

Space Weather

RESEARCH ARTICLE

10.1029/2019SW002416

Key Points:

- A simple method for deterministic or probabilistic forecasting of the geosynchronous >2 MeV electron flux exceeding set levels is examined
- Forecasts are based on the amount of time a single input variable exceeds a set threshold and do not include measured geosynchronous flux
- AL, SYM-H, and solar wind pressure provide the most skilful forecasts when analyzed using standard forecasting metrics

Supporting Information:

- Supporting Information S1

Correspondence to:

C. Forsyth,
colin.forsyth@ucl.ac.uk

Citation:

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

Received 24 NOV 2019

Accepted 4 JUN 2020

Accepted article online 10 JUN 2020

Forecasting GOES 15 >2 MeV Electron Fluxes From Solar Wind Data and Geomagnetic Indices

C. Forsyth¹ , C. E. J. Watt² , M. K. Mooney^{1,3} , I. J. Rae¹ , S. D. Walton¹ , and R. B. Horne⁴ 

¹Department of Space and Climate Physics, Mullard Space Science Laboratory, University College London, London, UK, ²Department of Meteorology, University of Reading, Reading, UK, ³Met Office, Exeter, Devon, UK, ⁴British Antarctic Survey, Cambridge, UK

Abstract The flux of > 2 MeV electrons at geosynchronous orbit is used by space weather forecasters as a key indicator of enhanced risk of damage to spacecraft in low, medium, or geosynchronous Earth orbits. We present a methodology that uses the amount of time a single input data set (solar wind data or geomagnetic indices) exceeds a given threshold to produce deterministic and probabilistic forecasts of the >2 MeV flux at GEO exceeding 1,000 or 10,000 $\text{cm}^{-2} \text{s}^{-1} \text{sr}^{-1}$ within up to 10 days. By comparing our forecasts with measured fluxes from GOES 15 between 2014 and 2016, we determine the optimum forecast thresholds for deterministic and probabilistic forecasts by maximizing the receiver-operating characteristic (ROC) and Brier skill scores, respectively. The training data set gives peak ROC scores of 0.71 to 0.87 and peak Brier skill scores of -0.03 to 0.32. Forecasts from AL give the highest skill scores for forecasts of up to 6 days. AL, solar wind pressure, or SYM-H give the highest skill scores over 7–10 days. Hit rates range over 56–89% with false alarm rates of 11–53%. Applied to 2012, 2013, and 2017, our best forecasts have hit rates of 56–83% and false alarm rates of 10–20%. Further tuning of the forecasts may improve these. Our hit rates are comparable to those from operational fluence forecasts, that incorporate fluence measurements, but our false alarm rates are higher. This proof-of-concept shows that the geosynchronous electron flux can be forecast with a degree of success without incorporating a persistence element into the forecasts.

Plain Language Summary Spacecraft that orbit the Earth 36,000 km above the equator take 24 hr to orbit the Earth, meaning they stay above the same point on the Earth's all the time. These "geosynchronous" orbits are incredibly useful, enabling satellites to have constant contact with the ground. As of 31 March 2019, over 500 spacecraft are in geosynchronous orbit (<https://www.ucsusa.org>). Geosynchronous orbit is also on the edge of one of the most hazardous regions of space around the Earth—the Van Allen Radiation Belts. These belts contain highly energetic particles capable of damaging spacecraft in geosynchronous orbit and are highly variable. As such, predicting when the radiation belts at geosynchronous orbit are dangerous is a key concern in space weather. In this study, we create forecasts of the radiation belts based on simple measurements of upstream and local conditions of Earth's space environment. These simple forecasts provide the probabilities of high-risk events at geosynchronous orbit over periods of up to 10 days, providing an interesting new mechanism for forecasting the conditions in near-Earth space as well as helping us to understand the factors that control the dynamics of the radiation belts.

1. Introduction

The risk to satellite operations from highly energetic charged particles in near-Earth space is a key design and operations driver for spacecraft in low, medium, or geosynchronous Earth orbits. The exact flux and energy levels at which a spacecraft becomes vulnerable to energetic charged particles is dependent on its design and the components used, but high fluxes of near-relativistic electrons have been shown to present an increased risk of anomalies or failures in electronic hardware (Baker et al., 1987; Iucci et al., 2005; Romanova et al., 2005; Wrenn & Sims, 1996; Wrenn & Smith, 1996; Wrenn et al., 2002). Typically, there is an increased occurrence of anomalies and phantom commands when the >2 MeV 1- or 2-day fluence exceeds $10^8 \text{cm}^{-2} \text{sr}^{-1}$, equivalent to an average electron flux of $\sim 500\text{--}1,000 \text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$. Space weather monitoring and forecasting organizations, such as NOAA and the Met Office, now issue alerts when the

©2020. The Authors.

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

measured electron flux in geosynchronous orbit exceeds $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ or the fluence forecast by models such as the Relativistic Electron Flux Model (REFM; based on Baker et al., 1990) exceeds $10^8 \text{ cm}^{-2} \text{ sr}^{-1}$.

Modeling the variations in the radiation belts based on physical processes is non-trivial. The complex interplay of particle injections from processes further out in the magnetosphere and the energization, diffusion and scattering of particles from their interactions with electromagnetic waves poses a significant challenge. A full physics-based description of the radiation belts requires knowledge of the ion and electron distribution functions across all pitch angles and at all energies, the wave populations that they generate and the subsequent coupling back of those wave populations into variations in those distribution functions. In practice, electron fluxes at an outer boundary and global maps of average wave power under different solar wind or geomagnetic conditions (e.g., Meredith et al., 2012; Usanova et al., 2012) are used as the inputs to physics-based models (e.g., Horne et al., 2013; Subbotin & Shprits, 2009). However, there is significant overlap in the distributions of wave power categorized by geomagnetic activity which may lead to uncertainties in the outputs of models based upon them (Murphy et al., 2016; Watt et al., 2017). Nevertheless, these models show some success in reconstructing particle fluxes in the radiation belts, with correlation coefficients of 0.72 to 0.99 and skill scores of 0.6 to 0.8 (e.g., Glauert et al., 2018). Such models can be operationally useful if the inputs controlling the wave fields can be forecast. For example, the SPACECAST model (Horne et al., 2013) uses Kp-derived average wave power models, with Kp forecasts of up to 3 days now available from NOAA.

An alternative approach to operationally modeling the radiation belts is to produce empirical models of the electron fluence at geosynchronous orbit based on measured parameters, typically using solar wind data as input (e.g., Baker et al., 1990; Balikhin et al., 2016). These models use solar wind data from several days previous to provide forecasts of the radiation belt fluence over the next 1–3 days. However, atypical changes in the electron flux at geosynchronous orbit, such as rapid drop-outs or increases, are not well predicted from the upstream inputs alone. To correct for this, these models compare themselves to the observed fluence and make adjustments to account for significant offsets between forecasts and observations. As such, these models require both the upstream parameters and in situ measurements to provide accurate forecasts.

In the following, we show a proof-of-concept for forecasts that predict whether the $>2 \text{ MeV}$ flux at geosynchronous orbit is likely to be above the levels at which alerts are issued by space weather monitoring agencies, rather than predicting the actual flux. These forecasts are based on individual solar wind parameters and geomagnetic indices. Data from 2014–2016 inclusive are used as a training data set for the forecasts, and the forecast performance is checked using data from 2012, 2013, and 2017. We examine whether geomagnetic or solar wind activity observed over 1–10 days can provide a skillful forecast of the flux in the radiation belts exceeding $1,000$ or $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ within the next 1–10 days using standard forecast verification techniques and discuss possible implications for space weather operators and our understanding of radiation belt dynamics.

2. Data and Methodology

2.1. Data

In this study, we examine the extent to which the $>2 \text{ MeV}$ flux observed by the GOES 15 spacecraft can be forecast using simple “threshold breaking” forecasts based on different upstream and geomagnetic parameters. The electron flux in geosynchronous orbit is measured by the Energetic Proton Electron and Alpha Detector Electron, part of the Space Environment Monitor, on the GOES 15 spacecraft. In this study we use the 1-min data from the E2 channel of the A sensor (westward facing) which captures the $>2 \text{ MeV}$ electron flux. We use data from 1 January 2014 and 31 December 2016 to construct our forecasts, encompassing the maximum of solar cycle 24 and its declining phase.

Within the training data from GOES 15 $>2 \text{ MeV}$ flux data, $\sim 28\%$ of the data points are marked as bad or missing data. For the majority of these data points, the adjacent good data fluxes are low, with a median value of $35 \text{ cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1}$. As such, we assume that missing data is the result of low fluxes and treat any missing data as a nonevent.

We construct forecasts of the $>2 \text{ MeV}$ flux using the AL and SYM-H geomagnetic indices, which provide indication of geomagnetic activity through enhancements predominantly in the auroral current systems

(AL) and the ring current (SYM-H). We also create forecasts using the solar wind velocity and ram pressure and the IMF B_z . The data used in this study were all obtained through the NASA CDAWeb data service.

2.2. Terminology

To aid the reader, we define a series of terms that will be used throughout to discuss our methodology, analysis, and results.

1. “Events” are when the GOES 15 >2 MeV electron flux exceeds $1,000 \text{ cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1}$ (corresponding to approximately the median flux level) or $10,000 \text{ cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1}$ (corresponding to approximately the 90th percentile of flux) for at least one data point (1 min) within the forecast window. Events for the flux exceeding these two different thresholds are examined separately;
2. “Forecast Window” is a period of between 1 and 10 days in length within which we are forecasting that an event may occur. We note that a 10-day forecast does not indicate that an event will occur in 10 days’ time, but instead that an event will occur within the next 10 days. As such, longer forecasts naturally encompasses the shorter forecast windows as well. Forecast Windows directly follow an Input Window;
3. “Input Window” is a period of between 1 and 10 days over which the Input Variable is assessed to provide a forecast. Input Windows directly precede Forecast Windows;
4. “Input Variable” is the variable used to create the forecast, which in this study is one of either AL, SYM-H, IMF B_z , solar wind pressure, or solar wind velocity;
5. “Forecast Threshold” is the value the Input Variable must exceed in order to count toward a forecast. AL, SYM-H, IMF B_z must be below the Forecast Threshold and solar wind pressure or velocity must be above the threshold;
6. “Break Duration” is the amount of time that the Input Variable exceeds the Forecast Threshold within the Input Window.

2.3. Creating Forecasts

Forecasts were created between 1 January 2014 and 31 December 2016 inclusive. For each day, we determined the break duration over the previous 1–10 days (the input windows) for input thresholds set to each percentile of the input variable (calculated over 1 January 2014 to 31 December 2016). Sets of deterministic forecasts were then derived from this by setting a positive forecast to be returned when the break threshold exceeded each decile of the forecast window length, while sets of probabilistic forecasts were derived by setting the break duration as a proportion of the input window to be the event probability (see below for details). For each day, we also determined whether the 1-min flux exceeded either $1,000$ or $10,000 \text{ cm}^{-2} \text{ s}^{-1} \text{ sr}^{-2}$ within the next 1–10 days (the forecast window).

