Probabilistic prediction of geomagnetic storms and the $K_p$ index

Shibaji Chakraborty$^{1,*}$ and Steven Karl Morley$^2$

$^1$ Bradley Department of Electrical & Computer Engineering, Virginia Tech, Blacksburg, 24061 VA, USA
$^2$ Space Science and Applications (ISR-1), Los Alamos National Laboratory, Los Alamos, 87545 NM, USA

Received 23 December 2019 / Accepted 4 July 2020

Abstract – Geomagnetic activity is often described using summary indices to summarize the likelihood of space weather impacts, as well as when parameterizing space weather models. The geomagnetic index $K_p$, in particular, is widely used for these purposes. Current state-of-the-art forecast models provide deterministic $K_p$ predictions using a variety of methods – including empirically-derived functions, physics-based models, and neural networks – but do not provide uncertainty estimates associated with the forecast. This paper provides a sample methodology to generate a 3-hour-ahead $K_p$ prediction with uncertainty bounds and from this provide a probabilistic geomagnetic storm forecast. Specifically, we have used a two-layered architecture to separately predict storm ($K_p \geq 5$) and non-storm cases. As solar wind-driven models are limited in their ability to predict the onset of transient-driven activity we also introduce a model variant using solar X-ray flux to assess whether simple models including proxies for solar activity can improve the predictions of geomagnetic storm activity with lead times longer than the L1-to-Earth propagation time. By comparing the performance of these models we show that including operationally-available information about solar irradiance enhances the ability of predictive models to capture the onset of geomagnetic storms and that this can be achieved while also enabling probabilistic forecasts.

Keywords: geomagnetic storms / $K_p$ forecasting / deep learning / LSTM / Gaussian process

1 Introduction

Modern electrical systems and equipment on the Earth such as navigation, communication, satellite and power grid systems can be affected by space weather (e.g., Choi et al., 2011; Qiu et al., 2015; Morley, 2019). The societal impact of space weather is increasing (Schrijver et al., 2015; Eastwood et al., 2017) and operational centers provide a range of predictions for end-users (Bingham et al., 2019), including geomagnetic storm predictions based on the $K_p$ index (Sharpe & Murray, 2017).

$K_p$ is a planetary 3-h averaged range index that describes the intensity of the magnetic disturbance (Bartels, 1949). $K_p$ starts from 0 (very quiet) to 9 (very disturbed) with 28 discrete values described by 0, 0”, 1”, 1, 1”, ..., 9”, 9 (Bartels, 1949). The National Oceanic and Atmospheric Administration (NOAA) Space Weather Prediction Center (SWPC) classifies geomagnetic activity in six levels, shown in Table 1, based on the ranges of $K_p$. In addition to being a forecast product in its own right, the $K_p$ index is widely used as an input to other magnetospheric models (e.g., Carberry, 2005) including some aimed at operational use (Obrien, 2009; Horne et al., 2013). There are three main categories of models for $K_p$ prediction: coupled physics-based models such as the Space Weather Modeling Framework (e.g., Töth et al., 2005; Haiducek et al., 2017); “traditional” empirical models (Newell et al., 2007; Luo et al., 2017); and machine-learning models (Costello, 1998; Boberg et al., 2000; Wing et al., 2005; Wintoft et al., 2017; Tan et al., 2018; Sexton et al., 2019).

Longer lead-time predictions are typically the domain of empirical (e.g., Luo et al., 2017) and machine-learning models (e.g., Sexton et al., 2019). Since the first neural network-based $K_p$ forecasting model proposed by Costello (Costello, 1998), many subsequent forecast models (Boberg et al., 2000; Wing et al., 2005; Tan et al., 2018) implement different variants of the neural network to improve the forecast accuracy. To date, none of these predictions have provided a probabilistic prediction, and very few attempted to characterize the uncertainty associated with the predicted $K_p$ value (see, e.g., Wintoft et al., 2017). With the development of new machine-learning techniques, recent $K_p$ and storm forecast models come with much higher accuracy, but few have separately examined the model performance under different ranges of geomagnetic storm conditions (see, e.g., Zhelavskaya et al., 2019). Recent work on $K_p$ prediction by (Shprits et al., 2019) highlighted the inherent limitation of solar wind-driven models for long lead-time...
predictions, and noted that “further improvements in long term modeling should include … empirical modeling driven by observations of the Sun”.

To assess the viability of moving beyond solar wind-driven models using operationally-available data, we also investigate the inclusion of solar X-ray flux data as a model parameter. Solar X-ray flux was chosen as recent studies have shown that these data can be used to forecast solar flare activity (Winter & Balasubramaniam, 2015) as well as solar energetic particle (SEP) events (Núñez, 2018). While the majority of large geomagnetic storms are caused by CMEs (e.g., Gonzalez et al., 2015) as well as solar energetic particle (SEP) events (Núñez, 2018), there is a long history of using plasma moments (number density and velocity) and the interplanetary magnetic field to describe the coupling of energy from the solar wind into the magnetosphere (e.g., Baker et al., 1981; Borovsky et al., 1998; Xu & Borovsky, 2015). More recently-developed coupling functions using solar wind parameters have been shown to have better correlations with geomagnetic indices (e.g., Newell et al., 2007) and clear physical origins (e.g., Borovsky, 2013). It is also clear that including a measure of the present state of the magnetosphere helps predict the evolution of global indices like \( K_p \) (Borovsky, 2014; Luo et al., 2017).

Based, in part, on these considerations we use just nine solar wind parameters and the historical \( K_p \) as model features (input parameters). The input parameters are chosen based on the studies done by the previous studies (Newell et al., 2007; Xu & Borovsky, 2015). These model features are listed in Table 2 along with the notation we use in this paper and any transformations applied prior to model training. For the solar wind data, we use 20 years of 1-min resolution values, starting from 1995, obtained from NASA’s OMNIWeb service (https://omniweb.gsfc.nasa.gov/ow_min.html). The 3-hourly \( K_p \) index is obtained from the GFZ German Research Centre for Geosciences at Potsdam (https://www.gfz-potsdam.de/en/kp-index/).

As solar wind structures are spatial in nature, the measured parameters are auto-correlated. Solar wind data from L1 monitors are point measurements and hence spatial structure along the Sun–Earth line manifests as a temporal correlation. Hence, we performed an auto-correlation analysis of all the solar wind parameters as presented in Figure 1. From the figure, we can conclude that, during solar minimum, most of the solar wind parameters are highly autocorrelated for a longer duration, while during solar maximum, the correlation coefficient drops within a few hours. This suggests more transients in solar wind during solar maximum than the solar minimum, consistent with observations (Richardson & Cane, 2012). All parameters selected as model features display auto-correlation, and most parameters decorrelate within 3 h, with solar wind speed being a notable exception. Indeed, at solar maximum, all parameters except solar wind speed decorrelate in less than 3 h. At solar minimum, the auto-correlation for the majority of parameters falls below 0.5 within 3 h. We, therefore, used 3-hourly averaged solar wind data to train our models, which has the added benefit of placing the input data on the same cadence as the predicted output. Note that the goal of this linear analysis (autocorrelation) is to describe the redundant information given in subsequent samples of individual solar wind parameters, rather than identifying time lags in the magnetospheric response to solar wind driving. In addition, we acknowledge that including nonlinear correlations may provide more robust estimates of the correlation scales, and could exhibit different behaviors (e.g., Wing et al., 2016; Johnson et al., 2018).

