Substorms_on_August_15_1968)
- Milan, S. E., Clausen, L. B. N., Coxon, J. C., Carter, J. A., Walach, M.-T., Laundal, K., & Anderson, B. J. (2017). Overview of solar wind magnetosphere atmosphere coupling and the generation of magnetospheric currents. *Space Science Reviews*, *206*(1–4), 547–573. <https://doi.org/10.1007/s11214-017-0333-0>
- Milan, S. E., Gosling, J. S., & Hubert, B. (2012). Relationship between interplanetary parameters and the magnetopause reconnection rate quantified from observations of the expanding polar cap. *Journal of Geophysical Research*, *117*(3), 3226. <https://doi.org/10.1029/2011JA017082>
- Murray, S. A., Bingham, S., Sharpe, M., & Jackson, D. R. (2017). Flare forecasting at the Met Office Space Weather Operations Centre. *Space Weather*, *15*(4), 577–588. <https://doi.org/10.1002/2016SW001579>
- Ngwira, C. M., Pulkkinen, A. A., Bernabeu, E., Eichner, J., Viljanen, A., & Crowley, G. (2015). Characteristics of extreme geoelectric fields and their possible causes: Localized peak enhancements. *Geophysical Research Letters*, *42*(17), 6916–6921. <https://doi.org/10.1002/2015GL065061>
- Ngwira, C. M., Sibeck, D., Silveira, M. V. D., Georgiou, M., Weygand, J. M., Nishimura, Y., & Hampton, D. (2018). A study of intense local dB/dt variations during two geomagnetic storms. *Space Weather*, *16*(6), 676–693. <https://doi.org/10.1029/2018SW001911>
- Nicolaou, G., Wicks, R. T., Rae, I. J., & Kataria, D. O. (2020). Evaluating the performance of a plasma analyzer for a space weather monitor mission concept. *Space Weather*, *18*(12), e2020SW002559. <https://doi.org/10.1029/2020SW002559>
- Oliveira, D. M., Arel, D., Raeder, J., Zesta, E., Ngwira, C. M., Carter, B. A., & Gjerloev, J. W. (2018). Geomagnetically induced currents caused by interplanetary shocks with different impact angles and speeds. *Space Weather*, *16*(6), 636–647. <https://doi.org/10.1029/2018SW001880>
- Owens, M., Riley, P., Lang, M., & Lockwood, M. (2019). Near-Earth solar wind forecasting using corotation from L5: The error introduced by heliographic latitude offset. *Space Weather*, *17*, 1105–1113. <https://doi.org/10.1029/2019SW002204>

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved from <http://jmlr.org/papers/v12/pedregosa11a.html>
- Perreault, P., & Akasofu, S.-I. (1978). A study of geomagnetic storms. *Geophysical Journal International*, *54*(3), 547–573. <https://doi.org/10.1111/j.1365-246X.1978.tb05494.x>
- Pulkkinen, A., Bernabeu, E., Eichner, J., Viljanen, A., & Ngwira, C. (2015). Regional-scale high-latitude extreme geoelectric fields pertaining to geomagnetically induced currents. *Earth, Planets and Space*, *67*(1), 93. <https://doi.org/10.1186/s40623-015-0255-6>
- Pulkkinen, A., Rastätter, L., Kuznetsova, M., Singer, H., Balch, C., Weimer, D., & Weigel, R. (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, A., Thomson, A., Clarke, E., & McKay, A. (2003). April 2000 geomagnetic storm: Ionospheric drivers of large geomagnetically induced currents. *Annales Geophysicae*, *21*(3), 709–717. <https://doi.org/10.5194/angeo-21-709-2003>
- Rae, I. J., Donovan, E. F., Mann, I. R., Fenrich, F. R., Watt, C. E. J., Milling, D. K., & Balogh, A. (2005). Evolution and characteristics of global Pc5 ULF waves during a high solar wind speed interval. *Journal of Geophysical Research*, *110*(A12), A12211. <https://doi.org/10.1029/2005JA011007>
- Rae, I. J., Watt, C. E. J., Murphy, K. R., Frey, H. U., Ozeke, L. G., Milling, D. K., & Mann, I. R. (2012). The correlation of ULF waves and auroral intensity before, during and after substorm expansion phase onset. *Journal of Geophysical Research*, *117*(A8), 168. <https://doi.org/10.1029/2012JA017534>
