Article:
Li, P., Wei, H.-L., Boynton, R.J. et al. (2026) Modeling and prediction of electron fluxes in GEO with NARMAX approach using data set with missing data points. Advances in Space Research. ISSN: 0273-1177

https://doi.org/10.1016/j.asr.2026.05.093

Accepted for publication by *Advances in Space Research* on 28 May 2026; Available online in the journal on 2 June 2026. DOI: 10.1016/j.asr.2026.05.093

Modeling and Prediction of Electron Fluxes in GEO with NARMAX Approach Using Data Set with Missing Data Points

Ping Li[1]*, Hua-Liang Wei[1]*, Richard J. Boynton[1], and Michael A. Balikhin[1]

[1]School of Electrical and Electronic Engineering, University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK

Abstract

NARMAX approach, originally developed for identifying models for complex nonlinear dynamic systems, has been used for space system and environment modelling and space weather forecasting for more than a decade. In this study, the NARMAX approach is employed to build predictive models for 1.5 MeV electron fluxes in geostationary Earth orbit (GEO). A challenge in modelling space weather is that there are many gaps due to missing data points in the available dataset for modelling, and we usually end up building the model with some imputed data to fill the gaps. The challenge is addressed in this paper. We propose to segment the data set with missing data point and reformulate the task as a modelling problem from multiple data sets and then introduce an effective model identification scheme that allows us to identify a single model directly from multiple datasets without resorting to any data interpolation or imputation techniques. The models are identified using two different model term selection methods and the prediction performance between models will then be compared to providing some guidelines for choosing modelling methods. We will also discuss best practices, limitations and caveats to consider when building prediction models from a given/or an available data set with missing data points.

1 Introduction

High levels of the energetic electrons can have harmful effect on modern technological systems, particularly satellites in geostationary Earth orbit (GEO) and may lead to temporary or permanent loss of system functions, it can also be hazards for humans in space (Reagan *et al*., 1983; Baker *et al*.,1987). The prior knowledge about fluences of energetic electrons can help to mitigate the adverse effect of the high fluxes of the energetic electrons, and this will require the reliable predictions of electron fluxes in space around the satellite orbits. The electron flux is known to evolve with the changing solar wind parameters. Over the last two decades, a lot of effort has been devoted to develop modelling methods for making reliable predictions using the observed solar wind parameters. The approaches used for modelling can roughly be divided into two categories, i.e. i) the first principles-based approach where the prediction model was derived from the first principles/or physical laws that govern the dynamical behaviours of geospace

* p.li@sheffield.ac.uk, w.hualiang@sheffield.ac.uk

environment (see *e.g*. Friedel *et al*. 2002); and ii) the data-driven approach where the prediction model was directly derived from the measured electron flux and solar wind parameter data.

The first principles-based approach can be viewed as a "white-box" method. To build model with such an approach, comprehensive knowledge on the physical mechanisms that govern the dynamics of space weather will be required. Due to the huge complexity of space environment evolution processes, it is not easy to obtain a reliable physics-based model for prediction. As such, and also as rapid increase in the computational power of the modern computers, the data-driven approach as an alternative has attracted much attention in recent two decades. This approach can be further divided into two categories, i.e. the machine learning-based method and system identification-based method. The machine-learning method makes use of recent development in artificial intelligence to build prediction models (see *e.g*. Wei *et al.* 2018, Sun *et al*. 2024, Zou *et al*. 2024, 2026). It is essentially a "black-box" method and the model is built by learning from large data sets without resort to the knowledge on the complicated physical mechanisms governing the space weather. One shortcoming of the model built with this approach is the lack of interpretability and, as the "black-box" suggested, the model is usually opaque to the end-users. System identification-based method (which can be viewed as a "grey-box" method), on the other hand, can build parametric models (see *e.g.* Boynton *et al.* 2015, Simms *et al*. 2023, 2024) and these models are transparent to the end-users, hence allow users to analyze the influences of different solar wind parameters or their combinations on the response variable (*i.e.* electron flux) as shown in Balikhin *et al*. (2011) and Boynton *et al*. (2013).

In this paper we will focus on system identification-based methods, in particular the NARMAX approach (Billings, 2013) to build models for 1.5 MeV electron flux prediction. The NARMAX method was initially developed for identifying models for complex nonlinear dynamic systems (Leontaritis & Billings, 1985) and has since been employed for modelling of complex systems in many fields. One of the main challenges in nonlinear dynamic model identification is the model structure detection. With NARMAX approach, the problem is solved by using a linear-in-the parameters parameterization for the model in conjunction with the orthogonal forward regression (OFR) (Billings *et al*. 1989, Li *et al*. 2013) for model term selection where the most significant model terms are identified by a stepwise regression procedure. This method has been applied for space weather forecasting, particularly electron flux modelling, in recent decade. For example, the NARMAX-OFR algorithm had been used for analyzing the influences of different solar wind parameters and their combinations on the electron flux and to determine the most influential coupling functions of solar wind parameters to be used in the forecasting models in Balikhin *et al*. (2011) and Boynton *et al*. (2013). In Wei *et al*. (2011) and Boynton *et al*. (2015), the NARMAX-OFR algorithm was used to identify models for forecasting electron fluxes at various energy levels. There are two popular criteria used in an OFR algorithm for model term selection, i.e. the error reduction ratio (ERR) and the predicted residual sums of squares (PRESS) statistic (Li *et al*. 2013), and the NARMAX modelling research for prediction of electron flux in the aforementioned studies were all carried out using the classical ERR-based OFR method. A challenge in 1.5 MeV electron flux modelling considered in this paper is that there are many missing data points in the available dataset for modelling and this challenge will be addressed in this paper. In addition to ERR-based OFR, the PRESS-based OFR will also be used to build the prediction model and the main contributions of this paper

include: i.) introduce an effective model identification scheme that can identify a model directly from a data set with missing data points without resorting to any data imputation techniques and this is achieved by reformulating the modelling task with missing data points as a problem of identifying a single model from multiple data sets and ii.) compare the prediction performance between models identified with two different model term selection methods (i.e. ERR-based OFR and PRESS-based OFR) to provide some guidelines for choosing modelling algorithms. We will also discuss the best practices, limitations and caveats to consider when building prediction models from a given/or an available data set with missing data points.