The start times of each of our forecasts are separated by 1 day. However, because most of the Input Windows and Forecast Windows are greater than 1 day in length, the calculated forecasts are not independent of one another; that is, some of the input data or forecast windows overlap with each other. In order to analyze independent forecasts, we extracted and analyzed subseries of forecasts that are separated by either the Input Window or Forecast Window duration, whichever is greater. For example, for forecasts with a 3-day Input Window and 5-day Forecast Window, we extracted a subseries of forecasts separated by 5 days. In order to use all the available data, this extraction was repeated by varying the start day and then taking the mean of the calculated verification metrics.

3. Deterministic Forecasts

Deterministic forecasts provide a binary prediction that an event will or will not occur. Analysis of these forecasts is typically based on contingency or truth tables. These show the number of times an event was: (A) predicted and occurred (hit); (B) predicted and did not occur (false alarm); (C) not predicted and occurred (miss); and (D) not predicted and did not occur (correct rejection). A number of measures of the validity or skill of the forecast can be calculated from these tables (for a discussion, see Hogan & Mason, 2012). Commonly used measures include the hit rate ($H=A/(A+C)$), false alarm rate ($F=B/(B+D)$), and the difference between them, known as the Peirce score (also known as Hanssen and Kuipers discriminant or True Skill score, Peirce, 1884).

The ability of a variable to distinguish between two different events or states, or the separation of the distributions of said variable for two different events or states, can be assessed using Receiver Operating Characteristics (ROC, e.g., Swets, 1988). In this study, we examine how well the distributions of the break durations are separated for periods of high or low energetic electron flux. This is done by generating a series of contingency tables by varying the value of the break duration for a given forecast threshold above which a positive forecast is made. In this study, this value is set to the deciles of the input window length. The hit rates and false alarm rates from these tables are then calculated and plotted against one another and the area under the curve is calculated, in our case using a trapezoidal sum with the end points fixed at [0,0] and [1,1]. This area is known as the ROC score and is a nonparametric two-sample statistic (e.g., Zweig & Campbell, 1993) that indicates the degree of separation of the distributions of break durations for events and nonevents. If the distributions of break duration are entirely separated, that is, there is a value of the break duration below which only nonevents occur and above which only events occur, then the ROC score is 1. If the distributions of the break duration entirely overlap, that is, break duration shows no skill in separating events and nonevents, then the ROC score is 0.5. We compare the ROC scores for different forecast thresholds to determine the optimum forecast threshold for which break duration can separate events and nonevents.

For a given input variable, the maximum ROC score for an input window/forecast window pair indicates the best input threshold for discriminating between high or low electron flux events. By comparing these maximum ROC scores across variables, we can then determine which variable provides the most skillful deterministic forecasts from our methodology. Given that there is likely to be some overlap in the distributions of break durations preceding events and nonevents, the best break duration can only be determined by a cost-benefit analysis by the forecast user. One simple form of this analysis is to find the break duration that returns the highest Peirce score, although different users may sacrifice higher hit rates for lower false alarm rates or vice versa. We note that the Peirce score from a ROC curve of forecasts with no skill will be 0 while the ROC score is 0.5. As such, the ranges of skilful ROC scores and Peirce scores differ by a factor of 2.

Figures 1 and 2 show matrices of the maximum average ROC scores calculated for each input window (X axis) and forecast window (Y axis) pair for forecasts of the >2 MeV flux exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ and $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$, respectively. In each panel, reds indicate maximum average ROC scores greater than 0.75 and blues indicate scores below this. For each forecast window length, the input window length that gives the highest maximum average ROC score is highlighted. Table 1 shows the variable that gives the highest maximum average ROC score for each forecast window length and the corresponding input window length and forecast threshold. ROC scores cannot indicate which break duration is best, so we report the hit rate, false alarm rate and break duration that give the highest Peirce score for the specified forecast threshold and input window length. For the interested reader, we have similar tables for each input variable in the supporting information.

The variation of the maximum average ROC score with input window and forecast window differs for each input variable. For AL and SYM-H forecasts of both the flux exceeding $1,000$ or $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$, the maximum average ROC score decreases with forecast window length, such that the highest ROC scores for any input window length are for forecasts windows of 1 day. Similarly, for a given forecast window, the ROC scores increase with input window length up to some maximum for an input window length of ~ 5 days for AL and $\sim 3\text{--}4$ days for SYM-H. The maximum average ROC scores from AL are greater than those for the other input variables for most input window and forecast window pairs. The exception to this is solar wind pressure, from which the maximum average ROC scores increase with input window length and forecast window length. The maximum average ROC scores from the solar wind pressure forecasts are greater than those from the AL forecasts when the sum of the forecast and input window lengths is greater than 16 for forecasts of flux exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ or 14 for forecasts of flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$.

The pattern for the solar wind input variables is distinctly different from that of the geomagnetic variables and indeed from each other. As described above, the maximum average ROC scores from solar wind pressure forecasts increase with forecast and input window length. The maximum average ROC scores from IMF B_z forecasts are low for all input windows and forecast window pairs, not exceeding a ROC score of 0.69. The maximum average ROC scores from forecasts by solar wind velocity have a distinct chevron pattern for forecasts of flux exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$. The peak maximum average ROC score occurs for

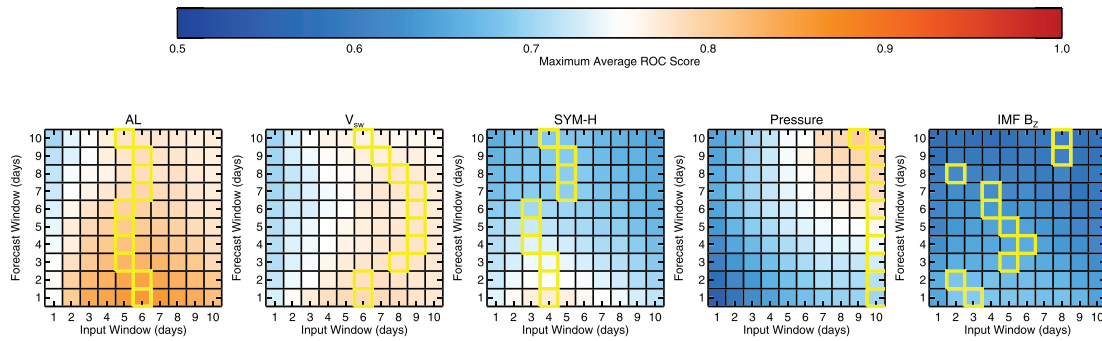


Figure 1. Matrices of the maximum average ROC scores for forecasts of electron flux exceeding $1,000 \text{ cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1}$ for different input and forecast window lengths. Results are shown for forecasts by AL, solar wind velocity, SYM-H, solar wind pressure, and IMF B_z . For each forecast window length, the maximum ROC score is highlighted by a yellow outline.

increasing input window lengths for forecast window lengths up to 5 days, then for decreasing input window lengths for forecast windows between 6 and 10 days. The maximum average ROC scores are relatively high for all input window and forecast window pairs, just below those for AL. This pattern is not repeated for the solar wind velocity forecasts of flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$, instead being qualitatively similar to the maximum average ROC scores from SYM-H forecasts. This is in line with the link between solar wind velocity and Dst (and by extension, SYM-H) during storms (e.g., Burton et al., 1975; Klimas et al., 1998). However, solar wind velocity and solar wind pressure show distinctly different patterns of maximum average ROC scores, with the ROC scores from solar wind velocity forecasts generally being higher than those from pressure forecasts. This implies that it is the solar wind coupling rather than the compression of the magnetosphere that is important for the enhancement of flux at geosynchronous orbit, likely due to the radiation belts being compressed to within geosynchronous orbit, and thus the flux at GEO dropping rather than increasing under high solar wind pressures.

For AL, SYM-H, and solar wind pressure, the input window lengths that give the highest maximum average ROC scores remain similar with increasing forecast window length. For AL, the highest maximum ROC scores come from input windows of 5 or 6 days for the lower flux forecasts and 3–6 days for the higher flux forecasts; for SYM-H, they come from input windows of 3–5 days for the lower flux forecasts and 2–4 days for the higher flux forecasts; for solar wind pressure, they predominantly come from input windows of 10 days for forecasts of both flux levels. For forecasts from the solar wind velocity and IMF B_z , the input windows that give the highest maximum ROC scores are more variable, although the input windows are generally shorter for the higher flux forecasts by solar wind velocity. As such, for forecasts of the electron flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$, less input data are required.

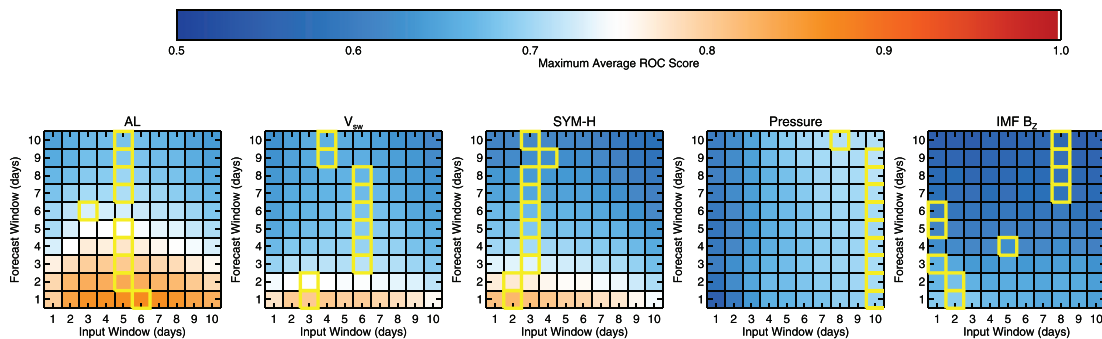


Figure 2. Matrices of the highest average ROC scores for forecasts of electron flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ for different input and forecast window lengths by AL, solar wind velocity, SYM-H, solar wind pressure and IMF B_z . Darker reds indicate higher ROC scores. For each forecast window length, the maximum ROC score is highlighted by a yellow outline.

