JGR Space Physics

RESEARCH ARTICLE
10.1029/2023JA031478

Evaluating Auroral Forecasts Against Satellite Observations Under Different Levels of Geomagnetic Activity

M. K. Mooney[1,2,3], C. Forsyth[2], M. S. Marsh[3], L. Bradley[4], T. Finnigan[4], F. Forde[4], F. Garrigan[4], C. Mancini-Tuffier[4], T. Mancini-Tuffier[4], E. Roberts[4], P. Vessoni[4], J. Powell[4], S. Clark[4], C. J. Lao[2], A. Smith[2,5], D. R. Jackson[3], S. Bingham[3], M. Sharpe[3], T. Hughes[3], G. Chisham[6], and S. Milan[1]

[1]University of Leicester, Leicester, UK, [2]Mullard Space Science Laboratory, University College London, Dorking, UK, [3]Met Office, Exeter, UK, [4]St Richard Reynolds Catholic College, London, UK, [5]Department of Mathematics, Physics and Electrical Engineering, Northumbria University, Newcastle Upon Tyne, UK, [6]British Antarctic Survey, Cambridge, UK

Key Points:

- We evaluate the spatial performance of the OVATION-Prime 2013 auroral flux model under different levels of geomagnetic activity
- The model is able to predict the equatorward extent of the auroral oval with increasing geomagnetic activity
- The model does not capture the poleward auroral boundary location or the polar cap expansion with increasing geomagnetic activity

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

M. K. Mooney,
michaela.mooney@leicester.ac.uk

Citation:

Mooney, M. K., Forsyth, C., Marsh, M. S., Bradley, L., Finnigan, T., Forde, F., et al. (2024). Evaluating auroral forecasts against satellite observations under different levels of geomagnetic activity. *Journal of Geophysical Research: Space Physics*, *129*, e2023JA031478. https://doi.org/10.1029/2023JA031478

Received 14 MAR 2023
Accepted 5 JUL 2024

Author Contributions:

Conceptualization: M. K. Mooney, C. Forsyth
Formal analysis: M. K. Mooney, C. Forsyth, L. Bradley, T. Finnigan, F. Forde, F. Garrigan, E. Roberts, P. Vessoni
Funding acquisition: C. Forsyth, M. S. Marsh, D. R. Jackson
Investigation: M. K. Mooney, C. Forsyth, L. Bradley, T. Finnigan, F. Forde, F. Garrigan, E. Roberts, P. Vessoni, J. Powell
Methodology: M. K. Mooney, C. Forsyth, M. Sharpe, T. Hughes
Project administration: J. Powell, S. Clark

Abstract The aurora and associated high energy particles and currents pose a space weather hazard to communication networks and ground-based infrastructure. Forecasting the location of the auroral oval forms an integral component of daily space weather operations. We evaluate a version of the OVATION-Prime 2013 auroral forecast model that was implemented for operational use at the UK Met Office Space Weather Operations Cent. Building on our earlier studies, we evaluate the ability of the OVATION-Prime 2013 model to predict the location of the auroral oval in all latitude and local time sectors under different levels of geomagnetic activity, defined by Kp. We compare the model predictions against auroral boundaries determined from IMAGE FUV data. Our analysis shows that the model performs well at predicting the equatorward extent of the auroral oval, particularly as the equatorward auroral boundary expands to lower latitudes for increasing Kp levels. The model performance is reduced in the high latitude region near the poleward auroral boundary, particularly in the nightside sectors where the model does not accurately capture the expansion and contraction of the polar cap as the open flux content of the magnetosphere changes. For increasing levels of geomagnetic activity ($Kp \geq 3$), the performance of the model decreases, with the poleward edge of the auroral oval typically observed at lower latitudes than forecast. As such, the forecast poleward edge of the auroral oval is less reliable during more active and hazardous intervals.

Plain Language Summary Enhanced auroral activity can be hazardous to technology and essential daily services at Earth. The aurora can cause disruption to communication networks including long-range radio communications and induce currents in the ground which can impact electricity supply networks. A version of the OVATION-Prime 2013 auroral forecast model is commonly used in space weather forecast centers to predict the occurrence and location of the aurora, providing advanced warning of possible disruption to stakeholder industries in the aviation, defense and energy sectors. In this study, we perform a detailed evaluation of the performance of this model by comparing the auroral forecasts against satellite observations of the aurora from the IMAGE satellite. Our analysis shows that the model performs well at predicting the location of the main auroral emission, particularly the extent of the auroral emission to lower latitudes with increasing levels of geomagnetic activity. However, the model performance is reduced at higher latitudes and does not accurately capture the auroral dynamics in this region.

1. Introduction

The aurora is the most visible aspect of space weather. The high-energy charged particles precipitating in the upper atmosphere that cause the bright auroral emission also disrupt high frequency radio communication networks used by emergency responders and aircraft for long-range communication and tracking (Cannon et al., 2013; Greenberg & LaBelle, 2002; Harang & Stroffregen, 1940; Jones et al., 2017; Moore, 1951; Redmon et al., 2018; Schrijver et al., 2015). Meanwhile, the associated auroral current systems can cause geomagnetically-induced currents which can damage electricity supply networks and increase the rate of corrosion in oil and gas pipelines (Boteler, 2019; Cannon et al., 2013; Eastwood et al., 2017; Erinmez et al., 2002; Freeman et al., 2019; Smith et al., 2019; Viljanen et al., 2006). Modeling and forecasting the location of the auroral oval can provide stakeholder industries, such as the defense, aviation and energy sectors, with advanced warning of possible disruption to mitigate the risk of space weather impact on essential services. In addition, forecasting the

Resources: C. Forsyth, J. Powell, C. J. Lao, A. Smith
Supervision: C. Forsyth, M. S. Marsh, J. Powell, S. Clark, D. R. Jackson, S. Bingham
Visualization: M. K. Mooney, L. Bradley, T. Finnigan, F. Forde, F. Garrigan, E. Roberts, P. Vessoni
Writing – original draft: M. K. Mooney
Writing – review & editing: M. K. Mooney, C. Forsyth, M. S. Marsh, J. Powell, S. Clark, C. J. Lao, A. Smith, D. R. Jackson, M. Sharpe, G. Chisham, S. Milan

occurrence of visible aurora is important for auroral tourism and is a key tool in promoting public awareness and engagement with space weather, through projects such as Aurorasaurus (MacDonald et al., 2015) and AuroraWatch UK (Case et al., 2017).

A number of empirical models have been developed to estimate the location of the aurora. Hardy et al. (1985) presented a statistical model of auroral particle precipitation under different levels of geomagnetic activity. Similarly, Carbary (2005) developed a statistical model of the auroral boundaries with Kp, determined from averaged ultra-violet (UV) satellite observations of the auroral oval from the Polar spacecraft. McGranaghan et al. (2021) have developed an electron precipitation model using machine learning methods. A version of the OVATION-Prime 2013 (OP-2013 Newell et al., 2014) auroral forecast model is currently used for daily operations by leading space weather forecasting centers including the U.S. National Oceanic Atmospheric Administration (NOAA) Space Weather Prediction Center (SWPC), the U.S. Department of Defense Space Weather Operations Center and the UK Met Office Space Weather Operations Center (MOSWOC). We note that this is only a small subset of auroral models, there are many more which have been developed or are currently under development.

Model evaluation is important at all stages of the research-to-operations development cycle of space weather forecast models. A number of verification studies have previously evaluated the performance of the earlier generation aurora forecast model, OVATION-Prime 2010 (OP-2010) (Kosar et al., 2018; Lane et al., 2015; Machol et al., 2012; Mitchell et al., 2013; Newell, Sotirelis, Liou, et al., 2010; Newell, Sotirelis, & Wing, 2010). There is no single absolute model evaluation technique and so different studies can reveal different capabilities and limitations in the model.

Newell, Sotirelis, Liou, et al. (2010) evaluated the auroral forecast by comparing the instantaneous and hourly-averaged predicted auroral power to the observed power estimated from Polar UVI data and found that the predicted auroral power correlated with the observed auroral power with correlation coefficients (r^2) of 56% and 58%, respectively. This validation study demonstrated that just over half of the observed variance in the auroral power can be forecast by the OP-2010 model but did not assess whether that power was predicted in the correct locations. Mitchell et al. (2013) evaluated the OP-2010 model in terms of how well the model captured the variance in the nightside auroral power, compared with that derived from Polar UVI observations. The results were compared against those of a similar model, OVATION SuperMAG (OVATION-SM) which is based on the OVATION model but includes additional ground magnetometer data from SuperMAG and the times of substorm onsets which aim to capture the auroral dynamics driven by processes internal to the magnetosphere. Mitchell et al. (2013) found that OP-2010 described 47% of the variance in the Polar UVI nightside auroral power while OVATION-SM described 71% of the nightside variance. This may suggest that OVATION-SM is better at nowcasting or forecasting the aurora but the SuperMAG data is not currently available in near real time. OVATION-SM also requires advanced warning of the time until the next substorm onset which is not currently well predicted (e.g., Maimaiti et al., 2019).

These studies provide some evaluation of the OP-2010 auroral model against observations but they are limited in terms of analyzing how well the model performs as a deterministic model predicting the occurrence of aurora. Extending this verification analysis, Machol et al. (2012) evaluated the OP-2010 model in terms of its suitability as a tool for forecasting visible nightside aurora. Machol et al. (2012) used binary event analysis to compare the nightside auroral forecast against observed auroral boundaries derived from a fixed brightness threshold of the nightside auroral emission in the Polar UVI data. Machol et al. (2012) used a number of verification statistics which are commonly used in weather forecast evaluation, including the hit rate (the proportion of correct positive forecasts out of the total positive observations of aurora), the false alarm ratio (the proportion of positive aurora forecasts which were not subsequently observed) and the accuracy (the proportion of correct positive and negative forecasts over the total number of forecasts). The verification analysis by Machol et al. (2012) showed that the model performs reasonably well compared to the observed auroral boundaries with a higher proportion of correct positive forecasts (hit rate = 0.58) than the proportion of false positive forecasts (false alarm ratio = 0.14). The overall accuracy of 0.86, with a maximum score of 1, is also high, however the accuracy statistic can be dominated by the number of correct negative forecasts in low latitude regions (i.e., forecasting that aurora will not occur and it does not occur in regions where the aurora almost never occurs).

Similarly to Machol et al. (2012), Mooney et al. (2021) used binary event analysis to perform a detailed evaluation of the performance of the operational version of the OP-2013 model implemented by MOSWOC. They compared the model output against observed auroral boundaries determined from the far-ultraviolet (FUV) auroral images from the IMAGE satellite between 2000 and 2002. The analysis utilized relative operating characteristic (ROC) curves and ROC scores to examine the global performance of the model as a deterministic forecast of the location of the aurora under different geomagnetic conditions and in a limited number of local time sectors. Mooney et al. (2021) also used reliability curves and Brier skill scores to examine the validity of the probabilities of aurora occurring forecast by the model. Overall, they found that the operational model performed well at predicting the location of the auroral oval with a ROC score of 0.82, however the performance was reduced in the dayside local time sectors (ROC score = 0.59) and during periods of higher geomagnetic activity (ROC score = 0.55 for Kp = 8). However, global ROC scores can hide regional variation in the model performance. As a probabilistic forecast, the operational model tends to under-predict the probability of aurora occurring.

