AGU PUBLICATIONS

Space Weather

RESEARCH ARTICLE
10.1002/2017SW001669

Measures of Model Performance Based On the Log Accuracy Ratio

S. K. Morley[1], T. V. Brito[1,2], and D. T. Welling[3]

[1]Space Science and Applications, Los Alamos National Laboratory, Los Alamos, NM, USA, [2]Now at Department of Physics, University of Helsinki, Helsinki, Finland, [3]Climate and Space Sciences and Engineering Department, University of Michigan, Ann Arbor, MI, USA

Key Points:
- The median symmetric accuracy and symmetric signed percentage bias are introduced to address some drawbacks of current metrics
- The spread of a multiplicative linear model can be robustly estimated using the log accuracy ratio
- The properties of the median symmetric accuracy and the symmetric signed percentage bias are demonstrated on radiation belt examples

Correspondence to:
S. K. Morley,
smorley@lanl.gov

Citation:
Morley, S. K., Brito, T. V., & Welling, D. T. (2018). Measures of model performance based on the log accuracy ratio. *Space Weather*, *16*, 69–88. https://doi.org/10.1002/2017SW001669

Received 23 MAY 2017
Accepted 22 DEC 2017
Accepted article online 3 JAN 2018
Published online 23 JAN 2018

Abstract Quantitative assessment of modeling and forecasting of continuous quantities uses a variety of approaches. We review existing literature describing metrics for forecast accuracy and bias, concentrating on those based on relative errors and percentage errors. Of these accuracy metrics, the mean absolute percentage error (MAPE) is one of the most common across many fields and has been widely applied in recent space science literature and we highlight the benefits and drawbacks of MAPE and proposed alternatives. We then introduce the log accuracy ratio and derive from it two metrics: the median symmetric accuracy and the symmetric signed percentage bias. Robust methods for estimating the spread of a multiplicative linear model using the log accuracy ratio are also presented. The developed metrics are shown to be easy to interpret, robust, and to mitigate the key drawbacks of their more widely used counterparts based on relative errors and percentage errors. Their use is illustrated with radiation belt electron flux modeling examples.

1. Introduction

The utility, or value, of any forecast model is determined by how well the forecast predicts the quantities being modeled. There exists, however, a wide range of metrics to assess forecast quality and a similarly wide range of views on just what a "good" forecast is (see, e.g., Jolliffe & Stephenson, 2011; Murphy, 1993; Thornes & Stephenson, 2001). One key measure of the quality of a forecast is in how much it deviates from the observation. Although a forecast is strictly a prediction of events that have not yet occurred, this work treats simulation results as a forecast, regardless of the time interval. For application to validation of a reanalysis model ("hindcasting") the model output corresponds to the forecast and the validation data correspond to the observation (see, e.g., Jolliffe & Stephenson, 2011).

Model validation in regimes where the data vary over a limited range typically uses metrics that have the same scale and units as the quantities being modeled. For example, Lundstedt et al. (2002) presented a forecast model for the *Dst* index and evaluated the performance of their model using distributions of the forecast error as well as examining the root-mean-square error (RMSE). Another example applying this type of metric in model validation is that of Glocer et al. (2009), who evaluated the impact of including the Polar Wind Outflow Model in the Space Weather Modeling Framework by examining the RMSE of the magnetic field strength and elevation angle at geosynchronous orbit. One clear benefit of metrics that have the same units as the data is that they are easy to interpret.

For data from different data sets or time periods, or that cover multiple scales, accuracy measures that are independent of the scale of the data (such as percentage errors) are often used. An example of such data is radiation belt electron fluxes. Although the variability in electron fluxes at a given location and energy can be large (e.g., Friedel et al., 2002; Selesnick & Blake, 1997), scale-dependent measures could still be appropriate. However, there can be several orders of magnitude difference between electron fluxes at $L \simeq 4$ and geosynchronous orbit, with each location displaying different levels of variability (e.g., Li et al., 2005; Morley et al., 2017; Reeves et al., 2011). Thus, comparing scale-dependent accuracy measures can be problematic. Similarly, the measurements across a single orbit of a satellite in a highly elliptical orbit cover regions that could be argued to be of different scale and dynamics (e.g., Reeves et al., 2013). Throughout this manuscript we use examples from, or based on, radiation belt electron flux, but the presented work is applicable to any type of data where accuracy and bias measures that are independent of the scale of the data are desirable.

One approach to giving more equal weight to errors across several orders of magnitude is to use metrics that are based on relative errors (Subbotin & Shprits, 2009; Zhelavskaya et al., 2016) or are otherwise scaled to normalize the errors (Athanasiu et al., 2003; Welling, 2010). Alternatively, the data themselves can be transformed through the application of a power function, such as taking logarithms or applying a Box-Cox transform (Wilks, 2006). By transforming the data this way (Francq & Menvielle, 1996; Osthus et al., 2014), the use of scale-dependent accuracy measures may be better justified, as well as application of methods that assume homoscedasticity (i.e., the variance does not depend on the independent variable) (Sheskin, 2007). It is important to note that transforming the data alters the scale and may invalidate the assumptions behind other analyses.

Estimates of accuracy and bias aim to describe aspects of forecast quality, and no single metric of accuracy (or bias) is meaningful across all situations. How the metric penalizes different magnitudes and directions of forecast error should be considered. Should errors of equal magnitude be penalized equally? Should an underestimate by a factor of 2 have the same penalty as an overestimate by a factor of 2? How does the penalty implied by the metric scale with the size of the error? Finally, is the metric sensitive to assumptions about how the forecast error is distributed?

This paper assumes a number of desirable properties for metrics of model performance: (1) The metrics must be meaningful for data that cover orders of magnitude, (2) underprediction and overprediction by the same factor should be penalized equally, (3) the metrics should be easy to interpret, and (4) the metrics should be robust to the presence of outliers and bad data. This list of desirable properties is not universal but is likely to be relevant to a number of space weather applications.

We will begin with a brief review of model performance metrics, before giving a more in-depth discussion of the mean absolute percentage error and some variants of that metric. We then introduce metrics based on the log of the accuracy ratio that satisfy the list of desirable properties: the median symmetric accuracy and the symmetric signed percentage bias. Through the use of simple examples, as well as a multiplicative linear model, we then illustrate the behavior and drawbacks of metrics based on the percentage error, as well as the new metrics described in this paper. We also demonstrate the use of the log accuracy ratio in robustly estimating the spread of the error distribution in a multiplicative noise model. Finally, we show two illustrative examples of electron radiation belt prediction in which we discuss the application of both new and commonly used metrics. The examples presented aim at characterizing the accuracy and bias for an end user or for tracking of overall model performance with time. Using accuracy and bias metrics for understanding how well a particular model captures particular physical processes, for example, requires a different approach, and we briefly discuss how model performance metrics might be used differently for this purpose.

2. Measures of Forecast Quality

Scalar accuracy measures describe the average correspondence between individual pairs of forecasts and observations (Murphy, 1993). Various metrics can be used for this (e.g., mean square error) (see, e.g., Déqué, 2011; Walther & Moore, 2005; Wilks, 2006), and a selection will be described later in this section and summarized in Table 1. Our discussion begins with the forecast error, ε

$$\varepsilon = y - x \tag{1}$$

where x denotes the observation and y denotes the predicted value. Thus, the forecast error is negative when the forecast under predicts and is positive for an overprediction. Usually, we have multiple (n) pairs of forecast and observation ((x_i, y_i), where $i = 1, \ldots, n$) so it is helpful to aggregate these errors and present summary statistics (the summary statistics can be aggregated over subsets of the data, as well as the full set.)

The forecast bias describes the difference between the average forecast and the average observation (Murphy, 1993). A standard measure of bias is the mean error (ME; cf. Table 1), defined as the arithmetic mean of the set of forecast errors. Forecasts that, on average, overestimate or underestimate the observed value display bias. A negative number indicates a systematic underprediction, whereas a positive bias would indicate a systematic overprediction.

It is assumed throughout this paper that the quantity of interest is scalar. A number of approaches could be used to measure accuracy and bias for vector quantities such as the geomagnetic field (see also Tsyganenko, 2013; Wilks, 2006), but a simple and intuitive approach would be to calculate model performance metrics

Table 1
A Summary of Key Metrics

Metric	Definition	Symmetry	Scale/Order dependent	Comments
			Error metrics	
ε	$y - x$	Y	Scale	Forecast error
Q	y/x	N	Order	Ratio; complement of forecast relative error
			Accuracy metrics	
MSE	$\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right)$	Y	Scale	Different units/scale; quadratic penalty
RMSE	$\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right)}$	Y	Scale	Same units as x, y; quadratic penalty
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert\varepsilon_i\rvert$	Y	Scale	Same units as x, y; linear penalty
MdAE	$M\lvert\varepsilon_i\rvert$	Y	Scale	Same units as x, y; linear penalty; robust and resistant
MAPE	$\frac{100}{n}\sum_{i=1}^{n}\left\lvert\frac{\varepsilon_i}{x_i}\right\rvert$	N	Order	Percentage; penalizes overprediction more heavily
sMAPE	$100\,\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - x_i}{(x_i + y_i)/2}\right\rvert$	Y	Order	Percentage; unintuitive normalization; handles $x = 0$
ζ	$100\left(e^{(M(\lvert\log_e(Q_i)\rvert))} - 1\right)$	Y	Order	Percentage; robust and resistant
			Bias metrics	
ME	$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$	Y	Scale	Same units as x, y
MPE	$\frac{100}{n}\sum_{i=1}^{n}\frac{\varepsilon_i}{x_i}$	N	Order	Percentage; penalizes overprediction more heavily
MdLQ	$M\log_e(Q_i)$	Y	Order	Different scale
SSPB	$100\ \mathrm{sgn}(\mathrm{MdLQ})\ (e^{(\lvert\mathrm{MdLQ}\rvert)} - 1)$	Y	Order	Percentage; robust and resistant

Note. The columns give, in order, the abbreviation or symbol of the metric (as used in the text), the definition, whether the penalty is symmetric, whether the metric is scale or order dependent, and selected attributes.

like those presented in this paper on the magnitudes of the quantity only. Additional metrics to quantify the angular difference would then be required (e.g., Brito & Morley, 2017).

Forecast skill quantifies the accuracy of a set of model predictions relative to a reference prediction (Jolliffe & Stephenson, 2011; Wilks, 2006). One common reference is the accuracy of using the sample's climatological mean. For the specific case of using the mean square error (see section 2.1) as our accuracy metric and the sample mean as our reference, the skill score is typically called the prediction efficiency (e.g., Osthus et al., 2014). While the skill score quantifies improvement over a reference model (in the chosen accuracy metric) and requires an accuracy metric be calculated, it does not convey information about the accuracy of any specific set of model predictions. In this paper we focus on quantifying accuracy and bias for a single set of model predictions and do not discuss model skill.

2.1. Metrics Based On Scale-Dependent Errors

Like the bias, accuracy measures typically begin with the forecast errors, ε_i but then transform the data so that the direction of difference is removed. This is typically done by either squaring the forecast error or taking the absolute value of the forecast error. The mean square error (MSE; cf. Table 1) takes the former approach, and it can be seen that the mean square error is analogous to the variance penalizing large errors more heavily than small errors. Squaring the errors leads to the units and scale being different from the forecast quantity, which makes the MSE difficult to interpret. Transforming MSE back to the original scale by taking the square root then gives the root-mean-square error (RMSE).