2 Methodology

2.1 NARMAX model

Assume the behaviour of a nonlinear dynamical system output (denoted by y) is dependent on or correlated to a total of *r* inputs denoted by $u_1, u_{2,} \cdots, u_r$. With the NARMAX approach, the relationship between the output y and the inputs $u_i, (i = 1,2, \cdots, r)$ can be represented by the following NARMAX model (Billings, 2013):

$$y(t) = f\Big(y(t-1), \cdots, y\big(t-n_y\big), u_1(t-1), \cdots, u_1(t-n_u), \cdots, u_r(t-1), \cdots, u_r(t-n_u), \cdots, e(t-1), \cdots, e(t-n_e)\Big) + e(t) \quad (1)$$

where $u_i(t), (i = 1,2, \cdots, r)$ and $y(t)$ are the values of system inputs and output observed at time instant t, $e(t)$ is the noise sequence, and n_y, n_u, n_e are the maximum time lags associated with output, inputs and noise respectively. $f(\cdot)$ is an unknown function that needs to be identified from the available observations. The noise signal $e(t)$ cannot be measured in real applications but can, in practice, be approximated with the model prediction error, i.e. $e(t) \approx y(t) - \hat{y}(t)$ where $\hat{y}(t)$ is the model prediction at time instant t. The moving averaging (MA) elements $e(t-1), \cdots, e(t-n_e)$ are used for noise estimation and model refinement in model building process, will be removed in prediction stage.

Various possibilities of parameterizing function $f(\cdot)$ in (1) exist and they can be classified into two categories based on the relationship between the model output and the adjustable model parameters $\boldsymbol{\theta}$, namely, nonlinear-in-the parameters and linear-in-the parameters. To simplify model identification, in particular to facilitate model structure determination or selection of a parsimonious model, linear-in-the parameters model is commonly used. Specifically with the NARMAX methodology, the model is constructed as a linear combination of some nonlinear basis functions (*e.g.* polynomials, radial basis functions, *etc.*) which depend on the vector $\boldsymbol{x}(t) = [y(t-1), \cdots, y\big(t-n_y\big), u_1(t-1), \cdots, u_1(t-n_u), \cdots, u_r(t-1), \cdots, u_r(t-n_u), \cdots, e(t-1), \cdots, e(t-n_e)]^T$ and the resulting model can be written as follows:

$$y(t) = \sum_{i=1}^{m} \theta_i p_i\big(\boldsymbol{x}(t)\big) + e(t) \quad (2)$$

where p_i, $p_i\big(\boldsymbol{x}(t)\big)$ and θ_i $(i = 1,2, \cdots, m)$ are the basis functions, regressors/or model terms and the associated parameters respectively. In this study, polynomial basis function will be used and the resulting polynomial NARMAX model is formed by the sum of the

monomials of $\boldsymbol{x}(t)$ with each model term $p_i(\boldsymbol{x}(t))$, $(i = 1,2,\cdots,m)$ being a monomial of some lagged inputs, outputs and prediction errors selected from the full candidate model term set of size M $(M)m)$ that includes all possible products of lagged inputs, outputs and errors up to the specified nonlinear degree.

2.2 Model term selection by OFR

A basic principle in practical nonlinear modeling is the parsimonious principle which requires that the model should be no more complex than is required to capture the underlying system dynamics. As described previously, with the NARMAX modeling methodology, the initial full regression model constructed by all the M terms in the full candidate model term set may be very complex and can contain many redundant model terms. Therefore, some model structure determination procedures will be needed so as to select the important terms to be included in the model and eliminate those that are deemed to be irrelevant. The NARMAX modeling methodology is centered around the belief that the determination of model structure is a crucial part of the identification process. With the linear-in-the-parameters model defined by (2), the problem is tackled with the orthogonal forward regression (OFR) algorithm (Billings *et al*. 1989, Li *et al*. 2013). The basic idea behind the OFR is to decouple the contributions of individual regressors (i.e. model terms) to the model's "goodness" measure by orthogonally decomposing the regression matrix so that we can evaluate the contribution of each individual model term to the model "goodness" measure separately. This will allow us to select individual model terms based on their contributions to the goodness measure and the core component in an OFR algorithm is the routine for stepwise orthogonalization of the regression matrix and forward selection of the relevant model terms.

The two most popular "goodness" measures used in an OFR algorithm for model term selection are the error reduction ratio (ERR) (Billings *et al*. 1989) and the predicted residual sums of squares (PRESS) statistic (Wang & Cluett 1996, Hong *et al*. 2003). With the ERR-based OFR algorithm, the model terms are selected based on their ability to explain the desired output variance, which is measured by the ERR, hence the model terms are selected by maximizing ERR. Whereas with the PRESS-based OFR algorithm, the model terms are selected based on their generalization capability, which is evaluated through leave-one-out cross validation (see, e.g. Efron & Tibshirani 1993). Briefly, the idea behind leave-one-out cross validation is as follows. Suppose that there are N data points in the training data set used for modelling, each time one data point in the training set is moved out and then predicted using the model estimated with the remaining $N-1$ data points in the training set. The difference between this prediction and the original data point that has been moved out is known as the predicted residual, which is the true prediction error because the data point used for calculating the prediction error has been set aside and not been used for estimating model. We can repeat this procedure across N data points in the training set and obtain N predicted residuals for calculating PRESS statistic. The model terms are then

Accepted for publication by *Advances in Space Research* on 28 May 2026; Available online in the journal on 2 June 2026. DOI: 10.1016/j.asr.2026.05.093

selected by minimizing PRESS error. A brief introduction to both ERR-based OFR and PRESS-base OFR for NARMAX model identification can be found in Li *et al*. (2013).

3 Data set description

A full description of the data to be used in this study for electron flux modelling is summarized in Table 1. The data set contains the observations of five variables, where electron flux y is the output of the system and the four solar wind parameters V, n, p and Bs are used as the system inputs. Note that our primary intention in this study is to build models for predicting the electron flux directly from the solar wind "source" parameters that can eventually be used in real time. As such, the magnetospheric indices (e.g. AE, Kp) are not considered, since they are strongly correlated with solar wind parameters and some indices, such as AE, is not available in real time. Exclusion of these indices will ensure the models identified remain purely and directly driven by upstream solar wind conditions and all the inputs are available in real time. It also avoids redundant feature/inputs and reduces correlation between input variables.

Table 1. Description of variables for modelling

Variable	Description
y	Daily averaged electron flux at 1.5MeV $(\text{electrons}/(\text{cm}^2 \cdot \text{sr} \cdot \text{s} \cdot \text{keV}))$
V	Solar wind velocity (km/s)
n	Solar wind density ($/\boldsymbol{cm^3}$)
p	Solar wind dynamic pressure (nPa)
Bs	Southward solar wind magnetic field (nT)

The observations are daily-average data covering the period from 1 Jan. 2018 to 31 Aug. 2024. There are intermittent data gaps in this data set caused by the missing data points in both electron flux and solar wind data, hence the entire data set were divided into 12 uninterrupted data segments with the total data points available for modelling being 2376 and the longest uninterrupted data segment being 849. As the chronological order of data points is important for dynamical signals, each uninterrupted data segment can be viewed as a single data set for model identification and they are summarized in Table 2, where the data lengths for each data segment are shown in the last column of the table. The data ranges (i.e. the maximum and the minimum values) for each of the five variables across 12 data segments are also given in the table and it was noticed that there were some extreme values for some inputs (highlighted with bold font) in data segment 12 that were a

bit far from the values in the other data segments. This information will be used for selection of data segments for model training, validation and performance analysis.