In this study, we present a more detailed evaluation of auroral forecasts from the version of OP-2013 that was being used operationally at the Met Office specifically examining the spatial performance of the model. We compare the auroral forecasts from the model against auroral boundaries derived by Chisham et al. (2022a) from global FUV images of the auroral oval obtained by the IMAGE satellite. We extend the analysis of Mooney et al. (2021) by using maps of ROC scores to evaluate the deterministic model performance in each grid cell with geomagnetic activity, extending the coverage of the validation to all latitudes and all local time sectors.

Our results show that the model performs well at predicting the equatorward extent of the auroral emission and accurately captures the widening of the auroral oval for mid to high levels of geomagnetic activity (Kp ≥ 3). However, the model performance is reduced at higher latitudes, particularly in the nightside sectors where the model does not capture the location of the poleward auroral boundary or the expansion of the polar cap as the open flux content increases for mid to high levels of geomagnetic activity (Kp ≥ 3).

2. Data

2.1. Auroral Forecast Data

The research in this study uses a version of the OP-2013 auroral forecast model that was used operationally in daily space weather forecasts at the UK MOSWOC until December 2020. A detailed description of the OP-2013 model and the Met Office implementation is provided in Mooney et al. (2021). Here we provide a short summary.

The OVATION-Prime and the upgraded OVATION-Prime 2013 auroral forecast models (Newell et al., 2014; Newell, Sotirelis, Liou, et al., 2010) predict the precipitating electron and proton auroral flux based on upstream solar wind conditions measured at L1. The model is based on a linear scaling between the electron and proton flux from the DMSP data with an empirically derived solar wind coupling function (Newell et al., 2007, 2009). In each model grid point, the particle flux is calculated as a function of season, the type of aurora (mono-energetic, broadband and diffuse electron aurora and ion aurora) and the linear scaling to the solar wind input (Newell et al., 2009). In the later OP-2013 version of the model, additional auroral data from the Global Ultraviolet Imager (GUVI) instrument onboard the TIMED satellite was included to improve the model performance at higher values of Kp, between Kp 5 and 8, compared to the performance of the predecessor version OP-2010 model (Newell et al., 2014). Additional upgrades to the model included further noise reduction and a smoother data interpolation in the post-midnight magnetic local time (MLT) sectors (Newell et al., 2014). The magnetic latitude (MLAT) range of the OP-2013 data spans 50°–89.5° and covers 24-hr of MLT, with a grid resolution of 0.25 MLT by 0.5° MLAT. We direct the interested reader to Newell et al. (2007, 2009), Newell, Sotirelis, and Wing (2010), and Newell et al. (2014) for full details of the OP-2010 and OP-2013 models.

The Met Office operational implementation of OP-2013 assessed in this study assumes a fixed 30-min propagation time for the solar wind measured at the L1 point to arrive at Earth. The fixed 30-min propagation time corresponds to a fast solar wind speed of approximately 800 km/s. Assuming a fast solar wind speed is common in space weather forecasting to provide a prediction of the space weather conditions at the earliest time of arrival of the solar wind measured at L1 to reach Earth. The aurora does not respond instantaneously to the upstream solar wind conditions and so to introduce some time history to the input conditions the operational implementation of the model requires 4-hr of solar wind data input to produce each forecast. A weighting is applied to the solar wind data with the most recent data having the heaviest weighting of 1. The weighting decreases linearly in increments of 0.25 such that the earliest data from 4-hr previous has the smallest weighting of 0.25. The OP-2013 auroral model outputs the estimated combined precipitating particle flux from all types of aurora at each grid point for the given upstream solar wind conditions. In this operational version, the flux in each grid point is linearly scaled into an estimated probability of aurora occurring. The linear conversion applied in each grid point is $probability = 14.4 \times flux + 18$, where the flux is the combined precipitating particle flux from all four types of aurora in each grid cell, with a maximum probability of 100%. Using this conversion from auroral flux to probability, we can estimate that a probability of ∼20% corresponds to approximately 0.5 erg cm^{-2} s^{-1}, ∼30% corresponds to approximately 1 erg cm^{-2} s^{-1} and ∼40% corresponds to approximately 1.5 erg cm^{-2} s^{-1}. The forecast probability output by the model in each grid point is interpreted as the probability of an observer seeing the visible aurora. The operational implementation provides an auroral forecast for both the northern and southern hemispheres 30 min ahead of the current time.

In this study, we use hindcasts of the output from the 30-min nowcast version of OP-2013 used at the Met Office using historic solar wind data for the period between May 2000 to October 2002 (Marsh & Mooney, 2021), not auroral forecasts that were issued in near real time by the Met Office.

2.2. Observational Auroral Boundary Data

In this study, we use auroral luminosity boundaries determined from global observations of the far-ultraviolet (FUV) auroral emission obtained by the NASA Imager for Magnetopause-to-Aurora Global Exploration (IMAGE) satellite as a ground truth observational data set to compare with the model forecast probability maps output from OP-2013.

The IMAGE satellite was operational between 2000 and 2005 in a precessing polar orbit with a perigee of 1,000 km and an apogee of 44,000 km (∼7 R_E) (Mende et al., 2000a, 2000b). IMAGE was spin-stabilized, rotating about its axis once every 2 min. Over the first 3 years of nominal operations between 2000 and 2002, the orbital apogee was situated over the northern hemisphere. The IMAGE satellite carried a far-ultraviolet (FUV) wideband imaging camera (WIC), sensitive to emission between 140 and 190 nm (Mende et al., 2000a, 2000b). Images were taken by the WIC instrument approximately every 2 min determined by the spin period of the satellite (Burch, 2000).

Using the IMAGE FUV data, Longden et al. (2010) identified the poleward and equatorward auroral luminosity boundaries in each MLT sector between May 2000 and October 2002 inclusive. The auroral boundaries were identified by fitting two functions to the latitudinal intensity profile in each MLT sector. One function is a single Gaussian function with a quadratic background and the other is a double Gaussian function with a quadratic background. Both functions are then quantitatively assessed using the reduced χ^2 goodness-of-fit statistic to determine which form provided a better fit in each sector. The poleward and equatorward auroral luminosity boundaries (PALBs and EALBs, respectively) are then identified from the best fitting function. In sectors where the single Gaussian function provides the better fit, the PALB/EALB is estimated as being offset poleward/equatorward of the center of the Gaussian peak by the full width half maximum (FWHM). Where the double Gaussian function is the better fit the PALB/EALB is estimated as being offset poleward/equatorward from the poleward/equatorward peak by the FWHM of that peak.

Recently, Chisham et al. (2022a) have produced an updated subset of the auroral boundaries identified by Longden et al. (2010) which removes potentially unreliable boundary identifications such as erroneous boundaries identified due to instrumental effects for example, bright pixels near the edge of the instrument field of view or boundaries identified during periods when there was an error in mapping the IMAGE FUV data to the ionosphere. In this study, we have used the updated Chisham et al. (2022a) subset of auroral boundary identifications. The full description of the method can be found in Longden et al. (2010) and the details of the corrections made to the data set in Chisham et al. (2022a).

2.3. Data Selection and Verification Method

In this study, we use the OP-2013 auroral forecasts spanning the period of May 2000 to October 2002 (Marsh & Mooney, 2021), coinciding with the available observational auroral boundary data from Chisham et al. (2022a). The observational auroral boundary data is only available for the northern hemisphere aurora and as such we limit our analysis to only evaluate the northern hemisphere auroral forecasts. To select independent forecast and observation pairs, we follow the same data selection process as Mooney et al. (2021). Each forecast requires 4-hr of input solar wind data, thus we down-sampled our forecast data set by selecting forecasts to have a 4-hr

Table 1
Table Detailing the Number of Forecast and Observation Pairs for Each Kp Level

Kp level	Number of forecast and observation pairs
1	496
2	502
3	342
4	158
5	53
6	32
7	14
8	3

separation between the previous forecast and the next forecast that is, a 4-hr resolution to ensure that the forecasts were independent of one another. We then match the observational ground truth auroral boundaries identified from a single IMAGE snapshot that were closest in time and within ±2.5 min of the 4-hr separated forecast time. This results in 2622 corresponding forecast and observation pairs.

We are particularly interested in evaluating the performance of OP-2013 under different levels of geomagnetic activity, defined by the Kp index, as higher geomagnetic activity poses a greater space weather risk to technology and daily services. We subdivide the data set of 2,622 independent forecast and observation pairs by Kp level. As Kp is a 3-hourly index, we select forecast and observation pairs which are at least 1 hour into the 3-hr window, to allow the magnetosphere and the aurora to respond to the level of geomagnetic activity. Table 1 shows the number of forecast and observation pairs for each Kp level. We note that due to using the updated Chisham et al. (2022a) subset of the original Longden et al. (2010) observed auroral boundaries, the 2262 forecast and observation pairs used in this study are a subset of the 3360 forecast and observation pairs used in the Kp analysis of Mooney et al. (2021). As the level of geomagnetic activity increases, the number of events in each Kp level decreases, particularly for the high Kp levels of 5 and above. There are only 3 forecast and observation pairs in the Kp = 8 category which limits any statistical conclusions that we can draw regarding the model's performance during very high levels of geomagnetic activity. Due to the limited number of forecast and observation pairs in the high Kp categories we combine all the forecast and observation pairs for Kp of 5 and above to evaluate the forecasts during high geomagnetic activity. Kp = 5 is the threshold for a G1 geomagnetic storm as defined by NOAA and MOSWOC and so the high activity category of Kp = 5–8 combines the Kp levels that are considered as a storm. We don't consider Kp = 9 events because the model is only valid up to Kp = 8 (Newell et al., 2014).