As we are concerned with estimating the accuracy of a forecast the decision of which error metric should be used depends on the relative cost of different errors. For example, if the error doubles, is this twice as bad? or is it more than twice as bad? Is an overestimate worse than an underestimate of the same magnitude? If we wish to reduce the penalty on large errors, we can use the mean absolute error (MAE). This is defined as the arithmetic mean of $|\varepsilon_i|$, as shown in Table 1. This metric is more resistant to outliers as it uses $|\varepsilon|$ rather

than ε^2. It may, therefore, be more appropriate in cases where the errors are not normally distributed, where outliers are present, or where large forecast errors are not required to be weighted more heavily.

Both the RMSE and MAE estimate the typical magnitude of error using the mean. As the mean is not a robust measure of central tendency, we can improve the robustness of our accuracy metric by using a common robust measure of location: the median. Aggregating over all i using the median function (M) gives us the median absolute error (MdAE; cf. Table 1).

A good summary of scale-dependent measures of accuracy and bias can be found in Walther and Moore (2005). As seen here, scale-dependent metrics imply that deviations of the same magnitude have equal importance at different magnitudes of the base quantity. For example, an error of $\varepsilon = 100$ is penalized equally at $x = 10^3$ and $x = 10^6$.

2.2. Metrics Based On Order-Dependent Errors

When measuring the accuracy of a prediction in an order-dependent manner, the magnitude of relative error is often used; it is defined as the absolute value of the ratio of the error to the actual observed value. When multiplied by 100, this gives the absolute percentage error (APE). This measure is generally only used when the quantity of interest is strictly positive, and we make this assumption throughout.

We first define the relative error, η,

$$\eta = \frac{y - x}{x} = \frac{\varepsilon}{x} \tag{2}$$

Following the discussion given in section 2.1, we then remove the direction of difference by taking $|\eta|$, the absolute relative error. Defining relative error with equation (2), we find the magnitude of relative error and convert to a percentage to obtain the absolute percentage error. We then aggregate over multiple prediction-observation pairs using the mean, giving us the mean absolute percentage error (MAPE):

$$\text{MAPE} = 100\frac{1}{n}\sum_{i=1}^{n}|\eta_i| \tag{3}$$

To assess the bias using a percentage error, we simply aggregate the relative errors using the mean and then convert to a percentage, giving us the mean percentage error (MPE; cf. Table 1). Other metrics based on the relative error or similar order-dependent errors are given in Table 1.

As seen here, order-dependent metrics such as relative and percentage errors imply that deviations of the same order have equal importance at different magnitudes of the base quantity. For example, an error of $\varepsilon = 100$ where $x = 10^3$ has an equal penalty to an error $\varepsilon = 1$ where $x = 10$; both give a relative error of 0.1 and thus a percentage error of 10%. Order-dependent metrics are meaningful for data that cover orders of magnitude, and percentage errors are easy to interpret, so measures such as MAPE satisfy both the first and third desirable qualities for measures of model performance.

3. Mean Absolute Percentage Error and Variants

MAPE is used in many different fields of research, from population research (e.g., Swanson et al., 2000) to business forecasting (e.g., Kohzadi et al., 1996), atmospheric science (e.g., Grillakis et al., 2013; Zheng & Rosenfeld, 2015), and space science (e.g., Reikard, 2011; Zhelavskaya et al., 2016). MAPE has also been used in validation of radiation belt models (Kim et al., 2012; Li et al., 2014; Tu et al., 2013), and these are discussed further in section 3.2. However, though meaningful in a wide range of situations and easy to interpret, MAPE is not without problems that may be important in any given application.

3.1. Some Problems With MAPE

The following problems have been noted by various authors:

1. MAPE becomes undefined when the true value is zero (Hyndman & Koehler, 2006).
2. MAPE is asymmetric with respect to overforecasting and underforecasting (Hyndman & Koehler, 2006; Makridakis, 1993; Tofallis, 2015).
3. APE is constrained to be positive, so its distribution is generally positively skewed (Hyndman & Koehler, 2006; Swanson et al., 2000).
4. MAPE is not resistant to outliers (Swanson et al., 2000; Tofallis, 2015).

Due to the first point, unless a physically reasonable approach can be determined to work with cases where $x = 0$, MAPE is not an appropriate metric where the quantity being predicted is likely to be zero (e.g., Tofallis, 2015). We also note that unless the data used are positive-valued ratio-level data (having a meaningful, nonarbitrary zero point) (Sheskin, 2007; Stevens, 1946), the APE has limited meaning (Hyndman & Anathasopoulos, 2014). For example, radiation belt fluxes are constrained to lie in the interval $[0, \infty)$ and the units of flux have a true zero; therefore, APE can be used for radiation belt flux predictions and model validation. Neither the *Kp* geomagnetic index (Menvielle & Berthelier, 1991) nor the Celsius temperature scale is ratio-level data (Stevens, 1946) (these are ordinal and interval data, respectively), and thus, metrics based on relative errors should not be used. Further discussion of zeros and measurement backgrounds is given in section 6.

To elaborate on the second point, a prediction of 1,000 where the observed value is 500 gives a different magnitude of error (100%) than a prediction of 500 where the observed value is 1,000 (50%). Underprediction is therefore less heavily penalized than overprediction, even if the order of the error is the same. Similarly, given $x = 10^5$ and two models $y_1 = 5 \times 10^4$ (a factor of 2 under prediction) and $y_2 = 1.75 \times 10^5$ (a factor of 1.75 overprediction), the APE for model 1 is 50% and the APE for model 2 is 75%; based on the APE or for aggregated measurements the MAPE, model 1 is deemed to be more accurate yet in many applications we would not wish to penalize the overprediction more heavily. MAPE, therefore, does not satisfy the second desirable property for a metric of model performance given earlier in this paper. Variants of MAPE have been proposed that mitigate this asymmetry (e.g., Flores, 1986; Makridakis, 1993) by normalizing the forecast error by the mean of x and y, for example,

$$\text{sMAPE} = 100\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - x_i}{(x_i + y_i)/2}\right| \tag{4}$$

The unconventional normalization in the relative error makes the resulting percentage error unintuitive in its interpretation, though this does address the cases where one of y or x is zero as well as mitigating the asymmetry of MAPE.

Regarding the third point, given that APE has a lower bound of zero but has no upper bound they are likely to be skewed positive. Take a case where the forecast errors are distributed approximately normally and are symmetric about the true value. By taking the absolute values, the distribution of APE is now highly skewed. By subsequently using the arithmetic mean, which is a poor measure of central tendency in skewed distributions, MAPE is prone to overstating the error (Swanson et al., 2011).

Finally, MAPE is easily affected by outliers as the mean has a breakdown point of zero (Hampel, 1974). Given a set of predictions with APE of [5,3,10,2,5,120] %, MAPE takes the value 24.16%; reducing the error on the final prediction from 120% to 30% reduces the MAPE to 9.17%. Therefore, any large errors due to, for example, bad data or late prediction of a large change will be heavily penalized by taking the arithmetic mean. This means that MAPE also fails to satisfy the fourth desirable property (robustness) given above. Swanson et al. (2000) describe a method for reducing the impact of outliers in which the distribution of APE is symmetrized, using a modification of the Box-Cox transform (Wilks, 2006). Specifically, they use (Swanson et al., 2000)

$$y(\lambda) = (x^{\lambda} - \lambda)/\lambda \text{ when } \lambda \neq 0 \tag{5}$$

$$y(\lambda) = \log_e(x) \text{ when } \lambda = 0 \tag{6}$$

and the optimal value of λ is found using maximum likelihood estimation. After finding the optimal value of λ, the absolute percentage errors are transformed using equations (5) and (6) and the mean of the transformed APEs (called MAPE-Transformed or MAPE-T) is used in place of MAPE. Though mitigating the impact of skewed distributions and outliers in estimating the mean, the value of MAPE-T is difficult to interpret as it no longer represents a percentage error. This was addressed by Coleman and Swanson (2007) (see also Swanson et al., 2011) in their presentation of MAPE-R (MAPE-rescaled), where MAPE-T is reexpressed in the original scale of the data by applying the inverse of the modified Box-Cox transform to MAPE-T.

3.2. Selected Applications of MAPE and Variants

As mentioned above, MAPE is used widely for model validation in many fields, including the space sciences. To predict the effective dose of galactic cosmic radiation received on transpolar aviation routes, Hwang et al. (2015) developed a model that forecasts the heliocentric potential (HCP) from a lagged time series of monthly sunspot number. The HCP is a required input for the Federal Aviation Administration's CARI-6 M software for dose estimation. Zhelavskaya et al. (2016) have developed a neural network to predict the frequency

of the upper-hybrid resonance to derive electron number densities in the inner magnetosphere, using Van Allen Probes electric field data. These authors used MAPE to assess the accuracy of their predictions, both in predicted frequency and predicted number density. We note that the electron number density, like radiation belt electron flux, is constrained to be positive and has a physically meaningful zero. Further, the electron number density can vary by orders of magnitude over a single orbit as well as at a fixed location due to dynamical processes. Hwang et al. (2015) and Zhelavskaya et al. (2016) calculated MAPE directly, without first transforming the data, and their reported percentage errors are therefore directly interpretable, though should still be interpreted keeping the drawbacks described in section 2.2 in mind. The effect of the asymmetry of MAPE is explored further in section 5.1.

Kim et al. (2012) used MAPE as the accuracy metric for comparing their model results with observations from the CRRES satellite. However, they defined MAPE using log-transformed data. This approach was subsequently used by Tu et al. (2013) and Li et al. (2014). In addition to the main drawbacks of MAPE described above, applying equation (3) to log-transformed data can be demonstrated to be incorrect (see, e.g., Morley, 2016). Effectively, replacing x_i and y_i in equation (3) with $\log_{10}(x_i)$ and $\log_{10}(y_i)$ means that the quantity being calculated is the arithmetic mean of $|\log_{x_i}(y_i/x_i)|$. This change of base renders the arithmetic mean meaningless, and if x_i is large, then the result will incorrectly be a very small error. It is worth noting that, for small errors, $\log_e(y_i/x_i)$ is an approximation of the relative error. Thus, when (y_i/x_i) is of order unity, $100\log_e(Q)$ gives an approximate percentage error. If all errors are small (i.e., all $(y_i/x_i) \sim 1$), then aggregating $|\log_e(y_i/x_i)|$ using a mean is a good estimate of MAPE.

