Probabilistic Geomagnetic Storm Forecasting via Deep Learning

Adrian Tasistro-Hart[1,2], Alexander Grayver[1], Alexey Kuvshinov[1]

[1]Institute of Geophysics, ETH Zürich, Sonneggstrasse 5, 8092 Zürich, Switzerland

[2]Department of Earth Science, University of California, Santa Barbara, CA 93106, USA

Key Points:

- We present a neural network architecture that utilizes both observations from the L1 point and solar disk, improving forecast reliability
- Our neural network architecture learns reliable estimates of uncertainty in multiple hour ahead forecasts
- Instead of the conventional disturbance storm time (Dst) index we forecast the external component of geomagnetic storms, Est

Corresponding author: Adrian Tasistro-Hart, adrian_tasistro-hart@ucsb.edu

Abstract

Geomagnetic storms, which are governed by the plasma magnetohydrodynamics of the solar-interplanetary-magnetosphere system, entail a formidable challenge for physical forward modeling. Yet, the abundance of high quality observational data has been amenable for the application of data-hungry neural networks to geomagnetic storm forecasting. Almost all applications of neural networks to storm forecasting have utilized solar wind observations from the Earth-Sun first Lagrangian point (L1) or closer and generated deterministic output without uncertainty estimates. Furthermore, forecasting work has focused on indices that are also sensitive to induced internal magnetic fields, complicating the forecasting problem with another layer of non-linearity. We address these points, presenting neural networks trained on observations from both the solar disk and the L1 point. Our architecture generates reliable probabilistic forecasts over Est, the external component of the disturbance storm time index, showing that neural networks can gauge confidence in their output.

Plain Language Summary

Geomagnetic storms are capable of damaging infrastructure like power grids and communication lines, motivating our need to forecast them. Solar phenomena produce geomagnetic storms, which occur when these phenomena reach Earth as bursts of the solar wind. Decades of satellite observations of both the solar wind near the Earth and of the Sun itself are promising for forecasting geomagnetic storms with algorithms known as neural networks. Several neural network architectures have been applied to geomagnetic storm forecasting, but their full potential remains unexplored. First, all existing neural networks have used measurements of the solar wind one hour upstream of the Earth or closer. While these observations are critical for understanding geomagnetic storm progression, from them it is nearly impossible to forecast more than an hour in advance. We include observations of the Sun itself, which reach Earth much faster than the solar wind, thereby including information for forecasting further in advance. Second, all existing neural networks have generated forecasts without uncertainty estimates, meaning that end-users (such as utilities or telecommunications companies) know little about forecast confidence. We present an architecture that generates estimates of uncertainty, and our results demonstrate that neural networks learn how confident to be in their forecasts.

1 Introduction

Mankind has experienced a number of blackouts caused by geomagnetically induced currents (GICs), which can result in millions of dollars of damages and leave millions without electricity (Bolduc, 2002; Love et al., 2018). The possibility of such disruptions has motivated the goal of forecasting GICs. All GICs in turn result from geomagnetic storms, which generate the variation in Earth's external field that induces GICs. The problem of forecasting GICs then amounts to forecasting geomagnetic storms. These storms result from the propagation of solar activity via the solar wind and its coupling to Earth's magnetosphere. Given abundant observational data of the solar wind and disk as well as of Earth's magnetic field, the application of data-hungry deep learning algorithms is suitable for the forecasting problem.

1.1 Geomagnetic storms

Geomagnetic storms have traditionally been quantified by indices such as the disturbance storm time (Dst in nT) or Kp (unitless) indices (e.g. Bartels et al. (1939)), both of which register deviations from the quiet time horizontal component of Earth's magnetic field. The basic mechanism of geomagnetic storm formation is the strengthening of Earth's ring current in response to changing solar wind conditions, and this strengthened current system generates a magnetic field that counters Earth's dipole, weakening it relative to quiet conditions (Daglis et al., 1999). The solar wind parameters most important for strengthening the ring current are its southward component of the inter-planetary magnetic field (IMF), velocity, and plasma density, which all positively impact storm amplitude (Daglis et al., 1999; Gonzalez et al., 1999; Wolf et al., 1997). All solar wind activity that generates significant, rapid fluctuation in Earth's external magnetic field poses a threat to ground-based conducting systems, such as power and communication lines, during geomagnetic storms.

1.2 Why deep learning?

Given the complexity of the underlying physics, which involves the magnetohydrodynamics (MHD) and plasma physics of propagating solar activity through the solar wind and its subsequent interaction with Earth's magnetosphere, a fully physical forward model of the system would be both computationally expensive and poorly constrained. At the

same time, given that we are aware of the important physical quantities responsible for geomagnetic storms, such a physical model is overkill for the problem of forecasting the low-order response of Earth's magnetic field to solar activity.

For this reason, the first approach to geomagnetic storm modeling took the form of simple empirical models that related the time rate of change of Dst to solar wind parameters. The pioneering work was a three-term deterministic model developed by Burton et al. (1975), but its simplicity, while elegant, often generates inaccurate forecasts. Subsequent modeling has attempted to improve accuracy by adding more degrees of freedom. For example, while obtaining more predictive power, Temerin & Li (2006) added almost a dozen more terms with significantly more complex functional forms, sacrificing the simplicity of the initial model. Neural networks (NNs), which form the backbone of deep learning, are the logical conclusion to the exercise of adding more and more heuristic functional forms, since the task of a NN is to learn the relevant functions rather than have them prescribed: even a single layer neural network with sufficient "neurons" is capable of approximating any continuous function to arbitrary precision (Leshno et al., 1993; Pinkus, 1999). However, given that this sufficient number of neurons in a single layer network is typically unknown and potentially intractable, workers have found success by instead adding layers of neurons rather than neurons themselves. This composition of layers, in which neurons in a given layer operate on the outputs of neurons from the preceding one, is coined "deep learning", and has met with unprecedented success in classification and regression problems. While still poorly understood beyond a heuristic sense, some workers hypothesize that deep neural networks are successful because many learning problems are outcomes of hierarchical and compositional processes, which deep networks can efficiently reproduce (Brahma et al., 2016; Lin et al., 2017). Furthermore, Lin et al. (2017) demonstrate how the properties of symmetry, locality, and polynomial log-probability in many natural processes are efficiently learned by even relatively shallow (i.e., consisting of a handful of hidden layers) neural networks.

