Journal of Atmospheric and Solar–Terrestrial Physics 218 (2021) 105624

Contents lists available at ScienceDirect

Journal of Atmospheric and Solar-Terrestrial Physics

journal homepage: www.elsevier.com/locate/jastp

RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics

Michael W. Liemohn [a,*], Alexander D. Shane [a], Abigail R. Azari [b], Alicia K. Petersen [a,c], Brian M. Swiger [a], Agnit Mukhopadhyay [a]

[a] Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI, USA
[b] Space Sciences Laboratory, University of California, Berkeley, CA, USA
[c] Now at Air Force Research Laboratory, Kirtland Air Force Base, Albuquerque, NM, USA

ARTICLE INFO

Keywords:
Data-model comparisons
Metrics
Fit performance
Event detection
Space weather
Magnetospheric physics
Forecasting

ABSTRACT

The magnetospheric physics research community uses a broad array of quantitative data-model comparison methods (metrics) when conducting their research investigations. It is often the case, though, that any particular study will only use one or two metrics, with the two most common being Pearson correlation coefficient and root mean square error (RMSE). Because metrics are designed to test a specific aspect of the data-model relationship, limiting the comparison to only one or two metrics reduces the physical insights that can be gleaned from the analysis, restricting the possible findings from modeling studies. Additional physical insights can be obtained when many types of metrics are applied. We organize metrics into two primary groups: 1) fit performance metrics, often based on the data-model value difference; and 2) event detection metrics, which use a discrete event classification of data and model values determined by a specified threshold. In addition to these groups, there are several major categories of metrics based on the aspect of the data-model relationship that the metric assesses: 1) accuracy; 2) bias; 3) precision; 4) association; 5) and extremes. Another category is skill, which is a measure of any of these metrics against the performance of a reference model. These can be applied to a subset of either the data or the model values, known as reliability and discrimination assessments. In the context of magnetospheric physics examples, we discuss best practices for choosing metrics for particular studies.

1. Introduction

The Earth's magnetic field not only protects our planet but also channels energy into geospace, serving both as a shield and a funnel. This magnetic field extends into near-Earth space and would continue indefinitely, except that other magnetic fields exist in the solar system to confine it, such as the interplanetary magnetic field (IMF) carried by the solar wind (Axford and Hines, 1961; Dungey, 1961). While canonical configurations of this magnetized bubble around the planet, known as the magnetosphere, exist (see recent reviews by Tanaka, 2007; Ganushkina et al., 2018), it is in a continuous dynamical state as the solar wind and IMF deform it, sometimes into extreme configurations (e.g., Tsurutani et al., 2003; Siscoe et al., 2006; Cid et al., 2015). Since Explorer 1, the first successful spacecraft with a scientific payload, measured energetic particles and discovered the existence of the radiation belts (Van Allen et al., 1958), scholars have been developing explanations for the observed phenomena within the magnetosphere. These descriptions are developed and tested with approaches including analytical theories, coupled suites of numerical models and data-based empirical models. In every case, they give values that can be compared against observations.

Methods of quantitatively comparing model output to observational data are called metrics. The magnetospheric physics research community uses many different data-model comparison formulas, several of which are shown in Fig. 1. Most space weather modeling studies include forecast verification (i.e., model validation) involving a range of sophistication levels for their data-model comparison analysis. Space weather research, which has an operational focus seeking to understand the present or future state of potential electromagnetic hazards in outer space, involves nowcasts, forecasts, and post-event analysis towards improved decision making regarding assets that are adversely affected by the space environment. For operational space weather usages of magnetospheric models, it is up to the user to decide what metrics best define their decision-making needs (e.g., Halford et al., 2019), and the

* Corresponding author.
E-mail address: liemohn@umich.edu (M.W. Liemohn).

https://doi.org/10.1016/j.jastp.2021.105624
Received 19 July 2020; Received in revised form 14 December 2020; Accepted 18 March 2021
Available online 1 April 2021

more metrics that are calculated, the more information that user will have for deciding a course of action. As reviewed by Morley (2020), most of the progress in metrics usage and adoption by the space research community has been with the intention of improving space weather predictions.

Compared to space weather, magnetospheric physics has taken a much slower path to adopting comprehensive metrics into research studies. Magnetospheric physics is defined here as the pursuit of fundamental knowledge about the magnetic-dominated space environment around Earth and other planets. While cogent reviews on the use and misuse of statistics in space physics have existed for decades (e.g., Reiff, 1990), advocating for root mean square error (RMSE) and the Pearson linear correlation coefficient (R), ubiquitous application of metrics for basic research in the field is a recent phenomenon. Efforts towards large-scale data-model comparison did not become the normal until the Geospace Environment Modeling (GEM) program's series of magnetospheric community "challenges," supported by the Community Coordinated Modeling Center (CCMC). The first of these, less than 20 years ago, was the GEM Reconnection Challenge (Birn and Hesse, 2001), which included only model-model comparisons for idealized input scenarios. This was the exception, not the rule, however, and nearly all challenges since then have involved user or CCMC simulations of selected intervals, usually one or more with particular geomagnetic activity of interest to that group. For instance, the second GEM challenge focused on substorms, with several modeling groups conducting studies of the events. The data-model comparisons were not that sophisticated, however, either being entirely qualitative (e.g., Raeder et al., 2001) or using a single metric, like RMSE (e.g., Ridley, et al., 2002). The Inner Magnetosphere/Storms Assessment Challenge (Liemohn, 2006) involved many data providers in the formulation of the task, but unfortunately did not mandate particular metrics to be used by all participants, so much of the scientific advancement was based on few, if any, metrics.

GEM challenges for magnetospheric physics began to be more quantitative only in the last decade. A significant factor in this shift was the change of how the challenges were conducted, with CCMC conducting the metrics assessments. There was a series of "challenge" studies on ground-based and satellite-based magnetic field comparisons, such as Pulkkinen et al. (2010) followed by more in-depth analysis by Rastätter et al. (2011) and Pulkkinen et al. (2011). One of the biggest assessments was that of Pulkkinen et al. (2013), providing the objective comparison of several large-scale models against ground-based magnetometer station dB/dt time series to aid in the selection of one code for transition to space weather forecasting operations. The GEM community increased its focus on global model validation against various data sets, including the Dst index (Rastätter et al., 2013), in situ plasma parameters (Honkonen et al., 2013), empirical models of magnetospheric characteristics (Gordeev et al., 2015), and local K index values (Glocer et al., 2016). Another GEM challenge, which focused on the interactions of the magnetosphere with the ionosphere, resulted in several studies, such as focusing on Poynting flux and Joule heating (Rastätter et al., 2016) and another on total electron content (Shim et al. (2017). In creating the virtual model repository, Ridley et al. (2016) used over 600 simulations available at the CCMC to examine over 2000 satellite passes through model output, calculating RMSE against the observed magnetic field and compiling this into a star rating for each code. One of the latest GEM challenges sought to predict the inner magnetospheric electron fluxes responsible for spacecraft charging (Yu et al., 2019), effectively using a wide array of data-model comparisons to conclude that the inner magnetospheric electric field is, most likely, the largest discriminator in determining electron flux in near-Earth space. This rich history shows that the magnetospheric physics community has started accepting data-model comparisons as a standard practice for advancing knowledge and modeling development, but the robustness of these studies is mixed. There are still studies produced very recently that compare with data but only conduct qualitative assessments rather than rigorous metrics calculations. Some of these model-focused papers discuss data comparisons but do not include any plots or values, such as the Damiano et al. (2018) study of electron distribution modification by kinetic scale field line resonances. Others include direct overplots of data with the model results but provide no quantitative assessments, such as the Yu et al. (2017) unveiling of a new self-consistent electric field calculation in their kinetic drift physics model. Other studies, like Kalegaev et al. (2019) and Poedts et al. (2020), discuss large-scale modeling efforts for space weather prediction, including extensive plans for model validation, but no specifics or details of metrics usage.

To advance the conversation on space weather modeling assessments, CCMC staff organized a conference on this topic in April 2017. Several reports resulted from this with recommendations for magnetospheric physics, such as Welling et al. (2018) advocating contingency table use for quantifying model prediction of ground-based dB/dt events, Liemohn et al. (2018b) issuing metrics guidelines for geomagnetic index prediction, and Zheng et al. (2019) offering a framework for energetic charge particle assessments. A second conference on this topic was recently held in February 2020 to continue community engagement with this issue. A similar effort has been underway in Europe, with Opgenoorth et al. (2019) strongly recommending robust validation techniques for comparison of large-scale coupled space weather models, including the development of a consensus set of metrics for community adoption.

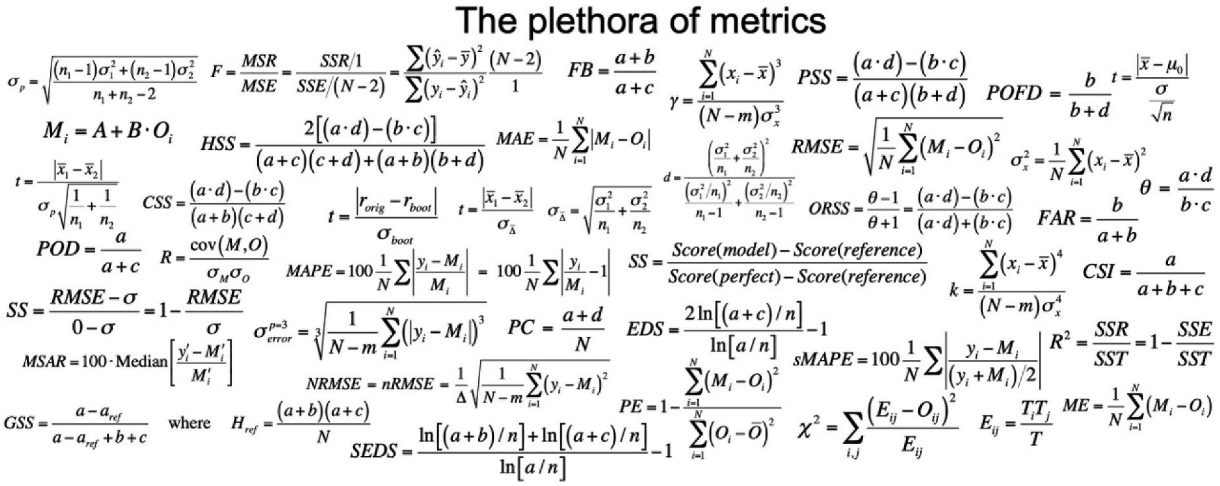

Fig. 1. Several of the metrics available for data-model comparisons.

Multiple metrics are highly advantageous for modern numerical modeling techniques. Due to the high dimensionality of most physics-based models, a single metric, designed to ascertain a particular facet of the data-model relationship, cannot fully describe the ability of the model to reproduce the observations and of outliers in particular. This concept of multi-faceted metric analysis has been understood for decades in the field of meteorology (e.g., Murphy, 1991), but has only recently been pointed out for magnetospheric physics studies (e.g., Kubo et al., 2017; Kubo, 2019). The approach of using many metrics can lead to additional inferences that affect the conclusions of scientific analyses. This is a recurrent theme throughout this review.

For magnetospheric scientific investigations, examining the full scope of metrics scores allows for additional knowledge discovery. Because the typical state of the magnetosphere is "quiet," those times deemed "active" are relatively rare. Extreme event behavior will, then, often not be represented very well in qualitative or single metric analyses because these events constitute only a small portion of the entire dataset. In a field where outlier detection and description are major aspects of scientific analyses, methods that focus outside of the canonical statistical focus on mean values is of great importance. Choosing only one or two metrics unnecessarily confines scientific advancements and effective operations with magnetospheric models. There is substantial usefulness in applying many metrics and, as discussed below, the magnetospheric physics community seems to be underutilizing the power of data-model comparison assessments.

The field of magnetospheric physics has entered a transitional time for data-model comparisons and quantitative metrics usage. As discussed by Liemohn et al. (2019), several factors are converging to make this happen. For one, the amount of data is dramatically increasing, in part due to better telemetry, smaller and more numerous satellites, and cheaper launch capabilities. This could lead to a new class of missions, enabling Magnetospheric Constellation and other distributed-arrays-in-space concepts (Kepko, 2018). Computational resources have also substantially improved, allowing for large-scale simulations with fine-scale resolution that merge the macroscopic and microscopic views of geospace physics. On the physical modeling side, the use of kinetic models in geospace system coupled simulations is becoming much more common, with an immense volume of code output readily available for comparison with particle velocity-space observations (see review by Wiltberger, 2015). Statistical modeling, for example machine learning techniques, have also pushed forward an interest in metrics. Machine learning techniques, with metrics analysis embedded in their creation, have been used for magnetospheric physics problems for decades (e.g., Lundstedt and Wintoft, 1994), but they have only recently gained widespread acceptance and use across this field (e.g., Camporeale, 2019) and should be paired with scientific understanding to optimize such algorithms (e.g., Azari et al., 2020; Swiger et al., 2020). The final component is an appreciation of data-model comparisons as a means to new scientific understanding, learning from our atmospheric science counterparts that the synthesis of forecast verification with fundamental research leads to better science (e.g., Folini, 2018).

Using a comprehensive metrics methodology will greatly improve the community's scientific return from data-model comparison studies. In this report, we review the basic groupings and categories of data-model comparison techniques and discuss the strengths and limitations of each of the common metrics. We then review the recent magnetospheric studies that have used these quantitative assessments and how that could be used in future magnetospheric physics investigations.

2. Organization of metrics

Because there are so many metrics from which to choose, it can be daunting for a researcher or model user to approach this vast list and pick those that are most appropriate for the desired assessment. Therefore, it is useful to organize the metrics according to how they are calculated and the quality that they assess. While there are many ways to cluster metrics, we discuss two, which we call groupings and categories. These are orthogonal definitions, with the former based on calculational method and the latter based on the characteristic of the data-model relationship that the metric examines. These groupings and categories make it easier to determine which metric is most appropriate for the planned assessment.

2.1. Metrics groupings: continuous versus discrete assessments

Metrics can be organized in many different ways; in this report we discuss one of the more common splits, dividing the metrics into two major groupings. These are based on how the metrics are calculated, regardless of the feature within the data-model comparison on which they focus.

The first grouping is called "fit performance" because these metrics use the exact values of both the data and model in their formulation. They are also called continuous metrics or regression metrics (e.g., Wilks, 2019). They are usually based on the difference between the data and model values, with most including a summation of these differences, often with extra mathematical operators included in the calculation, like squares, square roots, and absolute values. This is the grouping of the two most common metrics in use in magnetospheric physics, R and RMSE. Not all fit performance metrics include this difference of each data-model pairing, though; they might use the data and model values separately. That is, metrics based on probability distributions also fall within this grouping.

The other major grouping will be referred to as "event detection" and includes all metrics that classify each of the data and model values as either in or out of an "event state." This grouping is sometimes referred to as categorical metrics. Once the event state is determined, the exact value of either the data or model no longer matters, only the event status will be used in the metrics calculations. That is, the data-model scatter plot can be reduced to a 2×2 matrix, in which the elements depend on the event state of the data and model values. This table has many names (e.g., confusion matrix), but will be referred to as a contingency table in this report. The four numbers of this table also have many names, but we will use hits, false alarms, misses, and correct negatives (e.g., Joliffe and Stephenson, 2012). These values can be used in many different ways to understand the data-model relationship. Sometimes the data is collected as categorical values, other times the data are collected as continuous values and then converted to an event status by applying a threshold. This event identification threshold can be either the same or different for the data and model values, depending on the type of assessment being conducted and the needs of the eventual user of the information. Sweeping the event identification threshold setting can reveal the model's ability to detect events across a broad range of thresholds. Event detection is not limited to a dichotomous event status but could have three or more (e.g., n) states, making the contingency table into an n by n matrix. Event detection metrics represent a fundamentally different consideration toward data-model considerations, and ones in which outliers and/or events are the primary focus of comparison.

Fig. 2 shows schematics of the calculational emphasis for the fit performance and event detection groupings. Fit performance metrics use the individual M_i and O_i values, while event detection metrics count the points in each quadrant defined by the data and model event identification thresholds.

Fit performance metrics are by far the more common type of metric in use for magnetospheric space physics and space weather model assessments. They are, however, of limited use for investigations that focus on the rare "active times" of geospace. In this case, the methodology of event detection could be the better choice. Because each metric focuses on only a particular facet of model execution, using metrics from both groupings might be the optimal choice.