Table 1

Parameters That Provide the Highest Maximum Average ROC Scores for Each Forecast Window Length for Forecasts of Flux Exceeding $1,000 \text{ cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1}$ (upper rows) and $10,000 \text{ cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1}$ (lower rows)

Forecast window (days)	Input window (days)	Input variable	ROC score	Variable percentile (%)	Variable value	Max Peirce score	Break threshold (%)	Hit rate	False alarm rate	Mean hits	Mean misses	Mean false alarms	Mean true negatives
1	6	AL	0.855	32	-103.0	0.578	30	0.84	0.26	71	14	21	60
2	6	AL	0.844	32	-103.0	0.553	30	0.79	0.23	75	20	16	54
3	5	AL	0.828	39	-74.0	0.503	40	0.68	0.18	85	39	13	61
4	5	AL	0.816	35	-90.0	0.511	30	0.79	0.27	104	28	18	48
5	5	AL	0.807	28	-124.0	0.469	30	0.58	0.11	81	58	6	53
6	5	AL	0.796	35	-90.0	0.486	30	0.75	0.26	90	31	11	33
7	6	AL	0.791	36	-86.0	0.469	30	0.76	0.29	81	26	10	24
8	10	Press	0.789	36	1.7	0.463	30	0.76	0.29	58	18	6	15
9	10	Press	0.792	36	1.7	0.471	30	0.75	0.28	59	20	5	14
10	9	Press	0.803	34	1.8	0.494	30	0.70	0.20	56	24	3	14
1	6	AL	0.870	13	-249.0	0.599	20	0.73	0.13	20	7	18	120
2	5	AL	0.837	17	-205.0	0.540	20	0.80	0.26	34	8	41	116
3	5	AL	0.805	17	-205.0	0.484	20	0.74	0.26	36	12	39	111
4	5	AL	0.777	17	-205.0	0.421	20	0.68	0.26	38	17	37	106
5	5	AL	0.751	16	-215.0	0.372	20	0.59	0.22	37	25	30	107
6	3	AL	0.729	21	-171.0	0.324	20	0.68	0.35	38	18	38	70
7	5	AL	0.709	17	-205.0	0.294	20	0.56	0.27	30	23	24	65
8	10	Press	0.714	25	2.1	0.375	20	0.87	0.50	35	5	29	29
9	10	Press	0.713	26	2.0	0.374	20	0.89	0.51	38	4	29	27
10	8	Press	0.711	27	2.0	0.342	20	0.87	0.53	40	6	28	25

Note. The corresponding maximum mean Peirce score and mean contingency table parameters are also given.

The highest ROC scores for most input window/forecast window pairs come from AL, showing that the electron fluxes at GEO are strongly influenced by geomagnetic activity such as substorms, which are thought to provide seed and source particle injections (e.g. Horne, 2007; Jaynes et al., 2015). The higher ROC scores for SYM-H predicting electron fluxes above $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$, and the similarity to the ROC scores from forecasts from the solar wind velocity indicates that these fluxes are most likely during storm-like times when the ring current is enhanced. Given that AL and SYM-H are both enhanced during storms, it is difficult to decouple the effects of storms and substorms; however, these results hint that enhancements in the geosynchronous electron flux can occur within ~ 3 days following enhancements in SYM-H but within a longer period of ~ 5 days following substorm activity.

Table 1 shows that for both the lower and higher flux levels, the forecasts with the highest ROC scores and highest Peirce scores have hit rates between 56% and 89% and false alarm rates between 11% and 53%. The average hit rates for the forecasts with the highest ROC scores is 74% for both the higher and lower flux levels. The highest ROC scores and Peirce scores are achieved for forecast windows of 1 day. For both the lower and higher flux thresholds, the break threshold is a similar percentage of the input window for all forecast windows. Similarly, the input windows and input thresholds are similar for AL and solar wind pressure when they provide the highest ROC scores. The ROC scores and Peirce scores decrease with increasing forecast window. The forecast thresholds for electron flux being above $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ appear to be a relatively low activity level (AL $> \sim 100$ nT), but these correspond to the 28th–39th percentiles of AL. That is to say that if AL is in the lower tercile of activity for at least 30% of the previous 5–6 days, the flux at geosynchronous orbit is likely to be above $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$. The forecast thresholds for the electron flux being above $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ are lower, around the 17th percentile. Given that the input window lengths are similar for the corresponding forecast windows, the break durations for the higher flux forecasts are two thirds of those for forecasts of the lower flux level, indicating that a higher level of activity but for a shorter period of time is needed to predict the electron flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$. For both flux levels, solar wind pressure outperforms AL for forecast windows of 8–10 days with hit rates comparable to AL for short forecast windows; however, the false alarm rates are higher at $\sim 50\%$.

An alternative to using geomagnetic or solar wind conditions is to base forecasts on persistence: predicting that an event will occur within a set forecast window if it also occurred within a set input window. By varying

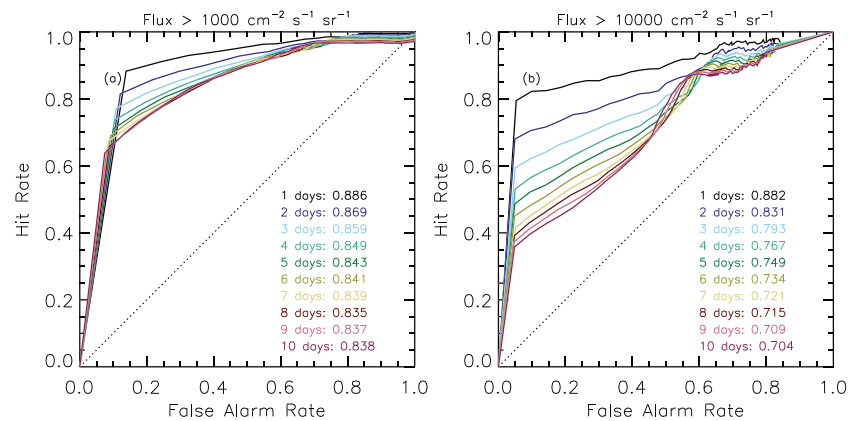


Figure 3. ROC curves (hit rate against false alarm rate) for persistence forecasts of the GOES >2 MeV electron flux exceeding (a) $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ and (b) $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$. Different color curves indicate the results for forecast windows of different lengths (1–10 days), as indicated by the plot key. Also given in the key are the corresponding ROC scores.

the input window length for each forecast window, we determine different hit rates and false alarm rates based on this principle and thus create a persistence ROC curve for each forecast window. Unlike the above, rather than generating a series of ROC curves for each forecast window/input window pair, we now can only create a single ROC curve for each forecast window. In order to give more complete curves, the input window range was extended to be up to 54 days. As above, independent forecasts are extracted from a complete set of forecasts and the average metrics are calculated.

Figure 3 shows the ROC curves for persistence forecasts with forecast windows of 1–10 days for (Figure 3a) the >2 MeV flux exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ and (Figure 3b) the >2 MeV flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$. Table 2 shows the maximum average ROC score for each forecast window along with the input window length that gives the highest Peirce score and the corresponding hit rates and false alarm rates.

The ROC scores from the persistence forecasts vary between 0.835 and 0.886 (Figure 3a) and 0.704 and 0.882 (Figure 3b), with the highest scores for forecast windows of 1 day or 22 days. The Peirce scores are also high, ranging from 49% to 74%, with hit rates of up to 91% and false alarm rates from 5% to 40%. The ROC scores from the persistence forecasts of flux exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ are higher than the corresponding deterministic forecasts by approximately 3 percentage points. In contrast, the ROC scores for persistence forecasts of the flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ tend to be lower than the corresponding deterministic forecast for forecast windows of up to 5 days. However, the Peirce scores of all the persistence forecasts are higher than the corresponding deterministic forecasts. The high ROC scores and hit rates from the majority of persistence forecasts is unsurprising when one considers variability of the radiation belts and the levels of flux being considered. The *e*-folding time for the radiation belt is of the order of 3–4 days (Borovsky & Steinberg, 2006) and the flux thresholds tested are close to the median and 90th percentiles. If the flux exceeds these thresholds by $\sim >30\%$, it will generally remain above that threshold for at least 1 day.

In summary, simple forecasts of whether the >2 MeV electron flux at geosynchronous orbit can be constructed by determining the amount of time various solar wind or geomagnetic activity parameters exceed a specified threshold. While persistence-based forecasts of the electron flux exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ have ROC scores that are consistently higher than those from our corresponding forecasts, for forecast windows of 2–5 days our forecasts of the flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ have higher ROC scores than the corresponding persistence-based forecasts, indicating that our deterministic forecasts can be more skilful than persistence. We note, however, that the persistence forecasts tend to have higher maximum Peirce scores. Overall, our deterministic forecasts with the highest ROC and Peirce scores have average hit rates of $\sim 74\%$ and false alarm rates of $\sim 27\%$.

3.1. Probabilistic Forecasts

Probabilistic forecasts provide an indication of the likelihood that a binary event will or will not happen and naturally account for some of the uncertainty in forecasting. For a perfect probabilistic forecast, an event

Table 2

The ROC Scores, Maximum Peirce Score, and Corresponding Parameters for Persistence Forecasts of the >2 MeV Flux Exceeding 1,000 (top) or 10,000 $\text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$ (bottom)

Forecast window days	ROC score	Input window days	Peircescore	Mean hit rate	Mean false alarm rate	Mean hits	Mean misses	Mean false alarms	Mean correct rejections
1	0.886	1	0.744	0.883	0.138	526	70	69	430
2	0.869	1	0.668	0.911	0.243	271	26	60	188
3	0.859	22	0.608	0.694	0.086	31	13	0	3
4	0.849	22	0.613	0.732	0.120	32	12	0	3
5	0.843	22	0.612	0.765	0.153	34	10	0	3
6	0.841	22	0.605	0.792	0.186	35	9	0	3
7	0.839	22	0.594	0.814	0.220	36	8	0	2
8	0.835	22	0.579	0.832	0.253	37	7	1	2
9	0.837	22	0.558	0.849	0.291	38	6	1	2
10	0.838	22	0.534	0.863	0.329	38	6	1	2
1	0.882	1	0.742	0.795	0.053	182	47	46	820
2	0.831	1	0.723	0.823	0.099	94	20	43	389
3	0.793	1	0.680	0.825	0.144	63	13	41	247
4	0.767	1	0.647	0.834	0.187	47	9	40	175
5	0.749	1	0.621	0.847	0.226	38	7	39	134
6	0.734	1	0.591	0.854	0.263	32	5	38	106
7	0.721	1	0.553	0.854	0.301	27	4	37	86
8	0.715	1	0.532	0.866	0.334	24	3	36	72
9	0.709	1	0.505	0.871	0.366	21	3	35	61
10	0.704	1	0.485	0.881	0.395	19	2	34	52

would occur 10% of the time that a probability of 10% was forecast. In a similar manner to the deterministic forecasts, we created probabilistic forecasts by examining the proportion of an input window that the input variable exceeds a forecast threshold. We arbitrarily set the probability of an event occurring to the break duration as a proportion of the input window. This assumes that a linear relationship between the amount of time a forecast variable is beyond some limit and the likelihood of an event occurring and does not account for the actual magnitude of the input variable or any more complex relationships.