Table 2. Data used for modelling and performance analysis

Seg No.	V (km/s)		n(cm^{-3})		p(nPa)		Bs(nT)		y		No. of data
	min	max	min	max	min	max	min	max	min	max	
1	289.2535	651.2968	1.77957	22.1635	0.81132	6.4604	0.071238	4.4842	-0.88918	1.8737	189
2	299.6977	602.9586	1.87756	18.0583	0.92701	5.078	0.024013	8.8988	-0.85701	1.9726	149
3	298.6745	623.0939	2.04018	19.633	0.79424	5.016	0.14121	2.7534	-0.6357	1.8094	99
4	268.0587	458.9028	3.41176	13.7844	1.1553	3.2304	0.08741	1.9289	-0.28099	0.9894	11
5	372.4963	495.2577	2.4345	7.032	0.90667	2.4626	0.85342	2.4664	0.1269	1.7028	12
6	296.8343	528.1686	2.53199	21.6149	0.86935	4.4627	0.000404	6.2388	-0.55603	1.4442	57
7	314.7619	597.5232	2.47398	15.9066	0.79668	4.2464	0.2263	2.5394	-0.91639	1.4755	41
8	298.3614	698.1846	2.16404	16.0551	0.87935	5.599	0.017462	2.6497	-0.82677	1.9974	196
9	275.3991	641.9162	0.99124	20.6585	0.22582	4.9282	0.002184	4.4818	-0.91562	2.0182	397
10	287.6011	779.5746	0.3212	32.243	0.15661	7.3669	0	8.517	-0.96007	1.8466	849
11	292.8708	696.0903	1.45861	21.9726	0.46197	8.9438	0.027979	5.989	-0.93634	1.6298	206
12	283.9701	879.0555	1.43925	27.2696	0.71118	**17.618**	0.011282	**14.239**	-1.1632	1.198	170

4 NARMAX models for forecasting and performance analysis

4.1 NARMAX model identification using a data set with missing data points

The NARMAX approach presented in Section 2 was employed to build models from the data described in the last section for 1.5MeV electron flux forecasting. With the original OFR algorithm for NARMAX model identification, a basic assumption on the data set used for modelling is that the data points are continuously collected in chronological order and the classical OFR algorithm will be able to identify a dynamic model from this single data set. However, due to the missing data points as mentioned in the previous Section, we will need to either identify a single uninterrupted data segment of enough length for the classical OFR

algorithm or we will end up with a problem identifying a single model from multiple data sets. If the number of missing data points is small (say less than 5), an ad hoc method by applying linear interpolation to the data gaps had been proposed to produce a data set of enough length for model identification with the classical OFR algorithms (Boynton et al. 2013). In this paper, we introduce an effective model identification scheme that allows us to identify a single model directly from multiple datasets to address the challenge caused by the data gaps due to the missing data points. The key to this model identification scheme is to use the criteria evaluated from the multiple datasets for model term selection. Specifically, with the classical OFR algorithm, a model term is selected by either maximizing the ERR (for ERR-based OFR) or minimizing the mean square PRESS error (for PRESS-based OFR) over the given single data set for modelling. Whereas, with the introduced identification scheme using, say, S data segments for modelling, a model term is selected by either maximizing the sum of weighted ERR (for ERR-based OFR) or minimizing the sum of the weighted mean square PRESS error (for PRESS-based OFR) across the S data segments used for modelling, where the S weights are determined by the lengths and observation noise levels of the S data segments respectively. The details of this identification scheme can be found in Li *et al.* (2013), and it can be used to the data set with any number of missing data points without resorting to any data interpolation or imputation techniques.

As the identified models are for one-day ahead forecasting, the feed-through inputs will not be considered in modelling, i.e. the input time lag $n_u \geq 1$ in the identified model (2). Previous modelling studies of electron flux at various energy levels in GEO (see *e.g.* Balikhin et al. 2011, Wei et al. 2011, Boynton et al. 2013, Boynton *et al.* 2015) have shown that models with a nonlinear degree of 2 generally work well and produce good predictions for logarithmic flux (log-flux) modelling and prediction. As for the choice of maximum time lags for the input and output variables (i.e. n_y and n_u), our previous experience and simulation results on the datasets used in this work suggested that $n_y = n_u = 3$ is among the top choices. Based on these, the maximum degree of the monomials $p_i(\boldsymbol{x}(t))$ in the models is set to 2 and the maximum time lag for inputs and output is set to 3 for NARMAX model identification in the following subsections, and with such a setting, the size of the full candidate model term set will be $M = 190$.

4.2 Forecasting performance assessment

In this paper, both ERR-based OFR and PRESS-based OFR are used for identifying models with different choices of data segment or combinations of data segments. The prediction performances of these models are then analyzed and compared so as to provide some guidelines for selection of modelling method and data segments to be used for building models for electron flux forecasting. Because the models are producing outputs (i.e. electron fluxes) that ideally should exactly match the observed values, the following three commonly used metrics are calculated (Boynton *et al*., 2015; Liemohn *et al*., 2018) to facilitate performance analysis and comparison:

- The coefficient of determination (also known as the prediction efficiency---PE) is the proportion of the variation in the dependent variable (in our case, the electron flux y) that is predictable by the model. It is a measure that quantifies the ability of the

model to predict the variation of the observed electron flux around the baseline model (in this study, the mean) and is calculated as follows:

$$PE = 1 - \frac{\sum_{k=1}^{K}\sum_{t=1}^{N_k}(y_t^{(k)} - \hat{y}_t^{(k)})^2}{\sum_{k=1}^{K}\sum_{t=1}^{N_k}(y_t^{(k)} - \bar{y}^{(k)})^2} \quad (3)$$

where, K is the total number of data segments for PE evaluation, $y_t^{(k)}$ denotes the observed electron flux at the time instant t for the kth data segment, $\bar{y}^{(k)}$ denotes the mean of the observed electron flux for the kth data segment, $\hat{y}_t^{(k)}$ denotes the predicted electron flux from model for the kth data segment. Note that, for $K > 1$, the PE calculated with equation (3) is the overall PE across K data segments. A PE of 1 indicates perfect agreement at all times, and PE less than or equal to zero indicates that the model does not provide useful predictions of the time variation of the observations.

- Correlation coefficient R which is used to indicate how well the model predicts the trends of electron flux. It is calculated as follows:

$$R = \frac{\sum_{k=1}^{K}\sum_{t=1}^{N_k}\left(y_t^{(k)} - \bar{y}^{(k)}\right)\left(\hat{y}_t^{(k)} - \bar{\hat{y}}_t^{(k)}\right)}{\sqrt{\left[\sum_{k=1}^{K}\sum_{t=1}^{N_k}\left(y_t^{(k)} - \bar{y}^{(k)}\right)^2\right]\left[\sum_{k=1}^{K}\sum_{t=1}^{N_k}\left(\hat{y}_t^{(k)} - \bar{\hat{y}}_t^{(k)}\right)^2\right]}} \quad (4)$$

  As indicated in Liemohn *et al.*, (2018), R should be above a user-defined threshold that means the specified requirement for the application. This is usually at least 0.5, perhaps even 0.7 or more, to convince users that the model is performing well.