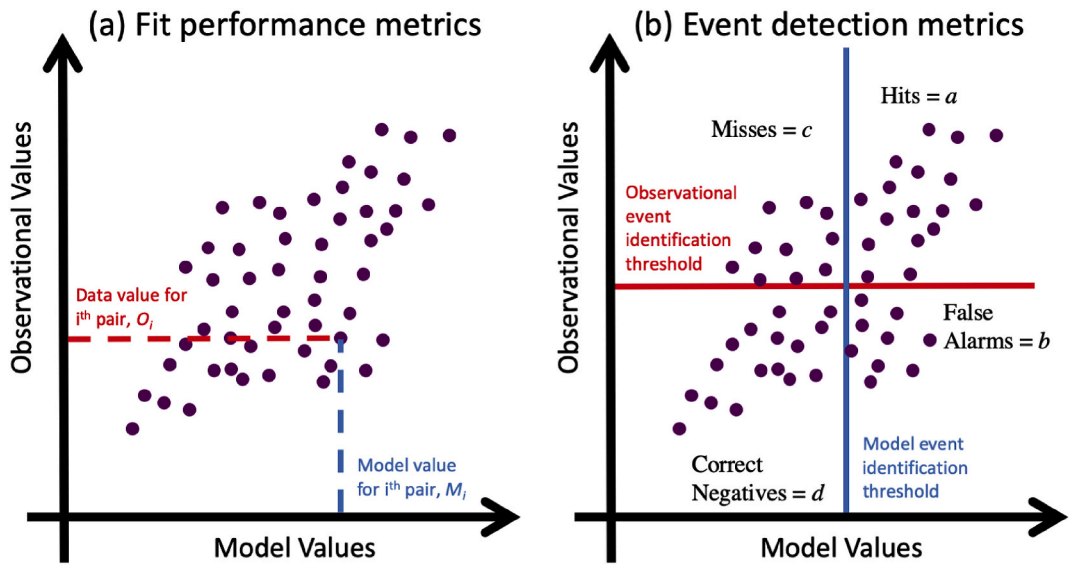

Fig. 2. Schematics of the calculational methodology for the two major metrics groupings, (a) fit performance metrics and (b) event detection metrics.

2.2. Metrics categories: analyzing particular facets of the data-model relationship

Independent of the method of calculation, metrics can be categorized by the property that they assess. There are numerous options for defining these categories; here we present some of the more common definitions. We divide the list into five primary categories and then discuss three derivative quantities in the next subsection.

2.2.1. Five major categories of metrics

The first and, arguably, foremost category is accuracy. This is defined as the closeness of the model values to the data values. Accuracy addresses the question of whether the model can exactly reproduce the data or not. If measures of accuracy are very good, then the comparison could, perhaps, stop here, as this indicates that other comparisons will, most likely reveal little in terms of information about model performance except to confirm the excellent accuracy values. When the measures of accuracy are less than perfect, however, then additional metrics from other categories are useful to better understand how the model is different from the data and therefore reveal a path for model improvement. While the metrics within the accuracy category quantify the overall quality of the data-model comparison, they cannot, by themselves, explain why the quality was good or bad.

Note that RMSE falls within the accuracy category. It is heartening that so many magnetospheric physicists choose this metric; it does a nice job at assessing the overall closeness of the model to the data values. However, RMSE might not be the optimal metric for a particular user's end needs. As discussed with the other categories below, RMSE cannot reveal if the model is systematically under or over predicting the observations; it cannot determine whether the difference is from many points being slightly off or because of a small cluster of outliers. Furthermore, comparing two models to the same data set and obtaining identical RMSE values does not mean that the models are identical in ability to reproduce the observations; it could be that the two models are completely different in their relationship to the data values, but this metric distills the comparison into a single number and does not provide any additional meaning.

To more robustly assess the data-model relationship, additional metrics from other categories are needed. The next category to be discussed is bias. Metrics in this category are comparisons of the centroids of the data and model values. Bias indicates if the model is systematically overestimating or underestimating the observations. A perfect bias score does not mean that the model values are identical to the data, but rather that the centroid of the model values matches the centroid of the data, whether that is average, median, or some other parameter. Bias is essentially a component of the accuracy calculation.

Precision is the next category, representing the complementary "other half" of accuracy. Metrics in this category seek to quantify the similarity in the clustering of the data and model values, after the centroid offsets have been removed. The equations for precision metrics, therefore, can be quite complicated. Similar to bias, a perfect precision value does not mean perfection. Since the bias is removed, a model with high precision could still have systematic differences between the model and the data. A possibly useful metaphor is that precision is temperature of a particle distribution, i.e., their random velocity, while bias is the bulk flow of the particles. In this metaphor, accuracy would then be the total velocity of the particles.

Another major category of metrics is association. Metrics in this category, including R, quantify the ability of the model to capture the up-and-down trends of the data. They often do not take into account how close the model values are to the data values; they usually only focus on whether the model is capable of increasing when the data increases or decreasing when it decreases. A perfect association metric indicates that the model captures this "motion" of the data but does not necessarily imply anything about accuracy. The inclusion of R along with RMSE in many data-model comparison studies makes for a good combination towards deciphering how the model behaves relative to the observations, but such a combination is only a beginning.

A final category to include in this set of common metrics is extremes. These are metrics that, while including all data-model pairs in the comparison, focus on the outliers of the distributions. As with other metrics, a perfect score for an extremes metric often does not indicate a perfect match to the data but rather that some aspect of the wings of the model distribution aligns with that of the observations. Because of the rarity of some modes of geospace activity, metrics from the extremes category could be particularly useful for magnetospheric studies.

2.2.2. Three major derivative categories of metrics

There are several metrics categories that use the formulas from other metrics in their definitions. In fact, the three to be discussed here could use metrics from any of the categories mentioned in the previous section. The three chosen for inclusion here are skill, discrimination, and reliability. Again, other metrics exist; this list is not meant to be exhaustive but rather a robust yet tractable listing of metrics categories that magnetospheric physicists could find particularly useful when conducting data-model comparisons.

The category of skill includes the type of calculation in which the metric value is compared against that from a reference model. While skill is often defined with an accuracy metric, any metric from one of the categories above can be used. Skill is usually defined as a skill score, which has a specific relationship between old and new model scores:

$$Skill\ Score = \frac{Metric(new\ model) - Metric(reference\ model)}{Metric(perfect\ value) - Metric(reference\ model)} \quad (1)$$

In (1), "new" represents the score of the metric value for the new model, "reference" is the metric value for the reference model, and "perfect" is the ideal data-model comparison for that metric. The perfect value for a skill score is, therefore, one. This is true regardless of the chosen metric. Note that, because of the subtraction in the numerator, a skill score of zero indicates the same skill for the new model as for the reference model, and a negative skill score denotes worse skill than the reference model. The reference model can be anything. Sometimes it is a previous model, other times it is based on the observations, and even occasionally it is "random chance; " there is no set rule on the definition of what constitutes a reference model.

The category of discrimination is unique in that these metrics use only a subset of the data range. Any metric can then be calculated for each data range subset. Subsets can be quite broad, splitting the data range into two segments, or quite fine, such as breaking the range into 20 bins and only using 5% of the total points for each subset. The term discrimination, therefore, is defined here as separation and identification. The most common calculation is to choose an accuracy metric, but this is not the only option. When applying a metric with discrimination subsetting, the interpretation of the result assesses the ability of the model to reproduce data only within each of the subsets. While it might seem strange to conduct a correlation coefficient calculation on only a quarter or a tenth of the total number of points, it is a useful examination of how well the model captures the up and down trends of the observations within that particular range of data values.

The category of reliability is the converse of discrimination; a reliability calculation examines portions of the full data-model comparison based on subsets of the model value range, not subsets of the data value range. Other than this switch over which number set the subsetting is performed, the calculational scheme is exactly the same as that for discrimination. The model value subset can be large, perhaps up to half the values, or very fine, with dozens or even hundreds of subset categories. All of the metrics categories above can be used on for reliability, assessing the model's ability to reproduce the data within a limited range of model values.

For event detection metrics, the subsetting ranges are easily defined as above and below the event identification threshold. If the observations are provided as categorical yes/no values, then this is indeed all that you can do with it. Because the contingency table only includes four values, some metrics within that grouping are designed specifically as discrimination or reliability metrics.

Subsetting is a powerful methodology for assessing particular features and aspects of the data-model relationship. This is useful not only for space weather applications, where the user might have specific needs for decision making purposes, but also for space physics investigations, where subsetting can reveal the influence of the inclusion or omission of physical processes on a limited and focused region of the data or model values.

3. Metrics equations

Given the above definitions of these metrics categories, each addressing certain aspects of the data-model relationship, we now present equations for metrics in each category within the two major groupings (continuous versus discrete formulations). As stated several times already, this is not a comprehensive list but one that includes the most common metrics found within magnetospheric physics studies, and the ones that could be of most value to magnetospheric physicists.

3.1. Common fit performance metrics

Fit performance metrics use the continuous nature of the data and model values (e.g., Wilks, 2019). There are several common terms that will be used in the definitions below. Model values will be denoted by M, with individual values with the number set listed as M_i. Observational values are given the variable O, with individual data points called out by O_i. The total number of pairs in the data-model set is N.

3.1.1. Accuracy

There are many fit performance metrics for accuracy. The most common accuracy fit performance metric is RMSE:

$$RMSE = \sqrt{\frac{1}{N-d}\sum_{i=1}^{N}(M_i - O_i)^2} \quad (2)$$

Here, d is the degrees of freedom in the model configuration. For a linear fit, $d = 2$. For a physics-based model, d can be hard to obtain. Fortunately, this value usually doesn't matter, because for nearly all magnetospheric physics data-model comparisons, $N \gg d$ so the influence of d on RMSE will be negligible.

The quadrature formulation in RMSE (squaring the differences and then taking the square-root) emphasizes the contribution of the larger differences. This is the same functional form as standard deviation σ_d, so it is directly comparable to that quantity. In fact, that is a way to give meaning to RMSE – comparisons against $\sigma_{d,O}$ of the observations, as well as $\sigma_{d,M}$ for the model values, provides context to this metric. An RMSE value that is smaller than both σ_d values is considered good.

Because of this similarity to s_d, RMSE is sometimes divided by $\sigma_{d,O}$ to create a normalized version, NRMSE. The normalization can, in fact, be anything, including the interquartile range or the full range of either the data or model values. Another way that is similar to the quality check mentioned above is to divide by the smaller of the two standard deviation values, min [$\sigma_{d,O}$, $\sigma_{d,M}$].

Instead of comparing against standard deviation, a common alternative is to forego the square root operation and leave the metric as mean square error:

$$MSE = \frac{1}{N-d}\sum_{i=1}^{N}(M_i - O_i)^2 \quad (3)$$

MSE can be directly compared with the variance of either the observations or the model values (or, even better, both). Note that because the values are left squared, the units are different than the original ones.

Rather than using the quadrature formalism, another often-used accuracy metric is the mean absolute error, MAE:

$$MAE = \frac{1}{N-d}\sum_{i=1}^{N}|M_i - O_i| \quad (4)$$

Like RMSE, MAE also has the same units as the original values. It is often smaller than RMSE, although if the model values are very close to the observed ones, then the reverse could be true.

One of the problems with RMSE, MSE, and MAE is that they favor the larger values within a set. If the data and model ranges only span an order of magnitude, then these metrics are fine. When the values have a larger range of several orders of magnitude, then the use of other metrics designed for measuring error for highly variable number sets are preferred. There are two common accuracy equations for this case, introduced to the space physics community by Morley et al. (2018a) specifically to address the wide-ranging nature of magnetospheric energetic particle flux measurements. The first is the symmetric mean absolute percentage error, SMAPE (Armstrong, 1978), which is a convenient replacement for MAE:

$$SMAPE = 100\frac{1}{N}\sum_{i=1}^{N}\left|\frac{O_i - M_i}{(O_i + M_i)/2}\right| \quad (5)$$

The division makes it a relative error, so an error of a factor of two contributes the same regardless of the magnitude of the values themselves. The average of the pair of values makes it symmetric. There is another version of this formula, MAPE, that simply divides by the model value M_i, which is sometimes used in magnetospheric physics studies but will undercount the relative error when the model is larger than the observation. SMAPE, however, also has the problem of being an average, and the positive definite nature of the absolute error calculation means that the distribution of error is often skewed right, rendering SMAPE susceptible to outliers. Therefore, another metric, median symmetric accuracy, was introduced that preserves the qualities of SMAPE but is based on the logarithm of the values and on the median of the error values rather than their mean:

$$MSA = 100\left(\exp\left[\text{Median}\left(\left|\ln\left(\frac{M_i}{O_i}\right)\right| \forall\, i\right)\right] - 1\right) \tag{6}$$

If the model matches the data perfectly, then the M/O ratio will always be one, the natural log will always yield zero, so the median value is zero and the exponential will yield one, and then the minus one at the end will set MSA to zero.

3.1.2. Bias

The next category, bias, has one metric that has been the dominant selection for data-model comparisons, similar to RMSE within the accuracy category. This standard metric within bias is the mean error:

$$ME = \overline{M} - \overline{O} \tag{7}$$

In (7), the bar over the values indicates the centroid or average of the values. This is usually taken as the arithmetic mean, but it could be median or is sometimes the geometric mean, $\overline{x}_{geom} = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N}$. Note that this is equivalent to finding the arithmetic mean of the logarithm of the values and then applying an exponential function after the subtraction of the means. Note that the ordering of the subtracted values is sometimes reversed, but this reverses the interpretation of ME. As written in (7), a negative ME indicates that the model systematically underpredicts the observations, while a positive value reveals that the model systematically overpredicts the observed data.

A similar issue arises with ME as with RMSE or MAE – when the values span orders of magnitude, ME favors those at the high end of the range. An alternative is to use the harmonic mean, which favors the low end of the range but still suffers from the same weighting issue. Morley et al. (2018a) suggested the use of a different bias metric for highly variable values, called the symmetric signed percentage bias, SSPB:

$$SSPB = 100\left(\text{sign}\left[\text{Median}\left(\ln\left(\frac{M_i}{O_i}\right) \forall\, i\right)\right]\right)\left(\exp\left[\left|\text{Median}\left(\ln\left(\frac{M_i}{O_i}\right) \forall\, i\right)\right|\right] - 1\right) \tag{8}$$

The "sign" in (8) indicates the + and – sign of the value but not the number itself. The last term in SSPB is quite similar to MSA but with one key difference, the absolute value is outside of the median operator within the exponential.

A final fit performance metric in the bias category is the median percentage error, MPE. This is simply the median value of (M-O)/O multiplied by 100. If the model is symmetrically balanced around the observed values, then the median of these differenced values will be zero. Writing it with M-O gives it the same interpretation as ME.

3.1.3. Precision

Precision is, unfortunately, rarely used in magnetospheric physics investigations, but there is one fit performance metric in this category that has received some usage in the CCMC modeling challenge results, modeling yield, YI, which is simply a ratio of the ranges:

$$YI = \frac{max(M) - min(M)}{max(O) - min(O)} \tag{9}$$

The interpretation of this is that values above one indicate that the model overpredicts the spread of the data, while values below one show that the model underpredicts the spread. YI is susceptible to outliers, though, because it only uses the maximum and minimum values from each of the number sets. A more robust measure of precision is based on the standard deviations of the M and O number sets, either their ratio,

$$P_{\sigma,\text{ratio}} = \frac{\sigma_M}{\sigma_O} \tag{10}$$

or their difference,

$$P_{\sigma,\text{diff}} = \sigma_M - \sigma_O \tag{11}$$

The first of these definitions, $P_{\sigma,\text{ratio}}$, has the same interpretation as YI, with values above and below unity indicating that the model over or under predicts the spread of the observations. The second variation, $P_{\sigma,\text{diff}}$, has the same interpretation as ME, with values above and below zero indicating that the model over or under predicts the spread of the data. If using (11), it is useful to compare it with both ME and either RMSE or MAE, as this provides some understanding of the relative contributions of systematic offset versus random spread in contributing to the accuracy.

3.1.4. Association

What if the model values are close to the observations in some part of the range but slowly drift away from the observed values elsewhere in the range? This is an indication that the model captures the up-down trends of data but perhaps is not that accurate at reproducing the exact values. This is quantified by metrics in the association category, and there is one metric that is by far the dominant choice, the Pearson linear correlation coefficient, R:

$$R = \frac{\sum\left(O_i - \overline{O}\right)\left(M_i - \overline{M}\right)}{\sqrt{\sum\left(O_i - \overline{O}\right)^2 \sum\left(M_i - \overline{M}\right)^2}} \tag{12}$$

Note that R is sometimes referred to as r or CC or even PCC. Because the observational and modeling values are never differenced from each other, only against their own mean values, the two number sets could have very different ranges, means, and spreads, but still result in a good correlation. The R metric can range in values between one and minus one; for a data-model comparison, the ideal value of R is one. A value of minus one would indicate a model that was quite bad at predicting the numerical value of the observations but nonetheless has predictive ability.

Correlation coefficient is a metric often interpreted with a p-value, a probability that the two number sets could have achieved that R value by random chance. Tradition has denoted a p-value of 0.05 (or, 5%) as significant. With many data points in the comparison, R might be relatively low but still be "statistically significant." The current statistics trend (e.g., Amrhein et al., 2019a) is to not assign thresholds to significance determinants, because they depend on context/application, i.e., 5% might be good for some problems but 1% or even 0.001% might be needed for others. In fact, the American Statistical Association issued a statement deemphasizing scholarly reliance on p-values (Wasserstein and Lazar, 2016). Regardless of the number of points, a "good" value of R should not only satisfy appropriate statistical significance but also be above some rule of thumb threshold for the problem being investigated. For many magnetospheric applications, rule of thumb values of a good R are 0.5 or even 0.7. The latter is chosen so that R^2, the coefficient of determination, will be at or above 0.5, indicating that more than half of the variance of the data is captured by the model. To reiterate, though, the situational context is critical for interpretation of a result and even these rule of thumb values could be too low or too high for a specific purpose.