In this study, binary event analysis was applied to evaluate the forecasts in each grid point in the MLT × MLAT grid, for each forecast probability threshold between 0% and 100%, in 10% increments. In each grid point, for each 10% forecast probability threshold we evaluate whether aurora was predicted to occur and whether it was observed and add this value to the truth table. By comparing the forecasts and observations in each grid cell, we do not need to define boundaries from the forecast data based on a fixed value of flux or probability. If the predicted auroral probability in the grid cell is greater than the probability threshold then the aurora was forecast to occur and we compare whether the forecast aurora occurred within or outside of the observed auroral boundaries to obtain the truth table. We repeat this for each 10% probability threshold. If the forecast probability of aurora occurring is equal to or greater than the probability threshold and aurora was also observed, it counts as a *hit* in our truth table. If the forecast probability of aurora occurring is equal to or greater than the probability threshold but aurora was not observed, it counts as a *false alarm*. If the forecast probability of aurora occurring is less than the probability threshold and aurora was observed, it counts as a *miss*. If the forecast probability of aurora occurring is less than the probability threshold but aurora was not observed, it counts as a *correct negative*. We repeat this for each forecast probability threshold, producing a truth table for each grid point, for each forecast probability threshold. These truth tables can be used to produce maps of each component of the truth table (i.e., hits, false alarms, misses, correct negatives) for each forecast probability threshold. Using the series of truth tables in each grid point, we calculate the hit rate (hits/(hits + misses)) and false alarm rate (false alarms/(false alarms + correct negative forecasts)) for each forecast probability threshold to produce a ROC curve (Mason, 1982; Swets, 1988; Swets et al., 1955) and obtain a ROC score for each grid point by calculating the proportional area under the ROC curve. ROC scores have a value between 0 and 1, with higher ROC scores between 0.5 and 1 indicating that the model is skillful at discriminating between aurora and no-aurora. A lower ROC score of less than 0.5 indicates that the model is not skillful and tends to predict the opposite to what is observed for example, predicts the aurora but the aurora does not occur. ROC scores are used to test model discrimination between binary yes/no forecasts that is, to test how well the model distinguishes between auroral occurrence or no occurrence. This allows us to test how well the OP-2013 model performs at predicting the location of the aurora. This analysis is repeated for forecast and observation pairs during different levels of Kp, providing detailed maps of the model performance in MLAT and local time, under different levels of geomagnetic activity. The ROC score maps allow us to quantitatively compare the model performance in different regions for example, higher latitudes versus lower latitudes, dawn versus dusk sectors as well as evaluating the performance of the model under different Kp levels. This analysis technique is the same as used in Mooney et al. (2021) but here applied to individual model grid cells rather than the forecasts as a whole. A full, detailed discussion of truth tables and ROC curves can be found in Wilks (2006) and in Mooney et al. (2021) and their Supplementary Information.

ROC curves have two fixed end points at (0, 0) and (1, 1). However, in some circumstances this leads to a slight quirk in our results. A very low probability threshold of 0% means that the aurora is always predicted to occur such that in the truth table, all the entries are either hits or false alarms, there are no missed or correct negative forecasts. This results in a hit rate = 1 and a false alarm rate = 1, creating a fixed point at (1, 1) on the ROC curve. Similarly, for a forecast threshold of 100%, no forecasts predicted by the model exceed 100%. This results in all the truth table entries being split across the missed or correct negative entries and the hit rate and false alarm rate tend toward zero, creating a second fixed point at (0, 0) on the ROC curve. The quirk in our results appears in regions where the aurora is rarely predicted and rarely observed to occur such as in the polar cap, in the lower latitude regions, below ~55° or in some dayside sectors during periods of low geomagnetic activity for example, Kp = 1. In these cases, the ROC curve tends toward a diagonal line between the fixed points at (0, 0) and (1, 1), with a ROC score of 0.5. A ROC score of 0.5 indicates that the model has no skill in these regions in discriminating between aurora and no aurora. However, the model is actually successful in consistently and correctly predicting a low to zero probability of aurora occurring in these regions.

2.4. The Suitability of ROC Scores

Many useful verification metrics can be derived from truth tables, which can be used to evaluate and quantitatively summarize the performance of a forecast model. The use of verification scores is becoming increasingly common across space physics, for example, to evaluate the performance of a space weather nowcasting/forecasting model and testing the performance of machine learning models (Camporeale, 2019; Liemohn et al., 2018, 2021; Morley, 2020; Morley et al., 2018). Three verification statistics which are commonly used in space physics include accuracy, ROC scores and precision-recall scores (Forsyth et al., 2020; Kalb et al., 2023; Machol et al., 2012; Mooney et al., 2021; Murray et al., 2017; Sharpe & Murray, 2017; Smith et al., 2020, 2021). In this section, we briefly compare these verification metrics applied to the evaluation of the auroral forecast model.

Similar to ROC curves, precision-recall (PR) curves are obtained by plotting the model precision against the hit rate (also referred to as recall). The precision is the fraction of positive forecasts that are hits (hits/(hits + false alarms)). Good models have both high precision and high recall such that the knee of the PR curve tends toward the top right hand corner of the plot, maximizing the area under the curve (AUC). Similarly to the ROC score, the proportional area under the PR curve (PR AUC) is often used as a summary value to quantify the model performance, with higher PR AUC scores indicating better model performance. PR AUC scores range between 0 and 1, with 1 representing perfect precision and recall. PR curves can be preferable over ROC curves when the data set is imbalanced that is, there are far fewer positive occurrences compared to negative occurrences. In such imbalanced data sets, ROC curves can be overly influenced by the number of correct rejections, resulting in a misleadingly inflated ROC score. PR curves can be normalized to account for the unbalanced data set (Boyd et al., 2012).

Accuracy is calculated directly from the truth table results. Accuracy is the proportion of correct forecasts (both hits and correct negatives) out of all the data in the truth table ((hits + correct negatives)/(hits + false alarms + misses + correct negatives)). Accuracy scores range between 0 and 1 with a maximum score of 1 indicating no false alarm or missed forecasts. Although the accuracy is calculated using all the coefficients in the truth table, it can be dominated by the number of correct rejections and hides the number of false alarms and missed forecasts. For example, a forecast with a 98% accuracy could have a hit rate of 50% and a precision of 50%.

To demonstrate the differences in these metrics for the evaluation of the OP-2013 model, Figures 1a–1c shows maps of the ROC scores, PR AUC scores for all forecast probability thresholds and accuracy for the 15% probability threshold calculated for each grid point for all of the independent forecast and observation pairs in the 2.5 year data set, for all geomagnetic conditions. In each panel, blue indicates a score >0.5, indicating regions where the model performance is better than a random chance model would predict the occurrence of the aurora. A ROC score of 1 is a perfect forecast and so the closer the ROC scores are to 1, the better the model is at correctly predicting whether the aurora will occur or not in that region. Red indicates a score <0.5, indicating poor model performance that is, worse than a random chance model and white indicates a score equal to 0.5. Gray cells

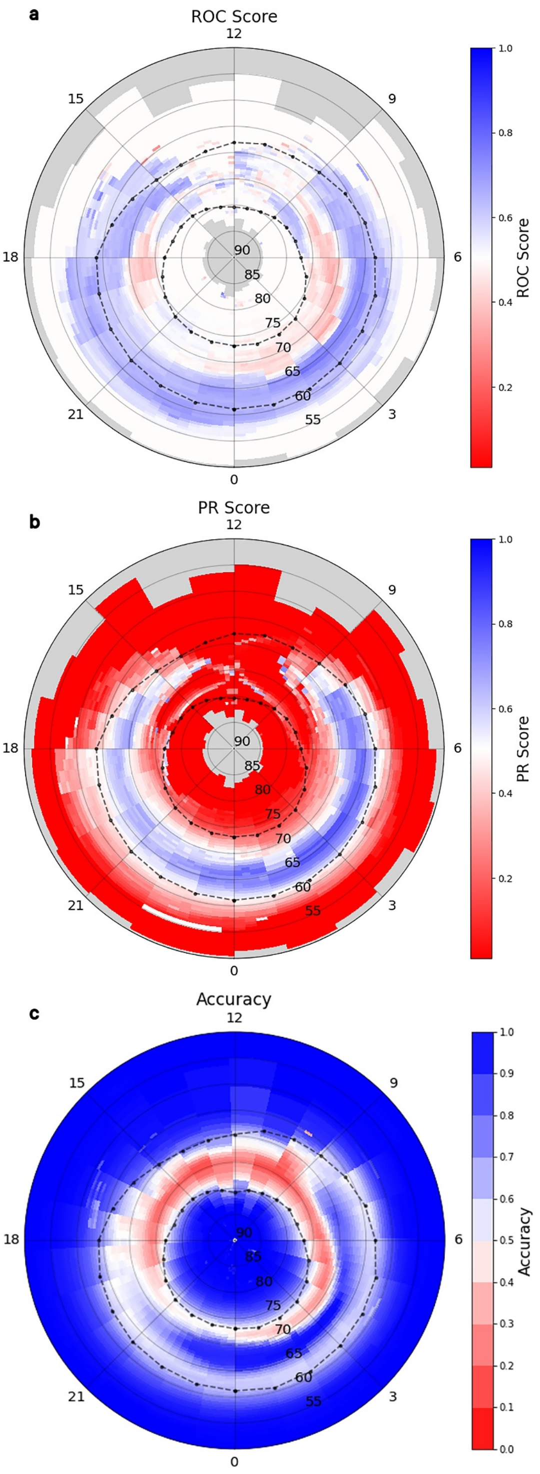

Figure 1. Panels (a)–(c) show maps of the relative operating characteristic (ROC) scores, normalized precision-recall area under the curve scores for all probability thresholds and accuracy for the 15% probability threshold, respectively, calculated for each grid point for each of the independent forecast and observation pairs in the 2.5 year data set. In each panel, blue grid cells show that the verification score is greater than 0.5, indicating that the model is skillful in these regions. Red grid cells show that the verification scores are less than 0.5, indicating poor model skill. White grid cells indicate a verification score equal to 0.5. The black dotted lines indicate the mean location of the observed poleward and equatorward auroral boundaries. Gray grid cells indicate regions where the ROC or PR scores are undefined.

indicate regions where the ROC or PR scores cannot be calculated because either the hit rate/recall, false alarm rate or precision is undefined. The black dotted lines indicate the mean location of the observed poleward and equatorward auroral boundaries over the 2.5 year period. The ROC and PR AUC scores are both calculated from the AUC which is plotted using the results from all forecast probability thresholds. In contrast, the accuracy is calculated from the truth table directly and as such, can only be presented for a single probability threshold. Mooney et al. (2021) found that forecast probabilities between 5% and 15% had the highest Peirce Score (Peirce, 1884), the largest difference between the hit rate and the false alarm rate indicating that the OP-2013 model performs best at discriminating between regions of aurora and no aurora, for forecast probabilities of 5%–15%. We have therefore chosen to focus on the results for the 15% probability threshold where appropriate in this study.

