

1     **Predicting geostationary 40-150 keV electron flux using**  
2     **ARMAX (an autoregressive moving average transfer**  
3     **function), RNN (a recurrent neural network), and**  
4     **logistic regression: a comparison of models**

5     **L. E. Simms<sup>1,2</sup>, N. Yu. Ganushkina<sup>1,3</sup>, M. Van der Kamp<sup>3</sup>, M. Balikhin<sup>4</sup>, M.**  
6     **W. Liemohn<sup>1</sup>**

7                     <sup>1</sup>University of Michigan, Ann Arbor, USA

8                     <sup>2</sup>Department of Physics, Augsburg University, Minneapolis, USA

9                     <sup>3</sup>Finnish Meteorological Institute, Helsinki, Finland

10                    <sup>4</sup>University of Sheffield, Sheffield, UK

11     **Key Points:**

- 12     • Regression models incorporating interaction and quadratic terms predict electron  
13     flux as well as neural network models
- 14     • The description of time series behavior by ARMAX models, while useful for hy-  
15     pothesis testing, is not necessary for prediction
- 16     • MLT as a predictor improves the models by describing changing flux levels and  
17     the differing influence of parameters over the diurnal period

---

Corresponding author: L. Simms, laurasim@umich.edu

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as doi: [10.1029/2022SW003263](https://doi.org/10.1029/2022SW003263).

This article is protected by copyright. All rights reserved.

**Abstract**

We screen several algorithms for their ability to produce good predictive models of hourly 40-150 keV electron flux at geostationary orbit (data from GOES-13) using solar wind, IMF, and geomagnetic index parameters that would be available for real time forecasting. Value-predicting models developed using ARMAX (autoregressive moving average transfer function), RNN (recurrent neural network), or stepwise-reduced regression produced roughly similar results. Including magnetic local time (MLT) as a categorical variable to describe both the differing levels of flux and the differing influence of parameters improved the models ( $r$  as high as 0.814; Heidke skill score as high as 0.663), however value-predicting models did a poor job at predicting highs and lows. Diagnostic tests are introduced (cubic fit to observation-prediction relationship and Lag1 correlation) that better assess predictions of extremes than single metrics such as RMSE, MAE, or MSA. Classifier models (RNN and logistic regression) were equally able to predict flux rise above the 75th percentile (Heidke skill score as high as 0.667). Logistic regression models were improved by the addition of multiplicative interaction and quadratic terms. Only predictors from 1 or 3 hours before were necessary and a detailed description of flux time series behavior was not needed. Stepwise selection of these variables trimmed non-contributing parameters for a more parsimonious and portable logistic regression model that predicted as well as neural network-derived models. We provide a logistic regression model (LL3: LogisticLag3) based on inputs measured 3 h previous, along with optimal probability thresholds, for future predictions.

**Plain Language Summary**

As high levels of electrons in the radiation belts can damage satellites, accurate forecasting would be a useful tool. Electron levels can be predicted using information from the solar wind, the interplanetary magnetic field, and indices measuring disturbances in Earth's magnetic field. We compare several algorithms to produce such models: regression and neural networks that depend on predictors at one or many previous time steps. We find that dependable predictions can be made from a regression model using predictors from only a single previous time step. More sophisticated neural network techniques are not necessary if interaction and nonlinear terms are introduced to the regression.

**1 Introduction**

Electrons in the radiation belts can cause both internal and surface charging of spacecraft (e.g., Lam et al., 2012; Loto'aniu et al., 2015), with internal charging mainly due to  $>100$  keV (kiloelectronVolt) electrons and surface charging to electrons below 100 keV. However, while daily averaged  $>100$  keV electron fluxes can be reasonably well predicted because they often result from geomagnetic storms (e.g., Subbotin & Shprits, 2009; Pakhotin et al., 2014; Glauert et al., 2014; Balikhin et al., 2016; Simms et al., 2016), the same is not true of  $<100$  keV electrons. Not only do these lower energy electrons result in the more damaging surface charging, they are also much more difficult to forecast (e.g., Koons et al., 2000; Choi et al., 2011; Matéo-Vélez et al., 2018). For LANL (Los Alamos National Laboratory) satellites, for example, it is the energy range of  $\sim 10$ -50 keV that is most important for surface charging (Thomsen et al., 2013; Matéo-Vélez et al., 2018). These lower energy electrons vary on time scales of minutes with their distribution depending on location in the magnetosphere, so daily/orbit averaging is not possible. Moreover, geomagnetic storms are not always predictive of keV electron enhancements, and surface charging events have been detected during even weak to moderate substorm activity (Matéo-Vélez et al., 2018; Ganushkina et al., 2021).

Electron fluxes at keV energies have been modeled with several techniques, including a first principle kinetic approach in several ring current simulations (e.g., Fok et al., 2014; Ganushkina et al., 2014; Chen et al., 2015; Jordanova et al., 2016), empirical mod-

68 els using different fittings (e.g., Roeder et al., 2005; Sicard-Piet et al., 2008; Ginet et al.,  
69 2013; Denton et al., 2015, 2016; Sillanpää et al., 2017), and multivariate approaches in-  
70 cluding conditional mutual information (Stepanov et al., 2021) and Nonlinear AutoRe-  
71 gressive Moving Average with eXogenous (NARMAX) inputs (Boynnton et al., 2013, 2016,  
72 2019). However, these empirical models may depend on only a few parameters. A wider  
73 array of input parameters could improve predictions of keV electron fluxes. Solar wind  
74 and IMF (Interplanetary Magnetic Field) parameters alone may produce reasonable pre-  
75 dictive models, with the advantage that these would be readily available for real time  
76 forecasting.

77 Several studies have examined the response of geosynchronous keV electron flux  
78 to solar wind parameters, with electron enhancements associated with pressure increases  
79 (Shi et al., 2009) or higher solar wind speed (Li et al., 2005; Kellerman & Shprits, 2012;  
80 Hartley et al., 2014). A combination of solar wind speed and the IMF  $B_z$  has been found  
81 to be predictive as well (Sillanpää et al., 2017), with lesser influence from the other two  
82 IMF components and solar wind density, temperature, and pressure (e.g., Li et al., 2005;  
83 Kellerman & Shprits, 2012; Ganushkina et al., 2019). This suggests that combinations  
84 of parameters, whether multiplicative or additive, may best predict flux, reflecting mul-  
85 tiple driving parameters (Denton et al., 2016). As keV electrons levels fluctuate on time  
86 scales of hours, better models may come from prediction parameters at a similar cadence.

87 Higher energy electrons (MeV; MegaelectronVolt), when daily averaged, have shown  
88 high correlations with solar wind parameters (wind speed and density either individu-  
89 ally or in combination) (e.g., Blake et al., 1997; Paulikas & Blake, 1979; Li et al., 2001;  
90 Reeves et al., 2011; Lyatsky & Khazanov, 2008; Balikhin et al., 2011). However, the hourly  
91 response may be much lower (Simms, Engebretson, & Reeves, 2022), and the physical  
92 influence of many solar wind drivers on even MeV electron flux may not be as high as  
93 these correlations suggest. Much of the solar wind influence may not be direct but in-  
94 stead mediated by waves and electron injections following substorms (e.g., (Simms et al.,  
95 2018a)), and simple correlations of solar wind parameters with electrons may be inflated  
96 by common cycles and trends if these commonalities are not removed via such methods  
97 as a differencing transformation or ARMAX modelling (Simms, Engebretson, & Reeves,  
98 2022). However, for prediction purposes, it may not be important that variables phys-  
99 ically drive keV electron flux, nor that the correlations are only due to mutual cycles.  
100 Highly correlated proxies may be sufficient for prediction, and more practical given their  
101 real-time availability. For keV electrons, the strongest solar wind correlates are some com-  
102 bination of velocity, density and pressure (Ganushkina et al., 2019; Simms, Ganushkina,  
103 et al., 2022). IMF  $B_z$ , while it does not show as high a correlation as solar wind veloc-  
104 ity, may still be a useful addition as it provides further information not present in the  
105 solar wind parameters alone. (The southward component of IMF ( $B_s$ ) may appear to  
106 be a more targeted version of this parameter and therefore likely of more predictive use,  
107 but we have found that  $B_s$  does not correlate better with flux than  $B_z$  itself, at least in  
108 hourly data (Simms, Ganushkina, et al., 2022)).

109 Geomagnetic indices are easy to obtain measures that have often been used in pre-  
110 diction models. Although there may be concern that ground-based indices (measured  
111 at ground magnetometers), may not represent conditions in the magnetosphere well, they  
112 are worth testing as possible predictors that contain, at least, some information that we  
113 do not have access to otherwise. Bearing in mind that they may be proxies of pertinent  
114 physical processes that all manifest as magnetic perturbation in a single number, we can  
115 still use these for prediction purposes. However, although the  $AE$  (Auroral Electrojet)  
116 index may be a reasonable measure of substorm activity that correlates well with keV  
117 electrons (Ganushkina et al., 2021) due to its ability to indicate electron injections, it  
118 is not useful for real time predictions because it is not published immediately. If we can-  
119 not use  $AE$ , two other indices,  $Kp$  (Planetary Kennziffer) and  $SymH$  (symmetric  
120 H-component of the ground magnetic field, or  $Dst$  (Disturbance Storm-Time), show sim-

121 ilarly high simple correlations with flux. Neither  $Kp$  nor  $SymH$  show as much associ-  
122 ation with flux as  $AE$  does when all 3 of these indices are included in the same analy-  
123 sis, but they may be a practical second choice for prediction purposes.  $Kp$  correlates well  
124 with 1-40 keV flux (e.g., Freeman, 1974; Korth et al., 1999; Thomsen et al., 2013; Den-  
125 ton et al., 2015, 2016). Its 3 hour cadence may make it too slow to measure quick changes  
126 in geomagnetic activity that may be associated with fast electron enhancements, but its  
127 inclusion in a prediction model may be helpful to measure the general background level  
128 of disturbance.  $SymH$  would be the obvious choice as it is reported at a 1-min cadence,  
129 but it is not currently available in real time for prediction purposes. Given this prob-  
130 lem,  $Dst$  may be the best geomagnetic index parameter to include in a prediction model  
131 as it is available in real time and at an hourly cadence (improving over the 3 h  $Kp$  ca-  
132 dence). As  $SymH$  is essentially the  $Dst$  index at finer time resolution (minute vs hourly),  
133 the choice of  $Dst$  over  $SymH$  should make no difference in the prediction of hourly elec-  
134 tron flux (Iyemori et al., 2010). It also may be useful to incorporate the solar energy flux  
135 (f10.7) even though it changes relatively slowly.

136 Previous work has also explored the effect of polynomial (Balikhin et al., 2011) and  
137 polynomial and multiplicative interaction terms (Simms et al., 2018b). The quadratic  
138 (square) and cubic terms of predictors can account for possible nonlinear effects that are  
139 not dealt with by log transformations, while multiplicative interaction terms describe the  
140 synergistic effects of variable pairs. Polynomial and multiplicative terms such as this will  
141 either be incorporated automatically by a neural network approach, if the algorithm finds  
142 them useful, or can be included as additional terms in ARMAX or regression models.

143 In this study, we explore the ability of several multivariable prediction model types  
144 to predict electron flux that have been used at various electron energies: neural networks  
145 (for example, Simms and Engebretson (2020); Koons and Gorney (1991); Ling et al. (2010);  
146 Freeman et al. (1998); Smirnov et al. (2020)); Ma et al. (2022); Katsavrias et al. (2022);  
147 Chu et al. (2021); Swiger et al. (2022)), autoregressive moving average time series trans-  
148 fer functions (ARMAX) (Balikhin et al., 2011; Boynton et al., 2013, 2015; Simms et al.,  
149 2018a; Simms & Engebretson, 2020), conventional regression (value-predicting) (Simms  
150 et al., 2014, 2016) , and logistic regression (which classifies predictions into groups) (Simms  
151 & Engebretson, 2020; Capman et al., 2019; Neter et al., 1990).

152 In the present paper, we explore the capabilities of three approaches, namely, re-  
153 current neural networks (RNN), ARMAX and conventional and logistic regression, to  
154 model hourly electron fluxes with energies of 40-150 keV as observed at geostationary  
155 GOES-13 satellite using solar wind, IMF, and geomagnetic indices as parameters. Mod-  
156 els may be either value-predicting (RNN, ARMAX, conventional regression), or predict  
157 the probability of being over a given threshold value (RNN, logistic regression). Either  
158 ARMAX or conventional regression values output can also be categorized as above or  
159 below a threshold (although they do not predict probability). We also take the oppor-  
160 tunity to briefly compare the power of several single-value metrics to distinguish between  
161 model prediction ability. (However, for a more comprehensive comparison see Liemohn  
162 et al. (2021).) We note that these single-value metrics are heavily weighted by mid-range  
163 values and are not well suited to assessing how well a model predicts the high electron  
164 fluxes that are of most interest. We propose several other assessment techniques, but this  
165 is not the main focus of this study which, instead, seeks to determine whether or not model  
166 predictions can be improved by various methods.

167 Section 2 gives a brief description of the GOES-13 MAGED data used in this study.  
168 Section 3 outlines the steps for building the models used in the study. Assessing and val-  
169 idation results of the three models' outputs over the GOES-13 MAGED data are pre-  
170 sented in Section 4 including predictions above the threshold for model comparison. Sec-  
171 tion 5 is devoted to the building of probability prediction models which can give more  
172 accurate predictions than models predicting flux values. The obtained results are dis-  
173 cussed in Section 6 and the conclusions are drawn in Section 7.

## 2 Data

We use hourly averaged electron fluxes (centered at midpoints of 40, 75, and 150 keV) from the geostationary GOES-13 satellite. Directional differential electron fluxes ( $cm^{-2}.s^{-1}.sr^{-1}.keV^{-1}$ ) from the nine collimated solid state telescopes of the MAGED instrument (e.g., Rowland and Weigel (2012)) each have a  $30^\circ$  full-angle conical field of view. We compute one omnidirectionally averaged flux (flight direction-integrated differential electron flux) for each of the energies using pitch angles calculated from the GOES Magnetometer 1 data following the method presented in Sillanpää et al. (2017) and Ganushkina et al. (2019). The GOES-13 MAGED data of electron fluxes and the data for the pitch angles of each telescope are available at

<https://www.ncei.noaa.gov/data/goes-space-environment-monitor/access/full/>

We use data covering 10 June 2013 - 6 August 2016 to build the models (the training set) and the 7 August 2016 - 12 December 2017 period for validation (the test set). There were minimal data gaps of only several hours during these time periods. These gaps were filled using linear interpolation between existing observations. This was necessary for the ARMAX models which require complete time series. Because the ARMAX models require a continuous time period for each of both the training and validation sets, cross-validation using a number of randomly selected sets out of the data is not possible. Therefore, to compare model performance on the same data, models were all built on the same training set and validated on the same withheld test set. Due to data availability, the models are built on observations from the solar cycle peak moving into the declining phase, but validation is performed on a withheld test set from further in the declining phase. This could potentially reduce the effectiveness of predictions if electron flux response to solar wind, IMF, and geomagnetic parameters were to vary over the solar cycle. While the average levels of these parameters vary over the solar cycle, we are unaware of any evidence suggesting that the flux response to a given level changes. However, given this possibility, further work should attempt validation of these models with periods during different phases of the solar cycle.

Solar wind parameters (solar wind velocity  $V$ , number density  $N$ , pressure  $P$ , the solar flux f10.7 index (*SolarFlux*), IMF  $B_z$  and electric field  $E_y$ , and magnetic indices ( $Kp$  and  $Dst$ ) were obtained from OMNIWeb

(<https://omniweb.gsfc.nasa.gov/form/dx1.html>) with 1 h resolution with data time-shifted to the bow shock nose.

We take  $\log_{10}$  of all variables  $\geq 0$ . Variables containing zero values which cannot be logged without creating missing values (i.e.,  $Kp$ ) were transformed by adding 1 to all values before the log transformation.  $B_z$  and  $E_y$ , as they have both positive and negative values, were not logged. A log transformation of electron flux data linearizes the relationship between predictors and response, allowing the use of techniques that assume this such as regression and neural networks (Simms, Ganushkina, et al., 2022). This transformation reduces skewness, inequality of variances among groups, and the non-normality of residual errors, all of which would make the use of linear models invalid. Examination of residual plots of the linear ARMAX and regression models (not shown) showed that this transformation fixed all these problems.

Because the dependent variable (electron flux) is log-transformed, these models will describe a nonlinear relationship between flux and all the variables: a power function relationship for those predictor variables that are also log-transformed, and an exponential function relationship for those predictor variables that are not logged. Subsequent to the log transformation, all variables were standardized by subtracting that series mean and dividing by its standard deviation. This creates unitless variables (Z-scores) for which regression coefficients (slopes) can be directly compared (Neter et al., 1990) but is also necessary for both efficient convergence and accuracy of prediction in neural networks (Alpaydin, 2014). An additional benefit is that scaling all output variables to the same

**Table 1.** Means and standard deviations used to calculate Z scores.

	Mean	Std Dev
log 40 keV Flux	4.5066	0.4152
log 75 keV Flux	4.2019	0.3954
log 150 keV Flux	3.6587	0.4241
log B	0.7458	0.186
Bz	0.0393	3.1674
Ey	-0.0186	1.3692
log N	0.7906	0.2826
log V	2.611	0.0868
log P	0.2362	0.2691
log (Kp+1)	0.4017	0.1995
Dst	-11.4803	17.4235
log Solar Flux	2.0805	0.0963

standard deviation allows direct comparison of metrics such as the RMSE between models and output variables that might otherwise show differences in the metrics only due to different scaling. The means and standard deviations are given so that readers can backtransform to the actual flux and predictor values if desired (Table 1).

ARMAX models were developed in IBM SPSS Statistics (formerly known as the Statistical Package for the Social Sciences). RNN and regression (as well as logistic regression) models were developed in MATLAB.