One other association metric in the fit performance category is the Spearman rank order correlation coefficient. To calculate this metric,

usually designated R_S or ROCC, the M and O values in (12) are replaced by the sorted rank order of each model value and observed value, respectively. It has the advantage of deemphasizing any outliers and exaggerating the influence of any near-centroid clustering of the values. Because the M and O values themselves are lost and replaced with their rank order, however, the assumed ideal linearity between the observations and model values is also obscured.

3.1.5. Extremes

Fit performance metrics that quantify a model's extreme values (the extremes category) are not often used in magnetospheric physics. There are, however, several methodologies for quantifying extremes. The first is the use of the cumulative probability distribution (CPD). An illustration of the relationship of a number set's probability distribution to its CPD is shown in the top panels of Fig. 3. The CPD is calculated for some value x along a number set by adding together the probability distribution (i.e., its histogram) bins from negative infinity up to x and then dividing by the total number of values in the entire number set, N. As x increases, CPD monotonically increases from 0 to 1. CPD could also be defined by starting at positive infinity and sweeping x towards lower values. Here, x is either the M or O number sets in the data-model comparison. The lower panel of Fig. 3 shows an example of this for the AL index, based on high-latitude magnetometers, comparing observed values (from OMNIWeb) with those from a model (WINDMI, solar wind interaction with the magnetosphere and ionosphere) for the year 2014.

Defining a measure of extremes can be accomplished by determining the value above or below which a particular small percentage of values, ε, is located. For a Gaussian distribution, for example, there are 5% of values below $\overline{x} - 1.96\sigma$, and 5% of values above $\overline{x} + 1.96\sigma$. Neither the data nor the model, however, is guaranteed to be Gaussian, so the specific values for the data and model number sets corresponding to this ε offset from either end of the full range could be determined:

$$CPD_{\Delta,\varepsilon} = M_{\varepsilon} - O_{\varepsilon} \tag{13}$$

$$CPD_{\Delta,1-\varepsilon} = M_{1-\varepsilon} - O_{1-\varepsilon} \tag{14}$$

These metrics evaluate the similarity or difference between these "extreme tail thresholds" of the two number sets. They are interpreted similarly to ME in (7), with values above zero indicating that the model has a higher value for this extreme cutoff and values below zero indicating that the model has a lower value for the cutoff.

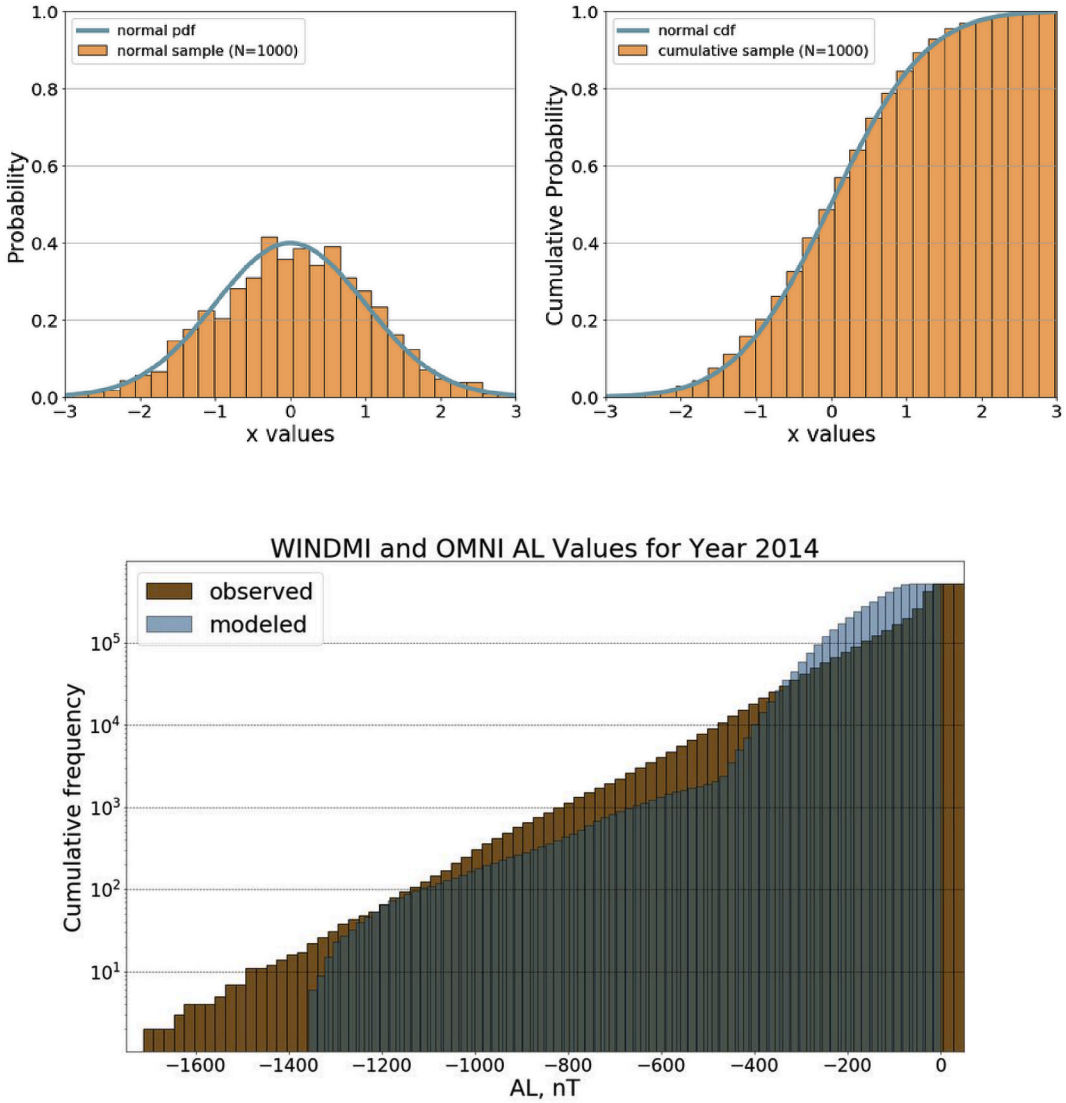

Fig. 3. The top row shows a histogram of values that closely matches a Gaussian probability distribution (shown in blue) and the corresponding CPDs for both the values and the normal distribution. The low panel shows CPD values of the AL index for the year 2014, with the observed CPD in orange and modeled AL values in blue.

Another way to plot the CPD values is the quantile-quantile plot. To create this tool, the data values and the model values are sorted, separately, in ascending order. This makes it similar to the CPD comparisons, but now the two separately-ordered distributions are plotted against each other, resulting in a monotonic line (or series of points). Usually, a unity-slope zero-offset line is also drawn for reference. If the two data sets match, then the points will fall on the diagonal reference line. Deviations from this reveal subsets of the data or model values where the comparison is not particularly good. Note, however, that this comparison has reshuffled the data-model pairings, so it assesses the model's ability to get the distribution of the observed values but not necessarily at the same time or place.

Another method of determining extremes is the comparison of the skew and kurtosis values for the observation and model distributions. Skew, often denoted as γ, involves the cubic differencing of a number set against its mean, relative to its standard deviation:

$$\gamma = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^3}{(N-d)\sigma_x^3} \tag{15}$$

Kurtosis involves the analogous quartic equation:

$$k = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^4}{(N-d)\sigma_x^4} \tag{16}$$

In both (15) and (16), the x represents either the M or O number sets. These values for each of the two number sets can then be compared, usually with a difference as was done in (7) above, for both skew:

$$\gamma_\Delta = \gamma_M - \gamma_O \tag{17}$$

and for kurtosis:

$$k_\Delta = k_M - k_O \tag{18}$$

For either (17) or (18), a value below zero means that the model is underpredicting the skew or kurtosis, respectively, of the observations. A value greater than zero indicates that the model overpredicts this feature of the data.

3.1.6. Skill

Skill is the final category with its own formulas. Any of the equations above can be used in (1) to define a fit performance skill score. The most common in use for magnetospheric physics, by far, is prediction efficiency, PE, based on MSE with the variance of the observations as the reference value (e.g., Murphy, 1988):

$$PE = 1 - \frac{\sum (M_i - O_i)^2}{\sum \left(O_i - \overline{O}\right)^2} \tag{19}$$

Because MSE has an ideal value of zero, if the model values exactly match the observations then PE will be one. A value of PE less than zero indicates that the model is worse than the average of the data at predicting the observations. Other metrics can be used in (1), but the most common are accuracy. Also, while PE uses a data-based reference model (i.e., the observational mean), skill scores can be calculated with a prior model, M^{old}, as the reference model. In this case the prediction efficiency formula, for example, becomes this:

$$SS_{MSE} = 1 - \frac{\sum (M_i - O_i)^2}{\sum \left(M_i^{old} - O_i\right)^2} \tag{20}$$

Rather than comparing PE values between the two models, this directly compares the two models against the data set of interest.

3.1.7. Subsetting: discrimination and reliability

As stated earlier, the subsetting categories, discrimination and reliability, do not have their own formulas in the fit performance grouping. Any of the formulas listed above can be applied in the discrimination or reliability context, subsetting either the data or model values, respectively. Fit performance metrics applied to only part of the full number set are powerful tools that allow for a detailed assessment of the particular ranges within the full data or model range where the model is particularly good (or, not good) at reproducing the observations. This is useful not only in the magnetospheric physics scenario but also in space weather applications. When conducting physics studies, it allows the researcher to focus usage of the model on those applications where it is most appropriate and focus model development on the aspects of the output that are most troublesome. When determining models for use in operational settings, a robust discrimination or reliability assessment allows the user to understand how best to incorporate the model output into their decision-making process.

3.2. Common event detection metrics

Let us now switch over to metrics for all of the categories within the event detection grouping. These are the metrics based on the categorical event status of the data and model, regardless of the actual values. Using event status, we define the contingency table from which many metrics can be calculated. Following Wilks (2019), the quadrants of hits, false alarms, misses, and correct negatives will be referred to as *a, b, c,* and *d*, respectively, in the equations below. Each metric is based on its application to one contingency table. We end this section with a discussion of methods for analysis with varying event state thresholds.

3.2.1. Accuracy

For the accuracy category, the goal of these metrics is to capture the model's ability to correctly identify the event status of the observations. Therefore, metrics in this category assess the "correct" quadrants of hits and correct negatives against the other two values of false alarms and misses. There are two primary metrics that fall within this definition. The first is proportion correct, PC:

$$PC = \frac{a+d}{N} \tag{21}$$

Remember that N is the total number of model-observation pairs in the comparison, so it is the sum of the quadrant values, $N = a+b+c+d$. Note that PC is sometimes multiplied by 100 and renamed percent correct, which unfortunately has the same initials of PC. When using PC, always be clear with the definition (with or without the multiplication by 100) to avoid confusion and misinterpretation. The second metric is the critical success index, CSI:

$$CSI = \frac{a}{a+b+c} \tag{22}$$

Note that CSI is sometimes called the threat score. A third commonly-used measure of accuracy is the F_1 score:

$$F_1 = \frac{2a}{2a+b+c} \tag{23}$$

The F_1 score is very close to CSI, with the same functional form but including an extra emphasis on the hits. The big difference between PC versus CSI or F_1 is the removal of d from both the numerator and denominator. This can be especially useful for the evaluation and interpretation of model quality if the number of correct negatives is large and that value dominates the contingency table. In this case, PC could indicate a very good accuracy for the model even if b and c are larger than a.

3.2.2. Bias

The bias category has one metric that is the regular choice for many studies, the frequency bias, FB:

$$FB = \frac{a+b}{a+c} \tag{24}$$

It is sometimes called the bias ratio. FB assesses the symmetry of the contingency table, with the hits acting as a regulating value in the equation. When the model systematically overpredicts the observations, then it will be in event state more often than the observations, resulting in more false alarms than misses and therefore an FB greater than one. If the model systematically underestimates the observations, then misses will be more numerous than false alarms and FB will be less than one. When the model predicts event state more often than not, then the inclusion of hits in both the numerator and denominator will lead to an FB close to one. Note that this metric could be exactly one even when b and c are larger than a; FB is not measuring accuracy, but rather bias.

3.2.3. Precision

The category of precision is a comparison of the amount of clustering of the model with respect to the amount of clustering in the observations. Since the exact values no longer are meaningful in event detection metrics (only the event status matters), there is no single metric that exactly matches this category in the even detection grouping. Note, however, that there is an event detection metric called precision. This metric is actually a metric within reliability so this discussion is being postponed until metrics for that category have been introduced. More on event detection precision is given below at the end of section 3.2.7.

3.2.4. Association

In the category of association, an evaluation of how well the model matches the observed trends, the metric most often assigned to this category is the odds ratio skill score, (ORSS):

$$ORSS = \frac{\theta - 1}{\theta + 1} = \frac{(a{\cdot}d) - (b{\cdot}c)}{(a{\cdot}d) + (b{\cdot}c)} \tag{25}$$

In (25), θ is the odds ratio, defined as $\theta = ad/bc$. ORSS is preferred over the odds ratio itself because it removes the false alarms times misses term from the denominator, preventing a possible division by zero. ORSS has the same properties as other skill scores, in that the perfect score is one and values below zero are worse than the reference model (here, random chance). If the model is correctly predicting the event status of the observations and $ad \gg bc$, then ORSS will be close to one. If the opposite is true and the model incorrectly predicts the event status more than it gets it right, then ORSS will be less than zero. Because of the inclusion of d in the equation, it should be noted that ORSS can be close to one even when the model is not particularly good at predicting events.

3.2.5. Extremes

This issue where some metrics are overwhelmed by a large number of correct negatives is addressed by the extremes category, which includes metrics that work well when events are rare (e.g., Provost and Fawcett, 2001; Haixiang et al., 2017). The most popular metric in this category is the symmetric extreme dependency score, SEDS:

$$SEDS = \frac{\ln\left(\frac{a+b}{N}\right) + \ln\left(\frac{a+c}{N}\right)}{\ln\left(\frac{a}{N}\right)} - 1 \tag{26}$$

All three natural logarithm operands in (26) are, by definition of N, less than one, rendering the output from the natural log function negative. The numerator terms include b and c in the operands, so these natural logs will be equal to or more negative than the denominator log value, ensuring that the first term on the right-hand side of (26) is always equal to or less than one. If $b = c = 0$, then SEDS will be one. If $a = b = c = d$, then SEDS will be zero. If a, b, and c are small relative to d, then SEDS increases towards 1. If a is smaller than b and c, then SEDS remains close to zero, but if a is the largest of these three, then SEDS approaches one. Rather than simply ignoring correct negatives, SEDS includes it in the denominators of the natural log functions. This accentuates the values when $a \ll d$, making SEDS a keen indicator of model quality when events are rare.

3.2.6. Skill

There are numerous skill scores that have been developed from the contingency table values. There are three that will be mentioned here, simply because of their occasional use in magnetospheric physics studies. The first, and arguably most widely known, is the Heidke skill score, HSS (e.g., Hogan and Mason, 2012):

$$HSS = \frac{2[(a{\cdot}d) - (b{\cdot}c)]}{(a+c)(c+d) + (a+b)(b+d)} \tag{27}$$

HSS measures the fraction of correctly predicted times after eliminating those predictions that would be correct purely from random chance. Note that the numerator is nearly identical to that of ORSS in (25), off by only a factor of two, but the denominator is completely different. Like other skill scores, a perfect score for HSS is one, and values of HSS less than zero mean that the model is worse than the reference model, which in this case is random guessing.

There are two other skill scores to list here. Another often-used skill metric is the Peirce skill score, PSS:

$$PSS = \frac{(a{\cdot}d) - (b{\cdot}c)}{(a+c)(b+d)} \tag{28}$$

PSS is also called the True Skill Statistic and the Hanssen-Kuiper Skill Score. PSS has exactly the same numerator as ORSS. The denominator is simpler than HSS, being the multiplication of the number of times the data are in the event state with the number of times it is not. The other metric in this category that is sometimes used is the Gilbert skill score:

$$GSS = \frac{a - a_{ref}}{a - a_{ref} + b + c} \quad where \quad a_{ref} = \frac{(a+b)(a+c)}{N} \tag{29}$$

GSS has the advantage of minimizing the influence of d because d only appears in the denominator of a_{ref}. When d is relatively large, then a_{ref} goes to zero and GSS asymptotes to CSI.