Other measures similar to MAPE have been proposed and applied in radiation belt modeling. For example, Subbotin and Shprits (2009) used a set of metrics based on what they called the normalized difference. The normalized difference was calculated for 2-D grids of simulation results. The equation can be given as (see Table 4 of Subbotin & Shprits, 2009)

$$\mathrm{ND}_i(f) = 100\frac{y_i(f) - x_i(f)}{\max(y_i(f) + x_i(f))/2} \tag{7}$$

where f denotes the additional dimension and i indicates the primary index variable for consistency with the rest of this manuscript. The results were then aggregated using the mean of $|\mathrm{ND}_i|$ to give the "average difference" and using the maximum of $|\mathrm{ND}_i|$ to give the "maximum difference." These are seen to be similar in construction to sMAPE (Makridakis, 1993) but using the maximum value of the means of each (forecast and observation) pair instead of simply using the mean of y and x. The average difference is identical to sMAPE when y_i and x_i are uniform in f. When varying in f, the interpretation becomes more difficult as the forecast error is not normalized to either the forecast, the observation, or even the mean of (x, y). The normalized difference and average difference have subsequently been used by Drozdov et al. (2017) to examine differences between different configurations of the Versatile Electron Radiation Belt model (Subbotin & Shprits, 2009). While Subbotin and Shprits (2009) provide descriptions of how to interpret these metrics and provide use cases for them on higher dimensionality data, they cannot easily be interpreted as measures of accuracy (as defined in section 2).

4. Introducing Robust, Symmetric Measures Based On the Log Accuracy Ratio

We now aim to describe two measures of model performance that satisfy the four desirable properties enumerated previously. We begin by defining the accuracy ratio, Q, as y/x, that is, the ratio of the predicted value to the observed value. The name "accuracy ratio" was coined by Tofallis (2015), who note that Q is the complement of the relative error ($\eta = Q - 1$) and so will have the same distribution as the relative error but shifted by one unit. Tofallis (2015) also showed that $\log_e(Q)$ is a superior accuracy measure to MAPE for data where the variance depends on the magnitude of the variable (as is often the case with space physics data, such as radiation belt electron fluxes; e.g., Morley et al., 2016; Reeves et al., 2011). The interested reader is also referred to Kitchenham et al. (2001) for a discussion of the accuracy ratio in measuring model performance. It is instructive to note that the log of the accuracy ratio is identical to the forecast error for log-transformed x and y.

We note that previous work on radiation belt electron data has used ratios of the observed to predicted values. Chen et al. (2007) defined the "PSD matching ratio," R, (see also Yu et al., 2014) as the ratio of phase

space densities, where the denominator is always the smaller of the two values. Here we generalize this to our prediction-observation pair (x, y)

$$\begin{aligned} x' &= x \text{ if } x < y \text{ else } y \\ y' &= y \text{ if } x < y \text{ else } x \\ R &= \frac{y'}{x'} \end{aligned} \tag{8}$$

The matching ratio R can be alternatively expressed using the accuracy ratio. Specifically, we use the fact that $\log(x/y) = -\log(y/x)$, and thus, $|\log(x/y)| = |\log(y/x)| = \log(y'/x')$. To transform this back to the original units and scale, we exponentiate

$$\begin{aligned} \log_e(R) &= \log_e(y'/x') \\ &= |\log_e(y/x)| \\ &= |\log_e(Q)| \\ R &= \exp(|\log_e(Q)|) \end{aligned} \tag{9}$$

Morley et al. (2016) used the accuracy ratio to compare electron fluxes computed from the Global Positioning System constellation with "gold standard" measurements from the Van Allen Probes mission. When presenting graphical summaries of these data, Morley et al. (2016) showed $\log_{10}(Q)$ "so that the ratios are symmetric both above and below 1." Taking the logarithm ensures that a factor of 3 difference between x and y is the same magnitude of error, regardless of the direction of error. However, even though log transforming the data will tend to symmetrize positively skewed distributions, the actual distributions of $\log_{10}(Q)$ may not be symmetric. For this reason, Morley et al. (2016) used the median of $\log_{10}(Q)$ as a measure of central tendency. This quantity also represents a robust measure of bias, though it suffers from a lack of intuitive interpretability. The effect of the transformation does not depend on the base of logarithm used here, although the interpretation of the exact value does depend on the base used.

4.1. Accuracy: Median Symmetric Accuracy

We propose a measure of accuracy derived from logarithms of the accuracy ratio. The specific aim is to mitigate many of the problems inherent in using MAPE (see section 3.1) and that maintain the interpretability of MAPE and satisfy all the desirable properties given at the end of section 1. Specifically, we follow the lead of Tofallis (2015) and Morley et al. (2016) in using $\log(Q)$ and modify our accuracy metric such that it is interpretable as a percentage error. We use the natural log in this presentation, but note that any base can be used as long as the antilog is found correctly. This metric was first suggested by Morley (2016), but we here expand on the derivation and meaning of this accuracy metric before testing the behavior of this metric.

We begin by taking the absolute values of $\log_e(Q)$. This transformation ensures that the metric is symmetric in the sense that switching the values of the predicted and observed value gives the same error (unlike MAPE). We then aggregate over all prediction-observation pairs using the median function and then exponentiate to return to the original units and scale.

$$\exp\left(M\left(\left|\log_e(Q_i)\right|\right)\right) \tag{10}$$

As the median function is an order statistic, this is equivalent to the median matching ratio. The resulting value has a lower bound of 1, so we subtract one such that our metric lies in the range $[0, \infty)$. This subtraction allows the interpretation as an unsigned (symmetric) fractional error, and multiplying this by 100 yields an equivalent percentage error.

$$\zeta = 100\left(\exp\left(M\left(\left|\log_e(Q_i)\right|\right)\right) - 1\right) \tag{11}$$

This metric, ζ, is therefore named the median symmetric accuracy (cf. Morley, 2016). We can see that for two prediction-observation pairs, $(1.7 \times 10^5, 10^5)$ and $(1.7 \times 10^2, 10^2)$, ζ is 70% in both cases; this is the same as the correct application of MAPE. Using log-transformed data gives absolute percentage errors of [4.6, 11.5]% and an incorrect estimate of MAPE as 8.1%. The results from ζ are also symmetric with respect to the reversal of the predictions and observations, in contrast with MAPE.

As noted previously, we specifically aim for a metric that is intuitive and can be interpreted as a percentage error. We now show that the median symmetric accuracy (ζ) is equivalent to the median percentage error, when the relative error is defined to always have the same direction.

Taking our predicted and observed values to be y and x, as defined previously, we can define y' to be the larger value and x' to be the smaller value. We now define a new "unsigned" forecast error, $\varepsilon' = y' - x'$, and thus a new unsigned relative error

$$\eta' = \frac{y' - x'}{x'} \tag{12}$$

It can be seen that η' is equal to $R - 1$ where R is the matching ratio defined in equation (8). Using equation (9) along with the fact that quantiles are preserved under monotonic transformations, we see that

$$\begin{aligned} \zeta &= 100\left(\exp\left(M\left(\left|\log_e(Q_i)\right|\right)\right) - 1\right) \\ &= 100\left(M\left(R_i - 1\right)\right) \\ &= 100\left(M\left(\eta_i'\right)\right) \end{aligned} \tag{13}$$

Thus, the median symmetric accuracy is equivalent to the median unsigned percentage error. In practice, this relationship is exact only when n is odd or when n is large. In the case of even n the median in equation (11) will give the geometric mean of the two central unsigned percentages, where equation (13) will give the arithmetic mean. This effect will only impact very small, even-valued n, and since the geometric mean of a lognormal distribution is equal to the median, we recommend using ζ as defined in equation (11).

The median symmetric accuracy mitigates the problems with asymmetric penalty and effects of outliers (problems 2 and 4 described in section 3.1) yet maintains interpretability. By using a robust and resistant measure of central tendency, we minimize the effect of the skewness of the distribution of absolute errors (problem 3). ζ therefore satisfies all four desirable properties listed at the start of this paper and mitigates several key problems of MAPE as an accuracy metric. We note that ζ is undefined when the smaller value in the forecast-observation pair is zero and returns to this point in section 6. The interpretation of this metric is that 50% of the unsigned percentage errors are smaller than ζ. If we interpret the median as being an indicator of the "typical" value in a distribution, then we can further say that ζ represents the typical unsigned percentage error.

4.2. Bias: Symmetric Signed Percentage Bias

The bias (mean error; cf. Table 1) gives values smaller than 0 for a systematic underprediction and values greater than 0 for a systematic overprediction. An order-dependent alternative should be interpretable in the same way. The physical meaning of the accuracy ratio is clear, making the median accuracy ratio an easily interpretable quantity (Morley et al., 2016; Rodriguez et al., 2017). However, it is centered on 1 and is not symmetric. Assuming that symmetry is a desirable property for our bias metric, then we can use the median log accuracy ratio (e.g., Morley, 2016; Morley et al., 2016). Underprediction will give a negative value of $M(\log(Q))$, and overprediction will give a positive value; an unbiased forecast will yield $M(\log(Q)) = 0$. This symmetry about zero then mirrors the more common measures of bias, the mean error, and mean percentage error. Due to the log transform, the choice of base affects the result and will determine the level of interpretability for any given data set. We therefore present a new measure of bias based on the log accuracy ratio.

Ideally, our bias metric should have the same desirable properties given in section 1, including an interpretable scale. To achieve this, we first estimate the magnitude of the bias by taking the absolute value of $M(\log(Q))$ (we use natural logarithms here for ease of notation), taking the antilog, and subtracting 1 so that the lower limit is zero. We then find the direction of the bias using the signum function and multiply by 100 to express as a percentage.

$$\mathrm{SSPB} = 100\ \mathrm{sgn}(M(\log_e(Q_i)))(\exp(|M(\log_e(Q_i))|) - 1) \tag{14}$$

The symmetric signed percentage bias (SSPB) can therefore be interpreted similarly to a mean percentage error but is not affected by the likely asymmetry in the distribution of percentage error and robustly estimates the central tendency of the error. As SSPB is based on relative errors, penalizes underprediction and overprediction equally, is robust, and is interpretable as a percentage, it meets all of our stated desirable properties.

5. Applications

To illustrate the use of the metrics described above, we generate a series of data, z, that we use as our ground truth. Figure 1a shows 80 keV electron flux data from the MagEIS instrument (Blake et al., 2013) on the Van Allen Probes mission (Mauk et al., 2013) as a function of time on 19–20 January 2014. We define a series, z,

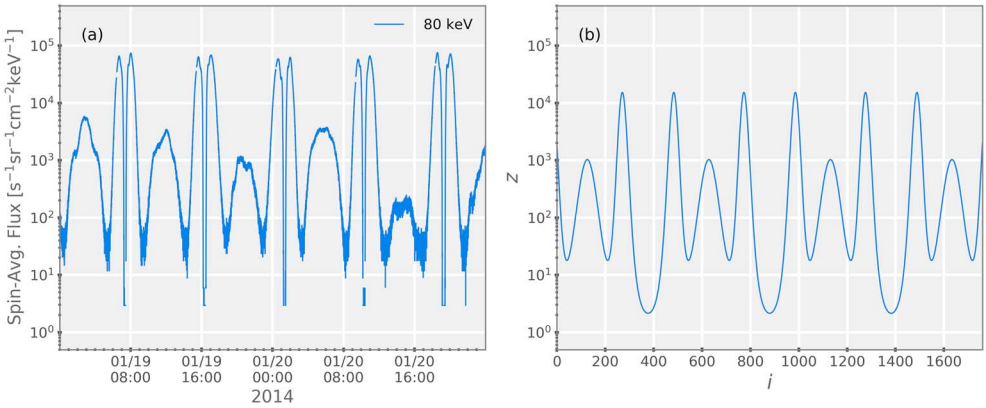

Figure 1. (a) Spin-averaged electron flux at 80 keV measured by the MagEIS instrument on the Van Allen Probes (RBSP A) satellite on 19–20 January 2014. (b) A time series constructed (equation (15)) to approximate the electron flux data for the purpose of illustrating the application of the metrics presented in this paper.

to approximate these data using a model that varies cyclically between very small and very large values, varying over approximately 5 orders of magnitude. This is shown in Figure 1b and is given by

$$\begin{aligned} g &= 10^{\sin(i)} + 10^{\cos(2i)-\sin(i)} \\ z &= 2^g \end{aligned} \tag{15}$$

We also define a noisy series derived from z that we can use as our imperfect "model." A multiplicative linear model is used here to compare several metrics. If we assume a counting process, such as measuring particle radiation, and ignore detection issues such as instrument dead time, then we can assume the process to be Poisson. As the mean of a Poisson process increases, so does the variance. That is, the error becomes larger as the expected value becomes larger. Note that as the mean of a Poisson distribution becomes large, the Poisson distribution can be well approximated by a Gaussian distribution.