1.3 Prior applications of deep learning to geomagnetic storm forecasting

Previous work with NNs has focused almost entirely on prediction of Dst or other indices of geomagnetic activity, such as the Kp and the auroral electrojet (AE) indices. Supplemental Table S1 provides a succinct review of the application of NNs to the fore-

casting of Dst (Andriyas & Andriyas, 2015; Bala & Reiff, 2012; Gleisner et al., 1996; Jankovičová et al., 2002; Kugblenu et al., 1999; Lazzús et al., 2017; Munsami, 2000; Pallocchia et al., 2006; Revallo et al., 2014; Sharifie et al., 2006; Stepanova et al., 2005; Stepanova & Pérez, 2000; Wei et al., 2007; Wu & Lundstedt, 1996, 1997). These studies have applied a variety of architectures and data sources, but in generating forecasts for Dst, most have used the basic solar wind parameter measurements as well as prior values of Dst. All previous studies applying NNs to Dst forecasting to our knowledge have utilized observations made at the Earth-Sun L1 point or closer, with the exception of Chakraborty & Morley (2020), who include solar x-ray fluxes. Furthermore, almost all studies to date using NNs to forecast Dst (or any other geomagnetic storm index) have been deterministic, generating predictions without any measure of uncertainty. Gruet et al. (2018) assess uncertainty in their forecasts via a Gaussian process model with fixed kernel parameters, and this process takes as input their deterministic NN forecasts. Chakraborty & Morley (2020) on the other hand use a deep long short term memory (LSTM) network to learn how to dynamically update the kernel parameters for a Gaussian process representation of the Kp index, which is how they generate probabilistic forecasts. Finally, while not utilizing neural networks, Gu et al. (2019) generate probabilistic forecasts of the auroral electrojet (AE) index by considering output from an ensemble of 100 nonlinear autoregressive models trained on independently resampled subsets of their data.

This work improves on previous advances by presenting the first application of probabilistic neural networks that explicitly generate measures of uncertainty in their output. Our networks are capable of learning how confident to be in their predictions, and in doing so improve forecast reliability. These networks take as input not just observations from the L1 point but also observations of radiative phenomena on the solar disk. Finally, instead of forecasting Dst, we focus on its external component, Est, which does not incorporate the effects of Earth's subsurface conductivity structure.

2 Data and Methods

2.1 Probabilistic Neural Network Architecture

Recently, recurrent architectures for time series regression have emerged that combine ridge functions with state vectors to create units with "memory". The most successful of these has been the long short-term memory architecture (LSTM), introduced

by Hochreiter & Schmidhuber (1997). The LSTM cell, as its name implies, uses new input data with both the previous output and previous internal state to update its internal state and generate new output (Supplement, Text S6). This architecture has been applied to Dst forecasting by Gruet et al. (2018), but, like all previous applications of neural networks to storm forecasting (summarized in the Supplement, Table S1), the network generated deterministic output with no prediction of forecast uncertainty.

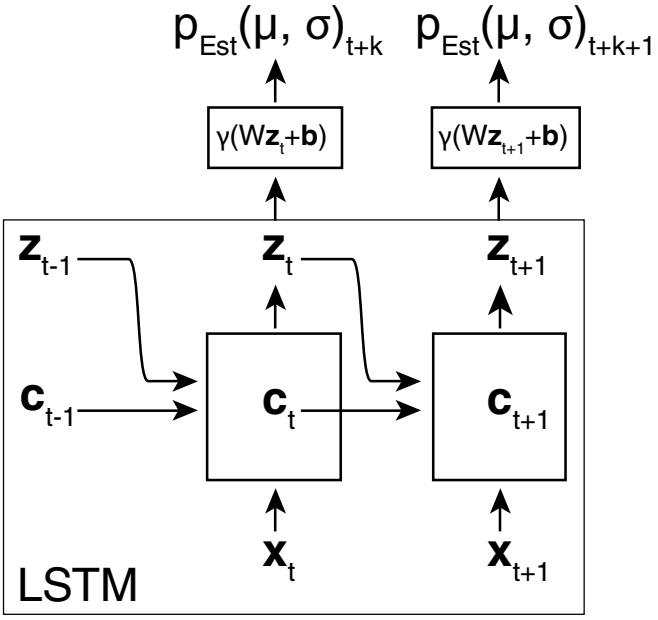

Figure 1. Schematic architecture for the deterministic network that learns parameters over an output distribution for Est. Our output distribution is a Gaussian. Two full time steps of network iteration are shown, with the portion of the network enclosing the LSTM cell labeled "LSTM".

We present an architecture (Figure 1) in which the NN learns to assess uncertainty in its own forecast, thereby generating probabilistic forecasts. The two basic layers utilized within this architecture are LSTM and dense layers. The former is described above, and the latter is an implementation of the so-called "fully connected hidden layer", which references the fact that each entry in $\mathbf{z}$ depends on all of the outputs from the preceding layer via W. That is, a dense layer that receives inputs $\mathbf{x} \in \mathbb{R}^n$ from a preceding network layer in turn generates an output vector $\mathbf{z} \in \mathbb{R}^m$ via the operation $\mathbf{z} = \gamma(W\,\mathbf{x} + \mathbf{b})$ with $W \in \mathbb{R}^{mxn}$ and $\mathbf{b} \in \mathbb{R}^m$, where n is the dimensionality of the preceding hid-

den layer, m is the dimensionality of the current hidden layer, and $\gamma(.)$ is a nonlinear activation function that acts element-wise.

Inputs into our NN architecture are fed directly to an LSTM cell, and outputs from the LSTM cell are fed through a series of fully-connected hidden layers. The outputs from the last hidden layer are parameters for an output distribution over Est. We choose to use a Gaussian output distribution and compare other alternatives in the Supplement (Text S3).

The simplest cost function in this probabilistic framework is precisely the output distribution itself evaluated as a likelihood of observed data y (i.e., Est at some time $t+k$) with respect to the distribution parameters generated from the given input:

$$C(\mathbf{x}, y) = -\log p\left(y|\mu(W, \mathbf{b}, \mathbf{x}), \sigma(W, \mathbf{b}, \mathbf{x})\right) \tag{1}$$

where the distribution parameters μ and σ depend on the network weights and biases and can thus be differentiated against them. However, when learning two-parameter distributions, the parameter for scale often introduces leniency in the output distribution, allowing the network to expand uncertainty in its forecast rather than move its estimate for the center (see supplement Text S3).

We found that utilizing a Gaussian output distribution with a regularized Gaussian likelihood as the cost function performed well for geomagnetic storm forecasting. Equation 2 shows the form for this modified log-likelihood in which $\alpha\,(y-\mu)^2 + \beta\,\frac{1}{\sigma^2}$ are the terms that we have added, introducing α and β as additional hyper-parameters. This regularization encourages the network to learn more reasonable estimates for μ, offsetting the normalization by σ^2, while also allowing the user to further incentivize ($\beta > 0$) or penalize ($\beta < 0$) expanding forecast uncertainty.

$$C_{\text{Gaussian, regularized}}(y, \mu, \sigma) = \log\left(\sqrt{2\,\pi}\,\sigma\right) + \frac{(y-\mu)^2}{2\,\sigma^2} + \alpha\,(y-\mu)^2 + \beta\,\frac{1}{\sigma^2} \tag{2}$$

Other approaches have been formulated for learning uncertainty via neural networks, such as Bayes-by-Backprop (Blundell et al., 2015), which represents uncertainty in the network weights rather than in its output. Our implementation of this approach was not useful for storm forecasting (Supplement Text S2).