3.2.7. Subsetting: discrimination and reliability

Unlike fit performance metrics, the discrimination and reliability categories of the event detection grouping have unique equations. This is because subsetting of the data and model number sets is essentially predefined by the event status. This natural breakpoint makes it convenient to use the contingency table values for defining equations specifically for discrimination and reliability. Discrimination subsets the number sets according to the observed events and non-events, and two mutually exclusive metrics naturally arise from the separation of the contingency table into these two halves. The first, based on observations in the event state, is the probability of detection, POD:

$$POD = \frac{a}{a+c} \tag{30}$$

POD is sometimes known as true positive rate, sensitivity, or recall. The second, which uses only the times when the observation was not an event, is the probability of false detection, POFD:

$$POFD = \frac{b}{b+d} \tag{31}$$

Note that they have different ideal values, with a perfect POD being one and a perfect POFD being zero. Equation (31) can be rewritten with an ideal value of one by replacing the b in the numerator with d, a metric known as the true negative rate or specificity. For reliability, the values are split into subsets according to the model values being in or out of event state. Again, two mutually exclusive metrics can be written from the contingency table values. One of these, for model values in the event state, is the false alarm ratio, FAR:

$$FAR = \frac{b}{a+b} \tag{32}$$

with b in the numerator, an ideal value of FAR is zero. This can be rewritten as the positive predictive value (PPV) by substituting the b in

the numerator with *a*. The second, using the values when the model is not in event state, is the miss ratio, MR:

$$MR = \frac{c}{c+d} \tag{33a}$$

The complement of MR is the negative predictive value, which replaces *c* in the numerator of (33) with correct negatives, *d*. There is one more metric in the reliability category, the forecast ratio, FR, which is the ratio of hits to false alarms (Weigel et al., 2006), providing a quick measure of the utility of a model at producing useful predictive information.

It was mentioned in section 3.2.3 that the event detection metrics for precision, the similarities of the spreads of the data and model, needed to wait until after the metrics listings for the subsetting categories. The metric known as precision is the same as PPV, the hits divided by hits plus false alarms. Therefore, event detection metrics in the precision category can be defined as POD and PPV because these assess the spread away from the hits quadrant into the two incorrect quadrants (false alarms, also known as type 1 errors, and misses, also called type 2 errors).

3.2.8. Moving the event identification thresholds

A final topic to mention is that all of the above quantities are defined for a single contingency table with a given event identification threshold for the data and an event identification threshold for the model. These two values could be different. In fact, the data could even be collected as categorical event status values, a unitless yes/no designation. It could be the case, then, that a user will want to specifically set the model's event identification threshold to optimize some aspect of event detection. All of the metrics above can be calculated with a sliding threshold of model event status. Examining the relationship of these metrics with model event threshold provides information on the model's ability to reproduce the observed events.

A special technique has been created that deserves mention, the relative (or receiver) operating characteristic (ROC) curve (e.g., Birdsall, 1973; Cook, 2007). Long used in terrestrial weather prediction (e.g., Mason, 1982), this is a graph of POD on the y axis against POFD on the x axis, with many values of these metrics created by sweeping through the full range of thresholds for the event status of the model. This results in a monotonic curve that moves from (1,1) for low thresholds to (0,0) for high thresholds. The area under the curve (AUC) of the ROC curve has been interpreted as a measure of how well the model captures the physical processes responsible for the observed event status (e.g., Handley and McNeil, 1982; Flasch et al., 2011).

Along these same lines, if the data are continuous, rather than categorical, and the model is trying to exactly reproduce the observed values, then another technique could be applied to the event detection metrics. Specifically, both thresholds, for model and data, can be simultaneously swept from low to high values. In this case, all points start as hits and eventually end up as correct negatives. All of the metrics above can be calculated with this process of sliding both event identification thresholds and a user can examine the results to better understand model quality.

Like the ROC curve, a special curve has been devised for this process of moving both thresholds together, the sliding threshold of observations for numeric evaluation (STONE) curve (Liemohn et al., 2020). Created with the same procedure of plotting POD against POFD, the STONE curve has many of the same features as the ROC curve, but with one key difference. With the ROC curve, the points stay above or below the observed event threshold line because this line doesn't shift. With the STONE curve, this line is sweeping just like the model event threshold, and the points can shift between the quadrants differently in the ROC curve case. Most notably, the STONE curve can have non-monotonicities whereas the ROC curve is always monotonic. These ripples and wiggles in the STONE curve reveal non-Gaussian features of the scatter plot of data-model points, which can be assessed in more detail using other metrics from both the event detection and fit performance groupings.

Note that some definitions of fit performance metrics to include, as part of the definition of this grouping, any and all calculations that use the continuous nature of the model values. In this case, when the model or observational event determination setting is swept through the possible thresholds, event detection metrics could be considered part of the fit performance metrics. Because of this ambiguity, we prefer not to expand the definition of fit performance metrics in this way, keeping the two groupings distinct.

3.3. Summarizing the metrics definitions

Table 1 provides a quick-reference summary of the major categories (listed in the first column, defined in the second column) and the metrics within these categories for the two major groupings (third and fourth columns for the continuous and discrete metrics, respectively). A study that only uses RMSE in the data-model comparison is quantitatively assessing the overall similarity of the values, which is a good start. More can be learned by using additional metrics, especially those from the other categories. If you are evaluating the quality of the model at identifying active times, then RMSE is actually the wrong metric and will reveal very little about the model's ability towards this purpose. It is useful to use several, even many, metrics when assessing the quality of a model, in order to robustly test its capabilities for the desired purpose.

4. Review of recent magnetospheric studies with metrics usage

The common metrics defined above have been used throughout the last several decades in magnetospheric physics studies. As stated in the Introduction, however, metrics were used rather infrequently until the last several years. Here, we provide a comprehensive review this recent usage of metrics for magnetospheric research and magnetospheric space weather applications, focusing on studies published in the last three years only. Also, to keep it tractable, the review is limited to magnetospheric physics metrics usage only, omitting papers that focus on either the thermosphere-ionosphere or the solar-heliosphere regions.

We present this discussion in the same order as in section 3, focusing first on those studies that only use fit performance metrics, followed by those that used only event detection metrics. A third subsection then covers studies that used metrics from both groupings in their analysis. In each subsection, we start from the magnetopause and work inward, covering global magnetospheric properties, and then back outward, covering specific particle populations.

Note that all of the studies mentioned below are "good" because they included at least one metric in their data-model comparison analysis. The discussion below only covers some of the many magnetospheric physics studies that consider data and models together.

Table 1
Summary of metrics within each grouping (last two columns) and category (rows).

Category	Feature Being Assessed	Common Fit Performance Metrics	Common Event Detection Metrics
Accuracy	Overall similarity of values	RMSE, MSE, MAE, SMAPE, MSA	PC, CSI, F1
Bias	Similarity of centroids	ME, SSPB	FB
Precision	Similarity of spreads	YI, P_σ,r, P_σ,d	PPV, POD
Association	Similarity of trends	R, RS	ORSS
Extremes	Reproducing outliers	CPDΔ,ε, CPDΔ,1-ε, γΔ, κΔ	SEDS
Skill	Quality relative to a reference model	PE, SSMSE	HSS, PSS, GSS
Discrimination	Data-range subsets	Any of the above	POD, POFD
Reliability	Model-range subsets	Any of the above	FAR, MR, FR

4.1. Fit performance metrics in magnetospheric physics

As indicated several times above, fit performance metrics are, by far, the more commonly used metrics grouping in magnetospheric physics and, indeed, across all of space physics and weather. Fig. 4 presents a composite of various visualizations of fit performance metrics. More detail about each panel will be given throughout this section.

For the magnetopause, Staples et al. (2020) is the only recent study that systematically used metrics to examine this structure. They chose a single accuracy metric, ΔR_{MP} (modeled minus observed magnetopause distance from Earth's center), for their analysis. They examined the response of an empirical magnetopause model during storm sudden commencements (SSC). It could be argued that they also applied discrimination methods to the analysis, as they separately considered specific types of solar wind structures and then separately considered different times relative to the SSC via superposed epoch analysis. Fig. 4b

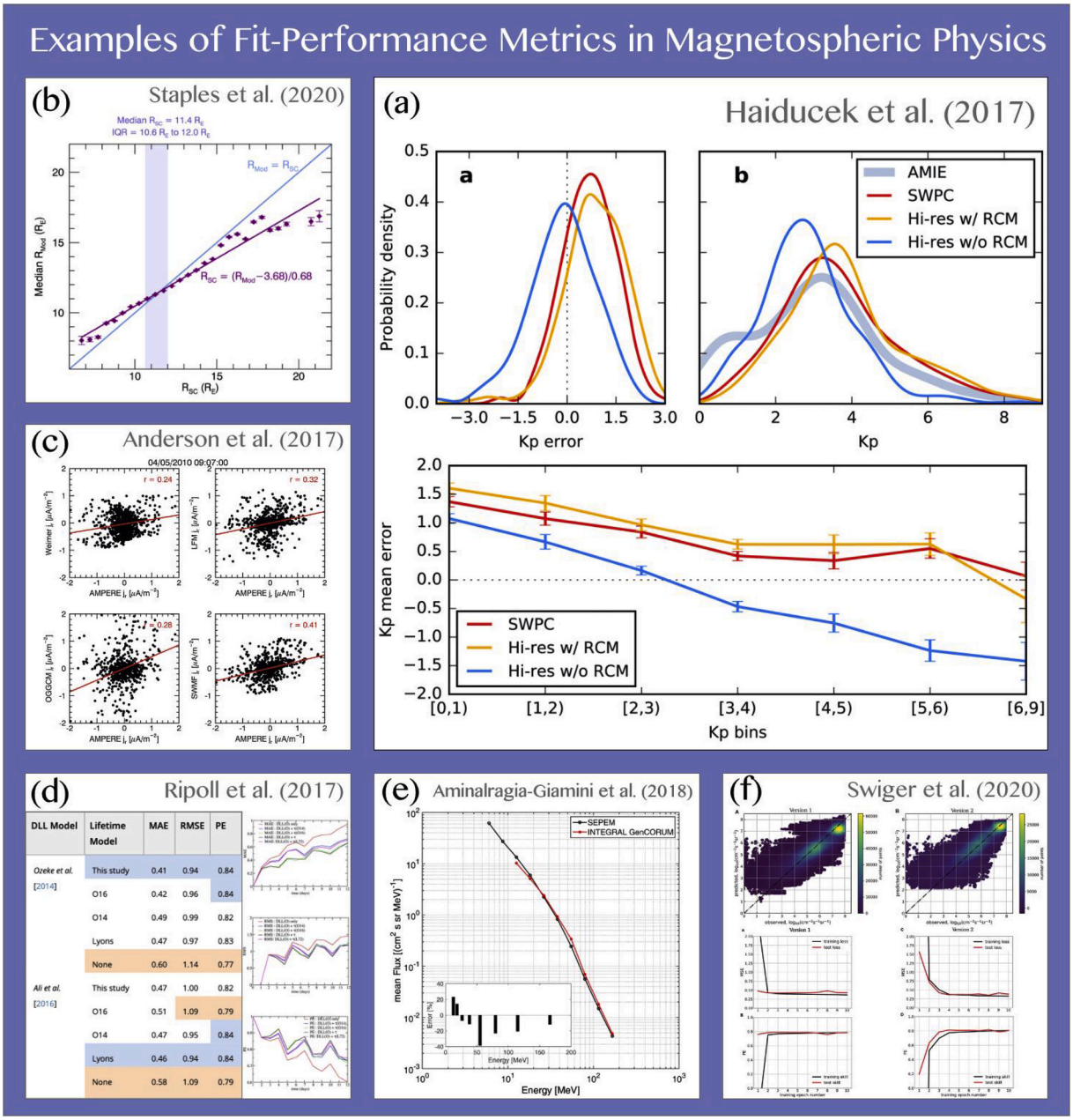

Fig. 4. Proficient examples of fit-performance metric analysis in magnetospheric physics community. (a) From Haiducek et al. (2017). (Upper) Probability density of Kp error and Kp itself for different global model configurations during 1–31 January 2005. (Lower) Mean error for each Kp bin. The ranges for each bin are denoted in the x axis labels in the form [Kpmin, Kpmax). (b) From Staples et al. (2020). Purple diamonds show the median standoff distance calculated by the Shue et al. (1998) model, R_{Mod}, corresponding to a spacecraft magnetopause crossing measured at a given standoff distance, R_{SC}. The error bars show the propagated error of the Shue et al. (1998) model. (c) From Anderson et al. (2017). Scatterplots of Jr from various global and empirical models versus AMPERE Jr together with the linear fit between them, and the linear regression coefficient, R. (d) From Ripoll et al. (2017). (Left) An assessment of the models accuracy through the computation of MAE, RMSE, and PE. (Right) An assessment of the model's accuracy through the evolution during the 12 days of the event (4–16 March) of the three global error indices. (e) From Aminalragia-Giamini et al. (2018). Mean spectra from all fluxes during the studied SPEs from SEPEM and unfolded in this work re-binned to SEPEM energies. The inset shows the percentage error at each energy bin. (f) From Swiger et al. (2020). (Top) Scatter density correlation of modeled vs. observed electron flux at all 17 energy channels for two versions of their model. The comparison for both models was performed on the data reserved for testing. (Bottom) Training and test loss and skill for the two different model configurations. Various panels show the loss and skill metrics after each epoch of training the models.

shows an example plot from this study, presenting the magnetopause standoff distance from the Shue et al. (1998) empirical model against the standoff distance measured by the chosen spacecraft. It shows that there is a particular standoff distance in which the model closely matches the observations. The data and model values are highly correlated though, because the offset between them is linear. This metrics usage was effective for the study they conducted.

Brito and Morley (2017) assessed the ability of five different empirical magnetospheric magnetic field models, presenting an optimization scheme that greatly improved the comparisons with data. They focused on only two metrics, the accuracy metric of RMSE and the bias metric of MPE, performed on a value τ that included both magnitude and angular direction of the local field. These are excellent choices that assess the closeness of the model value to the observed τ and, like the Staples et al. (2020) study, examined subsets of the observations based on geomagnetic activity conditions (in particular, sorted by Kp).

The large-scale magnetic field configuration is associated with magnetospheric currents, and two studies have considered the ability of numerical models to capture observed field-aligned current (FAC) patterns, as measured by low-Earth orbit satellites. Wiltberger et al. (2017) assessed the influence of a new electrojet turbulence model in their global magnetohydrodynamic (MHD) model and associated code suite, comparing the resulting FACs with several data sets. While the comparisons are mostly qualitative, the study mentions RMSE scores against some of the observations. Anderson et al. (2017) examined FAC patterns for several global MHD models as well as empirically-derived FAC models. In addition to many qualitative comparisons presented as scatterplot or line plot figures, they chose to use R as the quantitative assessment of these codes. The scatterplots for a selected time are shown in Fig. 4c. The data values, plotted along the x axis, are the same in each panel, and all of the points are from the same 2D map of FAC values, with the red lines showing the linear fit.

The large-scale electric currents in geospace cause magnetic field perturbations observed by ground-based magnetometer stations, and quite a few studies have included quantitative fit performance metrics calculations when assessing models that seek to predict or explain these perturbations. Starting with individual magnetometer comparisons, Castillo et al. (2017) assessed the ability of an empirical magnetic field model to reproduce midlatitude ground-based perturbations, using R from daily value sets as well as distributions of R for quiet and active days. The time derivative of such perturbations, dB/dt, was considered by Welling et al. (2017), who revisited the Pulkkinen et al. (2013) results as a function of activity level, showing that the underlying conductance models are being extrapolated beyond validity. They quantified this decrease in performance with a quantity they called relative error, which is MAE normalized by the observation value. For both of these, the use of a single metric was enough to make their main point – that the models are only modestly good for geomagnetically active intervals. Bentley et al. (2019) also used only a single fit performance metric, PE, for comparison of their magnetic wave model against ground-based magnetometer data, which was enough to show that the model was better than both the observational variance and 1-h persistence.

The vast majority of model comparisons with ground-based magnetometer data are not with individual station observations but rather with geomagnetic indices. Starting at low latitudes, quite a few models exist that attempt to reproduce the Dst index, or its 1-min counterpart, the SYMH index. Bashkar and Vichare (2019) used RMSE and R in the assessment of their neural network model. Lazzus et al. (2017) also have a neural network model for predicting Dst, for which they routinely rely on RMSE, MAE, and R for model assessment (Lazzus et al., 2019). Chandorkar et al. (2017) use an autoregressive exogenous modeling approach to predict Dst, also using these same three metrics. Similarly, Lethy et al. (2018) have another machine learning algorithm (a neutral network) for computing Dst, and, as above, use RMSE, MAE, and R. While these three metrics provide a well-rounded assessment of the data-model relationship, there is more that can be learned when additional metrics are included in the assessment. Specifically regarding Dst, Engel et al. (2019) conducted a comparison of a new version of their ring current drift physics model against SYMH, using not only these same three metrics but also SMAPE for additional accuracy assessment, ME and SSPB for bias, and two skill scores, PE, and another based on MAE. Moreover, they examined comparisons with observed hot plasma fluxes as well. While they only considered a single storm event, the richness of the metrics usage allowed them to thoroughly assess the quality of the new model version relative to the old one.