The nature of probabilistic forecasts means that an individual forecast is neither right nor wrong. For example, if a forecast gives a 90% probability that the fluxes will be high on the next day, but the fluxes are low, then the forecast was not necessarily wrong since 10% of the time a forecast of 90% is given, the fluxes should be low. As such, one cannot determine the components of a contingency table for a probabilistic forecast and a ROC score cannot be generated. Instead, probabilistic forecasts can be examined using Brier scores (Brier, 1950) or Brier skill scores. Brier scores are the mean squared error of the forecasts, calculated as

$$B = \frac{1}{N} \sum_{i=1}^N (f_i - e_i)^2 \quad (1)$$

where N is the total number of forecasts, f is the forecast probability (in the range 0 to 1). e is set to 1 if the event occurred or 0 if it did not. A perfect forecast will have a Brier score of 0. The Brier skill score compares the calculated Brier score with that from a reference forecast, typically a climatological forecast. The Brier skill score is then calculated as

$$BSS = 1 - \frac{B}{B_{\text{ref}}} \quad (2)$$

where B_{ref} is the Brier score of the reference forecast. The Brier skill score varies between $-\infty$ and 1, with a perfect forecast returning a Brier skill score of 1 and a forecast which shows no improvement over the reference forecast returning a score of 0. In this study, we used the climatology (average event occurrence rate for each forecast window length) to create reference forecasts and associated Brier skill scores.

Figure 4 shows matrices of the maximum average Brier skill scores calculated for each input window (X axis) and forecast window (Y axis) pair for forecasts of the >2 MeV flux exceeding $1,000 \text{ cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$, with darker

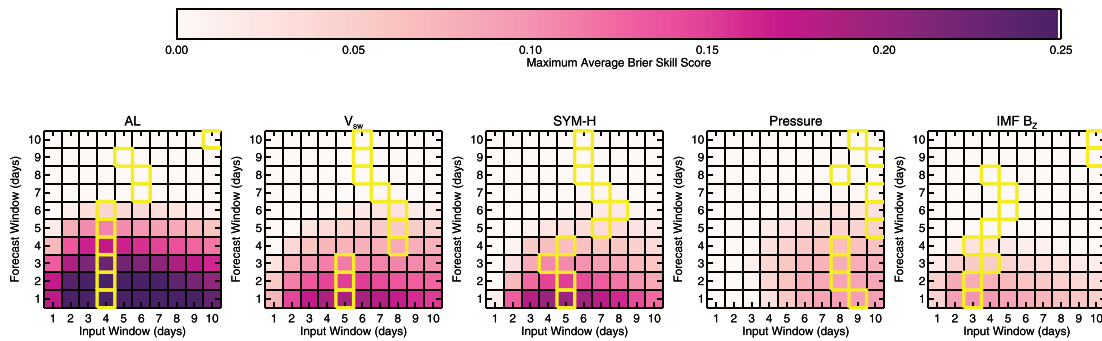


Figure 4. Matrices of the highest Brier skill scores for forecasts of electron flux exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ for different input and forecast window lengths by AL, solar wind velocity, SYM-H, solar wind pressure, and IMF B_z . Darker purples indicate higher Brier skill scores. For each forecast window length, the maximum Brier skill score is highlighted by a yellow outline.

colors indicating higher scores. Figure 5 shows similar results for forecasts of the $>2 \text{ MeV}$ flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$.

The overall trends in the Brier skill scores show some similarities to those in the ROC scores from the deterministic forecasts: forecasts from IMF B_z have low Brier skill scores for all forecast window and input window pairs; the Brier skill scores from forecasts using AL and SYM-H decrease with increasing forecast window; the maximum Brier skill score occurs at similar input windows for each forecast window for forecasts of the flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ by AL and SYM-H; the Brier skill scores from forecasts of the flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ using solar wind pressure increase with increasing length of forecast window and input window. For the higher flux forecasts, the Brier skill score drops off more rapidly with input window length compared to the lower flux forecasts, but the Brier skill scores from the lower flux forecasts drop off more rapidly with forecast window length than the higher flux forecasts. Unlike the ROC scores of the deterministic forecasts, the Brier skill scores from solar wind velocity forecasts are qualitatively similar to those of SYM-H for forecasts of the flux exceeding both $1,000$ and $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$. Furthermore, the highest Brier skill scores for solar wind pressure forecasts of the flux exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ are achieved for long input windows but short (<5 day) forecast windows.

For forecasts by AL and solar wind pressure, and for forecasts of the flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ by SYM-H and solar wind velocity, the input windows that give the highest maximum average Brier skill scores are similar. For AL, these are typically 3–4 days; for SYM-H they are 4–5 days; for solar wind velocity they are ~ 6 days, although the highest Brier skill scores are for input windows of 3–4 days; and for solar wind pressure they are 9–10 days. This is in contrast to the deterministic forecasts, in which SYM-H gives the highest ROC scores in the shortest input windows and indicates that auroral zone geomagnetic activity over a relatively short interval provides a good indicator of the likelihood of high fluxes at GEO.

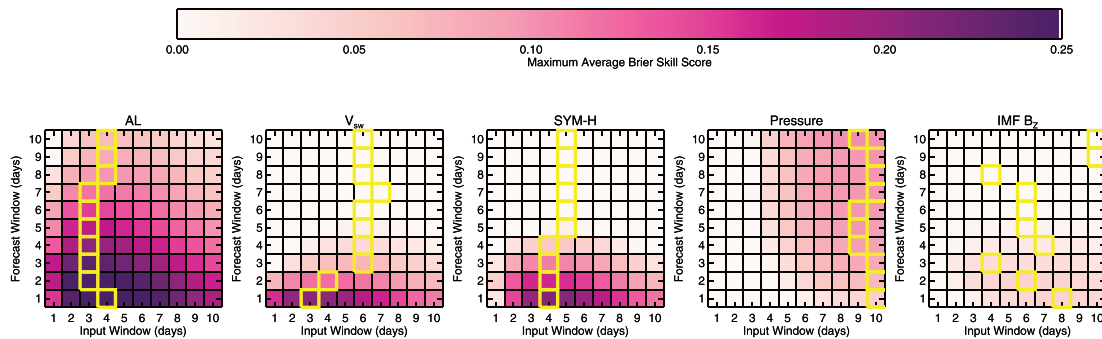


Figure 5. Matrices of the highest Brier skill scores for forecasts of electron flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ for different input and forecast window lengths by AL, solar wind velocity, SYM-H, solar wind pressure, and IMF B_z . Darker purples indicate higher Brier skill scores. For each forecast window length, the maximum Brier skill score is highlighted by a yellow outline.

Table 3
Parameters That Provide the Highest Maximum Average Brier Skill Scores for Each Forecast Window Length for Forecasts of Flux Exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ (Upper Rows) and $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ (Lower Rows)

Forecast window (days)	Input window (days)	Brier skill score	Variable	Variable percentile	Variable value (nT)	Brier score
1	4	0.32	AL	47	-50.0	0.17
2	4	0.27	AL	51	-42.0	0.18
3	4	0.23	AL	53	-39.0	0.18
4	4	0.17	AL	55	-36.0	0.18
5	4	0.11	AL	57	-34.0	0.19
6	4	0.04	AL	58	-32.0	0.19
7	7	0.01	SYM-H	76	0.0	0.18
8	6	0.00	SYM-H	78	1.0	0.17
9	6	-0.02	SYM-H	78	1.0	0.16
10	6	-0.03	SYM-H	81	2.0	0.16
1	4	0.28	AL	16	-215.0	0.10
2	3	0.27	AL	20	-178.0	0.12
3	3	0.24	AL	22	-163.0	0.14
4	3	0.20	AL	25	-142.0	0.16
5	3	0.18	AL	27	-130.0	0.18
6	3	0.15	AL	30	-113.0	0.19
7	3	0.12	AL	34	-94.0	0.21
8	4	0.10	AL	36	-86.0	0.22
9	10	0.10	Press	41	1.5	0.22
10	9	0.10	Press	43	1.5	0.22

Table 3 shows the variables and inputs that provide the best Brier skill scores for forecasting the probability of the $>2 \text{ MeV}$ flux exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ (top rows) or $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ (bottom rows) and for each forecast window. A set of tables which shows similar results for each input variable is provided in the supporting information. For both sets of forecasts, the Brier skill score decreases with increasing forecast window, indicating that our forecasts are tending toward forecasts based on the climatology. This is a natural consequence of our analysis. For our forecasts of the electron flux exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$, the Brier

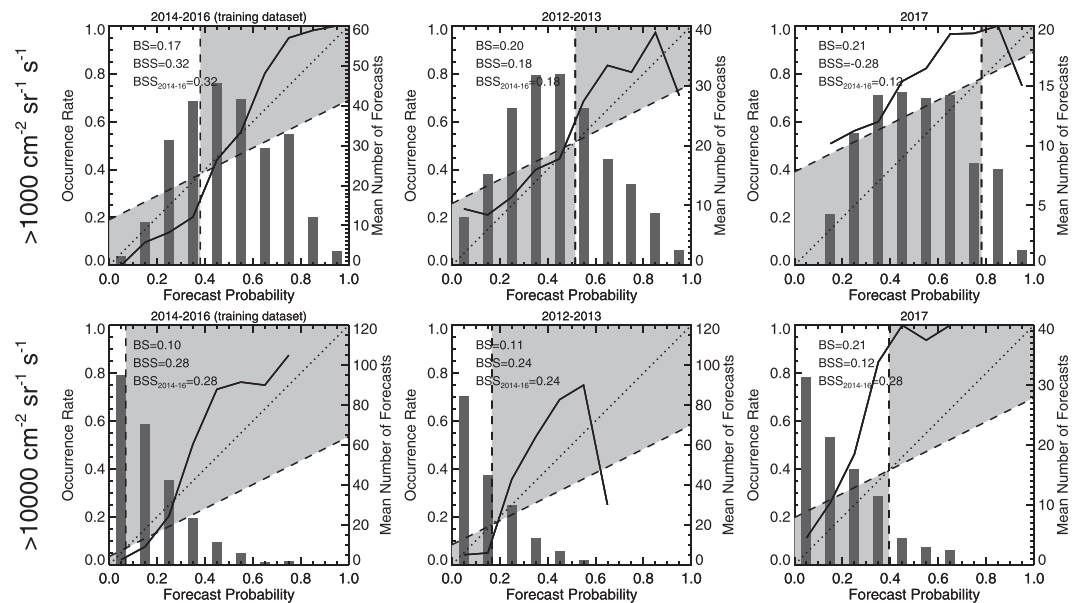


Figure 6. Reliability curves of the 1-day forecasts (occurrence against forecast probability) for 2014–2016, and for 2012–2013 and 2017. The forecast parameters are those given in Table 3. The vertical and horizontal dashed lines show the climatology of events for that period. The diagonal dashed line shows half the sum of the climatology and the probability. The histograms show the mean number of events in each category. The top and bottom rows show the reliability curves for forecasts of the flux exceeding $1,000$ and $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$, respectively.

score remains relatively constant with increasing forecast window length, indicating that the performance of our best forecasts does not degrade with increasing window length, but rather that the occurrence of flux exceeding this near-median level tends toward the climatology. While SYM-H provides the highest Brier skill scores for forecasts of the flux exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ within 7 or more days, these forecasts only perform as well as the climatology. In contrast, the Brier scores for forecasts of the flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ steadily increase with forecast window length, indicating that our best forecasts perform less well with increasing forecast window, but still perform better than the climatology for all forecast windows.