At a high level, the results for each of the three validation metrics in Figure 1 show similar results. The highest scores for each metric indicating good model performance occur from approximately the middle of the auroral oval and extend toward the equatorward auroral boundary on the nightside. The scores are reduced at higher latitudes in the nightside sectors from approximately the middle of the auroral oval to the average poleward auroral boundary and these lower scores extend round to the dawn and dusk sectors. The PR and accuracy scores are also reduced in the dayside sectors. The most notable differences in the three scores are in the polar cap and low latitudes regions. The ROC scores in this region tend toward 0.5, suggesting that the model has no skill in this region. The PR-AUC scores indicate very poor model performance in these regions and the accuracy scores indicate perfect model performance. The polar cap and low latitude regions are dominated by correct negative forecasts where generally the aurora is neither predicted nor observed to occur and that there is a low occurrence of aurora in these regions. The high proportion of correct negatives in these regions has a very different impact on each of the three scores. We note that at lower latitudes near the average equatorward auroral boundary, the high ROC scores extend to slightly lower latitudes compared to the latitudinal extent of the high PR-AUC and accuracy scores. This suggests that the ROC scores may be slightly over-inflated in this region due to the higher number of correct negative forecasts.

In summary, there are many different verification metrics that can be determined from the truth table data and used to evaluate model performance. In this study, we focus on ROC scores as we are interested in the balance between hits and false alarm forecasts although we are aware that ROC scores might be slightly inflated in regions where there are a high number of correct negative forecasts. Accuracy is not considered to be a particularly informative verification score because it is heavily influenced by the correct negative forecasts. In this study it is also less useful as the accuracy score can only be calculated for a single probability threshold, unlike for the ROC and PR AUC scores which account for the model performance under all probability thresholds. PR scores are used particularly for unbalanced or skewed data sets, for example, where there are far more positive than negative forecasts. The forecasts of the auroral occurrence are skewed differently in different regions. For example, in the polar cap and at very low latitudes aurora is rarely predicted to occur whereas in certain parts of the main auroral oval, the aurora will be predicted to occur more frequently than not. It may be useful to perform a similar analysis using PR scores to evaluate the performance of OP-2013 for different levels of geomagnetic activity for comparison, however this is outside the scope of this work.

3. Results

In the following section, we present the results of our evaluation of the OP-2013 model for different levels of geomagnetic activity. We present the detailed ROC score maps for Kp = 1–4 and the combined results for Kp = 5–8 alongside maps of the truth table statistics (hits, false alarms, misses and correct negatives) for the 15% forecast probability threshold. Similar figures Kp = 5, 6 and combined results for Kp = 7 and 8 are included in Supporting Information S1. We have also included maps of the truth table statistics (hits, false alarms, misses and correct negatives) for various probability thresholds between 0% and 100% to show how the results vary as the probability threshold is increased.

3.1. Low Geomagnetic Activity, Kp = 1–2

Figures 2a and 2f show the ROC score map for Kp = 1 and Kp = 2 events respectively with the color scale indicating the ROC score for each grid point. Blue indicates a ROC score greater than 0.5, indicating regions where the model has high discrimination and tends to correctly forecast whether the aurora will occur or not in

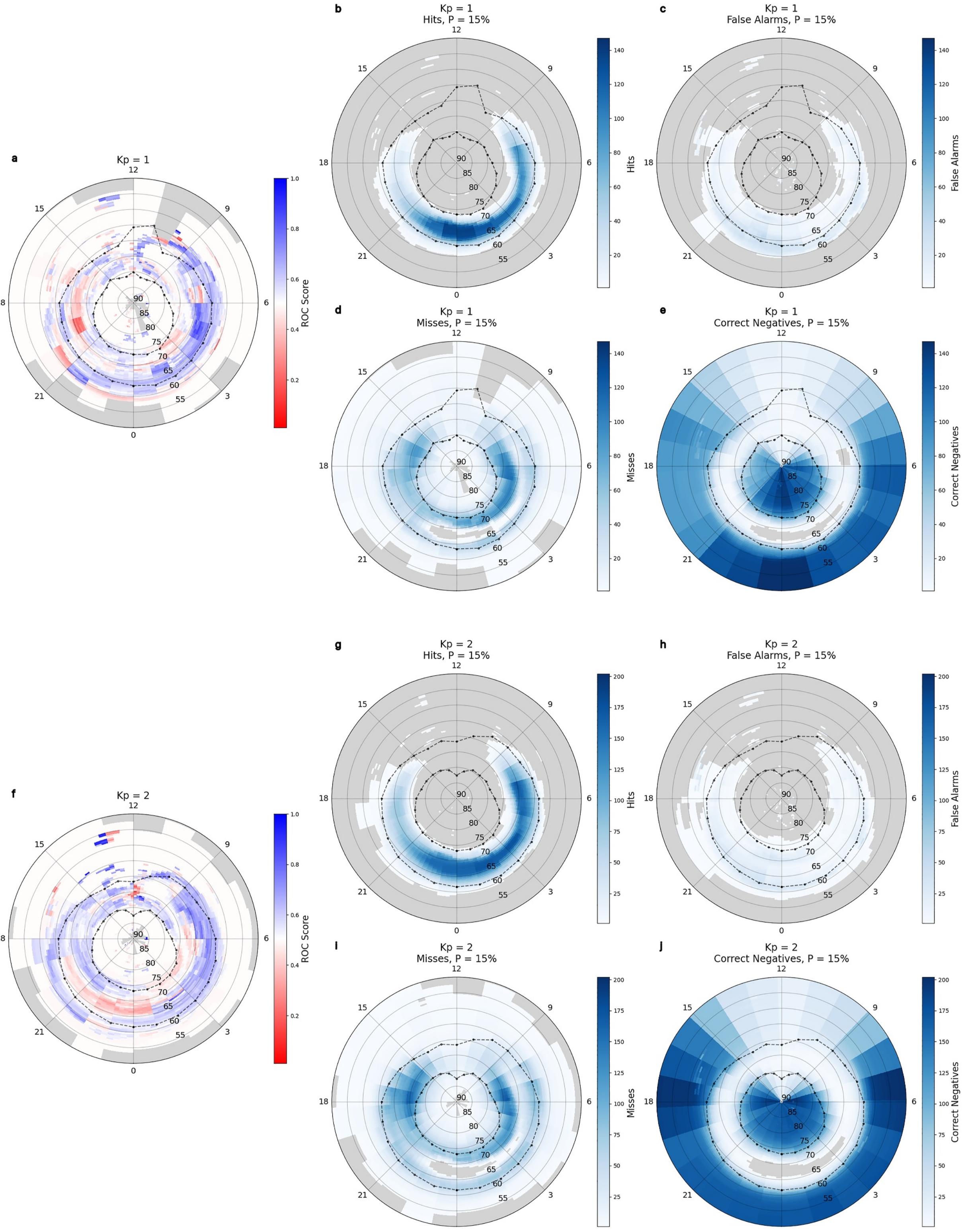
a
Kp = 1
ROC Score
b
Kp = 1
Hits, P = 15%
Hits
c
Kp = 1
False Alarms, P = 15%
False Alarms
d
Kp = 1
Misses, P = 15%
Misses
e
Kp = 1
Correct Negatives, P = 15%
Correct Negatives
f
Kp = 2
ROC Score
g
Kp = 2
Hits, P = 15%
Hits
h
Kp = 2
False Alarms, P = 15%
False Alarms
i
Kp = 2
Misses, P = 15%
Misses
j
Kp = 2
Correct Negatives, P = 15%
Correct Negatives
12
15
9
18
6
21
3
0
90
85
80
75
70
65
60
55

Figure 2.

these regions. Red indicates a ROC score of less than 0.5 where the model tends to forecast the opposite to what is observed and white indicates a ROC score equal to 0.5. The black dotted lines indicate the mean location of the observed poleward and equatorward auroral boundaries during Kp = 1 and Kp = 2 events. Panels b–e and g–j in Figure 2, show the maps of hits, false alarms, misses and correct rejections, respectively, for a forecast probability of 15%. Gray grid cells indicate that there are no hits, false alarms, misses or correct rejections in that grid cell. Mooney et al. (2021) found that the OP-2013 forecast model had an overall ROC score of 0.83 across all latitudes, local time sectors and all levels of geomagnetic activity. Evaluating the model performance by local time (defined as 3 hr of MLT centered on 12, 06, 18, and midnight), the model performance varied from 0.59 in the noon sectors to 0.86 in the midnight sectors. For low to mid levels of geomagnetic activity (Kp = 1–4) the overall ROC scores were around 0.82. The overall ROC scores decreased for higher levels of Kp, reducing to 0.55 for Kp = 8 events. In this analysis we evaluate the model performance and calculate the ROC scores of individual grid cells. The ROC scores for individual grid cells are typically lower than the overall ROC score calculated by Mooney et al. (2021) however we can use the ROC scores results of Mooney et al. (2021) as a benchmark to define approximate ROC scores that would indicate good versus poor model performance. In this analysis, we consider ROC scores above the overall ROC score from Mooney et al. (2021) ≥ 0.8 to indicate very good performance. ROC scores between ∼0.6–0.8 to be reasonably good performance and ROC scores of ∼0.5–0.6 to be adequate performance, that is, marginally better than a random chance model. ROC scores of 0.5 indicate that the model has the same level of skill as a random chance prediction model therefore in this analysis we consider ROC scores of ≤0.5 (indicated in white and red) as poor model performance. We focus our discussion on the nightside sectors between 18 and 06 MLT.

First we consider the results for the Kp = 1 events. Within the average observed auroral boundaries, there is a slight dawn-dusk asymmetry in the model performance. In Figure 2a grid cells in the post-midnight to dawn local time sectors, (01 and 09 MLT) between latitudes of 65°–70° show reasonably good ROC scores typically above 0.6. The higher ROC scores in this region correspond to the high number of hits for a forecast probability threshold of 15% observed in Figure 2b and indicate that typically the model is correctly predicting the occurrence of the aurora in this region. On the dusk side grid cells between 16 and 22 MLT in latitudes between 65° and 70° show low ROC scores of 0.5 or less. Figure 2b shows a reasonably high number of hits in this region for the 15% probability threshold however, there are also quite a high number of missed forecasts here which will reduce the model hit rate. These results are only for a single probability threshold. As the forecast probability threshold was increased (see Supporting Information S1), the number of missed forecasts in this region increased and the number of hits decreased, reducing the ROC score for these grid cells. The lower number of hits in this region as the probability threshold increases indicates that the model does not predict high probabilities of aurora occurring for solar wind conditions which result in low Kp = 1 geomagnetic conditions. Grid cells between approximately 70°–75° near the mean observed poleward auroral boundary show a high number of missed forecasts and cells between 60° and 65° near the mean observed equatorward auroral boundary show a higher number of missed and false alarm forecasts. Reduced model performance in these regions near the auroral boundaries are to be expected as a slight latitudinal offset between the extent of the forecast and observed aurora would result in an incorrect prediction. In the very high latitude regions in the polar cap (above 75° on the nightside and above 80° on the dayside) and at very low latitudes of less than 60° on the nightside and 65°–70° on the dayside, the ROC scores tend toward 0.5. In these regions, the results are dominated by the high number of correct rejections (panel d) where the aurora is generally not predicted to occur and is rarely observed. As such, the model performs well at not predicting the occurrence of the aurora in this region but has low skill since the aurora is unlikely to occur.