**Table 1.** Table showing different categories of geomagnetic storm and associated \( K_p \) levels. The categorization is done based on the intensity of the geomagnetic storm following the NOAA SWPC scales.

| Storm level | \( K_p \) range          |
|-------------|-------------------------|
| \( G_0 \)   | \( K_p < 5 \)           |
| \( G_1 \)   | \( 5 \leq K_p < 6 \)    |
| \( G_2 \)   | \( 6 \leq K_p < 7 \)    |
| \( G_3 \)   | \( 7 \leq K_p < 8 \)    |
| \( G_4 \)   | \( 8 \leq K_p < 9 \)    |
| \( G_5 \)   | \( K_p > 9 \)           |

**2 Data analysis & model architecture**

Here we describe the data sets and the model architecture used in this study. Specifically, we present the basic characteristics of the data sets and a brief justification of our choices of model features. In addition, we provide a short introduction to the technical terms and metrics used to evaluate the model performance. Finally, we describe the construction of our predictive models and note some strengths and weaknesses of our approach.
As the Kp index is quasi-logarithmic and essentially categorical (Mayaud, 1980), we converted the reported Kp values to decimal numbers using \( k/C6 \) following Tan et al. (2018).

In this study, we will also test the effectiveness of including a previously-unused set of solar data for Kp prediction. That is, we will introduce our modeling approach using a standard construction with only upstream solar wind measurements to drive the model. We will then train a second model that differs only in the inclusion of X-ray flux in the model features.

As described above, the idea behind using the solar X-ray data is that the upstream solar wind carries little to no information about transients that are moving towards Earth. Advanced notice of transients from in-situ L1 measurements is therefore limited to periods typically less than 1 h. By including solar data, even a coarse measure such as the X-ray flux, we aim to demonstrate that additional information about the likelihood of transients can be included in the model and improve forecasts with a lead time longer than the L1-to-Earth propagation time.

In other words, we treat the X-ray flux data as a proxy for the magnetic complexity of the Sun and anticipate that including this data will allow the model to predict the arrival of CMEs earlier than a model-driven purely by solar wind measurements. We obtain the GOES X-ray flux data from NOAA’s National Centers for Environmental Information (https://satdat.ngdc.noaa.gov/sem/goes/data/).

The GOES X-ray sensors have two channels that measure solar irradiance in two wavebands, namely, hard (0.05–0.4 nm) and soft (0.1–0.8 nm) X-rays. In this study, we followed Winter & Balasubramaniam (2015) in using a background term and a flux ratio derived from the two GOES wavebands. The X-ray background \( B/Wm^2 \) has been computed as the smoothed minimum flux in a 24-h window preceding each 1-min GOES soft X-ray flux observation. In a recent study, the X-ray background parameter was found to describe the solar activity cycle better than the daily F10.7 parameter (Winter & Balasubramaniam, 2014). The X-ray flux ratio \( R \) has been calculated by taking

\[
\text{Box-Cox} \rightarrow B(x) = \text{sgn}(x) \times \log_e x \quad \text{and} \quad \text{Erf} \rightarrow \text{erf}(x) = \text{sgn}(x) \times \sqrt{2}\text{erf}^{-1} \left( 1 - 2e^{-ax^2} \right).
\]

### Table 2

| Features | Symbol | Units | Transformation | Model |
|----------|--------|------|----------------|-------|
| Anti-sunward component of IMF | \( B_x \) | nT | Box-Cox | All |
| Transverse component of IMF | \( B_T = \sqrt{B_y^2 + B_z^2} \) | nT | Box-Cox | All |
| IMF clock angle | \( \theta_x = \text{tan}^{-1} \left( \frac{B_y}{B_z} \right) \) | radian | Box-Cox | All |
| Bulk velocity | \( v = \sqrt{v_x^2 + v_y^2 + v_z^2} \) | km s\(^{-1}\) | Box-Cox | All |
| Number density | \( n \) | m\(^{-3}\) | Box-Cox | All |
| Electron temperature | \( T \) | K | Box-Cox | All |
| Dynamic pressure | \( P_{\text{dyn}} \) | Nm\(^{-2}\) | Box-Cox | All |
| Mach number | \( M_a \) | 1 | Box-Cox | All |
| Plasma beta | \( \beta \) | 1 | Erf | All |
| Historical Kp | \( K_{p}^{(-3)} \) | 1 | Box-Cox | All |
| Background X-ray | \( B \) | Wm\(^{-2}\) | None | LSTM\(_s\), dGPR\(_s\) |
| X-ray flux ratio | \( R \) | 1 | None | LSTM\(_s\), dGPR\(_s\) |

**Fig. 1.** Auto-correlation functions of different solar-wind parameters during: (a) solar minima, and (b) solar maxima.
the ratio of hard X-ray flux over soft X-ray flux, and provides a measure of the temperature of the flare emission (García, 1994). Kay et al. (2003) showed that flares associated with CMEs tended to have lower temperatures, at a given intensity level, than flares without CMEs. Thus both the soft X-ray flux and the flux ratio, $R$, are important for determining the likelihood of an eruptive event. Further, Michalek (2009) showed a good correlation between the energy of the CME and the peak of the X-ray flare. Finally, recent studies showed that the X-ray flux ratio is a good predictor for extreme solar events (Kahler & Ling, 2018). As X-rays propagate from the Sun to the Earth at the speed of light, there will, of course, be a time lag associated with the arrival at Earth of any related geomagnetic activity due to associated coronal mass ejections. In this preliminary work to demonstrate the utility of including solar X-ray flux data in a $K_p$ prediction model, we assume a constant time lag that we apply to the derived X-ray products $B$ and $R$. Figure 2 presents a time-lagged cross-correlation analysis of $B$ and $R$ with other solar wind parameters to highlight any time lags between these two data sets. The correlation analysis shows that the X-ray background ($B$) parameter is significantly correlated with many solar wind parameters at lags around 48 h. The correlations between the X-ray flux ratio ($R$) and the solar wind parameters are smaller and less consistent across solar wind parameters. A lack of clear, or strong, linear correlations with solar wind parameters at a given fixed lag does not necessarily indicate that the parameter $R$ is confounding. Better lag estimates could be obtained using nonlinear analysis (e.g., Wing et al., 2016; Johnson et al., 2018), however, models used in this study can extract nonlinear relationships. We therefore expect nonlinear relationships in the dataset to be captured by the proposed models.

Transit times of coronal mass ejections can range from less than 20 h to more than 80 h (e.g., Schwenn et al., 2005), however faster CMEs tend to be more geoeffective (Gonzalez et al., 1999; Srivastava & Venkatakrishnan, 2002). Hence we, therefore, apply a constant time lag of 48 h to the X-ray flux derived parameters, consistent with a typical travel time from the Sun to Earth of geomagnetic-storm associated interplanetary CMEs (Srivastava & Venkatakrishnan, 2004). Although we note that in future work it may be useful to explore the effects of choosing this lag over a different fixed lag, or even use of a variable lag.

2.2 Technical definitions & metrics for model evaluation

In this subsection, we define the technical terms and the metrics that we use in the latter part of this study to evaluate and compare the models’ performance. Good overviews of model performance metrics and model validation methodologies targeted at space weather have been given by Morley et al. (2018a), Liemohn et al. (2018), and Camporeale (2019). For binary event analysis, we define correct “yes” events as True Positives (denoted as $a$). Similarly, we call incorrect “yes” events False Positives ($b$), incorrect “no” events are False Negatives ($c$), and correct “no” events are True Negatives ($d$).