- Root mean square error (RMSE) is a statistical metric widely used in forecasting to quantify prediction error. It calculates the square root of the average of squared differences between observations and model predictions:

$$RMSE = \sqrt{\frac{1}{\sum_{k=1}^{K} N_k}\sum_{k=1}^{K}\sum_{t=1}^{N_k}(y_t^{(k)} - \hat{y}_t^{(k)})^2} \quad (5)$$

4.3 Results from models identified with a single data segment

Model identification was carried out first using a single data segment, and the data segment 10 was used in this case because it is the longest single data segment available for modelling from Table 2. The NARMAX models defined by equation (2) were identified with both ERR-based OFR and PRESS-based OFR and then used to produce one-day ahead forecasts of electron fluxes over the time intervals across all data segments in Table 2. The model identified with the ERR-based OFR contains 14 model terms (i.e. $m = 14$) and the model identified with the PRESS-based OFR contains 13 model terms (i.e. $m = 13$). They are summarized in Table 3.

Accepted for publication by *Advances in Space Research* on 28 May 2026; Available online in the journal on 2 June 2026. DOI: 10.1016/j.asr.2026.05.093

Table 3. Model terms from both PRESS-based OFR and ERR-based OFR using single data segment 10 only

PRESS-based OFR			ERR-based OFR		
Term No $\boldsymbol{i}$	θ_i	Model terms $(p_i(\boldsymbol{x}(t)))$	Term No i	θ_i	Model terms $(p_i(\boldsymbol{x}(t)))$
1	-0.00026849	$y(t-1)V(t-1)$	1	-0.00031571	$y(t-1)V(t-1)$
2	9.1298e-05	$V(t-1)p(t-2)$	2	9.0011e-05	$V(t-1)p(t-2)$
3	0.0045594	$n^2(t-1)$	3	0.0027363	$n^2(t-1)$
4	0.00023719	$V(t-1)Bs(t-1)$	4	0.00080223	$V(t-1)Bs(t-1)$
5	1.0541	$y(t-1)$	5	-0.0697	$y(t-1)p(t-1)$
6	-0.13508	$y(t-1)Bs(t-1)$	6	1.1468	$y(t-1)$
7	-0.26125	$n(t-1)$	7	-0.00046686	$V(t-2)Bs(t-1)$
8	4.2396e-07	$V^2(t-1)$	8	-0.21304	$n(t-1)$
9	-0.028287	$y(t-2)n(t-1)$	9	0.00050507	$V(t-1)n(t-1)$
10	0.00073785	$V(t-1)n(t-1)$	10	-0.033078	$p^2(t-1)$
11	-0.0004137	$V(t-2)p(t-1)$	11	-0.12638	$y(t-1)Bs(t-1)$
12	0.057047	$y(t-2)Bs(t-2)$	12	-0.020102	$y(t-2)n(t-1)$
13	-0.015797	$n(t-1)p(t-1)$	13	0.047664	$y(t-2)Bs(t-2)$
			14	-0.011037	$Bs^2(t-1)$

As can be seen in Table 3, seven common model terms have been picked out by both PRESS-based OFR and ERR-based OFR, in particular, the first four most significant model terms picked out by both methods are exactly the same (as shown in the first four term rows in Table 3). To compare the prediction performance of two models, the prediction efficiency, correlation

coefficient and RMSE defined by equations (3), (4) and (5) for one-day ahead forecasting from two models are evaluated over all data segments and the results are summarized in Table 4.

Table 4. Prediction efficiency, correlation coefficient and RMSE of two models identified using a single data segment 10 only

Seg No. k	Prediction efficiency (PE)		Correlation coefficient (R)		RMSE		Training/testing
	PRESS-OFR	ERR-OFR	PRESS-OFR	ERR-OFR	PRESS-OFR	ERR-OFR	
1	0.84285	0.84583	0.92824	0.93075	0.27454	0.27192	testing
2	0.85197	0.84313	0.92736	0.92384	0.28939	0.2979	testing
3	0.83249	0.83504	0.91762	0.91962	0.25109	0.24917	testing
4	0.37623	0.40306	0.7471	0.7581	0.30947	0.30274	testing
5	0.35718	0.37469	0.6908	0.69701	0.39246	0.38708	testing
6	0.70218	0.67175	0.85406	0.83812	0.27725	0.29107	testing
7	0.90776	0.92213	0.95745	0.96666	0.23414	0.21513	testing
8	0.86068	0.85803	0.93262	0.93277	0.26281	0.2653	testing
9	0.87381	0.87654	0.93886	0.94121	0.26666	0.26376	testing
10	0.80344	0.80605	0.89635	0.8978	0.29095	0.28901	**training**
11	0.69315	0.70635	0.83505	0.84269	0.31821	0.31129	testing
12	0.37494	**-0.1586**	0.69695	0.55901	0.44301	0.60314	testing
overall	0.80224	0.77718	0.89951	0.88602	0.29755	0.31583	

The first column in Table 4 indicates the data segment number as shown in Table 2, the last column in Table 4 indicates which data segment was used for model identification (training) and the last row of the table shows the overall PE, R and RMSE across all $K(=12)$ data segments calculated with equations (3), (4) and (5). As can be seen from Table 4, the models from both PRESS-based OFR and ERR-based OFR have comparable prediction performance (in terms of PE, R and RMSE) over most of the individual data segments. Figure 1 depicts

Accepted for publication by *Advances in Space Research* on 28 May 2026; Available online in the journal on 2 June 2026. DOI: 10.1016/j.asr.2026.05.093

the direct comparison between the observed electron flux (black) and the one-day ahead forecasting electron flux (red) from the models for data segment 8.

