# The STONE curve: A ROC-derived model performance assessment tool

Michael W. Liemohn[1,1], Abigail R. Azari[1,1], Natalia Yu Ganushkina[2,2], and Lutz Rastatter[3,3]

[1]University of Michigan-Ann Arbor
[2]Finnish Meteorological Institute
[3]NASA GSFC

November 30, 2022

## Abstract

A new model validation and performance assessment tool is introduced, the sliding threshold of observation for numeric evaluation (STONE) curve. It is based on the relative operating characteristic (ROC) curve technique, but instead of sorting all observations in a categorical classification, the STONE tool uses the continuous nature of the observations. Rather than defining events in the observations and then sliding the threshold only in the classifier/model data set, the threshold is changed simultaneously for both the observational and model values, with the same threshold value for both data and model. This is only possible if the observations are continuous and the model output is in the same units and scale as the observations; the model is trying to exactly reproduce the data. The STONE curve has several similarities with the ROC curve – plotting probability of detection against probability of false detection, ranging from the (1,1) corner for low thresholds to the (0,0) corner for high thresholds, and values above the zero-intercept unity-slope line indicating better than random predictive ability. The main difference is that the STONE curve can be nonmonotonic, doubling back in both the x and y directions. These ripples reveal asymmetries in the data-model value pairs. This new technique is applied to modeling output of a common geomagnetic activity index as well as energetic electron fluxes in the Earth's inner magnetosphere. It is not limited to space physics applications but can be used for any scientific or engineering field where numerical models are used to reproduce observations.

1

1 **The STONE curve: A ROC-derived model performance assessment tool**
2

3 **Michael W. Liemohn,[1] Abigail R. Azari,[1] Natalia Yu. Ganushkina,[1,2] and Lutz Rastätter[3]**

4 [1]Department of Climate and Space Sciences and Engineering, University of Michigan, Ann
5 Arbor, MI.

6 [2]Finnish Meteorological Institute, Helsinki, Finland

7 [3]Community Coordinated Modeling Center, NASA Goddard Space Flight Center, Greenbelt,
8 MD

9

10 Corresponding author: Michael Liemohn (liemohn@umich.edu)
11

14 **Key Points:**

15    • A new event-detection-based metric for model performance appraisal is given with
16      sliding thresholds in both observational and model values

17    • The new metric is like the relative operating characteristic curve but uses continuous
18      observational values, not just categorical status

19    • The new metric is used on real-time model predictions of common geomagnetic activity
20      parameters, demonstrating its features and strengths
21

22 **AGU Index Terms:**

23    • 1984        Statistical methods: Descriptive (4318)

24    • 4318        Statistical analysis (1984, 1986)

25    • 7924        Forecasting (1922, 2722, 4315)

26    • 0550        Model verification and validation

27    • 9820        Techniques applicable in three or more fields

28

29 **Keywords:**

30 ROC curve, STONE curve, data-model comparison, model validation, forecasting
31

32    **Abstract**

33    A new model validation and performance assessment tool is introduced, the sliding threshold of
34    observation for numeric evaluation (STONE) curve. It is based on the relative operating
35    characteristic (ROC) curve technique, but instead of sorting all observations in a categorical
36    classification, the STONE tool uses the continuous nature of the observations. Rather than
37    defining events in the observations and then sliding the threshold only in the classifier/model
38    data set, the threshold is changed simultaneously for both the observational and model values,
39    with the same threshold value for both data and model. This is only possible if the observations
40    are continuous and the model output is in the same units and scale as the observations, i.e., the
41    model is trying to exactly reproduce the data. The STONE curve has several similarities with the
42    ROC curve – plotting probability of detection against probability of false detection, ranging from
43    the (1,1) corner for low thresholds to the (0,0) corner for high thresholds, and values above the
44    zero-intercept unity-slope line indicating better than random predictive ability. The main
45    difference is that the STONE curve can be nonmonotonic, doubling back in both the x and y
46    directions. These ripples reveal asymmetries in the data-model value pairs. This new technique is
47    applied to modeling output of a common geomagnetic activity index as well as energetic electron
48    fluxes in the Earth's inner magnetosphere. It is not limited to space physics applications but can
49    be used for any scientific or engineering field where numerical models are used to reproduce
50    observations.

51

52    **Plain Language Summary**

53    Scientists often try to reproduce observations with a model, helping them explain the
54    observations by adjusting known and controllable features within the model. They then use a
55    large variety of metrics for assessing the ability of a model to reproduce the observations. One
56    such metric is called the relative operating characteristic (ROC) curve, a tool that assesses a
57    model's ability to predict events within the data. The ROC curve is made by sliding the event-
58    definition threshold in the model output, calculating certain metrics and making a graph of the
59    results. Here, a new model assessment tool is introduced, called the sliding threshold of
60    observation for numeric evaluation (STONE) curve. The STONE curve is created by sliding the
61    event definition threshold not only for the model output but also simultaneously for the data
62    values. This is applicable when the model output is trying to reproduce the exact values of a
63    particular data set. While the ROC curve is still a highly valuable tool for optimizing the
64    prediction of known and pre-classified events, it is argued here that the STONE curve is better
65    for assessing model prediction of a continuous-valued data set.
66

67

## 1. Introduction

68

69 Numerical models are a fundamental feature of research in the natural sciences. Models
70 are often used to explain strange and interesting features in an archival data set in order to assess
71 the physical processes responsible for that observational signature. They are also used for
72 prediction, using some estimate of future initial and boundary conditions to determine the state
73 of the system, or even a particular observational quantity, ahead of time. These are typical uses
74 of models in every discipline of Earth and space sciences.

75 There exists a large collection of metrics to assess the goodness of fit for these models to
76 a particular data set. These metrics, for the most part, can be sorted into several major groupings,
77 two of which are fit performance metrics and event detection metrics (e.g., Wilks, 2011; Joliffe
78 and Stephenson et al., 2012; Liemohn, McCollough, et al., 2018). The former group, also called
79 continuous metrics, is based on differencing each data-model value pair and includes many well-
80 known assessment equations such as root mean square error, correlation coefficient, mean error,
81 and prediction efficiency (e.g., Hogan and Mason, 2012; Morley et al., 2018). The second group,
82 also called categorical metrics, is based on categorizing the observations into events and non-
83 events and then assessing a model's ability to reproduce this classification. This is done through
84 a contingency table (also commonly called a confusion matrix) in which each data-model pair
85 gets two designations: determining if the observation is in the event state or not and similarly if
86 the model value is in the event state or not. The similarity or difference of the data and model
87 values is irrelevant, only the event/non-event designation matters. This second group includes
88 other well-known assessment equations such as the probability of detection, false alarm rate,
89 frequency bias, and Heidke skill score (see, e.g., Muller et al., 1944; Wilks, 2011).