ROC curve = A graphical diagnostic illustration of a binary classifier.

\[
\text{AUC} = \text{Area under the ROC curve.}
\]

Probability of detection (PoD) = \( \frac{a}{a + c} \)

Probability of false detection (PoFD) = \( \frac{b}{b + d} \)

False alarm ratio (FAR) = \( \frac{b}{a + b} \)

Bias = \( \frac{a + b}{a + c} \)

Critical success index (CSI) = \( \frac{a}{a + b + c} \)

Matthews correlation coefficient (MCC) = \( \frac{a \times d - b \times c}{\sqrt{(a + b)(a + c)(b + d)(c + d)}} \)

Heidke skill score (HSS) = \( \frac{2 \times (ad - bc)}{(a + c)(c + d) + (a + b)(b + d)} \)

Note that a perfect forecast will have HSS, MCC, PoD, Bias, CSI, $R^2$ equal to 1 and FAR equal to 0. Unskilled or
random forecasts will have HSS, MCC, PoD, CSI, $R^2$ of 0, FAR of 1.

For assessing numerical predictions of the $K_p$ index we use:

- **Root mean square error (RMSE)** ($\varepsilon$) = $\sqrt{\frac{1}{n} \sum (x_i - \hat{x}_i)^2}$
- **Mean absolute error (MAE)** ($|\hat{d}|$) = $\frac{1}{n} \sum |x_i - \hat{x}_i|$  
- **Coefficient of determination ($R^2$)** = $1 - \frac{\sum (x_i - \hat{x}_i)^2}{\sum (x_i - \bar{x})^2}$

where, $x_i$, $\hat{x}_i$, and $n$ are observations, predicted output and a total number of observations, respectively. A perfect model will have zero RMSE and MAE, while a coefficient of determination of 1.

### 2.3 Model development

Figure 3 presents the distribution of 22 years of $K_p$ values, where the log-scaled frequency of occurrence is shown on the vertical axis and the $K_p$ value is shown on the horizontal axis. From the figure, it is evident that most of the events are distributed between $[0, 5^-]$, a relatively small number of events are distributed between $[5^-, 7^+]$, and there are very few extreme events $\geq 7$.

Following the NOAA G-scale for geomagnetic storms, we choose $K_p \geq 5^-$ as the threshold for “storm-time”, marked with a vertical line in Figure 3. Using this division, storm-time comprises approximately one-twentieth of the data set. This number continues to drop rapidly as $K_p$ increases. If we take the ratio of more extreme events ($K_p \geq 8^+$) versus non-storm events, the number drops to $\approx \frac{1}{216}$. This effect is known as data imbalance (e.g., Estabrooks et al., 2004) and can lead to significant errors in a model fit to the data without accounting for the imbalance (see, e.g., Shprits et al., 2019).

Both oversampling and undersampling techniques have been used to address data imbalance (Estabrooks et al., 2004; Shprits et al., 2019), and both methods help develop models with a better predictive skill (Yap et al., 2014). Choosing a single regression model to predict $K_p$ across quiet and disturbed geomagnetic conditions will likely not provide an optimal forecast unless the data imbalance is addressed. Here we take a similar approach to Tan et al. (2018), Boberg et al. (2000) and minimize the data imbalance by separating the problem into two different categories: storm-time and non-storm. This then leads to the first step in our model architecture, where we use a classifier to determine quiet or active conditions, and subsequently use a probabilistic regressor to predict the $K_p$ value.

In this study, we have used a deterministic long-short term memory (LSTM) neural network (Hochreiter & Schmidhuber, 1997). An LSTM is a type of recurrent neural network, where a “memory cell” is used to store information between time steps (see Tan et al., 2018, and references therein). LSTM neural networks are a special type of recurrent neural network, well-suited to time series data analysis, and they require continuous data for training. However, we encounter missing IMF and solar wind data values issue. To handle the missing data issue we use the interpolation method described in Tan et al. (see Sect. 2.1 of 2018, and references therein). The method used by Tan et al. (2018) is appropriate for relatively small (up to 12 samples) data gaps, for larger data gaps we discarded the samples. As with other types of neural networks, an LSTM can be used as either a regressor or a classifier. The layout of our LSTM classifier is shown schematically in Figure 4. Panel 4a presents a schematic of a single LSTM unit. The LSTM unit consists of input, output, forget gates, and memory cell ($C_t$) (see Hochreiter & Schmidhuber, 1997, and references therein). Panel 4b illustrates the overall architecture of the classifier, where the central layer is comprised of LSTM units. Panel 4c shows the implementation as layers in the Keras model. The “N” in the input/output shape of the model blocks shows the number of time samples, which can be varied at run-time. However, as described later in this section we use a 27-day window at 3-hourly cadence, therefore $N = 216$.

As described in the previous paragraph LSTM classifier comprises an input layer with 10 neurons, a hidden layer with 100 neurons, and an output layer of 2 neurons. The MSE on the validation data reduced as the number of neurons in our hidden layer increased, but adding more than ~100 LSTM units led to smaller improvements. We therefore chose to retain only 100 LSTM units in our hidden layer. To help reduce overfitting and increase the generalizability of the model we included dropout regularization. The input shape of the data is therefore $(N \times 10 \times 3)$, where $N = 216$, 10 and 3 are the number of data points, the number of input parameters (see Table 2), and the time-lagged units (the input vector at times $t$, $t - 1$, and $t - 2$), respectively. Hence, the input shape for one data point is $10 \times 3$. Note, the input shape for one data point can also be $12 \times 3$, based on the choice of model ($\mu_{\text{OMNI}}$ or $\mu_{\text{OMNI}}^*$, see the following paragraphs for details). To ensure that the classifier performance can be generalized beyond the training data, we split available data into two categories: training/validation and testing. For training and validation we used the intervals 1995–2000 and 2011–2014. To mitigate the effects of data imbalance we used a random downsampling strategy to balance the storm-time and no-storm intervals. After downsampling (from 29,200 to 5716 points), we split the data into “training” and “validation” sets and train the classifier, where the validation data is used for tuning the model parameters and comprises 20% of the training set. Data from 2001–2010 was reserved for out-of-sample testing of the predictions (26,645 points). The rationale behind using the above mentioned periods for train/validation and test the model is to increase the model performance throughput and reduce the model bias. Both train and test periods consist of different solar maximum and minimum data.
to capture solar cycle dynamics and testing with out-of-sample data ensures that the model generalizes well to unseen data.

Following Tan et al. (2018) we employ a two-layer architecture, where we use separate regression models to predict $K_p$ under quiet or active geomagnetic conditions. Unlike Tan et al. (2018) we use probabilistic regressors. The model structure is shown in Figure 5. The prediction made by the classifier is used to determine which regressor is going to be selected. As the primary aim of this work is to produce a probabilistic prediction of $K_p$, we chose regression models that output a distribution rather than a single value. We used semi-parametric Deep Gaussian Process Regression, commonly known as deepGPR, to build the regressors. DeepGPR (dGPR) is a Gaussian Process model with neural network architecture. A Gaussian Process is a random process that follows a multivariate normal distribution. Specifically, dGPR (Al-Shedivat et al., 2017) is a Gaussian Process Regression (Rasmussen & Williams, 2006) method, which uses a deep LSTM network to optimize the hyperparameters of the Gaussian Process kernels.