Table 3 also shows that the forecast threshold increases with increasing forecast window length. Considering the forecast threshold as a percentile of the input variable, this varies between 47% and 81% for the lower flux forecasts and between 16% and 43% for the higher flux forecasts. However, the input window length does not show a similar increasing trend, instead staying relatively constant for a given input variable. The implication of this is that greater activity is required to forecast over shorter periods than longer periods (noting that smaller AL and SYM-H values indicate higher activity levels).

In summary, simple probabilistic forecasts of the $>2 \text{ MeV}$ flux at geosynchronous orbit exceeding $1,000$ or $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ based on the duration that an input variable exceeds a set threshold can provide skillful forecasts that perform better than the climatology. The mean square error of these forecasts is between 0.1 and 0.22. AL provides the best input variable, with extended periods of moderate activity forecasting the $>2 \text{ MeV}$ flux exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ and shorter periods of greater activity forecasting the $>2 \text{ MeV}$ flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$.

4. Forecast Verification

In the above, we have used data from 2014 to 2016 inclusive to produce simple forecasts of the $>2 \text{ MeV}$ flux at geosynchronous orbit exceeding either $1,000$ or $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$. We now examine the performance of these forecasts by applying them to data from 2012, 2013, and 2017 which was not used in the forecast training data sets.

4.1. Verifying Probabilistic Forecasts

Reliability plots or attributes diagrams (Hsu & Murphy, 1986) show the occurrence rate of events against the grouped forecast probabilities. The accompanying histograms show the number of forecasts within each probability bin. The shaded region, delimited by the climatology (\bar{o} , horizontal and vertical lines) and a diagonal line given by $Y=(X + \bar{o})/2$ indicates those results that contribute positively to the Brier skill score, weighted by their relative occurrence. The results from a perfect forecast will lie along the one-to-one line. To generate the reliability curves in Figure 6, we created forecasts of the flux exceeding $1,000$ or $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ between 2012 and 2017 inclusive using the forecast parameters with the highest Brier skill scores in Table 3. The forecasts between 2014 and 2016 are our training set. As above, we extracted subseries of independent forecasts and calculated the mean occurrence from these subseries in each of the probability bins. Similarly, the histograms show the mean number of events in each bin. In each panel, the climatology used to define the shaded area is the average occurrence of events within the time period of the plot. The forecast Brier scores are given in the panels, along with the Brier skill scores calculated using the climatology within that period as a reference (BSS) or using the climatology of the training data as a reference ($BSS_{2014-2016}$).

The reliability curves from our training data set largely fall within the gray-shaded region, consistent with their low Brier score and positive Brier skill scores. For both flux levels, the forecasts tend to over-predict occurrence below the climatology and underpredict occurrence above the climatology, but the trend runs close to the one-to-one line. For the higher flux level forecasts (Figure 6, lower panels), there are fewer events in the probability bins in which the occurrence rate differs from the forecast probability the most. For example, in 2017 the occurrence rate in the 0.5–0.6 and 0.6–0.7 forecast probability bins was 1.0, deviating from the forecast probability by 30–50%. However, on average there were fewer than five events in the 0.5–0.6 and 0.6–0.7 probability bins compared to 23 in the 0.1–0.2 probability bin, for which the occurrence rate deviated from the probability by only 10%. Given that the Brier score is the mean of the difference of the forecast probability and occurrence (Equation 1), the Brier score is naturally weighted towards the most

common probabilities. As such, the deviation of the occurrence rates from the higher probabilities does not strongly influence the Brier score.

The reliability curves from the application of our forecasts to 2012–2013 and 2017 show that the forecasts perform less well in these years, tending to under-predict the occurrence of high flux events. The Brier skill scores calculated using the concurrent climatologies are much lower than the Brier skill scores for the training data set. In contrast, the Brier skill scores calculated using the climatology of the training data set remain positive, indicating that the forecasts outperform forecasts based on the climatology of the training data. This implies that the underlying year-to-year variability in the climatologies of the fluxes exceeding 1,000 or 10,000 $\text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$ is not well captured by a single set of forecast parameters. However, the fact that all the reliability curves increase with increasing probability and that the training data provides good Brier skill scores would suggest that this forecast methodology can be tuned to give skilful forecasts of high-flux intervals.

4.2. Verifying Deterministic Forecasts

Figure 7 shows the daily maximum fluxes measured by GOES-15 color-coded by the 1 day deterministic forecast of the flux exceeding 1,000 or 10,000 $\text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$ using the parameters set out in Table 1. Blues indicate a correct forecast, oranges indicate a miss, greens indicate a false alarm and whites indicate a correct nonevent forecast. The top panels show the results for 2012, the middle panel shows the results for 2013, and the bottom panels show the results for 2017. The mean hits, false alarms, misses, and correct rejections for the forecasts are shown by each panel. As for the training data set, we calculate the forecast metrics for each year by extracting multiple subseries of independent forecasts and report the means of metrics of these subseries. Given that the optimum forecasts use a 6-day input window, each subseries contains 60 forecasts. Overall, the forecasts perform relatively well, with a majority of blue or white coloured days. For forecasts of the flux exceeding 1,000 $\text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$, there are a greater number of misses than false alarms and these show a slight tendency to occur following the correct forecast of high fluxes. Given that our forecasts do not predict how high the fluxes get and do not include any decay times, these misses may arise due to the persistence of fluxes above the median level. The hit rates for the forecasts of the flux exceeding 1,000 $\text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$ in 2012, 2013, and 2017 were 0.72, 0.63, and 0.72, respectively, somewhat lower than the training data (see Table 1). Similarly, the false alarm rates were 0.26, 0.14, and 0.25, respectively, comparable to the training data. The forecasts of the flux exceeding 10,000 $\text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$ had hit rates of 0.64, 0.57, and 0.65 in 2012, 2013, and 2017, respectively, again somewhat lower than from the training data set, and false alarm rates of 0.11, 0.08, and 0.09, slightly lower than the training data. The mean level of flux during the false alarms for forecasts of the flux exceeding 10,000 $\text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$ was 3,693, 3,759 and 4,419 $\text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$ in 2012, 2013, and 2017, respectively, thus these false alarms tend to occur when the flux is elevated, but does not reach the 10,000 $\text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$ level.

The differences in the performances of both the deterministic and probabilistic forecasts between the training data and 2012, 2013, and 2017 indicate that the parameters that provide the highest ROC, Peirce, and Brier skill scores are somewhat variable. Figure 6 shows that the climatology of the occurrence of high-flux events varies, with the flux having been more frequently above 1,000 $\text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$ in 2017. Given that our methodology does not use measured values of the flux at GEO as an input to the forecasts, they do not account for this underlying variation in the climatology. This may be corrected for by using, for example, the previous year as training data for the current year and regenerating the forecast parameters more frequently.

5. Discussion

We have demonstrated that the amount of time a single variable exceeds a set forecast threshold can be used to generate either deterministic or probabilistic forecasts of the >2 MeV electron flux at GEO being above 1,000 or 10,000 $\text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$. The optimum forecast thresholds are determined by maximizing standard forecasting metrics for deterministic and probabilistic forecasts. By comparing these metrics for different input variables, we find that AL, SYM-H, and solar wind pressure provide the most skilful forecasts of high-flux events.

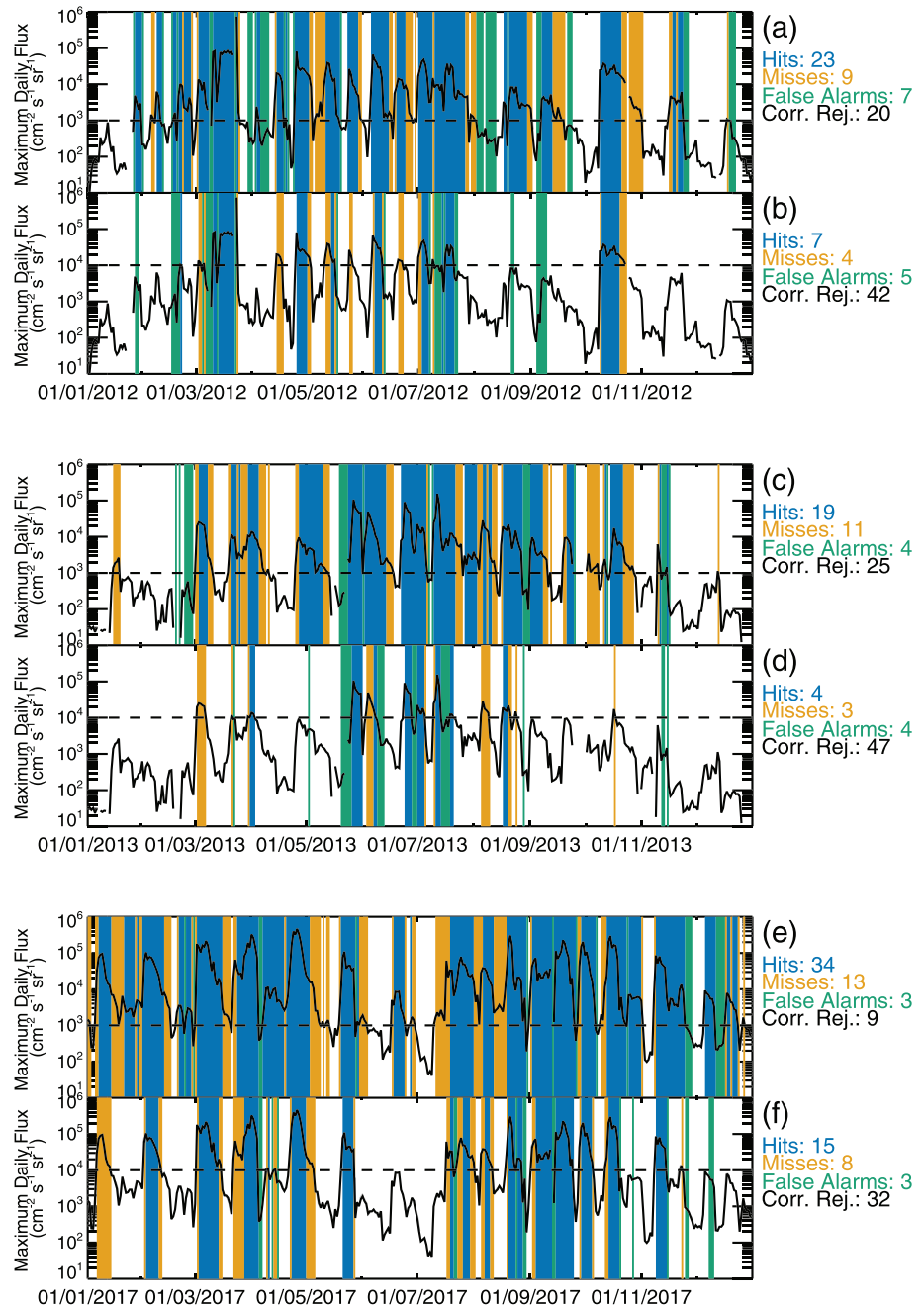


Figure 7. Plots of the daily maximum GOES-15 >2 MeV flux in 2012 (top), 2013 (middle), and 2017 (bottom) color coded by deterministic forecasts of the flux exceeding either 1,000 (a, c, and e) or 10,000 $\text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$ (b, d, and f). Blues indicates a correct forecast, oranges indicate a miss, greens indicate a false alarm and whites indicate a correct rejection. The dashed lines in each panel indicate the level we are forecasting the flux to exceed. The mean number of hit, misses, false alarms, and correct rejections from the subseries of independent forecasts are given beside each panel.