Shifting poleward, the Kp index is a commonly-used parameter of perturbation variability, compiled from midlatitude station data. Wintoft et al. (2017) presented a new version of their neural network model for predicting Kp, using RMSE and R for the comparative assessment. Similarly, Sexton et al. (2019) presented a neural network model for Kp, also using RMSE and R as the quantitative comparison. Shprits et al. (2019) also developed a neural network model for Kp, comparing it against persistence and historical Kp values for different forecast lead times. They only used RMSE. With all of these models, the use of only one or two metrics is better than purely qualitative comparisons but it does not explore the details of how the model output is similar to or different from the data values. In contrast, Zhelavskaya et al. (2019) developed yet another neural network model for Kp prediction, using not only RMSE and R but also included a table with the additional metrics of MAE, ME, and PE as a function of hours of lead time for prediction. With this extra information about model performance, they take a detailed look at the Kp features for which their model performs the best, and why.

A few studies have used fit performance metrics to examine several geomagnetic indices simultaneously. Andriyas and Andriyas (2017) used multivariate relevance vector machines to predict 1-h lead times of not only Dst but also the high-latitude indices of AL and PC. Examining 177 storms, they use RMSE as well as PE in their analysis. Gopinath et al. (2018) took a dynamical system approach to magnetospheric physics, developing a model of both Dst and the high-latitutde index of AE, using R as their quantitative assessment tool. Wintoft and Wik (2018) showed a new version of their neural network models for Dst and Kp, using the same two that they usually use, RMSE and R, but now including ME and quantile-quantile plots for bias and extremes comparisons, respectively. An even better fit performance assessment with multiple indices is that of Haiducek et al. (2017), who assessed a month of MHD and associated model output against SYMH, Kp, and AL. Using RMSE, NRMSE, and ME for the basic assessment, they also considered histograms of relative error as a function of subsetted data range, i.e., a discrimination analysis, to reveal that the model is quite good at reproducing quiet and moderate activity but decreases in quality for strong geomagnetic activity. Examples of the Haiducek et al. (2017) metrics visualizations are shown in Fig. 4a, presenting probability distributions and mean error values of the observed and modeled Kp values.

These magnetic field perturbations are registered not only by ground-based magnetometers but also by satellite-based sensors, leading to measurements of plasma waves with periods from milliseconds to minutes. A few wave parameter models have been developed and compared against observations with fit performance metrics. Aryan et al. (2017) have models for chorus waves and plasmaspheric hiss, which they compared against data using RMSE. Saikin et al. (2018) compared wave amplitudes from linear theory with spacecraft data, using relative difference (i.e., normalized MAE) as their quantitative metric.

Like the geomagnetic index modelers, inner magnetospheric plasma and particle modelers have a strong record of using quantitative metrics in their assessments. Starting with the highest energies, i.e., the radiation belt particles, over a dozen studies in the last three years have used fit performance metrics. The majority of these studies use first-principles physics-based models, such as Li et al. (2017), who compared their radial diffusion model of energetic electrons with Van Allen Probes data using difference of the log of the flux values. This same metric was used

by several others, such as Castillo et al. (2019) in their comparison of model output with Van Allen Probes and geosynchronous data and by Zhu et al. (2019) to test the improvement from their new wave diffusion coefficients in their radiation belt model. RMSE is also a common choice for radiation belt modelers, as demonstrated by its use as the quantitative metric in both Woodroffe et al., (2018) and Jordanova et al. (2018). Some first-principles modeling studies use multiple fit performance metrics. Ripoll et al. (2017), for instance, in their examination of the efficacy of many different radial diffusion coefficient settings, conducted comparisons with RMSE, MAE, and PE. Fig. 4d shows plots of these three quantities against time in days from March 4, 2013 for several different model runs. Similarly, Ma et al. (2018) used MSA, median log accuracy ratio, and normalized difference.

A watershed in quantitative fit performance metrics for radiation belt modeling occurred with the publication of Morley et al. (2018a), who introduced the space physics community to several metrics based on the logarithm of the values, in particular SMAPE and SSPB. As discussed above, such metrics are especially useful for highly variable values that span several orders of magnitude, as occurs for energetic electrons in near-Earth space. That study presented an initial usage of these new metrics for their radiation belt model output, demonstrating the effectiveness of these new assessment tools. They immediately came into usage by other groups, with both Glauert et al., 2018 and Yu et al. (2019) adopting SSPB in their studies a few months later, along with other fit performance metrics.

Machine learning, artificial intelligence, and assimilative approaches are common for radiation belt modeling. Developers users of such codes have regularly adopted fit performance metrics into their assessments. Aminalragia-Giamini et al. (2018) used MAPE to quantitatively examine the quality of output from their artificial intelligence model for radiation. Fig. 4e shows the resulting flux from this model against corresponding observations, along with the percent error at each energy in the lower left inset. Most of these studies, however, choose several metrics. Wei et al. (2018), for example, used RMSE, R, and PE to assess their neural network model of geosynchronous high-energy electron flux, and Pires de Lima et al. (2020) used these same three metrics for their neural network – RMSE in the creation of the model and R and PE for testing and validation. Boynton et al. (2019) used a nearly identical set of metrics, except they swapped out RMSE for MSE. Finally, Coleman et al. (2018), in diagnosing the quality of their outer radiation belt nowcasting model, not only used RMSE and PE but also, applying the generic skill score formula in (1), calculated MSE-based skill scores (i.e., like PE) comparing the new model against a previous model and against persistence.

Fit performance metrics have also been used with the hot (~keV) plasma of the inner magnetosphere and plasma sheet. Specifically, Katus et al. (2017) used RMSE to compare their hot ion temperature model, based on energetic neutral atom imagery, against in situ temperature values from the Geotail spacecraft. Wang et al. (2017), in creating their support vector machine for predicting plasma sheet temperatures, used R and NRMSE in training and testing the model. A more substantial usage of these metrics is that of Yu et al. (2019) for their spacecraft surface charging GEM challenge results, in which several models were compared against available data sets, using RMSE for accuracy, YI and SSPB for bias, R for association, and PE for skill. While it was for one value (integrated electron flux from 10 to 50 keV) for one storm event, this robust set of metrics allowed for a more complete examination of the models, including follow-up discussion on the differences between the model assessment values. Swiger et al. (2020) created a neural network model of plasma sheet keV electron fluxes, using a wide range of nine different fit performance metrics – MSE, MAE, MSA, SMAPE, ME, YI, SSPB, R, and PE. They showed that the inclusion of physical understanding in the setup configuration of the neural network yielded better model results, with 7 of the 9 metrics improving even though the training data set was smaller. Fig. 4f is an example of their results for the two model configurations considered, presenting a color heat map of the observed fluxes against the modeled fluxes along with the training and test set MSE and PE scores as a function of training epoch number.

Thermal plasma models of the plasmasphere have also been the subject of fit performance metrics evaluations. He et al. (2017) collected many data sets into a new plasmapause model, then used RMSE and the Spearman rank order RS for comparison with the reserved test data. Several neural networks have been created for plasmapause location and plasmaspheric density, such as Zhelavskaya et al. (2017), who used RMSE, Chu et al. (2017), who use RMSE and R, and Zheng et al. (2019), who used RMSE, normalized RMSE, and R. The topside ionosphere, which is directly connected to and controlled by magnetospheric dynamics, has also been explored in this manner, like Adebiyi et al. (2019), who compared topside densities against other nearby data, using RMSE and R to quantify the relationships. Studies here include the CEDAR challenge results of Shim et al. (2017) and Shim et al. (2018), who used relative difference, RMSE, YI, and R, the assessment of a multifluid transport model by Swoboda et al. (2017), employing RMSE and relative RMSE. There are also the climatological studies of Perlongo et al. (2018), who used NRMSE and PE, and that of Tsagouri et al. (2018), who included a number of metrics – RMSE, ME, R, and mean relative error (another metric for bias).

The unique study of Borovsky& Denton (2018) explored the idea of a composite scalar index for the entire magnetosphere, not just from ground-based magnetometer data but also from in situ satellite values. They only use R for their assessment, but they subset the values according to various solar wind input parameters and solar cycle phases, making it a discrimination assessment as well.

A final study to mention here is that of Rastätter et al. (2019), who presented their new analysis software available at the CCMC. Before, the available tools only examined the model output, and users had to download the digital values to conduct their own data-model comparisons. The new interface includes many built-in comparison tools, including MAPE, R, SSPB, and PE for magnetospheric model evaluation against satellite data. More metrics are regularly being incorporated into these CCMC capabilities with their periodic updates to the online version of the code.

4.2. *Event detection metrics in magnetospheric physics*

Fewer magnetospheric studies use event detection metrics, but there are some in the recent past that focus entirely on this type of comparison for assessing a model against observations. Fig. 5 shows a composite of various plots of event detection metrics. More information about the specific panels of this figure is given throughout this section and the next.

Two recent studies are from governmental space weather forecasting centers. Sharpe and Murray (2017) provide the details of the latest capabilities of the UK Met Office Space Weather Operations Center, including forecast verification analysis. Validation of their geospace storm predictions focused on ROC curves and decile reliability diagrams for capturing “G1 level and above” storm activity, showing assessments for 1, 2, 3, and 4-day forecasts. Similarly, Podladchikova et al. (2018) analyzed 5 years of StormFocus service output, using POD and FAR to show the abilities of their storm prediction models.

The most comprehensive global magnetospheric model assessment using event detection metrics was that of Haiducek et al. (2020), who were focused on substorm identification in long-baseline simulation results. They created a data-based and corresponding model-based “substorm score” from signatures they identified that represent properties of substorm expansion phase onset. Using this list of observed and modeled events, they calculated PC, HSS, POD, and POFD to assess the ability of the model, concluding that it has significant skill at capturing substorm timing. The resulting waiting times from the observations and the model results are shown in Fig. 5b, as a function of the threshold used to identify a substorm event. The metrics essentially quantify how well the curves in the bottom set of panels matches those in the upper set

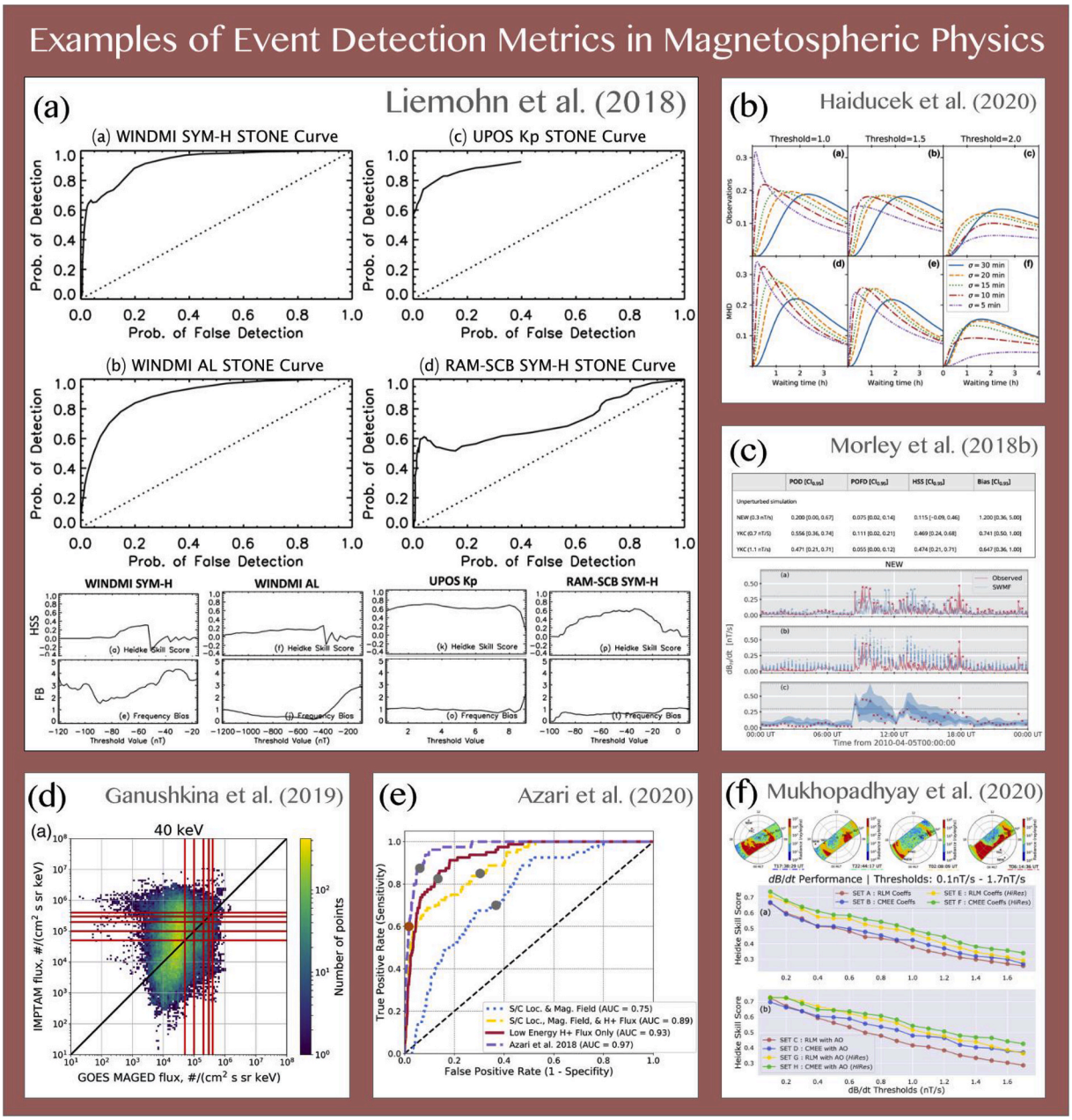

Fig. 5. Proficient examples of event detection metric analysis in magnetospheric physics community. (a) From Liemohn et al. (2018b). (Top) STONE curves for the comparisons of WINDMI SYM-H, WINDMI AL, UPOS Kp, and RAM-SCB SYM-H. Event performance metrics for the comparisons of the aforementioned four datasets. (b) From Haiducek et al. (2020). Distributions of substorm waiting times for a range of identification thresholds and kernel widths used in the identification procedure. (a–c) Observed waiting time distributions. (d–f) MHD waiting time distributions. (a, d) Threshold = 1.0; (b, e) Threshold = 1.5; (c, f) Threshold = 2.0. (c) From Morley et al. (2018b). (Top) Event analysis metrics of dB/dt values for Newport (NEW) and Yellowknife (YKC) stations with different thresholds. (Bottom) Observed and simulated dB_H/dt for the Newport (NEW) magnetic observatory. (d) From Ganushkina et al. (2019). Scatter plots of GOES MAGED electron fluxes versus modeled fluxes by IMPTAM for 40 keV overplotted with population density of the data, together with thresholds used for binary event analysis marked by red lines. (e) From Azari et al. (2020). ROC curves from several logistical regression algorithms as well as from a manually created event identification method. (f) From Mukhopadhyay et al. (2020). Impact of changes to the auroral conductance on dB/dt predictions. (Top) Expansion of the auroral oval as seen through DMSP F16 auroral radiance maps and the magnetometer stations at Yellowknife (YKC) and Newport (NEW). (Bottom) Heidke Skill Score (HSS) performance of SWMF simulation variants with different conductance models at ascending dB/dt predictions for all events listed in Pulkkinen et al. (2013).

of panels.

A few studies have used event detection with geomagnetic indices. Savani et al. (2017) created a model that relates solar wind to Kp and then tested its ability to reproduce high-Kp space weather events. They used PC, CSI, FB, PSS, POD, and FAR in this assessment, comparing against other Kp prediction capabilities for a variety of solar wind input conditions, robustly demonstrating that their new model is quite good at predicting active time intervals. Similarly, Maimaiti et al. (2019) used a wide array of event detection metrics (PC, TPR, MR, PPV, F_1 score, and ROC curves) to validate their neural network model for predicting SML.

A few studies have used event detection metrics with inner magnetospheric magnetic wave power and radiation belt electron flux models. Balasis et al. (2019) created a machine learning routine for wave event recognition, using PC for their quantitative assessment. Capman et al. (2019) divided magnetospheric wave power into three bands for logistic regression modeling of relativistic electrons, using PC, POD, and TNR

(1-POFD) to demonstrate the high quality of the model. Even more recently, Simms and Engebretson (2020) created a recurrent neural network model for relativistic electrons, trained on 5 years of data and tested on 2 other years of observations, exploring the assessment with POD, ROC curves, and its integral value, AUC.

A final study to mention here is that of Azari et al. (2020) regarding energetic ions in Saturn's magnetosphere, who argued that machine learning algorithms greatly benefit from the inclusion of known physical relationships. They used a combination of HSS scores, ROC curves, and probability density curves, including eye-catching explanatory visualizations of the modeling technique continuum and the Saturn space environment. The study showed that their science informed event identification algorithm from Azari et al. (2018) performs better than other more complex logistic regression models, as shown in Fig. 5, and obtains the same performance as a data intensive random forest model.