An ordinary linear model has a number of assumptions, one of which is that the data are homoscedastic, that is, the variance of the data is assumed to be constant. Particle fluxes are well known to display unequal variance. Specifically, the variance increases as the flux increases. The log transformation is variance stabilizing, so to ensure that the variance of our error term scales with the estimated flux value, we assume a Gaussian error distribution in log(flux). Then our estimate of the flux ($\hat{z}$) can be modeled as the true flux (z) plus an error term (Γ). This model is thus illustrative of the particle flux use case.

$$\log_e(\hat{z}) = \log_e(z) + \Gamma \tag{16}$$

$$\hat{z} = z\exp(\Gamma) \tag{17}$$

$$\hat{z} = z\exp(\sigma\upsilon + \epsilon) \tag{18}$$

where Γ represents our error distribution, υ represents a random variate drawn from a standard normal distribution, and σ is the standard deviation of the error distribution. To model a systematic bias in the error, we include ϵ; if $\epsilon = 0$, then the Gaussian error is centered on $\log(z)$.

5.1. Symmetry and Robustness Properties

Taking our series z, we first apply simple noise models in which we apply a constant offset of a factor of 2; we use both $2z$ and $z/2$. We then derive a third noisy series where each point i is randomly chosen to be either $2z_i$ or $z_i/2$. We then calculate MAPE, sMAPE, and ζ. As expected, ζ gives the same answer (= 100%) in each of the three cases. By contrast, MAPE gives answers of 50% ($2z$), 100% ($z/2$), and 74.3% (random) and sMAPE gives answers of 66.$\dot{6}$% in each case. While ζ and sMAPE both penalize overprediction and underprediction equally, MAPE represents an equal order of error differently depending on the direction of the error. Of these metrics, only ζ consistently gives the intuitive answer that a factor of 2 difference is a 100% error.

We now turn to the performance of each metric on a more realistic case, $(x, y) = (z, \hat{z})$. Series $\hat{z}$ is described by equation (18) and is displayed as a time series in Figure 2a. These data are displayed versus z in Figure 2b. The inset panels show zoomed areas to illustrate the scale of the noise in $\hat{z}$.

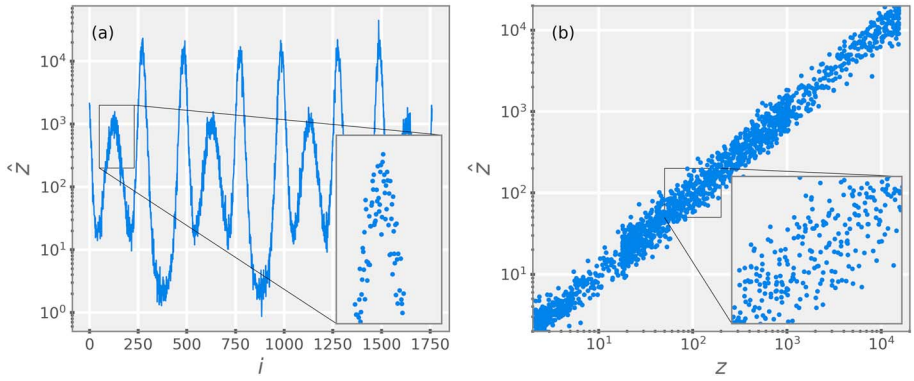

Figure 2. Series $\hat{z}$ plotted as (a) a function of index i and (b) a scatter plot against z. Each panel has an inset window expanding a small section of the displayed area to better show the scatter in the points.

Figure 3 shows probability distributions for different error estimates for the case of z as our observation and $\hat{z}$ as our prediction. The vertical dashed lines mark the median of each distribution, and the vertical solid lines mark the arithmetic mean. Figure 3a shows the distribution of the percentage error. It can be seen clearly that this distribution is asymmetric. Taking the absolute values gives the distribution of APE, shown in Figure 3b. The probability distribution of $\log_e(Q)$ is shown in Figure 3c and can be seen to be both centered near zero and symmetric. Taking the absolute values gives the distribution of the symmetric accuracy ($|\log_e(Q)|$), which is shown in Figure 3d. The median symmetric accuracy (ζ) is 22.71%, and the MAPE is 24.33%. Taking the median of $\log_e(Q)$ and applying equation (14) gives the symmetric signed percentage bias (SSPB) as −1.1%, while inspection of Figure 3a shows that the mean percentage error (MPE) is 5.04%.

We illustrate the "rescaled" MAPE of Swanson et al. (2011) in Figure 4. Figure 4a shows the distribution of APE: this panel is identical to Figure 3b. We then apply the modified Box-Cox transform of Swanson et al. (2000)

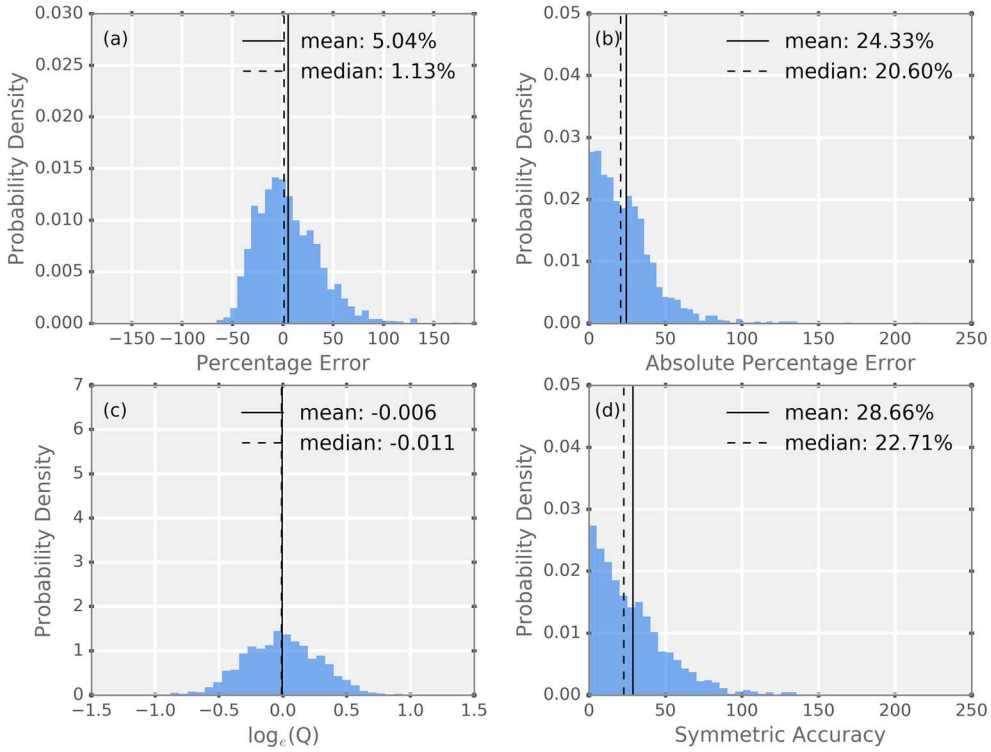

Figure 3. Probability distributions of (a) percentage error, (b) absolute percentage error, (c) $\log_e(Q)$, and (d) symmetric accuracy for $(x, y) = (z, \hat{z})$. Mean values for the presented distributions are marked with solid vertical lines, and median values are indicated by dashed vertical lines.

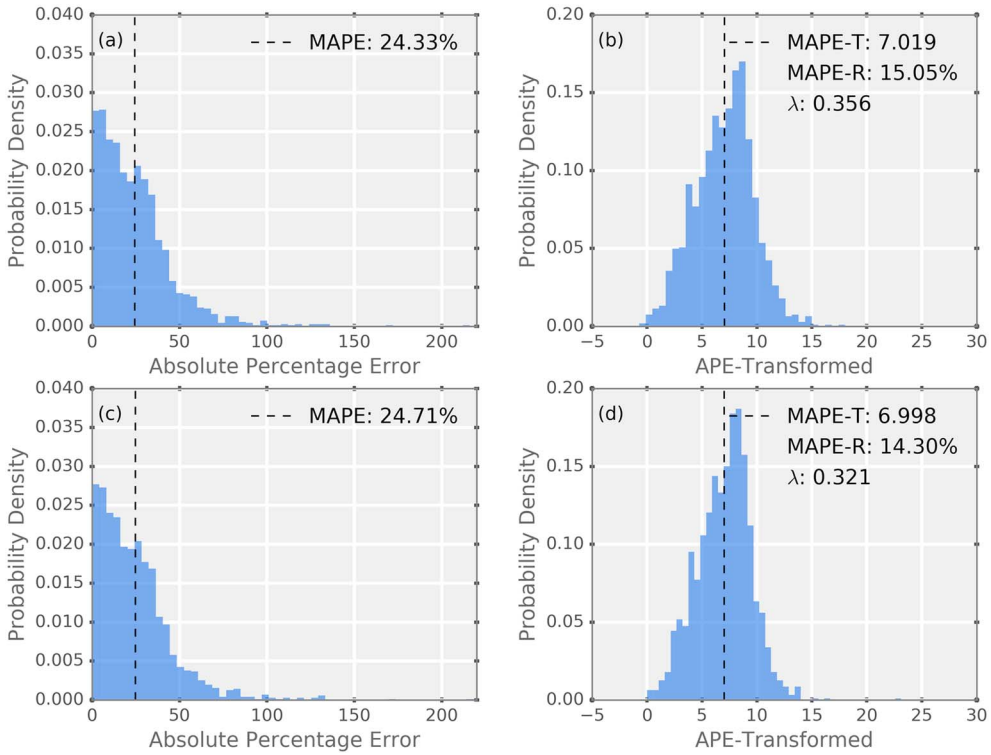

Figure 4. Probability distributions of (a and c) absolute percentage error (APE) and (b and d) Box-Cox transformed APE. The vertical dashed lines on Figures 4a and 4c and 4b and 4d represent MAPE and MAPE-T, respectively. Figures 4b and 4d are annotated with the values of MAPE-R and the value of λ from the Box-Cox transform. Figures 4a and 4b show results using series (z, $\hat{z}$), wherein Figures 4c and 4d show results where 10% of the points in the Gaussian noise model have been replaced by outliers from a Gaussian of standard deviation 8σ.

to these data to get APE-transformed. This distribution is shown in Figure 4b, and MAPE-T is calculated as the mean of this symmetrized distribution of APEs. Finally, we calculate MAPE-R by applying the inverse of the modified Box-Cox transform to MAPE-T (Coleman & Swanson, 2007; Swanson et al., 2011):

$$\mathrm{MAPE-R} = ((\lambda)\,(\mathrm{MAPE-T} + 1))^{\frac{1}{\lambda}} \tag{19}$$

For this example we see that MAPE-R is calculated as 15.03%. This value depends critically on λ, which will vary with the exact distribution of APE. The value is difficult to interpret as the rescaling effectively weights the different magnitudes of APE differently (see Swanson et al., 2011), and comparisons between models are not straightforward.