Tables 1 and 3 show that some of the parameters and skill scores of our forecasts with the highest ROC or Brier skill scores show little variation. For example, the AL-based deterministic forecasts of the electron flux exceeding 10,000 $\text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$ have ROC scores of 0.71–0.87 from variable percentiles of 13–21% and Peirce scores of 0.29–0.6 from break thresholds of 20%. This is a natural consequence of our methodology. Our methodology gives the highest skill scores for forecasts of the flux being high within 1 day into the future. Longer forecasts must also include this first day, thus the success of longer forecasts is dominated by

success in the near-term. However, within our training data, we do find successful longer term forecasts that are not successful due to high fluxes in the short term. For example, there are 48 successful 6-day forecasts of the flux exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ when the flux was not high in the first 3-days of the forecast window.

5.1. Comparison With Other Forecasts

Earth's radiation belt is heavily populated by a range of spacecraft, all of which have some risk exposure due to the abundance of high-energy particles in this region. Empirically, spacecraft are at an enhanced risk of anomalies due to internal charging effects when the 2-day $>2 \text{ MeV}$ electron fluence exceeds $10^8 \text{ cm}^{-2} \text{ sr}^{-1}$, corresponding to a mean flux of $\sim 500 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ (Baker et al., 1987; Iucci et al., 2005; Romanova et al., 2005; Wrenn & Smith, 1996; Wrenn & Sims, 1996; Wrenn et al., 2002). Both NOAA and the Met Office issue "Electron Event Alerts" to satellite operators when the observed $>2 \text{ MeV}$ flux exceeds $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ or when the 72-hr fluence predicted by the Relativistic Electron Flux Model exceeds $10^9 \text{ cm}^{-2} \text{ sr}^{-1}$. However, this only allows for nowcasts of the flux level or relies on upstream solar wind data and in-situ flux measurements. In this study, we have shown that it is possible to forecast the likelihood of the fluxes reaching these levels and thus an alert being issued within up to a 10-day interval based on simple measurements of either AL, SYM-H, or solar wind pressure and with hit rates of up to 84%.

A variety of radiation belt models have been developed to predict the current or near-future state of the radiation belts. These typically fall into two categories: physics-based semiempirical models that typically impose a measured particle distribution at an outer boundary and use empirical models of various wave populations to provide the magnetospheric conditions to solve a series of transport equations and thus forecast the fluxes across the whole of the radiation belts (e.g., Glauert et al., 2014a, 2014b; Fok et al., 2008; Li et al., 2001; Subbotin & Shprits, 2009; Turner & Li, 2008); or empirical and machine learning models which use a combination of geomagnetic and upstream solar wind conditions and recently measured radiation belt fluxes to predict flux levels for a few days into the future (e.g., Baker et al., 1990; Boynton et al., 2013, 2015, 2016; Chen et al., 2019; Coleman et al., 2018; Shin et al., 2016; Wei et al., 2018). In these models, a key component is the flux at the outer boundary or the measured flux at geosynchronous orbit. Our forecasts have three key differences to these models: our forecasts only predict whether the flux will exceed some value rather than predicting the actual flux that will be observed; our forecasts do not use in situ measurements of the flux in any form; our forecasts predict over a much longer time range. However, given the way that our forecast windows have been constructed, we do not predict when within a forecast window a flux threshold will be exceeded. As such, our forecasts may be a useful supplement to other operational models since they can operate in the absence of any upstream or in-situ measurements and provide some indication as to the possible upcoming hazardous levels of energetic electron flux.

Balikhin et al. (2016) presented contingency tables for the REFM and SNB3GEO models forecasting the 1-day $>2 \text{ MeV}$ electron fluence exceeding 10^8 or $10^9 \text{ cm}^{-2} \text{ sr}^{-1}$ between 2 March 2012 and 31 December 2013. While these forecasts are not directly comparable to our own, comparing their performance is informative. From these tables, the hit rates for these forecasts were 66% (86/129), and 57% (4/7), respectively for the REFM and 82% (106/129) and 57% (4/7), respectively, for SNB3GEO, with the bracketed numbers indicating the number of hits and the sum of the hits and misses. From our training data set, the hit rates of our 1-day forecasts of flux exceeding 1,000 or $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ were 83% and 71% respectively. Restricting our forecasts to the same period as Balikhin et al. (2016), we get hit rates of 67% (239/357) and 52% (63/121) for forecasts of the flux exceeding 1,000 and $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$, respectively. As such, our forecasts, which predict the $>2 \text{ MeV}$ flux exceeding 1,000 or $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ have similar hit rates to forecasts of the 1-day fluence exceeding 10^8 or $10^9 \text{ cm}^{-2} \text{ sr}^{-1}$, although notably there are a greater number of events for our forecasts to predict. Similarly, we calculate that the false alarm rates were 5% (22/532) and 1% (7/654), respectively, for the REFM and 6% (33/532) and 0.3% (2/654), respectively, for SNB3GEO. In comparison, the false alarm rates for our forecasts were higher at 21% (66/313) and 10% (56/549).

Since both the REFM and SNB3GEO models incorporate corrections based on in-situ measurements of the $>2 \text{ MeV}$ fluxes, it is informative to compare their hit rates and false alarm rates with those of a persistence forecast. Over the same period as the Balikhin et al. (2016) study, a 1-day persistence forecast of the fluence exceeding $10^8 \text{ cm}^{-2} \text{ sr}^{-1}$ had a hit rate of 72% (93/129) and a false alarm rate of 7% (36/530). Similarly, a 1-day forecast of the fluence exceeding $10^9 \text{ cm}^{-2} \text{ sr}^{-1}$ had a hit rate of 57% (4/7) and a false alarm rate of

0.5% (3/652). The hit rates for the lower fluence level are higher than the REFM, but 10% lower than SNB3GEO, but the false alarm rates were within a few percentage points. For the higher fluence level, the hit rates and false alarm rates were similar for the REFM, SNB3GEO, and persistence forecasts. This would suggest that the persistence aspect of the REFM and SNB3GEO tends to dominate forecasts of high fluence. By definition, our forecasts have no persistence aspect.

Comparing the hit rates of our forecasts between the training data set, 2012–2013 and 2017 we see that, while our forecast methodology has some merits, the forecast parameters have a level of instability. The Reliability Curves in Figure 6 show that the climatology of the flux exceeding 1,000 or 10,000 $\text{cm}^{-2} \text{sr}^{-1} \text{s}^{-1}$ varied across the three different time periods, becoming notably higher in 2017. Given that our forecasts do not use the measured flux in anyway, they cannot account for this other than by recalculating the forecast parameters. We note that the parameters for the SNB3GEO model were determined from data between 11 July 2004 and 11 October 2005 (Boynton et al., 2015) and this forecast performs relatively well during 2012–2013, especially at the low fluence level. As such, there may be some underlying variability in the radiation belts which is not adequately captured by geomagnetic activity or solar wind driving over a few days, but can be accounted for by including some level of persistence.

To our knowledge, our methodology is the first to provide a probabilistic forecast of the geosynchronous >2 MeV electron flux. As per our deterministic forecasts, AL proved to provide the highest Brier skill scores using inputs over the previous 3–5 days. The Brier skill scores, which are a measure of the improvement over the climatology, were above 0.2 for forecast windows of up to 3 days, a level comparable to some terrestrial weather forecasts (e.g., Barnston et al., 2010; Stefanova & N. Krishnamurti, 2002). In order to calculate the forecast probabilities, we made the arbitrary assumption that the duration that the forecast threshold was broken within the input window was equivalent to the probability of an event occurring. The mean squared errors in the forecasts were between 10–20%, driven by the forecasts overestimating occurrence rates below the climatology and underestimating occurrence rates above the climatology, which would indicate a slightly non-linear relationship between break duration and the probability of an event. In principle, the reliability diagrams in Figure 6 can be used as a calibration curve for the forecasts to correct for this, but more complex relations could be examined. For example, Mourenas et al. (2019) showed that enhancements in the radiation belt occur after the integrated Ap index was above 800nT hr. Further forecasts could be based upon this integrated measure or equivalently the average of a geomagnetic index within a given input window.

The methodology used to verify the forecasts is based on that commonly used in terrestrial weather forecasting. These methods have been used in other aspects of space weather forecasting (e.g., Murray et al., 2017; Sharpe & Murray, 2017) and are widely applicable within this field.

5.2. Radiation Belt Drivers

The >2 MeV electron flux at geosynchronous orbit can vary by several orders of magnitude (e.g., Paulikas & Blake, 1979). These variations are driven by the interactions between the solar wind and Earth's magnetosphere but can be complex as different parts of the system respond in different ways at different times and these responses can be coupled to each other. As a result, geomagnetic activity can arise through near-direct driving by large-scale structures in the solar wind resulting in geomagnetic storms (e.g., Borovsky & Denton, 2006), or as an integrate-and-fire response to solar wind coupling, resulting in geomagnetic substorms Freeman and Morley (2004). While the global response of the radiation belts to geomagnetic storms is repeatable (Murphy et al., 2018), the variations at different locations in the radiation belts are more complex (Reeves et al., 2003). This response is complicated by the fact that the spatial location of closed electron drift paths are strongly dependent on the magnetic field configuration: during geomagnetic storms, a compression of the magnetosphere means that geosynchronous orbit is closer to the outer edge of the magnetosphere and that the radiation belts are located closer to the Earth. The radiation belts can also be enhanced by lower-level geomagnetic activity such as substorms (Horne & Thorne, 1998; Horne, Thorne, Glauert, et al., 2005; Horne, Thorne, Shprits, et al., 2005; Jaynes et al., 2015; Li et al., 2007; Summers et al., 1998), with 50% of substorm intervals followed by an increase in the number of electrons in the radiation belts within 24–48 hr (Forsyth et al., 2016). Overall, the response of the fluxes of electrons at different energies and different L^* shells shows stronger correlations with measures of geomagnetic activity than with the solar wind driver (Zhao et al., 2017). The integrated geomagnetic activity can be a good indicator of likely

radiation belt enhancements (Mourenas et al., 2019). As such, both internal and external drivers influence the state of the radiation belts and fluxes at set locations in the magnetosphere.