The results for Kp = 2 shown in Figure 2 panels f–j show a similar result to the Kp = 1 results. The dawn-dusk asymmetry in the model performance is more visible. MLT sectors between 19 and 01 MLT and 65°–70° within the mean observed auroral boundaries show a band of low (<0.5) ROC scores. Panel g shows a reasonably high number of hits in this region, however the number of hits are distinctly lower than similar latitudes in the dawn

Figure 2. Maps of model performance for Kp = 1 and Kp = 2 events. Panels (a) and (f) show the relative operating characteristic (ROC) score map for Kp = 1 and Kp = 2 events respectively. Blue grid cells show that the ROC score is greater than 0.5. Red grid cells show that the ROC scores are less than 0.5, indicating that the model is not skillful in these regions. White grid cells indicate a ROC score equal to 0.5. Gray grid cells indicate regions where the ROC scores cannot be calculated because either the hit rate or false alarm rate is undefined. The black dotted lines indicate the mean location of the observed poleward and equatorward auroral boundaries during Kp = 1 and Kp = 2 events. Panels (b–e) and (g–j) show maps of the hits, false alarms, misses and correct negatives, respectively, for a forecast probability threshold of 15%.

sectors. Similarly to Kp = 1, as the forecast probability threshold was increased (see Supporting Information S1), the number of missed forecasts in this region increased and the number of hits decreased, reducing the ROC score for these grid cells. In the post-midnight to dawn sectors in 02–07 MLT, there is a band of low ROC scores between approximately 70°–75° which are dominated by missed forecasts. As with Kp = 1 events, reduced model performance in regions near the auroral boundaries due to slight latitudinal offsets between the extent of the forecast and observed aurora.

3.2. Mid-Level Geomagnetic Activity, Kp = 3–4

Figure 3 is in the same format as Figure 2 for Kp = 3 and Kp = 4 events. For the Kp = 3 events in panels a–e, the model performs well at lower latitudes between approximately 58°–65° with ROC scores (>0.5) at this latitude across all nightside MLT sectors (18–06 MLT) with higher ROC scores of ≥0.6 in the dusk sectors. However, at higher latitudes between approximately 65°–75° the model is performing poorly with ROC scores typically ≤0.5. As with the low Kp events, panel c shows a high number of missed forecasts near the poleward boundary. As the forecast probability threshold was increased (see Supporting Information S1) the number of missed forecasts begins to dominate over the number of hits between approximately 65°–75°.

For each increasing level of Kp, both the forecast and observed aurora cover a wider latitudinal extent. This is particularly evident when comparing the observed auroral boundaries for Kp = 4 and Kp = 1. The observed auroral boundaries are located at about 58° and 73° in the midnight sector for Kp = 4 events compared to 64° and 74° in the midnight sector during Kp = 1 events.

The results of Kp = 4 appear to show a slightly improved model performance across the nightside sectors and latitudes. However, we note from Table 1 that the number of Kp = 4 events is less than half the number of Kp = 3 events. Figure 3g shows a wide band of hits between approximately 58° and 76° in the nightside sectors where the model correctly predicted that the aurora would occur. Within the observed auroral boundaries, the ROC scores are higher than for the Kp = 3 events. Generally, the observed equatorward auroral boundary matches reasonably well with the equatorward edge of the high ROC scores. There remain some slightly lower ROC scores of less than 0.5 near the observed poleward auroral boundaries, particularly in the post-midnight to dawn MLT sectors (02–09 MLT). We also note low ROC scores at the equatorward boundary between 55° and 60° in afternoon to premidnight sectors (16–00 MLT). The lower ROC scores observed near the observed auroral boundary correspond to a higher number of misses in these regions, as observed in Figure 3i, consistent with a slight discrepancy between the forecast equatorward extent of the aurora and the observed equatorward auroral boundary.

On the dayside, there is a region of low ROC scores predominantly in the post-noon sectors (13–16 MLT) within the observed auroral boundaries corresponding to a region of slightly higher missed forecasts in Figure 3i.

The results for Kp = 4 may suggest that the linear correlation between the auroral precipitation and the upstream solar wind parameters provides a reasonably good approximation for the location of the auroral oval for this Kp level, particularly in terms of predicting the equatorward latitudinal extent of the aurora and the location of the aurora on the nightside. However we caution that the increased performance of the model for Kp = 4 compared to Kp = 3 may be due to the comparatively lower number of events for Kp = 4. Overall, the results are similar for both the Kp = 3 and Kp = 4 events with the model performing better at lower latitudes within the average observed auroral boundaries and a reduced performance between the middle of the auroral oval to the poleward boundary.

3.3. High Geomagnetic Activity, Kp = 5–8

Figure 4 is in the same format as Figures 2 and 3 for the combined Kp = 5–8 events. Both the predicted and observed auroral ovals are expanded to lower latitudes and cover a wider latitudinal extent. Panel (a) shows that the low ROC scores of less than 0.5 cover a wider portion of the auroral oval from approximately 60°–70° in nightside sectors between 21 and 03 MLT. The low ROC scores extend round to 10 and 14 MLT in the dawn and afternoon regions respectively. The observed poleward auroral boundary is expanded equatorward to latitudes between 70° and 75° in the nightside sectors between 18 and 06 MLT. At high latitudes above ~70° panel c shows a high proportion of false alarm forecasts where the aurora is being predicted but not observed indicating the observed poleward auroral boundary is typically more expanded and located at lower latitudes than predicted by the model. In contrast, at lower latitudes ~55° near the equatorward auroral boundary panel d shows a higher

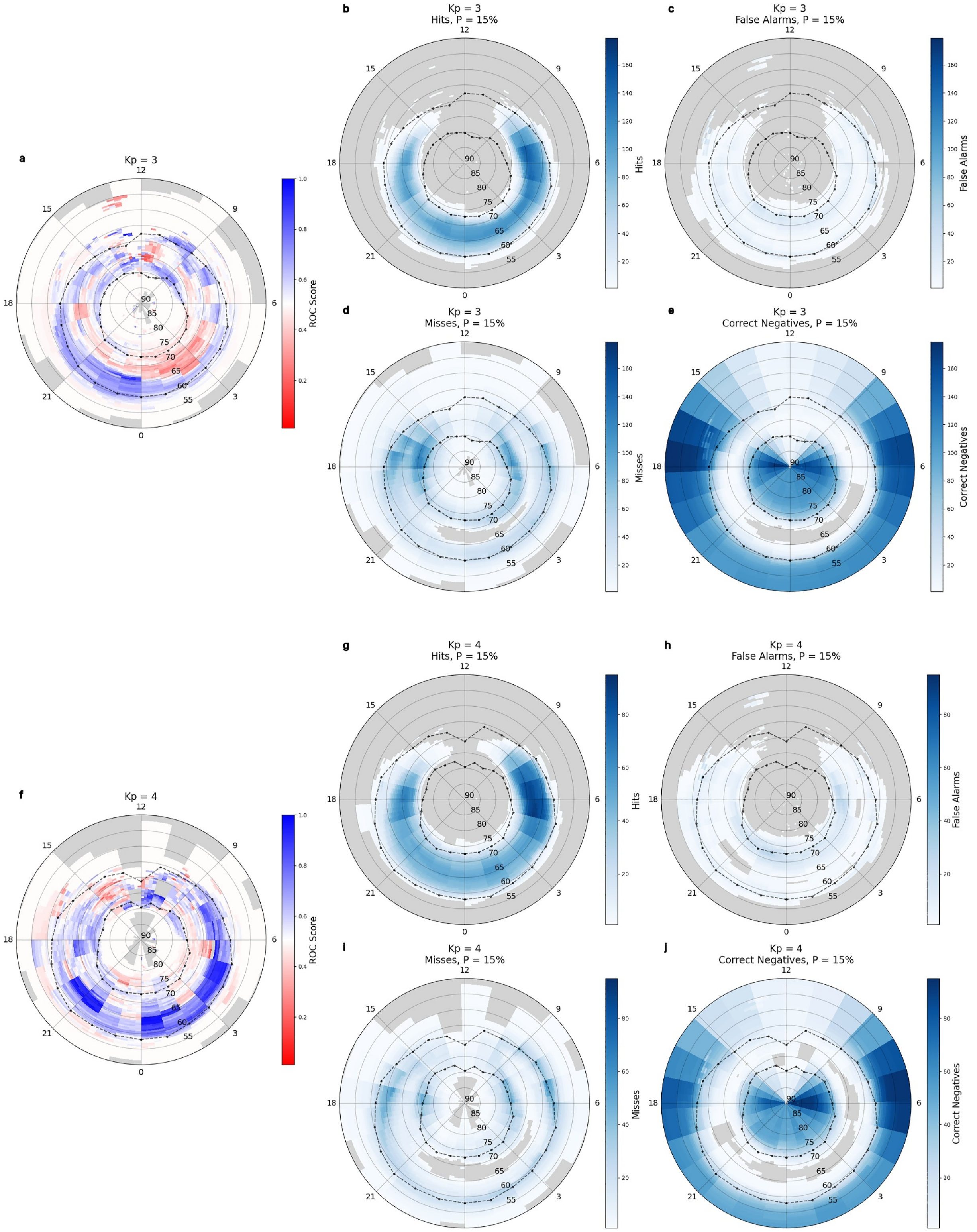
a
Kp = 3
b
Kp = 3
Hits, P = 15%
c
Kp = 3
False Alarms, P = 15%
d
Kp = 3
Misses, P = 15%
e
Kp = 3
Correct Negatives, P = 15%
f
Kp = 4
g
Kp = 4
Hits, P = 15%
h
Kp = 4
False Alarms, P = 15%
i
Kp = 4
Misses, P = 15%
j
Kp = 4
Correct Negatives, P = 15%
12
15
9
18
6
21
3
0
90
85
80
75
70
65
60
55
ROC Score
1.0
0.8
0.6
0.4
0.2
Hits
False Alarms
Misses
Correct Negatives
160
140
120
100
80
60
40
20

Figure 3.