We now increase the weight of the tails in our noise model. To do this, we randomly select 10% of the indices, i, for series $\hat{z}$ and recalculate $\hat{z}_i$ with a value of σ that is 8 times larger. Figure 5 shows results for the present case where $\hat{z}$ has been contaminated by a much broader error distribution. Figure 5a shows the distribution of the percentage error. Comparing Figure 5a to Figure 3a shows that the distributions are visually very similar. The resulting distribution of APE is shown in Figure 5b. The probability distribution of $\log_e(Q)$ for the contaminated series is shown in Figure 5c, and the distribution of absolute values ($|\log_e(Q)|$) is shown in Figure 5d. In this case, ζ is almost unchanged at 22.79% and the MAPE is slightly different at 24.71%. The SSPB still estimates the bias as −1.1%, and the MPE has increased very slightly to 5.32%.

Having now added a contaminating distribution, we recalculate MAPE-T and MAPE-R, shown in Figures 4c and 4d. The inclusion of outliers increases the weight of the tail of the distribution, and hence, the modified Box-Cox transform has a different λ. This leads to a different rescaling of APE and in this case a MAPE-R (14.3%) that is lower than the case without outliers (15.05%). In this case the sensitivity of MAPE-R to the transform leads us to the incorrect conclusion that the error has decreased. This test clearly illustrates that values of MAPE-R for different samples are not necessarily comparable in a meaningful way and that interpreting MAPE-R is difficult, at best.

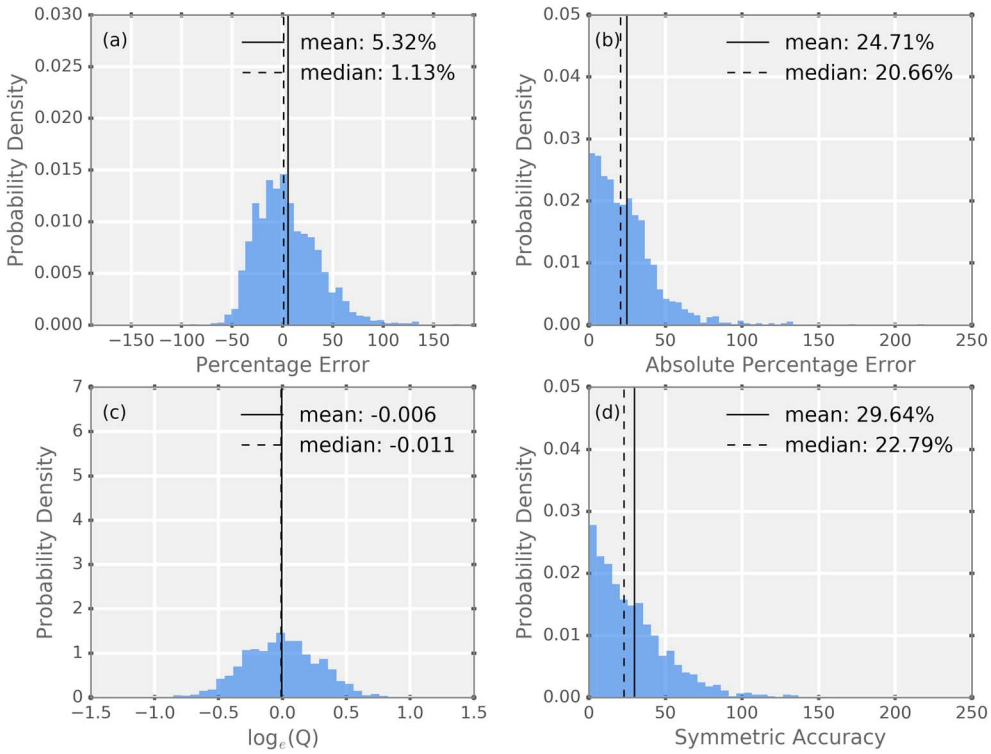

Figure 5. Same as Figure 3 where 10% of the points in the Gaussian noise model have been replaced by outliers from a Gaussian of standard deviation 8σ.

5.2. Estimating σ for a Multiplicative Linear Model

Previous authors have also used errors based on the forecast errors in log flux (e.g., Ginet et al., 2013; O'Brien & McPherron, 2003; Weiss et al., 1997). While this may simply seem like a convenient transformation to make metrics like the RMSE scale independent, it can be demonstrated to have a clear meaning. Specifically, in the case of an unbiased error distribution the RMSE is an estimator of the standard deviation of a Gaussian error distribution. The estimated standard deviation is defined as

$$\hat{\sigma} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} \tag{20}$$

and can be compared to the root-mean-square error (see Table 1). The RMSE of log flux therefore estimates the standard deviation for a multiplicative linear model in which the error is Gaussian in log space; we estimate σ for our multiplicative linear model using the RMSE where $\epsilon = \log_e(z) - \log_e(\hat{z})$. Due to the log transformation, ϵ is now simply $\log_e(Q)$.

We can also estimate σ robustly using $\log_e(Q)$. Calculating the median absolute error of $\epsilon = \log_e(z) - \log_e(\hat{z})$ is equivalent to calculating the median of $|\log_e(Q)|$. Above we estimate the standard deviation using the RMSE; similarly, we here estimate the median absolute deviation (MAD) using $M(|\log_e(Q)|)$. The median absolute deviation provides a consistent estimator of the standard deviation by

$$\hat{\sigma} = b\,\mathrm{MAD} \tag{21}$$

where b is a scale factor that is distribution dependent. To scale MAD for consistency with σ for a Gaussian distribution, we set $b = 1.4826$ (e.g., Rousseeuw & Croux, 1993).

An alternative measure for the spread of a distribution has been presented by Rousseeuw and Croux (1993). Their S_n estimator has been shown to be very robust, among other desirable properties.

$$S_n = c\,M_i(M_j(|x_i - x_j|)) \tag{22}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, n$. The outer median is defined to be the low median, given by the order statistic of rank $(n + 1)/2$, so that for an even number of data points the lower of the two central values

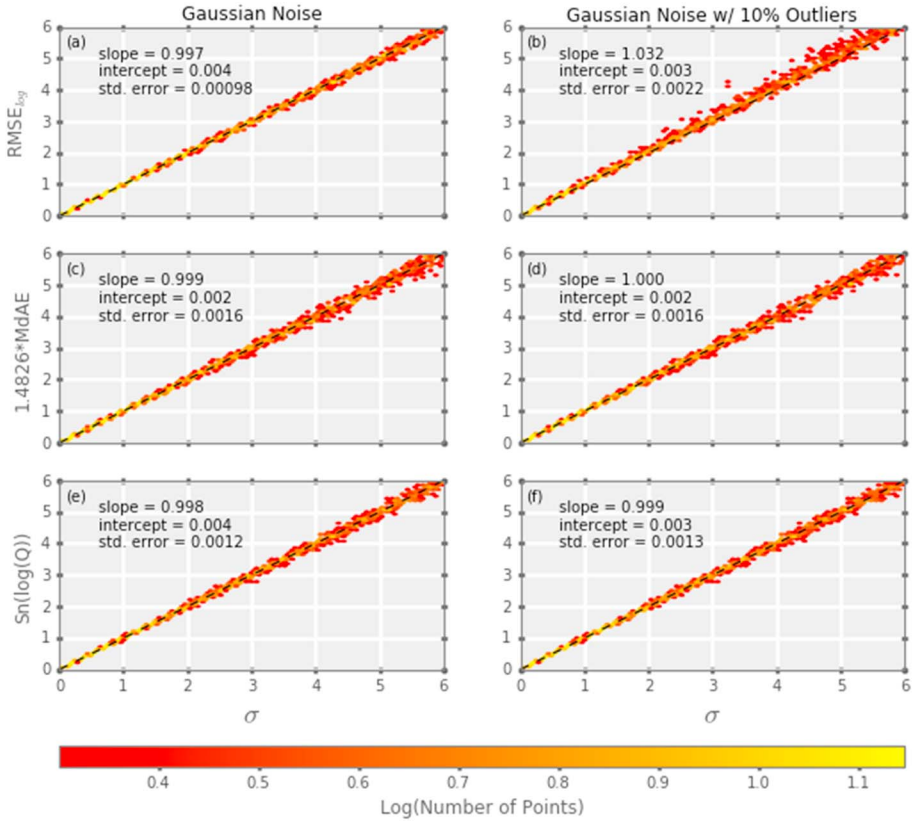

Figure 6. Two-dimensional histograms of σ versus estimated σ. Figures 6a, 6c, and 6e show the estimates using the $\hat{z}$ with a Gaussian error distribution. Figures 6b, 6d, and 6f show the estimates where 10% of the points have been replaced with $\hat{z}$ with errors from a much broader Gaussian. (a, b) Estimates of σ using the RMS of $\log_e(Q)$. (c, d) Estimates using the median of $|\log_e(Q)|$. (e, f) Estimates using the S_n estimator. Each panel has a black dashed line marking $y = x$.

is always taken. The inner median is defined to be the high median, given by the order statistic of rank $(n/2)+1$, so in the case of an even number of data points the higher of the two central values is always taken. In the case of an odd number of data points the high and low medians are identical and in all cases the high and low medians are actual data points, whereas a standard median of an even-length series is given as the arithmetic mean of the two central values and is not guaranteed to be an actual value in the data set. S_n provides an unbiased estimate of σ for a Gaussian distribution when $c = 1.1926$ (Rousseeuw & Croux, 1993). S_n is not referenced to a measure of location and is therefore suitable for use with asymmetric distributions. We will also estimate σ using S_n, where x is given by $\log_e(Q)$.

We generate a series $\hat{z}$ for $\sigma = (0, 6)$ in steps of 0.005. For each value of σ we estimate it using each of the above methods. Figure 6 shows 2-D histograms of σ against (a) the RMSE of $\log_e(Q)$, (b) the MdAE of $\log_e(Q)$, and (c) the S_n of $\log_e(Q)$. The color of each cell shows the density of points. The annotations give the slope, intercept, and standard error of a linear fit to the data. For reference, each panel has a black dashed line marking $y = x$. In the case of a single Gaussian error distribution all the metrics estimate σ consistently. The standard error of the estimate using the median absolute error is slightly larger than the other two methods, with RMSE having the lowest standard error (the linear fit uses ordinary least squares and hence will minimize this quantity). The S_n estimator provides the best estimate of σ. When we include additional noise, the performance of the RMSE is noticeably worsened, and $S_n(\log_e(Q))$ remains a good estimator of σ for the dominant noise model.

6. Zero-Valued Predictions or Observations

The metrics developed in this work that have not addressed the problem that measures based on relative error become undefined when zeros are present. In practice, we note that there is always a measurement threshold. In the radiation belts the measured electron flux at very high energies (several MeV) is typically near instrument background levels. If a count of zero is recorded in a detector, that does not mean zero flux.