For all implementations, training, and testing of neural networks, we use Python wrappers for the learning framework TensorFlow (Abadi et al., 2015). This framework is capable of representing neurons and the functional operations relating them as well

as numerically computing the relevant gradients to train the network. TensorFlow provides an implementation of the high level deep learning Keras API (`https://keras.io/`), which allows for modular construction of networks from the layers described above. We also make use of the recently introduced TensorFlow Probability library, which provides a straightforward means of adding probability distributions as layers, allowing outputs from previous layers to be used as parameters for the distribution layers. These layers are compiled into a model that contains all the operations of the entire network as well as the particular cost functions and optimizers that dictate the learning process for given training inputs and outputs. We use the Adam optimizer for gradient update steps (Kingma & Ba, 2014). All neural network configuration and training parameters are listed in Supplement Text S4.

2.2 Output Data

Most geomagnetic storm forecasting thus far has emphasized prediction of Dst. However, Dst is actually a sum of internal and external components (Equation 3), and the internal component, Ist, is generated by currents induced in Earth's subsurface by variation in the external component, Est (Maus & Weidelt, 2004) ,

$$Dst(t) = Ist(t) + Est(t) \tag{3}$$

and

$$Ist(t) = \int_{-\infty}^{t} Q(t-\tau)\, Est(\tau) d\tau, \tag{4}$$

where $Q(t-\tau)$ is the impulse response that depends on subsurface electrical conductivity σ, assuming that $\sigma \equiv \sigma(r)$ varies only along radial direction. This decomposition is easier to express in frequency domain, in which $Q(t)$ becomes the transfer function that relates internal and external components such that $\tilde{Ist}(\omega) = \tilde{Q}(\omega)\tilde{Est}(\omega)$. Subsequently, Est and Ist can be computed as (Maus & Weidelt, 2004):

$$\tilde{Est}(\omega) = \frac{1}{1+\tilde{Q}(\omega)}\, \tilde{Dst}(\omega) \tag{5}$$

$$\tilde{Ist}(\omega) = \frac{\tilde{Q}(\omega)}{1+\tilde{Q}(\omega)}\, \tilde{Dst}(\omega) \tag{6}$$

with knowledge of $\tilde{Q}(\omega)$ and observations of Dst. For more details about this decomposition and the induction transfer functions, including generalization to three-dimensional conductivity models, the reader is referred to Grayver et al. (2020); Maus & Weidelt (2004); Olsen et al. (2005).

The problem of forecasting Dst is then actually two separate problems: the first is forecasting Est, and the second is learning Earth's induction response, Ist. With a suitable model of Earth's subsurface conductivity structure, however, knowledge of Est is sufficient to reconstruct Ist and thereby Dst. Furthermore, because the external field is what responds to magnetospheric activity anyway, it is more natural to forecast Est than Dst. Therefore, we generate forecasts of Est, and the data accessed from NOAA were generated following the methodology of Maus & Weidelt (2004) .

This approach will be increasingly important as we attempt to forecast higher order structure in Earth's external field. Est captures only the first zonal (dipole) component of external magnetic field variability, but significant variation exists on shorter spatial scales, where the interaction with local conductivity structures becomes more important and complicated (Kelbert, 2020). Given that the ultimate goal of geomagnetic storm forecasting is to forecast GICs, it is important to note that GICs themselves depend strongly on local conductivity structures and local external magnetic field variability (Olsen & Kuvshinov, 2004; Püthe et al., 2014). The first step to forecasting these higher order external field coefficients is forecasting a single external field coefficient, Est, which is the focus of this work.

2.3 Input Data

Two basic observation types relevant to geomagnetic storm forecasting have been made for the past few decades: the first includes measurements of the solar wind made in situ around the L1 point in the Earth-Sun system, and the second are measurements made directly of the solar disk and corona. These two data streams provide related but temporally disjoint information. Radiative phenomena on the solar disk take under nine minutes to be observed at Earth, while the solar wind requires two to five days to propagate from the solar disk to the L1 point. The L1 point is only 1.5×10^6 km from Earth, however, which is approximately one hour travel time at typical solar wind speeds (mean solar wind speed is roughly 440 km s^{-1} from the OMNI dataset).

Thus, while solar wind measurements near Earth's magnetosphere are ultimately the most relevant quantities for accurate geomagnetic storm forecasting, using only observations from the L1 point limits the forecast time to roughly an hour (Shprits et al., 2019). On the other hand, solar activity is ultimately responsible for all geomagnetic storms,

but identifying which phenomena on the solar disk have the potential to cause geomagnetic storms and predicting the storm lag times and amplitudes resulting from those phenomena are not trivial tasks. Observations from the solar disk include measurements of coronal mass ejections (CMEs) around the perimeter of the disk, images of the solar surface at various wavelengths, and surface radiative fluxes.

Input data come from three sources: the OMNI, GOES, and SOHO LASCO CME datasets. All details on data preprocessing are briefly discussed in the supplement (Text S2).

The low (hourly) resolution OMNI data include several solar wind and solar observations, which we extracted for the years 1995-2018. During this time interval, measurements of the solar wind (SW) come from the WIND, IMP8, Geotail, and ACE satellite missions . The quantities that we use as input data from this dataset are the three components of the interplanetary magnetic field in geocentric solar magnetospheric (GSM) coordinates, SW velocity, SW particle density, SW temperature, and SW longitude and latitude incident on Earth's magnetosphere. Because the OMNI dataset contains observations from near-Earth (e.g., IMP8) and L1 (e.g., WIND, ACE) spacecraft, observations are propagated to the bow shock by adding time shifts that account for the spacecraft location and solar wind flow. These time shifts are included in the publicly available dataset, and we utilize the observation time stamps as they are reported.

The GOES mission provided two time series of x-ray fluxes integrated over the solar disk. One series covered the wavelengths from 0.5-4 Å, and the other series covered wavelengths 1-8 Å. Given that these measurements vary over orders of magnitude, we take their logarithm as input. These data were reliably available from 1986 onwards.

The LASCO SOHO CME database provides a catalogue of CMEs observed around the perimeter of the solar disk, with five basic quantities estimated for each event: central position angle, angular width, speed, mass, and kinetic energy. Three estimates of speed are reported in the catalogue, all of which we include as training input. We only consider CMEs for which all data fields are reported, and during hours with multiple events, we take only the event with the largest estimated kinetic energy. The estimates of mass and kinetic energy varied over orders of magnitude, so we instead take their logarithms as input. The database contains measurements from 1996-2018.

We did not attempt to time shift observations of the solar disk to the bow shock as is done for satellite observations at the L1 point in the OMNI dataset. This time lag between the solar disk and Earth's magnetosphere is precisely a learning problem of great interest that NN's may be able to help solve.