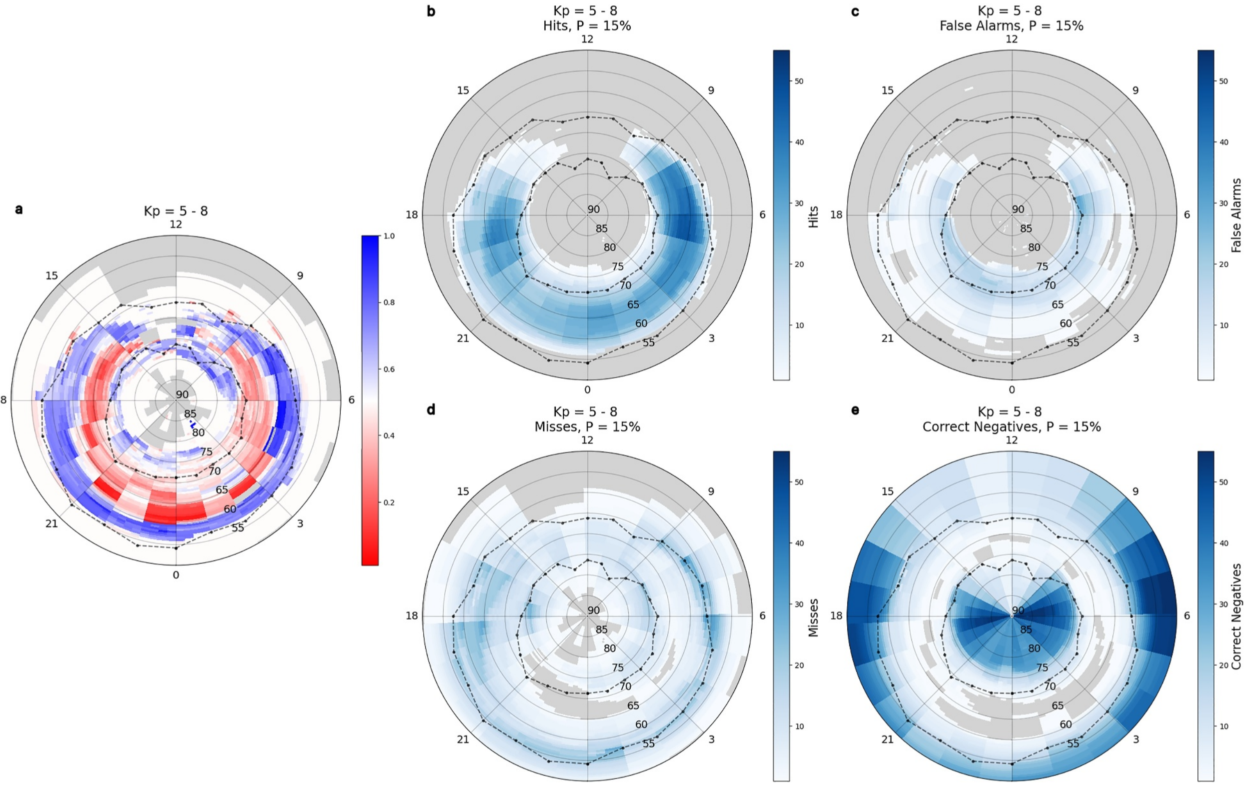

Figure 4. Panel (a) shows the relative operating characteristic (ROC) score map for the combined Kp = 5–8 events. Blue grid cells show that the ROC score is greater than 0.5. Red grid cells show that the ROC scores are less than 0.5, indicating that the model is not skillful in these regions. White grid cells indicate a ROC score equal to 0.5. Gray grid cells indicate regions where the ROC scores cannot be calculated because either the hit rate or false alarm rate is undefined. The black dotted lines indicate the mean location of the observed poleward and equatorward auroral boundaries during Kp = 5–8 events. Panels (b–e) show maps of the hits, false alarms, misses, and correct negatives, respectively, for a forecast probability threshold of 15%.

proportion of missed forecasts. These missed forecasts are likely due to small discrepancies between the predicted and observed equatorward auroral boundaries. The ROC scores in this region remain reasonably high (>0.6) indicating that generally the model is still performing well in this region and captures the equatorward extent of the equatorward auroral boundary to lower latitudes for higher levels of geomagnetic activity.

For Kp = 5–8 we note that within the main auroral oval there is a wide band of low ROC scores (<0.5). From the maps of the truth table results included in the Supporting Information S1, the cells with low ROC scores are dominated by a high number of hits for probability thresholds of ~75% and below and dominated by missed forecasts above this threshold. This indicates that the aurora is typically always observed to occur in these regions for Kp levels between 5 and 8. However, in this region there are also a very small number (typically less than 5) false alarm events but importantly there are essentially no correct negatives. This results in the false alarm rate tending toward a value of 1 for most probability thresholds and thus the false alarm rate tends to exceed the hit rate for most probability thresholds. As a result, the ROC scores in these regions are significantly low. We note that this result is likely due to the limited number of events in the Kp = 5–8 category which results in a highly unbalanced truth tables. ROC analysis is not particularly suited to highly unbalanced data sets and so the model is likely performing better in this region than the ROC scores suggest. With a higher number of events, the truth

Figure 3. Maps of model performance for Kp = 3 and Kp = 4 events. Panels (a) and (f) show the relative operating characteristic (ROC) score map for Kp = 3 and Kp = 4 events respectively. Blue grid cells show that the ROC score is greater than 0.5. Red grid cells show that the ROC scores are less than 0.5, indicating that the model is not skillful in these regions. White grid cells indicate a ROC score equal to 0.5. Gray grid cells indicate regions where the ROC scores cannot be calculated because either the hit rate or false alarm rate is undefined. The black dotted lines indicate the mean location of the observed poleward and equatorward auroral boundaries during Kp = 3 and Kp = 4 events. Panels (b–e) and (g–j) show maps of the hits, false alarms, misses and correct negatives, respectively, for a forecast probability threshold of 15%.

table would likely be more balanced and the resulting ROC scores would be more representative of the model performance.

4. Discussion

4.1. Summary of the Model Performance

In this analysis, we have used maps of ROC scores to evaluate the performance of the OP-2013 auroral forecast model. The ROC score maps allow us to quantitatively compare the model performance in different latitudes and local times as well as for different levels of geomagnetic activity.

For increasing levels of geomagnetic activity, the auroral oval widens covering a larger latitudinal extent and the equatorward auroral boundary expands to lower latitudes. Both of these features were well captured by the model. The higher ROC scores near the equatorward boundary could suggest that the location of the equatorward auroral boundary is well organized by the upstream solar wind driving of the magnetosphere and it has been found to correspond well with the low latitude boundary of solar wind-driven ionospheric plasma convection (Greenwald et al., 2002). However, we note that the ROC scores in this region may be slightly inflated due to the higher number of correct negative forecasts, reducing the false alarm rate.

For all levels of geomagnetic activity, the model showed a reduced performance at higher latitudes from the middle of the auroral oval to the poleward boundary in the nightside MLT sectors. Reduced ROC scores near the observed auroral boundaries are to be expected as a slight offset between the forecast and the observed auroral oval would result in an incorrect prediction, reducing the ROC score. However, the low ROC scores are observed from the middle of the auroral oval to the poleward boundary. For lower levels of geomagnetic activity (Kp = 1–3), the reduced performance at higher latitudes was dominated by missed forecasts. The proportion of missed forecasts increased for higher probability levels which indicates that the model does not predict the occurrence of the aurora at higher latitudes with a high probability for low Kp values but the aurora does occur. For higher levels of geomagnetic activity of Kp = 5–8, there are a higher proportion of false alarm forecasts near the poleward auroral boundary which indicates that the model incorrectly predicts that the aurora will occur at higher latitudes but it is not observed. The high number of false alarm forecasts at higher latitudes for high Kp levels indicates that the model does not capture the expanding polar cap to lower latitudes. The poleward auroral boundary is particularly dynamic and responds to changes in flux content due to the upstream solar wind driving and internal processes such as substorms. During substorms the poleward auroral boundary also deviates from its typical oval shape particularly near substorm onset (Mooney et al., 2020). This operational version of OP-2013 is not able to capture substorm occurrence and associated auroral dynamics as it is entirely solar wind driven and has no input data to identify substorm onsets. However, the expansion of the polar cap and poleward auroral boundary due to the increase in the magnetospheric open flux content is largely driven by the solar wind-magnetospheric coupling. The fact that the OP-2013 auroral forecast model doesn't capture this expansion is unexpected and may be due to a limitation in the underlying model.

4.2. Model Coefficients

In this section, we examine the underlying model coefficients in detail to understand the reduced performance observed at higher latitudes near the poleward edge of the auroral emission.

The OP-2013 auroral forecast model predicts the auroral flux in each grid cell as a function of magnetic local time, MLAT, season, the type of aurora (diffuse, monoenergetic, wave, and ion aurora) and the solar wind-magnetospheric coupling (Newell et al., 2009; Newell, Sotirelis, & Wing, 2010). A linear function is then defined in each grid point which scales the predicted flux by the gradient of the linear function (termed the flux multiplier here) and the *y*-axis intercept of the linear function (termed the flux offset here). Figure 5 shows the flux offset and multiplier coefficients for spring in panels a and b for the predecessor version of the model, OP-2010, which form the linear scaling to predict the auroral particle precipitation flux. Spring is defined as 90 days centered on the spring equinox (Newell, Sotirelis, & Wing, 2010). We have used the OP-2010 version of the model because it is freely available online. The exact values of the coefficients may differ for the updated OP-2013 version but we expect the sign of the coefficients in each grid cell to be broadly similar across the two versions. The coefficients are shown for the diffuse aurora which dominate the auroral particle flux (Newell et al., 2009). We note that Figure 5 is only intended to demonstrate how the underlying coupling between the level

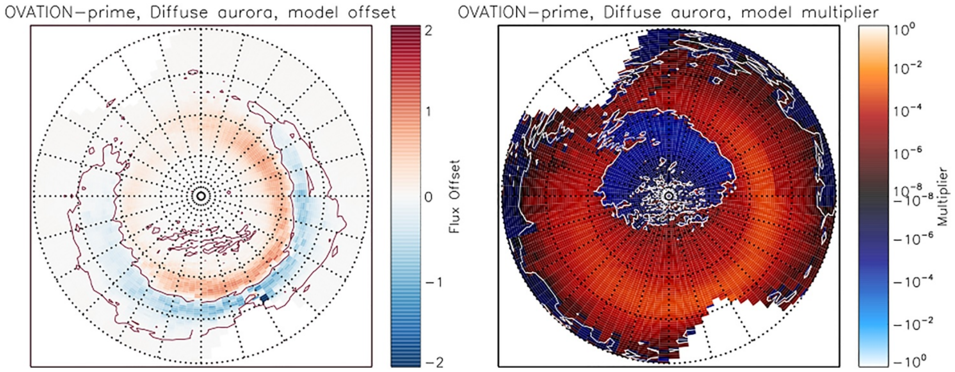

Figure 5. Maps showing the OP-2010 model coefficients of the linear scaling of the upstream solar wind-magnetospheric driving to the predicted auroral flux. Panel (a) shows the flux offset (linearly fitted intercept) and panel (b) shows the $\log_{10}$ of the multiplier (linearly fitted gradient) for the diffuse auroral component of the model for the spring season. On both maps, reds indicate positive values of the offset or multiplier, blues indicate negative values of the offset or multiplier and the contour indicates the 0 level.

of solar wind driving and the predicted auroral flux is implemented by the model. Therefore we don't include maps of the model coefficients for each type of aurora or for each season.