90 A feature of the event detection metrics is that the model does not have to cover the same
91 range or even have the same units as the observations. The model could be anything that might
92 predict the event state of the observations. Furthermore, the observations do not have to be a
93 continuous-valued real number set, but could be pre-categorized into events and non-events (or a
94 multi-level classification). The model could be a continuous-valued real number set or a discrete-
95 valued categorized set. When the data or model happens to be a continuous-valued real number
96 set, then a threshold value for event identification is chosen, a threshold value that could be
97 different between the observational events and the modeled events.

98 An event detection metric that is often used for weather prediction (e.g., Mason, 1982),
99 psychology (e.g., Swets, 1972), medical clinical trials (e.g., Ekelund, 2011), and machine
100 learning (e.g., Fawcett et al., 2006) is the relative (or, originally, receiver) operating
101 characteristic (ROC) curve (see review by Carter et al., 2016). This is an assessment tool that can
102 be applied when the model values are continuous-valued real numbers, using not just one event
103 identification threshold but many. The method is to sweep the event definition threshold for the
104 model values from low to high, calculating two specific metrics, the probability of detection
105 (POD) and the probability of false detection (POFD), and plotting these two arrays against each
106 other. The threshold that yields the location on the ROC curve closest to the upper left corner
107 (high POD and low POFD) can be considered a possible "best setting" for event prediction by
108 this model. This is not the only location for an optimum pick of a final threshold along an ROC
109 curve. Often the final choice will depend on the application and problem specific details. For
110 example, recent developments have discussed the use of skill scores for different solar and space

111 physics applications (e.g., Bobra & Couvidat 2015) and their location on ROC diagrams (e.g.,
112 Manzato, 2007; Azari et al., 2018). A further detailed discussion on skill scores and their relation
113 to ROC diagrams can be found within Manzato (2005). An integral quantity sometimes used
114 from the ROC curve is the area under the curve (AUC), which is an overall measure of goodness
115 of fit for the model to the observational events across all of the possible model value event
116 identification thresholds.

117 　　　The ROC curve has recently been used quite often in the Earth and space sciences to
118 assess model performance at detecting events in an observational data set. It is used regularly in
119 the atmospheric sciences, such as for regional ozone ensemble forecasting (e.g., Delle Monache
120 et al., 2006), deciphering the microphysical properties of clouds (e.g., Gabriel et al., 2009), and
121 forecasting summer monsoons over India (e.g., Borah et al., 2013). Earth scientists also employ
122 the ROC curve for a diverse set of modeling activities, including the distribution of rock glaciers
123 (e.g., Brenning et al., 2007), assessing triggering mechanisms of earthquake aftershocks (e.g.,
124 Meade et al., 2017), and snow slab instability physics (e.g., Reuter & Schweizer, 2018). This
125 also includes land-air interactions, such as mapping of expected ash cloud locations after
126 eruptions (e.g., Stefanescu et al., 2014), modeling rainfall-induced landslides (e.g.,
127 Anagnostopoulos et al., 2015), and statistically forecasting extreme corn losses in the eastern
128 United States (Mathieu & Aires, 2018). The fields of space and planetary science have also
129 started to employ this technique, such as for oblique ionogram retieval algorithm assessment
130 (Ippolito et al., 2016), identifying energetic particle flux injections at Saturn (e.g., Azari et al.,
131 2018), magnetic activity prediction (e.g., Liemohn, McCollough, et al., 2018), and identifying
132 solar flare precursors (e.g., Chen et al., 2019). In short, the ROC curve has become an essential
133 tool, among many that can and should be applied, for model assessment across many natural
134 science disciplines.

135 　　　The ROC curve, however, only assesses the model's ability to predict a single
136 observational event identification threshold. While this is desirable if the data were pre-classified
137 as events or non-events, this imposes a simplification of the data set when the observations are
138 also continuous-valued real numbers. That is, the ROC curve does not test the model's ability to
139 predict events across the full range of the data. A family of ROC curves can be produced using
140 different data-value event identification thresholds (and sweeping the model-value event
141 identification threshold to produce each ROC curve), which is acceptable if the model is only
142 being used to maximize the prediction of events. If the model, however, is trying to reproduce
143 the exact values of the observations, then it is useful to conduct an assessment for which the data
144 and model have the same threshold setting. The ROC curve, unfortunately, cannot easily test the
145 model's ability to reproduce the observed events at the same threshold setting, sweeping through
146 all possible event identification thresholds.

147 　　　There exists a need for a new metric. Like the ROC curve, this new metric should test a
148 model's ability to predict observed events across the full range of possible model-value event
149 identification settings, but rather than using a single observational event categorization, it should
150 sweep through the same range of event identification thresholds as used for the model. Such a
151 metric is proposed below, called the sliding threshold of observation numeric evaluation, or
152 STONE, curve. This is based on the ROC curve but includes the desirable features described
153 above. The work then presents an application of the STONE curve to two space physics data
154 sets, the prediction of a geomagnetic activity index and energetic electron fluxes in near-Earth

155    space. Similarities and differences between the ROC and STONE curves are discussed, as well
156    as the interpretive meaning of features in the STONE curve.