Finally, previous observations of Est were included as input while forecasting future values. In total then data is available roughly from 1996 through 2018. Of these 22 years, we take 18 years (approximately 158,000 hours) as training data, and 4 years for testing data (approximately 35,000 hours).

2.4 Evaluating network performance

In most geomagnetic storm forecasting to date, forecast accuracy is assessed by metrics such as the root mean square error and the Pearson correlation coefficient between forecasted and observed data. However, these statistics are generally not compared to those of a null hypothesis, for example persistence forecasting in which the best estimate for any time in the future is simply the last observed value. Due to the auto-correlative nature of the Est time series, persistence forecasting actually generates deceptively high correlations and low errors (Figure 3, Supplement Figure S9) (Shprits et al., 2019). In fact, all the networks in Supplement Table S1 either underperform or barely outperform persistence forecasting as quantified by these two metrics. Furthermore, these metrics are computed over both quiet and disturbed times, while the ability to predict storms is the task of interest. Other than refining consideration of these metrics to only storm main phases, it is not obvious what metric best evaluates forecasting performance for models with deterministic output.

With probabilistic networks, however, reliability curves provide a useful and easily interpretable metric to evaluate forecast performance. Each curve corresponds to a threshold in Est, and the axes compare the observed probability of exceeding that threshold compared to the predicted probability. A perfectly reliable forecast would generate curves that fall on the 1:1 line through the origin for all thresholds. If the observed reliability curve plots over the 1:1 line, then the forecast underestimates the occurrence of events exceeding that threshold, while if the reliability curve plots under the 1:1 line, the forecast is conservative and overestimates the occurrence of events exceeding the given threshold. Because these statistics are computed for thresholds in Est, the reliability as-

sessment method by construction evaluates storm time forecasting separately from quiet time forecasting. We utilize reliability curves to assess forecast performance.

Computing reliability curves requires binning data by intervals of predicted exceedance probabilities, which means that empirical statistics for infrequent, large storms will be less well constrained than smaller storms, particularly at large forecast probabilities. To assess uncertainty in the reliability curve computation, we use bootstrap resampling of forecasted and observed threshold exceedances to compute confidence intervals over observed threshold exceedances for a given bin of predicted threshold exceedance. Furthermore, we compute consistency intervals that indicate for the amount of data in each bin the spread in forecasting skill that one might anticipate from a perfectly reliable forecast (Bröcker & Smith, 2007).

3 Results and Discussion

In general, our architecture is capable of learning meaningful and reliable measures of uncertainty in its forecasts, and our forecasts outperform the basic benchmark of persistence forecasting. We focus in our discussion on the performance of the six-hour ahead probabilistic forecasts because this forecast window is long enough that information from the L1 point is insufficient to forecast storm arrivals, allowing us to evaluate whether our network is able to leverage information from the solar disk.

3.1 Storm Case Studies

We present and discuss in this section six hour ahead forecasts for three different storms: two caused by CMEs (Figure 2A & B) and one resulting from a co-rotating interaction region (CIR) (Figure 2C). One to six hour ahead forecasts, as well as detailed comparison of the networks trained with only L1 data and both L1 and solar disk inputs, can be found in the Supplement.

The first storm, in March of 2015, provides a prototypical example of a geomagnetic storm caused by a magnetic cloud emitted by the mass ejection visible as the spike in CME energy at the beginning of March 15 (Figure 2A, note that the axis is orders of magnitude) (Patel et al., 2019). The energetic mass ejection is associated with a peak in integrated x-ray fluxes, followed two days later by a relatively large geomagnetic storm beginning on March 17. The storm main phase is associated with a sustained southward