4.3. Robust metrics usage in magnetospheric physics

The studies mentioned above have focused on one or the other of the two major groupings of metrics, either fit performance or event detection. Several recent magnetospheric studies have employed both approaches to their assessments. Half of these have focused on geomagnetic indices. The neural network model of Tan et al. (2018) for Kp prediction underwent comparisons against observations using RMSE, MAE, and R from the fit performance grouping as well as high-Kp event detection quantification using POD and POFD and a skill value known as the F1 score. A neural network for Dst was developed by Gruet et al. (2018), using not only RMSE, MAE, and R but also POD and POFD as a function of model event identification threshold (i.e., a tabular ROC curve) and a reliability diagram, which plots observed event occurrence rate against the modeled rate as a function of model threshold. Two other papers that considered geomagnetic indices with both metrics groupings are Liemohn et al. (2018a) and Liemohn et al. (2018b). The former examined the real-time nowcasting performance of a global modeling suite against Dst while the latter laid out guidelines for metrics assessments of new index prediction models, including example calculations on three different state-of-the-art codes. The usage of metrics in these two studies spans nearly everything listed in section 3 above. Fig. 5a shows a few of the plots from Liemohn et al. (2018b), displaying both the STONE curves in the upper panels as well as HSS and FB as a function of event identification threshold setting. The threshold settings for which each model is good and those thresholds when the model is less accurate.

Two studies have examined modeling capabilities at reproducing dB/dt observed by ground-based magnetometers. Morley et al. (2018b) conducted 40 global magnetospheric simulations with perturbed solar wind inputs around the baseline observed values, quantifying the impact of this variability on the resulting magnetic field at Earth's surface. They used RMSE, MAE, and ME against SYMH during the selected intervals as well as PC, FB, HSS, POD, and POFD to assess event detection in dB/dt in high-latitude magnetometers. Their study robustly quantified confidence intervals around the baseline metrics values, allowing operational users to better identify true space weather events from a spurious input value. Fig. 5c presents a summary of their key findings, showing the data-model comparison in the lower panels and four different event detection metric scores in the upper panel. Similarly, Mukhopadhyay et al. (2020) tested the space weather event prediction capabilities of their model with an even broader range of fit performance and event detection metrics, allowing them to probe which aspects of the code lead to improved dB/dt event identification as well as identifying additional avenues for model development. A visualization of their results is included in Fig. 5f, showing auroral oval observations in the top panel as context to accompany the HSS values in the lower panels.

Two final studies to mention here are investigations of energetic particles in the inner magnetosphere. With their new radiation belt model, Chen et al. (2019) compared against 3.5 years of Van Allen Probes data using PE along with CSI, FB, POD, and POFD for flux enhancements, subdividing the analysis by radial distance, demonstrating that the model has a potentially useful forecast capability out to one day ahead. In addition, Ganushkina et al. (2019) compared their nowcasting of keV-energy electron fluxes against three energy channels of geosynchronous observations, considering not only MSA, SSPB, and R but also HSS for detecting high-flux events. Fig. 5d is a scatterplot of the model output for 40 keV electron fluxes against the observed values, with the red lines showing event identification thresholds used in the analysis. The extra analysis allows them to better discuss and explain discrepancies, exploring the underlying assumptions of the model configuration.

5. Synthesis: metrics best practices

Some magnetospheric physics and magnetospheric space weather studies conduct data-model comparisons in a purely qualitative manner, supplying two plots next to each other or overplotting observations with the simulation output. While this provides a general impression of the goodness of the model and has led to many new physical insights through the history of magnetospheric physics research, it is not rigorous and does not invoke confidence in the veracity of the code. That is, this might be useful for physics when the field is in an initial phase of discovery on a particular topic, but this is not sufficient for detailed physical analysis or space weather operational decision-making.

When only using RMSE, R, or some other singular metric, the study is greatly improved over a purely qualitative analysis because it now includes a quantitative data-model comparison. The use of only one metric, however, only tests one aspect of the relationship. That is, RMSE only tests the overall accuracy, with an emphasis on outlier data-model differences. Similarly, R only tests the model's ability to reproduce up-down trends of the data, saying nothing about the closeness of the model values to the observations. RMSE alone does not indicate if the model values are systematically or randomly off from the data, only the extent of difference. The use of bias and precision measures provide the additional context. Furthermore, RMSE alone will not reveal a subset range of the data or model where the model might be particularly good or bad. In addition, neither RMSE nor R indicate if the model is good at capturing events in the observations. Evaluating a model or forecasting method using only a single metric and then making improvements to optimize for that single metric alone can lead to improvements in one area at the cost of another important feature. This might be what is needed for that particular usage of the model, but the model should not be considered validated for any other use.

The lesson to learn from this presentation of metrics and exploration of magnetospheric studies that used quantitative metrics in their data-model comparisons is that additional insights can be gleaned when additional metrics are included in the analysis. In a research study, the model is being used to assess a particular hypothesis, so that aspect of the model needs to be validated against observations to show that it is solving the right equations for the problem of interest. The model output could greatly exceed the spatial or complexity scope of the stated problem, but if the model has not been tested for the particular feature of interest, then this needs to be done prior to using it for the proposed analysis.

A similar argument can be used for space weather operations with magnetospheric models. To trust the model results to the point of making decisions based on code output, the aspects of the code most relevant to that decision process need to be thoroughly vetted. This might involve one or two fit performance metrics, like RMSE or R, but most likely it should involve a large variety of metrics considering that aspect of model output from many angles. Such in-depth assessment of model capabilities are necessary to substantiate any subsequent decision making based on the model. Furthermore, validation for one usage does not mean that the model can be applied for other purposes. This is the AUL concept of Halford et al. (2019) – when the user or application of a

model changes, that model reverts to a lower AUL and must be assessed again in a manner that adequately evaluates its capabilities for the new usage.

The above metrics can be applied to any set of model output and corresponding observations. The most intuitive usage is with time-series values, such as a ground-based magnetometer index or particle flux at a given energy. Even for this usage, it should be noted that comparisons with indices have their own concerns, such as those raised by Liemohn et al. (2018b). The metrics could also be applied to spatially distributed values, in one or more dimensions, such as ΔB from many magnetometer stations at a particular time (e.g., Mukhopadhyay et al., 2020), a spatial array of field-aligned currents (e.g., Anderson et al., 2017), or a remotely-sensed image of magnetospheric plasma, such as those of the cold and hot ions (e.g., Burch, 2000; McComas et al., 2009a, 2009b). The multidimensional aspect of data can also extend into velocity space (across an energy-pitch angle grid, for example) or using time in addition to one or more of these other phase space dimensions, such as an energy-time spectrogram of particle fluxes or a latitude-time keogram of auroral emissions. The metrics above can be used with any of these, but multi-dimensional comparisons open the possibility of additional metrics. For example, Liemohn et al. (2006) used several such techniques, including the local time of the peak in energetic neutral atom images and the radial distance of the plasmapause. Even more sophisticated options exist, such as those of Uritsky et al. (2002), who calculated the spatial extent of auroral bright regions and integrated their dynamically-changing area over time. Multidimensional data-model comparison techniques deserve their own comprehensive review.

It is good practice to conduct uncertainty calculations. A value calculated from a metric is difficult to interpret without context, and uncertainties around the base value is one way to provide that. An issue regarding uncertainty is the number of data-model pairs in the calculation. Especially with fine-scale subsetting, there is a risk that the number of points within the subset could become small, therefore the uncertainty of that metric value could become large and obscure the meaning and interpretation of the metric values.

Uncertainties can be placed on all of the metrics above. Morley et al. (2018b) provide one example of this, conducting many model simulations with the inputs perturbed around the baseline values, then using the perturbed-input metrics values to obtain a confidence interval around the metrics values with the baseline inputs. Another way to obtain error bars is the bootstrap method (e.g., Efron and Tibshirani, 1993), in which the data-model pairs are randomly sampled, with replacement, to create a "new" set of data-model pairs, from which metrics can be calculated. Doing this random selection with replacement hundreds of times yields a distribution of metrics values, from which a standard deviation can be calculated and applied to the original metric score. In addition to these two methods, a thorough discussion of uncertainties on event detection metrics values is given by Hogan and Mason (2012).

A final best practice is the use of quantitative tests to compare a metric value against its perfect score or against that metric score from another data-model comparison (either the same model for a different data set or a different model for the same data set). Many such tests exist, and the appropriate test should be used for the hypothesis being assessed. Here, we will list two of the most common tests. The first is the Welch's *t*-test, which assesses the similarity of two values (A_1 and A_2) with unequal sample sizes (n_1 and n_2) and unequal variances (σ_1^2 and σ_2^2):

$$t = \frac{|A_1 - A_2|}{\sigma_{\overline{\Delta}}} \quad where \quad \sigma_{\overline{\Delta}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{33b}$$

This Welch's *t* statistic can be converted into a p-value probability that the two values came from the same population, knowing the degrees of freedom *d*, which is a function of the sample sizes and variances:

$$d_W = \frac{\sigma_{\overline{\Delta}}{}^4}{\frac{\left(\sigma_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(\sigma_2^2/n_2\right)^2}{n_2 - 1}} \tag{34}$$

If the *t* statistic is low (below ~1.5 to 3, depending on d_W) and the probability is high, then this indicates a good chance that the two metric scores, A_1 and A_2, are "sampling the same population" – that is, the metrics values can be considered close enough to be "the same." If the *t* statistic is high and the probability is low, then this implies a statistically significant difference between A_1 and A_2. An example would be if the metric A were *R* and the new version of your model yielded a higher value than your old version. If this difference is deemed statistically significant by a Welch's *t*-test, then you have quantitative evidence that the new version is better with respect to the "similarity of trends" in the model and data values (quoting the association definition from Table 1).

The other assessment of model performance we will mention is the *F* ratio, which is calculated as the mean square regression over mean square error:

$$F = (N - d)\frac{\sum \left(M_i - \overline{O}\right)^2}{\sum \left(M_i - O_i\right)^2} \tag{35}$$

This is a comparison of the error of the model to the observational mean against the error of the model to the individual observations. In (35), *d* is the model degrees of freedom used in many of the fit performance metrics. The *F* ratio can be converted into a p-value probability for statistical significance, which usually requires a value above ~250 (depending on *N* – *d*). If the probability is low, then the model is a good fit to the data. Because it is based on MSE, this is a test of accuracy. That is, the *F* ratio assesses the significance of the "overall similarity of the model to the data.

As stated above, these probabilities for *t* and *F* should be used with caution. Wasserstein et al. (2019) strongly suggest dropping the implied "statistical significance" at the 0.05 threshold as an indicator of hypothesis confirmation or rejection. In an even stronger statement, Hurlbert et al. (2019) recommend journals disallow the use of the terms "statistically significant" and "statistical significance." The reporting of p-values is still encouraged, but now should be used as one of many indicators subject to interpretation. Finally, because magnetospheric physics relies on measurements of the natural environment and thus is not conducive to controlled and repeatable experiments, differing p-values between studies does not imply contradictory results. As Amrhein et al. (2019b) put it, "there is no replication crisis if we don't expect replication." It is recommended to use these significance tests as guides for assessing the quality of a data-model comparison, taking into account all of the other caveats, limitations, and constraints of the measurement techniques and numerical setup.

6. Conclusions

This review explored the historical usage of data-model comparison metrics in magnetospheric physics studies and space weather forecasting of magnetospheric quantities. After a general introduction, the groupings, categories, and types of metrics were presented, listing their strengths with respect to probing some aspect of the data-model relationship. A detailed examination of the recent magnetospheric studies using metrics was presented, followed by a discussion of the main lessons to be learned about limited or comprehensive assessments.

The main points can be summarized as follows:

1. The field of magnetospheric physics adopted robust metrics usage only within the last 10 years or so; prior to this, metrics usage was sporadic at best and most studies included only qualitative comparisons.

2. The advent of GEM challenges greatly increased metrics usage, in particular the involvement of CCMC to orchestrate large-scale comparisons involving many models and data sets.
3. Two metrics groupings exist, called here fit performance and event detection; the former usually focuses the exact values of the data and observations, in particular their difference, while the latter converts the exact values into event status.
4. Metrics can be divided into eight commonly used categories – accuracy, bias, precision, association, extremes, skill, discrimination, and reliability. Fit performance and event detection metrics exist within each category, sometimes many.
5. Within the last 3 years, dozens of magnetospheric space physics and space weather studies have used multiple metrics to assess the data-model relationship for the chosen model usage, demonstrating that the field is growing in its acceptance of metrics usage as a standard practice in this field.
6. Because each metric was designed to test only one limited aspect of the data-model relationship, it is highly advantageous to conduct a robust suite of metrics calculations to validate the usefulness of the model for its specific usage in the new study being conducted.
7. Uncertainties can be calculated on all metrics, which can be particularly helpful for operational usage of magnetospheric models in space weather forecast decision making.

The space weather community has largely embraced the use of metrics in their usage of magnetospheric models in operational settings. With this review, it is advocated to the magnetospheric physics community to also fully adopt metrics usage as a standard practice. Each metric is designed to test only a specific aspect of the data-model relationship and therefore can only yield a limited assessment. The physical insights to be gained from consideration and analysis with additional metrics will yield additional scientific impact from each study. As presented and discussed above, there is a zoo of metrics available for our use. We urge the magnetospheric physics community, and indeed the entire space plasma physics field, to visit the zoo.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments and Data

The authors would like to thank the US government for sponsoring this research, in particular research grants from NASA (NNX17AB87G, NNX16AQ04G, 80NSSC17K0015) and NSF (1663770). This study received partial funding from the European Union Horizon 2020 Research and Innovation Programme under grant agreement 870452 (PAGER). A. Azari's contributions are based on work supported by the NSF Graduate Research Fellowship Program (DGE 1256260), A. Mukhopadhyay's contributions are based on work supported by the NASA Future Investigator fellowship 80NSSC18K1120. B. Swiger's contributions were partially supported by the NASA Future Investigator fellowship number 80NSSC20K1504. Data for Fig. 3 is available at the University of Michigan Deep Blue Data Repository, https://doi.org/10.7302/Z25T3HQC. Figures in section 4 are reused with permission.

References

Adebiyi, S.J., Adeniyi, J.O., Reinisch, B.W., Adebesin, B.O., Ikubanni, S.O., Adimula, I.A., et al., 2019. Variation of digisonde-derived scale height during quiet and disturbed geomagnetic conditions over an African equatorial station. Radio Sci. 54, 552–560. https://doi.org/10.1029/2018RS006762.

Aminalragia-Giamini, S., Papadimitriou, C., Sandberg, I., Tsigkanos, A., Jiggens, P., Evans, H., Rodgers, D., Daglis, I.A., 2018. Artificial intelligence unfolding for space radiation monitor data. Journal of Space Weather and Space Climate 8, A50. https://doi.org/10.1051/swsc/2018041.

Amrhein, V., Greenland, S., McShane, B., 2019a. Scientists rise up against statistical significance. Nature 567, 305–307. https://doi.org/10.1038/d41586-019-00857-9.

Amrhein, Valentin, Trafimow, David, Greenland, Sander, 2019b. Inferential statistics as descriptive statistics: there is No replication crisis if we don't expect replication. Am. Statistician 73, 262–270. https://doi.org/10.1080/00031305.2018.1543137 sup.1.

Anderson, B.J., Korth, H., Welling, D.T., Merkin, V.G., Wiltberger, M.J., Raeder, J., Barnes, R.J., Waters, C.L., Pulkkinen, A.A., Rastaetter, L., 2017. Comparison of predictive estimates of high-latitude electrodynamics with observations of global-scale Birkeland currents. Space Weather 15, 352–373. https://doi.org/10.1002/2016SW001529.

Andriyas, T., Andriyas, S., 2017. Use of multivariate relevance vector machines in forecasting multiple geomagnetic indices. J. Atmos. Sol. Terr. Phys. 154, 21–32. https://doi.org/10.1016/j.jastp.2016.11.002.

Armstrong, S.J., 1978. Long Range Forecasting. Wiley, New York, USA.

Aryan, H., Sibeck, D.G., Kang, S.-B., Balikhin, M.A., Fok, M.-C., Agapitov, O., Komar, C. M., Kanekal, S.G., Nagai, T., 2017. CIMI simulations with newly developed multiparameter chorus and plasmaspheric hiss wave models. J. Geophys. Res. Space Physics 122, 9344–9357. https://doi.org/10.1002/2017JA024159.

Axford, W.I., Hines, C.O., 1961. A unifying theory of high-latitude geophysical phenomena and geomagnetic storms. Can. J. Phys. 39, 1433.

Azari, A., Liemohn, M.W., Jia, X., Thomsen, M.F., Mitchell, D.G., Sergis, N., Rymer, A., Hospodarsky, G., Paranicas, C., Vandegriff, J., 2018. Interchange injections at Saturn: statistical survey of energetic H^+ sudden flux intensifications. Journal of Geophysical Research Space Physics 123, 4692–4711. https://doi.org/10.1002/2018JA025391.

Azari, A.R., Lockhart, J., Liemohn, M.W., Jia, X., 2020. Incorporating physical knowledge into machine learning for planetary space physics. Frontiers in Astronomy and Space Sciences 7, 36. https://doi.org/10.3389/fspas.2020.00036.

Balasis, Georgios, Aminalragia-Giamini, Sigiava, Papadimitriou, Constantinos, Daglis, Ioannis A., Anastasiadis, Anastasios, Haagmans, Roger, 2019. A machine learning approach for automated ULF wave recognition. Journal of Space Weather and Space Climate 9, A13. https://doi.org/10.1051/swsc/2019010.