- Rajput, V. N., Boteler, D. H., Rana, N., Saiyed, M., Anjana, S., & Shah, M. (2020). Insight into impact of geomagnetically induced currents on power systems: Overview, challenges and mitigation. *Electric Power Systems Research*, 106927. <https://doi.org/10.1016/j.epsr.2020.106927>
- Ranjan, C. (2020). *Understanding deep learning application in rare event prediction*. Retrieved from [https://books.google.co.in/books/about/Understanding\\_Deep\\_Learning.html?id=80MlzgEACA AJ&redir\\_esc=y](https://books.google.co.in/books/about/Understanding_Deep_Learning.html?id=80MlzgEACA AJ&redir_esc=y)
- Reiff, P. H., Daou, A. G., Sazykin, S. Y., Nakamura, R., Hairston, M. R., Coffey, V., et al. (2016). Multispacecraft observations and modeling of the 22/23 June 2015 geomagnetic storm. *Geophysical Research Letters*, *43*(14), 7311–7318. <https://doi.org/10.1002/2016GL069154>
- Richardson, I. G. (2018). *Solar wind stream interaction regions throughout the heliosphere* (Vol. 15) (No. 1). Springer. <https://doi.org/10.1007/s41116-017-0011-z>
- Rodger, C. J., Mac Manus, D. H., Dalzell, M., Thomson, A. W. P., Clarke, E., Petersen, T., & Divett, T. (2017). Long-term geomagnetically induced current observations from New Zealand: Peak current estimates for extreme geomagnetic storms. *Space Weather*, *15*(11), 1447–1460. <https://doi.org/10.1002/2017SW001691>
- Rogers, N. C., Wild, J. A., Eastoe, E. F., Gjerloev, J. W., & Thomson, A. W. P. (2020). A global climatological model of extreme geomagnetic field fluctuations. *Journal of Space Weather and Space Climate*, *10*, 5. <https://doi.org/10.1051/swsc/2020008>
- Russell, C. T., Ginskey, M., Petrinec, S., & Le, G. (1992). The effect of solar wind dynamic pressure changes on low and mid-latitude magnetic records. *Geophysical Research Letters*, *19*(12), 1227–1230. <https://doi.org/10.1029/92GL01161>
- Shinbori, A., Tsuji, Y., Kikuchi, T., Araki, T., Ikeda, A., Uozumi, T., et al. (2012). Magnetic local time and latitude dependence of amplitude of the main impulse (MI) of geomagnetic sudden commencements and its seasonal variation. *Journal of Geophysical Research*, *117*(8), 8322. <https://doi.org/10.1029/2012JA018006>
- Shore, R. M., Freeman, M. P., & Gjerloev, J. W. (2018). An empirical orthogonal function reanalysis of the northern polar external and induced magnetic field during solar cycle 23. *Journal of Geophysical Research: Space Physics*, *123*(1), 781–795. <https://doi.org/10.1002/2017JA024420>
- Siciliano, F., Consolini, G., Tozzi, R., Gentili, M., Giannattasio, F., & De Michelis, P. (2020). Forecasting SYMH index: A comparison between long short term memory and convolutional neural networks. *Space Weather*, *19*, e2020SW002589. <https://doi.org/10.1029/2020SW002589>
- Smith, A. W., Forsyth, C., Rae, I. J., Rodger, C. J., & Freeman, M. P. (2021). The impact of sudden commencements on ground magnetic field variability: Immediate and delayed consequences. *Space Weather*, *19*, e2021SW002764. <https://doi.org/10.1029/2021SW002764>
- Smith, A. W., Freeman, M. P., Rae, I. J., & Forsyth, C. (2019). The influence of sudden commencements on the rate of change of the surface horizontal magnetic field in the United Kingdom. *Space Weather*, *17*, 1605–1617. <https://doi.org/10.1029/2019SW002281>
- Smith, A. W., Rae, I. J., Forsyth, C., Oliveira, D. M., Freeman, M. P., & Jackson, D. R. (2020). Probabilistic forecasts of storm sudden commencements from interplanetary shocks using machine learning. *Space Weather*, *18*(11), e2020SW002603. <https://doi.org/10.1029/2020SW002603>
- Smith, A. W., Rae, I. J., Forsyth, C., Watt, C. E. J., & Murphy, K. R. (2020). On the magnetospheric ULF wave counterpart of substorm onset. *Journal of Geophysical Research: Space Physics*, *125*(4). <https://doi.org/10.1029/2019JA027573>