### 3 Building Value-Predicting Models: ARMAX, RNN, and Regression

One of the more popular classes of prediction model algorithm are neural networks. As we are working with time series data, we have chosen a type specific to this type of data: an LSTM RNN model (Long Short Term Memory- Recurrent Neural Network) (Hochreiter & Schmidhuber, 1997). This type of model uses an input sequence (e.g., we use the 48 h previous of each predictor variable), with the LSTM layer “learning” the long term time dependencies between time steps. Pathways can also be “forgotten” if they are determined to contain little information, giving a more parsimonious and less overfitted model. This model type can produce either values or classification output depending on the output layer chosen. This allows us to compare output validation to either a value-output model (such as ARMAX or conventional regression) or to a probability (classification) model such as logistic regression (see below). RNN models (or any neural network) automatically test more than just the main effects of each predictor variable. The algorithm will also test multiplicative interactions between variables and polynomial terms, describing the nonlinear relationships more completely. We also attempt to refine the predictions from the RNN models by creating a different model for each magnetic local time (MLT).

ARMAX models incorporate terms to describe the time series behavior of the dependent variable (autoregressive (AR) and moving average (MA) terms), as well as exogenous predictor variables, the transfer function (represented by X) (Hyndman & Athana-

sopoulos, 2018; Simms et al., 2019). The AR and MA terms are chosen to represent the cyclical behavior such as the daily variations in flux due to the satellite orbit. Predictor variables can be limited to the standard main effect of each parameter, or, additionally, include terms to describe the polynomial response (such as in the NARMAX models of Balikhin et al. (2010) and Boynton et al. (2011)), or synergistic action between predictors (multiplicative interaction effects), or decay terms to describe the influence from time steps in the past. The response variable can also be differenced, a transformation where each observation is subtracted from itself (e.g.  $y_t - y_{t-1}$ ), in which case the model would be called an ARIMAX model. However, we did not find this to be a necessary transformation for this data once the time series was described with appropriate AR and MA terms. The ARMAX or ARIMAX model formulation is useful for removing cycles that can result in spurious correlations between variables and therefore avoiding erroneous conclusions about the physical driving of a system (Simms, Engebretson, & Reeves, 2022). However, the method has also been suggested as a means to better predict electron flux (Balikhin et al., 2011; Boynton et al., 2013, 2015). The output of an ARMAX model will be values, although these can be categorized if classification is desired. In the ARMAX models, we include both an influence term (from 1 h previous) and a decay term for each solar wind, IMF, and magnetospheric input variable (Hyndman & Athanasopoulos, 2018). This incorporates the continuing effect of each variable over previous hours.

Multiple regression can also be used to predict electron flux. In the simplest case, regression models can be a model of main effects. However, polynomial terms for each predictor, and the multiplicative terms between them, can also be entered to describe the variation more fully. There is also the possibility of entering predictors from many previous time steps, similar to the RNN procedure. This may lead to overfitting and an unnecessarily complicated model, but, similar to the RNN “forgetting” of inessential pathways, stepwise regression can be applied to a logistic regression model to remove predictors that do not contribute explanatory power. In this method, predictors are added or removed one by one, checking whether this improves the model at each step. The stepwise procedure is an improvement over backward elimination used previously by Camporeale et al. (2022) as it also incorporates forward selection, giving variables the opportunity to be selected at a later stage if they were eliminated prematurely. While the indiscriminate use of stepwise procedures to identify physical drivers is problematic (Whittingham et al., 2006; Smith, 2018), the same concern does not apply when developing predictive models. With a prediction model we are only concerned with the result (the prediction) and not whether the variables used to make that prediction are physically meaningful. Logistic regression, if given the same variables to work with and if reduced by stepwise regression, may find essentially the same relationships as RNN and thus be just as good at prediction. However, a regression model will be the most portable of these three model types as the coefficients can be easily printed or coded without the need for the end user to have access to the particular software the model was developed in.

Both ARMAX and RNN empirical models would appear to have an advantage over more conventional multiple regression models. RNN has the ability to incorporate predictor values from many previous time steps while with ARMAX models the time behavior of the dependent variable is modeled using AR and MA terms. This modelling of past behavior or associations would, hopefully, improve the predictions. However, it is possible that electron flux, particularly below 200 keV, is not dependent on the long term states of the magnetosphere or solar wind, or that these states are long lasting enough that correlation to just a few previous time steps holds enough information to create an accurate prediction. If that is the case, then conventional regression models should perform just as well for predictions.

We incorporate a number of solar wind, IMF, and magnetospheric variables, the only constraint being that they must be available in real time for predictions. We also, to some of the ARMAX and regression models, add a variable to identify MLT. In pre-

vious work, ARMAX models have been built for each MLT (Boynton et al., 2019). We modify this approach by providing a model for the entire time series, but with MLT as a categorical variable. In practice, this is actually a set of 23 indicator variables (0 or 1), one less than the number of hours. The coefficients of these add or subtract to the constant term of the model to describe variations in flux related to MLT. We also add multiplicative interaction terms between each of the other (continuous) variables and these indicator variables. These interaction coefficients describe how the slope or association of that continuous variable with flux changes over the 24 h of the day. The use of indicator variables essentially creates a different model for each level of the categorical variable but makes more effective use of the available information in the data.

We train each of the model types above to predict electron flux values (value-predicting models). These can be validated by correlating predictions with observations in the test dataset. Another method of evaluation is to identify flux events (e.g.,  $\geq 75th$  or  $90th$  percentiles) and categorize output predictions into above or below these cut offs. The ability of the model to distinguish event from non-event can then be assessed with a Heideke skill score (see below for calculation details) which compares predictions to a null hypothesis of random assignment to classes.

## 4 Assessing the Value-Predicting Models

### 4.1 Assessing the 40 keV ARMAX models

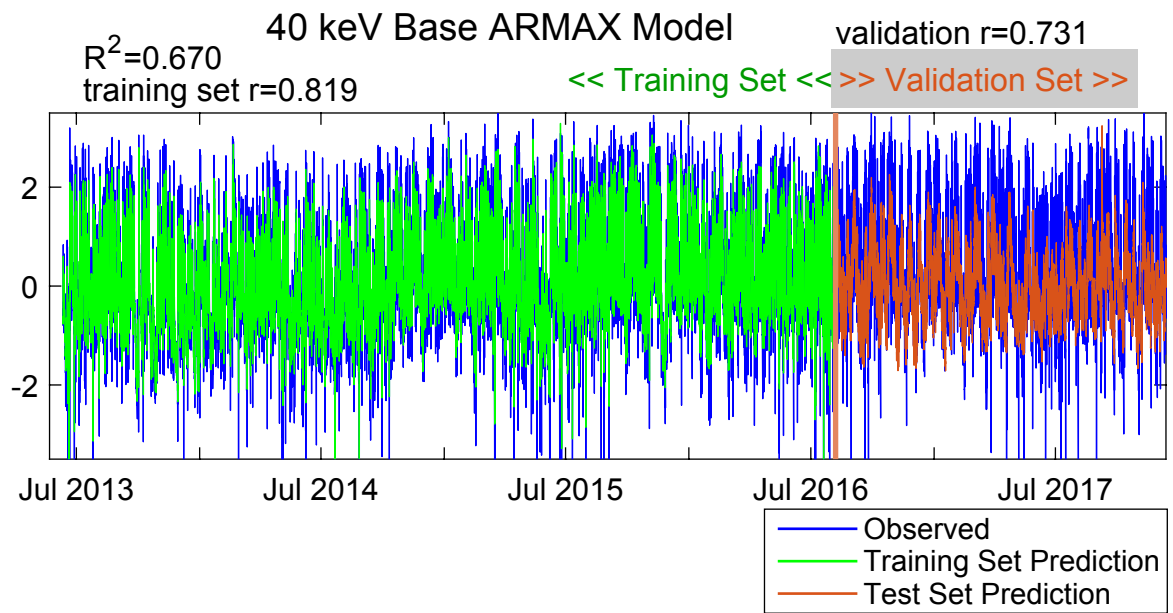
We start our analysis with predictions of 40 keV electron fluxes. Figure 1 presents the predictions of the observed (shown in blue) 40 keV electron fluxes using the base ARMAX model over the training set (shown in green) and the validation period (shown in orange). (To facilitate comparison of the influence of predictors with widely different units, we use Z-scores obtained by subtracting the mean of each series and dividing by its standard deviation.) The base ARMAX model validation  $r = 0.731$  appears to give a reasonable fit to observations in the validation (test) set. The training set  $r = 0.819$ , when squared, gives an  $R^2 = 0.670$ , showing that a reasonable fraction (67.8%) of the variability is captured by the ARMAX model. (The  $R^2$  is mathematically equivalent to the prediction efficiency, or PE, used in some other work.) However, a timeplot of observed and expected points reveals that the model does a poor job of predicting the high and low extremes in the validation set compared to the training set (Figure 1).

Figure 2 shows a more detailed view of just the validation period. Predicted values from the base ARMAX model do follow the general rise and fall of observed flux (Figure 2a; further metric scores are given in Figure 4 and Table 2. However, this expanded view shows more clearly that the base model has more trouble predicting high values than low values. Adding Lag 1 flux (40 keV flux from the previous hour) as a predictor provides some improvement (Figure 2b; validation  $r = 0.822$ ), but we will discuss below why this is not an optimal approach.

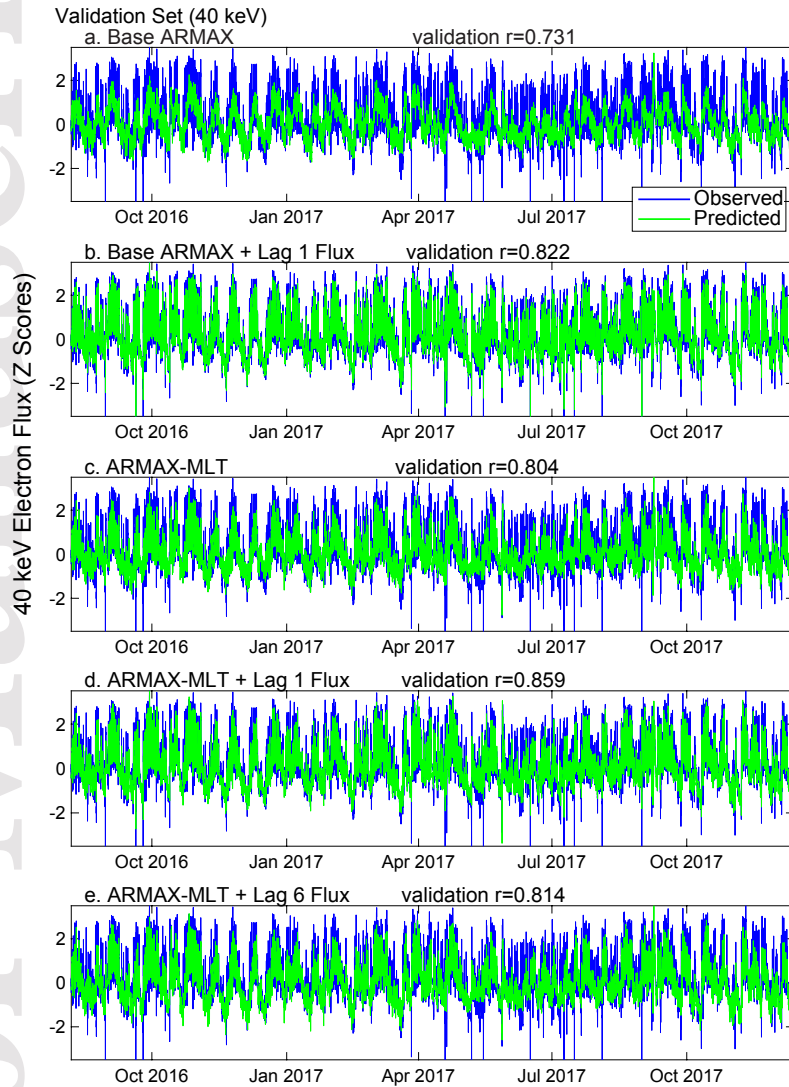
The model is improved in both validation correlation (0.804) and apparent ability to predict highs and lows by adding MLT as a categorical variable to account both for the differing levels of flux and the differing response of flux to the other predictors throughout the diurnal period (Figure 2c; ARMAX-MLT model). An improvement in validation correlation (0.859) can be achieved by adding 40 keV flux from 1 h previous (Lag 1 flux) as a predictor to the ARMAX-MLT model (Figure 2d). Less improvement (validation  $r = 0.814$ ) is seen if 40 keV flux 6 h previous (Lag 6 flux) is added instead (Figure 2e).

However, further assessment of these models at a finer scale reveals that adding previous flux from an hour before as an explanatory variable, while increasing the validation correlation, causes predictions to lag behind observations. Figure 3 shows the predictions over one week of the validation period (6-12 December 2016). The base ARMAX





**Figure 1.** Predictions from the Base ARMAX model (40 keV) over the training set (predictions in green) and the validation period (predictions in orange). Flux is converted to unitless Z-scores.



**Figure 2.** Predictions over the validation period (40 keV) from a. Base ARMA model, b. ARMAX-MLT, c. ARMAX-MLT with flux at lag 6 added as a predictor, d. ARMAX-MLT with flux at lag 1 added as a predictor, e. ARMAX-MLT with flux at lag 6 added as a predictor. Flux is converted to unitless Z-scores. Further metrics are given in Figure 4 and Table 2.

356 model prediction tracks the pattern of the observations, although always lower at the  
 357 peaks (Figure 3a). Adding the Lag 1 flux to the base ARMAX model appears to track  
 358 the height of the peaks better and to improve the validation correlation (0.822), but vi-  
 359 sually we can see that these predictions lag behind by 1 h (Figure 3b). This results in  
 360 predictions that appear very good, but only an hour after we already know what the flux  
 361 was. More quantitatively, we can assess models for this delay problem by comparing the  
 362 same time vs. the Lag 1 validation correlation. In the model including Lag 1 flux, the  
 363 correlation between observations and the prediction 1 h later (0.978) is much higher than  
 364 between observation and prediction from the same time step (0.822). In contrast, the  
 365 Lag 1 validation for the base ARMA model is lower than the same time validation. Note  
 366 that this delay in prediction is only due to the introduction of Lag 1 flux, as all other  
 367 parameters are the same between these two models.

368 Over this one week, the ARMAX-MLT model tracks the observed peaks somewhat  
 369 better, without a delay (the overall Lag 1 validation is lower than the same time vali-  
 370 dation) (Figure 3c). Adding Lag 1 flux to the ARMAX-MLT model does not as obvi-  
 371 ously introduce a delay in this one week, but over the entire test period, the Lag 1 va-  
 372 lidation correlation is still slightly higher than the same time validation (Figure 3d). Al-  
 373 though the ARMAX-MLT+Lag1Flux model does appear to somewhat improve the abil-  
 374 ity to reproduce the peaks, we cannot guarantee that flux from one hour previous would  
 375 be available for real time forecasting. However, the more important shortcoming is that  
 376 we are likely most interested in those occasions when flux rises sharply and unexpect-  
 377 edly. The very predictions we are most interested in are the ones that will fail to be pre-  
 378 dicted until an hour after the occurrence.

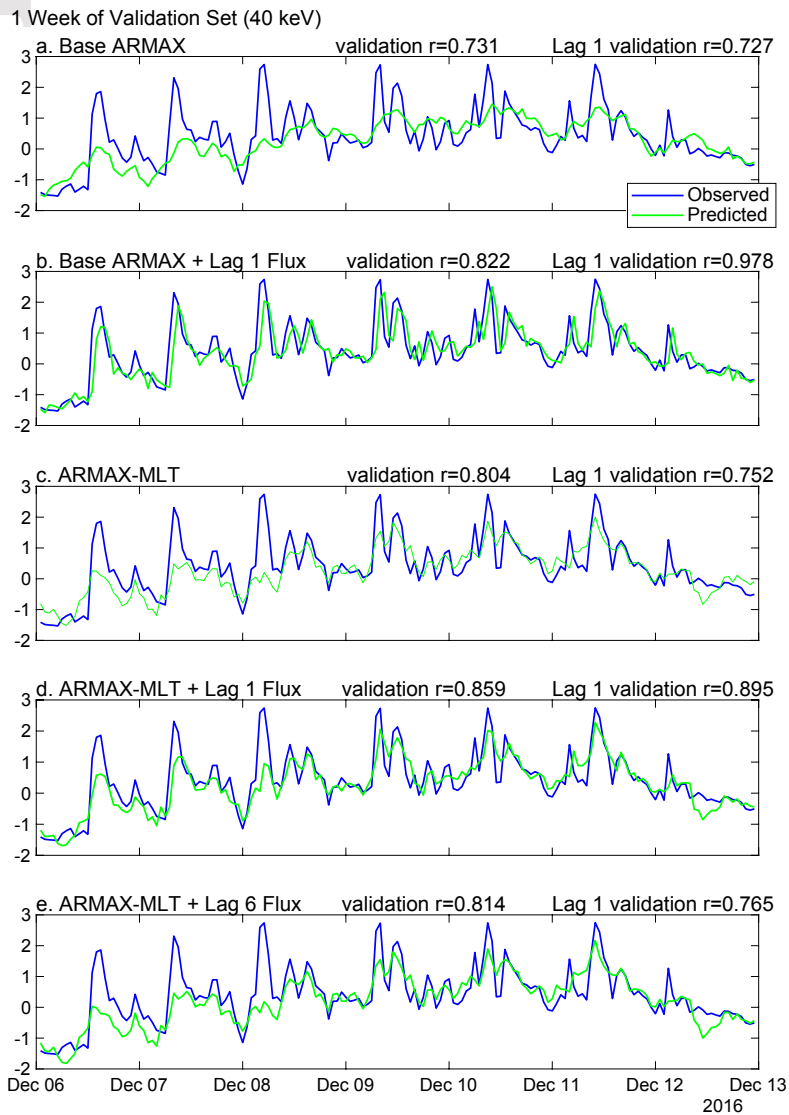
379 We consider a compromise model, the ARMAX-MLT+Lag6Flux model (Figure 3e),  
 380 in hopes that this will predict peaks better but without the disadvantage of a delay. How-  
 381 ever, while this does not show a delay in predictions (same time validation  $r = 0.814$   
 382 vs Lag 1 validation  $r = 0.765$ ), the overall validation correlation of 0.814 is not much  
 383 above the ARMAX-MLT model alone (0.804), and peak prediction is not improved.

384 Thus, of the ARMAX models at 40 keV, the ARMAX-MLT is the best model in  
 385 that it correlates reasonably well with observations and does not show a delay. However,  
 386 we would hope for a model that is able to predict the peaks better.