The scale in Figure 5b is a logarithmic scale with -10^0 to -10^{-8} shown in red and $+10^0$ to $+10^{-8}$ shown in blue. Panel (b) shows that the flux multiplier is positive between approximately 52° and 80° on the nightside and approximately 52° and 70° on the dayside. A positive flux multiplier indicates that some amount of auroral flux is always predicted in this region, even for low levels of solar wind driving. The negative flux multiplier values in the polar cap and at low latitudes indicate where the auroral flux is not predicted by the model. Panel (a) shows the flux offset in each grid point. Between ~65° and 70° in the nightside sectors and between ~70° and 75° in the dayside sectors, the flux offset is positive indicating that the aurora would be predicted in this region, even for low levels of solar wind driving. At lower equatorward latitudes, there is a band of negative flux offset values which would result in auroral flux only being predicted in this region by the model when the level of solar wind-magnetospheric driving exceeds the negative flux offset.

From the model coefficients shown in Figure 5, we can see that during high levels of solar wind driving (expected to correspond to high Kp levels), the auroral flux predicted by the model will be high in all regions where the flux multiplying coefficient is positive including the high latitude nightside region. However, in reality, the polar cap and poleward auroral boundaries expand equatorward to lower latitudes during higher levels of solar wind-magnetospheric coupling as the magnetospheric open flux content increases (Cowley & Lockwood, 1992). Auroral emission is not typically observed in this region. The expanding polar cap is not accurately captured by the model resulting in auroral emission being predicted in the high latitude nightside region more often than it is observed hence the high number of false alarm forecasts.

In contrast, during periods of lower solar wind driving corresponding to low Kp levels, the poleward auroral boundary is contracted to higher latitudes where the flux offset is positive but smaller (~<1). If the solar wind driving is low the product of the solar wind driving and the flux multiplier is small and so the auroral flux will be predicted with a low to zero probability by the model, resulting in a higher number of missed forecasts. This effect is larger at higher probability thresholds because the model does not predict high probabilities of aurora occurring for low solar wind driving/low Kp levels reducing the ROC scores in this region.

4.3. Operational Implementation

We note that in this analysis we used the OP-2013 auroral forecast model as it was set up for operational use in the Met Office. This operational version of the model assumes a fixed 30-min (~800 km/s) solar wind propagation time from the point of observation at L1 to the Earth. Future generations of solar wind driven auroral forecast models might consider the choice of solar wind speed for example, using the real time solar wind speed or producing the auroral forecast for a range of solar wind speeds (i.e., three distinct solar wind speeds to produce an ensemble forecast) to assess whether this impacts the model performance, however as the aurora is not directly driven by the external solar wind, the choice of solar wind speed may not make a significant difference.

Finally we note that the OP-2013 auroral model calculates southern hemisphere fluxes by using the seasonally determined model coefficients time-shifted by half a year. However, asymmetries in latitudinal and longitudinal displacements as well as differences in auroral intensities have been observed between the two hemispheres (Hu et al., 2017; Laundal & Østgaard, 2009; Østgaard et al., 2004, 2005) that may not be accounted for by this shift. The asymmetries are thought to occur mainly due to the upstream IMF orientation as well as asymmetries in the inter hemispheric field-aligned currents. Substantial improvements in modeling asymmetries between the two auroral ovals requires a better understanding of exactly how and when auroral asymmetries occur which requires consistent simultaneous auroral observations in both hemispheres. Future auroral models, and similar model evaluations, should aim to move away from the assumption of symmetrical auroral ovals and try to capture asymmetries between the two hemispheres.

5. Conclusions

In this study, we have performed a detailed evaluation of a version of the OP-2013 auroral forecast model to assess its performance in MLAT and local time under different levels of geomagnetic activity. The auroral hindcasts were compared against the observed auroral boundaries determined from IMAGE FUV WIC global observations of the auroral oval. Overall the model captures the expansion of the equatorward auroral boundary to lower latitudes and the widening latitudinal extent of the auroral oval under mid to high levels of geomagnetic activity ($Kp \geq 3$). This result could suggest that the equatorward auroral boundary is well organized by the upstream solar wind driving of the magnetosphere. However, the model performance is significantly reduced between the middle of the auroral oval to the poleward boundary, particularly for Kp = 3–8. The model does not capture the expansion of the poleward auroral boundary to lower latitudes as the open magnetospheric flux content increases for higher levels of geomagnetic activity. The poleward auroral boundary is particularly dynamic and responds to changes in the magnetospheric open flux content which is driven by both the upstream solar wind conditions and internal magnetospheric such as substorms. This version of the OP-2013 forecast model is entirely solar wind driven and has no input data to identify the substorm onsets and as such it is unable to capture substorm occurrence and associated auroral dynamics. The occurrence of substorm activity is not currently understood well enough to develop a predictive forecast model of substorm onsets. The reduction in model performance is also partly due to the underlying coefficients within the model which linearly scale the upstream solar wind-magnetospheric coupling to predicted auroral flux in each grid cell.

The results of this verification analysis can be used to help space weather forecasters to understand the limitations of the model, how well the model agrees with the extent of the observed auroral boundaries on average during different levels of geomagnetic activity, and where the model generally performs well or poorly. Future generations of auroral forecast models should aim to better capture the dynamics of the poleward auroral boundary and internal magnetospheric processes including substorms.

Data Availability Statement

The auroral hindcast dataset was provided by the Met Office and is publicly available at: http://doi.org/10.5281/zenodo.4653288 (Marsh & Mooney, 2021). The model output was produced using Advanced Composition Explorer (ACE) solar wind data, provided by the National Oceanic Atmospheric Administration (NOAA). The Ovation Prime 2013 code was provided to the Met Office by Rodney Viereck of the NOAA Space Weather Prediction Center and the original Ovation Prime code was developed by Patrick Newell and colleagues at Johns Hopkins University. The IMAGE auroral boundary data for the northern hemisphere (2000–2002) are freely available at https://doi.org/10.5285/fa592594-93e0-4ee1-8268-b031ce21c3ca (Chisham et al., 2022b). The IMAGE FUV data were provided courtesy of the instrument PI Stephen Mende (University of California, Berkeley). We thank the PI, the IMAGE mission, and the IMAGE FUV team for data usage and processing tools. IMAGE FUV data is archived at: https://cdaweb.gsfc.nasa.gov/pub/data/image/fuv/. Kp index data was provided by ISGI GFZ Potsdam. Archive Kp index data is available at: https://spaceweather.gfz-potsdam.de/products-data/nowcasts/nowcast-kp-index/downloads. The OVATION-Prime 2010 version of the model is freely available at https://sourceforge.net/projects/ovation-prime/.

Acknowledgments

MKM was supported by a UK NERC NPIF studentship 1926376. MKM and SEM were supported by the Science and Technology Facilities Council (STFC), UK, Grant ST/W00089X/1. GC was supported by the British Antarctic Survey Polar Science for Planet Earth Programme, funded by the UK Natural Environment Research Council (NERC). This work was carried out in partnership with the Met Office.

References

Boteler, D. H. (2019). A 21st century view of the March 1989 magnetic storm. *Space Weather*, *17*(10), 1427–1441. https://doi.org/10.1029/2019SW002278

Boyd, K., Santos Costa, V., Davis, J., & Page, C. D. (2012). Unachievable region in precision-recall space and its effect on empirical evaluation. In *Proceedings of the international conference on machine learning, 2012* (p. 349).

Burch, J. L. (2000). IMAGE mission overview. *Space Science Reviews*, *91*, 1–14. https://doi.org/10.1007/978-94-011-4233-5_1

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

Cannon, P., Angling, M., Barclay, L., Curry, C., Dyer, C., Edwards, R., et al. (2013). *Extreme space weather: Impacts on engineered systems and infrastructure (technical report)*. Royal Academy of Engineering.

Carbary, J. F. (2005). A Kp-based model of auroral boundaries. *Space Weather*, *3*(10), S10001. https://doi.org/10.1029/2005SW000162

Case, N. A., Marple, S. R., Honary, F., Wild, J. A., Billett, D. D., & Grocott, A. (2017). AuroraWatch UK: An automated aurora alert system. *Earth and Space Science*, *4*(12), 746–754. https://doi.org/10.1002/2017EA000328

Chisham, G., Burrell, A. G., Thomas, E. G., & Chen, Y. J. (2022a). Ionospheric boundaries derived from auroral images. *Journal of Geophysical Research (Space Physics)*, *127*(7), e30622. https://doi.org/10.1029/2022JA030622

Chisham, G., Burrell, A. G., Thomas, E. G., & Chen, Y. J. (2022b). Ionospheric boundaries derived from IMAGE satellite mission data (May 2000–October 2002)—VERSION 2.0 [Dataset]. *Journal of Geophysical Research: Space Physics*, *127*(7), e2022JA030622. https://doi.org/10.5285/fa592594-93e0-4ee1-8268-b031ce21c3ca

Cowley, S. W. H., & Lockwood, M. (1992). Excitation and decay of solar wind-driven flows in the magnetosphere-ionosphere system. *Annales Geophysicae*, *10*(1–2), 103–115.

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

Erinmez, I. A., Kappenman, J. G., & Radasky, W. A. (2002). Management of the geomagnetically induced current risks on the national grid company's electric power transmission system. *Journal of Atmospheric and Solar-Terrestrial Physics*, *64*(5–6), 743–756. https://doi.org/10.1016/S1364-6826(02)00036-6

Forsyth, C., Watt, C. E. J., Mooney, M. K., Rae, I. J., Walton, S. D., Marsh, M., & Albert, J. (2020). Forecasting GOES 15 > 2MeV electron fluxes from solar wind data and geomagnetic indices. *Space Weather*, *18*(8), e2019SW002416. https://doi.org/10.1029/2019SW002416

Freeman, M. P., Forsyth, C., & Rae, I. J. (2019). The influence of substorms on extreme rates of change of the surface horizontal magnetic field in the United Kingdom. *Space Weather*, *17*(6), 827–844. https://doi.org/10.1029/2018SW002148

Greenberg, E. M., & LaBelle, J. (2002). Measurement and modeling of auroral absorption of HF radio waves using a single receiver. *Radio Science*, *37*(2), 1022. https://doi.org/10.1029/2000RS002550

Greenwald, R. A., Shepherd, S. G., Sotirelis, T. S., Ruohoniemi, J. M., & Barnes, R. J. (2002). Dawn and dusk sector comparisons of small-scale irregularities, convection, and particle precipitation in the high-latitude ionosphere. *Journal of Geophysical Research (Space Physics)*, *107*(A9), 1241. https://doi.org/10.1029/2001JA000158

Harang, L., & Stroffregen, W. (1940). Echoversuche auf Ultrakurzwellen. *Hochfreq Elektroakust*, *55*, 105–108.