**2. Method of Calculation**

158        The calculation of a STONE curve is rather similar to that of a ROC curve, with one
159    major exception – both thresholds slide together, incrementing the two event identification
160    thresholds simultaneously so that the same threshold value is used for both the data and the
161    model at each setting from low to high across the range. Because this tool is for continuous-
162    valued observations and model results, for which an "event" is an arbitrary designation, there
163    does not have to be a pre-defined event threshold in the observations. In fact, it is desired that the
164    model match the observations for all levels of "event" definition. Therefore, in the STONE tool,
165    the two thresholds move together. This is illustrated in Figure 1, showing an arbitrary data set
166    plotted against a model output that is trying to reproduce these values.
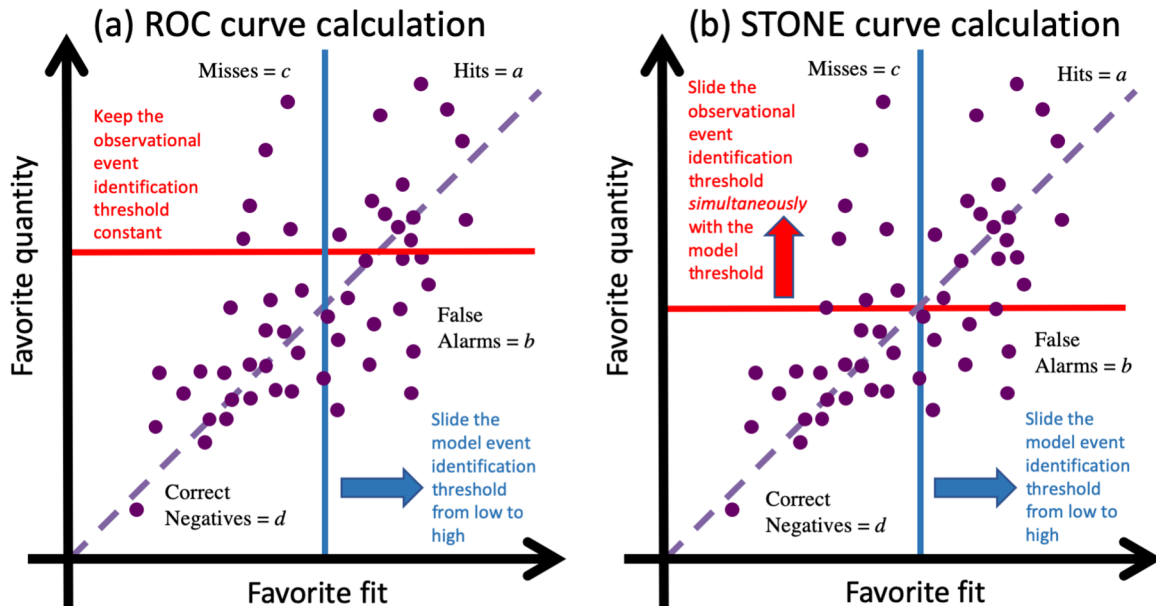
167



**Figure 1.** Idealized examples of how to calculate (a) the ROC curve and (b) the STONE curve. In (a), only the blue curve shifts while the red curve is fixed at some level. In (b), both the red and blue thresholds shift together. As these lines shift, data points are converted from one quadrant to another. The purple dashed curve is the zero-intercept unity-slope line, for reference.

173

174        Figure 1a shows the calculation scenario for the ROC curve, with the event identification
175    threshold for the observations set to a fixed value and the threshold for the model results
176    sweeping from low to high values. Annotations label the four quadrants of the chart, as defined
177    by these two thresholds. As the model threshold changes, the points in the chart change quadrant.
178    Specifically, two shifts occur: points in the "hits" quadrant (variable $a$) move to the "misses"
179    quadrant ($c$) and points in the "false alarms" quadrant ($b$) move to the "correct negatives"
180    quadrant ($d$).

181        The ROC curve is defined from two metrics in the "discrimination" category (Murphy &
182   Winkler, 1987) of data-model comparison techniques. Discrimination metrics are assessments
183   that only use a portion of the data values within a specified range (and the corresponding model
184   values). For event detection metrics, the usual practice is to use the event state of the
185   observations to define the subsets of the data. In particular, the ROC curve uses POD and POFD,
186   which have the following formulas:

187
$$POD = \frac{a}{a+c} \tag{1}$$

188
$$POFD = \frac{b}{b+d} \tag{2}$$

189   Where $a$, $b$, $c$, and $d$ are point counts from the quadrants in the scatter plot. It is seen that these
190   two formulas are mutually exclusive, POD only uses the hits and misses quadrants while POFD
191   only uses the false alarms and correct negatives quadrants. Because the data threshold remains
192   fixed for the ROC curve, the points either contribute to POD or POFD, regardless of the model
193   threshold designation. For a very low model threshold setting, all of the points are in either the
194   hits or false alarms quadrants, which sets both POD and POFD to one. As the model threshold is
195   increased, points are converted from hits to misses and from false alarms to correct negatives,
196   which monotonically decreases POD and POFD. For a very high model threshold, all of the
197   points will then be misses or correct negatives, and both POD and POFD will be zero.

198        Figure 1b shows the calculation scenario for the STONE curve. In this situation, both
199   event identification thresholds move simultaneously. The four quadrants are still defined as with
200   the ROC curve, but with both thresholds changing, the shift of points from one quadrant to
201   another is not so simple. For a very low threshold setting, nearly all points will be hits and
202   perhaps a few will be false alarms. Thus, like the ROC curve, the STONE curve also begins in
203   the (1,1) corner of POFD-POD space (assuming a "low" starting threshold value). Also similarly,
204   for a very large threshold setting, nearly all points will be correct negatives and perhaps a few
205   will be misses, with the STONE curve ending in the (0,0) corner of POFD-POD space. Another
206   similarity is that false alarms are converted into correct negatives as the threshold setting
207   increases.

208        The big difference between the ROC and STONE curve calculations, however, is that as
209   the event identification threshold increases, a hit event can shift to any of the other three
210   quadrants. If it is far above the data threshold but close to the model threshold, then the threshold
211   increase will cause the point to shift from being a hit to a miss. If it is close to the data threshold
212   but far away from the model threshold, then it will shift from being a hit to being a false alarm. If
213   it is close to both thresholds, then there is a chance it will cross both lines during the incremental
214   shift and jump from the hits regions to the correct negatives zone. Only the first of these three
215   moves (hits to misses) occurs with the ROC curve calculation. In addition, misses are shifting to
216   become correct negatives as the observational threshold is incremented to higher values, another
217   move that is not part of the ROC curve calculation. The behavior of the POD and POFD values
218   as a function of threshold, therefore, are not intuitively known and the STONE curve does not
219   have to be monotonic between its (1,1) and (0,0) endpoints.

## 3. Application of the STONE tool

220

221     With this definition for the STONE curve, it can be used on a few example data-model
222   comparisons to illustrate the similarities and differences with the ROC curve. Here, two
223   comparisons will be shown. The first is for a model prediction of a geomagnetic activity index,
224   originally presented by Liemohn, Ganushkina, et al. (2018), and the second is for energetic
225   electrons in near-Earth space, originally presented by Ganushkina et al. (2019).