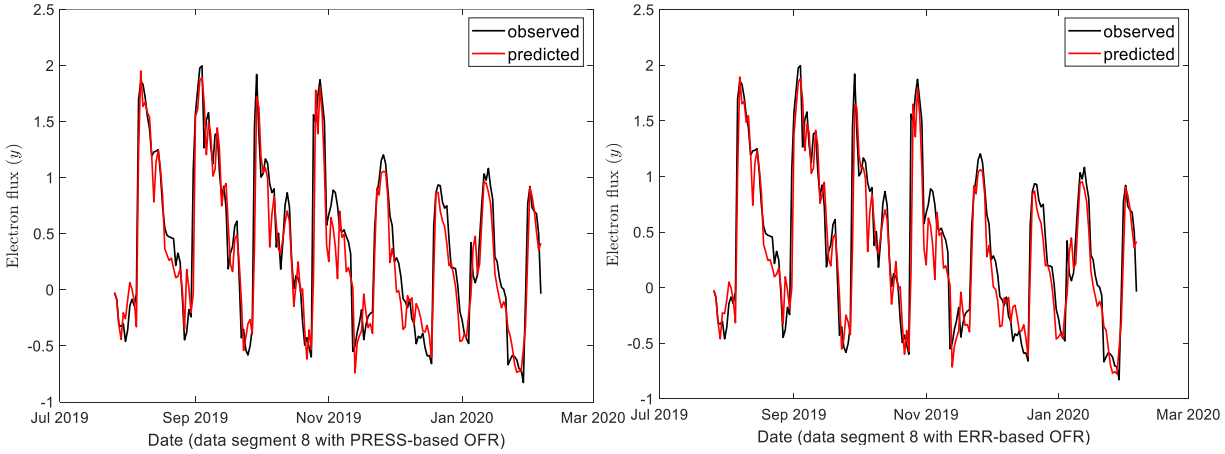

Figure 1. Comparison between observed electron flux (black) and the one-day ahead predictions (red) from two models identified with the single data segment 10 only for data segment 8 (date span from 26 July 2019 to 6 February 2020)

The red line in left figure shows the one-day ahead forecast from the model identified by PRESS-based OFR and the red line in right figure shows the one-day ahead forecast from the model identified by ERR-based OFR. Similar prediction performance of both models (note that both models were identified with the single data segment 10 only) can be observed in Figure 1. However, the difference in performance indices between two models for data segment 12 is significantly bigger than that for the other data segments as can be seen in Table 4 and the model from ERR-based OFR even has a negative prediction efficiency (highlighted with bold font in Table 4). This is essentially caused by some extreme input values in data segment 12 as mentioned in Table 2. These extreme inputs were far from the training inputs in data segment 10 used for building the models. To illustrate this, time series of the four input variables in data segment 12 are shown in Figure 2 below, and the direct

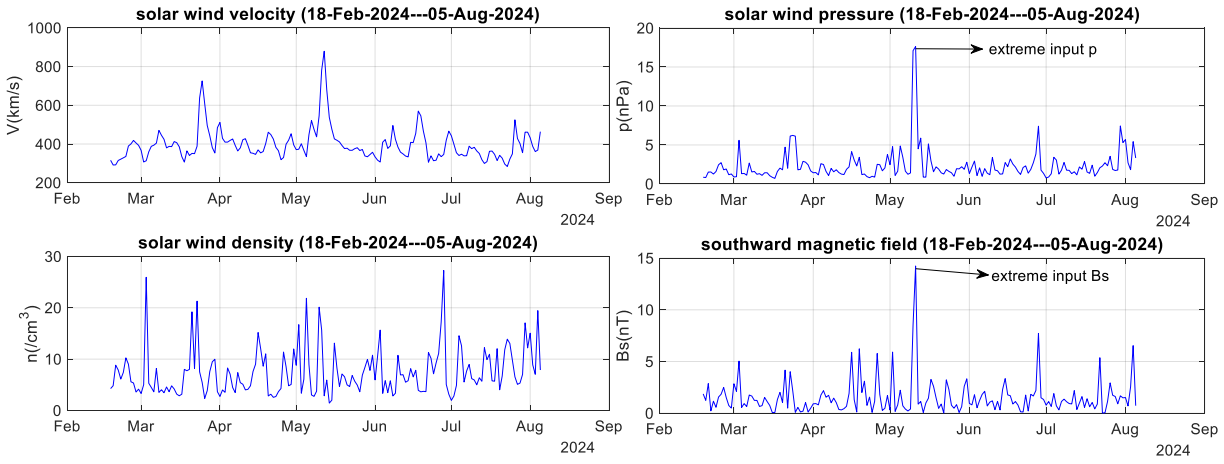

Figure 2. Time series of four input variables in data segment 12

comparison between the observed electron flux (black) and the one-day ahead forecasting electron flux (red) from the models for data segment 12 are shown in Figure 3.

The extreme input values appeared around 11 May 2024 for input variables p (solar wind dynamic pressure) and Bs (southward solar wind magnetic field) as shown in Figure 2 and the large prediction errors can be observed around the same date in Figure 3. It appears that

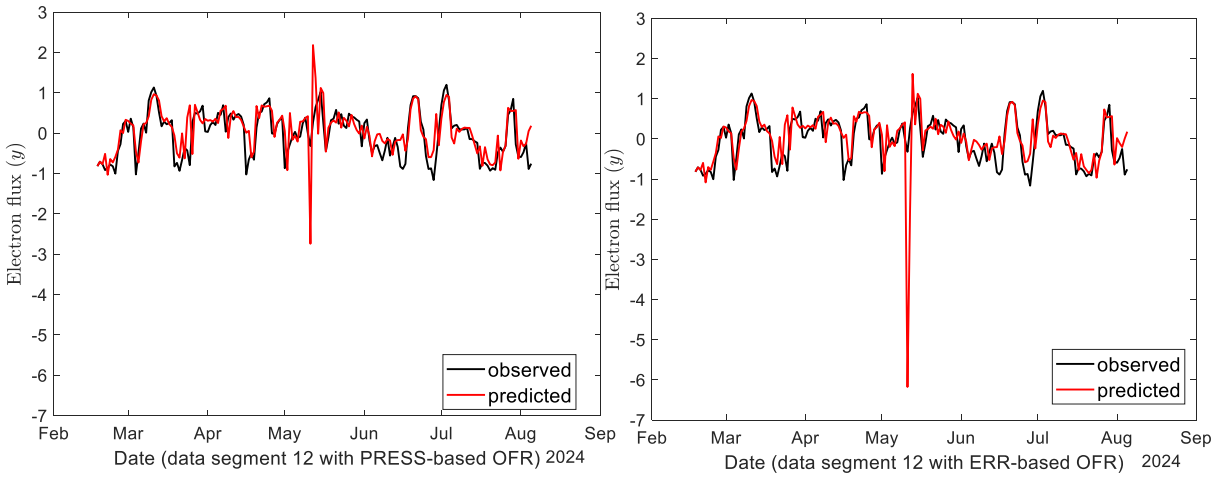

Figure 3. Comparison between observed electron flux (black) and the one-day ahead predictions (red) from two models identified with the single data segment 10 only for data segment 12 (date span from 18 February 2024 to 5 August 2024)

the model identified with PRESS-based OFR predicts better than the model identified with ERR-based OFR in the case where forecasting inputs are far from the training inputs as shown in Figure 3, where the prediction error around 11 May 2024 in the left graph of Figure 3 (from PRESS-based OFR) is much less than that in the right graph of Figure 3 (from ERR-based OFR). This is attributed to the emphasis being placed on the different aspects of the performance of the identified models by these two OFR algorithms. As described in Section 2, with the ERR-based OFR, the model terms are selected by maximizing the ERR which is equivalent to minimizing the mean squared residuals of the identified model over the training data (i.e., the data used for estimating the model), hence, the emphasis is placed on the model's quality of fit over the training data. Whereas with the PRESS-based OFR algorithm, the model terms are selected by minimizing a PRESS statistic which is a measure of the model prediction error on the data that has not been used for model estimation within the training data set; hence the emphasis is placed on the future prediction or generalization capability of the identified model. As such, we could expect that the models from PRESS-based OFR will predict better than the models from ERR-based OFR in the cases that the forecasting inputs are not seen in the training data set.