Hardy, D. A., Gussenhoven, M. S., & Holeman, E. (1985). A statistical model of auroral electron precipitation. *Journal of Geophysics Research*, *90*(A5), 4229–4248. https://doi.org/10.1029/JA090iA05p04229

Hu, Z.-J., Yang, Q.-J., Liang, J.-M., Hu, H.-Q., Zhang, B.-C., & Yang, H.-G. (2017). Variation and modeling of ultraviolet auroral oval boundaries associated with interplanetary and geomagnetic parameters. *Space Weather*, *15*(4), 606–622. https://doi.org/10.1002/2016SW001530

Jones, J., Sanders, S., Davis, B., Hedrick, C., Mitchell, E. J., & Cox, J. M. (2017). Research to operations transition of an auroral specification and forecast model. In *Advanced Maui optical and space surveillance (AMOS) technologies conference* (p. 94).

Kalb, V., Kosar, B., Collado-Vega, Y., & Davidson, C. (2023). Aurora detection from nighttime lights for Earth and space science applications. *Earth and Space Science*, *10*(1), e2022EA002513. https://doi.org/10.1029/2022EA002513

Kosar, B. C., MacDonald, E. A., Case, N. A., Zhang, Y., Mitchell, E. J., & Viereck, R. (2018). A case study comparing citizen science aurora data with global auroral boundaries derived from satellite imagery and empirical models. *Journal of Atmospheric and Solar-Terrestrial Physics*, *177*, 274–282. https://doi.org/10.1016/j.jastp.2018.05.006

Lane, C., Acebal, A., & Zheng, Y. (2015). Assessing predictive ability of three auroral precipitation models using DMSP energy flux. *Space Weather*, *13*(1), 61–71. https://doi.org/10.1002/2014SW001085

Laundal, K. M., & Østgaard, N. (2009). Asymmetric auroral intensities in the Earth's Northern and Southern hemispheres. *Nature*, *460*(7254), 491–493. https://doi.org/10.1038/nature08154

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

Liemohn, M. W., Shane, A. D., Azari, A. R., Petersen, A. K., Swiger, B. M., & Mukhopadhyay, A. (2021). RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics. *Journal of Atmospheric and Solar-Terrestrial Physics*, *218*, 105624. https://doi.org/10.1016/j.jastp.2021.105624

Longden, N., Chisham, G., Freeman, M. P., Abel, G. A., & Sotirelis, T. (2010). Estimating the location of the open-closed magnetic field line boundary from auroral images. *Annales Geophysicae*, *28*(9), 1659–1678. https://doi.org/10.5194/angeo-28-1659-2010

MacDonald, E. A., Case, N. A., Clayton, J. H., Hall, M. K., Heavner, M., Lalone, N., et al. (2015). Aurorasaurus: A citizen science platform for viewing and reporting the aurora. *Space Weather*, *13*(9), 548–559. https://doi.org/10.1002/2015SW001214

Machol, J. L., Green, J. C., Redmon, R. J., Viereck, R. A., & Newell, P. T. (2012). Evaluation of OVATION Prime as a forecast model for visible aurorae. *Space Weather*, *10*(3), S03005. https://doi.org/10.1029/2011SW000746

Maimaiti, M., Kunduri, B., Ruohoniemi, J. M., Baker, J. B. H., & House, L. L. (2019). A deep learning-based approach to forecast the onset of magnetic substorms. *Space Weather*, *17*(11), 1534–1552. https://doi.org/10.1029/2019SW002251

Marsh, M. S., & Mooney, M. K. (2021). OVATION-PRIME-2013 Met Office nowcast verification dataset [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.4653288

Mason, I. (1982). A model for assessment of weather forecasts. *Australian Meteorological Magazine*, *30*, 291–303.

McGranaghan, R. M., Ziegler, J., Bloch, T., Hatch, S., Camporeale, E., Lynch, K., et al. (2021). Toward a next generation particle precipitation model: Mesoscale prediction through machine learning (a case study and framework for progress). *Space Weather*, *19*(6), e02684. https://doi.org/10.1029/2020SW002684

Mende, S. B., Heetderks, H., Frey, H. U., Lampton, M., Geller, S. P., Abiad, R., et al. (2000a). Far ultraviolet imaging from the IMAGE spacecraft. 2. Wideband FUV imaging. *Space Science Reviews*, *91*(1/2), 271–285. https://doi.org/10.1023/A:1005227915363

Mende, S. B., Heetderks, H., Frey, H. U., Lampton, M., Geller, S. P., Habraken, S., et al. (2000b). Far ultraviolet imaging from the IMAGE spacecraft. 1. System design. *Space Science Reviews*, *91*(1/2), 243–270. https://doi.org/10.1023/A:1005271728567

Mitchell, E. J., Newell, P. T., Gjerloev, J. W., & Liou, K. (2013). OVATION-SM: A model of auroral precipitation based on SuperMAG generalized auroral electrojet and substorm onset times. *Journal of Geophysical Research (Space Physics)*, *118*(6), 3747–3759. https://doi.org/10.1002/jgra.50343

Mooney, M. K., Forsyth, C., Rae, I. J., Chisham, G., Marsh, M. S., Jackson, D. R., et al. (2020). Examining local time variations the gains and losses of open magnetic flux during substorms. *Journal of Geophysics Research: Space Physics*, *125*(4), e2019JA027369. https://doi.org/10.1029/2019JA027369

Mooney, M. K., Marsh, M. S., Forsyth, C., Sharpe, M., Hughes, T., Bingham, S., et al. (2021). Evaluating auroral forecasts against satellite observations. *Space Weather*, *19*(8), e02688. https://doi.org/10.1029/2020SW002688

Moore, R. K. (1951). A VHF propagation phenomenon associated with aurora. *Journal of Geophysics Research*, *56*(1), 97–106. https://doi.org/10.1029/jz056i001p00097

Morley, S. K. (2020). Challenges and opportunities in magnetospheric space weather prediction. *Space Weather*, *18*(3), e02108. https://doi.org/10.1029/2018SW002108

Morley, S. K., Welling, D. T., & Woodroffe, J. R. (2018). Perturbed input ensemble modeling with the space weather modeling framework. *Space Weather*, *16*(9), 1330–1347. https://doi.org/10.1029/2018SW002000

Murray, S. A., Bingham, S., Sharpe, M., & Jackson, D. R. (2017). Flare forecasting at the Met Office space weather operations centre. *Space Weather*, *15*(4), 577–588. https://doi.org/10.1002/2016SW001579

Newell, P. T., Liou, K., Zhang, Y., Sotirelis, T., Paxton, L. J., & Mitchell, E. J. (2014). OVATION Prime-2013: Extension of auroral precipitation model to higher disturbance levels. *Space Weather*, *12*(6), 368–379. https://doi.org/10.1002/2014SW001056

Newell, P. T., Sotirelis, T., Liou, K., Lee, A. R., Wing, S., Green, J., & Redmon, R. (2010). Predictive ability of four auroral precipitation models as evaluated using Polar UVI global images. *Space Weather*, *8*(12), S12004. https://doi.org/10.1029/2010SW000604

Newell, P. T., Sotirelis, T., Liou, K., Meng, C. I., & Rich, F. J. (2007). A nearly universal solar wind-magnetosphere coupling function inferred from 10 magnetospheric state variables. *Journal of Geophysics Research (Space Physics)*, *112*(A1), A01206. https://doi.org/10.1029/2006JA012015

Newell, P. T., Sotirelis, T., & Wing, S. (2009). Diffuse, monoenergetic, and broadband aurora: The global precipitation budget. *Journal of Geophysics Research (Space Physics)*, *114*(A9), A09207. https://doi.org/10.1029/2009JA014326

Newell, P. T., Sotirelis, T., & Wing, S. (2010). Seasonal variations in diffuse, monoenergetic, and broadband aurora. *Journal of Geophysics Research (Space Physics)*, *115*(A3), A03216. https://doi.org/10.1029/2009JA014805

Østgaard, N., Mende, S. B., Frey, H. U., Immel, T. J., Frank, L. A., Sigwarth, J. B., & Stubbs, T. J. (2004). Interplanetary magnetic field control of the location of substorm onset and auroral features in the conjugate hemispheres. *Journal of Geophysical Research (Space Physics)*, *109*(A7), A07204. https://doi.org/10.1029/2003JA010370

Østgaard, N., Tsyganenko, N. A., Mende, S. B., Frey, H. U., Immel, T. J., Fillingim, M., et al. (2005). Observations and model predictions of substorm auroral asymmetries in the conjugate hemispheres. *Geophysical Research Letters*, *32*(5), L05111. https://doi.org/10.1029/2004GL022166

Peirce, C. S. (1884). The numerical measure of the success of predictions. *Science*, *4*(93), 453–454. https://doi.org/10.1126/science.ns-4.93.453.b

Redmon, R. J., Seaton, D. B., Steenburgh, R., He, J., & Rodriguez, J. V. (2018). September 2017's geoeffective space weather and impacts to Caribbean radio communications during Hurricane response. *Space Weather*, *16*(9), 1190–1201. https://doi.org/10.1029/2018SW001897

Schrijver, C. J., Kauristie, K., Aylward, A. D., Denardini, C. M., Gibson, S. E., Glover, A., et al. (2015). Understanding space weather to shield society: A global road map for 2015–2025 commissioned by COSPAR and ILWS. *Advances in Space Research*, *55*(12), 2745–2807. https://doi.org/10.1016/j.asr.2015.03.023

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

Smith, A. W., Forsyth, C., Rae, I. J., Garton, T. M., Bloch, T., Jackman, C. M., & Bakrania, M. (2021). Forecasting the probability of large rates of change of the geomagnetic field in the UK: Timescales, horizons, and thresholds. *Space Weather*, *19*(9), e02788. https://doi.org/10.1029/2021SW002788

Smith, A. W., Freeman, M. P., Rae, I. J., & Forsyth, C. (2019). The influence of sudden commencements on the rate of change of the surface horizontal magnetic field in the United Kingdom. *Space Weather*, *17*(11), 1605–1617. https://doi.org/10.1029/2019SW002281

Smith, A. W., Rae, I. J., Forsyth, C., Oliveira, D. M., Freeman, M. P., & Jackson, D. R. (2020). Probabilistic forecasts of storm sudden commencements from interplanetary shocks using machine learning. *Space Weather*, *18*(11), e02603. https://doi.org/10.1029/2020SW002603

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Swets, J. A., Tanner, W. P., & G, B. T. (1955). *The evidence for a decision-making theory of visual detection (volume 40; technical report)*. Engineering Research Institute, University of Michigan.

Viljanen, A., Pulkkinen, A., Pirjola, R., Pajunpää, K., Posio, P., & Koistinen, A. (2006). Recordings of geomagnetically induced currents and a nowcasting service of the Finnish natural gas pipeline system. *Space Weather*, *4*(10), S10004. https://doi.org/10.1029/2006SW000234

Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences* (2nd ed.). Elsevier.