226     ### 3.1. Predicting a geomagnetic activity index

227     Liemohn, Ganushkina, et al. (2018) compared the output from experimental real-time
228   simulations of the Space Weather Modeling Framework (SWMF) against the disturbance storm-
229   time index, Dst (Rostoker et al., 1972). The SWMF is a collection of space physics numerical
230   models simulating the Sun-Earth space environment (Toth et al., 2012), and in many other
231   planetary environments (e.g., Jia et al., 2012; Ma et al., 2013; Dong et al., 2014; Liemohn et al.,
232   2017). This geospace environment simulation has a very similar setup to that of Pulkkinen et al.
233   (2013), using the Block Adaptive Tree Roe-type Upwind Scheme (BATS-R-US)
234   magnetohydrodynamic model coupled to the Rice Convection Model (RCM) and the Ridley
235   Ionosphere Model (RIM). Real-time solar wind and interplanetary magnetic field input was
236   taken from the Advanced Composition Explorer (ACE) satellite. The simulated Dst time series
237   from the SWMF was calculated with the method from Yu et al. (2010) and compared against the
238   real-time version of the Dst index as produced by the Kyoto World Data Center for
239   Geomagnetism. The interval of comparison spans from 19 April 2015 until 17 July 2017, which
240   is 27 months of 1-hour resolution measurements and corresponding model output values (just
241   under 300,000 data-model pairs).

242     Figure 2a shows a scatter plot of the SWMF Dst values against the observed Dst values.
243   While the individual points are analyzed as unique contributions, they are binned to produce the
244   colored curves on the plot, demarking contours of 50 points within a 5-by-5 nT grid. Note that,
245   because Dst is near zero for quiet times and shifts to negative values during storm times, events
246   are defined as values below (i.e., more negative) a chosen threshold. As defined by Gonzalez et
247   al. (1994), a typical designation for the Dst index measuring a storm situation is -30 nT or below
248   for a weak storm and -50 nT or below for a moderate storm, so these two settings are used for the
249   ROC curve observational threshold setting. These two thresholds are indicated in Figure 2a as
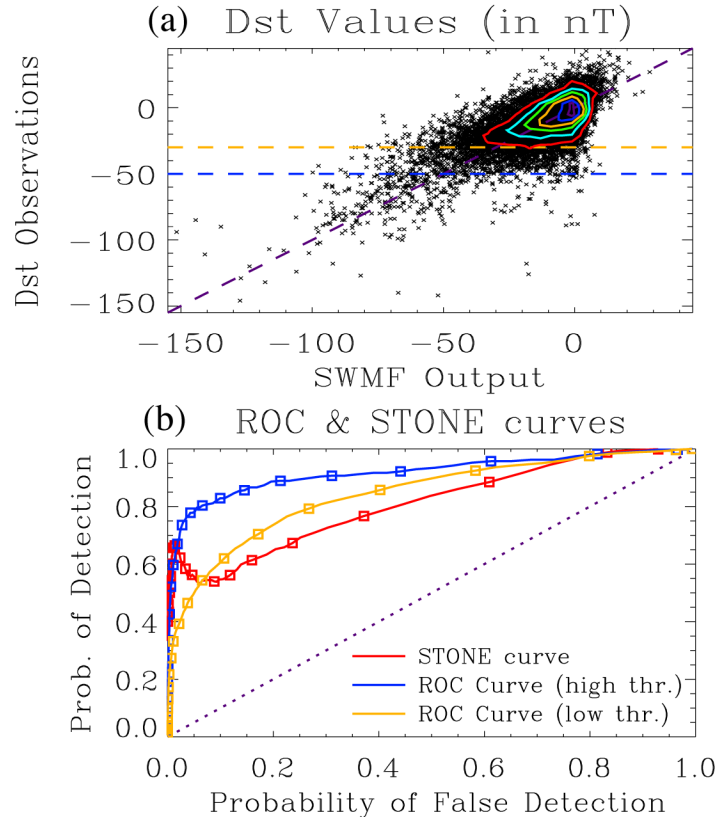250   horizontal dashed lines.
251

**Figure 2.** (a) Scatter plot of the observed real-time Dst time series (y-axis values) against a prediction Dst time series from the SWMF (x-axis values). The contours are drawn every 50 points per 5x5 nT bin. Also drawn are horizontal dashed lines at the ROC event thresholds of -30 and -50 nT, with events defined as the points below these lines. A purple dashed zero intercept unity slope line is also drawn, for reference. (b) STONE (red) and ROC curves (blue for -50 nT, orange for -30 nT observed event threshold) calculated from the scatter plot. Symbols are shown along all three curves at every 5 nT threshold increment. The diagonal dotted line with zero intercept and unity slope is shown for reference.

The ROC and STONE curves are calculated as follows and shown in Figure 2b. To create a ROC curve, the model threshold setting is initially set to +10 nT and then swept in 1 nT increments to -120 nT. The data threshold for events is held fixed, at -50 nT for the blue curve and -30 for the orange curve. To create STONE curve (red line), this same model threshold variation is followed, but the data threshold is also swept from +10 to -120 nT. Symbols are shown along each of the plots every 5 nT of threshold increment.

Some features of Figure 2b should be noted. It is seen that the ROC curves monotonically increase from (0,0) to (1,1). The ROC curve with a -50 nT event threshold is well above the zero-intercept, unity-slope line (the diagonal purple dotted line on Figure 2b), indicating that the model is reasonably good at reproducing moderate and stronger storm events recorded by the real-time Dst index. The closest approach to the upper left corner occurs at a threshold of -37 nT for the -50 nT threshold ROC curve and -17 nT for the -30 nT ROC curve, which indicates that the model somewhat underpredicts the strength of such storms.

275    The STONE curve lies both above and below these two ROC curves, depending on the
276    threshold. The STONE curve is coincident with each ROC curve at the locations where the ROC
277    curve model threshold setting is equal to the observational threshold setting (-30 nT for the
278    orange curve, -50 nT for the blue curve). They cross elsewhere, too, such as in the low-threshold
279    (i.e., a threshold of near and above zero) region in the upper right region of the plot. It is seen
280    that the STONE curve is not monotonic but includes a local maximum and local minimum at the
281    "high threshold" settings (minimum at -28 nT threshold and maximum at -52 nT threshold). The
282    nonmonotonicity is because POD increases at these threshold values. An increase in POD is
283    achieved by more points leaving the misses quadrant than leaving from the hits quadrant.

284    This is better understood by considering the distribution of points beyond a few threshold
285    choices. Figure 3 shows histograms of the points above a particular data or model threshold
286    setting. In particular, three threshold settings are displayed – -30 nT, -40 nT, and -50 nT –
287    showing the points at "higher" (more negative) Dst values in both the data and model (left and
288    right columns, respectively). For Figure 3a, the counts are for all points below some horizontal
289    line of an event identification threshold setting of the observations. For Figure 3b, the counts are
290    for all points to the left of some event identification threshold setting for the model values. The
291    calculated skew for these distributions is listed in each panel.