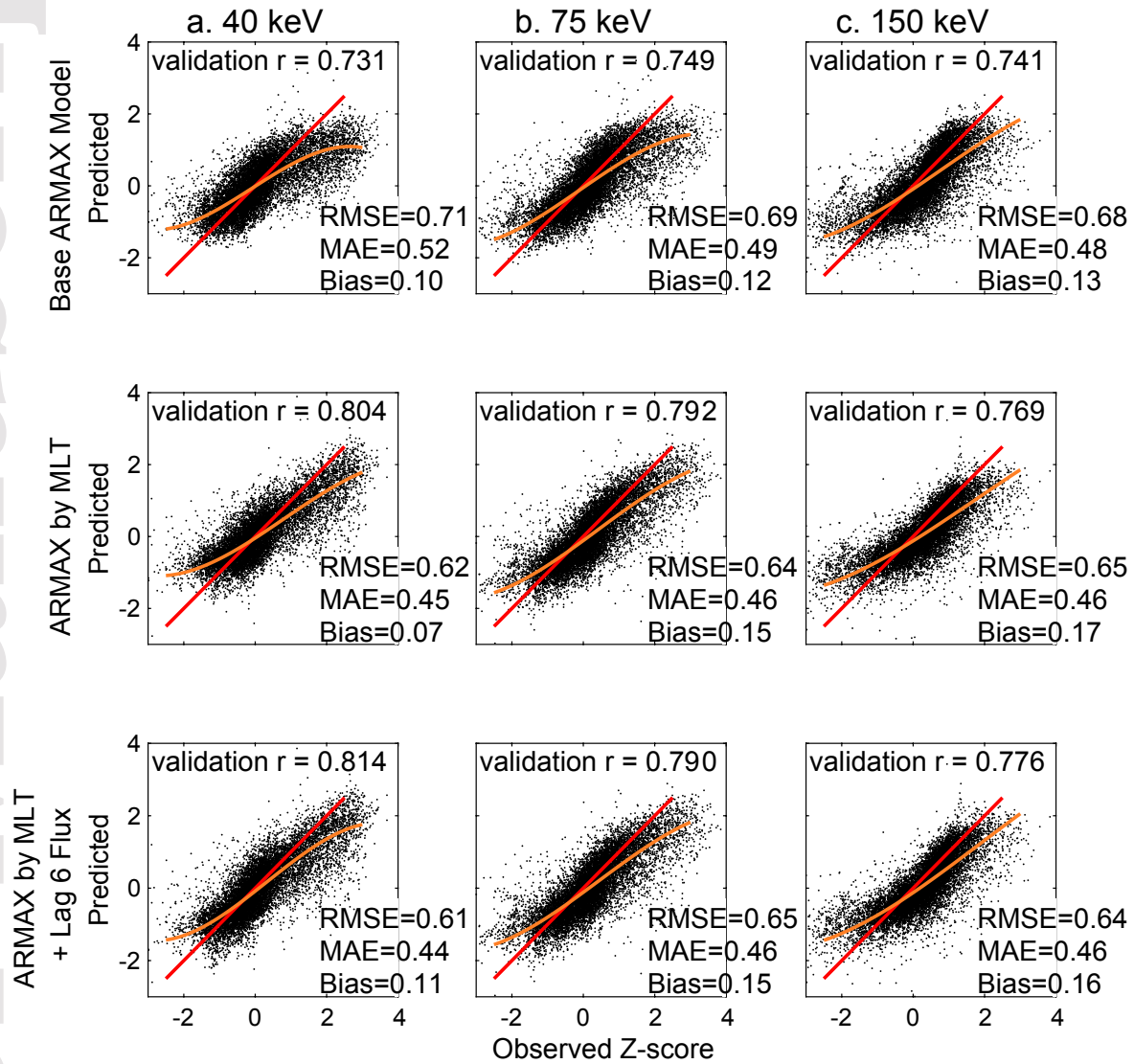
## 387 4.2 Validation of 40-150 keV ARMAX Models

388 Scatterplots of observations vs. predictions from several ARMAX models give more  
 389 information (Figure 4). We use scatterplots, rather than 2-D density plots, because the  
 390 most important information (the over and under prediction of the low and high flux) is  
 391 contained in the areas of lowest point density. Using 40 (column of three plots in Fig-  
 392 ure 4a), 75 (column of three plots in Figure 4b) and 150 keV (column of three plots  
 393 in Figure 4c) electron fluxes, only over the validation period, we show a number of di-  
 394 agnostics. In each plot, the red line shows the ideal 1:1 relationship between prediction  
 395 and observation. Many of the scatterplots not only show a great deal of scatter around  
 396 this idealized 1:1 line, but also non-linearity between observed and predicted values. Points  
 397 above the red line in the lower left or below the red line in the upper right represent pre-  
 398 dictions that failed to reach the lows and highs, respectively, in the observed data. The  
 399 orange line is the cubic fit to these deviations and can be used to roughly assess how se-  
 400 rious this problem is for each model. In the best case, the orange cubic line would lie on  
 401 the 1:1 line, showing a model that reproduced the peaks and valleys well. The base AR-  
 402 MAX model at 40 keV is a model that does particularly poorly at this task. The orange  
 403 cubic fit in this case veers radically from the 1:1 red line. Both valley and peak magni-  
 404 tudes are severely under-predicted.

405 We report RMSE (root mean square error, or the standard deviation of the pre-  
 406 diction vs observation residuals over the test set), and the MAE (mean absolute error,



**Figure 3.** Predictions over 1 week of the validation period (40 keV) showing how the models incorporating Lag 1 flux lag behind observations. a. Base ARMA model, b. Base ARMA with lag 1 flux, c. ARMAX-MLT, d. ARMAX-MLT with flux at lag 1, e. ARMAX-MLT with flux at lag 6. Flux is converted to unitless Z-scores. Lag 1 validation  $r$  correlates current observations with the prediction one hour previous. Models with Lag 1 flux as a predictor have higher Lag 1 validation correlation than same time validation correlation.



**Figure 4.** Scatterplots of predictions vs. observations over the full validation period and all three energies (a. 40 keV, b. 75 keV, c. 150 keV). Row 1: Base ARMA model, Row 2: ARMA model split by MLT (ARMAX-MLT), Row 3: ARMAX-MLT with lag 6 flux added as an additional predictor. Red line shows the ideal 1:1 correspondence between predictions and observations. Orange line gives the cubic fit to the actual prediction-observation relationship. Flux is converted to unitless Z-scores.

**Table 2.** Median symmetric accuracy (MSA) and symmetric signed percentage bias (SSPB) metrics for the value-predicting models (built on Z-score transformed data).

	40 keV		75 keV		150 keV	
	MSA	SSPB	MSA	SSPB	MSA	SSPB
Base ARMAX	5.70%	-0.20%	4.10%	-0.70%	4.00%	-1.50%
ARMAX-MLT	5.00%	-0.80%	3.80%	-1.20%	3.70%	-1.40%
ARMAX-MLT+Lag6	5.00%	-0.80%	3.80%	-1.20%	3.70%	-1.90%
REG-MLT	5.10%	-0.40%	3.90%	-0.80%	4.00%	-1.10%
Base RNN	5.40%	-0.70%	3.70%	-0.10%	3.30%	-0.80%
RNN-MLT	4.90%	-0.10%	3.70%	-0.004%	3.20%	-0.10%

407 or the average of the absolute differences between prediction and observation). The MAE  
408 is less sensitive to outliers, as the differences are not squared. (It is the same measure  
409 as the MAD, or mean absolute deviation.) Bias is the average of the differences of pre-  
410 diction and observation, without taking the absolute value. This measures whether pre-  
411 diction tend to lie above or below the observations. All these measures will indicate bet-  
412 ter fit when they are lower in magnitude (Hyndman & Athanasopoulos, 2018). Note that  
413 the standard deviation of the datasets is the same for all models as all use the same data  
414 (with standard deviation = 1 as these are Z-scores). Additionally, we calculate median  
415 symmetric accuracy (MSA) and symmetric signed percentage bias (SSPB) metrics (Ta-  
416 ble 2) (Morley et al., 2018). These two metrics appear to improve on the commonly used  
417 MAPE (mean absolute percentage error, which we do not use) by reducing the influence  
418 of outliers. However, this assumes that outliers do not carry relevant information, and  
419 in the case of our models, it is the points lying outside the main cloud that are of most  
420 interest to us, both because they lie in the regions of most interest (high or low flux) and  
421 because it is important to flag these areas where the predictions fail. Metrics based on  
422 the median of the error (MSA) reduce the influence of the error outliers, the rarer situ-  
423 ations where the model performs the worst. RMSE, based on the mean of the errors,  
424 is more influenced by outliers and is therefore the more appropriate metric for assess-  
425 ing model failures. As MSA and SSPB use the ratio of prediction to observation rather  
426 than the difference, they may be better for data where the error variance increases or  
427 decreases with magnitude (heteroscedasticity) (Tofallis, 2015). However, as we have al-  
428 ready dealt with the increasing error variance problem by taking the log and then the  
429 Z-score of the variables, the use of these metrics is redundant in that regard.

430 We should also point out a potential difficulty of MSA and SSPB with our partic-  
431 ular data set. As these two metrics take the log of the ratio between observation and pre-  
432 diction, which can be negative if the data is transformed by log or Z-score, it was nec-  
433 essary to further transform the data by moving both observation and prediction above  
434 zero by adding the magnitude of the lowest observation. It is unclear to us if this trans-  
435 formation changed the behavior of this metric. However, the inability of the MSA and  
436 SSPB to distinguish between our value-predicting models (see Table 2) mean that these  
437 metrics are of limited use in this situation beyond demonstrating that the mid-range bulk  
438 of observation-prediction pairs are forecast well.

439 At all three flux energies, the ARMAX-MLT model, with higher validation correla-  
440 tion (0.769–0.804) and slightly lower RMSE, MAE, and MSA, improves on the base  
441 ARMAX models ( $r = 0.731 - 0.749$ ). However, the ability to predict highs and lows  
442 correctly (comparing orange cubic line to red 1:1 line) is only slightly improved. The dif-  
443 ferences between ARMAX-MLT and ARMAX-MLT+Lag6 are not appreciable. The clear  
444 choice, for simplicity, would be the ARMAX-MLT model.

### 4.3 Validation of the Value-Predicting RNN Models

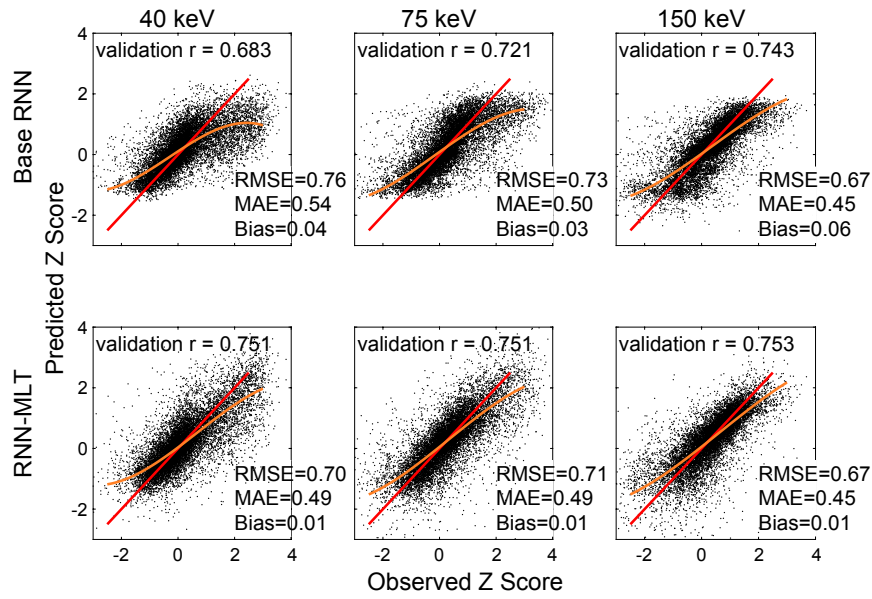
We would hope that a neural network model, by efficiently utilizing all available information in the data, would produce a prediction that both improved on the validation correlation and was more successful at predicting highs and lows. By building a time series dependent model using RNN (which we base on the previous 48 hours of predictor values), we also expected this would model the behavior of flux over time as well the ARMAX models. However, despite how the RNN model should include both this time dependent behavior and possible nonlinear information, we find no or, at best, only modest improvements in validation  $r$  (0.751–0.753), compared to the ARMAX models (Figure 5). Creating a separate model for each MLT does result in a small improvement over the base RNN model, but the metrics (validation  $r$ , RMSE, MAE, and MSA) are not much different from those of the ARMAX-MLT model.

### 4.4 Validation of the Reduced Polynomial/Interaction Regression Models

We are also interested in whether a parsimonious, more portable model, could be produced by ignoring the time series behavior of flux and focusing solely on the linear and nonlinear associations of flux with the predictors. This could work well if the time behavior of electron flux was highly dependent on the time behavior of the predictors. Nonlinear associations could take the form of multiplicative interactions (flux responding to each predictor differently depending on the levels of the other predictors), or some sort of polynomial response (quadratic, cubic, etc.). The possibility also exists that a long time stretch of predictors is not needed to produce a reasonable model. Predictors from a single or several hours before may be sufficient.

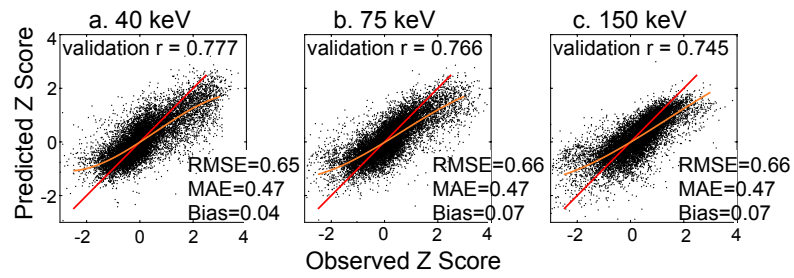
We created models containing all multiplicative interaction terms between predictors from one hour previous, as well as linear, quadratic, and cubic terms for each. We also included MLT as a categorical variable, including its interaction terms with the continuous variables. We label these as the REG-MLT models. However, adding all these terms resulted in unstable models that would be unsuitable for predictive purposes, so we further used stepwise regression (Neter et al., 1990) to trim the models down to the terms that best described flux. The stepwise process removed all cubic terms, leaving only linear, quadratic, and multiplicative interaction terms. These models, with validation correlation of  $r = 0.745 - 0.777$ , do only slightly worse at prediction, comparing the validation correlations, than the ARMAX-MLT models (Figure 6). As with the ARMAX-MLT models, there is also a tendency to underpredict the highs and over predict the lows, as evidenced by the cubic fit (orange line). RMSE, MAE, MSA, and bias measures are also similar to the ARMAX-MLT model. The advantage of this REG-MLT approach, however, is that these models would be more easily implemented by other users as all that is needed, after converting the data to standardized Z-scores, are the coefficient terms of the regression. While a neural network is also just a set of coefficients, the number of terms needed will be much lower for the stepwise-reduced REG-MLT model if statistically non-significant parameters are removed. This means that future prediction would be dependent on fewer inputs. Extraction of these coefficients is also easier as the output from a regression model in most statistical packages is simply the labelled coefficients. In the specific models we present here, there also appears to be little predictive advantage in including terms that describe time behavior, either the AR and MA terms of the ARMAX models or the 48 h measurements leading up to the flux prediction in the RNN models.

The validation correlation coefficients of the value-predicting models are all fairly close, but at 40 keV, the ARMAX-MLT model has the highest correlation with observations in the test set (Figure 7). At 150 keV, however, the 95% confidence intervals overlap so closely that there is no statistical difference between the models.

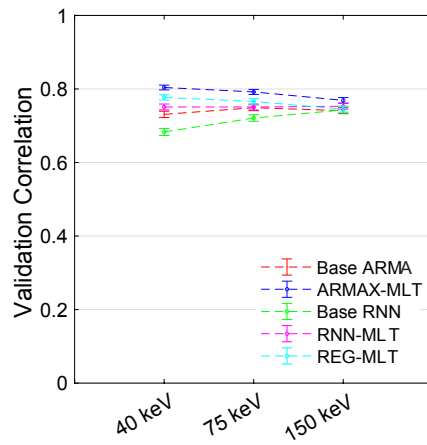


**Figure 5.** Scatterplots of predictions vs. observations for RNN models over the full validation period and all three energies (40 keV - 150 keV). First row: Base RNN model, second row: RNN-MLT. Red line shows the ideal 1:1 correspondence between predictions and observations. Orange line gives the cubic fit to the actual prediction-observation relationship. Flux is converted to unitless Z-scores.





**Figure 6.** Scatterplots of predictions vs. observations for the REG-MLT model over the full validation period and all three energies (a. 40 keV, b. 75 keV, c. 150 keV). Red line shows the ideal 1:1 correspondence between predictions and observations. Orange line gives the cubic fit to the actual prediction-observation relationship. Flux is converted to unitless Z-scores.



**Figure 7.** Validation correlation coefficients of the value-predicting models (base ARMAX, ARMAX-MLT, base RNN, RNN-MLT, and REG-MLT). Although 95% confidence intervals around each correlation are small, there is little practical difference between the models using this metric.