There remains a nonzero probability of a finite flux. A model predicting anywhere between zero and a defined threshold level should not be penalized. We propose that when the observed value, or the predicted value, falls below the defined measurement threshold for the predictand, the value is fixed to the threshold. That is, a very low, but finite, model prediction (below the observable threshold) when the instrument count rate is zero does not get penalized. While other authors have used approaches like the sMAPE metric to address this, we aim to preserve the interpretability of the metrics while considering the physical meaning of a zero measurement or prediction. This approach will not be universally appropriate, and other approaches to measuring accuracy (such as thresholding and applying categorical metrics) should be considered.

In the illustrative example above, which represents the electron flux measured at a satellite traversing the radiation belts (e.g., Van Allen Probes) at a relatively low energy, zero-valued predictions or observations are likely to be rare and use of a lower threshold in calculating performance metrics is likely to be justified. A forecast or observation that is often zero raises the likelihood of overstating prediction quality by this method. For the case of measuring solar energetic protons of >10 MeV the observations are typically at or near background. In this case, because the transient enhancements are relatively rare a constant prediction of zero (or of the background) would give an excellent accuracy but fails to predict the event of interest. For assessing the accuracy and bias of models for rare events, different approaches should be considered. For example, the data could be converted to categorical forecasts and the accuracy and bias calculated from the contingency table (Wilks, 2006). Probabilistic approaches that account for the probability of observing a value above the observing threshold could also be used.

7. Sample Applications: Predicting Electron Flux and Fluence

We illustrate the use of ζ and SSPB with two simple cases that are illustrative of possible space weather applications. We assume that a spacecraft operator (or stakeholder) is interested in predicting relativistic electron flux or fluence at a specific spacecraft. First, we present the case of predicting electron flux at a satellite in a highly elliptical, near-equatorial orbit, using a model that simulates a larger domain. The satellite orbit thus represents a sparse trajectory through the model domain. We then present the case of predicting daily electron fluence at geosynchronous orbit, using a model that predicts exactly this quantity. It will be clear that no single metric captures the full relationship between model and observation. For predictands that vary over orders of magnitude and where overprediction or underprediction by the same factor should be penalized equally, ζ and SSPB give robust and easily interpretable results. Other commonly used metrics penalize the errors differently and can be hard to interpret. Full presentations of model validation are beyond the scope of this work, and we use these examples as illustrative case studies. For rigorous model validation, much longer time periods should be used, covering a wide range of conditions, as well as performing quantitative comparison across the model domain. Further comments on the use of summary metrics, especially for higher dimensionality data, are given in section 8.

7.1. Predicting Electron Flux Along an Orbit

In this first simple case, we require a 1-D time series of the electron flux at a given location and to quantify the model performance we are interested in summarizing the model accuracy and bias for the simulation interval. We use data from MagEIS as our observation and output from the Dynamic Radiation Environment Assimilation Model (DREAM) (Reeves, 2011; Reeves et al., 2012) as our prediction. The configuration of DREAM used for this simulation is a 1-D radial diffusion model that uses an ensemble Kalman filter for data assimilation, with a source term whose amplitude is estimated as part of the assimilation process (see section 4.4 of Reeves et al., 2012). As part of an ongoing validation study of DREAM, the month of January 2014 was run with input data from the Synchronous Orbit Particle Analyzer (Belian et al., 1992) on three Los Alamos geosynchronous satellites (1994-084, LANL-01A and LANL-04A). A virtual satellite was flown through the model output along the trajectory of the Van Allen Probes RBSP A satellite, where apogee is inside geosynchronous orbit, and the omnidirectional, differential number flux at 1.07 MeV was calculated.

Presenting only this short interval, with limited dynamics, ensures that the aspects of model performance displayed through this interval are not masked by a large number of data points or varying model performance as time and conditions change. We first describe the model performance qualitatively and then calculate a range of metrics. The interpretation of these metrics will then be placed in the context of the qualitative description, so that the behavior of these metrics can be compared and discussed. Figure 7a shows the omnidirectional flux measured by MagEIS on RBSP A (blue) and the flux at the same location predicted

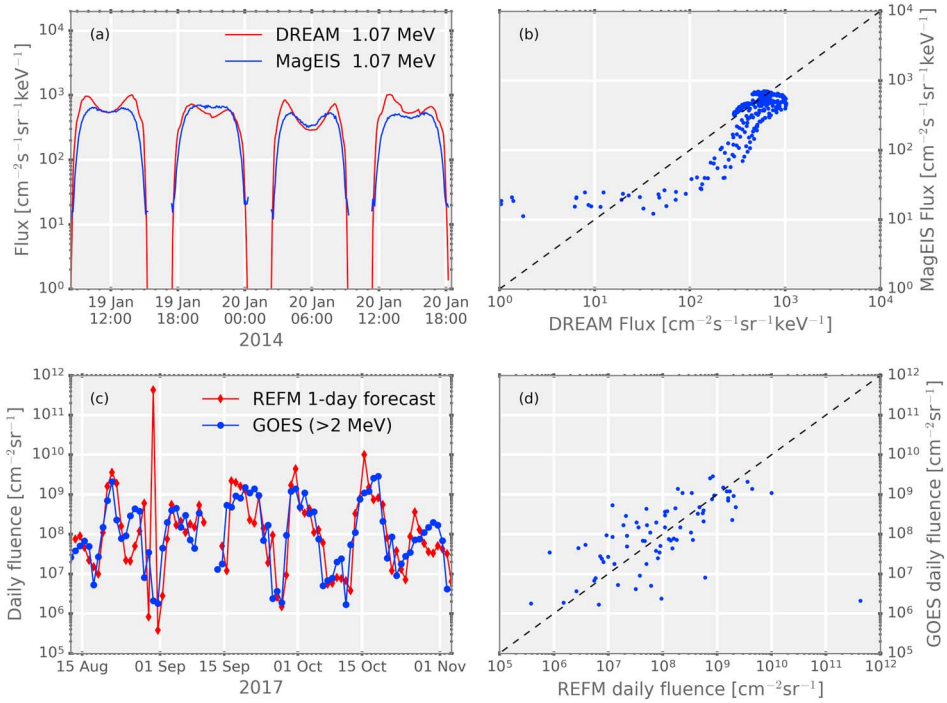

Figure 7. Comparisons of model flux or fluence to observations. Figures 7a and 7b show a comparison of omnidirectional electron flux from the MagEIS instrument on RBSP A with predicted 1.07 MeV electron flux at that orbit from DREAM. Times where the Van Allen Probes orbit was outside the domain of the DREAM run have been removed. (a) The fluxes as a function of time. (b) A scatterplot of the DREAM fluxes and MagEIS fluxes, with $y = x$ marked by a black dashed line. (c, d) The same format as Figures 7a and 7b but showing daily fluence at >2 MeV from GOES compared with the 1 day ahead forecast from the REFM model.

by DREAM (red). Times when the orbit of RBSP A was outside the model domain have been masked from both time series and removed from this analysis. It can be seen that the fluxes are qualitatively similar and that variation in fluxes covers orders of magnitude. Figure 7b shows a scatterplot of the observed and predicted flux. The abscissa is the flux predicted by DREAM, and the ordinate is the flux observed by MagEIS. A black dashed line corresponding to $y = x$ has been added to the plot.

Inspection of Figure 7a shows that at high fluxes, near the apogee of the Van Allen Probes orbit, the errors are typically smaller but DREAM tends to slightly overpredict. Due to the slower orbital speed near apogee, the majority of data points fall in this region. For this short time interval, DREAM consistently overestimates the flux as the satellite more rapidly moves between apogee and perigee. As the inner boundary of the model domain is approached, the MagEIS flux reaches a point of inflection while the DREAM flux continues to fall thereby causing DREAM to underestimate the flux. During this interval there is minimal temporal variation throughout the radiation belt and the bulk of the variation seen along the RBSP A orbit is due to its sampling of a minimally varying spatial structure of the radiation belt. Applying the metrics defined in this paper, we calculate that ζ is 34.6% and the SSPB is 21.1%. The interpretation of these metrics is that half of the forecast errors are smaller than a factor of 1.35 and that the median forecast error is an overestimate by 21.1%.

For comparison, we have calculated the other accuracy and bias metrics discussed above. The MAPE is 65.59%, which is higher than ζ due to two main factors: the tendency of DREAM to overpredict the flux results in a larger penalty, although this by itself would tend to make MAPE similar to ζ rather than exceeding it; the mean error is much larger than the median due to the strongly asymmetric distribution of forecast errors. The same reasons lead to a mean percentage error of 50.7%. Calculating sMAPE gives 44.4%. Bearing the caveats of section 3.2 in mind, we also calculate the MAPE of the log-transformed flux. This results in an accuracy of 12.4%, and visual inspection of Figure 7 clearly shows that the typical forecast error is somewhat larger than this; the MAPE of log flux would also be much smaller if we converted to differential flux per MeV. We can also use scale-dependent measures to assess the accuracy and bias. The RMSE for this prediction

is 202 $cm^{-2}s^{-1}sr^{-1}keV^{-1}$, and the mean error is 111 $cm^{-2}s^{-1}sr^{-1}keV^{-1}$. While the RMSE and mean error are not incorrect, they more heavily weight large magnitudes of deviation, which are actually the smaller relative errors in this situation. The RMSE of log-transformed flux is 0.38, which is an estimate of σ for a Gaussian noise model; however, it is clear that the errors are not normally distributed in log space. The spread of the error distribution is robustly estimated as $Sn(\log_e(Q)) = 0.21$, which is significantly smaller than the estimate using RMSE of log flux.

7.2. Predicting Daily Electron Fluence at Geosynchronous Orbit

For this example we show the daily >2 MeV fluence from GOES and the prediction of that same quantity using the REFM model (based on Baker et al., 1990), as reported by NOAA's Space Weather Prediction Center. Figure 7c shows the daily >2 MeV fluence measured by GOES (blue) and the prediction for that day from REFM (red). It can be seen that the fluences are qualitatively similar and that variation in daily fluence covers orders of magnitude. Figure 7d shows a scatterplot of the observed and predicted fluence. Inspection of Figure 7 shows that there is no clear systematic behavior in the errors over this interval. The data files for the displayed interval have fill values for both the observed fluence and the predictions for 10–12 September 2017. These are excluded from the plotting and the analysis. In addition, the 1 day ahead prediction from 29 August 2017 is a significant overestimate and appears as a significant outlier in Figure 7d. Applying the metrics defined in this paper, we calculate that ζ is 180.4% and the SSPB is −11.6%. The interpretation of these metrics is that half of the forecast errors are smaller than a factor of 2.8 and that the median forecast error is an underestimate by 11.6%. The MAPE and MPE are dominated by the outlier and are both about 2.63×10^5%. We therefore exclude this point from the rest of our analysis. On excluding the outlier we find that ζ and SSPB have changed only slightly, at 177.7% and −15.4% respectively.