292    In Figure 3a, it is evident, both qualitatively from the histograms and quantitatively from
293    the skew values, that the distribution of model output values is significantly changing across
294    these three observational threshold settings. For the more negative threshold, there are far fewer
295    model values between zero and -50 nT. That is, across these threshold settings, many of the
296    points in the misses quadrant were converted into correct negatives. In Figure 3b, the three
297    distributions have essentially the same shape, with a large negative skew. These distributions do
298    not undergo the same systematic alteration in their shape the way that the distributions in Figure
299    3a did. Putting these two features together, it means that more misses were removed than hits,
300    and so POD increased as the STONE threshold was swept to more negative values between -30
301    nT and -50 nT. This resulted in a nonmonotonic wiggle in the STONE curve at these thresholds.
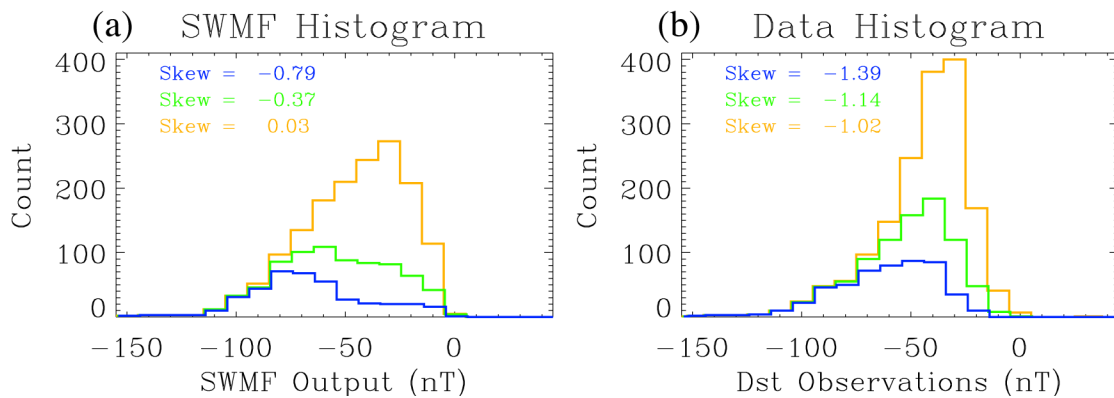
302



**Figure 3.** (a) Histogram of model values for all data values below three different thresholds: -30
nT (orange curve), -40 nT (green curve), and -50 nT (blue curve). (b) Histogram of data values
for all model values below the same three thresholds. The bin sizes for each histogram is 10 nT.
The calculated skew for each distribution is listed in each plot.

309 **3.2. Predicting energetic electrons in near-Earth space**

310       Ganushkina et al. (2019) compared real-time output from the inner magnetosphere
311 particle transport and acceleration model (IMPTAM) with measurements from the
312 magnetosphere electron detector (MAGED) on the geosynchronus orbiting environmental
313 satellites (GOES) in geostationary orbit at 6.62 Earth radii geocentric distance over the American
314 sector (Rowland & Weigel, 2012; Sillanpaa et al., 2017), specifically, with data from GOES-13,
315 -14, and -15. IMPTAM, initially developed by Ganushkina et al. (2001) and used regularly for
316 investigating the physics of plasma sheet electron transport (e.g., Ganushkina et al., 2013, 2014),
317 has been running in a real-time operational mode since February 2013, first in Europe and then a
318 mirror site at the University of Michigan. Ganushkina et al. (2015) made an initial comparison of
319 these model output values against a few months of GOES data, while Ganushkina et al. (2019)
320 provided a far more robust validation analysis of the model, covering over 18 months
321 (September 20, 2013 through March 31, 2015). It is this second interval that will be used again
322 for this study.

323       Figures 4a and 4b show two scatter plots comparing the IMPTAM and GOES electron
324 differential number fluxes at 40 keV. The colored contours show the point density, with a new
325 curve every 50 points within a bin (defined, for these contours, with 10 bins per decade in both
326 the data and model values). Figure 4a presents the full data set while Figure 4b only shows the
327 comparison for those values in the 03 to 09 magnetic local time (MLT) range, the region found
328 by Ganushkina et al. (2019) to have a "good comparison" between the data and model values. On
329 each of these plots, two observational event thresholds are shown as the horizontal dashed lines,
330 drawn at $5 \times 10^4$ and $2 \times 10^5$ electrons $cm^{-2}$ $s^{-1}$ $sr^{-1}$ $keV^{-1}$ in green and blue, respectively.
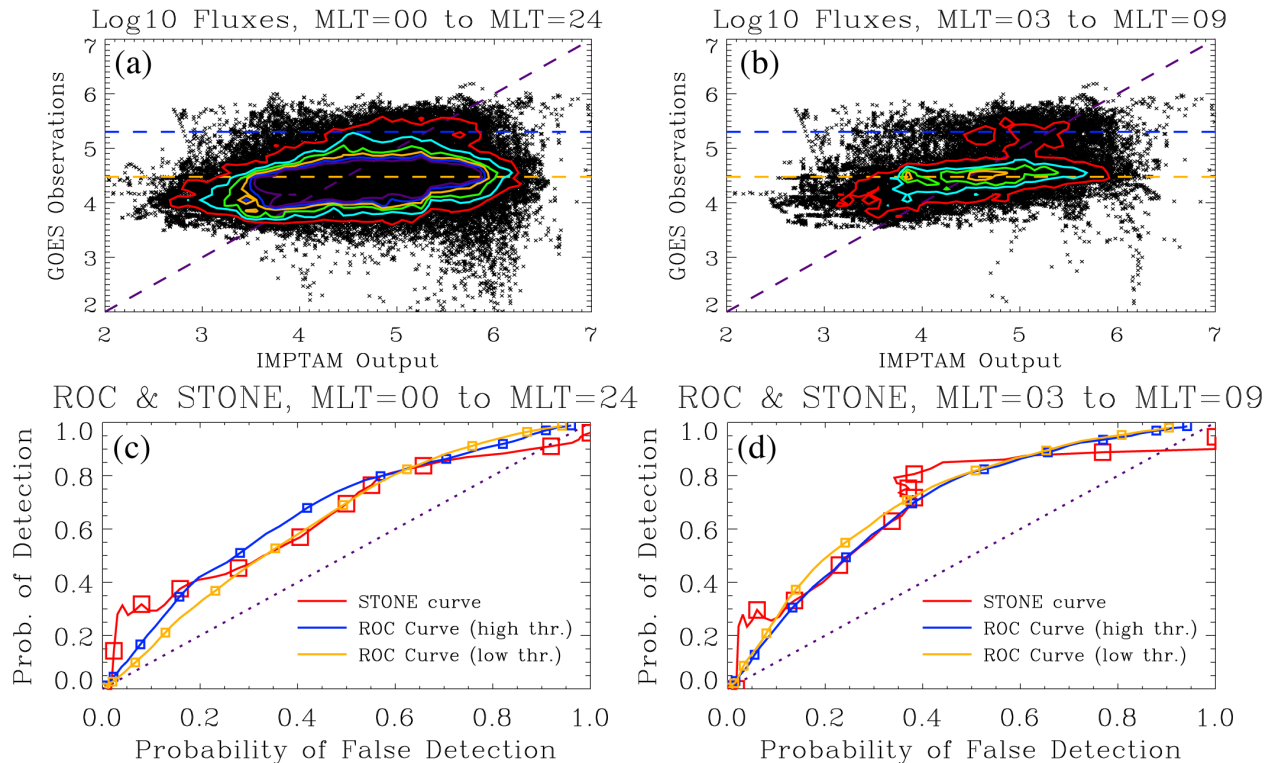331