Our results show that, in general, geomagnetic indices provide better forecasts of the >2 MeV flux exceeding either $1,000$ or $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ than solar wind parameters, but that the forecast verification parameters (ROC scores, Brier skill scores) can be similar. This is in keeping with recent studies of the variability of radiation belt fluxes. Zhao et al. (2017) showed correlation coefficients of ~ -0.5 between AL or solar wind velocity and the flux of $1,000$ MeV/G particles at an L^* of 6 but that across other L^* shells, the correlation with AL was generally higher. Other studies have shown there is greater mutual information between the radiation belt seed populations and geomagnetic parameters than with solar wind parameters (Tang et al., 2017). However, that is not to say that the solar wind has no influence. Studies such as Balikhin et al. (2011), Boynton et al. (2013), and Wing et al. (2016) have shown that the coupling between the radiation belts and the solar wind driver are nonlinear and may have different time lags but there is mutual information between them. The complex response of the magnetosphere to both external driving and internal conditions complicates forecasting the radiation belts based on upstream conditions alone. For example, the occurrence of substorms, which are thought to provide the necessary particle injections to energize the radiation belts (Forsyth et al., 2016; Horne, 2007; Jaynes et al., 2015), appears to be somewhat random, possibly occurring as an integrate-and-fire process (Freeman & Morley, 2004). For such events, the processing of the solar wind driver by the magnetosphere, which is reflected in geomagnetic activity, is a critical component in determining changes to the radiation belts.

In this study, we have opted to use AL as an input variable to our forecasts, as opposed to the AE index, which is commonly used as a parameterization for radiation belt model inputs (e.g., Meredith et al., 2012). AE is the difference between the AL index, which is designed to capture the strength of the westward electrojet and is always negative, and AU, which is designed to capture the eastward electrojet and is always positive. Given that substorm injections occur on the nightside in conjunction with the formation of the substorm current wedge and associated enhancement in the westward electrojet, we consider that AL is a more appropriate input to a forecast of the fluxes at the outer edge of the radiation belts.

In order to predict fluxes above $1,000$ or $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$, our deterministic forecasts require that AL drops below -74 or -171 nT, respectively. These are not particularly extreme values of AL, corresponding to the 36th and 17th percentiles. Superposed epoch analysis of substorms shows that, on average, AL just reaches -200 nT at the end of the expansion phase, but activity during the growth phase brings AL below -100 nT (Juusola et al., 2011). We also note that the thresholds for SYM-H to predict geosynchronous flux enhancements are above -10 nT (see supporting information), 45 percentage points above the -50 nT level that is taken to indicate storm activity (Gonzalez et al., 1994). Our results thus show that enhancements in flux at geosynchronous orbit follow periods of ongoing substorm activity, but activity at a level that would not necessarily be classed as a geomagnetic storm.

6. Conclusions

We have examined whether various geomagnetic indices and solar wind parameters can be used to forecast whether the >2 MeV flux from GOES 15 will exceed either $1,000$ or $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ within a forecast window of up to 10 days. These forecasts are based on the amount of time forecasting variables exceed a set threshold during a window of time prior to the forecasting time and rely entirely on conditions and observations external to the radiation belts. We used 2014–2016 inclusive as a training set for our forecasts. Our results show that the AL index provides the best forecasts and that

1. the deterministic forecasts of the flux exceeding $1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ that gives the highest Peirce score requires $AL < -103$ nT for 43 hr in the preceding 6 days;
2. the deterministic forecasts of the flux exceeding $10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ that gives the highest Peirce score requires $AL < -249$ nT for 29 hr in the preceding 6 days;
3. the best probabilistic forecasts of the flux $> 1,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ are based on the proportion of the preceding 4 days that $AL < -50$ nT;
4. the best probabilistic forecasts of the flux $> 10,000 \text{ cm}^{-2} \text{ sr}^{-1} \text{ s}^{-1}$ are based on the proportion of the preceding 4 days that $AL < -215$ nT.

Our simple deterministic forecasts have hit rates of up to 89% with high ROC scores (>0.8) and high Peirce scores (up to 0.57); however, our false alarm rates are also high (10–53%). Our deterministic forecasts do not perform as well as persistence forecasts of high-flux events. Compared to the results of Balikhin et al. (2016), our forecasts give slightly lower hit rates and higher false alarm rates than forecasts of high fluence; however, this may be improved with an optimised forecast tuning methodology. However, our forecasts are entirely independent of measurements of the electron flux at GEO. Overall, this demonstrates that forecasting the high energetic electron fluxes at geosynchronous orbit over periods of greater than a few days is possible with relatively high success rates based on observations solely from geomagnetic activity or upstream measurements.

Data Availability Statement

GOES flux and OMNI data used in this study were obtained from NASA's CDAWeb (<https://cdaweb.gsfc.nasa.gov>) using the SPDFCDAWEBCHOOSE IDL interface. GOES fluence was downloaded from this site (ftp://ftp.swpc.noaa.gov/pub/indices/old_indices/). The forecast data were created using the methods that are fully described in the text of this article and reproducible.

Acknowledgments

C. F. is supported by NERC Independent Research Fellowship NE/N014480/1. M. K. M. is supported by a UK NERC NPIF studentship 1926376. This research was supported by the Natural Environment Research Council (NERC) Highlight Topic Grant “Rad-Sat” NE/P017185/1 (MSSL), NE/P017274/1 (UoR), and NE/P01738X/1 (BAS) and NERC SWIMMR Grant NE/V002554/1. We thank the instrument teams, spacecraft operators, and data archives for their efforts in providing these data. Some of the figures within this paper were produced using IDL color-blind-friendly color tables (see Wright, 2017).

References

- Baker, D. N., Belian, R. D., Higbie, P. R., Klebesadel, R. W., & Blake, J. B. (1987). Deep dielectric charging effects due to high-energy electrons in earth's outer magnetosphere. *Journal of Electrostatics*, *20*(1), 3–19. [https://doi.org/10.1016/0304-3886\(87\)90082-9](https://doi.org/10.1016/0304-3886(87)90082-9)
- Baker, D. N., McPherron, R. L., Cayton, T. E., & Klebesadel, R. W. (1990). Linear prediction filter analysis of relativistic electron properties at 6.6 R_E . *Journal of Geophysical Research*, *95*(A9), 15,133–15,140. <https://doi.org/10.1029/JA095iA09p15133>
- Balikhin, M. A., Boynton, R. J., Walker, S. N., Borovsky, J. E., Billings, S. A., & Wei, H. L. (2011). Using the narmax approach to model the evolution of energetic electrons fluxes at geostationary orbit. *Geophysical Research Letters*, *38*, L18105. <https://doi.org/10.1029/2011GL048980>
- Balikhin, M. A., Rodriguez, J. V., Boynton, R. J., Walker, S. N., Aryan, H., Sibeck, D. G., & Billings, S. A. (2016). Comparative analysis of NOAA REFM and SNB³GEO tools for the forecast of the fluxes of high-energy electrons at GEO. *Space Weather*, *14*, 22–31. <https://doi.org/10.1002/2015SW001303>
- Barnston, A. G., Li, S., Mason, S. J., DeWitt, D. G., Goddard, L., & Gong, X. (2010). Verification of the first 11 years of IRI's seasonal climate forecasts. *Journal of Applied Meteorology and Climatology*, *49*(3), 493–520. <https://doi.org/10.1175/2009JAMC2325.1>
- Borovsky, J. E., & Denton, M. H. (2006). Differences between cme-driven storms and CIR-driven storms. *Journal of Geophysical Research*, *111*, A07S08. <https://doi.org/10.1029/2005JA011447>
- Borovsky, J. E., & Steinberg, J. T. (2006). The “calm before the storm” in CIR/magnetosphere interactions: Occurrence statistics, solar wind statistics, and magnetospheric preconditioning. *Journal of Geophysical Research*, *111*, A07S10. <https://doi.org/10.1029/2005JA011397>
- Boynton, R. J., Balikhin, M. A., & Billings, S. A. (2015). Online NARMAX model for electron fluxes at GEO. *Annales Geophysicae*, *33*(3), 405–411. <https://doi.org/10.5194/angeo-33-405-2015>
- Boynton, R. J., Balikhin, M. A., Billings, S. A., Reeves, G. D., Ganushkina, N., Gedalin, M., et al. (2013). The analysis of electron fluxes at geosynchronous orbit employing a NARMAX approach. *Journal of Geophysical Research: Space Physics*, *118*, 1500–1513. <https://doi.org/10.1002/jgra.50192>
- Boynton, R. J., Balikhin, M. A., Sibeck, D. G., Walker, S. N., Billings, S. A., & Ganushkina, N. (2016). Electron flux models for different energies at geostationary orbit. *Space Weather*, *14*(10), 846–860. <https://doi.org/10.1002/2016SW001506>
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- Burton, R. K., McPherron, R. L., & Russell, C. T. (1975). An empirical relationship between interplanetary conditions and DST. *Journal of Geophysical Research*, *80*(31), 4204–4214. <https://doi.org/10.1029/JA080i031p04204>
- Chen, Y., Reeves, G. D., Fu, X., & Henderson, M. (2019). PreMevE: New predictive model for megaelectron-volt electrons inside Earth's outer radiation belt. *Space Weather*, *17*(3), 438–454. <https://doi.org/10.1029/2018SW002095>
- Coleman, T., McCollough, J. P., Young, S., & Rigler, E. J. (2018). Operational nowcasting of electron flux levels in the outer zone of Earth's radiation belt. *Space Weather*, *16*(5), 501–518. <https://doi.org/10.1029/2017SW001788>
- Fok, M.-C., Horne, R. B., Meredith, N. P., & Glauert, S. A. (2008). Radiation belt environment model: Application to space weather nowcasting. *Journal of Geophysical Research*, *113*, A03S08. <https://doi.org/10.1029/2007JA012558>
- Forsyth, C., Rae, I. J., Murphy, K. R., Freeman, M. P., Huang, C.-L., Spence, H. E., et al. (2016). What effect do substorms have on the content of the radiation belts? *Journal of Geophysical Research: Space Physics*, *121*, 6292–6306. <https://doi.org/10.1002/2016JA022620>
- Freeman, M. P., & Morley, S. K. (2004). A minimal substorm model that explains the observed statistical distribution of times between substorms. *Geophysical Research Letters*, *31*, L12807. <https://doi.org/10.1029/2004GL019989>
- Glauert, S. A., Horne, R. B., & Meredith, N. P. (2014a). Simulating the Earth's radiation belts: Internal acceleration and continuous losses to the magnetopause. *Journal of Geophysical Research: Space Physics*, *119*, 7444–7463. <https://doi.org/10.1002/2014JA020092>
- Glauert, S. A., Horne, R. B., & Meredith, N. P. (2014b). Three-dimensional electron radiation belt simulations using the bas radiation belt model with new diffusion models for chorus, plasmaspheric hiss, and lightning-generated whistlers. *Journal of Geophysical Research: Space Physics*, *119*, 268–289. <https://doi.org/10.1002/2013JA019281>
- Glauert, S. A., Horne, R. B., & Meredith, N. P. (2018). A 30-year simulation of the outer electron radiation belt. *Space Weather*, *16*, 1498–1522. <https://doi.org/10.1029/2018SW001981>
- Gonzalez, W. D., Joselyn, J. A., Kamide, Y., Kroehl, H. W., Rostoker, G., Tsurutani, B. T., & Vasyliunas, V. M. (1994). What is a geomagnetic storm? *Journal of Geophysical Research*, *99*(A4), 5771–5792. <https://doi.org/10.1029/93JA02867>
- Hogan, R. J., & Mason, I. B. (2012). Deterministic forecasts of binary events. In *Forecast verification* (pp. 31–59). New York: John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781119960003.ch3>