**Table 3.** Metrics of the REG-MLT models calculated from a. Z-score standardized flux, b. log10 flux, and c. original flux units.

	a. Z-score of			b. Log10 Flux			c. Untransformed		
	Log10 Flux			Flux			Flux		
	40 keV	75 keV	150 keV	40 keV	75 keV	150 keV	40 keV	75 keV	150 keV
RMSE	0.65	0.66	0.66	0.27	0.26	0.28	70944	33235	7118
MAE	0.49	0.47	0.47	0.19	0.19	0.2	30997	13314	3320
Bias	0.04	0.07	0.06	0.02	0.03	0.02	16719	7590	1640
MSA	5.10%	3.90%	4.00%	3.20%	3.30%	4.00%	39.00%	37.90%	40.70%
SSPB	-0.40%	-0.80%	-1.10%	-0.30%	-0.60%	-1.00%	-0.01%	-0.10%	0.13%
validation r	0.777	0.766	0.746	0.777	0.766	0.746	0.777	0.766	0.746
Flux Std Dev	1.026	1.023	0.983	0.426	0.405	0.417	95060	40636	9025
RMSE/StdDev	0.6335	0.6452	0.6714	0.6338	0.6421	0.6715	0.7463	0.8179	0.7887

Using the REG-MLT models, we take the opportunity to compare metrics for standardized (Z-score of log10 flux) and backtransformed non-standardized (Log10 flux), and the completely backtransformed flux data (Table 3). First, the RMSE is strongly influenced by the standard deviation of the response variable. The standard deviation of the three (log10) electron energies are all near 0.4 while that of the Z-score fluxes are roughly 1. (The standard deviation of the standardized test set is close to but not exactly 1 because the original standardization values were calculated from the training set.) Because the RMSE scales with the standard deviation, the unstandardized RMSE metrics are less than half the RMSE of the standardized output. This gives the erroneous impression that the unstandardized flux produces a better prediction. However, the untransformed flux data, with standard deviation in the  $10^4$  range, has a similarly large RMSE. This clearly demonstrates that we should not directly compare the RMSE metric between models using differently scaled data. We can standardize the RMSE by dividing by the standard deviation, as shown in the table (Liemohn et al., 2021). The other metrics are also affected by the scaling difference. The MSA, although it does not scale strictly linearly with the standard deviation, still shows a huge difference among the same predictions that merely differ in units, but the MSA is not as easily scaled to the standard deviation. Only the validation correlation is unaffected by changes in standard deviation or units. However, all these metrics are weighted heavily by the accurate prediction of mid-range values, missing the high and low values that are of most interest.

#### 4.5 Prediction above a Threshold

Another method of comparing model accuracy is to determine how often models correctly predict a flux rise over a certain threshold, with correct predictions being true positives (TP) and true negatives (TN) and incorrect predictions being false positives (FP) and false negatives (FN). These four categories can be used to calculate an accuracy rate:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

The true positive rate ( $TPR = TP/(TP+FN)$ ) gives the rate at which surpassing the threshold is correctly predicted. The true negative rate ( $TNR = TN/(TN + FP)$ ) is the proportion of the time the model correctly predicts that the flux will stay

below the threshold. The false positive rate ( $FPR = 1 - TNR$ ) is the misses due to predicting over the threshold when the observation stays below, and the false negative rate ( $FNR = 1 - TPR$ ) the rate at which it is predicted flux will stay below the threshold when, in fact, it goes above it (Yerushalmy, 1947).

The TP, TN, FP, FN, and ACC can be compared directly, or used to calculate the Heidke Skill Score (HSS) (Heidke, 1926). A score below zero will be obtained if the model predicts less well than chance alone. Scores closer to 1 show greater accuracy in prediction, with a score of 1 representing a perfect prediction:

$$HSS = \frac{2(TP \times TN - FP \times FN)}{[(TP + FN)(FN + TN) + (TP + FP)(FP + TN)]} \quad (2)$$

We calculate the HSS for several of the above models, predicting above either the 75th or 90th percentile (Table 4). Based on the HSS, the REG-MLT model appears to be more accurate than any of the ARMAX models, but less accurate than the RNN models. This is a different ranking of models than that obtained from the validation correlations where the ARMAX models did better. However, these differences are small and are likely not of much consequence. In general, but not always, these models perform somewhat better in predicting flux above the 75th percentile than above the 90th percentile.

These skill scores are higher than those obtained by Ganushkina et al. (2019) for flux events predicted by the IMPTAM (Inner Magnetosphere Particle Transport and Acceleration Model), the highest skill score being 0.17 for the 40 keV electrons at roughly the 75th percentile. Although, this previous study was predicting events at 10 minutes, a more difficult task, much of this difference in skill scores may be due to our use of a strictly empirical model and the incorporation of MLT both as a predictor in its own right and as a modifier of the other variables which may behave differently at different times or locations. A NARMAX model predicting daily averages of higher energy electron flux ( $\geq 2$  MeV) achieved a Heidke skill score of 0.738 (Balikhin et al., 2016). Our ARMAX, REG-MLT, and RNN results for hourly lower energy electrons are somewhat lower than this, reflecting the that the prediction of both hourly flux and lower energy flux are more difficult task.

## 5 Building Probability Prediction Models: RNN and Logistic Regression

If we are interested in predicting above a threshold (e.g., a certain percentile) we may find that a probability model, rather than a value-predicting model, may give us more accuracy. Both regression and neural network models, such as RNN, can be made to output the probability of being above a threshold rather than a specific value. A regression of this sort is called logistic regression (Neter et al., 1990; Berkson, 1944). We classify flux observations and predictions  $\geq 75th$  percentile as an event. Those less than this cut off are a non-event. (Note that this percentile is not the same as the probabilities discussed in the next paragraph.) Previously, this approach (used for daily predictions) was found to be more accurate at predicting events than value-predicting multiple regression or ARMAX models (Simms & Engebretson, 2020). We create several models using these algorithms: RNN-MLTclass (a series of RNN models, one for each MLT, classifying predictions into the event and non-event groups); LogisticLag1, LogisticLag3, and LogisticLag6 (all including MLT and each using predictors from 1, 3, or 6 hours before the flux measurement).

But at what probability do we forecast an event? A threshold of probability = 0.5 can be used as a default value (equal probability of either outcome), but there is generally a more optimal cut off point which can be determined from either a ROC (receiver

operating characteristic) curve (Fawcett, 2006; Liemohn et al., 2020, 2022), or from a precision-recall curve, which is thought to often provide better accuracy for rare events (Saito & Rehmsmeier, 2015). A ROC curve plots  $TPR$  vs  $FPR$  and highlights the ability of the model to distinguish between the two classes. Alternatively, precision ( $TP/(TP+FP)$ ) vs. recall ( $TPR$ ) plots better highlight the ability to accurately detect events, particularly if they are rare. The threshold point of either curve can be "tuned" to find the optimal cut off for distinguishing between classes by finding the maximum accurate separation between the classes.

RNN naturally incorporates past values of predictors if needed (we have chosen to use up to 48 hours from past predictor values). If these past values are important, we expect that a logistic regression based on only one previous time step would not do as well as an RNN model. Although we could enter many time steps into the logistic regression, we found that more than one either resulted in a model that could not converge, or provided very little further explanation of the variance. Using a stepwise procedure, as in the REG-MLT model above, cubic terms were not found to be useful. The variables in the logistic models, therefore, included only a single previous time step (Lag 1, 3, or 6), and whichever main effects, interactions, and quadratic terms were chosen as influential by the stepwise procedure.

We report the usual true and false positive rates (TPR or hit rate, and FPR or false alarm ratio), true and false negative rates (TNR and FNR), along with the AUC (area under the ROC curve), the Heidke skill score (HSS), Matthews correlation coefficient (MCC), and the critical success index (CSI) all at the optimal threshold determined from a precision-recall curve (Table 5). (See Chakraborty and Morley (2020) for CSI and MCC calculations.) The AUC is often used to compare models. A larger area under the ROC curve corresponds to a model that better differentiates between classes, with a value of 1 being completely accurate discrimination. All the AUC values, for both logistic and RNN models, are similar (0.850 - 0.907). Based on the AUC values alone, we could conclude that all models are performing well and roughly equally, but other metrics may also be considered. The HSS, MCC, and CSI measure somewhat different attributes: the improvement over random forecasts (HSS: Heidke (1926)), a measure of correlation of classes unaffected by unbalanced data (MCC: Chicco and Jurman (2020)), or a measure weighted to give more value to warnings for rare events (CSI: Schaefer (1990)). The Heidke skill scores lie within a small range (0.557 - 0.667), so although the 95% confidence intervals from 10,000 bootstrap resamplings of the data (of size 7000 points) of the HSS for each of these models is small, there is still considerable overlap between models (Figure 8). (Overlap of the 95% confidence intervals is equivalent to finding no difference between means in a t-test.) The ranges of MCC (0.558 - 0.673) and CSI (0.537 - 0.684) are similarly narrow and therefore also do not provide much discrimination between models. We do not show the bootstrap assessments of the MCC and CSI variation as they were also very small. (Bootstrapping randomly resamples, with replacement, from the original sample to obtain a standard error needed to create confidence intervals or perform hypothesis tests for statistics for which the underlying distribution is not otherwise known (Efron & Tibshirani, 1986))

Although the LogisticLag6 model consistently performs worse in all these metrics, there is no particular advantage to the RNN-MLTclass models over the LogisticLag1 or LogisticLag3. Again, this suggests that the RNN algorithm is not adding anything more than what we have incorporated into the logistic models. We may as well use the more portable model (the logistic model). Given that the LogisticLag1 and LogisticLag3 perform much the same, we would choose the LogisticLag3 (referred to as LL3 from now on) as the more practical model. Predictions could be made up to 3 h ahead instead of the 1 h required by all the other models we present.

For this one model (LL3; 150 keV) we show the determination of the optimal probability threshold using both the ROC curve and the precision-recall curve (Figure 9). With

625 the ROC curve, the optimal point on the curve is closest to the upper left corner (9a).  
 626 (We determined this using the `perfcuve` function in MATLAB.) This gives an optimal  
 627 probability threshold of 0.480, close to the "default" of 0.5. For the precision-recall curve,  
 628 the optimal threshold is chosen by maximizing (Saito & Rehmsmeier, 2015):

$$2 \times \frac{\textit{Precision} \times \textit{Recall}}{\textit{Precision} + \textit{Recall}} \quad (3)$$

629 This gives an optimal threshold of 0.313 (Figure 9b). Although the overall accu-  
 630 racy (ACC of Equation 1) is slightly higher with the ROC optimal threshold (0.832 (Fig-  
 631 ure 9c) vs. 0.821 (Figure 9d)), the ACC can be quite high only because it identifies true  
 632 negatives correctly without any correct identification of true events. For this reason, the  
 633 ACC is often not the best measure. The Heidke skill score shows no improvement using  
 634 the optimal threshold. It is 0.639 at the optimal point chosen by the precision-recall  
 635 curve, but 0.646 at the point chosen by the ROC curve. There is some advantage in using  
 636 the precision-recall curve point as there is an improvement in both the correctly pre-  
 637 dicted events (TPR) and a reduction in the number of missed events (FNR). There may  
 638 be some tolerance for the increased false positives if it improves the true positive rate.

### 639 5.1 Predicting High Flux after Periods of Low Flux

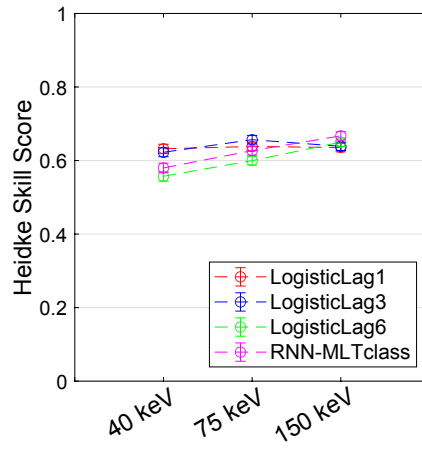
640 If we choose the LL3 model as the most practical, predicting when flux will be above  
 641 the 75th percentile 3 h later, there is one further test of its abilities we should make. As  
 642 all these models are based on a dataset dominated by one quiet hour leading into an-  
 643 other quiet hour, it would not be surprising if the ability to predict sudden rises were  
 644 limited by the overwhelming number of quiet data points. Furthermore, all our identi-  
 645 fied models predict whether high flux will occur in the future without regard to the cur-  
 646 rent status (low vs. high flux). It is a different task to predict a sudden rise in flux from  
 647 a low level versus a persistence of high flux. This model type should be assessed for its  
 648 ability to do that. In fact, we may find that a model made specifically for this situation  
 649 would do a better job.

650 We created one more model (LogisticLag3red) which uses as its training set only  
 651 those hours preceded by low flux. This removes all times when a high flux hour is pre-  
 652 ceded by high flux and gives us a method of predicting the specific case where flux rises  
 653 from a lower level. We also test the ability of the original LL3 model to perform this same  
 654 task by validating it only on the hours of the test set data which are preceded by low  
 655 flux. This is a more difficult task: less than 7% of this reduced test set are hours of high  
 656 flux following low flux. However, we find that these two models behave similarly, with  
 657 the original LL3 model performing better (at 40 and 150 keV) or not much worse (150  
 658 keV) as measured by the Heidke skill score if the optimal threshold is moved to account  
 659 for the lower percentage of observations in the event class (Table 6). The coefficients of  
 660 the original LL3 (or LogisticLag3) model (Tables 7- 9) can therefore be used to predict  
 661 this particular situation, if the revised optimal threshold coefficients are used (Table 10b),  
 662 without the need for a specialized model for this situation. As expected, the HSS for these  
 663 predictions are lower than the overall HSS from the LL3 model.

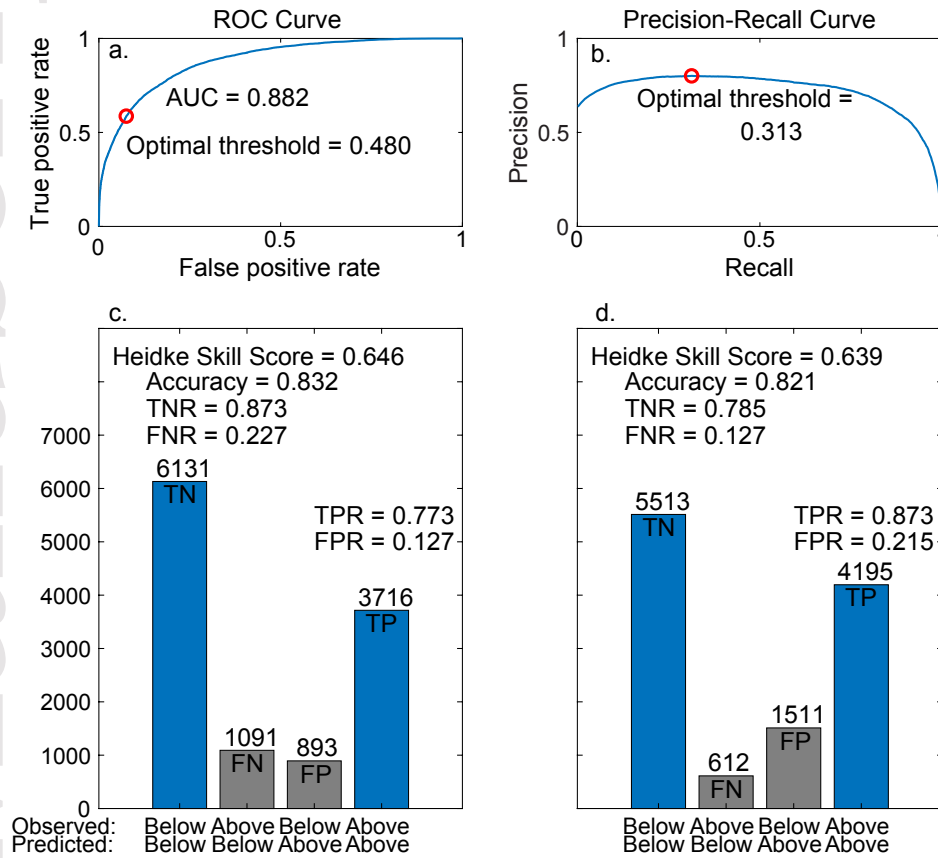
### 664 5.2 Using the LL3 Model for Prediction

665 We report the coefficients of the LL3 models (Table 7, Table 8, Table 9) for use  
 666 in future predictions. Inputs into these models must be Z-scores of the solar wind, IMF,  
 667 and geomagnetic index predictors, with logs taken as described above. For future ref-  
 668 erence, we use the LL3 (LogisticLag3) acronym for this model.

669 The MLT variable set consists of 23 indicator variables, one less than the number  
 670 of MLT values ( $MLT_0 - MLT_{22}$ ). The variable is discretized such that, for exam-



**Figure 8.** Heidke skill scores of the classifier models (LogisticLag1, LogisticLag3, LogisticLag6, and RNNclass-MLT). There is little difference between the models using this metric. Bootstrap confidence intervals (95%) are small.



**Figure 9.** Determining the optimal probability threshold (red circle) for the logistic Lag 3 (150 keV) model. a. ROC curve, b. precision-recall curve, c. classification of test set using the optimal threshold from ROC curve, d. classification of test set using optimal threshold from the precision-recall curve. Model predicts whether observation will be above or below the 75th percentile. AUC = area under the ROC curve, TPR = true positive rate, FPR = false positive rate, TNR = true negative rate, FNR = false negative rate.



**Table 4.** Heidke skill scores for the ARMAX, stepwise regression (including MLT), and RNN regression models evaluated for their ability to predict flux above the 75th and 90th percentiles.

		40 keV	75 keV	150 keV
Base ARMAX	75th %ile	0.586	0.597	0.622
	90th %ile	0.386	0.488	0.535
ARMAX-MLT	75th %ile	0.615	0.626	0.635
	90th %ile	0.521	0.541	0.562
ARMAX-MLT+Lag6	75th %ile	0.613	0.624	0.634
	90th %ile	0.518	0.546	0.566
REG-MLT	75th %ile	0.625	0.636	0.619
	90th %ile	0.570	0.591	0.549
Base RNN	75th %ile	0.527	0.613	0.654
	90th %ile	0.276	0.499	0.546
RNN-MLT	75th %ile	0.602	0.638	0.663
	90th %ile	0.589	0.569	0.609

671 ple,  $MLT = 0$  includes all observations where  $MLT = 0 - 0.99$ . An indicator variable  
672 is given the value of 1 for the hour it represents and 0 for all other hours. In other words,  
673 an observation at  $MLT = 0$  will have  $MLT\_0 = 1$ , with all other  $MLT\_$  variables = 0. An  
674 observation at  $MLT = 23$  is 0 in all  $MLT\_$  indicator variables.

675 The multiplicative interaction terms are produced by multiplying the relevant vari-  
676 ables. Any multiplication between a numeric variable and an MLT indicator variable will  
677 result in 0 for all cases except the hour of the MLT.

678 Predictions are made with the usual regression equation:

$$Y_{pred} = b_0 + b_1 X_1 + \dots + b_n X_n \quad (4)$$

679 where  $b_0$  is the constant term and each variable is multiplied by its corresponding  
680 coefficient ( $b_1, \dots, b_n$ ), but the output must be converted back to probabilities:

$$Prob = \frac{e^{(b_0 + b_1 X_1 + \dots + b_n X_n)}}{1 + e^{(b_0 + b_1 X_1 + \dots + b_n X_n)}} \quad (5)$$

681 Once these probabilities are calculated, assignment to classes (above or below the  
682 75th percentile) is accomplished by comparing to the optimal thresholds of Table 10a  
683 (Neter et al., 1990).