332

333 **Figure 4.** Scatter plot comparing GOES and IMPTAM 40 keV electron differential number
334 fluxes (log base 10 of electrons cm$^{-2}$ s$^{-1}$ sr$^{-1}$ keV$^{-1}$) for (a) all MLTs and (b) the 03-09 MLT
335 range. Color contours are shown every 50 points per bin (10 bins per decade in both data and
336 model). The horizontal dashed lines show the ROC thresholds of 3x10$^4$ and 2x10$^5$. A purple
337 dashed zero intercept and unity slope line is shown for reference. The lower panels show STONE
338 curves (red) and ROC curves (blue for 2x10$^5$ and orange for 3x10$^4$) for (c) the full MLT
339 comparison and (d) the 03-09 MLT range. Symbols are shown every factor of 2 increase in
340 threshold value. The diagonal dotted line with zero intercept and unity slope is shown for
341 reference.

342

343       Figures 4c and 4d show the ROC and STONE curves for these two data-model
344 comparisons, the full set with values at all MLTs and the subset from 03 to 09 MLT,
345 respectively. In both Figures 4c and 4d, the STONE curve again has a nonmonotonic shape at
346 high threshold settings (above 4x10$^5$). Like the similar case for the Dst STONE curve in Figure
347 2b, this shows that, for these thresholds, more points are being removed from the misses
348 quadrant than being removed from the hits quadrant.

349       Figure 4d has another unusual feature in the STONE curve, seen as a nonmonotonicity in
350 the x-axis values. This is from the POFD values increasing with increasing threshold (rather than
351 decreasing, as they always do with a ROC curve). This is occurring for thresholds between 1x10$^4$
352 and 4x10$^4$, just as the STONE curve crosses the orange ROC curve. Considering equation (2)
353 above, the correct negatives in the denominator are always increasing with increasing threshold,
354 as points convert to this quadrant from any of the other three quadrants. For POFD to increase,
355 the false alarms had to increase faster than the correct negative point count. This is seen in Figure
356 4b as the points have a horizontal peak (highlighted by the flat, elongated color contours). Many
357 points are being converted from the hits quadrant into the false alarms quadrant and, for these
358 threshold settings, this conversion to false alarms outpaces the conversion of points into the
359 correct negatives quadrant. This results in a ripple in the STONE curve at these thresholds.

360       Figures 4c and 4d, the STONE curve is quite close to the two ROC curves, which are
361 very similar to each other. This can be understood from the "flatness" of the cloud of points in
362 the scatter plots in Figures 4a and 4b. The points are not well aligned with the zero intercept and
363 unity slope line, revealing less than perfect agreement between the observations and model
364 output. However, in terms of physics-based real-time modeling of near-Earth magnetospheric
365 electron fluxes, this is actually quite good, arguably the best that is currently available. This
366 means that all ROC curves will be close to each other, as any observational event identification
367 threshold will have a relatively similar transfer of points between the quadrants. However,
368 because the model is trying to exactly reproduce the observed flux values, the STONE curve can
369 be calculated, and this new curve includes several nonmonotonicities. The wiggles and ripples in
370 the STONE curve reveal thresholds where the distribution of points, in either the vertical or
371 horizontal direction, are asymmetric, bi-modal, or otherwise non-Gaussian. The ROC curves
372 cannot reveal this kind of information about the distribution of points in the scatter plot the way
373 that the STONE curve can.

374 **4. Discussion**

375     The STONE curve introduced above is a new tool for assessing the ability of a model
376 with a continuous-valued output to exactly match a continuous-valued data set. As illustrative
377 example usages, it was applied to two recently-published data-model comparisons, a prediction
378 of the disturbance storm-time index Dst and a prediction of energetic electron fluxes in near-
379 Earth space.

380     The STONE curve is quite similar to the ROC curve. It is based on the same contingency
381 table calculations of POD and POFD, plotting these two values against each other for a range of
382 event threshold settings. Like the ROC curve, it starts at (1,1) for low threshold settings and
383 moves to (0,0) for high threshold settings. Also like the ROC curve, being above the zero-
384 intercept, unity-slope line indicates a prediction that is better than random chance. Curves are
385 better when they are closer to the upper left corner in POFD-POD space, and a common choice
386 for the best optimization point along along a ROC or STONE curve is that closest to this corner
387 as this point reveals the best model threshold setting for optimizing discrimination performance.
388 That is, both curves reveal a possible best model threshold setting for event prediction, the ROC
389 curve revealing the best settings for a specified observational event identification threshold and
390 the STONE curve revealing the best setting against the an identically defined observational
391 event. Of course, this is "best" only if discrimination is what should be optimized for the
392 particular application. A different threshold settings might be most favorable if other
393 considerations outweigh discrimination, such as minimizing false alarms or maximizing a
394 particular skill score.

395     Another similarity is that the integral area of the ROC curve, AUC, is equally applicable
396 to the STONE curve. AUC, a synthesis of the entire threshold-setting range into a single number,
397 indicates the quality of the chosen model to predict the events identified in the observational data
398 (see the detailed explanation of AUC in Fawcett (2006) or Ekelund (2011)). Being an integrated
399 quantity, AUC is a complementary metric to the "best model threshold setting for event
400 prediction" mentioned in the preceding paragraph because AUC uses information from all model
401 threshold settings, even those with POFD-POD coordinates far from the "best setting" upper-left
402 corner of the graph. Comparing AUCs for several STONE curves (i.e., using different models
403 against the same data set) will provide a quantitative assessment of which of the models has the
404 best system-level predictive capability against that data set. It could be that the model with the
405 highest AUC is not the model with a point along its STONE curve closest to (0,1) in POFD-POD
406 space. Such a case reveals that the first model, with the higher AUC, has the best model physics
407 for reproducing the data set as a whole, but that the second model is actually best at predicting
408 events with a particular threshold setting. Because it is calculated the same way, AUC can be
409 used to compare STONE curves just like it is for ROC curves.