- Smith, A. W., Rae, I. J., Forsyth, C., Watt, C. E. J., Murphy, K. R., & Mann, I. R. (2020). Diagnosing the time-dependent nature of magnetosphere-ionosphere coupling via ULF waves at substorm onset. *Journal of Geophysical Research: Space Physics*, *125*(11), e2020JA028573. <https://doi.org/10.1029/2020JA028573>
- Southwood, D. J., & Kivelson, M. G. (1990). The magnetohydrodynamic response of the magnetospheric cavity to changes in solar wind pressure. *Journal of Geophysical Research*, *95*(A3), 2301. <https://doi.org/10.1029/JA095iA03p02301>
- Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science of Science*, *240*(4857), 1285–1293. <https://doi.org/10.1126/science.3287615>
- Takeuchi, T., Araki, T., Viljanen, A., & Watermann, J. (2002). Geomagnetic negative sudden impulses: Interplanetary causes and polarization distribution. *Journal of Geophysical Research*, *107*(A7), 1096. <https://doi.org/10.1029/2001JA900152>
- 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>
- Tasistro-Hart, A., Grayver, A., & Kuvshinov, A. (2021). Probabilistic geomagnetic storm forecasting via deep learning. *Journal of Geophysical Research: Space Physics*, *126*(1). <https://doi.org/10.1029/2020JA028228>
- Thomas, S. R., Fazakerley, A., Wicks, R. T., & Green, L. (2018). Evaluating the skill of forecasts of the near-earth solar wind using a space weather monitor at L5. *Space Weather*, *16*(7), 814–828. <https://doi.org/10.1029/2018SW001821>
- 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>
- Thomson, A. W., McKay, A. J., Clarke, E., & Reay, S. J. (2005). Surface electric fields and geomagnetically induced currents in the Scottish Power grid during the 30 October 2003 geomagnetic storm. *Space Weather*, *3*(11). <https://doi.org/10.1029/2005sw000156>
- Tóth, G., Meng, X., Gombosi, T. I., & Rastätter, L. (2014). Predicting the time derivative of local magnetic perturbations. *Journal of Geophysical Research: Space Physics*, *119*(1), 310–321. <https://doi.org/10.1002/2013JA019456>

- Tsurutani, B. T., & Hajra, R. (2021). The Interplanetary and magnetospheric causes of geomagnetically induced currents (GICs) 10A in the Mäntsälä Finland Pipeline: 1999 through 2019. *Journal of Space Weather and Space Climate*, 11, 23. <https://doi.org/10.1051/swsc/2021001>
- Turnbull, K. L., Wild, J. A., Honary, F., Thomson, A. W. P., & McKay, A. J. (2009). Characteristics of variations in the ground magnetic field during substorms at mid latitudes. *Annales Geophysicae*, 27(9), 3421–3428. <https://doi.org/10.5194/angeo-27-3421-2009>
- Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2), 22–30. <https://doi.org/10.1109/MCSE.2011.37>
- Viljanen, A., Nevanlinna, H., Pajunpää, K., & Pulkkinen, A. (2001). Time derivative of the horizontal geomagnetic field as an activity indicator. *Annales Geophysicae*, 19(9), 1107–1118. <https://doi.org/10.5194/angeo-19-1107-2001>
- Viljanen, A., Pirjola, R., Prácer, E., Ahmadzai, S., & Singh, V. (2013). Geomagnetically induced currents in Europe: Characteristics based on a local power grid model. *Space Weather*, 11(10), 575–584. <https://doi.org/10.1002/swe.20098>
- Viljanen, A., Tanskanen, E. L., & Pulkkinen, A. (2006). Relation between substorm characteristics and rapid temporal variations of the ground magnetic field. *Annales Geophysicae*, 24(2), 725–733. <https://doi.org/10.5194/angeo-24-725-2006>
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., & Contributors, S. (2020). SciPy 1.0: Fundamental Algorithms For Scientific Computing in Python. *Nature Methods*, 17, 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- Voronkov, I., Donovan, E. F., & Samson, J. C. (2003). Observations of the phases of the substorm. *Journal of Geophysical Research*, 108(A2), 1073. <https://doi.org/10.1029/2002JA009314>