For example, if the classifier predicts a geomagnetic storm then regressor dGPR_s is selected. Otherwise regressor dGPR_q is used. At each time step, the dGPR is retrained using the interval of preceding data (the training window), and thus our regressors are dynamic non-linear models. Dynamic models do not need to assume a constant functional relationship between the model covariates (e.g., solar wind drivers) and response ($K_p$). Static models implicitly combine the effects of any potentially time-varying relationships in the error terms or average over the effects in the estimation of model coefficients (see, e.g., Osthus et al., 2014). By using a relatively short, local training window, the data is used more efficiently and computational complexity is reduced. For training and validation of the dGPR-based dynamic model, including training window selection, we used the mean squared error (MSE) as our loss function. The quiet-time and storm-time regressors each use different training windows, the lengths of which were selected to minimize the training error using the mean squared error (MSE) in $K_p$, as the loss function. Figure 6 shows one example of how the MSE varies with the training window (in days) for predictions over two months during 1995. It can be clearly seen that a training window of $\approx 27$ days is optimal at this time, as this captures recurrent structure in the solar wind such as corotating interaction regions (Richardson & Cane, 2012). While the deep architecture helps to capture the nonlinear trends in data to provide better accuracy, the Gaussian process mappings are used to provide the error distribution with a mean predicted $K_p$. The two dGPR regressors are different in terms of the length of the training window used for forecasting. The dGPR module dedicated for non-storm, or quiet, conditions has a 27-day training window.
window, whereas the dGPR module for storm conditions uses a
14-day training window.

One of the difficulties in predicting the “events” – i.e., the
storm-time \( K_p \) values – is that these are typically driven by solar
wind transients, which include interplanetary CMEs and coro-
tating interaction regions (CIRs) (see, e.g., Kilpua et al., 2009;
Zhang et al., 2018), with the largest storms driven by CMEs
(Borovsky & Denton, 2006). The in-situ solar wind measure-
ments from an L1 monitor do not convey the information
required to predict the occurrence of these transients for a
3-hour-ahead prediction of \( K_p \), or for longer prediction horizons.
For this reason, we perform a preliminary investigation in which
we include information that may encode the likelihood of CME
eruption. Following Winter & Balasubramaniam (2014) we use
X-ray flux data from the NOAA GOES platforms as a measure
of possible solar activity.

To test whether the inclusion of a proxy for solar activity
improves our ability to predict storm-time \( K_p \), we constructed
two different prediction systems. The first was trained only on
OMNI solar wind parameters (\( \mu^{\text{OMNI}} \)). The second added the
X-ray background (\( B \)) and the flux ratio (\( R \)) as extra model
parameters (\( \mu^{\text{OMNI}+} \)). When using the X-ray data we add
\( B \) and \( R \) as model features to the LSTM classifier as well as
the storm-time regressor. Note that we do not use the X-ray data
for the quiet-time regressor. Both the models are validated and
evaluated against 10 years of \( K_p \) (2001–2010), in addition to a
specific storm-time validation using 38 intervals listed in
Table 3.

### 3 Results

In this section, we present model forecasts and quantitative
comparison of predicted \( K_p \), comparing the models with and
without the GOES X-ray data. We further describe a simple
method to exploit the uncertainty bounds of the predicted \( K_p \)
to provide a probabilistic geomagnetic storm prediction. Finally,
we analyze the performance of the probabilistic \( K_p \) prediction
models.

#### 3.1 \( K_p \) forecast models

As the first step, we present 3-h ahead predicted \( K_p \) during
two months of 2004. Panels (a) and (b) of Figure 7 shows
predicted \( K_p \) with a 95% confidence interval using models
\( \mu^{\text{OMNI}} \) and \( \mu^{\text{OMNI}+} \), respectively. The horizontal dashed black
line shows the storm versus non-storm demarcation line. The
root mean square error (\( \varepsilon \)) and mean absolute error (\( |\delta| \)) for this
2 month interval are given each panel as annotations. In the
figure, we have discretized the \( K_p \) values and the 95% confi-
dence interval bounds by rounding them to the nearest valid
\( K_p \) values (see Sect. 4.2 of Morley et al., 2018b). We have
chosen this time interval as an example to showcase the ability
of the model to capture the dynamics of both storm-time and
quiet-time \( K_p \). Examining Figure 7, it is visually apparent that
both the models can capture the change in \( K_p \) during the transi-
tions into, and out of, storm-time. The error metrics given in
each panel suggest that models \( \mu^{\text{OMNI}} \) and \( \mu^{\text{OMNI}+} \) are compar-
able in their performance. However, a more detailed analysis is
required to allow us to draw firm conclusions and assess the
impact of including the X-ray flux data. Specifically, as the in-
tent of including the X-ray flux is to better capture storm-time
transients, we need validation methods that allow us to deter-
mine the performance as a function of activity level.

We have conducted a thorough head-to-head test of two \( K_p \)
forecast models, \( \mu^{\text{OMNI}} \) and \( \mu^{\text{OMNI}+} \), using predictions across our
validation data set (from January 2001 through December 2010). We also compare the model predictions for 38 storm-
time intervals (listed in Table 3). Summary statistics for the
different models are presented in Table 4. When comparing
the models across the full validation set, the error metrics are
nearly identical between the model variants. The RMSE,
MAE and correlation coefficient for both of our models show
similar performance to the models of Boberg et al. (2000) and
Balu & Reiff (2012). On the full data set our model does not
perform as well as that of Tan et al. (2018). However, in addi-
tion to generating a probabilistic forecast, we seek to answer
the question of whether including GOES X-ray data provide a
meaningful improvement in the prediction of storm-time \( K_p \)
intervals. Looking at the same metrics for the 38 storm-time
intervals, a different picture emerges. The model variant incor-
porating X-ray flux data (\( \mu^{\text{OMNI}+} \)) outperforms the standard
model by a substantial margin. The RMSE of \( \mu^{\text{OMNI}+} \) is 0.9
and the MAE is 0.67, showing that typical storm-time \( K_p \) pre-
dictions are within 1 \( K_p \) interval of observation. Of particular
importance is that the correlation coefficient for \( \mu^{\text{OMNI}+} \) is
0.93, whereas \( \mu^{\text{OMNI}} \) is 0.87 relative to the performance across the full validation
set. Here we note that the correlation coefficient can be consid-
ered a measure of potential performance, as it neglects condi-
tional and unconditional bias (Murphy, 1995).

To graphically display the model bias across these two val-
idation sets, Figure 8 plots observed \( K_p \) against predicted \( K_p \). Panel (a) shows the comparison across the full 10-year valida-
tion set and panel (b) shows the comparison for the 38 storm
intervals. Black and blue colors represent predicted \( K_p \) using
\( \mu^{\text{OMNI}} \) and \( \mu^{\text{OMNI}+} \), respectively. The circles give the mean pre-
dicted \( K_p \) and the vertical lines represent 1-\( \sigma \) error bars associ-
ated with that predicted \( K_p \) level. As above, the predictions from
both models are comparable when we use the full validation set
(Fig. 8a) and do not account for the activity level. For \( K_p \leq 3 \)
The predictions show little to no bias; the mean predicted \( K_p \) is
nearly identical to the observed value during quiet times. At higher levels of geomagnetic activity, we see a clear tendency for the mean predicted $K_p$ to be lower than the observation. That is, high values of $K_p$ tend to be underpredicted by both models. Comparing the two models shows a slight improvement in the bias at higher $K_p$ values using the $\mu_{\text{OMNI}}$ model, but the most visible improvement using this display format is in the smaller error bars on the $\mu_{\text{OMNI}}$ predictions.