4.4 Results from models identified using multiple data segments

The results presented in previous subsection show that the inputs in the training data set used for model identification could have significant influence on the prediction performance of the identified model and the predicted behaviors from the model will be the most relevant with the inputs in training data set used for building the model. Therefore, to improve the robustness of prediction of the identified model, we need to ensure that the

Accepted for publication by *Advances in Space Research* on 28 May 2026; Available online in the journal on 2 June 2026. DOI: 10.1016/j.asr.2026.05.093

input values in the training data set cover all possible input ranges that may appear in real prediction applications. To this end, we will try to identify models using multiple data segments, specifically the data segments 10 and 12, in this subsection. The new model identification scheme introduced in Subsection 4.1 was used to build a single model directly from these two data segments (i.e. $S = 2$). Two models were built with ERR-based OFR and PRESS-based OFR respectively with the new identification scheme. The model identified with PRESS-based OFR contains 7 model terms and the model identified with ERR-based OFR contains 13 model terms. The first three most significant model terms from these two models were exactly the same as those from the models identified with the single data segment 10 (as shown in the first three model term rows in Tabel 3). These two models were then used for making one-day ahead predictions across all $K(= 12)$ data segments as before. The prediction performances (in terms of prediction efficiency, correlation coefficient and RMSE) of two models are summarized in Table 5 below.

Accepted for publication by *Advances in Space Research* on 28 May 2026; Available online in the journal on 2 June 2026. DOI: 10.1016/j.asr.2026.05.093

Table 5. Prediction efficiency, correlation coefficient and RMSE of two models identified using PRESS-based OFR and ERR-based OFR with two data segments 10 and 12

Seg No. k	Prediction efficiency ***(PE)***		Correlation coefficient ***(R)***		*RMSE*		Training/testing
	PRESS-OFR	ERR-OFR	PRESS-OFR	ERR-OFR	*PRESS-OFR*	*ERR-OFR*	
1	0.78766	0.81531	0.89714	0.92231	*0.31912*	*0.29762*	testing
2	0.80805	0.82947	0.90655	0.92077	*0.32953*	*0.3106*	testing
3	0.74693	0.77694	0.87443	0.89362	*0.30861*	*0.28974*	testing
4	0.43754	0.38653	0.70022	0.75749	*0.29386*	*0.3069*	testing
5	0.21502	0.36902	0.64304	0.70344	*0.43369*	*0.38883*	testing
6	0.70813	0.68978	0.85057	0.85504	*0.27447*	*0.28297*	testing
7	0.83129	0.87161	0.92035	0.93979	*0.31665*	*0.27623*	testing
8	0.78537	0.82874	0.88947	0.92061	*0.3262*	*0.29139*	testing
9	0.84977	0.8651	0.92632	0.93803	*0.29096*	*0.27571*	testing
10	0.74271	0.78175	0.86369	0.88473	*0.33287*	*0.30658*	**training**
11	0.65677	0.68693	0.81091	0.8308	*0.33654*	*0.32141*	testing
12	0.57921	0.66464	0.77986	0.82918	*0.36348*	*0.3245*	**training**
overall	0.76422	0.79715	0.8778	0.8995	*0.32489*	*0.30135*	

It can be seen from Table 5, both models have reasonable and similar prediction performance across 12 data segments (though overall, it appears that the model from ERR-based OFR performed slightly better than that from PRESS-based OFR in this case). In particular, the prediction performance of these two models for data segment 12 improved significantly in comparison to the models identified previously with data segment 10 only. To show these, the direct comparison between the observed electron flux (black) and the one-

Accepted for publication by *Advances in Space Research* on 28 May 2026; Available online in the journal on 2 June 2026. DOI: 10.1016/j.asr.2026.05.093

day ahead forecasting electron flux (red) from these two models for data segments 8 and 12 are shown in Figures 4 and 5 respectively.

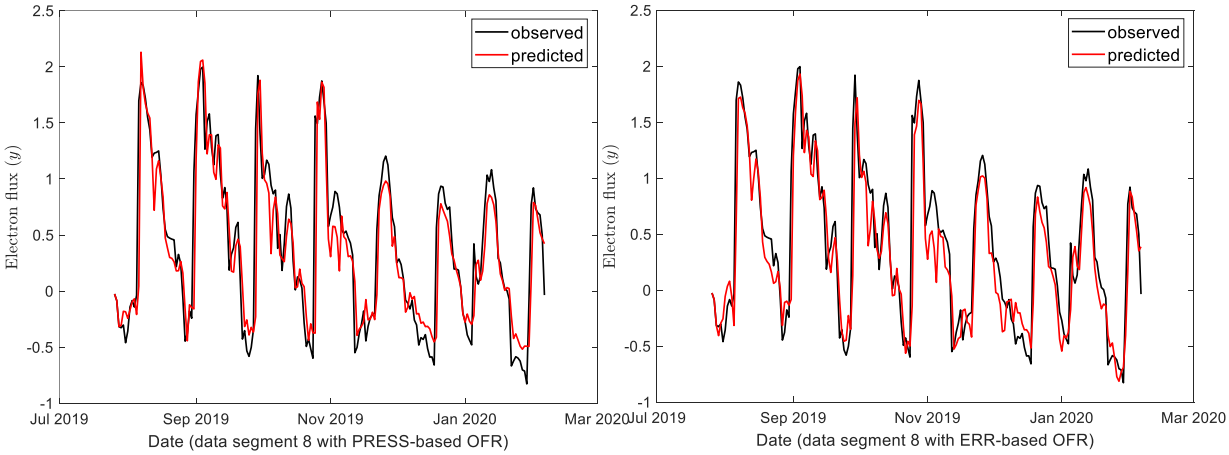

Figure 4. Comparison between observed electron flux (black) and the one-day ahead predictions (red) from two models identified with two data segments 10 and 12 for data segment 8 (date span from 26 July 2019 to 6 February 2020)