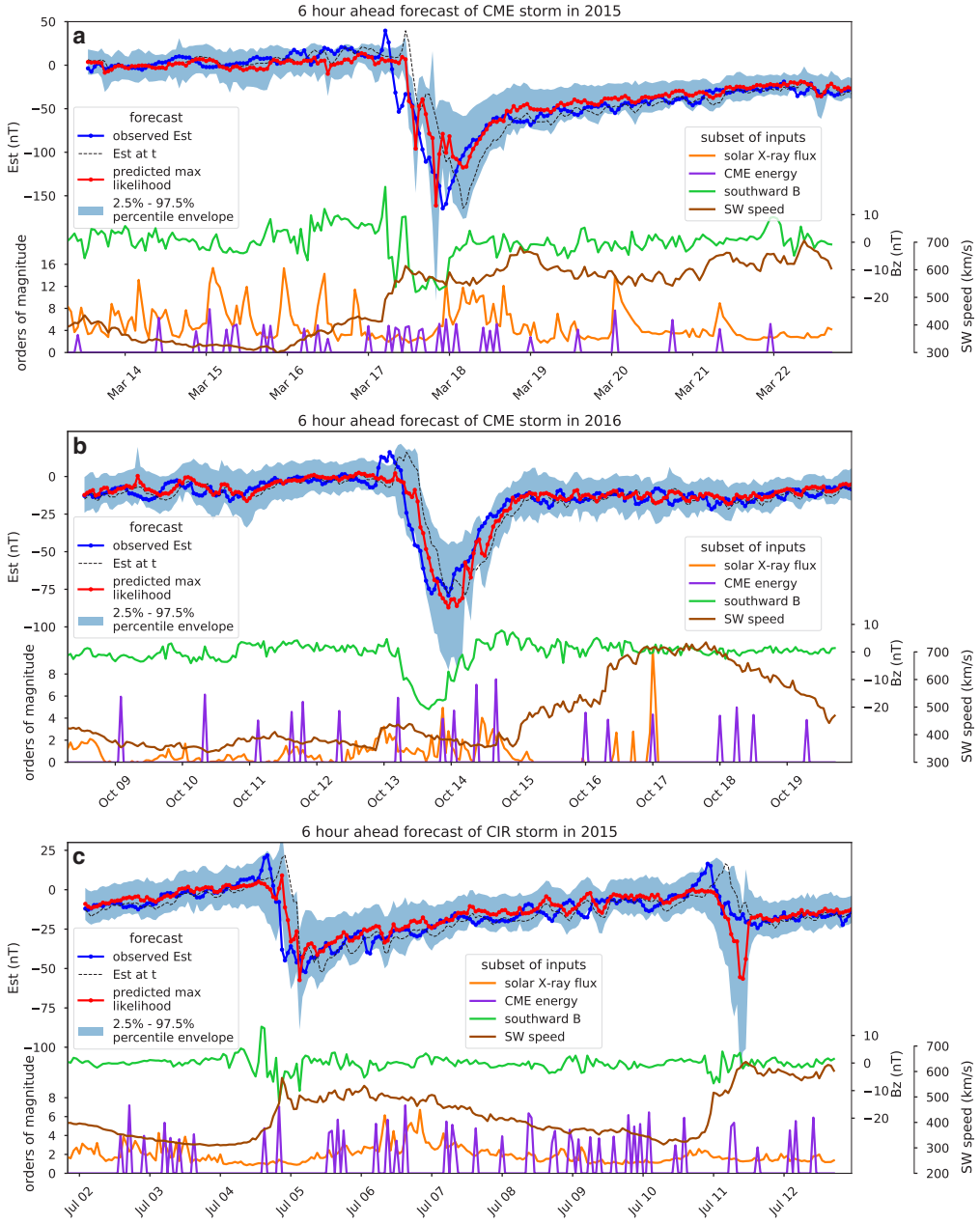

Figure 2. Six-hour-ahead probabilistic forecasts on testing data for two CME storms and one CIR storm, as identified by Patel et al. (2019); Shen et al. (2017). Order of magnitude variability in x-ray flux (long channel is plotted) is exaggerated by ten times. The black dashed line for Est is a persistence forecast. **(a)** Severe CME geomagnetic storm (min Dst=-222 nT). Large amplitude variability in the southward component of the IMF might be responsible for forecast variability as the storm enters the main phase. **(b)** Intense CME geomagnetic storm (min Dst=-104 nT). **(c)** Intense CIR geomagnetic storm (min SYM-H=-87 nT). Order of magnitude variability in x-ray flux here is only exaggerated by five times.

IMF of roughly -20 nT and roughly doubled solar wind speeds. In this situation, given the clear connection between the mass ejection and the ensuing storm, we would expect a successfully trained network to be able to expand uncertainty in its forecast as conceivable storm arrival times approach, reflecting an understanding of the causal association between activity on the solar disk and geomagnetic storms. Yet, forecast uncertainties only expand when disturbed solar wind reaches the L1 point. At that time, the network becomes aware of the storm arrival and adjusts its output by dropping Est forecasts and increasing forecast uncertainty (Figure 2A). The same is true for the smaller amplitude CME storm of October 13, 2016 in Figure 2B, where forecast uncertainty only grows as soon as the storm arrives at the L1 point. This storm is associated with the CME visible on October 10 (Patel et al., 2019), so the occurrence of other CMEs of similar magnitude (e.g. on October 9 and 11) demonstrates non-uniqueness that illustrates why the network struggles to identify geoeffective solar activity from the provided inputs.

The final storm on July 4, 2015 was chosen because it corresponds to a CIR (Shen et al., 2017), as evidenced by a lack of sustained, southward IMF, a step increase in solar wind velocity, and relatively low amplitude storm-time Est (Figure 2C). The nature of CIR storms differs from those originating from CMEs (Zhang et al., 2007), so we sought to investigate if the forecast for CIR storms differs from that for CME storms. Again, for the storm on July 4, the network is unable to preemptively expand forecast uncertainties in response to information from solar disk, demonstrating that inputs from the L1 point dominate the forecast. On July 11, the network mistakenly forecasts a storm main phase, likely in response to the increased solar wind speed that did not actually generate a substantial main phase.

In all cases, the six hour ahead forecast fails to capture storm onset, during which the network's forecasts tend to lag observations by the forecast length (thereby more closely tracking the persistence curve) until the storm arrives at the L1 point. At that point, the forecast begins to deviate from persistence as the network knows that a storm is occurring. This inability to predict storm onset indicates that the network is unable to utilize observations from the solar disk for storm arrival, which remains an open challenge.

Recovery is generally well-predicted, and forecasts deviate from persistence, meaning that the network is not just taking the last observed Est value for its next forecast. Unlike previous results, our network is capable of generating meaningful estimates of un-

certainty in its forecasts. In all cases, once the network detects the possible onset of a geomagnetic storm, it expands its forecast uncertainty, generally maintaining observed Est values within the 95% forecast confidence interval and providing reliable multiple hour ahead forecasts (see Supplement Text S5 for one to six hour ahead forecasts). After storm main phases, forecast uncertainties decrease during the generally well-predicted recovery phase. Given that the recovery phase is dictated by the internal dynamics of the ring current decay (and thereby independent of the solar wind state) (Daglis, 2007), its predictability is reasonable. Thus, our network exhibits forecast uncertainties that are consistent with where one would anticipate the greatest uncertainty in geomagnetic storm development with information from the L1 point, namely, the storm onset and main phase.

3.2 Conventional Metrics of Forecast Skill

In terms of the conventional forecasting skill assessments (i.e., forecast-observation Pearson correlation coefficients and root-mean-square errors, RMSE's) for one to six hour ahead forecasts, our networks outperform all previous neural network forecasts for all forecasts lengths (Figure 3). However, given that persistence forecasting for Est results in higher correlation coefficients and lower RMSE's than for persistence forecasting of Dst, our improvements should not be compared with previous results for forecasting Dst but with persistence forecasting of Est. When considering all observations, we slightly underperform persistence forecasting of Est in terms of the correlation coefficient, while outperforming in terms of RMSE, which is consistently lower. During storm times, however, our forecasting skill is much better than persistence forecasting in both metrics at all forecasting windows. The networks with both L1 and solar disk inputs always outperform networks with only L1 inputs when evaluated over both quiet and storm times (Figure S9). However, the difference in skill is small, and when considering only storm times, the five to six hour ahead forecasts of networks with only L1 inputs achieve lower RMSE's. These results again indicate that information from the solar disk does not significantly improve storm forecast skill. Finally, for particularly large storms exceeding Est $\leq$ -200 nT, a forecast saturation effect is observed (Figure S9), similar to that that occurs at smaller values with different cost functions (Supplement Text S3). This effect can be partially mitigated by fine-tuning of the cost function to further facilitate the fore-

casting of rare storms. However, since there is only a handful of such events in the data, this behaviour of the network is natural.

3.3 Forecast Reliability

Reliabilities for four different storm thresholds generally overlap with the consistency intervals for each bin, demonstrating that our network generates reliable forecasts (Figure 4). For threshold of -75 and -100 nT, forecasted exceedance probabilities in the range of 0.7-0.9 tended to slightly underestimate observed exceedance rates, which is consistent with the observation that storm onsets remain difficult to predict exactly.

Notably, the regularization of the cost function for a Gaussian output distribution significantly improves forecast reliability. Networks trained with unregularized Gaussian and Gumbel output distributions (Supplement Text S3) are unable move the location parameters of their forecasts during large amplitude storms, preferring instead to expand forecast uncertainty, meaning that peak storm times, while often within the 95% confidence interval, are only predicted at extremely low exceedance probabilities. This behavior explains why the reliability curves lack data to bin at high forecast probabilities and furthermore why storms are underestimated for lower exceedance probabilities (Figure 4).