Table 4 presents a summary of overall fit statistics, both for both model, for the 2001–2010 data set as well as the storm-time data set. On both data sets there is an improvement in the fit when adding the X-ray flux data to the model. Because the correlations have the observations in common, we test whether the improvement is significant using a test for correlated correlation coefficients (Steiger, 1980; Revelle, 2020), where we use the effective number of samples ($n_{\text{eff}}$ where $n_{\text{eff}} < n$) estimated by correcting for the lag-1 autocorrelation (see, e.g., Wilks, 2006). The improvements in $r$ for both models are statistically significant, with a $p$-value of $4.8 \times 10^{-5}$ for the 2001–2010 data set and $p < 10^{-5}$ for the storm-time data set. Given the results presented in both Table 4 and Figure 8, we conclude that including X-ray flux information provides information that improves $K_p$ prediction accuracy during geomagnetically active intervals.

However, this analysis only uses the mean prediction and neglects the fact that we have used a probabilistic regressor. So, how can we use the uncertainties provided by the dGPR prediction and how well do our probabilistic predictions perform?

### 3.2 Probabilistic storm forecast

Here we describe how we exploit the uncertainty in predicted $K_p$ to provide a probabilistic geomagnetic storm forecast using the SW-driven model $\mu_{\text{OMNI}}$ as an example. Figure 9 illustrates the probabilistic prediction of both the $K_p$ index and of geomagnetic storm occurrence. Figure 9a shows a “traffic light” display which gives the probability of $K_p$ exceeding 5 $-$, which we use to delineate storm-time, following the NOAA G-scale of geomagnetic storms (refer Table 1). The color represents the
likelihood of storm activity: green is \( Pr / C_20 \leq 0.3 \), yellow is \( 0.3 < Pr / C_20 \leq 0.66 \), and red marks intervals where the probability of geomagnetic storm conditions exceeds 0.66. Note that, \( Pr \) is the probability of geomagnetic storm, using the NOAA SWPC definition of \( K_p / C_5 \geq 5 / C_0 \) as the threshold for geomagnetic storm.

Figure 9b presents 10 days (21st–31st July 2004) of 3-h ahead predictions of \( K_p \), using OMNI model (cf. Fig. 7b). The horizontal dashed line is at \( K_p = 5 / C_0 \) and the vertical bar marks the time of the prediction shown in Figure 9c. Figure 9c illustrates how to calculate the probability of geomagnetic storm from the predicted distribution of \( K_p \). The blue curve gives the output Gaussian probability density function from the dGPR regressor while the blue and red vertical lines represent the mean prediction and the observed \( K_p \). The vertical dashed black line marks \( K_p = 5 / C_0 \) and the integral of the shaded area is the probability of exceeding the threshold.

In a non-probabilistic model, we would simply have a binary outcome of storm or non-storm. Following this method, we see that the prediction of a probability distribution gives us both uncertainty bounds on our prediction of the actual \( K_p \) as well as the probability of exceeding a given threshold.

To assess the probabilistic storm prediction (i.e., the probability of exceeding a threshold), we will examine binary event forecasts. For this we convert each probabilistic prediction into a prediction of “event” or “non-event”. For this, we need to choose a probability threshold. As our predicted probability distribution is Gaussian, the mean prediction is also the 50th percentile. Simply ignoring the predicted distribution and using the mean value is equivalent to using a threshold of 0.5. We can similarly convert the observed \( K_p \) to a series of events and non-events, and can subsequently determine whether the prediction was a true positive, false positive, false negative or true negative (see Sect. 2.2).
Figure 10 shows $K_p$ predictions using the $\mu_{OMNI}$ model during a geomagnetic storm at the end of March 2001. One forecast from each of the true positive (TP), false positive (FP), false negative (FN), and true negative (TN) categories are shown by the vertical lines. Figure 10 is in a similar format to Figure 9, except that panel (c) is omitted. We use the simpler model for this graphic to illustrate the main effect that our $\mu_{OMNI}$ model aims to combat. Specifically, the FP and FN cases are occurring in a specific pattern. At the start of the storm, we see a false negative and at the end of the storm, we see a false positive. This is typical for solar wind-driven models for predictions that are further ahead than the lead time given by the solar wind. In this case, we have a 3-h lead time to our prediction and so the model has no information to capture the sudden onset of the geomagnetic activity. By the next prediction, the model has “caught up” and now correctly predicts a very high likelihood of storm-time. The inverse is seen, though perhaps less clearly, at the trailing edge of the storm. The model is unable to predict the cessation of storm-level geomagnetic activity, although we do note that the uncertainty bounds include a non-negligible likelihood of $K_p < 5$.

The model prediction is, therefore, lagging the observation. While this figure shows one example where predicted $K_p$ is lagging the observations by 3 h, most of the storm-time predictions are lagging the observations by at least one 3-h prediction window. This implies that the model $\mu_{OMNI}$ has insufficient information to capture the imminent arrival of a solar wind transient from the L1 data alone, and the prediction is likely to be strongly persistent (giving high weight to the previous value of $K_p$) in the model. By including the X-ray data in the $\mu_{OMNI+}$ model as a proxy for the likelihood of CME occurrence, we aim to improve storm-time predictions and hopefully combat this “lag” effect.

3.3 Comparison of probabilistic predictions

In the previous sections, we described a method to use the uncertainty bound to predict probabilistic storm forecasts. Here we are going to compare the predictions using models ($\mu_{OMNI}$ and $\mu_{OMNI+}$), using the different metrics defined in Section 2. We begin with the receiver operating characteristic (ROC) curve. The ROC curve is calculated from the probability of detection (PoD) and the probability of false detection (PoFD) over a range of decision thresholds. If we make a decision threshold of $Pr = 0$, then all predictions lie above it and thus every time is predicted as an event and the PoD and PoFD are both 1. Conversely, taking a decision threshold of $Pr = 1$ leads to no events being predicted, thus PoD and PoFD are both zero.

Figure 11 presents the ROC curves calculated for both the $\mu_{OMNI}$ and $\mu_{OMNI+}$ models, using different $K_p$ threshold values. The solid lines are the ROC curves from model $\mu_{OMNI}$ and the dashed lines are the ROC curves from model $\mu_{OMNI+}$. Thresholds of $K_p = 2$, 4 and 6 are shown in red, black and blue, respectively. We also use the area under the ROC curve (abbreviated as AUC) as a summary measure to compare the performance of
our models (cf. Bradley, 1997), and the AUC for each ROC curve is given in the figure legend. For the lower \( K_p \) threshold values are shown (\( K_p = 2 \) or 4) the curves are similar and the AUC are correspondingly similar. For the higher \( K_p \) threshold value (\( K_p = 6 \)) the ROC curves visibly diverge across a broad range of decision thresholds. The AUC for \( \mu^{\text{OMNI}} \) is higher than that for \( \mu^{\text{OMNI}} \). This provides qualitative support for the hypothesis that the inclusion of GOES X-ray data has improved the performance of our \( K_p \) model for high geomagnetic activity.

As the aim of including the X-ray flux data was to potentially provide information relevant to predicting the intervals of higher \( K_p \), with a longer lead time, we also test the difference between the AUC for \( \mu^{\text{OMNI}} \) and \( \mu^{\text{OMNI}} \), when \( K_p \geq 6 \). Because the same test data is used for both ROC curves – the two blue curves in Figure 11 – we use DeLong’s nonparametric test for the area under correlated ROC curves (DeLong et al., 1988; Sun & Xu, 2014) as implemented in the pROC package (Robin et al., 2011). A two-sided test yielded (\( Z = -8.27; p < 2.2 \times 10^{-16} \)) showing that the visual difference between the ROC curves for the two models is statistically significant. This confirms the qualitative analysis presented above and supports the hypothesis that including even simple proxies for solar activity can improve the prediction of geomagnetic activity with lead times greater than the L1-to-Earth transit time.