As above, we have again calculated a range of accuracy and bias metrics (after excluding the fill values and the outlier). The MAPE is 276.1%, which is higher than ζ due to a few points with larger errors dominating the mean. For the same reason the mean percentage error is 210.5%, suggesting a mean overestimate of around a factor of 3. Note, however, that the SSPB is −15.4%, showing that most forecasts in this interval actually underpredict slightly. Looking at the other metrics, we see that sMAPE = 94.9%, which would incorrectly imply that the typical error is less than a factor of 2. As before, we bear the caveats of section 3.2 in mind and calculate the MAPE of the log-transformed flux. This results in an accuracy of 6.7%, which is clearly not representative of the actual forecast errors. Looking at the scale-dependent measures for accuracy and bias, we see that the RMSE for this prediction is 1.20×10^9 $cm^{-2}sr^{-1}$ and the mean error is 1.66×10^8 $cm^{-2}sr^{-1}$. While not technically incorrect, these metrics do not clearly communicate how well the model actually performs. The RMSE of log-transformed flux is 0.66, and $Sn(\log_e(Q)) = 0.69$ suggesting that the errors are close to normally distributed in log space. We note that we cannot compare any of these metrics to the performance statistics supplied by NOAA as they provide skill relative to three reference forecasts (sample mean, persistence, and recurrence) and do not explicitly give estimates of accuracy or bias.

8. Quantifying and Understanding Model Performance

We note that for simplicity we have used 1-D time series examples throughout and that our example illustrates the use of accuracy and bias metrics to summarize model performance. Calculating summary metrics, aggregated across all data, is useful for the scenarios described in section 7. This approach would not, however, allow a model developer to fully understand where or why their higher-dimensional model is inaccurate. For this use case, different approaches will likely be required.

For example, Schiller et al. (2017) investigated the differences between two radiation belt simulations (where their output was PSD(μ = const, K = const, L^*, t)) of the same interval using several methods; each method employed illustrates a different aspect of model performance. The difference between their two model runs was in the loss and transport terms: model 1 used event-specific terms and model 2 used statistical models to obtain the loss and transport terms. To understand where the model runs differ, and by how much, Schiller et al. (2017) present $\log_{10}(Q)$ as a function of time and L^* (see their Figure 8c). This visualizes the relative difference between the model runs in a 2-D slice of their model domain, allowing them to diagnose where and when the models differ.

Schiller et al. (2017) additionally quantify the performance of each model run by validating against phase space density measured at satellites from the Time History of Events and Macroscale Interactions during

Substorms (THEMIS) mission. The THEMIS satellites trace trajectories through the model domain and hence only sample part of the model space. To quantify the accuracy of each of their model runs, as a function of time, they calculate RMSE (between model and THEMIS observation) aggregated over all L^* and over 15 min windows in time. Their model accuracy is then quantified by reporting the RMSE as a function of time. This model validation approach mirrors the situation presented in section 7. The model performance over the full interval could be summarized using ζ and SSPB as described above and could be displayed as a function of time by aggregating over subsets of the data similar to Schiller et al. (2017).

As mentioned previously, Subbotin and Shprits (2009) have developed metrics aimed at understanding where and when differences between models exist. These metrics are typically applied to subsets of the model domain. For example, to compare 2-D slices of PSD(L^*,t) at constant μ and K they use ND (cf. equation (7). This metric is similar to sMAPE in that the normalization uses the mean of x and y, but the normalization factor is constant for any given time and is given by $\max(y_i(f)+x_i(f))/2$ where the maximum value is taken over all L^* at a given time. An additional example of the ND metric being applied to characterize model performance over a 2-D domain was given by Drozdov et al. (2017), who compared Van Allen Probes electron flux data (binned in L^* and time) with simulation output. They note that they use ND for this as "[i]t emphasizes how well the simulation can reproduce the flux peaks and flux profiles around the maximum. In case of the comparison between two simulations, it indicates the difference in the heart of the radiation belt and excludes the areas of the low flux values, such as the slot region to avoid comparison of very small numbers." Thus, while the absolute value of ND may not be intuitive, it has demonstrated utility in understanding model performance from a physical perspective.

The metrics presented in this paper can be applied to higher-dimensional data by, for example, aggregating across particular dimensions of the data. For quantitative analysis of higher-dimensional data other metrics for data-model comparison have been developed (see, e.g., Ch. 7 of Wilks, 2006) that have not been discussed in this paper. For properly characterizing the performance of a model, the particular meaning of performance metrics and the intended use (overall accuracy for customer, diagnosing deficiencies in model physics, etc.) should be considered. Derived quantities can also help understand model performance, such as the location of the peak in PSD in a radiation belt model. We reiterate our earlier statement that no single metric captures the full relationship between model and observation. In the cases of comparing 2-D (or higher-dimensional) domains the metrics presented in this paper could be used, with appropriate aggregation over subsets of the domain but may not be appropriate for answering the questions posed by the model developer. Summary metrics aggregated over all data may also be desirable in these cases so that overall model performance can be assessed in tandem with localization of any model errors.

9. Summary

In situations where observed (or modeled) data can vary over orders of magnitude, we identify four desirable properties for accuracy and bias metrics: (1) The metrics must be meaningful for data that cover orders of magnitude, (2) underprediction and overprediction by the same factor should be penalized equally, (3) the metrics should be easy to interpret, and (4) the metrics should be robust to the presence of outliers and bad data. We have reviewed a number of commonly used model performance metrics and have illustrated the ways in which these metrics do not display the given desirable properties. We have presented new measures of accuracy and bias and demonstrated that they satisfy all listed desirable properties. The metrics discussed in this paper are summarized in Table 1.

The new metrics presented in this work are interpretable as percentages but are designed to address known problems with standard metrics based on percentage errors. To address these drawbacks while still preserving the interpretability of MAPE, we present an accuracy measure based on the logarithm of the accuracy ratio. This measure can be interpreted as a percentage error but does not penalize overprediction and underprediction differently. This accuracy metric is called the median symmetric accuracy (cf. Morley, 2016), ζ, which is defined as

$$\zeta = 100\left(\exp\left(M\left(\left|\log(Q)\right|\right)\right) - 1\right)$$

In this paper we have shown that ζ is equivalent to the median unsigned percentage error and we have demonstrated its performance relative to other accuracy metrics similar to MAPE, showing that it satisfies the listed desirable properties. To provide a measure of bias that also satisfies the listed desirable

properties, we derive and describe the Symmetric Signed Percentage Bias (SSPB) which is also based on the log accuracy ratio.

$$\mathrm{SSPB} = 100\ \mathrm{sgn}(\mathrm{MdLQ})(\exp(|\mathrm{MdLQ}|) - 1)$$

Metrics based on ratios, including relative errors, can be undefined where zeros are present, and we suggest that in some cases a threshold related to the limits of measurement capability could be applied to both prediction and observation for the purposes of assessing model accuracy and bias.

We have also shown how the log accuracy ratio is related to the standard deviation of a multiplicative linear model and use robust estimators of the spread of $\log(Q)$ to estimate σ in a multiplicative linear model. We recommend the use of $S_n(\log_e(Q))$ for this purpose, where S_n is a robust measure of spread first described by Rousseeuw and Croux (1993).

In cases where accuracy and bias metrics are required that equally penalize errors of the same order—typically predictands that span many orders of magnitude, such as radiation belt fluxes—we recommend the median symmetric accuracy and the symmetric signed percentage bias. These new metrics are easily interpreted and address some of the known problems associated with more standard approaches based on relative errors and percentage errors. We have illustrated the use of these metrics with a simple example of predicting electron flux along a satellite orbit. We have discussed some additional considerations required for more complicated use cases.

Acknowledgments
This work was performed under the auspices of the U.S. Department of Energy. S. K. M. and T. V. B. acknowledge support from the Laboratory Directed Research and Development (LDRD) program, projects 20150127ER and 20150033DR. D. T. W. acknowledges support from LDRD20150033DR. The DREAM output used in this work is available on request from the corresponding author. GOES fluence data and REFM predictions were obtained from NOAA's Space Weather Prediction Center at http://services.swpc.noaa.gov. Analysis and plotting used the publicly available SpacePy library. SpacePy is available at http://sourceforge.net/p/spacepy. S. K. M. thanks Paul O'Brien for discussions motivating some of the presented work.

References

Athanasiu, M. A., Pavlos, G. P., Sarafopoulos, D. V., & Sarris, E. T. (2003). Dynamical characteristics of magnetospheric energetic ion time series: Evidence for low dimensional chaos. *Annales Geophysicae*, *21*, 1995–2010. https://doi.org/10.5194/angeo-21-1995-2003

Baker, D. N., McPherron, R. L., Cayton, T. E., & Klebesadel, R. W. (1990). Linear prediction filter analysis of relativistic electron properties at 6.6 R_E. *Journal of Geophysical Research*, *95*, 15,133–15,140. https://doi.org/10.1029/JA095iA09p15133

Belian, R. D., Gisler, G. R., Cayton, T., & Christensen, R. (1992). High-*Z* energetic particles at geosynchronous orbit during the Great Solar Proton Event Series of October 1989. *Journal of Geophysical Research*, *97*, 16,897–16,906. https://doi.org/10.1029/92JA01139

Blake, J., Carranza, P., Claudepierre, S., Clemmons, J., Crain Jr., W. R., Dotan, Y., … Zakrzewski, M. (2013). The Magnetic Electron Ion Spectrometer (MagEIS) instruments aboard the Radiation Belt Storm Probes (RBSP) spacecraft. *Space Science Reviews*, *179*(1–4), 383–421. https://doi.org/10.1007/s11214-013-9991-8

Brito, T. V., & Morley, S. K. (2017). Improving empirical magnetic field models by fitting to in situ data using an optimized parameter approach. *Space Weather*, *15*, 1628–1648. https://doi.org/10.1002/2017SW001702

Chen, Y., Friedel, R. H. W., Reeves, G. D., Cayton, T. E., & Christensen, R. (2007). Multisatellite determination of the relativistic electron phase space density at geosynchronous orbit: An integrated investigation during geomagnetic storm times. *Journal of Geophysical Research*, *112*, A11214. https://doi.org/10.1029/2007JA012314

Coleman, C. D., & Swanson, D. A. (2007). On MAPE-R as a measure of cross-sectional estimation and forecast accuracy. *Journal of Economic and Social Measurement*, *32*, 219–233.

Déqué, M. (2011). Deterministic forecasts of continuous variables. In I. T. Jolliffe, & D. B. Stephenson (Eds.), *Forecast verification* (pp. 77–94). Chichester, UK: John Wiley. https://doi.org/10.1002/9781119960003.ch5

Drozdov, A. Y., Shprits, Y. Y., Aseev, N. A., Kellerman, A. C., & Reeves, G. D. (2017). Dependence of radiation belt simulations to assumed radial diffusion rates tested for two empirical models of radial transport. *Space Weather*, *15*(1), 150–162. https://doi.org/10.1002/2016SW001426

Flores, B. E. (1986). A pragmatic view of accuracy measurement in forecasting. *Omega*, *14*(2), 93–98. https://doi.org/10.1016/0305-0483(86)90013-7

Francq, C., & Menvielle, M. (1996). A model for the *Am* (*Km*) planetary geomagnetic activity index and application to prediction. *Geophysical Journal International*, *125*, 729–746. https://doi.org/10.1111/j. 1365-246X.1996.tb06020.x

Friedel, R. H. W., Reeves, G. D., & Obara, T. (2002). Relativistic electron dynamics in the inner magnetosphere—A review. *Journal of Atmospheric and Solar-Terrestrial Physics*, *64*, 265–282. https://doi.org/10.1016/S1364-6826(01)00088-8

Ginet, G. P., O'Brien, T. P., Huston, S. L., Johnston, W. R., Guild, T. B., Friedel, R., … Su, Y.-J. (2013). AE9, AP9 and SPM: New models for specifying the trapped energetic particle and space plasma environment. *Space Science Reviews*, *179*, 579–615. https://doi.org/10.1007/s11214-013-9964-y

Glocer, A., Tóth, G., Gombosi, T., & Welling, D. (2009). Modeling ionospheric outflows and their impact on the magnetosphere, initial results. *Journal of Geophysical Research*, *114*, A05216. https://doi.org/10.1029/2009JA014053

Grillakis, M. G., Koutroulis, A. G., & Tsanis, I. K. (2013). Multisegment statistical bias correction of daily GCM precipitation output. *Journal of Geophysical Research: Atmospheres*, *118*, 3150–3162. https://doi.org/10.1002/jgrd.50323

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*(346), 383–393.