While some improvement in forecast reliability for smaller magnitude storms (Est thresholds of -50 and -75 nT) does seem to result from the incorporation of data from the solar disk (Supplement Figure S7), the preceding discussion and the result that forecast behavior does not qualitatively change by adding solar disk inputs (Supplement Text S5) indicates that we are unable to successfully utilize observations from the solar disk to forecast storm arrival and amplitude. This shortcoming suggests that the information necessary for identifying geoeffective solar activity is lacking in the training data, and/or that the network architecture is inadequate for utilizing these data. For instance, the x-ray fluxes are integrated over the entire solar disk, but peaks in these fluxes can often be associated with flare events, which themselves often occur simultaneously with geoeffective mass ejections (Tobiska et al., 2013). Larger, more central flares are associated with larger geomagnetic storms, so adding time series of flare occurrences with locations on the solar disk would complement the input series of x-ray fluxes and CMEs (Tobiska et al., 2013). Futhermore, the CME dataset only includes ejections visible around

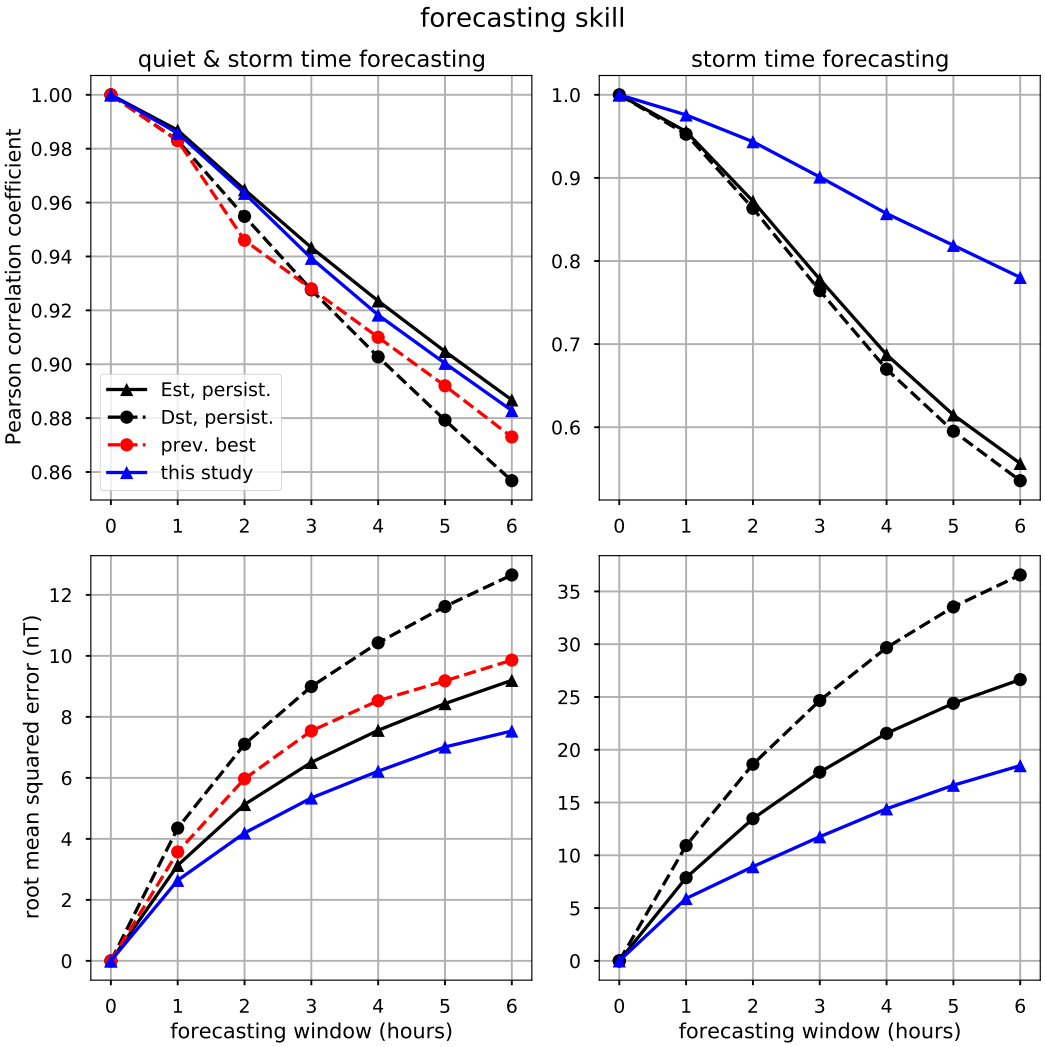

Figure 3. Conventional metrics of forecasting skill for networks with both solar disk and L1 inputs (see Figure S9 for networks with only L1 inputs). The first row shows the Pearson correlation coefficient between the forecasted Est (mean, for our study) or Dst and the observed Est or Dst. The second row shows the corresponding root mean squared errors. In the columns, we show these quantities for all observations (first column) and only storm-time observations (second column). The black lines show the metrics for persistence forecasts of Est and Dst, and the red line (available only for all observations) shows the best reported performance from NN forecasts (see Table S1 for references).

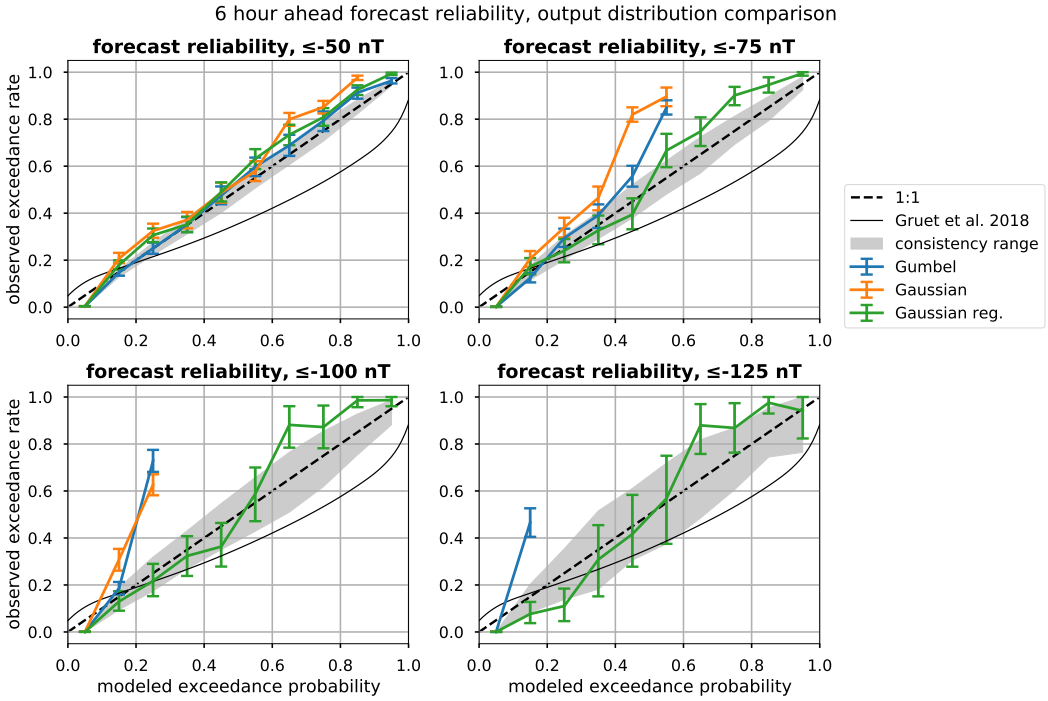

Figure 4. Reliability curves for the networks with a Gumbel output and cost (blue), Gaussian output and cost (orange), and Gaussian output with regularized cost (green). All curves are for a six hour ahead forecast for four Est thresholds in eleven bins. Exceedance was taken in the negative sense, i.e., taking values less than or equal to the given threshold. Error bars show the 2.5-97.5 percentile range from bootstrapped resampling (number of bootstrapped samples was 1000) within the bins of forecasted exceedance probability. The envelopes show the 2.5-97.5 percentile range from consistency resampling of a perfectly reliable forecast, demonstrating the conceivable range in reliable forecasts given the number of data in each bin (Bröcker & Smith, 2007). The regularized Gaussian network is the most reliable of the three. Also shown is the reliability curve from Gruet et al. (2018) for their six hour ahead Dst forecast, for which the exceedance threshold is unspecified.