Figure 12 also explores activity-dependent model performance. Using \( Pr = 0.5 \) as our decision threshold again, we calculate a range of performance metrics (described in Sect. 2.2) while varying the \( K_p \) threshold used to define an “event” (see also Liemohn et al., 2018). In each of the panels, the black markers show the results for \( \mu^{\text{OMNI}} \) and the blue markers show the results for \( \mu^{\text{OMNI}} \). The error bars show the 95% confidence interval estimated using 2000 bootstrap resamplings (Morley, 2018; Morley et al., 2018b). Panel (a) shows the threshold-dependent Heidke skill score (HSS), which measures model accuracy relative to an unskilled reference. Panel (b) shows the Matthews Correlation Coefficient, which can be interpreted in a similar way to a Pearson correlation. Panel (c) shows the probability of detection (PoD), also called hit rate or true positive rate. Panel (d) shows the false alarm ratio (FAR), which is the fraction of predicted events that did not occur. The ideal FAR is therefore zero. Panel (e) shows the critical success index (CSI), which can be interpreted as an accuracy measure after removing true negatives from consideration. Finally, panel (f) displays the frequency bias, where an unbiased forecast predicts the correct number of events and non-events and scores 1. As a reminder, the metrics displayed in panels a, b, c and e are positively-oriented, where 1 constitutes a perfect score. FAR (panel d) is negatively-oriented and a perfect model has an FAR of zero. The metrics shown in panels a–e have an upper bound of 1, and this is marked by the red dashed line. In every measure, the performance between the two models is indistinguishable at low values of \( K_p \) – which, as we recall, constitutes the vast majority of geomagnetic conditions – but as the threshold for identifying an event increases we clearly see improved performance from the \( \mu^{\text{OMNI}} \) model. While the confidence intervals substantially overlap for these scores we note that parameter estimates with overlapping confidence intervals can be significantly different (Afshartous & Preston, 2010). In other words, while non-overlapping confidence intervals are likely to show that the performance metrics are significantly different, the inverse is not necessarily true. Due to a variety of factors, we cannot assess the significance of the improvement in all performance metrics presented here. Among these are the fact that the metrics are correlated with each other, and we would need to correct for the effect of multiple significance tests. We have instead noted throughout this work where we were able to test for significance and described the consistent improvement in performance metrics. Although the improvement in these metrics is modest, we again conclude that adding the GOES X-ray flux data improves the model’s ability to predict geomagnetically active times.

Finally, we assess the reliability of the probability distributions generated by our dGPR models. In this context, reliability assesses the agreement between the predicted probability and the mean observed frequency. In other words, if the model predicts a 50% chance of exceeding the storm threshold, is that prediction correct 50% of the time?

Figure 13 presents a reliability diagram of the observed probability of a geomagnetic storm (for different \( K_p \) threshold levels, i.e. 2, 4, 6) plotted against the forecast probability of a geomagnetic storm. The top row – panels (a.1) and (a.2) – presents reliability diagram for models \( \mu^{\text{OMNI}} \) and \( \mu^{\text{OMNI}} \), respectively. In this figure red, blue and black lines represent geomagnetic storm thresholds of \( K_p = 2, 4 \) and 6 respectively. A perfectly reliable forecast should lie on the \( x = y \) line (black dashed line). For smaller chances of geomagnetic storms, both forecast models are reliable in their probabilistic predictions. As the predicted probability increases so do the tendency for the predicted probability to be higher than the observed probability. That is, the model tends to over-forecast slightly. Comparing the reliability of \( \mu^{\text{OMNI}} \) to that of \( \mu^{\text{OMNI}} \), we see similar results for activity thresholds of \( K_p = 2 \) and 4. However, the \( \mu^{\text{OMNI}} \) model predictions for a storm-time threshold of \( K_p = 6 \) are slightly more reliable than for its simpler counterpart.

The panels in the bottom row of Figure 13 are histograms showing the frequency of forecasts in each probability bin, also known as refinement distributions. These indicate the sharpness of the forecast, or the ability of the forecast to predict extremes.
in event probability. For example, a climatological mean forecast would have no sharpness and a deterministic model (i.e., a prediction with a delta function probability distribution) would be perfectly sharp, only ever predicting probabilities of zero or one. Ideally, a model would have both sharpness and reliability in its predicted probabilities. The refinement distributions presented here show that both $\mu^{\text{OMNI}}$ and $\mu^{\text{OMNI+}}$ display sharpness, with local peaks near probabilities of zero and one.

Both models exhibit high sharpness, which can be interpreted as the confidence of the model in its event prediction (Wilks, 2006). Further, both models perform similarly for lower $K_p$ thresholds. The $\mu^{\text{OMNI+}}$ model, when using an event threshold of $K_p \geq 6$, has slightly improved calibration over the $\mu^{\text{OMNI}}$. The addition of the solar X-ray flux data has consistently improved performance when assessing its performance in a deterministic setting, and here is shown to improve the calibration of the model at high activity levels without impact to the sharpness of the model. This analysis further supports the performance of our probabilistic model and confirms that the GOES X-ray data adds value to our $K_p$ prediction model.

### 4 Discussion

In this study, we presented a novel, probabilistic, geomagnetic storm forecast model that predicts 3 h ahead. Our model structure combined an LSTM classifier and dynamically-trained deep Gaussian processes to generate predictions of $K_p$ with an associated probability distribution. To test whether a simple, operationally-available data set could improve predictions of geomagnetic storm times, we trained two variants of our model: the first used only solar wind data and historical values of $K_p$; the second added X-ray fluxes from the NOAA GOES satellites, as a proxy for magnetic complexity at the Sun. Using a variety of model validation methods, we have confirmed that including the X-ray data enhances the performance of the forecast model at times of high geomagnetic activity. Due to the low number of samples (at high $K_p$ levels) for model testing, many measures of model performance suggest an improvement in the model performance at high activity levels but statistical significance could not be demonstrated. Significance tests of the improvement in the correlation coefficients and the change of the ROC AUC
show that there is a quantified, statistically-significant improvement in the model performance when GOES X-ray flux data is included. In this section, we further analyze the performance metrics and compare them with prior studies.

Although exact comparisons should not be made as we use different data sets for model validation, we place our results in the context of previous work. In comparison with some earlier models (e.g., Boberg et al., 2000; Wing et al., 2005; Bala & Reiff, 2012) our models typically perform well, with an RMSE of 0.77. The performance, as measured by RMSE, is not as good as the RMSE for the 3 h-ahead predictions of Zhelavskaya et al. (2019) (RMSE = 0.67) and Tan et al. (2018) (RMSE = 0.64). To assess the performance of their model when predicting geomagnetic storm intervals (defined as $K_p \geq 5$), Tan et al. (2018) has calculated the F1-score. This binary event metric is the harmonic mean of the precision and recall, using the nomenclature of machine learning literature. Using terminology from statistical literature, the precision is perhaps better known as the positive predictive value and represents the fraction of predicted positives that were correct. Similarly, the recall is the probability of detection and represents the fraction of observed events that were correctly predicted. The F1-score for the Tan et al. (2018) model was 0.55. Our initial model ($\mu^{\text{OMNI}}$) gave an F1-score of 0.56, while our model including the solar X-ray flux data ($\mu^{\text{OMNI}}$) gave an F1-score of 0.6.