410     A key difference between the STONE and ROC curves is that the STONE curve can have
411 nonmonotonicities. These features, which can be wiggles with respect to either POD or POFD,
412 reveal features of the model prediction of events that are not easily extracted from a ROC curve.
413 This makes the STONE curve somewhat like a fit performance metric, even though it is an
414 event-detection metric that disregards the difference between the data-model pairs.

415     The nonmonotonicities in the STONE curve reveal information about the distribution of
416 points in the data-model comparison. Specifically, they show the existence of an asymmetry,
417 perhaps a non-Gaussian point spread like a skewed or bimodal distribution, for the pairs above

418    that threshold setting. Combined with a histogram or even fit-performance data-model
419    comparison formulas for this subset of either the data or model values, the nature of this
420    distribution can be explored.

421          Why not just start out by calculating fit performance metrics on these subsets? The
422    answer is because the subset of interest would not have been known; the STONE curve revealed
423    the thresholds where the distribution had a changing or non-Gaussian distribution. That is, it
424    could be used to optimize the fit performance analysis by identifying the subset of the data or
425    model that should be considered in more detail. Also, the STONE curve includes information not
426    just within a subset of the data (discrimination) or a subset of the model (reliability), but includes
427    information about the entire data-model comparison set, because POD and POFD use all data-
428    model pairs in the point counting in the quadrants. For one of the specific examples in the
429    manuscript: continuous metrics will tell the user very little about the SWMF's ability to predict
430    magnetic storms of -50 nT or less. A ROC curve is far more suited to this, and a STONE curve
431    one step farther, revealing the ability of the model to predict Dst levels below any threshold
432    (which could be accomplished by a large family of ROC curves). No continuous metric that does
433    this type of assessment. If the detection of events is desired, then the STONE curve is an
434    advantageous assessment tool in addition to standard continuous fit performance metrics.

435          A useful follow-on study to this would be a detailed analysis of the features of the
436    STONE curve to the underlying distribution of points in the data-model scatter plot. That is, by
437    assuming known two-dimensional distributions of points of several different shapes and
438    parameter settings, the connection between the distribution and the resulting features in the
439    STONE curve can be isolated. Such an in-depth assessment of the STONE curve is beyond this
440    initial description and illustrative usage of this metrics tool and is left as a future project.

441          A key feature of the STONE curve is that it reveals the threshold (or range of thresholds)
442    for which the model does best at reproducing similarly-defined events in the data. A single ROC
443    curve cannot do this because it uses a fixed threshold for identifying events in the observations.
444    When the data are continuously-values and the model is seeking to reproduce these exact values,
445    then it is useful to examine the event detection capability of the model at the same threshold
446    settings between data and model. A single ROC curve doesn't do this, except at one threshold
447    setting. The STONE curve, therefore, is a better assessment tool for models that are trying to
448    predict the exact value of a data set.

449          The ROC is still a highly useful tool for event prediction and this study does not seek to
450    replace it with the STONE curve. Indeed, the ROC curve is optimal for categorical data sets
451    where the observations have been pre-classified as events and non-events. In this case, the
452    STONE curve cannot be used because the data and model are on different scales, the former
453    being a binary yes-no designation and the latter being either a real number range or its own
454    categorical designation. The ROC curve can handle this difference in units while the STONE
455    curve cannot.

456          The two example data-model comparisons to which the STONE curve was applied are
457    both from space physics. The first was an evaluation between a physics-based model of
458    geospace, running in real time, with the real-time version of the Dst index, a measure of
459    geospace activity (see its comparison with other similar indices in Katus & Liemohn (2013)).
460    Many models exist for the prediction of Dst (see the review by Liemohn, McCollough, et al.,
461    2018), with some models doing exceptionally well at reproducing the observed time series.

462    While this chosen model for this comparison is arguably the best physics-based model for
463    reproducing Dst (see, for comparison, the solar cycle storm-interval Dst comparison of Liemohn
464    & Jakowski (2008)), it is not the best model available at predicting this index. In fact, many
465    empirical models are substantially better at capturing the storm intervals of Dst. The second
466    example was a comparison of a physics-based model of energetic electron fluxes in the near-
467    Earth magnetosphere, running in real time, with real-time observations from a geosynchronous
468    spacecraft. Magnetospheric charged particle fluxes are notoriously difficult to reproduce with
469    physics-based modeling approaches (see, e.g., Morley et al., 2018), and even empirical models
470    reduce the problem to remove the fast temporal dynamics, averaging over a day (e.g., Li, 2004)
471    or an hour (e.g., Boynton et al., 2019). That is, these two examples represent state-of-the-art
472    physics-based approaches to space weather nowcasting, but are not the best predictions of these
473    two quantities across the field.

474        It is worth stating here that there are many other metrics in existence for evaluating a
475    scatter plot of data-model values like that shown in Figure 1. No one metric equation or
476    technique does everything; each was designed to assess only a specific aspect of the relationship.
477    That is, neither the ROC curve nor the STONE curve should be used as the sole assessment tool
478    for a model against a particular data set. In practice, many metrics, from both the continuous fit-
479    performance grouping and from the categorical event-detection grouping, should be applied to
480    examine the quality of the model from a number of perspectives.

481        It should be mentioned that this is not the first application of sliding both the
482    observational and model event identification threshold. As one example of this, in their
483    presentation and initial usage of the extreme dependency score (EDS), Stephenson et al. (2008)
484    simultaneously moved both thresholds. Events become rarer with increasing threshold and that
485    study examined the relationship of EDS as a function of this rarity – moving both thresholds
486    together, as is done here for the STONE curve.

487        A final note to make here is that this is not the first usage of the STONE curve. Both
488    Liemohn, Ganushkina, et al. (2018) and Liemohn, McCollough, et al. (2018) used STONE
489    curves in the plots labeled as ROC curves. It is clear that these panels are mislabeled because
490    nonmonotonicities are seen in these lines.