the rim of the solar disk, while geoeffective ejections occur towards the center. Thus, only centralized ejections that also emit an observable lobe beyond the rim of the solar disk could be reliably associated with geomagnetic storms, potentially rendering the CME database largely irrelevant for the problem of geomagnetic storm forecasting. Finally, integrated solar x-ray flux peaks from flares have been empirically related to solar wind speeds and geomagnetic storm amplitudes, thereby providing a means of learning lag times between solar activity and storm arrivals (Tobiska et al., 2013). However, LSTM networks struggle with learning lag times (Gers et al., 2002), so the network architecture we have utilized is not amenable to this task.

4 Conclusions

This work has demonstrated a NN architecture capable of learning reliable measures of uncertainty in its forecasts of geomagnetic storms. Learning uncertainty in NN output results in more useful probabilistic forecasts than learning uncertainty in the NN parameters, and the choice of output distribution and cost function has a large impact on the resulting reliability of the trained network. Specifically, adding regularizing terms in the likelihood cost function improves the forecast reliability by incentivizing networks to forecast more reasonable mean values rather than simply increasing forecast uncertainty.

These neural networks utilize as inputs observations from both the solar disk and L1 point, slightly improving forecast reliability and skill with respect to networks trained only with L1 inputs. However, storm arrival and amplitude forecasting did not substantially improve from the inclusion of these data. Thus, leveraging time series of observations of the solar disk, which are often sparse, remains an open problem, and future network architectures must be carefully designed to utilize these data sources.

Acknowledgments

This work was partially supported by the ESA through the Swarm DISC project. The low (hourly) resolution OMNI data were obtained from the GSFC/SPDF OMNIWeb interface (`https://omniweb.gsfc.nasa.gov`). We acknowledge use of the LASCO SOHO CME catalog (`https://cdaw.gsfc.nasa.gov/CME_list/`), which is generated and maintained at the CDAW Data Center by NASA and The Catholic University of America in cooperation with the Naval Research Laboratory. SOHO is a project of international co-

operation between ESA and NASA. We acknowledge use of GOES mission x-ray flux data, accessed from `https://satdat.ngdc.noaa.gov/sem/goes/data/avg/`. Finally, we acknowledge use of the EST-IST-DST dataset from NOAA accessed from `https://www.ngdc.noaa.gov/geomag/est_ist.shtml`. All data and analysis presented in this study are available as Jupyter notebooks at `https://doi.org/10.5281/zenodo.3751682`.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* Retrieved from `https://www.tensorflow.org/` (Software available from tensorflow.org)

Andriyas, T., & Andriyas, S. (2015). Relevance vector machines as a tool for forecasting geomagnetic storms during years 1996–2007. *Journal of Atmospheric and Solar-Terrestrial Physics*, *125*, 10–20.

Bala, R., & Reiff, P. (2012). Improvements in short-term forecasting of geomagnetic activity. *Space Weather*, *10*(6).

Bartels, J., Heck, N. H., & Johnston, H. F. (1939). The three-hour-range index measuring geomagnetic activity. *Journal of Geophysical Research*, *44*(4), 411. Retrieved from `http://doi.wiley.com/10.1029/TE044i004p00411` doi: 10.1029/TE044i004p00411

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*.

Bolduc, L. (2002). GIC observations and studies in the hydro-Québec power system. *Journal of Atmospheric and Solar-Terrestrial Physics*, *64*(16), 1793 - 1802. Retrieved from `http://www.sciencedirect.com/science/article/pii/S1364682602001281` (Space Weather Effects on Technological Systems) doi: https://doi.org/10.1016/S1364-6826(02)00128-1

Brahma, P. P., Wu, D., & She, Y. (2016). Why Deep Learning Works: A Manifold Disentanglement Perspective. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(10), 1997–2008. doi: 10.1109/TNNLS.2015.2496947

Bröcker, J., & Smith, L. A. (2007, jun). Increasing the Reliability of Reliability Diagrams. *Weather and Forecasting*, *22*(3), 651–661. Retrieved from `http://journals.ametsoc.org/doi/abs/10.1175/WAF993.1` doi: 10.1175/WAF993.1

Burton, R. K., McPherron, R., & Russell, C. (1975). An empirical relationship between interplanetary conditions and Dst. *Journal of geophysical research*, *80*(31), 4204–4214.

Chakraborty, S., & Morley, S. K. (2020, jul). Probabilistic prediction of geomagnetic storms and the K{$_p$} index. *Journal of Space Weather and Space Climate*, *10*, 36. Retrieved from `http://arxiv.org/abs/2007.02733https://www.swsc-journal.org/10.1051/swsc/2020037` doi: 10.1051/swsc/2020037

Daglis, I. A. (2007, feb). Ring Current Dynamics. *Space Science Reviews*, *124*(1-4), 183–202. Retrieved from `http://link.springer.com/10.1007/s11214-006-9104-z` doi: 10.1007/s11214-006-9104-z

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.

Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2002). Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research*, *3*(1), 115–143. doi: 10.1162/153244303768966139

Gleisner, H., Lundstedt, H., & Wintoft, P. (1996). Predicting geomagnetic storms from solar-wind data using time-delay neural networks. *Annales Geophysicae*, *14*, 679.

Gonzalez, W. D., Tsurutani, B. T., & De Gonzalez, A. L. C. (1999). Interplanetary origin of geomagnetic storms. *Space Science Reviews*, *88*(3-4), 529-562.

Grayver, A. V., Kuvshinov, A. V., & Werthmüller, D. (2020). Time-domain modelling of 3-D Earth's and planetary EM induction effect in ground and satellite observations. *in prep.*.

Gruet, M., 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.