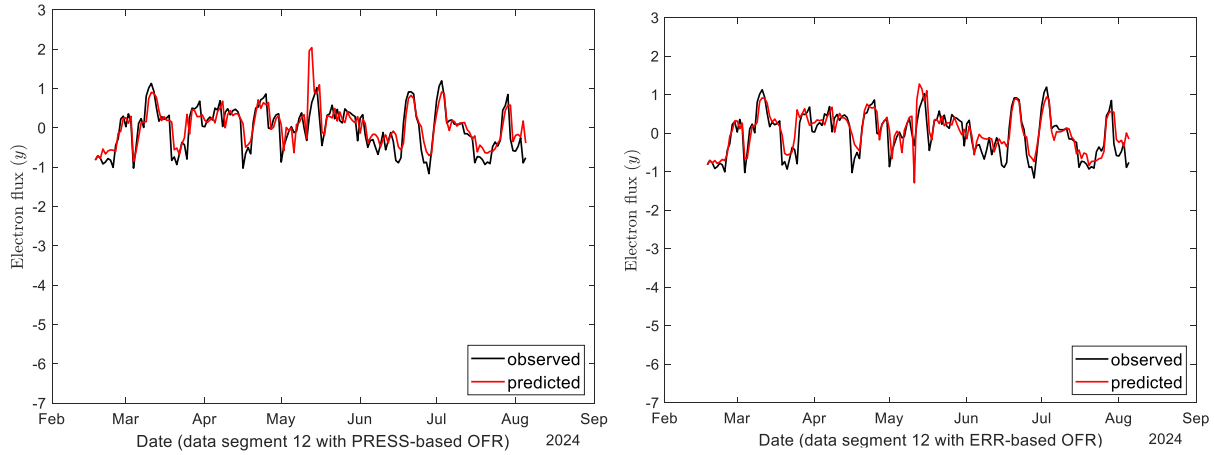

Figure 5. Comparison between observed electron flux (black) and the one-day ahead predictions (red) from two models identified with two data segments 10 and 12 for data segment 12 (date span from 18 February 2024 to 5 August 2024)

As can be seen from Figures 4 and 5, two models identified using PRESS-based OFR and ERR-based OFR respectively with two data segments 10 and 12 have similar prediction performance and the prediction errors at the extreme input data points (around 11 May 2024) are much smaller in this case as shown in Figure 5 in comparison with the previous case shown in Figure 3.

5 Discussion and Conclusions

NARMAX approach provides a powerful tool to build interpretable and parsimonious models for space weather forecasting. In comparison with the models built using other data-driven machine learning (ML) methods, the NARMAX models are transparent to end users, hence, in addition to predicting the future behaviour of the system, the NARMAX model can also provide useful information for the end-users to understand the mechanisms governing

the system's behaviour. It also requires less computation power to make prediction, and this makes them particularly suitable for real-time application. As such, NARMAX approach has found its way in space systems and environment modelling and been used for space weather forecasting over a decade. Two main objectives of this study are: i.) to address the issue of missing data which is a challenge frequently encountered by the modelers of space environment; ii.) to compare prediction performance of the models obtained with different NARMAX modelling strategies so as to provide some guidance on the best practices for building forecasting models from a given/or an available data set with missing data points.

To address the issue of missing data, we proposed to segment data set by the gaps due to missing data points first and each data segment can be viewed as an individual data set for modelling. We then re-formulated modelling task as a problem of identifying a single model from multiple data sets. To this end, a new model identification scheme that allows us to identify a single model from multiple datasets was introduced. With the new model identification scheme, we can identify a model directly from a data set with missing data points without resorting to any data imputation and interpolation techniques. This greatly simplifies the modelling process and extends the applicability of NARMAX approach for space environment modelling.

Two model term selection methods, i.e. PRESS-based OFR and ERR-based OFR, have been used in this study for NARMAX model identification with both single and multiple data segments. The prediction performance of the identified models was compared and some best practice advice, and caveats are summarized as follows:

- Models from PRESS-based OFR tend to predict better than models from ERR-based OFR in case where forecasting inputs are not present in training inputs as shown in Section 4.3. This would suggest that the PRESS-based OFR could be a better choice for model identification if the training data is limited and the input values in the training data set for modelling may not cover all possible ranges of the inputs that may appear in real forecast application.

- The diversity of excitation is important for model identification and richer excitation (inputs) in the training data set will help to improve the robustness of the predictions of the identified model as demonstrated in Section 4.4. This suggests that the training data or data segments need to be chosen such that the training inputs should be "rich" enough, containing a wide range of amplitudes and frequencies that cover all possible inputs that could be presented in real forecasting applications. For our particular application of predicting electron flux with the solar wind as input driver, this requires that the selected data segments should contain all types of solar wind structure that can present in the real forecasting application.

Note that no normalization was performed over input data in this study, since no evidence was found from simulation and our previous studies (see e.g. Wei *et al*. 2011) that normalization could obviously improve the prediction performance. Using the same training data, the models identified with the PRESS-based OFR were not the same as those identified with ERR-based OFR as shown in the previous Sections (though the most significant model terms picked out by both methods were exactly the same). This is because the two methods

put emphasis on different aspects of the performance of the identified model as explained in Section 4.3. However, it needs to be pointed out that, if the training data were generated from a true model, both PRESS-based OFR and ERR-based OFR should pick out the same model terms provided that the initial full candidate model term set (of size M) contains all the model terms in the true model that generated the data and the length of data is large enough as shown in Li et al. (2013). Nevertheless, the models from PRESS-based OFR often provide better generalization under unseen conditions in comparison with the models from ERR-based OFR. This is largely attributed to the fact that the PRESS-based OFR minimizes the out-of-sample error (which is the true prediction error) through leave-one-out cross validation. This contrasts with ERR-based OFR that minimizes the in-sample residual sum of squares over train data through maximizing ERR, which prone to overfitting, where the models may perform well on training data but poorly on unseen data. As such, the PRESS-based OFR will usually produce a more parsimonious model (i.e. a model with less model terms) than the ERR-based OFR, and we could expect that the models from PRESS-based OFR may predict better than the models from ERR-based OFR in the scenarios where the pattern of inputs for forecasting has not been shown in the training input set, whereas the models from ERR-based OFR may perform better than the models from PRESS-based OFR in the cases where all patterns of inputs for forecasting have been presented within the training input set used for modelling.

It also needs to be pointed out that, in this study, we treated all available data segments as equally important for modelling and analyzed the prediction performance of an identified model over all data segments. However, as indicated in Liemohn et al (2018), models are typically created with potential users in mind and each of those potential or real users may have specific needs for output prediction performance. For example, some might want to assess models against active times associated with some or a list of events only and may not care about the performance during quiet times. In such a case, selecting or placing more weight on the data segments over the active times for model identification and performance comparison may be more appropriate.

In this study, our attention was essentially focused on point prediction. Further research aiming to address the problem of quantifying uncertainty in the predictions from a NARMAX model is currently being carried out by the authors.

Acknowledgments

The authors would like to thank the UK Science and Technology Facilities Council (STFC) for supporting this work through grant ST/Y001524/1. The solar wind and electron flux data were sourced from the National Oceanic and Atmospheric Administration (NOAA) through https://data.ngdc.noaa.gov/platforms/solar-space-observing-satellites/goes/goes16/l1b/seis-l1b-mpsh_science/, which are gratefully acknowledged.