## 684 6 Discussion

685 Our exploration of various algorithms for creating predictive models shows that value  
686 predicting models (multiple regression, ARMAX, and RNN) have difficulty predicting  
687 the more extreme high and low fluxes of keV electrons at geostationary orbit. These mod-  
688 els can be improved by including MLT, both as a term that describes the differing lev-  
689 els of flux over the 24 h period as well as how the influence of other predictors varies at

**Table 5.** Accuracy metrics for the classification models: logistic and RNN-MLTclass.

Model	TPR	FPR	TNR	FNR	AUC	HSS	MCC	CSI
40 keV								
RNN-MLTclass	0.777	0.173	0.827	0.223	0.904	0.580	0.583	0.564
LogisticLag1	0.813	0.154	0.846	0.187	0.907	0.632	0.636	0.606
LogisticLag3	0.822	0.167	0.833	0.178	0.900	0.623	0.628	0.600
LogisticLag6	0.716	0.151	0.849	0.284	0.85	0.557	0.558	0.537
75 keV								
RNN-MLTclass	0.846	0.193	0.807	0.154	0.877	0.627	0.633	0.629
LogisticLag1	0.847	0.184	0.816	0.153	0.891	0.638	0.644	0.637
LogisticLag3	0.828	0.155	0.845	0.172	0.903	0.657	0.659	0.647
LogisticLag6	0.817	0.195	0.805	0.183	0.875	0.600	0.605	0.605
150 keV								
RNN-MLTclass	0.875	0.192	0.808	0.125	0.773	0.667	0.673	0.684
LogisticLag1	0.846	0.197	0.803	0.154	0.871	0.636	0.639	0.657
LogisticLag3	0.873	0.215	0.785	0.127	0.882	0.639	0.646	0.664
LogisticLag6	0.855	0.191	0.809	0.145	0.883	0.650	0.654	0.668

TPR: True Positive Rate (accurate prediction of event); FPR: False Positive Rate (false prediction of an event); TNR: True Negative Rate (accurate prediction of non-event); FNR: False Negative Rate (false prediction of non-event). AUC: Area under the ROC curve.

**Table 6.** Classification model accuracy metrics when limited to periods when flux could increase above the 75th percentile.

Model	TPR	FPR	TNR	FNR	AUC	HSS
LogisticLag3red						
Training set limited to starting low flux hours						
40 keV	0.728	0.190	0.810	0.272	0.854	0.357
75 keV	0.791	0.190	0.810	0.209	0.866	0.372
150 keV	0.611	0.075	0.925	0.389	0.886	0.412
LogisticLag3						
Training set includes all hours						
40 keV	0.653	0.139	0.861	0.347	0.900	0.395
75 keV	0.716	0.141	0.859	0.284	0.903	0.410
150 keV	0.743	0.163	0.837	0.257	0.882	0.361

690 each hour (by the use of multiplicative interaction terms with the MLT variables). The  
691 addition of multiplicative interactions between inputs and of quadratic terms is helpful  
692 in the regression model. (Presumably, these are also added by the RNN algorithm al-  
693 though it takes some effort to determine this.) However, even with these additional terms  
694 highs and lows are still under reported. Adding Lag 1 flux as a predictor may appear  
695 to rectify the issue, but this results in predictions that lag an hour behind. This prob-  
696 lem has also been found in a Kp predictive model. While adding historical Kp as a pre-  
697 dictor improved the model, it resulted in missing rapid changes (Chakraborty & Mor-  
698 ley, 2020). The addition of Lag 6 flux to our model, while it does not cause predictions  
699 to lag, also does not solve the problem of under-predicted highs and lows.

700 The prediction of high flux following low flux is the most challenging task, but also  
701 of the most practical importance. Models trained on these full datasets can give excel-  
702 lent prediction of the status quo (either the common mid-range values or persisting flux)  
703 because they tend to predict best what they are most heavily trained on. To better pre-  
704 dict just the start of high flux events, models can be built just on those days or hours  
705 when low flux could potentially rise to high flux (Simms & Engebretson, 2020). How-  
706 ever, an increase from low to high flux is still a somewhat rare event (much less than 50%:  
707 3049 out of 28921 hours or 10.5% of this subset) and without larger training sets, these  
708 empirical models may still struggle with predicting high flux events unless the relation-  
709 ship between flux increase and predictors is very strong. In fact, the high variability in  
710 flux response to the variables most often used for prediction might suggest that we are  
711 missing important parameters or processes that drive high flux events. One obvious can-  
712 didate in our particular models is substorm activity as we cannot include the *AE* index  
713 if we intend to use a model for real time prediction. However, studies that have included  
714 the *AE* (see below) do not achieve better prediction.

715 The regression models (both conventional and logistic) are able to include main ef-  
716 fects, multiplicative interactions, and quadratic and higher polynomial terms as needed.  
717 Thus, they could potentially produce a model very similar to that chosen by the RNN  
718 algorithm, with the exception that only a limited number of previous lags can be included  
719 before the model becomes intractable and too burdened with overly correlated predic-  
720 tors. The RNN models we created could, potentially, use up to 48 h of previous infor-  
721 mation, but limiting the number of previous hours in the regression models did not lower  
722 their predictive ability below that of the RNN models. We therefore conclude that the  
723 relevant prediction information is contained in the variables measured just an hour or  
724 up to 3 h before.

725 While cycling behavior may obscure the physical relationships of various processes,  
726 this may not be an issue with prediction models. For example, the nuisance diurnality  
727 of flux measurements from geosynchronous satellites may create misleading correlations.  
728 Models seeking to understand the physical drivers should account for this behavior, by  
729 such methods as ARMAX (autoregressive moving average transfer function) modelling  
730 or, at least, differencing of data to remove cycles (Simms, Engebretson, & Reeves, 2022),  
731 but perfectly serviceable prediction models can be produced even if these spurious cor-  
732 relations are not removed. Consequently, if ARMAX models are not needed to remove  
733 the cycling behavior, predictive models can be built from neural networks or simply from  
734 regression models. However, we must not make the mistake of interpreting an influen-  
735 tial parameter in these latter methods as evidence that it is a physical driver of flux. We  
736 must recognize that the high correlation seen between some variables and flux in mod-  
737 els that do not correct for co-cycling behavior and trends can only be interpreted as that  
738 parameter being a good proxy for the physical environment, not that that particular vari-  
739 able is the physical influence that drives electron flux. Bearing this in mind, when cre-  
740 ating prediction models, we can choose an algorithm for model construction based on  
741 such constraints as predictive ability, ease of determination, and portability to other users  
742 rather than considerations of which parameters are physically responsible for driving flux.

743 We do want to introduce as many variables as needed to describe the behavior of  
744 flux, not merely choose the few that are most important. It is true that we are not inter-  
745 ested in a cluttered model where there are many essentially duplicate variables, but  
746 previous work on drivers suggests that most available variables have some statistically  
747 significant association with flux even when other parameters are accounted for. Even if  
748 this apparent influence is small, there is little reason to discard a statistically significant  
749 variable. (The same does not hold true for a model seeking to answer questions about  
750 which variables are most likely to be drivers. As competition between variables can have  
751 large effects on both coefficient estimation and statistical testing, there are some sets of  
752 variables that should not be considered together in the same model.)

753 Model diagnostics should be geared to investigating the most important model fail-  
754 ures. For value-predicting models, single metrics may not be the best choice. RMSE, MAE,  
755 bias, MSA, and SSPB showed only small differences and therefore were not able to dif-  
756 ferentiate between models well. Heavy weighting on the most numerous prediction-observation  
757 pairs in the middle range leads these metrics to discount serious errors in the more im-  
758 portant, but less abundant, low and high flux ranges. This is the reason why we do not  
759 present 2-D density plots as they overlook the rarer, but more important, deviations from  
760 the model. Single metrics that seek to discount values farther from the predictions of the  
761 model, particularly if they are rarer, are even more misleading. The very low MSA val-  
762 ues that we obtain (compared to RMSE and MAE) show that this outlier-protected met-  
763 ric is missing much of the reduced prediction ability that we need to assess. We also show  
764 that both the mean-based RMSE and the median-based MSA are highly influenced by  
765 the standard deviation of the response variable. Comparing these metrics over variables  
766 of different standard deviations is therefore meaningless. Response variables should be  
767 normalized to the same standard deviation before metric comparisons are made.

768 In this study, we have developed several techniques for assessing poor fit of the value-  
769 predicting models that focus on the areas of most interest: the high and low fluxes. The  
770 first, the cubic fit line to the prediction-observation relationship, is a visual technique.  
771 Inspection of this line immediately tells us that the fit to highs and lows is much worse,  
772 for example, in the base RNN model than in the RNN-MLT model. The validation cor-  
773 relation only tells us that the fit is slightly worse and gives no indication of where the  
774 problem lies or how serious it is for our particular needs. Our second diagnostic, the Lag1  
775 validation correlation, assesses whether the predictions lag behind the observations. While  
776 lagging in the mid-range values is not as problematic, the missing of large changes in flux  
777 until the hour after they have happened is more troubling. If the predictions correlate  
778 better with observations from the hour before than with the hour being predicted, this  
779 tells us that the model will be missing the events that we are most interested in. While  
780 including persistence (Lag1 flux) as a predictor creates an extreme case, it is possible  
781 that other predictors could produce the same behavior so this should be checked in fu-  
782 ture models.

783 Validation correlations of the 3 value-predicting model types are reasonable and  
784 all of about the same magnitude (ARMAX-MLT: 0.731 - 0.814; RNN-MLT: 0.683 - 0.753;  
785 REG-MLT: 0.745 - 0.777). There is no apparent advantage to any of these model build-  
786 ing algorithms over the others in predictive ability. This suggests that each of these al-  
787 gorithms is accessing similar information from the predictors and that the choice of al-  
788 gorithm to build a model could depend mainly on accessibility. For a value-predicting  
789 model, a regression equation (the REG-MLT model) would be the obvious choice as there  
790 would be no need to transport the more complicated coefficients of an ARMAX or RNN  
791 model if the model were to be placed on another system.

792 The validation correlations we obtain are higher than the 0.67 validation correla-  
793 tion found for a previous hourly 40 keV ARMAX model (averaged over the individual  
794 MLT models) (Boynton et al., 2019). Reasons for this improvement in our ARMAX model  
795 may be that we include a decay term and that we create one, more efficient, model with

796 indicator variables to identify the individual MLTs instead of a series of models for each  
797 MLT.

798 However, validation correlations are not the only way to assess such models. We  
799 may be interested in a model's ability to predict when flux will go over a certain thresh-  
800 old, such as the 75th percentile. In this case, we could evaluate models with the Heidke  
801 skill score, comparing the classification success against that of random assignment to cat-  
802 egories. Using this metric, all 3 of these value predicting models perform similarly (Hei-  
803 dke skill scores at the 75th percentile of 0.615 - 0.635 (ARMAX-MLT), 0.602 - 0.663 (RNN-  
804 MLT), and 0.619 - 0.636 (REG-MLT)). All do better than random assignment and all  
805 are better than the output of the IMPTAM model where the highest skill score was 0.17  
806 for 40 keV electrons at roughly the 75th percentile ((Ganushkina et al., 2019)). Although,  
807 this previous study was predicting events at 10 minutes, a more difficult task, much of  
808 this difference in skill scores may be due to our use of a strictly empirical model and the  
809 incorporation of MLT.

810 An ARMAX model (*SNB<sup>3</sup>GEO*) predicting daily averages of higher energy elec-  
811 tron flux ( $\geq 2$  MeV at L=6.6) achieved a Heidke skill score of 0.738 (Balikhin et al., 2016).  
812 Our hourly models achieved a lower HSS, in part because hourly prediction is a more  
813 difficult task (Simms, Engebretson, & Reeves, 2022), and in part because correlations  
814 between predictors and lower energy flux are weaker. In our results, the 150 keV flux was  
815 usually somewhat better predicted than the 40 keV flux, so it is not surprising that 2  
816 MeV electrons would be more easily predicted than the lower energies.

817 Another daily regression model (at L=5.2) using several solar wind parameters gave  
818 a correlation between model prediction and observation of 0.854 at  $> 900$  keV, although  
819 it is not clear if this is a true validation correlation (on a reserved test set) or a corre-  
820 lation on the training data which would naturally be quite high (Katsavrias et al., 2022).  
821 In this same study, a daily prediction neural network model built on data over a wide  
822 range of energies and L shells (33 keV - 4.062 MeV; L 2.6-5.6) visually showed good cor-  
823 respondence with observations from a test set, while prediction at L=6.6, on an out of  
824 sample dataset, appeared considerably worse. The same problem we have experienced  
825 of under predicting the high values appears at low electron energies, along with over pre-  
826 diction of the low values at  $> 0.8$  MeV.

827 The MERLIN model, a neural network model predicting 120-600 keV electrons, shows  
828 a Spearman's rho validation correlation of 0.8 (at 120 keV), however this model incor-  
829 porates AE as a parameter, making it less useful for real time prediction (Smirnov et al.,  
830 2020)). This model also shows some difficulties with under predicting the highs and over  
831 predicting the lows. Although the validation correlation is reasonably high, this mostly  
832 represents accurate predictions in the less critical middle range. The ORIENT-M model,  
833 a neural network model for 50 keV - 1 MeV electrons, uses *AE* as a predictor as well (Ma  
834 et al., 2022). The  $R^2$  0.45 - 0.7 on the withheld test dataset (1 month of data near L  
835 6 for the 54 keV electrons) is therefore also not a good comparison to our model which  
836 does not incorporate the real time unavailable *AE*.

837 Predictions above or below a threshold can also be obtained from a dedicated clas-  
838 sifier model such as logistic regression or a classifier RNN model. Our classifier models  
839 provided little improvement to the Heidke skill score metric above that of the value-predicting  
840 models and were also not much different from each other (RNN-MLTclass: 0.580 - 0.667;  
841 LL3: 0.623 - 0.657).

842 While the RNN models were given 48 h of previous variables to work with, the lo-  
843 gistic regression models were given only predictors from 1, 3, or 6 h before. (Various com-  
844 binations of 1 with 2, 3, 6 did not provide any improvement.) While the RNN models  
845 could, in principle, have used any combination of polynomial and interaction terms from  
846 any of the 48 h, we found that the stepwise procedure on the logistic models did not choose

847 any polynomial terms above a quadratic. Thus, while we did not investigate the details  
848 of which terms the RNN models chose, we feel confident that a single predictor lag (Lag  
849 1 and Lag 3 worked best) and no polynomials above a squared term are sufficient to de-  
850 scribe this data.

851 Classifier models may be a better choice if the goal is to predict when an event (flux  
852 above a threshold) will occur. Classifier neural networks may predict as well as logistic  
853 regression models, but the latter provide the most portability. The LogisticLag3 model  
854 is preferred only because it would allow a 3 h lead in prediction time while not sacrific-  
855 ing any predictive power (as would be the case for the Lag 6 model). We therefore present  
856 the coefficients for the LL3 models (40-150 keV) in Tables 4-6 for future predictions.

857 We provide probability thresholds for classification using the LL3 model (Table 7)  
858 determined from optimizing precision vs. recall for two possible prediction scenarios: 1)  
859 predicting a flux over the 75th percentile from any flux level (high or low), and 2) pre-  
860 dicting a rise in flux over the 75th percentile from a lower flux level. We found that it  
861 was not necessary to produce another model for predictions in the second scenario, that  
862 the optimal prediction could be obtained simply by moving the probability threshold.  
863 We assumed equal costs of missing the prediction of an event (a false negative) vs. pre-  
864 dicting an event that did not happen (a false positive).

865 Finally, we note that what we have produced here are prediction models, not mod-  
866 els showing physical dependence. Although there is overlap between which variables (and  
867 in what form) best predict flux and which may be physical drivers of flux, the variables  
868 chosen by the optimal prediction models are not necessarily those that have a physical  
869 influence on flux. First, we have limited our variables to those that can be accessed in  
870 real time so as to obtain useful predictions. While substorms (represented by the *AE*  
871 index) may correlate well with flux this is not a useful variable for a working prediction  
872 model as *AE* is not available in real time. Second, as much of the correlation between  
873 predictor variables and flux is the result of common cycles (e.g., the diurnal cycle due  
874 to satellite position and the 27 d solar cycle; Simms, Engebretson, and Reeves (2022)),  
875 a good predictor may not be a driver at all. (For an investigation into the driving role  
876 of various parameters see Simms, Ganushkina, et al.) That the ARMAX models produced  
877 no better predictions than models derived by other means (RNN and regression) sug-  
878 gests that the description of the time behavior of flux can be accomplished either with  
879 AR and MA parameters or simply by using the co-cycling predictor variables, just so long  
880 as we have no reason to separate out the time behavior independent of these other vari-  
881 ables. ARMAX modelling, therefore, is best suited to exploring actual physical relation-  
882 ships between flux and possible drivers, but does not give this model type any advan-  
883 tage in producing a predictive model. By the same argument, the coefficients of the re-  
884 gression models presented here are not any more interpretable in a physical sense than  
885 the hidden coefficients of a neural network, as we have not accounted for inflated cor-  
886 relations due to common cycles.

887 However, there is a more fundamental problem in confusing prediction models with  
888 models used to test hypotheses about physical relationships. If the model selection method  
889 depends on sorting through many possible models to find the "best" (this includes tech-  
890 niques such as neural networks and stepwise regression), the probability of rejecting a  
891 true null hypothesis becomes almost 100% and any conclusions based on this will be mean-  
892 ingless (Mundry & Nunn, 2009; Hurvich & Tsai, 1990; Whittingham et al., 2006). The  
893 result of this "many-models-choose-best" approach is that we have no firm basis to say  
894 anything about the probability that a particular variable chosen by this type of model  
895 is actually influential. This means that any attempt to determine the drivers of electron  
896 flux from the terms chosen by a many-models-choose-best approach is misguided. Even  
897 if the model predicts well, there is no basis for inferring that the particular variables cho-  
898 sen in the model selection phase have an actual physical influence.