Gu, Y., Wei, H. L., Boynton, R. J., Walker, S. N., & Balikhin, M. A. (2019). System Identification and Data-Driven Forecasting of AE Index and Prediction Uncertainty Analysis Using a New Cloud-NARX Model. *Journal of Geophysical Research: Space Physics*, *124*(1), 248–263. doi: 10.1029/2018JA025957

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

539 Jankovičová, D., Dolinskỳ, P., Valach, F., & Vörös, Z. (2002). Neural network-
540 based nonlinear prediction of magnetic storms. *Journal of atmospheric and solar-*
541 *terrestrial physics*, *64*(5-6), 651–656.
542 Kelbert, A. (2020). *The Role of Global/Regional Earth Conductivity Models in*
543 *Natural Geomagnetic Hazard Mitigation* (Vol. 41) (No. 1). Springer Netherlands.
544 Retrieved from `https://doi.org/10.1007/s10712-019-09579-z` doi: 10.1007/
545 s10712-019-09579-z
546 Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv*
547 *preprint arXiv:1412.6980*.
548 Kugblenu, S., Taguchi, S., & Okuzawa, T. (1999). Prediction of the geomagnetic
549 storm associated Dst index using an artificial neural network algorithm. *Earth*
550 *Planets Space*, *51*(307-313).
551 Lazzús, J., Vega, P., Rojas, P., & Salfate, I. (2017). Forecasting the Dst index using
552 a swarm-optimized neural network. *Space Weather*, *15*(8), 1068–1089.
553 Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993). Multilayer feed-
554 forward networks with a nonpolynomial activation function can approximate
555 any function. *Neural Networks*, *6*(6), 861 - 867. Retrieved from `http://`
556 `www.sciencedirect.com/science/article/pii/S0893608005801315` doi:
557 https://doi.org/10.1016/S0893-6080(05)80131-5
558 Lin, H. W., Tegmark, M., & Rolnick, D. (2017, sep). Why Does Deep and Cheap
559 Learning Work So Well? *Journal of Statistical Physics*, *168*(6), 1223–1247. Re-
560 trieved from `http://link.springer.com/10.1007/s10955-017-1836-5` doi: 10
561 .1007/s10955-017-1836-5
562 Love, J. J., Lucas, G. M., Kelbert, A., & Bedrosian, P. A. (2018). Geoelectric
563 Hazard Maps for the Mid-Atlantic United States: 100 Year Extreme Values
564 and the 1989 Magnetic Storm. *Geophysical Research Letters*, *45*(1), 5–14. doi:
565 10.1002/2017GL076042
566 Maus, S., & Weidelt, P. (2004). Separating the magnetospheric disturbance mag-
567 netic field into external and transient internal contributions using a 1D con-
568 ductivity model of the Earth. *Geophysical Research Letters*, *31*(12), 2–5. doi:
569 10.1029/2004GL020232
570 Munsami, V. (2000). Determination of the effects of substorms on the storm-
571 time ring current using neural networks. *Journal of Geophysical Research: Space*

Physics, *105*(A12), 27833–27840.

Olsen, N., & Kuvshinov, A. (2004). Modeling the ocean effect of geomagnetic storms. *Earth, planets and space*, *56*(5), 525–530.

Olsen, N., Sabaka, T. J., & Lowes, F. (2005). New parameterization of external and induced fields in geomagnetic field modeling, and a candidate model for igrf 2005. *Earth, planets and space*, *57*(12), 1141–1149.

Pallocchia, G., Amata, E., Consolini, G., Marcucci, M., & Bertello, I. (2006). Geomagnetic Dst index forecast based on IMF data only. *Annales Geophysicae*, *24*(3), 989–999.

Patel, K., Singh, A., Singh, S. B., & Singh, A. K. (2019). Causes responsible for intense and severe storms during the declining phase of Solar Cycle 24. *Journal of Astrophysics and Astronomy*, *40*(1), 1–9. Retrieved from `https://doi.org/10.1007/s12036-018-9569-7` doi: 10.1007/s12036-018-9569-7

Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta numerica*, *8*, 143–195.

Püthe, C., Manoj, C., & Kuvshinov, A. (2014). Reproducing electric field observations during magnetic storms by means of rigorous 3-D modelling and distortion matrix co-estimation. *Earth, Planets and Space*, *66*(1), 162.

Revallo, M., Valach, F., Hejda, P., & Bochníček, J. (2014). A neural network Dst index model driven by input time histories of the solar wind-magnetosphere interaction. *Journal of Atmospheric and Solar-Terrestrial Physics*, *110*, 9–14.

Sharifie, J., Lucas, C., & Araabi, B. N. (2006). Locally linear neurofuzzy modeling and prediction of geomagnetic disturbances based on solar wind conditions. *Space Weather*, *4*(6).

Shen, X. C., Hudson, M. K., Jaynes, A. N., Shi, Q., Tian, A., Claudepierre, S. G., ... Sun, W. J. (2017). Statistical study of the storm time radiation belt evolution during Van Allen Probes era: CME- versus CIR-driven storms. *Journal of Geophysical Research: Space Physics*, *122*(8), 8327–8339. doi: 10.1002/2017JA024100

Shprits, Y. Y., Vasile, R., & Zhelavskaya, I. S. (2019). Now-casting and predicting the kp index using historical values and real-time observations. *Space Weather*, *17*. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018SW002141` doi: 10.1029/2018SW002141

Stepanova, M., Antonova, E., & Troshichev, O. (2005). Prediction of Dst variations from polar cap indices using time-delay neural network. *Journal of atmospheric and solar-terrestrial physics*, *67*(17-18), 1658–1664.

Stepanova, M., & Pérez, P. (2000). Autoprediction of Dst index using neural network techniques and relationship to the auroral geomagnetic indices. *GEOFISICA INTERNACIONAL-MEXICO-*, *39*(1), 143–146.

Temerin, M., & Li, X. (2006). Dst model for 1995–2002. *Journal of Geophysical Research: Space Physics*, *111*(A4).

Tobiska, W., Knipp, D., Burke, W., Bouwer, D., Bailey, J., Odstrcil, D., ... Bowman, B. (2013). The Anemomilos prediction methodology for Dst. *Space Weather*, *11*(9), 490–508.

Wei, H.-L., Zhu, D.-Q., Billings, S. A., & Balikhin, M. A. (2007). Forecasting the geomagnetic activity of the Dst index using multiscale radial basis function networks. *Advances in Space Research*, *40*(12), 1863–1870.

Wolf, R., Freeman Jr, J., Hausman, B., Spiro, R., Hilmer, R., & Lambour, R. (1997). Modeling convection effects in magnetic storms. *Magnetic Storms*, *98*, 161–172.

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.

Wu, J.-G., & Lundstedt, H. (1997). Geomagnetic storm predictions from solar wind data with the use of dynamic neural networks. *Journal of Geophysical Research: Space Physics*, *102*(A7), 14255–14268.

Zhang, J., Richardson, I., Webb, D., Gopalswamy, N., Huttunen, E., Kasper, J., ... others (2007). Solar and interplanetary sources of major geomagnetic storms (Dst $\leq$ -100 nT) during 1996–2005. *Journal of Geophysical Research: Space Physics*, *112*(A10).