Recent studies mainly focused on the predictive skill of the $K_p$ forecast models, whereas, in this paper, we aim to provide a probabilistic prediction of $K_p$ without compromising the predictive skill of the model. We have further demonstrated that including a simple, operationally-available proxy for the likelihood of solar activity improves the prediction of geomagnetic storms. The inability of $K_p$ prediction models to predict larger storms ($K_p \geq 5$) well from L1 solar wind data has previously been discussed in the literature (see, e.g., Zhelavskaya et al., 2019), and this work shows that including solar X-ray flux can directly improve the prediction of high levels of geomagnetic activity. In this work we found that including solar X-ray flux in our model features reduces the overall RMSE by 0.01, from 0.78 to 0.77. At the same time the correlation coefficient increased by a small but statistically significant amount (from 0.82 to 0.83). Importantly, for the storm-time test data the RMSE was reduced by 0.58, from 1.48 to 0.9, and the correlation coefficient increased from 0.69 to 0.75. For details of the results and the significance testing, see Table 4 and Section 3.1. Similarly, we note that analyzing the area under the ROC curve shows a significant improvement in the probabilistic predictions of $K_p$, for high activity levels, when X-ray flux is included (see Sect. 3.3). These comparisons show that inclusion of solar flux can enhance the storm time forecasting capability without diminishing the performance during less active periods.

Although we present only one sample model architecture, we use this to highlight a straightforward method by which uncertainty bounds can be predicted using machine-learning models, and also improve predictions of intervals of high geomagnetic activity. This clear demonstration that the X-ray flux data meaningfully improves our prediction of geomagnetic storms strongly suggests that future work including solar data sources are a promising way to extend the meaningful forecast horizon for high levels of geomagnetic activity. However, other operationally-available data sources exist that are likely to carry more information about magnetic complexity at the Sun (e.g., solar magnetograms; Arge et al., 2010), and hence using these will further improve the prediction of both the CME- and non-CME-driven geomagnetic activity. Further work is planned to investigate this work as well as incorporating other recent advances that will help improve the fidelity of our predictive model.

We also note that Zhelavskaya et al. (2019) explored methods for selecting model features and reducing a large number of candidate features to a smaller selection of those that are most important. Relative to their work we use a small number of features, all of which were selected based on a physically-motivated interpretation and subject to the constraint of being products generally available operationally. Applying their feature selection methods and further developing the architecture, such as using convolution neural network to process and extract CME features from 2-dimensional solar X-ray and magnetogram data, would likely yield immediate improvements in the model performance.

5 Conclusion

The two main objectives addressed by this work were to: 1. build a probabilistic $K_p$ forecast model; and 2. test whether the inclusion of operationally-available proxies for solar activity could improve the prediction of geomagnetic storms (using $K_p$, following the NOAA G-scale).

We presented a two-layer architecture, using an LSTM neural network to predict the likelihood of storm-time and deep Gaussian Process Regression models to predict the value of $K_p$, including uncertainty bounds. We then explored these uncertainty bounds to provide a probabilistic geomagnetic storm forecast. Our analysis demonstrates that this architecture can be used to build probabilistic space weather forecast models with good performance.

Further, we tested two variants of our model that differed only in the input parameters (“features”). The first used only upstream solar wind data and the second added solar X-ray flux data from the GOES spacecraft. Analysis of the predictions and the errors, for both the values of $K_p$ as well as the probability of exceeding a threshold in $K_p$, showed that the addition of X-ray flux data resulted in significant model performance improvements during geomagnetically active periods. The model using X-ray flux data had a significantly higher correlation coefficient on the storm-time test data (increased from 0.69 to 0.75), with other performance metrics showing improvements. The RMSE on the storm-time data set decreased from 1.48 to 0.9. This improvement in model performance was also seen across all contingency table-based metrics, with improvements in skill and reductions in false alarms. Similarly, the probabilistic predictions were shown to be significantly better by testing the difference in the area under the ROC curve. The probabilistic predictions were shown to be well-calibrated and sharp.

Adding solar X-ray flux data to empirical or machine-learned models can add useful information about transient solar activity, improving the 3-h ahead prediction of the $K_p$ index for high geomagnetic activity levels. Although including this relatively simple data set increased the accuracy of the forecast, supporting the suggestion that X-ray data is a reasonable proxy for solar magnetic activity, our model still shows lags.
in predicting large geomagnetic storms. Capturing uncertainty, providing probabilistic predictions and improving our ability to capture transient behavior are all within reach with modern tools and do not require sacrificing model predictive performance. We hope that future work continues to bring together recent advances in feature selection (e.g., Zhelavskaya et al., 2019), model design to accommodate probabilistic prediction, and more complex solar data sources such as solar magnetograms, to provide accurate forecasting of strong geomagnetic activity with longer lead times.

Acknowledgements. SC thanks to the Space Science and Applications group and the Center for Space and Earth Science (CSES) at Los Alamos National Laboratory for organizing and supporting the Los Alamos Space Weather Summer School. Portions of this work by SKM were performed under the auspices of the US Department of Energy and were partially supported by the Laboratory Directed Research and Development (LDRD) program, award number 20190262ER. The authors wish to acknowledge the use of the OMNI solar wind data, available at https://omniweb.gsfc.nasa.gov/ow.html. The $K_p$ index and GOES X-ray datasets were accessed through the GFZ-Potsdam website https://www.gfz-potsdam.de/en/kp-index/ and NOAA FTP server https://satdat.ngdc.noaa.gov/sem/goes/data/, respectively. The majority of analysis and visualization was completed with the help of free, open-source software tools, notably: Keras (Chollet, 2015); matplotlib (Hunter, 2007); IPython (Perez & Granger, 2007); pandas (McKinney, 2010); Scapyce (Morley et al., 2011); PyForecastTools (Morley, 2018); and others (see, e.g., Millman & Aivazis, 2011). The code developed during this work is available at https://github.com/shibaji7/Codebase_Kp_Prediction. The editor thanks two anonymous reviewers for their assistance in evaluating this paper.

References

Afshartous D, Preston RA. 2010. Confidence intervals for dependent data: Equating non-overlap with statistical significance. *Comput Stat Data Anal* 54(10): 2296–2305. https://doi.org/10.1016/j.csda.2010.04.011.

Al-Shedivat M, Wilson AG, Saatchi Y, Hu Z, Xing EP. 2017. Learning scalable deep kernels with recurrent structure. *J Mach Learn Res* 18(82): 1–37. URL: http://jmlr.org/papers/v18/16-498.html.

Arge CN, Henney CJ, Koller J, Compeau CR, Young S, Mackenzie D, Fay A, Harvey JW. 2010. Air force data assimilative photospheric flux transport (ADAPT) model. *Twelfth Int Sol Wind Conf* 1216: 343–346. https://doi.org/10.1063/1.3395870.

Baker DN, Hones EW, Payne JB, Feldman WC. 1981. A high time resolution study of interplanetary parameter correlations with AE. *Geophys Res Lett* 8(2): 179–182. https://doi.org/10.1029/GL008i002p00179.

Bala R, Reiff P. 2012. Improvements in short-term forecasting of geomagnetic activity. *Space Weather* 10(6): S06001. https://doi.org/10.1029/2012SW000779.

Bartels JR. 1949. The standardized index, $K_s$ and the planetary index, $K_p$. *IATME 97(12b)*: 0001.

Bingham Suzy, Murray Sophie A, Guerrero Antonio, Glover Alexi, Thorn Peter. 2019. Summary of the plenary sessions at European Space Weather Week 15: Space weather users and service providers working together now and in the future. *J Space Weather Space Clim* 9: A32. https://doi.org/10.1051/swsc/2019031.