## 7 Conclusions

1. We screen several algorithms for producing value-predicting models of hourly 40 - 150 keV electron flux at geostationary orbit: ARMAX, RNN, and regression. These methods produce roughly similar models, as measured by validation correlations and the Heidke skill score.

2. Classifier models (RNN and logistic regression) are somewhat better at predicting a flux event (flux rising above the 75th percentile) than value-predicting models. A model built by logistic regression using only variables from one previous time step predicts as well as one built by a neural network using 48 h of previous data. We choose the LL3 model because it is more parsimonious and more portable than an RNN model.

3. Although the prediction of high flux following a low flux hour is both the most difficult task as well as the most important, we were able to produce a reasonable prediction model for this special case merely by changing the probability threshold of the LL3 model.

4. Two new diagnostic tests are introduced to assess value-predicting models: the cubic fit to the observation-prediction relationship, to visually assess the degree to which high and low flux is under or over predicted, and the Lag1 correlation which determines the degree to which predictions may lag behind and miss rapid changes in flux. Additionally, to focus attention on the model failures instead of successes, we plot observations vs. predictions as scatterplots instead of 2-D density plots, the latter of which tend to discount the rarer but more important deviations from the model.

5. A "good" metric is one which focuses attention on the ability of a model to predict the cases of most interest, not simply one which produces a low value. We find that single metrics such as RMSE, MAE, bias, MSA, and SSPB are all too influenced by the bulk of well-predicted, mid-range values to differentiate between models that do better at predicting extreme values of flux. However, median-based metrics, such as the MSA, may be even less useful as they weight large deviations less heavily, giving an unreasonably reassuring picture of model effectiveness at predicting the outliers we are most interested in. We note that RMSE uses the prediction error mean, while the MSA (and SSPB) use the prediction error median. Use of the median in the MSA reduces the influence of prediction error outliers in the metric, thus it is of less use in identifying model failures than the RMSE is.

6. Single-value metrics such as RMSE (or MSA) which use the difference (or ratio) between observation and prediction are highly influenced by the standard deviation of the response variable. It is therefore meaningless to compare the RMSE or MSA across variables with different standard deviations. The Z-score transformation (obtained by subtracting the mean and dividing by the standard deviation) normalizes variables to a standard deviation of 1, making a comparison across models or datasets more useful.

7. Parameters are chosen for their availability, not solely due to their high correlation with flux. This is because we are interested in a useful model, rather than a model with the highest validation correlation. Some highly correlated variables (such as the *AE* index) are not used because they would not be available in real time when predictions are needed.

8. The addition of MLT to these models, describing both the changing level of flux over the 24 h period as well as the change in predictor influence at each hour, improves predictions.

9. Value-predicting models do a poor job at predicting the highs and lows although mid-range prediction is very good. The relative rarity of data points in the areas of most interest (very high flux) results in models that will miss these events much of the time.

948 The addition of flux from the previous hour as an input variable appears to fix this prob-  
949 lem, but produces predictions that lag behind observations.

950 10. The addition of multiplicative interactions between the predictor variables, as  
951 well as quadratic terms, improves predictions. Cubic terms had no effect.

952 11. Predictors from a full 48 h before are not needed. Variables measured in a sin-  
953 gle hour (1-3 h before the flux observation) are sufficient for a reasonable prediction. The  
954 time behavior of flux does not need to be described (i.e., with an ARMAX model) to pro-  
955 duce reasonable predictions.

956 12. We provide coefficients and optimal probability thresholds to predict a flux rise  
957 above the 75th percentile 3 h in advance using a logistic regression model. (LL3) The  
958 logistic regression model was chosen for its portability to other systems. The 3 h time  
959 frame was chosen to provide a good balance between early warning and best prediction.

960 13. The best predictive model does not necessarily tell us anything about the phys-  
961 ical relationship of each solar wind, IMF, or geomagnetic parameter with flux. If hypothe-  
962 ses about these relationships are to be explored, the model approach should be targeted  
963 to that goal instead of merely improving prediction scores.



**Table 7.** Coefficients of the 40 keV LL3 (Logistic Lag at 3 h) prediction model.

		MLT Associated Coefficients:				
		MLT	$B_z \times MLT$	$V \times MLT$	$K_p \times MLT$	
Constant	-1.482					
B	-0.104					
$B_z$	-0.542	MLT_0	0.275	-0.009	0.046	0.047
V	0.418	MLT_1	0.375	0.040	0.040	-0.052
P	0.036	MLT_2	0.963	0.158	0.142	-0.088
$K_p$	1.297	MLT_3	1.081	0.022	-0.014	0.121
Dst	-0.359	MLT_4	1.316	-0.185	-0.037	0.137
$B^2$	-0.092	MLT_5	1.079	-0.200	-0.145	0.500
$V^2$	-0.117	MLT_6	1.254	-0.586	0.098	0.932
$B \times B_z$	0.065	MLT_7	1.203	-0.496	0.322	0.374
$B \times V$	-0.096	MLT_8	1.029	-0.591	0.112	1.035
$B \times P$	-0.065	MLT_9	0.650	-0.417	0.176	0.944
$B_z \times V$	-0.083	MLT_10	0.679	-0.485	0.022	0.480
$B_z \times K_p$	0.229	MLT_11	0.359	-0.214	-0.318	0.896
$V \times K_p$	0.272	MLT_12	-0.291	-0.093	-0.133	0.973
$P \times K_p$	-0.224	MLT_13	-0.387	-0.237	-0.080	0.288
$K_p \times Dst$	0.184	MLT_14	-0.752	-0.127	-0.366	0.347
		MLT_15	-1.182	-0.113	-0.335	0.571
		MLT_16	-1.319	-0.050	-0.304	0.282
		MLT_17	-1.973	-0.225	-0.292	0.360
		MLT_18	-2.563	-0.301	-0.198	0.753
		MLT_19	-2.169	-0.274	-0.176	0.538
		MLT_20	-1.632	-0.070	-0.067	0.614
		MLT_21	-1.572	0.111	0.079	0.635
		MLT_22	-0.473	-0.060	0.231	0.016

**Table 8.** Coefficients of the 75 keV LL3 (Logistic Lag at 3 h) prediction model.

		MLT Associated Coefficients:					
		MLT	$B \times MLT$	$Bz \times MLT$	$V \times MLT$	$Kp \times MLT$	
Constant	-1.986						
B	-0.438						
Bz	1.961	MLT_0	0.061	0.141	-0.34	0.33	-0.397
Ey	2.104	MLT_1	0.21	-0.07	-0.053	0.036	-0.149
V	0.787	MLT_2	0.488	-0.153	0.048	0.093	0.127
P	-0.085	MLT_3	0.863	-0.077	-0.114	-0.1	0.179
Kp	1.432	MLT_4	1.199	0.167	-0.332	0.07	-0.074
Dst	-0.797	MLT_5	1.355	0.17	-0.397	-0.111	0.201
SolarFlux	-0.081	MLT_6	1.545	0.045	-0.485	0.037	0.689
$B^2$	-0.07	MLT_7	1.663	0.118	-0.485	0.229	0.463
$V^2$	-0.236	MLT_8	1.526	0.134	-0.579	-0.057	1.32
$P^2$	-0.081	MLT_9	1.29	0.422	-0.388	0.023	0.501
$Kp^2$	0.499	MLT_10	1.131	0.495	-0.494	-0.103	0.32
$B \times Bz$	0.118	MLT_11	0.793	0.318	-0.328	-0.046	0.991
$B \times V$	-0.18	MLT_12	0.727	0.253	-0.103	-0.032	0.551
$B \times Kp$	-0.349	MLT_13	0.456	0.444	-0.256	-0.024	0.164
$B \times Dst$	-0.156	MLT_14	0.269	0.323	-0.028	-0.253	0.325
$B \times SolarFlux$	0.084	MLT_15	0.086	0.348	0.042	-0.266	0.16
$Bz \times V$	0.399	MLT_16	-0.011	0.482	-0.056	-0.262	-0.168
$Bz \times P$	-0.325	MLT_17	-0.342	0.483	0.032	-0.476	-0.043
$Bz \times Dst$	-0.17	MLT_18	-0.611	0.451	-0.147	-0.211	-0.298
$Ey \times P$	-0.263	MLT_19	-0.484	0.215	-0.02	-0.234	-0.161
$Ey \times Dst$	-0.124	MLT_20	-0.621	0.446	-0.072	0.113	-0.193
$V \times P$	0.071	MLT_21	-0.838	0.358	0.044	0.03	0.152
$V \times Kp$	0.226	MLT_22	-0.592	0.392	0.026	-0.07	-0.114
$V \times SolarFlux$	-0.12						
$P \times Kp$	-0.183						
$P \times Dst$	0.133						
$P \times SolarFlux$	-0.093						
$Kp \times Dst$	0.506						

**Table 9.** Coefficients of the 150 keV LL3 (Logistic Lag at 3 h) prediction model.

		MLT Associated Coefficients:			
			MLT	Bz X MLT	Kp X MLT
Constant	-1.865				
B	-0.533				
Bz	0.773	MLT_0	-0.395	-0.039	-0.105
Ey	0.718	MLT_1	-0.400	0.305	0.045
N	-0.146	MLT_2	-0.048	0.246	0.086
V	1.130	MLT_3	0.166	0.273	0.189
Kp	0.309	MLT_4	0.656	0.078	0.282
Dst	-1.098	MLT_5	0.954	-0.053	0.250
SolarFlux	-0.260	MLT_6	1.400	-0.242	0.556
B <sup>2</sup>	-0.121	MLT_7	1.603	-0.097	0.579
Bz <sup>2</sup>	-0.186	MLT_8	1.674	-0.095	0.676
N <sup>2</sup>	-0.094	MLT_9	1.619	-0.066	0.699
V <sup>2</sup>	-0.164	MLT_10	1.562	0.073	0.778
Kp <sup>2</sup>	0.522	MLT_11	1.474	-0.010	0.786
Dst <sup>2</sup>	-0.057	MLT_12	1.260	0.072	0.574
SolarFlux <sup>2</sup>	0.009	MLT_13	1.086	0.106	0.445
B:Ey	-0.111	MLT_14	0.819	0.223	0.461
B:N	0.093	MLT_15	0.552	0.230	0.398
B:V	-0.098	MLT_16	0.287	0.281	0.366
B:Kp	-0.102	MLT_17	0.219	0.426	0.342
B:Dst	-0.155	MLT_18	-0.174	0.276	0.091
B:SolarFlux	0.185	MLT_19	-0.363	0.240	0.258
Bz:Ey	-0.121	MLT_20	-0.380	0.152	0.146
Bz:Kp	-0.652	MLT_21	-0.598	0.200	0.207
Bz:Dst	-0.325	MLT_22	-0.705	0.225	0.194
Ey:Kp	-0.325				
Ey:Dst	-0.228				
N:Kp	-0.225				
N:Dst	0.303				
N:SolarFlux	-0.073				
V:Kp	-0.182				
V:Dst	0.182				
V:SolarFlux	-0.183				
Kp:Dst	0.403				
Dst:SolarFlux	0.147				

**Table 10.** Optimal thresholds for the LL3 prediction models, a. Using entire test set, b. Using only low flux hours that could rise above 75th percentile.

Test set:	a. All hours	b. Starting low flux hours
40 keV	0.34	0.17
75 keV	0.308	0.235
150 keV	0.326	0.351

## Acknowledgments

The work at the University of Michigan was partly funded by National Aeronautics and Space Administration grants NNX17AI48G, 80NSSC20K0353, and National Science Foundation grant 1663770. The contributions by M. van de Kamp and N. Ganushkina were also partly supported by the Academy of Finland (grant 339329). We thank the reviewers and editor (Dr. S. Morley) for their helpful suggestions.

## Data Availability

The GOES-13 MAGED data and GOES Magnetometer 1 data used in the present study are available at <https://www.ngdc.noaa.gov/stp/satellite/goes/dataaccess.html>. Solar wind parameters and magnetic indices were obtained from OMNIWeb:

<https://omniweb.gsfc.nasa.gov/form/dx1.html>.

All models developed for this paper are available at doi:10.5281/zenodo.7520424.

## References

- Alpaydin, E. (2014). *Introduction to machine learning. 3*. Cambridge, MA, USA: MIT Press.
- Balikhin, M. A., Boynton, R. J., Billings, S. A., Gedalin, M., Ganushkina, N., Coca, D., & Wei, H. (2010). Data based quest for solar wind-magnetosphere coupling function. *Geophysical Research Letters*, *37*(24). doi: <https://doi.org/10.1029/2010GL045733>
- 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*(18). doi: <https://doi.org/10.1029/2011GL048980>
- Balikhin, M. A., Rodriguez, J., R.J., B., Walker, S., Aryan, H., Sibeck, D., & Billings, S. (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. doi:10.1002/2015SW001303
- Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, *39*(227), 357-365.
- Blake, J. B., Baker, D. N., Turner, N., Ogilvie, K. W., & Lepping, R. P. (1997). Correlation of changes in the outer-zone relativistic-electron population with upstream solar wind and magnetic field measurements. *Geophysical Research Letters*, *24*(8), 927-929. doi: <https://doi.org/10.1029/97GL00859>
- Boynton, R. J., Amariutei, O. A., Shprits, Y. Y., & Balikhin, M. A. (2019). The system science development of local time-dependent 40-keV electron flux models for geostationary orbit. *Space Weather*, *17*(6), 894-906. doi: <https://doi.org/10.1029/2018SW002128>
- Boynton, R. J., Balikhin, M. A., & Billings, S. A. (2015). Online NARMAX model for electron fluxes at GEO. *Annales Geophysicae*, *33*, 405-411. doi:10.5194/angeo-33-405-2015
- Boynton, R. J., Balikhin, M. A., Billings, S. A., Reeves, G. D., Ganushkina, N.,

- 1005 Gedalin, M., . . . Walker, S. N. (2013). The analysis of electron fluxes at  
 1006 geosynchronous orbit employing a NARMAX approach. *Journal of Geophysical*  
 1007 *Research: Space Physics*, 118(4), 1500-1513. doi: [https://doi.org/10.1002/](https://doi.org/10.1002/jgra.50192)  
 1008 [jgra.50192](https://doi.org/10.1002/jgra.50192)
- 1009 Boynton, R. J., Balikhin, M. A., Billings, S. A., Wei, H. L., & Ganushkina, N.  
 1010 (2011). Using the NARMAX OLS-ERR algorithm to obtain the most  
 1011 influential coupling functions that affect the evolution of the magneto-  
 1012 sphere. *Journal of Geophysical Research: Space Physics*, 116(A5). doi:  
 1013 <https://doi.org/10.1029/2010JA015505>
- 1014 Boynton, R. J., Balikhin, M. A., Sibeck, D. G., Walker, S. N., Billings, S. A., &  
 1015 Ganushkina, N. (2016). Electron flux models for different energies at geosta-  
 1016 tionary orbit. *Space Weather*, 14, 846-860. doi: [doi:10.1002/2016SW001506](https://doi.org/10.1002/2016SW001506)
- 1017 Camporeale, E., Wilkie, G. J., Drozdov, A. Y., & Bortnik, J. (2022). Data-driven  
 1018 discovery of fokker-planck equation for the earth's radiation belts electrons us-  
 1019 ing physics-informed neural networks. *Journal of Geophysical Research: Space*  
 1020 *Physics*, 127(e2022JA030377). doi: [10.1029/2022JA030377](https://doi.org/10.1029/2022JA030377)
- 1021 Capman, N. S. S., Simms, L. E., Engebretson, M. J., Clilverd, M. A., Rodger, C. J.,  
 1022 Reeves, G. D., & et al. (2019). Comparison of multiple and logistic regres-  
 1023 sion analyses of relativistic electron flux enhancement at geosynchronous orbit  
 1024 following storms. *Journal of Geophysical Research: Space Physics*, 124. doi:  
 1025 [10.1029/2019JA027132](https://doi.org/10.1029/2019JA027132)
- 1026 Chakraborty, S., & Morley, S. K. (2020). Probabilistic prediction of geomagnetic  
 1027 storms and the Kp index. *Journal of Space Weather and Space Climate*, 10,  
 1028 36.
- 1029 Chen, M. W., Lemon, C. L., Orlova, K., Shprits, Y., Hecht, J., & Walterscheid,  
 1030 R. L. (2015). Comparison of simulated and observed trapped and precipitating  
 1031 electron fluxes during a magnetic storm. *Geophysical Research Letters*, 42(20),  
 1032 8302-8311. doi: <https://doi.org/10.1002/2015GL065737>
- 1033 Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation co-  
 1034 efficient (MCC) over F1 score and accuracy in binary classification evaluation.  
 1035 *BMC Genomics*, 21(6). doi: <https://doi.org/10.1186/s2864-019-6413-7>
- 1036 Choi, H. S., Lee, J., Cho, K. S., Kwak, Y. S., Cho, I. H., Park, Y. D., . . . Lee, D. K.  
 1037 (2011). Analysis of GEO spacecraft anomalies: Space weather relationships.  
 1038 *Sp. Weather*, 9(5), 1-12. doi: [10.1029/2010SW000597](https://doi.org/10.1029/2010SW000597)
- 1039 Chu, X., Ma, D., Bortnik, J., Tobiska, A., W. K. and Cruz, Bouwer, S. D., & et  
 1040 al. (2021). Relativistic electron model in the outer radiation belt using  
 1041 a neural network approach. *Space Weather*, 19(e2021SW002808). doi:  
 1042 [10.1029/2021SW002808](https://doi.org/10.1029/2021SW002808)
- 1043 Denton, M. H., Henderson, M. G., Jordanova, V. K., Thomsen, M. F., Borovsky,  
 1044 J. E., Woodroffe, J., . . . Pitchford, D. (2016). An improved empirical model of  
 1045 electron and ion fluxes at geosynchronous orbit based on upstream solar wind  
 1046 conditions. *Space Weather*, 14(7), 511-523. doi: [10.1002/2016SW001409](https://doi.org/10.1002/2016SW001409)
- 1047 Denton, M. H., Thomsen, M. F., Jordanova, V. K., Henderson, M. G., Borovsky,  
 1048 J. E., Denton, J. S., . . . Hartley, D. P. (2015). An empirical model of elec-  
 1049 tron and ion fluxes derived from observations at geosynchronous orbit. *Space*  
 1050 *Weather*, 13(4), 233-249. doi: [10.1002/2015SW001168](https://doi.org/10.1002/2015SW001168)
- 1051 Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confi-  
 1052 dence intervals, and other measures of statistical accuracy. *Statistical Science*,  
 1053 1, 54-77. doi: [10.1214/SS/1177013815](https://doi.org/10.1214/SS/1177013815)
- 1054 Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*,  
 1055 27(8), 861-874. doi: <https://doi.org/10.1016/j.patrec.2005.10.010>
- 1056 Fok, M.-C., Buzulukova, N. Y., Chen, S.-H., Glocer, A., Nagai, T., Valek, P., &  
 1057 Perez, J. D. (2014). The comprehensive inner magnetosphere-ionosphere  
 1058 model. *Journal of Geophysical Research: Space Physics*, 119(9), 7522-7540.  
 1059 doi: [10.1002/2014JA020239](https://doi.org/10.1002/2014JA020239)