Bhaskar, Ankush, Vichare, Geeta, 2019. Forecasting of SYMH and ASYH indices for geomagnetic storms of solar cycle 24 including St. Patrick's day, 2015 storm using NARX neural network. Journal of Space Weather an Space Climate 9, A12. https://doi.org/10.1051/swsc/2019007.

Bentley, S.N., Watt, C.E.J., Rae, I.J., Owens, M.J., Murphy, K., Lockwood, M., Sandhu, J. K., 2019. Capturing uncertainty in magnetospheric ultralow frequency wave models. Space Weather 17, 599–618. https://doi.org/10.1029/2018SW002102.

Birdsall, T.G., 1973. The Theory of Signal Detectability: ROC Curves and Their Character. PhD dissertation, Department of Electrical and Computer Engineering, University of Michigan.

Birn, J., Hesse, M., 2001. Geospace Environment Modeling (GEM) magnetic reconnection challenge: resistive tearing, anisotropic pressure and Hall effects. J. Geophys. Res. 106 (A3), 3737–3750. https://doi.org/10.1029/1999JA001001.

Borovsky, J.E., Denton, M.H., 2018. Exploration of a composite index to describe magnetospheric activity: reduction of the magnetospheric state vector to a single scalar. J. Geophys. Res.: Space Physics 123, 7384–7412. https://doi.org/10.1029/2018JA025430.

Boynton, R.J., Amariutei, O.A., Shprits, Y.Y., Balikhin, M.A., 2019. The system science development of local time-dependent 40-keV electron flux models for geostationary orbit. Space Weather 17, 894–906. https://doi.org/10.1029/2018SW002128.

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.

Burch, J.L., 2000. IMAGE mission overview. Space Sci. Rev. 91, 1.

Camporeale, E., 2019. The challenge of machine learning in Space Weather: nowcasting and forecasting. Space Weather 17, 1166–1207. https://doi.org/10.1029/2018SW002061.

Capman, N.S.S., Simms, L.E., Engebretson, M.J., Clilverd, M.A., Rodger, C.J., Reeves, G. D., et al., 2019. Comparison of multiple and logistic regression analyses of relativistic electron flux enhancement at geosynchronous orbit following storms. J. Geophys. Res.: Space Physics 124, 10246–10256. https://doi.org/10.1029/2019JA027132.

Castillo, Angelica M., Shprits, Yuri Y., Ganushkina, Natalia, Alexander, Drozdov, Aseev, Nikita, Wang, Dedong, Dubyagin, Stepan, 2019. Simulations of the inner magnetospheric energetic electrons using the IMPTAM-VERB coupled model. J. Atmos. Sol. Terr. Phys. 191 https://doi.org/10.1016/j.jastp.2019.05.014.

Castillo, Yvelice, Alexandra Pais, Maria, Fernandes, João, Ribeiro, Paulo, Morozova, Anna L., Fernando, J., Pinheiro, G., 2017. Geomagnetic activity at Northern Hemisphere's mid-latitude ground stations: how much can be explained using TS05 model. J. Atmos. Sol. Terr. Phys. 38–53. https://doi.org/10.1016/j.jastp.2017.11.002. Volumes 165–166.

Chandorkar, M., Camporeale, E., Wing, S., 2017. Probabilistic forecasting of the disturbance storm time index: an autoregressive Gaussian process approach. Space Weather 15, 1004–1019. https://doi.org/10.1002/2017SW001627.

Chen, Y., Reeves, G.D., Fu, X., Henderson, M., 2019. PreMevE: new predictive model for megaelectron-volt electrons inside Earth's outer radiation belt. Space Weather 17, 438–454. https://doi.org/10.1029/2018SW002095.

Chu, X., et al., 2017. A neural network model of three-dimensional dynamic electron density in the inner magnetosphere. J. Geophys. Res. Space Physics 122, 9183–9197. https://doi.org/10.1002/2017JA024464.

Cid, C., Saiz, E., Guerrero, A., Palacios, J., Cerrato, Y., 2015. A Carrington-like geomagnetic storm observed in the 21st century. Journal of Space Weather and

Space Climate 5, A16. https://doi.org/10.1051/swsc/2015017 doi: 10.1051/swsc/2015017. Retrieved from.
Coleman, T., McCollough, J.P., Young, S., Rigler, E.J., 2018. Operational nowcasting of electron flux levels in the outer zone of Earth's radiation belt. Space Weather 16, 501–518. https://doi.org/10.1029/2017SW001788.
Cook, N.R., 2007. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 115, 928–935. https://doi.org/10.1161/CIRCULATIONAHA.106.672402.
Damiano, P.A., Chaston, C.C., Hull, A.J., Johnson, J.R., 2018. Electron distributions in kinetic scale field line resonances: a comparison of simulations and observations. Geophys. Res. Lett. 45, 5826–5835. https://doi.org/10.1029/2018GL077748.
Dungey, J.W., 1961. Interplanetary magnetic field and the auroral zones. Phys. Rev. Lett. 6, 47.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap, 436pp. Chapman and Hall, New York.
Engel, M.A., Morley, S.K., Henderson, M.G., Jordanova, V.K., Woodroffe, J.R., Mahfuz, R., 2019. Improved simulations of the inner magnetosphere during high geomagnetic activity with the RAM-SCB model. J. Geophys. Res.: Space Physics 124, 4233–4248. https://doi.org/10.1029/2018JA026260.
Flach, P., Hernández-Orallo, J., Ferri, C., 2011. A coherent interpretation of AUC as a measure of aggregated classification performance. Proceedings of the 28th International Conference on Machine Learning. Bellevue, WA.
Folini, Doris, 2018. Climate, weather, space weather: model development in an operational context. Journal of Space Weather & Space Climate 8, A32. https://doi.org/10.1051/swsc/2018021.
Ganushkina, N. Yu, Liemohn, M.W., Dubyagin, S., 2018. Current systems in the Earth's magnetosphere. Rev. Geophys. 56 (2), 309–332. https://doi.org/10.1002/2017RG000590.
Ganushkina, N.Y., Sillanpää, I., Welling, D.T., Haiducek, J., Liemohn, M., Dubyagin, S., Rodriguez, J.V., 2019. Validation of Inner Magnetosphere Particle Transport and Acceleration Model (IMPTAM) with long-term GOES MAGED measurements of keV electron fluxes at geostationary orbit. Space Weather 17, 687–708. https://doi.org/10.1029/2018SW002028.
Glauert, S.A., Horne, R.B., Meredith, N.P., 2018. A 30-year simulation of the outer electron radiation belt. Space Weather 16, 1498–1522. https://doi.org/10.1029/2018SW001981.
Glocer, A., et al., 2016. Community-wide validation of geospace model local K-index predictions to support model transition to operations. Space Weather 14, 469–480. https://doi.org/10.1002/2016SW001387.
Gopinath, Sumesh, Santhosh Kumar, G., Prince, P.R., 2018. Non-extensive statistical analysis on solar activity dependence of magnetospheric dynamics. J. Atmos. Sol. Terr. Phys. 167, 96–106. https://doi.org/10.1016/j.jastp.2017.11.011.
Gordeev, E., Sergeev, V., Honkonen, I., Kuznetsova, M., Rastätter, L., Palmroth, M., Janhunen, P., Tóth, G., Lyon, J., Wiltberger, M., 2015. Assessing the performance of community-available global MHD models using key system parameters and empirical relationships. Space Weather 13, 868–884. https://doi.org/10.1002/2015SW001307.
Gruet, M.A., Chandorkar, M., Sicard, A., Camporeale, E., 2018. Multiple-hour-ahead forecast of the Dst index using a combination of long short-term memory neural network and Gaussian process. Space Weather 16, 1882–1896. https://doi.org/10.1029/2018SW001898.
Haiducek, J.D., Welling, D.T., Ganushkina, N.Y., Morley, S.K., Ozturk, D.S., 2017. SWMF global magnetosphere simulations of January 2005: geomagnetic indices and cross-polar cap potential. Space Weather 15, 1567–1587. https://doi.org/10.1002/2017SW001695.
Haiducek, J.D., Welling, D.T., Morley, S.K., Ganushkina, N.Y., Chu, X., 2020. Using multiple signatures to improve accuracy of substorm identification. J. Geophys. Res.: Space Physics 125, e2019JA027559. https://doi.org/10.1029/2019JA027559.
Halford, A., Kellerman, A., Garcia-Sage, K., Klenzing, J., Carter, B., McGranaghan, R., Guild, T., Cid, C., Henney, C., Ganushkina, N., Burrell, A., Terkildsen, M., Thompson, B.J., Pulkkinen, A., McCollough, J., Murray, S., Leka, K.D., Fung, S., Bingham, S., Walsh, B., Liemohn, M., Bisi, M., Morley, S., Welling, D., 2019. Application Usability Levels: a framework for tracking project product progress. Journal of Space Weather and Space Climate 9, A34. https://doi.org/10.1051/swsc/2019030.
Haixiang, Guo, Li, Yijing, Shang, Jennifer, Gu, Mingyun, Huang, Yuanyue, Gong, Bing, 2017. Learning from class-imbalanced data: review of methods and applications. Expert Syst. Appl. 73, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035.
Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36. https://doi.org/10.1148/radiology.143.1.7063747.
Hogan, R.J., Mason, I.B., 2012. Deterministic forecasts of binary events. In: Jolliffe, I.T., Stephenson, D.B. (Eds.), Forecast Verification: A Practitioner's Guide in Atmospheric Science, second ed. John Wiley, Ltd, Chichester, UK, pp. 31–60. https://doi.org/10.1002/9781119960003.ch3. chap. 3.
He, F., Zhang, X.-X., Lin, R.-L., Fok, M.-C., Katus, R.M., Liemohn, M.W., Gallagher, D.L., Nakano, S., 2017. A new solar wind-driven global dynamic plasmapause model: 2. Model and validation. J. Geophys. Res. Space Physics 122, 7172–7187. https://doi.org/10.1002/2017JA023913.
Honkonen, I., Rastätter, L., Grocott, A., Pulkkinen, A., Palmroth, M., Raeder, J., Ridley, A.J., Wiltberger, M., 2013. On the performance of global magnetohydrodynamic models in the Earth's magnetosphere. Space Weather 11, 313–326. https://doi.org/10.1002/swe.20055.
Hurlbert, Stuart H., Levine, Richard A., Utts, Jessica, 2019. Coup de Grâce for a tough old bull: "statistically significant" expires. Am. Statistician 73, 352–357. https://doi.org/10.1080/00031305.2018.1543616 sup.1.
Jolliffe, I.T., Stephenson, D.B., 2012. Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley-Blackwell, Hoboken, NJ.
Jordanova, V.K., Delzanno, G.L., Henderson, M.G., Godinez, H.C., Jeffery, C.A., Lawrence, E.C., Morley, S.K., Moulton, J.D., Vernon, L.J., Woodroffe, J.R., Brito, T. V., Engel, M.A., Meierbachtol, C.S., Svyatsky, D., Yu, Y., Tóth, G., Welling, D.T., Chen, Y., Haiducek, J., Markidis, S., Albert, J.M., Birn, J., Denton, M.H., Horne, R.B., 2018. Specification of the near-Earth space environment with SHIELDS. J. Atmos. Sol. Terr. Phys. 177, 148–159. https://doi.org/10.1016/j.jastp.2017.11.006.
Kalegaev, Vladimir, Panasyuk, Mikhail, Myagkova, Irina, Shugay, Yulia, Vlasova, Natalia, Barinova, Wera, Beresneva, Evgenia, Bobrovnikov, Sergey, Eremeev, Valery, Dolenko, Sergey, Nazarkov, Ilya, Nguyen, Minh, Prost, Arnaud, 2019. Monitoring, analysis and post-casting of the Earth's particle radiation environment during February 14–March 5, 2014. Journal of Space Weather and Space Climate 9, A29. https://doi.org/10.1051/swsc/2019029.
Katus, R.M., Keesee, A.M., Scime, E., Liemohn, M.W., 2017. Storm time equatorial magnetospheric ion temperature derived from TWINS ENA flux. J. Geophys. Res. Space Physics 122, 3985–3996. https://doi.org/10.1002/2016JA023824.
Kepko, L., 2018. Magnetospheric Constellation: Leveraging Space 2.0 for Big Science. IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, pp. 285–288. https://doi.org/10.1109/IGARSS.2018.8519475.
Kubo, Yuki, 2019. Why do some probabilistic forecasts lack reliability? Journal of Space Weather & Space Climate 9, A17. https://doi.org/10.1051/swsc/2019016.
Kubo, Yûki, Mitsue Den, Ishii, Mamoru, 2017. Verification of operational solar flare forecast: case of regional warning center Japan. Journal of Space Weather & Space Climate 7, A20. https://doi.org/10.1051/swsc/2017018.
Lazzús, J.A., Vega, P., Rojas, P., Salfate, I., 2017. Forecasting the *Dst* index using a swarm-optimized neural network. Space Weather 15, 1068–1089. https://doi.org/10.1002/2017SW001608.
Lazzús, J.A., Vega-Jorquera, P., Palma-Chilla, L., Stepanova, M., Romanova, N.V., 2019. *Dst* index forecast based on ground-level data aided by bio-inspired algorithms. Space Weather 17, 1487–1506. https://doi.org/10.1029/2019SW002215.
Lethy, A., El-Eraki, M.A., Samy, A., Deebes, H.A., 2018. Prediction of the *Dst* index and analysis of its dependence on solar wind parameters using neural network. Space Weather 16, 1277–1290 https://doi.org/10.1029/2018SW001863.
Li, Z., Hudson, M., Patel, M., Wiltberger, M., Boyd, A., Turner, D., 2017. ULF wave analysis and radial diffusion calculation using a global MHD model for the 17 March 2013 and 2015 storms. Journal of Geophysical Research Space Physics 122, 7353–7363. https://doi.org/10.1002/2016JA023846.
Liemohn, M.W., 2006. Introduction to the special section on "results of the national science foundation geospace environment modeling inner magnetosphere/storms assessment challenge. J. Geophys. Res. 111 https://doi.org/10.1029/2006JA011970. A11S01.
Liemohn, M., Ganushkina, N.Y., De Zeeuw, D.L., Rastaetter, L., Kuznetsova, M., Welling, D.T., et al., 2018a. Real-time SWMF at CCMC: assessing the Dst output from continuous operational simulations. Space Weather 16, 1583–1603. https://doi.org/10.1029/2018SW001953.
Liemohn, M.W., McCollough, J.P., Jordanova, V.K., Ngwira, C.M., Morley, S.K., Cid, C., et al., 2018b. Model evaluation guidelines for geomagnetic index predictions. Space Weather 16, 2079–2102. https://doi.org/10.1029/2018SW002067.
Liemohn, M.W., Keesee, A.M., Kepko, L., Moldwin, M.B., 2019. Instigators of future change in magnetospheric physics. In: Solar/Heliosphere 2: Magnetospheres in the Solar System, Accepted 5 June 2019, Manuscript # 2018-Oct-CH-0891.
Liemohn, M.W., Azari, A.R., Ganushkina, N.Y., Rastätter, L., 2020. The STONE curve: a ROC-based model performance assessment tool. Earth and Space Science 7, e2020EA001106. https://doi.org/10.1029/2020EA001106.
Lundstedt, H., Wintoft, P., 1994. Prediction of geomagnetic storms from solar wind data with the use of a neural network. Ann. Geophys. 12 (1), 19–24. https://doi.org/10.1007/s00585-994-0019-2.
Ma, Q., Li, W., Bortnik, J., Thorne, R.M., Chu, X., Ozeke, L.G., et al., 2018. Quantitative evaluation of radial diffusion and local acceleration processes during GEM challenge events. J. Geophys. Res.: Space Physics 123, 1938–1952. https://doi.org/10.1002/2017JA025114.
Mason, I.B., 1982. A model for assessment of weather forecasts. Aust. Meteorol. Mag. 30, 291–303.
McComas, D.J., Allegrini, F., Baldonado, J., Blake, B., Brandt, P.C., Burch, J., Clemmons, J., Crain, W., Delapp, D., DeMajistre, R., Everett, D., Fahr, H., Friesen, L., Funsten, H., Goldstein, J., Gruntman, M., Harbaugh, R., Harper, R., Henkel, H., Holmlund, C., Lay, G., Mabry, D., Mitchell, D., Nass, U., Pollock, C., Pope, S., Reno, M., Ritzau, S., Roelof, E., Scime, E., Sivjee, M., Skoug, R., Sotirelis, T.S., Thomsen, M., Urdiales, C., Valek, P., Viherkanto, K., Weidner, S., Ylikorpi, T., Young, M., Zoennchen, J., 2009a. The two wide-angle imaging neutral-atom spectrometers (TWINS) NASA mission-of-opportunity. Space Sci. Rev. 142 (1–4), 157–231. https://doi.org/10.1007/s11214-008-9467-4.
McComas, D.J., Allegrini, F., Bochsler, P., Bzowski, M., Collier, M., Fahr, H., Fichtner, H., Frisch, P., Funsten, H.O., Fuselier, S.A., Gloeckler, G., Gruntman, M., Izmodenov, V., Knappenberger, P., Lee, M., Livi, S., Mitchell, D., Moebius, E., Moore, T., Pope, S., Reisenfeld, D., Roelof, E., Scherrer, J., Schwadron, N., Tyler, R., Wieser, M., Witte, M., Wurz, P., Zank, G., 2009b. IBEX – interstellar boundary explorer. Space Sci. Rev. 146, 11–33. https://doi.org/10.1007/s11214–009–9499–4.
Morley, S.K., Brito, T.V., Welling, D.T., 2018a. Measures of model performance based on the log accuracy ratio. Space Weather 16, 69–88. https://doi.org/10.1002/2017SW001669.
Morley, S.K., Welling, D.T., Woodroffe, J.R., 2018b. Perturbed input ensemble modeling with the space weather modeling framework. Space Weather 16, 1330–1347. https://doi.org/10.1029/2018SW002000.