Boberg F, Wintoft P, Lundstedt H. 2000. Real time $K_p$ predictions from solar wind data using neural networks. *Phys Chem Earth Part C Sol Terr Planet Sci* 25(4): 275–280. https://doi.org/10.1016/S1464-1917(00)00116-7.

Borovsky JE. 2013. Physical improvements to the solar wind reconnection control function for the Earth’s magnetosphere. *J Geophys Res Space Phys* 118(5): 2113–2121. https://doi.org/10.1002/jgra.50110.

Borovsky JE. 2014. Canonical correlation analysis of the combined solar wind and geomagnetic index data sets. *J Geophys Res Space Phys* 119(7): 5364–5381. https://doi.org/10.1002/2013JA019607.

Borovsky JE, Denton MH. 2006. Differences between CME-driven storms and CIR-driven storms. *J Geophys Res Space Phys* 111(A7): A07S08. https://doi.org/10.1029/2005JA011447.

Borovsky JE, Thomsen MF, Elphic RC, Cayton TE, McComas DJ. 1998. The transport of plasma sheet material from the distant tail to geosynchronous orbit. *J Geophys Res Space Phys* 103(A9): 20297–20331. https://doi.org/10.1029/97JA03144.

Bradley AP. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn* 30(7): 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.

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.

Carbary JF. 2005. A $K_p$-based model of auroral boundaries. *Space Weather* 3(10): S10001. https://doi.org/10.1029/2005SW000162.

Choi H-S, Lee J, Cho K-S, Kwak Y-S, Cho I-H, Park Y-D, Kim Y-H, Baker DN, Reeves GD, Lee D-K. 2011. Analysis of GEO spacecraft anomalies: Space weather relationships. *Space Weather* 9(6): S06001. https://doi.org/10.1029/2010SW000597.

Chollet F. 2015. Keras. https://keras.io.

Costello K.A. 1998. Moving the rice MSFM into a real-time forecast mode using solar wind driven forecast modules. *PhD Thesis*. URL: https://scholarship.rice.edu/handle/1911/19251.

DeLong ER, DeLong DM, Clarke-Pearson DL. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 44(3): 837–845. https://doi.org/10.2307/2531595.

Duney JW. 1961. Interplanetary magnetic field and the auroral zones. *Phys Rev Lett* 6(2): 47–48. https://doi.org/10.1103/PhysRevLett.6.47.

Eastwood JP, Biffis E, Hapgood MA, Green L, Bisi MM, Bentley RD, Wicks R, Micknall L-A, Gibbs M, Burnett C. 2017. The Economic impact of space weather: Where do we stand? *Risk Anal* 37(2): 206–218. https://doi.org/10.1111/risa.12765.

Estabrooks A, Jo T, Japkowicz N. 2004. A multiple resampling method using solar wind driven forecast modules. *PhD Thesis*. URL: https://doi.org/10.1023/A:1005160129098.

Fay A, Harvey JW. 2010. Air force data assimilative photospheric flux transport (ADAPT) model. *Twelfth Int Sol Wind Conf* 1216: 343–346. https://doi.org/10.1063/1.3395870.

Garcia HA. 1994. Temperature and emission measure from GOES soft X-ray measurements. *Sol Phys* 154(2): 275–308. https://doi.org/10.1007/BF00681100.

Gonzalez WD, Tsurutani BT, Chua de Gonzalez AL. 1999. Interplanetary origin of geomagnetic storms. *Space Sci Rev* 88(3): 529–562. https://doi.org/10.1023/A:100516129098.

Haiduczek JD, Welling DT, Ganushkina NY, Morley SK, Ozturk DS. 2017. SWMF global magnetosphere simulations of January 2005: Geomagnetic indices and cross-polar cap potential. *Space Weather* 15(12): 1567–1587. https://doi.org/10.1002/2017SW001695.

Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Comput* 9(8): 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
Observations. Space Weather 17(8): 1219–1229. https://doi.org/10.1029/2018SW002141.

Srivastava N, Venkatakrishnan P. 2002. Relationship between CME speed and geomagnetic storm intensity. Geophys Res Lett 29(9): 1-1–1-4. https://doi.org/10.1029/2001GL013597.

Srivastava N, Venkatakrishnan P. 2004. Solar and interplanetary sources of major geomagnetic storms during 1996–2002. J Geophys Res Space Phys 109(A10): A10103. https://doi.org/10.1029/2003JA010175.

Steiger JH. 1980. Tests for comparing elements of a correlation matrix. Psychol Bull 87(2): 245–251. https://doi.org/10.1037/0033-2909.87.2.245.

Sun X, Xu W. 2014. Fast implementation of DeLong’s algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Sig Proc Lett 21(11): 1389–1393. https://doi.org/10.1109/LSP.2014.2337313.

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.

Töth G, Sokolov IV, Gombosi TI, Chesney DR, Clauer CR, et al. 2005. Space weather modeling framework: A new tool for the space science community. J Geophys Res Space Phys 110(A12): A12226. https://doi.org/10.1029/2005JA011126.

Wilks DS. 2006. Statistical methods in the atmospheric sciences, 2nd edn, Academic Press. https://www.elsevier.com/books/statistical-methods-in-the-atmospheric-sciences/wilks/978-0-12-385022-5.

Wing S, Johnson JR, Jen J, Meng C-I, Sibeck DG, Bechtold K, Freeman J, Costello K, Balikhin M, Takahashi K. 2005. Kp forecast models. J Geophys Res Space Phys 110(A4): A04203. https://doi.org/10.1029/2004JA010500.

Wing S, Johnson JR, Camporeale E, Reeves GD. 2016. Information theoretical approach to discovering solar wind drivers of the outer radiation belt. J Geophys Res Space Phys 121(10): 9378–9399. https://doi.org/10.1002/2016JA022711.

Winter LM, Balasubramaniam KS. 2014. Estimate of solar maximum using the 1–8Å geostationary operational environmental satellites X-ray measurements. Astrophys J 793(2): L45. https://doi.org/10.1088/2041-8205/793/2/45.

Winter LM, Balasubramaniam K. 2015. Using the maximum X-ray flux ratio and X-ray background to predict solar flare class. Space Weather 13(5): 286–297. https://doi.org/10.1002/2015SW001170.

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. J Space Weather Space Clim 7: A29. http://doi.org/10.1051/SWSC/2017027.

Wu DJ, Feng HQ, Chao JK. 2008. Energy spectrum of interplanetary magnetic flux ropes and its connection with solar activity. Astron Astrophys 480(1): L9–L12. https://doi.org/10.1051/0004-6361:20079173.

Xu F, Borovsky JE. 2015. A new four-plasma categorization scheme for the solar wind. J Geophys Res Space Phys 120(1): 70–100. https://doi.org/10.1002/2014JA020412.

Yap BW, Rani KA, Rahman HAA, Fong S, Khairudin Z, Abdullah NN. 2014. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In: Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), Herawan T., Deris M.M., Abawajy J. (Eds.), Springer Singapore, Singapore, pp. 13–22.

Zhang J, Blanco-Cano X, Nitta N, Srivastava N, Mandrini CH. 2018. Editorial: Earth-affecting solar transients. Sol Phys 293: 80. https://doi.org/10.1007/s11207-018-1302-9.

Zhelavskaya IS, Vasile R, Shprits YY, 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.

Zhou G, Wang J, Cao Z. 2003. Correlation between halo coronal mass ejections and solar surface activity. A&A 397(3): 1057–1067. https://doi.org/10.1051/0004-6361:20021463.