- 1060 Freeman, J. W. (1974). Kp dependence of plasma sheet boundary. *Journal of Geo-*  
 1061 *physical Research*, *79*, 4315.
- 1062 Freeman, J. W., O'Brien, T. P., Chan, A. A., & Wolf, R. A. (1998). Energetic  
 1063 electrons at geostationary orbit during the november 3-4, 1993 storm: Spa-  
 1064 tial/temporal morphology, characterization by a power law spectrum and,  
 1065 representation by an artificial neural network. *Journal of Geophysical Re-*  
 1066 *search: Space Physics*, *103*(A11), 26251-26260. doi: [https://doi.org/10.1029/](https://doi.org/10.1029/97JA03268)  
 1067 [97JA03268](https://doi.org/10.1029/97JA03268)
- 1068 Ganushkina, N. Y., Liemohn, M. W., Amariutei, O. A., & Pitchford, D. (2014).  
 1069 Low-energy electrons (5-50 keV) in the inner magnetosphere. *Journal*  
 1070 *of Geophysical Research: Space Physics*, *119*(1), 246-259. doi: [10.1002/](https://doi.org/10.1002/2013JA019304)  
 1071 [2013JA019304](https://doi.org/10.1002/2013JA019304)
- 1072 Ganushkina, N. Y., Sillanpää, I., Welling, D., Haiducek, J., Liemohn, M.,  
 1073 Dubyagin, S., & Rodriguez, J. V. (2019). Validation of Inner Magnetosphere  
 1074 Particle Transport and Acceleration Model (IMPTAM) with long-term GOES  
 1075 MAGED measurements of keV electron fluxes at geostationary orbit. *Space*  
 1076 *Weather*, *17*(5), 687-708. doi: <https://doi.org/10.1029/2018SW002028>
- 1077 Ganushkina, N. Y., Swiger, B., Dubyagin, S., Matéo-Vélez, J.-C., Liemohn, M. W.,  
 1078 Sicard, A., & Payan, D. (2021). Worst-case severe environments for sur-  
 1079 face charging observed at LANL satellites as dependent on solar wind and  
 1080 geomagnetic conditions. *Space Weather*, *19*(9), e2021SW002732. doi:  
 1081 <https://doi.org/10.1029/2021SW002732>
- 1082 Ginet, G. P., O'Brien, T. P., Huston, S. L., Johnston, W. R., Guild, T. B., Friedel,  
 1083 R., ... Su, Y.-J. (2013). AE9, AP9 and SPM: new models for specifying  
 1084 the trapped energetic particle and space plasma environment. *Space Science*  
 1085 *Reviews*, *179*, 579-615. doi: [10.1007/s11214-013-9964-y](https://doi.org/10.1007/s11214-013-9964-y)
- 1086 Glauert, S. A., Horne, R. B., & Meredith, N. P. (2014). Three-dimensional elec-  
 1087 tron radiation belt simulations using the BAS radiation belt model with  
 1088 new diffusion models for chorus, plasmaspheric hiss, and lightning-generated  
 1089 whistlers. *Journal of Geophysical Research: Space Physics*, *119*(1), 268-289.  
 1090 doi: <https://doi.org/10.1002/2013JA019281>
- 1091 Hartley, D. P., Denton, M. H., & Rodriguez, J. V. (2014). Electron number density,  
 1092 temperature, and energy density at GEO and links to the solar wind: A simple  
 1093 predictive capability. *Journal of Geophysical Research: Space Physics*, *119*(6),  
 1094 4556-4571. doi: <https://doi.org/10.1002/2014JA019779>
- 1095 Heidke, P. (1926). Measures of success and goodness of wind force forecasts by the  
 1096 gale-warning service. *Geogr. Ann.*, *8*, 301-349. doi: [https://doi.org/10.1080/](https://doi.org/10.1080/20014422.1926.11881138)  
 1097 [20014422.1926.11881138](https://doi.org/10.1080/20014422.1926.11881138)
- 1098 Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural compu-*  
 1099 *tation*, *9*(8), 1735-1780. doi: <https://doi.org/10.1162/neco.1997.9.8.1735>
- 1100 Hurvich, C. M., & Tsai, C. (1990). The impact of model selection on inference in  
 1101 linear regression. *The American Statistician*, *44*(3), 214-217. doi: [10.1080/](https://doi.org/10.1080/00031305.1990.10475722)  
 1102 [00031305.1990.10475722](https://doi.org/10.1080/00031305.1990.10475722)
- 1103 Hyndman, R., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*.  
 1104 Heathmont, Victoria, Australia: OTexts.
- 1105 Iyemori, T., Takeda, M., Nose, M., Odagi, Y., & Toh, H. (2010). Mid-latitude geo-  
 1106 magnetic indices ASY and SYM for 2009 (provisional). *Internal Report of Data*  
 1107 *Analysis Center for Geomagnetism and Space Magnetism, Kyoto University,*  
 1108 *Japan.*
- 1109 Jordanova, V. K., Tu, W., Chen, Y., Morley, S. K., Panaitescu, A.-D., Reeves,  
 1110 G. D., & Kletzing, C. A. (2016). RAM-SCB simulations of electron trans-  
 1111 port and plasma wave scattering during the October 2012 double-dip storm.  
 1112 *Journal of Geophysical Research: Space Physics*, *121*(9), 8712-8727. doi:  
 1113 <https://doi.org/10.1002/2016JA022470>
- 1114 Katsavrias, C., Aminalragia-Giamini, S., Papadimitriou, C., Daglis, I. A., Sandberg,

- 1115 I., & Jiggins, P. (2022). Radiation belt model including semi-annual varia-  
 1116 tion and solar driving (Sentinel). *Space Weather*, 20(e2021SW002936). doi:  
 1117 10.1029/2021SW002936
- 1118 Kellerman, A. C., & Shprits, Y. Y. (2012). On the influence of solar wind conditions  
 1119 on the outer-electron radiation belt. *Journal of Geophysical Research: Space*  
 1120 *Physics*, 117(A5). doi: <https://doi.org/10.1029/2011JA017253>
- 1121 Koons, H. C., & Gorney, D. J. (1991). A neural network model of the relativistic  
 1122 electron flux at geosynchronous orbit. *Journal of Geophysical Research: Space*  
 1123 *Physics*, 96(A4), 5549-5556. doi: <https://doi.org/10.1029/90JA02380>
- 1124 Koons, H. C., Mazur, J. E., Selesnick, R. S., Blake, J. B., Fennell, J. F., Roeder,  
 1125 J. L., & Anderson, P. C. (2000). *The impact of the space environment on space*  
 1126 *systems* (Vol. AFRL-VS-TR-20001578).
- 1127 Korth, H., Thomsen, M. F., Borovsky, J. E., & McComas, D. J. (1999). Plasma  
 1128 sheet access to geosynchronous orbit. *Journal of Geophysical Research: Space*  
 1129 *Physics*, 104(A11), 25047-25061. doi: <https://doi.org/10.1029/1999JA900292>
- 1130 Lam, H.-L., Boteler, D. H., Burlton, B., & Evans, J. (2012). Anik-E1 and E2 satel-  
 1131 lite failures of January 1994 revisited. *Space Weather*, 10(10). doi: 10.1029/  
 1132 2012SW000811
- 1133 Li, X., Baker, D. N., Temerin, M., Reeves, G., Friedel, R., & Shen, C. (2005). En-  
 1134 ergetic electrons, 50 keV to 6 MeV, at geosynchronous orbit: Their responses  
 1135 to solar wind variations. *Space Weather*, 3(4). doi: <https://doi.org/10.1029/2004SW000105>
- 1136
- 1137 Li, X., Temerin, M., Baker, D. N., Reeves, G. D., & Larson, D. (2001). Quantita-  
 1138 tive prediction of radiation belt electrons at geostationary orbit based on solar  
 1139 wind measurements. *Geophysical Research Letters*, 28(9), 1887-1890. doi:  
 1140 <https://doi.org/10.1029/2000GL012681>
- 1141 Liemohn, M. W., Adam, J. G., & Ganushkina, N. Y. (2022). Analysis of fea-  
 1142 tures in a sliding threshold of observation for numeric evaluation (STONE)  
 1143 curve. *Space Weather*, e2022SW003102. doi: <https://doi.org/10.1029/2022SW003102>
- 1144
- 1145 Liemohn, M. W., Azari, A. R., Ganushkina, N. Y., & Rastätter, L. (2020). The  
 1146 STONE Curve: A ROC-Derived Model Performance Assessment Tool. *Earth*  
 1147 *and Space Science*, 7(8), e2020EA001106. doi: <https://doi.org/10.1029/2020EA001106>
- 1148
- 1149 Liemohn, M. W., Shane, A. D., Azari, A. R., Petersen, A. K., Swiger, B. M., &  
 1150 Mukhopadhyay, A. (2021). Rmse is not enough: Guidelines to robust data-  
 1151 model comparisons for magnetospheric physics. *Journal of Atmospheric*  
 1152 *and Solar-Terrestrial Physics*, 218, 105624. doi: <https://doi.org/10.1016/j.jastp.2021.105624>
- 1153
- 1154 Ling, A. G., Ginet, G. P., Hilmer, R. V., & Perry, K. L. (2010). A neural network-  
 1155 based geosynchronous relativistic electron flux forecasting model. *Space*  
 1156 *Weather*, 8(9). doi: <https://doi.org/10.1029/2010SW000576>
- 1157 Loto'aniu, T. M., Singer, H. J., Rodriguez, J. V., Green, J., Denig, W., Biesecker,  
 1158 D., & Angelopoulos, V. (2015). Space weather conditions during the  
 1159 Galaxy 15 spacecraft anomaly. *Space Weather*, 13(8), 484-502. doi:  
 1160 10.1002/2015SW001239
- 1161 Lyatsky, W., & Khazanov, G. V. (2008). Effect of solar wind density on relativistic  
 1162 electrons at geosynchronous orbit. *Geophysical Research Letters*, 35(3). doi:  
 1163 <https://doi.org/10.1029/2007GL032524>
- 1164 Ma, D., Chu, X., Bortnik, J., Claudepierre, S. G., Tobiska, W. K., Cruz, A., & et al.  
 1165 (2022). Modeling the dynamic variability of sub-relativistic outer radiation belt  
 1166 electron fluxes using machine learning. *Space Weather*, 20(e2022SW003079).  
 1167 doi: 10.1029/2022SW003079
- 1168 Matéo-Vélez, J.-C., Sicard, A., Payan, D., Ganushkina, N., Meredith, N. P., & Sil-  
 1169 lanpää, I. (2018). Spacecraft surface charging induced by severe environments

- at geosynchronous orbit. *Space Weather*, 16. doi: 10.1002/2017SW001689
- 1170 Morley, S. K., Brito, T. V., & Welling, D. T. (2018). Measures of model performance  
1171 based on the log accuracy ratio. *Space Weather*, 16, 69–88. doi: [https://doi](https://doi.org/10.1002/2017SW001669)  
1172 [.org/10.1002/2017SW001669](https://doi.org/10.1002/2017SW001669)  
1173
- 1174 Mundry, R., & Nunn, C. (2009). Stepwise model fitting and statistical inference:  
1175 turning noise into signal pollution. *American Naturalist*, 173(1), 119–123.
- 1176 Neter, J., Kutner, M. H., & Wassermann, W. (1990). *Applied linear statistical mod-*  
1177 *els, 3rd ed.* Homewood, Illinois, USA: Irwin.
- 1178 Pakhotin, I. P., Drozdov, A. Y., Shprits, Y. Y., Boynton, R. J., Subbotin, D. A., &  
1179 Balikhin, M. A. (2014). Simulation of high-energy radiation belt electron fluxes  
1180 using NARMAX-VERB coupled codes. *Journal of Geophysical Research: Space*  
1181 *Physics*, 119(10), 8073–8086. doi: <https://doi.org/10.1002/2014JA020238>
- 1182 Paulikas, G., & Blake, J. (1979). Effects of the solar wind on magnetospheric  
1183 dynamics: Energetic electrons at the synchronous orbit. In *Quantitative mod-*  
1184 *eling of magnetospheric processes* (p. 180–202). American Geophysical Union  
1185 (AGU). doi: <https://doi.org/10.1029/GM021p0180>
- 1186 Reeves, G. D., Morley, S. K., Friedel, R. H. W., Henderson, M. G., Cayton, T. E.,  
1187 Cunningham, G., ... Thomsen, D. (2011). On the relationship between  
1188 relativistic electron flux and solar wind velocity: Paulikas and Blake re-  
1189 visited. *Journal of Geophysical Research: Space Physics*, 116(A2). doi:  
1190 <https://doi.org/10.1029/2010JA015735>
- 1191 Roeder, J. L., Chen, M. W., Fennell, J. F., & Friedel, R. (2005). Empirical models of  
1192 the low-energy plasma in the inner magnetosphere. *Space Weather*, 3(12). doi:  
1193 <https://doi.org/10.1029/2005SW000161>
- 1194 Rowland, W., & Weigel, R. S. (2012). Intracalibration of particle detectors on a  
1195 three-axis stabilized geostationary platform. *Space Weather*, 10(11). doi: 10  
1196 [.1029/2012SW000816](https://doi.org/10.1029/2012SW000816)
- 1197 Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative  
1198 than the ROC plot when evaluating binary classifiers on imbalanced datasets.  
1199 *PLOS ONE*, 10(3).
- 1200 Schaefer, J. T. (1990). The critical success index as an indicator of warning skill.  
1201 *Weather and Forecasting*, 5(4), 570. doi: 10.1175/1520-0434
- 1202 Shi, Y., Zesta, E., & Lyons, L. R. (2009). Features of energetic particle ra-  
1203 dial profiles inferred from geosynchronous responses to solar wind dy-  
1204 namic pressure enhancements. *Annales Geophysicae*, 27(2), 851–859. doi:  
1205 [10.5194/angeo-27-851-2009](https://doi.org/10.5194/angeo-27-851-2009)
- 1206 Sicard-Piet, A., Bourdarie, S., Boscher, D., Friedel, R. H. W., Thomsen, M., Goka,  
1207 T., ... Koshiishi, H. (2008). A new international geostationary electron  
1208 model: IGE-2006, from 1 keV to 5.2 MeV. *Space Weather*, 6(7). doi:  
1209 <https://doi.org/10.1029/2007SW000368>
- 1210 Sillanpää, I., Ganushkina, N. Y., Dubyagin, S., & Rodriguez, J. V. (2017). Elec-  
1211 tron fluxes at geostationary orbit from GOES MAGED data. *Space Weather*,  
1212 15(12), 1602–1614. doi: 10.1002/2017SW001698
- 1213 Simms, L. E., & Engebretson, M. (2020). Classifier neural network models predict  
1214 relativistic electron events at geosynchronous orbit better than multiple re-  
1215 gression or ARMAX models. *Journal of Geophysical Research: Space Physics*,  
1216 125(5), e2019JA027357. doi: <https://doi.org/10.1029/2019JA027357>
- 1217 Simms, L. E., Engebretson, M., Clilverd, M., Rodger, C., Lessard, M., Gjerloev, J.,  
1218 & Reeves, G. (2018a). A distributed lag autoregressive model of geostationary  
1219 relativistic electron fluxes: Comparing the influences of waves, seed and source  
1220 electrons, and solar wind inputs. *Journal of Geophysical Research: Space*  
1221 *Physics*, 123, 3646–3671. doi: <https://doi.org/10.1029/2017JA025002>
- 1222 Simms, L. E., Engebretson, M., & Reeves, G. (2022). Removing diurnal sig-  
1223 nals and longer term trends from electron flux and ULF correlations: a  
1224 comparison of spectral subtraction, simple differencing, and ARIMAX