- Walach, M., & Grocott, A. (2019). SuperDARN observations during geomagnetic storms, geomagnetically active times, and enhanced solar wind driving. *Journal of Geophysical Research: Space Physics*, 124(7), 5828–5847. <https://doi.org/10.1029/2019JA026816>
- Webb, D. F., & Howard, T. A. (2012). Coronal mass ejections: Observations. *Living Reviews in Solar Physics*, 9(1), 3. <https://doi.org/10.12942/lrsp-2012-3>
- Welling, D. T., Ngwira, C. M., Opgenoorth, H., Haiducek, J. D., Savani, N. P., Morley, S. K., & Liemohn, M. (2018). Recommendations for next generation ground magnetic perturbation validation. *Space Weather*, 16(12), 1912–1920. <https://doi.org/10.1029/2018SW002064>
- 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: Space Physics*, 110(A4), A04203. <https://doi.org/10.1029/2004JA010500>
- Wintoft, P., Viljanen, A., & Wik, M. (2016). Extreme value analysis of the time derivative of the horizontal magnetic field and computed electric field. *Annales Geophysicae*, 34(4), 485–491. <https://doi.org/10.5194/angeo-34-485-2016>
- Wintoft, P., & Wik, M. (2018). Evaluation of Kp and Dst predictions using ACE and DSCOVR solar wind data. *Space Weather*, 16(12), 1972–1983. <https://doi.org/10.1029/2018SW001994>
- 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>
- Wintoft, P., Wik, M., & Viljanen, A. (2015). Solar wind driven empirical forecast models of the time derivative of the ground magnetic field. *Journal of Space Weather and Space Climate*, 5, A7. <https://doi.org/10.1051/swsc/2015008>
- Wu, J.-G., & Lundstedt, H. (1996). Prediction of geomagnetic storms from solar wind data using Elman Recurrent Neural Networks. *Geophysical Research Letters*, 23(4), 319–322. <https://doi.org/10.1029/96GL00259>
- Yousef, W. A., Wagner, R. F., & Loew, M. H. (2005). Estimating the uncertainty in the estimated mean area under the ROC curve of a classifier. *Pattern Recognition Letters*, 26(16), 2600–2610. <https://doi.org/10.1016/j.patrec.2005.06.006>
- Yue, C., Zong, Q. G., Zhang, H., Wang, Y. F., Yuan, C. J., Pu, Z. Y., et al. (2010). Geomagnetic activity triggered by interplanetary shocks. *Journal of Geophysical Research*, 115(A5), 105. <https://doi.org/10.1029/2010JA015356>
- Zhang, J. J., Wang, C., Sun, T. R., Liu, C. M., & Wang, K. R. (2015). GIC due to storm sudden commencement in low-latitude high-voltage power network in China: Observation and simulation. *Space Weather*, 13(10), 643–655. <https://doi.org/10.1002/2015SW001263>
- Zhang, X. Y., Zong, Q. G., Wang, Y. F., Zhang, H., Xie, L., Fu, S. Y., et al. (2010). ULF waves excited by negative/positive solar wind dynamic pressure impulses at geosynchronous orbit. *Journal of Geophysical Research*, 115(10). <https://doi.org/10.1029/2009JA015016>
- Zhelavskaya, I. S., Aseev, N. A., & Shprits, Y. Y. (2021). A combined neural network and physics based approach for modeling plasmasphere dynamics. *Journal of Geophysical Research: Space Physics*, 126, e2020JA028077. <https://doi.org/10.1029/2020JA028077>
- Zhelavskaya, I. S., Vasile, R., Shprits, Y. Y., Stolle, C., & Matzka, J. (2019). Systematic analysis of machine learning and feature selection techniques for prediction of the Kp index. *Space Weather*, 17(10), 1461–1486. <https://doi.org/10.1029/2019SW002271>
- Zweig, M. H., & Campbell, G. (1993). *Receiver-operating characteristic (ROC) Plots: A fundamental evaluation tool in clinical medicine* (Vol. 39; Tech. Rep. No. 4). Retrieved from [http://www.floppybunny.org/robin/web/virtualclassroom/dss/articles/roc%20curves/roc\\_as\\_evaluation\\_tool\\_zweig\\_campbell\\_1993.pdf](http://www.floppybunny.org/robin/web/virtualclassroom/dss/articles/roc%20curves/roc_as_evaluation_tool_zweig_campbell_1993.pdf)