Morley, S.K., 2020. Challenges and opportunities in magnetospheric space weather prediction. Space Weather 18, e2018SW002108. https://doi.org/10.1029/2018SW002108.

Mukhopadhyay, A., Welling, D.T., Liemohn, M.W., Ridley, A.J., Chakraborty, S., Anderson, B.J., 2020, 18. Space Weather, e2020SW002551. https://doi.org/10.1029/2020SW002551. Conductance model for extreme events: Impact of auroral conductance on space weather forecasts.

Murphy, A.H., 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. Mon. Weather Rev. 116, 2417–2424. https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.

Murphy, A.H., 1991. Forecast verification: its complexity and dimensionality. Mon. Weather Rev. 119, 1590.

Opgenoorth, Hermann J., Wimmer-Schweingruber, Robert F., Anna, Belehaki, Berghmans, David, Hapgood, Mike, Hesse, Michael, Kauristie, Kirsti, Lester, Mark, Jean, Lilensten, Messerotti, Mauro, Temmer, Manuela, 2019. Assessment and recommendations for a consolidated European approach to space weather – as part of a global space weather effort. Journal of Space Weather and Space Climate 9, A37. https://doi.org/10.1051/swsc/2019033.

Perlongo, N.J., Ridley, A.J., Cnossen, I., Wu, C., 2018. A year-long comparison of GPS TEC and global ionosphere-thermosphere models. J. Geophys. Res.: Space Physics 123, 1410–1428. https://doi.org/10.1002/2017JA024411.

Pires de Lima, R., Chen, Y., Lin, Y., 2020. Forecasting megaelectron-volt electrons inside Earth's outer radiation belt: PreMevE 2.0 based on supervised machine learning algorithms. Space Weather 18, e2019SW002399. https://doi.org/10.1029/2019SW002399.

Podladchikova, Tatiana, Petrukovich, Anatoly, Yermolaev, Yuri, 2018. Geomagnetic storm forecasting service StormFocus: 5 years online. Journal of Space Weather and Space Climate 8, A22. https://doi.org/10.1051/swsc/2018017.

Poedts, Stefaan, Kochanov, Andrey, Lani, Andrea, Scolini, Camilla, Verbeke, Christine, Hosteaux, Skralan, Chané, Emmanuel, Herman, Deconinck, Mihalache, Nicolae, Diet, Fabian, Heynderickx, Daniel, De Keyser, Johan, De Donder, Erwin, Crosby, Norma B., Echim, Marius, Rodriguez, Luciano, Vansintjan, Robbe, Verstringe, Freek, Mampaey, Benjamin, Horne, Richard, Glauert, Sarah, Jiggens, Piers, Keil, Ralf, Glover, Alexi, Grégoire Deprez, Luntama, Juha-Pekka, 2020. The virtual space weather modelling Centre. Journal of Space Weather and Space Climate 10, 14. https://doi.org/10.1051/swsc/2020012.

Provost, F., Fawcett, T., 2001. Robust classification for imprecise environments. Mach. Learn. 42, 203–231. https://doi.org/10.1023/A:1007601015854.

Pulkkinen, A., Rastätter, L., Kuznetsova, M., Hesse, M., Ridley, A., Raeder, J., Singer, H. J., Chulaki, A., 2010. Systematic evaluation of ground and geostationary magnetic field predictions generated by global magnetohydrodynamic models. J. Geophys. Res. 115, A03206. https://doi.org/10.1029/2009JA014537.

Pulkkinen, A., et al., 2011. Geospace environment modeling 2008–2009 challenge: ground magnetic field perturbations. Space Weather 9, S02004. https://doi.org/10.1029/2010SW000600.

Pulkkinen, A., et al., 2013. Community-wide validation of geospace model ground magnetic field perturbation predictions to support model transition to operations. Space Weather 11, 369–385. https://doi.org/10.1002/swe.20056.

Raeder, J., McPherron, R.L., Frank, L.A., Kokubun, S., Lu, G., Mukai, T., Paterson, W.R., Sigwarth, J.B., Singer, H.J., Slavin, J.A., 2001. Global simulation of the geospace environment modeling substorm challenge event. J. Geophys. Res. 106 (A1), 381–395. https://doi.org/10.1029/2000JA000605.

Rastätter, L., Kuznetsova, M., Vapirev, A., Ridley, A., Wiltberger, M., Pulkkinen, A., Hesse, M., Singer, H.J., 2011. Geospace environment modeling 2008–2009 challenge: geosynchronous magnetic field. Space Weather 9, S04005. https://doi.org/10.1029/2010SW000617.

Rastätter, L., Kuznetsova, M.M., Glocer, A., Welling, D., Meng, X., Raeder, J., et al., 2013. Geospace environment modeling 2008–2009 challenge: D_{st} index. Space Weather 11, 187–205. https://doi.org/10.1002/swe.20036.

Rastätter, L., Shim, J.S., Kuznetsova, M.M., Kilcommons, L.M., Knipp, D.J., Codrescu, M., Fuller-Rowell, T., Emery, B., Weimer, D.R., Cosgrove, R., et al., 2016. GEM-CEDAR challenge: Poynting flux at DMSP and modeled Joule heat. Space Weather 14, 113–135. https://doi.org/10.1002/2015SW001238.

Rastätter, L., Wiegand, C., Mullinix, R.E., MacNeice, P.J., 2019. Comprehensive assessment of models and events using library tools (CAMEL) framework: time series comparisons. Space Weather 17, 845–860. https://doi.org/10.1029/2018SW002043.

Reiff, P.H., 1990. The use and misuse of statistics in space physics. J. Geomagn. Geoelectr. 42, 1145–1174. https://doi.org/10.5636/jgg.42.1145.

Ridley, A.J., Hansen, K.C., Tóth, G., De Zeeuw, D.L., Gombosi, T.I., Powell, K.G., 2002. University of Michigan MHD results of the geospace global circulation model metrics challenge. J. Geophys. Res. 107 (A10), 1290. https://doi.org/10.1029/2001JA000253.

Ridley, A.J., De Zeeuw, D.L., Rastätter, L., 2016. Rating global magnetosphere model simulations through statistical data-model comparisons. Space Weather 14, 819–834. https://doi.org/10.1002/2016SW001465.

Ripoll, J.-F., Santolík, O., Reeves, G.D., Kurth, W.S., Denton, M.H., Loridan, V., Thaller, S.A., Kletzing, C.A., Turner, D.L., 2017. Effects of whistler mode hiss waves in March 2013. J. Geophys. Res. Space Physics 122, 7433–7462. https://doi.org/10.1002/2017JA024139.

Saikin, A.A., Jordanova, V.K., Zhang, J.C., Smith, C.W., Spence, H.E., Larsen, B.A., Reeves, G.D., Torbert, R.B., Kletzing, C.A., Zhelavskaya, I.S., Shprits, Y.Y., 2018. Comparing simulated and observed EMIC wave amplitudes using in situ Van Allen Probes' measurements. J. Atmos. Sol. Terr. Phys. 177, 190–201. https://doi.org/10.1016/j.jastp.2018.01.024.

Savani, N.P., Vourlidas, A., Richardson, I.G., Szabo, A., Thompson, B.J., Pulkkinen, A., Mays, M.L., Nieves-Chinchilla, T., Bothmer, V., 2017. Predicting the magnetic vectors within coronal mass ejections arriving at Earth: 2. Geomagnetic response. Space Weather 15, 441–461. https://doi.org/10.1002/2016SW001458.

Sexton, Ernest Scott, Nykyri, Katariina, Ma, Xuanye, 2019. Kp forecasting with a recurrent neural network. Journal of Space Weather and Space Climate 9, A19. https://doi.org/10.1051/swsc/2019020.

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

Shim, J.S., Rastätter, L., Kuznetsova, M., Bilitza, D., Codrescu, M., Coster, A.J., et al., 2017. CEDAR-GEM challenge for systematic assessment of Ionosphere/thermosphere models in predicting TEC during the 2006 December storm event. Space Weather 15, 1238–1256. https://doi.org/10.1002/2017SW001649.

Shim, J.S., Tsagouri, I., Goncharenko, L., Rastaetter, L., Kuznetsova, M., Bilitza, D., et al., 2018. Validation of ionospheric specifications during geomagnetic storms: TEC and foF2 during the 2013 March storm event. Space Weather 16, 1686–1701. https://doi.org/10.1029/2018SW002034.

Shprits, Y.Y., Vasile, R., Zhelavskaya, I.S., 2019. Nowcasting and predicting the *Kp* index using historical values and real-time observations. Space Weather 17, 1219–1229. https://doi.org/10.1029/2018SW002141.

Shue, J.-H., Song, P., Russell, C.T., Steinberg, J.T., Chao, J.K., Zastenker, G., Vaisberg, O. L., Kokubun, S., Singer, H.J., Detman, T.R., Kawano, H., 1998. Magnetopause location under extreme solar wind conditions. J. Geophys. Res. 103 (17), 691–17700. https://doi.org/10.1029/98JA01103.

Simms, L.E., Engebretson, M.J., 2020. Classifier neural network models predict relativistic electron events at geosynchronous orbit better than multiple regression or ARMAX models. J. Geophys. Res.: Space Physics 125, e2019JA027357. https://doi.org/10.1029/2019JA027357.

Siscoe, G., Cooker, N., Clauer, C.R., 2006. *Dst* of the Carrington storm of 1859. Adv. Space Res. 38, 173–179. https://doi.org/10.1016/j.asr.2005.02.102.

Staples, F.A., Rae, I.J., Forsyth, C., Smith, A.R.A., Murphy, K.R., Raymer, K.M., et al., 2020. Do statistical models capture the dynamics of the magnetopause during sudden magnetospheric compressions? J. Geophys. Res.: Space Physics 125, e2019JA027289. https://doi.org/10.1029/2019JA027289.

Swiger, B., Liemohn, M.W., Ganushkina, N., 2020. Improvement of plasma sheet neural network accuracy with inclusion of physical information. Frontiers Astronomy and Space Sciences 7, 42. https://doi.org/10.3389/fspas.2020.00042.

Swoboda, J., Semeter, J., Zettergren, M., Erickson, P.J., 2017. Observability of ionospheric space-time structure with ISR: a simulation study. Radio Sci. 52, 215–234. https://doi.org/10.1002/2016RS006182.

Tan, Y., Hu, Q., Wang, Z., Zhong, Q., 2018. Geomagnetic index *Kp* forecasting with LSTM. Space Weather 16, 406–416. https://doi.org/10.1002/2017SW001764.

Tanaka, T., 2007. Magnetosphere-ionosphere convection as a compound system. Space Sci. Rev. 133, 1–72. https://doi.org/10.1007/s11214-007-9168-4.

Tsagouri, I., Goncharenko, L., Shim, J.S., Belehaki, A., Buresova, D., Kuznetsova, M.M., 2018. Assessment of current capabilities in modeling the ionospheric climatology for space weather applications: foF2 and hmF2. Space Weather 16. https://doi.org/10.1029/2018SW002035, 1930– 1945.

Tsurutani, B.T., Gonzalez, W.D., Lakhina, G.S., Alex, S., 2003. The extreme magnetic storm of 1–2 September 1859. J. Geophys. Res. 108 (A7), 1268. https://doi.org/10.1029/2002JA009504.

Uritsky, V.M., Klimas, A.J., Vassiliadis, D., Chua, D., Parks, G., 2002. Scale-free statistics of spatiotemporal auroral emissions as depicted by POLAR UVI images: dynamic magnetosphere is an avalanching system. J. Geophys. Res. 107 (A12), 1426. https://doi.org/10.1029/2001JA000281.

Van Allen, J.A., Ludwig, G.H., Ray, E.C., McIlwain, C.E., 1958. Observation of high intensity radiation by satellites 1958 Alpha and Gamma. J. Jet Propuls. 28, 588–592.

Wasserstein, Ronald L., Lazar, Nicole A., 2016. The ASA statement on *p*-values: context, process, and purpose. Am. Statistician 70 (2), 129–133. https://doi.org/10.1080/00031305.2016.1154108.

Wasserstein, Ronald L., Schirm, Allen L., Lazar, Nicole A., 2019. Moving to a world beyond "$p < 0.05$". Am. Statistician 73, 1–19. https://doi.org/10.1080/00031305.2019.1583913 sup.1.

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.

Welling, D.T., Anderson, B.J., Crowley, G., Pulkkinen, A.A., Rastätter, L., 2017. Exploring predictive performance: a reanalysis of the geospace model transition challenge. Space Weather 15, 192–203. https://doi.org/10.1002/2016SW001505.

Welling, D.T., Ngwira, C.M., Opgenoorth, H., Haiducek, J.D., Savani, N.P., Morley, S.K., et al., 2018. Recommendations for next-generation ground magnetic perturbation validation. Space Weather 16, 1912–1920. https://doi.org/10.1029/2018SW002064.

Wilks, D.S., 2019. Statistical Methods in the Atmospheric Sciences, fourth ed. Academic Press, Oxford.

Wiltberger, M., 2015. Review of global simulation studies of effect of ionospheric outflow on magnetosphere-ionosphere system dynamics. In: Keiling, A., Jackman, C. M., Delamere, P.A. (Eds.), *Magnetotails In the Solar* System. https://doi.org/10.1002/9781118842324.ch22.

Wiltberger, M., et al., 2017. Effects of electrojet turbulence on a magnetosphere-ionosphere simulation of a geomagnetic storm. Journal of Geophysical Research Space Physics 122, 5008–5027. https://doi.org/10.1002/2016JA023700.

Wintoft, P., Wik, M., 2018. Evaluation of *Kp* and *Dst* predictions using ACE and DSCOVR solar wind data. Space Weather 16. https://doi.org/10.1029/2018SW001994, 1972– 1983.

Wintoft, Peter, Wik, Magnus, Matzka, Jürgen, Shprits, Yuri, 2017. Forecasting Kp from solar wind data: input parameter study using 3-hour averages and 3-hour range values. Journal of Space Weather and Space Climate 7, A29. https://doi.org/10.1051/swsc/2017027.

Woodroffe, J.R., Brito, T.V., Jordanova, V.K., Henderson, M.G., Morley, S.K., Denton, M. H., 2018. Data-optimized source modeling with the backwards liouville test–kinetic method. J. Atmos. Sol. Terr. Phys. 177, 125–130. https://doi.org/10.1016/j.jastp.2017.09.010.

Yu, Y., Jordanova, V.K., Ridley, A.J., Toth, G., Heelis, R., 2017. Effects of electric field methods on modeling the midlatitude ionospheric electrodynamics and inner magnetosphere dynamics. J. Geophys. Res.: Space Physics 122 (5), 5321–5338.

Yu, Y., Rastätter, L., Jordanova, V.K., Zheng, Y., Engel, M., Fok, M.-C., Kuznetsova, M.M., 2019. Initial results from the GEM challenge on the spacecraft surface charging environment. Space Weather 17, 299–312. https://doi.org/10.1029/2018SW002031.

Zhelavskaya, I.S., Shprits, Y.Y., Spasojevic, M., 2017. Empirical modeling of the plasmasphere dynamics using neural networks. J. Geophys. Res.: Space Physics 122 (11). https://doi.org/10.1002/2017JA024406, 227– 11,244.

Zhelavskaya, I.S., Vasile, R., Shprits, Y.Y., Stolle, C., Matzka, J., 2019. Systematic analysis of machine learning and feature selection techniques for prediction of the Kp index. Space Weather 17, 1461–1486. https://doi.org/10.1029/2019SW002271.

Zheng, Y., Ganushkina, N.Y., Jiggens, P., Jun, I., Meier, M., Minow, J.I., et al., 2019. Space radiation and plasma effects on satellites and aviation: quantities and metrics for tracking performance of space weather environment models. Space Weather 17, 1384–1403. https://doi.org/10.1029/2018SW002042.

Zheng, Z., Lei, J., Yue, X., Zhang, X., He, F., 2019. Development of a 3-D plasmapause model with a back-propagation neural network. Space Weather 17, 1689–1703. https://doi.org/10.1029/2019SW002360.

Zhu, Hui, Shprits, Yuri Y., Spasojevic, M., Drozdov, Alexander Y., 2019. New hiss and chorus waves diffusion coefficient parameterizations from the Van Allen Probes and their effect on long-term relativistic electron radiation-belt VERB simulations. J. Atmos. Sol. Terr. Phys. 193, 105090. https://doi.org/10.1016/j.jastp.2019.105090.