- 1225 models. *Journal of Geophysical Research*, 127(2), e2021JA030021. doi:  
1226 <https://doi.org/10.1029/2021JA030021>
- 1227 Simms, L. E., Engebretson, M. J., Clilverd, M. A., Rodger, C. J., & Reeves,  
1228 G. D. (2018b). Nonlinear and synergistic effects of ULF Pc5, VLF Cho-  
1229 rus, and EMIC waves on relativistic electron flux at geosynchronous orbit.  
1230 *Journal of Geophysical Research: Space Physics*, 123(6), 4755-4766. doi:  
1231 <https://doi.org/10.1029/2017JA025003>
- 1232 Simms, L. E., Engebretson, M. J., Pilipenko, V., Reeves, G. D., & Clilverd, M.  
1233 (2016). Empirical predictive models of daily relativistic electron flux at  
1234 geostationary orbit: Multiple regression analysis. *Journal of Geophysical*  
1235 *Research: Space Physics*, 121(4), 3181-3197. doi: [https://doi.org/10.1002/](https://doi.org/10.1002/2016JA022414)  
1236 [2016JA022414](https://doi.org/10.1002/2016JA022414)
- 1237 Simms, L. E., Engebretson, M. J., Rodger, C. J., Gjerloev, J. W., & Reeves, G. D.  
1238 (2019). Predicting lower band chorus with autoregressive-moving average  
1239 transfer function (ARMAX) models. *Journal of Geophysical Research: Space*  
1240 *Physics*, 124, 5692 - 5708.
- 1241 Simms, L. E., Ganushkina, N. Y., van de Kamp, M., Liemohn, M. W., & Dubya-  
1242 gin, S. (2022). Using ARMAX models to determine the drivers of 40-  
1243 150 keV GOES electron fluxes. *Journal of Geophysical Research*, 127(9),  
1244 e2022JA030538. doi: 10.1029/2022JA030538
- 1245 Simms, L. E., Pilipenko, V., Engebretson, M. J., Reeves, G. D., Smith, A. J., &  
1246 Clilverd, M. (2014). Prediction of relativistic electron flux at geostationary  
1247 orbit following storms: Multiple regression analysis. *Journal of Geophysical*  
1248 *Research: Space Physics*, 119(9), 7297-7318. doi: [https://doi.org/10.1002/](https://doi.org/10.1002/2014JA019955)  
1249 [2014JA019955](https://doi.org/10.1002/2014JA019955)
- 1250 Smirnov, A. G., Berrendorf, M., Shprits, Y. Y., Kronberg, E. A., Allison, H. J.,  
1251 Aseev, N. A., & et al. (2020). Medium energy electron flux in earth's  
1252 outer radiation belt (MERLIN): a machine learning model. *Space Weather*,  
1253 18(e2020SW002532). doi: 10.1029/2020SW002532
- 1254 Smith, G. (2018). Step away from stepwise. *Journal of Big Data*, 5(32). doi:  
1255 <https://doi.org/10.1186/s40537-018-0143-6>
- 1256 Stepanov, N. A., Sergeev, V. A., Sormakov, D. A., Andreeva, V. A., Dubyagin,  
1257 S. V., Ganushkina, N., & et al. (2021). Superthermal proton and electron  
1258 fluxes in the plasma sheet transition region and their dependence on solar  
1259 wind parameters), journal = *Journal of Geophysical Research: Space Physics*,  
1260 volume = 126, e2020JA028580, doi = 10.1029/2020JA028580.
- 1261 Subbotin, D. A., & Shprits, Y. Y. (2009). Three-dimensional modeling of the ra-  
1262 diation belts using the Versatile Electron Radiation Belt (VERB) code. *Space*  
1263 *Weather*, 7(10). doi: <https://doi.org/10.1029/2008SW000452>
- 1264 Swiger, B. M., Liemohn, M. W., Ganushkina, N. Y., & Dubyagin, S. (2022). Ener-  
1265 getic electron flux predictions in the near-earth plasma sheet from solar wind  
1266 driving. *Space Weather*, 20(e2022SW003150). doi: 10.1029/2022SW003150
- 1267 Thomsen, M. F., Henderson, M. G., & Jordanova, V. K. (2013). Statistical prop-  
1268 erties of the surface-charging environment at geosynchronous orbit. *Space*  
1269 *Weather*, 11(5), 237-244. doi: 10.1002/swe.20049
- 1270 Tofallis, C. (2015). A better measure of relative prediction accuracy for model selec-  
1271 tion and model estimation. *Journal of the Operational Research Society*, 66(8),  
1272 1352-1362. doi: <https://doi.org/10.1057/jors.2014.103>
- 1273 Whittingham, M., Stephens, P., Bradbury, R., & Freckleton, R. (2006). Why do we  
1274 still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecol-*  
1275 *ogy*, 75(5), 1182-9. doi: 10.1111/j.1365-2656.2006.01141.x.
- 1276 Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagno-  
1277 sis, with special reference to X-Ray techniques. *Public Health Reports*, 62(40),  
1278 1432-1449. doi: 10.2307/4586294

## Figure Captions

Figure 1. Predictions from the Base ARMA model (40 keV) over the training set (predictions in green) and the validation period (predictions in orange). Flux is converted to unitless Z-scores.

Figure 2. Predictions over the validation period (40 keV) from a. Base ARMA model, b. ARMAX-MLT, c. ARMAX-MLT with flux at lag 6 added as a predictor, d. ARMAX-MLT with flux at lag 1 added as a predictor, e. ARMAX-MLT with flux at lag 6 added as a predictor. Flux is converted to unitless Z-scores. Further metrics are given in Figure 4 and Table 2.

Figure 3. Predictions over 1 week of the validation period (40 keV) showing how the models incorporating Lag 1 flux lag behind observations. a. Base ARMA model, b. Base ARMA with lag 1 flux, c. ARMAX-MLT, d. ARMAX-MLT with flux at lag 1, e. ARMAX-MLT with flux at lag 6. Flux is converted to unitless Z-scores. Lag 1 validation  $r$  correlates current observations with the prediction one hour previous. Models with Lag 1 flux as a predictor have higher Lag 1 validation correlation than same time validation correlation.

Figure 4. Scatterplots of predictions vs. observations over the full validation period and all three energies (a. 40 keV, b. 75 keV, c. 150 keV). Row 1: Base ARMA model, Row 2: ARMA model split by MLT (ARMAX-MLT), Row 3: ARMAX-MLT with lag 6 flux added as an additional predictor. Red line shows the ideal 1:1 correspondence between predictions and observations. Orange line gives the cubic fit to the actual prediction-observation relationship. Flux is converted to unitless Z-scores.

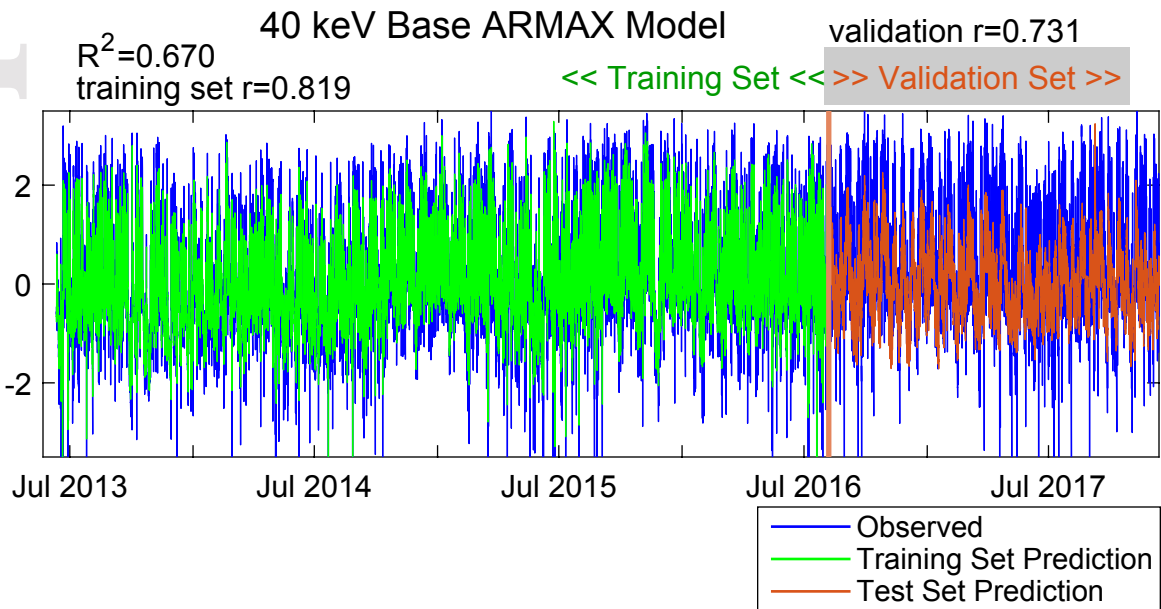
Figure 5. Scatterplots of predictions vs. observations for RNN models over the full validation period and all three energies (40 keV - 150 keV). First row: Base RNN model, second row: RNN-MLT. Red line shows the ideal 1:1 correspondence between predictions and observations. Orange line gives the cubic fit to the actual prediction-observation relationship. Flux is converted to unitless Z-scores.

Figure 6. Scatterplots of predictions vs. observations for the REG-MLT model over the full validation period and all three energies (a. 40 keV, b. 75 keV, c. 150 keV). Red line shows the ideal 1:1 correspondence between predictions and observations. Orange line gives the cubic fit to the actual prediction-observation relationship. Flux is converted to unitless Z-scores.

Figure 7. Validation correlation coefficients of the value-predicting models (base ARMAX, ARMAX-MLT, base RNN, RNN-MLT, and REG-MLT). Although 95% confidence intervals around each correlation are small, there is little practical difference between the models using this metric.

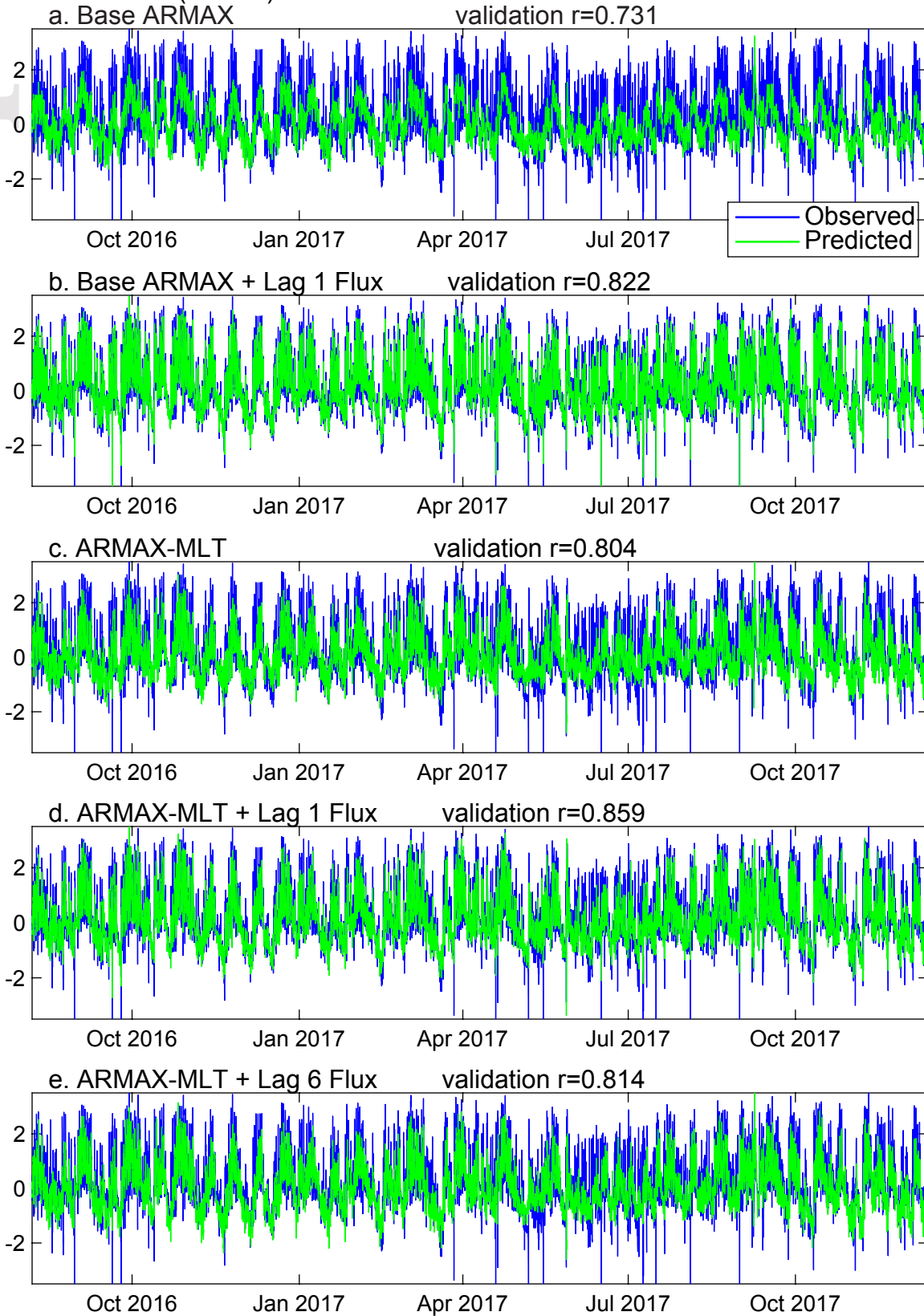
Figure 8. Heidke skill scores of the classifier models (LogisticLag1, LogisticLag3, LogisticLag6, and RNNclass-MLT). There is little difference between the models using this metric. Bootstrap confidence intervals (95%) are small.

Figure 9. Determining the optimal probability threshold (red circle) for the logistic Lag 3 (150 keV) model. a. ROC curve, b. precision-recall curve, c. classification of test set using the optimal threshold from ROC curve, d. classification of test set using optimal threshold from the precision-recall curve. Model predicts whether observation will be above or below the 75th percentile. AUC = area under the ROC curve, TPR = true positive rate, FPR = false positive rate, TNR = true negative rate, FNR = false negative rate.

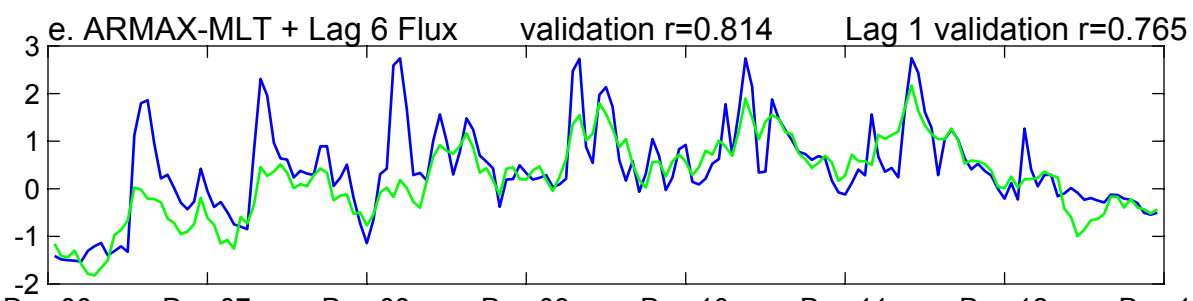
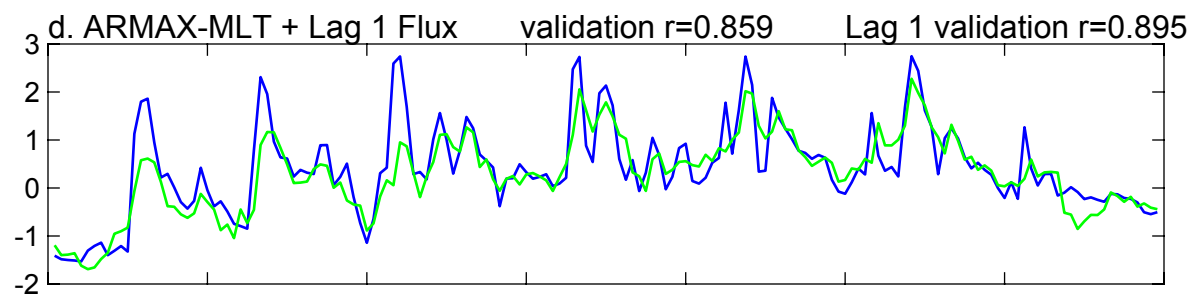
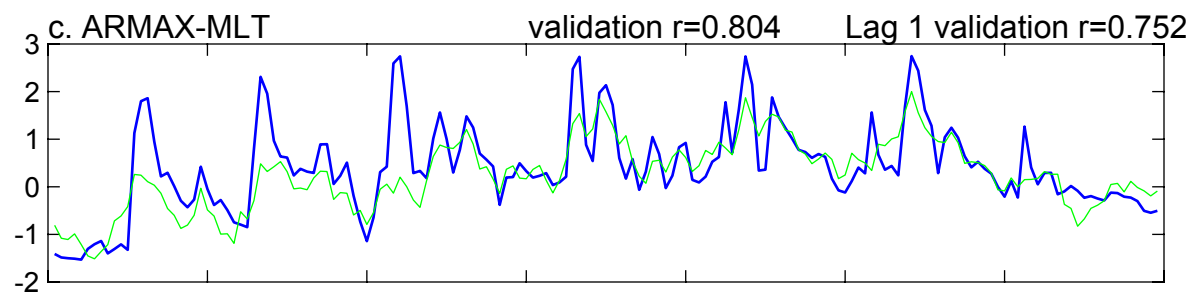
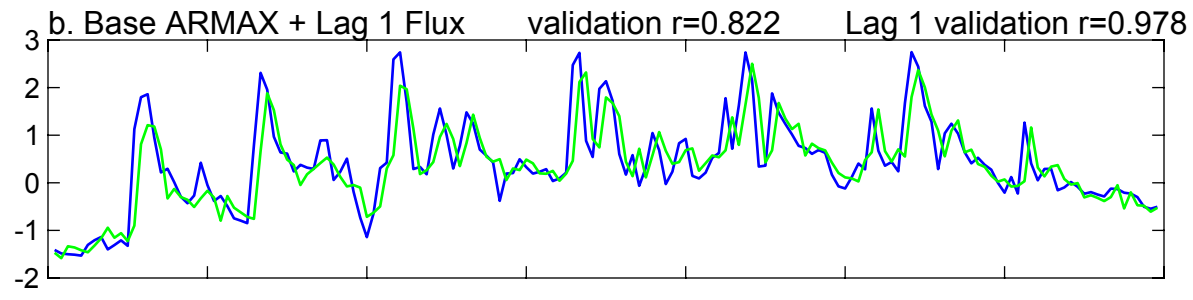
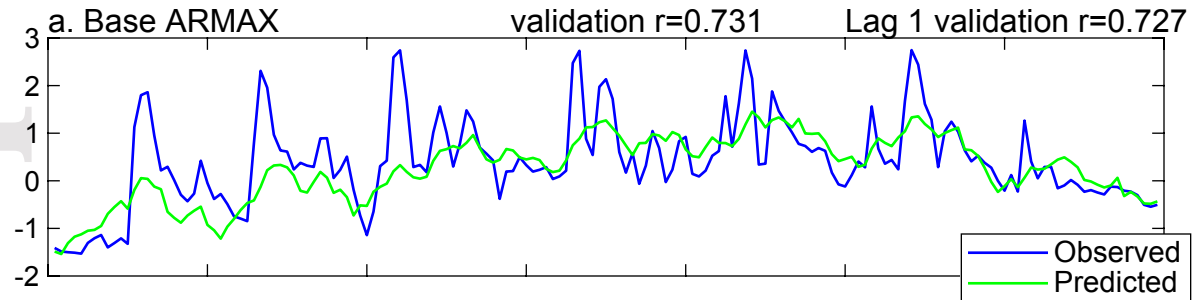


Validation Set (40 keV)

40 keV Electron Flux (Z Scores)



1 Week of Validation Set (40 keV)



Dec 06 Dec 07 Dec 08 Dec 09 Dec 10 Dec 11 Dec 12 Dec 13  
2016