Hwang, J., Kim, K. C., Dokgo, K., Choi, E., & Kim, H. P. (2015). Heliocentric potential (HCP) prediction model for nowscast of aviation radiation dose. *Journal of Astronomy and Space Science*, *22*(1), 39–44. https://doi.org/10.5140/JASS.2015.32.1.39

Hyndman, R. J., & Anathasopoulos, G. (2014). *Forecasting: Principles and practice*. Melbourne, Australia: OTexts.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Jolliffe, I. T., & Stephenson, D. B. (2011). Introduction. In I. T. Jolliffe, & D. B. Stephenson (Eds.), *Forecast verification* (pp. 1–9). Chichester, UK: John Wiley. https://doi.org/10.1002/9781119960003.ch1

Kim, K.-C., Shprits, Y., Subbotin, D., & Ni, B. (2012). Relativistic radiation belt electron responses to GEM magnetic storms: Comparison of CRRES observations with 3-D VERB simulations. *Journal of Geophysical Research*, *117*, A08221. https://doi.org/10.1029/2011JA017460

Kitchenham, B. A., Pickard, L. M., MacDonell, S. G., & Shepperd, M. J. (2001). What accuracy statistics really measure [software estimation]. *IEE Proceedings – Software*, *148*(3), 81–85. https://doi.org/10.1049/ip-sen: 20010506

Kohzadi, N., Boyd, M. S., Kermanshahi, B., & Kaastra, I. (1996). A comparison of artificial neural network and time series models for forecasting commodity prices. *Neurocomputing*, *10*(2), 169–181. https://doi.org/https://doi.org/10.1016/0925-2312(95)00020-8

Li, X., Baker, D. N., Temerin, M., Reeves, G., Friedel, R., & Shen, C. (2005). Energetic electrons, 50 keV to 6 Mev, at geosynchronous orbit: Their responses to solar wind variations. *Space Weather*, *3*, S04001. https://doi.org/10.1029/2004SW000105

Li, Z., Hudson, M., & Chen, Y. (2014). Radial diffusion comparing a THEMIS statistical model with geosynchronous measurements as input. *Journal of Geophysical Research: Space Physics*, *119*, 1863–1873. https://doi.org/10.1002/2013JA019320

Lundstedt, H., Gleisner, H., & Wintoft, P. (2002). Operational forecasts of the geomagnetic *Dst* index. *Geophysical Research Letters*, *29*(24), 2181. https://doi.org/10.1029/2002GL016151

Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, *9*(4), 527–529. https://doi.org/10.1016/0169-2070(93)90079-3

Mauk, B., Fox, N., Kanekal, S., Kessel, R., Sibeck, D., & Ukhorskiy, A. (2013). Science objectives and rationale for the Radiation Belt Storm Probes mission. *Space Science Reviews*, *179*(1–4), 3–27. https://doi.org/10.1007/s11214-012-9908-y

Menvielle, M., & Berthelier, A. (1991). The K-derived planetary indices: Description and availability. *Reviews of Geophysics*, *29*(3), 415–432. https://doi.org/10.1029/91RG00994

Morley, S. K. (2016). Alternatives to accuracy and bias metrics based on percentage errors for radiation belt modeling applications (Tech. Rep. LA-UR-16-24592). Los Alamos, NM: Los Alamos National Laboratory. https://doi.org/10.2172/1260362

Morley, S. K., Sullivan, J. P., Carver, M. R., Kippen, R. M., Friedel, R. H. W., Reeves, G. D., & Henderson, M. G. (2017). Energetic particle data from the global positioning system constellation. *Space Weather*, *15*, 283–289. https://doi.org/10.1002/2017SW001604

Morley, S. K., Sullivan, J. P., Henderson, M. G., Blake, J. B., & Baker, D. N. (2016). The Global Positioning System constellation as a space weather monitor: Comparison of electron measurements with Van Allen Probes data. *Space Weather*, *14*(2), 76–92. https://doi.org/10.1002/2015SW001339

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, *8*(2), 281–293. https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2

O'Brien, T. P., & McPherron, R. L. (2003). An empirical dynamic equation for energetic electrons at geosynchronous orbit. *Journal of Geophysical Research*, *108*(A3), 1137. https://doi.org/10.1029/2002JA009324

Osthus, D., Caragea, P. C., Higdon, D., Morley, S. K., Reeves, G. D., & Weaver, B. P. (2014). Dynamic linear models for forecasting of radiation belt electrons and limitations on physical interpretation of predictive models. *Space Weather*, *12*, 426–446. https://doi.org/10.1002/2014SW001057

Reeves, G. D. (2011). DREAM: An integrated space radiation nowcast system for natural and nuclear radiation belts, *Proceedings of the Advanced Maui Optical and Space Surveillance Technologies Conference (AMOS) Maui, HI, September 14–17, 2011* (p. E2). Retrieved from http://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-11-06307

Reeves, G. D., Morley, S. K., Friedel, R. H. W., Henderson, M. G., Cayton, T. E., Cunningham, G., … Thomsen, D. (2011). On the relationship between relativistic electron flux and solar wind velocity: Paulikas and Blake revisited. *Journal of Geophysical Research*, *116*, A02213. https://doi.org/10.1029/2010JA015735

Reeves, G. D., Chen, Y., Cunningham, G. S., Friedel, R. W. H., Henderson, M. G., Jordanova, V. K., … Zaharia, S. (2012). Dynamic radiation environment assimilation model: DREAM. *Space Weather*, *10*, S03006. https://doi.org/10.1029/2011SW000729

Reeves, G. D., Spence, H. E., Henderson, M. G., Morley, S. K., Friedel, R. H. W., Funsten, H. O., … Niehof, J. T. (2013). Electron acceleration in the heart of the Van Allen radiation belts. *Science*, *341*(6149), 991–994. https://doi.org/10.1126/science.1237743

Reikard, G. (2011). Forecasting space weather: Can new econometric methods improve accuracy? *Advances in Space Research*, *47*(12), 2073–2080. https://doi.org/https://doi.org/10.1016/j.asr.2011.03.037

Rodriguez, J. V., Sandberg, I., Mewaldt, R. A., Daglis, I. A., & Jiggens, P. (2017). Validation of the effect of cross-calibrated goes solar proton effective energies on derived integral fluxes by comparison with stereo observations,. *Space Weather*, *15*, 290–309. https://doi.org/10.1002/2016SW001533

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*(424), 1273–1283. https://doi.org/10.1080/01621459.1993.10476408

Schiller, Q., Tu, W., Ali, A. F., Li, X., Godinez, H. C., Turner, D. L., … Henderson, M. G. (2017). Simultaneous event-specific estimates of transport, loss, and source rates for relativistic outer radiation belt electrons. *Journal of Geophysical Research: Space Physics*, *122*, 3354–3373. https://doi.org/10.1002/2016JA023093

Selesnick, R. S., & Blake, J. B. (1997). Dynamics of the outer radiation belt. *Geophysical Research Letters*, *24*(11), 1347–1350. https://doi.org/10.1029/97GL51409

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). New York: Chapman and Hall/CRC.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Subbotin, D. A., & Shprits, Y. Y. (2009). Three-dimensional modeling of the radiation belts using the versatile electron radiation belt (verb) code. *Space Weather*, *7*, S10001. https://doi.org/10.1029/2008SW000452

Swanson, D. A., Tayman, J., & Barr, C. F. (2000). A note on the measurement of accuracy for subnational demographic estimates. *Demography*, *37*(2), 193–201.

Swanson, D. A., Tayman, J., & Bryan, T. M. (2011). MAPE-R: A rescaled measure of accuracy for cross-sectional subnational population forecasts. *Journal of Population Research*, *28*(2), 225–243. https://doi.org/10. 1007/s12546-011-9054-5

Thornes, J. E., & Stephenson, D. B. (2001). How to judge the quality and value of weather forecast products. *Meteorological Applications*, *8*(3), 307–314. https://doi.org/10.1017/S1350482701003061

Tofallis, C. (2015). A better measure of relative prediction accuracy. *Journal of the Operational Research Society*, *66*(8), 1352–1362.

Tsyganenko, N. A. (2013). Data-based modelling of the Earth's dynamic magnetosphere: A review. *Annales Geophysicae*, *31*(10), 1745–1772. https://doi.org/10.5194/angeo-31-1745-2013

Tu, W., Cunningham, G. S., Chen, Y., Henderson, M. G., Camporeale, E., & Reeves, G. D. (2013). Modeling radiation belt electron dynamics during GEM challenge intervals with the DREAM3D diffusion model. *Journal of Geophysical Research: Space Physics*, *118*, 6197–6211. https://doi.org/10.1002/jgra.50560

Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, *28*(6), 815–829. https://doi.org/10.1111/j.2005.0906-7590.04112.x

Weiss, L. A., Thomsen, M. F., Reeves, G. D., & McComas, D. J. (1997). An examination of the Tsyganenko (T89a) field model using a database of two-satellite magnetic conjunctions. *Journal of Geophysical Research*, *102*, 4911–4918. https://doi.org/10.1029/96JA02876

Welling, D. T. (2010). The long-term effects of space weather on satellite operations. *Annales Geophysicae*, *28*(6), 1361–1367. https://doi.org/10.5194/angeo-28-1361-2010

Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences* (2nd ed.). Oxford: Academic Press.

Yu, Y., Koller, J., Jordanova, V. K., Zaharia, S. G., Friedel, R. W., Morley, S. K., … Spence, H. E. (2014). Application and testing of the *L** neural network with the self-consistent magnetic field model of RAM-SCB. *Journal of Geophysical Research: Space Physics*, *119*, 1683–1692. https://doi.org/10.1002/2013JA019350

Zhelavskaya, I. S., Spasojevic, M., Shprits, Y. Y., & Kurth, W. S. (2016). Automated determination of electron density from electric field measurements on the Van Allen Probes spacecraft. *Journal of Geophysical Research: Space Physics*, *121*, 4611–4625. https://doi.org/10.1002/2015JA022132

Zheng, Y., & Rosenfeld, D. (2015). Linear relation between convective cloud base height and updrafts and application to satellite retrievals. *Geophysical Research Letters*, *42*, 6485–6491. https://doi.org/10.1002/2015GL064809