- Horne, R. B. (2007). Acceleration of killer electrons. *Nature Physics*, 3, 590 EP. <https://doi.org/10.1038/nphys703>
- Horne, R. B., Glauert, S. A., Meredith, N. P., Boscher, D., Maget, V., Heynderickx, D., & Pitchford, D. (2013). Space weather impacts on satellites and forecasting the Earth's electron radiation belts with spacecast. *Space Weather*, 11, 169–186. <https://doi.org/10.1002/swe.20023>
- Horne, R. B., & Thorne, R. M. (1998). Potential waves for relativistic electron scattering and stochastic acceleration during magnetic storms. *Geophysical Research Letters*, 25(15), 3011–3014. <https://doi.org/10.1029/98GL01002>
- Horne, R. B., Thorne, R. M., Glauert, S. A., Albert, J. M., Meredith, N. P., & Anderson, R. R. (2005). Timescale for radiation belt electron acceleration by whistler mode chorus waves. *Journal of Geophysical Research*, 110, A03225. <https://doi.org/10.1029/2004JA010811>
- Horne, R. B., Thorne, R. M., Shprits, Y. Y., Meredith, N. P., Glauert, S. A., Smith, A. J., et al. (2005). Wave acceleration of electrons in the Van Allen radiation belts. *Nature*, 437, 227–30. <https://doi.org/10.1038/nature03939>
- Hsu, W.-R., & Murphy, A. H. (1986). The attributes diagram a geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, 2(3), 285–293. [https://doi.org/10.1016/0169-2070\(86\)90048-8](https://doi.org/10.1016/0169-2070(86)90048-8)
- Iucci, N., Levitin, A. E., Belov, A. V., Eroshenko, E. A., Ptitsyna, N. G., Villoresi, G., et al. (2005). Space weather conditions and spacecraft anomalies in different orbits. *Space Weather*, 3, S01001. <https://doi.org/10.1029/2003SW000056>
- Jaynes, A. N., Baker, D. N., Singer, H. J., Rodriguez, J. V., Loto'aniu, T. M., Ali, A. F., et al. (2015). Source and seed populations for relativistic electrons: Their roles in radiation belt changes. *Journal of Geophysical Research: Space Physics*, 120, 7240–7254. <https://doi.org/10.1002/2015JA021234>
- Juusola, L., Ostgaard, N., Tanskanen, E., Partamies, N., & Snekvik, K. (2011). Earthward plasma sheet flows during substorm phases. *Journal of Geophysical Research*, 116, A10228. <https://doi.org/10.1029/2011JA016852>
- Klimas, A. J., Vassiliadis, D., & Baker, D. N. (1998). Dst index prediction using data-derived analogues of the magnetospheric dynamics. *Journal of Geophysical Research*, 103(A9), 20,435–20,447. <https://doi.org/10.1029/98JA01559>
- Li, W., Shprits, Y. Y., & Thorne, R. M. (2007). Dynamic evolution of energetic outer zone electrons due to wave-particle interactions during storms. *Journal of Geophysical Research*, 112, A10220. <https://doi.org/10.1029/2007JA012368>
- Li, X., Temerin, M., Baker, D. N., Reeves, G. D., & Larson, D. (2001). Quantitative prediction of radiation belt electrons at geostationary orbit based on solar wind measurements. *Geophysical Research Letters*, 28(9), 1887–1890. <https://doi.org/10.1029/2000GL012681>
- Meredith, N. P., Horne, R. B., Sicard-Piet, A., Boscher, D., Yearby, K. H., Li, W., & Thorne, R. M. (2012). Global model of lower band and upper band chorus from multiple satellite observations. *Journal of Geophysical Research*, 117, A10225. <https://doi.org/10.1029/2012JA017978>
- Mourenas, D., Artemyev, A. V., & Zhang, X.-J. (2019). Impact of significant time-integrated geomagnetic activity on 2-MeV electron flux. *Journal of Geophysical Research: Space Physics*, 124, 4445–4461. <https://doi.org/10.1029/2019JA026659>
- Murphy, K. R., Mann, I. R., Rae, I. J., Sibeck, D. G., & Watt, C. E. J. (2016). Accurately characterizing the importance of wave-particle interactions in radiation belt dynamics: The pitfalls of statistical wave representations. *Journal of Geophysical Research: Space Physics*, 121, 7895–7899. <https://doi.org/10.1002/2016JA022618>
- Murphy, K. R., Watt, C. E. J., Mann, I. R., Jonathan Rae, I., Sibeck, D. G., Boyd, A. J., et al. (2018). The global statistical response of the outer radiation belt during geomagnetic storms. *Geophysical Research Letters*, 45, 3783–3792. <https://doi.org/10.1002/2017GL076674>
- Murray, S. A., Bingham, S., Sharpe, M., & Jackson, D. R. (2017). Flare forecasting at the Met Office Space Weather Operations Centre. *Space Weather*, 15, 577–588. <https://doi.org/10.1002/2016SW001579>
- Paulikas, G. A., & Blake, J. B. (1979). Effects of the solar wind on magnetospheric dynamics: Energetic electrons at the synchronous orbit. In *Quantitative modeling of magnetospheric processes* (pp. 180–202). Washington, DC: American Geophysical Union. <https://doi.org/10.1029/GM021p0180>
- Peirce, C. S. (1884). The numerical measure of the success of predictions. *Science*, ns-4(93), 453–454. <https://doi.org/10.1126/science.ns-4.93.453-a>
- Reeves, G. D., McAdams, K. L., Friedel, R. H. W., & O'Brien, T. P. (2003). Acceleration and loss of relativistic electrons during geomagnetic storms. *Geophysical Research Letters*, 30(10), 1529. <https://doi.org/10.1029/2002GL016513>
- Romanova, N. V., Pilipenko, V. A., Yagova, N. V., & Belov, A. V. (2005). Statistical correlation of the rate of failures on geosynchronous satellites with fluxes of energetic electrons and protons. *Cosmic Research*, 43(3), 179–185. <https://doi.org/10.1007/s10604-005-0032-6>
- Sharpe, M. A., & Murray, S. A. (2017). Verification of space weather forecasts issued by the Met Office Space Weather Operations Centre. *Space Weather*, 15, 1383–1395. <https://doi.org/10.1002/2017SW001683>
- Shin, D.-K., Lee, D.-Y., Kim, K.-C., Hwang, J., & Kim, J. (2016). Artificial neural network prediction model for geosynchronous electron fluxes: Dependence on satellite position and particle energy. *Space Weather*, 14, 313–321. <https://doi.org/10.1002/2015SW001359>
- Stefanova, L., & N. Krishnamurti, T. (2002). Interpretation of seasonal climate forecast using Brier skill score, the Florida State University superensemble, and the AMIP-I dataset. *Journal of Climate*, 15, 537–544. [https://doi.org/10.1175/1520-0442\(2002\)015h0537:IOSCFUj2.0.CO;2](https://doi.org/10.1175/1520-0442(2002)015h0537:IOSCFUj2.0.CO;2)
- Subbotin, D. A., & Shprits, Y. Y. (2009). Three-dimensional modeling of the radiation belts using the versatile electron radiation belt (VERB) code. *Space Weather*, 7, S10001. <https://doi.org/10.1029/2008SW000452>
- Summers, D., M. Thorne, R., & Xiao, F. (1998). Relativistic theory of wave-particle resonant diffusion with application to electron acceleration in the magnetosphere. *Journal of Geophysical Research*, 103, 20,487–20,500. <https://doi.org/10.1029/98JA01740>
- Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240(4857), 1285–1293. <https://doi.org/10.1126/science.3287615>
- Tang, C. L., Wang, Y. X., Ni, B., Zhang, J.-C., Reeves, G. D., Su, Z. P., et al. (2017). Radiation belt seed population and its association with the relativistic electron dynamics: A statistical study. *Journal of Geophysical Research: Space Physics*, 122, 5261–5276. <https://doi.org/10.1002/2017JA023905>
- Turner, D. L., & Li, X. (2008). Quantitative forecast of relativistic electron flux at geosynchronous orbit based on low-energy electron flux. *Space Weather*, 6, S05005. <https://doi.org/10.1029/2007SW000354>
- Usanova, M. E., Mann, I. R., Bortnik, J., Shao, L., & Angelopoulos, V. (2012). THEMIS observations of electromagnetic ion cyclotron wave occurrence: Dependence on AE, SYMH, and solar wind dynamic pressure. *Journal of Geophysical Research*, 117, A10218. <https://doi.org/10.1029/2012JA018049>
- Watt, C. E. J., Rae, I. J., Murphy, K. R., Anekallu, C., Bentley, S. N., & Forsyth, C. (2017). The parameterization of wave-particle interactions in the outer radiation belt. *Journal of Geophysical Research: Space Physics*, 122, 9545–9551. <https://doi.org/10.1002/2017JA024339>
- Wei, L., Zhong, Q., Lin, R., Wang, J., Liu, S., & Cao, Y. (2018). Quantitative prediction of high-energy electron integral flux at geostationary orbit based on deep learning. *Space Weather*, 16, 903–916. <https://doi.org/10.1029/2018SW001829>
- Wing, S., Johnson, J. R., Camporeale, E., & Reeves, G. D. (2016). Information theoretical approach to discovering solar wind drivers of the outer radiation belt. *Journal of Geophysical Research: Space Physics*, 121, 9378–9399. <https://doi.org/10.1002/2016JA022711>

- Wrenn, G. L., Rodgers, D. J., & Ryden, K. A. (2002). A solar cycle of spacecraft anomalies due to internal charging. *Annales Geophysicae*, 20(7), 953–956. <https://doi.org/10.5194/angeo-20-953-2002>
- Wrenn, G. L., & Sims, A. J. (1996). Internal charging in the outer zone and operational anomalies. In *Radiation belts: Models and standards* (pp. 275–278). Washington, DC: American Geophysical Union. <https://doi.org/10.1029/GM097p0275>
- Wrenn, G. L., & Smith, R. J. K. (1996). Probability factors governing esd effects in geosynchronous orbit. *IEEE Transactions on Nuclear Science*, 43(6), 2783–2789. <https://doi.org/10.1109/23.556867>
- Wright, P. J. (2017). ColourBlind: A collection of IDL colour-blind-friendly colour tables. Zenodo. Retrieved from <https://github.com/PaulJWright/ColourBlind> doi: 10.5281/zenodo.840393.
- Zhao, H., Baker, D. N., Jaynes, A. N., Li, X., Elkington, S. R., Kanekal, S. G., et al. (2017). On the relation between radiation belt electrons and solar wind parameters/geomagnetic indices: Dependence on the first adiabatic invariant and l^* . *Journal of Geophysical Research: Space Physics*, 122, 1624–1642. <https://doi.org/10.1002/2016JA023658>
- Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4), 561–577.