References

Reagan, J. B., Meyerott, R. E., Gaines, E. E., Nightingale, R. W., Filbert, P. C., & Imhof, W. (1983). Space charging currents and their effects on spacecraft systems, *Electrical Insulation*, *IEEE Transactions on, EI*-**18**, 354–365, **DOI:** 10.1109/TEI.1983.298625

Baker, D., Belian, R., Higbie, P., Klebesadel, R., & Blake, J. (1987) Deep dielectric charging effects due to high-energy electrons in earth's outer magnetosphere, *J. Electrostat*., **20**, 3–19. https://doi.org/10.1016/0304-3886(87)90082-9

Friedel, R., G. Reeves, & Obara, T. (2002), Relativistic electron dynamics in the inner magnetosphere – a review, *J. Atmos. Sol. Terr. Phys*., **64**(2), 265–282, doi:10.1016/S1364-6826(01)00088-8

O'Brien, T. P., Lorentzen, K. R., Mann, I. R., Meredith, N. P., Blake, J. B., Fennell, J. F., Looper, M. D., Milling, D. K., & Anderson, R. R. (2003) Energization of relativistic electrons in the presence of ulf power and mev microbursts: Evidence for dual ulf and vlf acceleration, *J. Geophys. Res*., **108**, 1329, doi:10.1029/2002JA009784.

Li, X. L., M. Temerin, D. N. Baker, G. D. Reeves, & Larson, D. (2001). Quantitative prediction of radiation belt electrons at geostationary orbit based on solar wind measurements, *Geophys, Res. Lett*., **28**(9), 1887–1890. https://doi.org/10.1029/2000GL012681

Wei, L., Zhong, Q., Lin, R., Wang, J., Liu, S., & Cao, Y. (2018). Quantitative prediction of high energy electron integral flux at geostationary orbit based on deep learning. *Space Weather*, **16**, 903–916. https://doi.org/10.1029/2018SW001829

Sun, X., Wang, D., Drozdov, A., Lin, R., Smirnov, A., Shprits, Y., Liu, S., Luo, B., & Luo, X. (2024), A modeling study of ≥2 MeV electron fluxes in GEO at different prediction time scales based on LSTM and transformer networks, *J. Space Weather Space Clim*. **14**, 25. https://doi.org/10.1051/swsc/2024021

Zou, Z., Zhang, L., Zuo, P., San, W., Yuan, Q., Zhu, B., & Hu, J. (2024), Global prediction of sub-relativistic and relativistic electron fluxes in the geosynchronous orbit using artificial neural networks, *Physics of Fluids*. **36**, 126638. https://doi.org/10.1063/5.0245593

Zou, Z., Zhang, L., Zuo, P., Wang, C., San, W., Hu, J., Shen, G., Sun, Y., & Hou, D. (2026), Predicting >2MeV electron fluxes in geosynchronous orbit using multiple satellite measurements, *Physics of Fluids*. **38**, 026608. https://doi.org/10.1063/5.0325198

Boynton, R. J., Balikhin, M. A., & Billings, S. A. (2015), Online NARMAX model for electron fluxes at GEO. *Ann. Geophys*., **33** 405-411. doi:10.5194/angeo-33-405-2015

Simms, L. E., Engebretson, M. J., & Reeves, G. D. (2023). Determining the timing of driver influences on 1.8–3.5 MeV electron flux at geosynchronous orbit using ARMAX methodology and stepwise regression. *Journal of Geophysical Research: Space Physics*, **128**, e2022JA030963. https://doi.org/10.1029/2022JA030963

Simms, L. E., Ganushkina, N. Y., & Dubyagin, S. (2024). Comparing influences of solar wind, ULF waves, and substorms on 20 eV–2 MeV electron flux (RBSP) using ARMAX models. *Journal of Geophysical Research: Space Physics,* **129**, e2024JA032458. https://doi.org/10.1029/2024JA032458

Billings, S. A. (2013), *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*, Wiley, Hoboken, N. J. DOI:10.1002/9781118535561

Leontaritis, I. J., & Billings, S. A. (1985). Input-Output Parametric Models for Nonlinear Systems, Part I: Deterministic Nonlinear Systems; Part II: Stochastic Nonlinear Systems. *Int. J. Control*, **41**(1), 303–344. https://doi.org/10.1080/0020718508961129

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, *Geophys. Res. Lett*., **38**, L18105, doi:10.1029/2011GL048980.

Wei, H.-L., Billings, S. A., Surjalal Sharma, A., Wing, S., Boynton, R. J., & Walker, S. N. (2011), Forecasting relativistic electron flux using dynamic multiple regression models, *Ann. Geophys*., **29**, 415–420, doi:10.5194/angeo-29-415-2011.

Boynton, R. J., M. A. Balikhin, S. A. Billings, H. L. Wei, & N. Ganushkina (2011), Using the NARMAX OLS-ERR algorithm to obtain the most influential coupling functions that affect the evolution of the magnetosphere, *J. Geophys. Res*., **116**(A5), A05218, doi:10.1029/2010JA015505.

Boynton, R. J., Balikhin, M. A., Billings, S. A., Reeves, G. D., Ganushkina, N., Gedalin, M., Amariutei, O. A., Borovsky, J. E., & Walker, S. N. (2013), The analysis of electron fluxes at geosynchronous orbit employing a narmax approach, *J. Geophys. Res.- Space Physics*, **118**, 1500–1513. https://doi.org/10.1002/jgra.50192

Boynton, R. J., Balikhin, M. A., & Billings, S. A. (2015), Online NARMAX model for electron fluxes at GEO, *Ann. Geophys.,* **33**, 405–411, www.ann-geophys.net/33/405/2015/doi:10.5194/angeo-33-405-2015.

Billings, S. A., Chen, S., & Korenberg, M. J. (1989), Identification of MIMO Non-Linear Systems Using a Forward Regression Orthogonal Estimator," *Int. J. Control*, **49**, pp. 2157–2189. https://doi.org/10.1080/00207178908559767

Li, P., H.-L. Wei, S. A. Billings, M. A. Balikhin, & R. Boynton (2013), Nonlinear model identification from multiple data sets using an orthogonal forward search algorithm, *J. Comput. Nonlinear Dyn.,* **8**(4), 041001, doi:10.1115/1.4023864.

Wang, L., & Cluett, W. R., (1996), Use of PRESS Residuals in Dynamic System Identification, *Automatica*, **32**(5), pp. 781–784. https://doi.org/10.1016/0005-1098(96)00003-9

Hong, X., Sharkey, P. M., & Warwick, K., (2003), Automatic Nonlinear Predictive Model-Construction Algorithm Using Forward Regression and the PRESS Statistic, *IEEE Proc. Control Theory Appl*., **150**(3), pp. 245–254. https://doi.org/10.1049/ip-cta:20030311

Efron, B., and Tibshirani, R. J., (1993), *An Introduction to the Bootstrap*, Chapman and Hall, London. https://doi.org/10.1201/9780429246593

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