491    **5. Conclusions**

492        A new data-model comparison assessment tool has been introduced, described, used, and
493    interpreted – the sliding threshold of observations for numeric evaluation curve. Based on the
494    relative operating characteristic curve, the STONE curve is created by plotting POD against
495    POFD for a wide range of threshold settings. The main difference with the ROC curve is that the
496    STONE curve requires the data to be continuous-valued real numbers and the model to be
497    attempting to reproduce these exact values. The threshold is moved not only for the model, as is
498    done for the ROC curve, but also for the observational event identification threshold setting,
499    which is moved simultaneously with the model threshold setting.

500        The STONE curve has many features in common with the ROC curve with one large
501    exception – it can have nonmonotonicities in both the POD and POFD values. For the ROC
502    curve, the points shift within the quadrants defining POD or within the quadrants used to define
503    POFD, but not between these two mutually exclusive regions. The ROC curve is, therefore,
504    always monotonic, sweeping from (1,1) to (0,0) in POFD-POD space. For the STONE curve, the

505  motion of the observational threshold moves points from the POD regions to the POFD regions,
506  allowing for these nonmonotonic features in the STONE curve.

507  These wiggles and ripples, however, reveal information about the underlying distribution
508  of points in the data-model scatter plot. Specifically, if the distribution is shifted, asymmetric, or
509  bi-modal, the STONE curve will have a nonmonotonicity. Further investigation of the
510  distribution, through a histogram, skew calculation, or other metric assessment, can reveal the
511  true nature of the data-model comparison for this threshold setting.

512  It is hoped that the STONE curve becomes a useful data-model comparison tool. It has
513  been used with two space weather applications in this study but these are purely illustrative
514  examples. A dozen studies using ROC curves across the Earth and space sciences were given in
515  the Introduction above. Some of these studies were based on observations that were pre-
516  classified yes/no as events or not, and so the ROC curve is the proper tool for assessing the
517  model's ability to predict those events. Some of these studies, however, and others like them, are
518  based on models trying to exactly predict the observed data values, in which case the STONE
519  curve might be a useful assessment tool. For any continuous-valued model trying to reproduce
520  the exact numbers of a continuous-valued data set, the STONE curve can be calculated, perhaps,
521  as shown for the two examples here, revealing additional information about the data-model
522  comparison than can be obtained from the ROC curve alone. The STONE curve is a general
523  purpose metric for use whenever a model is trying to exactly reproduce a continuous-valued data
524  set. It can be used with both archival observations as well as for assessment of real-time
525  nowcasting across the full breadth of science and engineering disciplines.

526

546

547 **References**

548 Anagnostopoulos, G. G., Fatichi, S., & Burlando, P. (2015). An advanced process-based
549      distributed model for the investigation of rainfall-induced landslides: The effect of
550      process representation and boundary conditions. *Water Resources Research*, *51*, 7501–
551      7523. https://doi.org/10.1002/2015WR016909

552 Azari, A. R., Liemohn, M. W., Jia, X., Thomsen, M. F., Mitchell, D. G., Sergis, N., et al. (2018).
553      Interchange injections at Saturn: Statistical survey of energetic H+ sudden flux
554      intensifications. *Journal of Geophysical Research: Space Physics*, 123, 4692– 4711.
555      https://doi.org/10.1029/2018JA025391

556 Bobra, M. G., & Couvidat, S. (2015). Solar flare prediction using SDO/HMI vector magnetic
557      field data with a machine learning algorithm. The Astrophysical Journal, 789, 135.
558      http://dx.doi.org/10.1088/0004-637X/798/2/135

559 Borah, N., Sahai, A. K., Chattopadhyay, R., Joseph, S., Abhilash, S., & Goswami, B. N. (2013).
560      A self-organizing map–based ensemble forecast system for extended range prediction of
561      active/break cycles of Indian summer monsoon. *Journal of Geophysical Research -
562      Atmospheres*, *118*, 9022– 9034. https://doi.org/10.1002/jgrd.50688

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

566 Brenning, A., Grasser, M., & Friend, D. A. (2007). Statistical estimation and generalized
567      additive modeling of rock glacier distribution in the San Juan Mountains, Colorado,
568      United States. *Journal of Geophysical Research – Solid Earth*, 112, F02S15.
569      https://doi.org/10.1029/2006JF000528

570 Carter, J. V., J. Pan, S. N. Rai, & S. Galandiuk (2016). ROC-ing along: Evaluation and
571      interpretation of receiver operating characteristic curves. *Surgery, 159*(6), 1638-1645,
572      https://doi.org/10.1016/j.surg.2015.12.029

573 Chen, Y., Manchester, W. B., Hero, A. O., Toth, G., Dufumier, B., Zhou, T., et al. (2019).
574      Identifying solar flare precursors using time series of SDO/HMI Images and SHARP
575      Parameters. *Space Weather*, 17, 1404– 1426. https://doi.org/10.1029/2019SW002214

576  Delle Monache, L., Hacker, J. P., Zhou, Y., Deng, X., & Stull, R. B. (2006). Probabilistic
577      aspects of meteorological and ozone regional ensemble forecasts. *Journal of Geophysical
578      Research - Atmospheres*, *111*, D24307. https://doi.org/10.1029/2005JD006917

579 Dong, C., S. W. Bougher, Y. Ma, G. Toth, A. F. Nagy, & Najib, D. (2014). Solar wind
580      interaction with Mars upper atmosphere: Results from the one-way coupling between the
581      multifluid MHD model and the MTGCM model. *Geophysical Research Letters*, *41*,
582      2708–2715. https://doi.org/10.1002/2014GL059515

583 Ekelund, S. (2011). ROC Curves – What are They and How are They Used?, *Point of Care,
584      11*(1), 16-21. https://doi.org/10.1097/POC. 0b013e318246a642

585 Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861-874.
586      https://doi.org/10.1016/j.patrec.2005.10.010

587  Gabriel, P., Barker, H. W., O'Brien, D., Ferlay, N., & Stephens, G. L. (2009). Statistical
588      approaches to error identification for plane-parallel retrievals of optical and
589      microphysical properties of three-dimensional clouds: Bayesian inference. *Journal of*
590      *Geophysical Research - Atmospheres*, *114*, D06207.
591      https://doi.org/10.1029/2008JD011005

592  Ganushkina N. Yu., T. I. Pulkkinen, V. F. Bashkirov, D. N. Baker, & X. Li (2001). Formation of
593      intense nose structures. *Geophysical Research Letters, 28*, 491-494.

594  Ganushkina, N. Y., Amariutei, O. A., Shprits, Y. Y., & Liemohn, M. W. (2013). Transport of the
595      plasma sheet electrons to the geostationary distances. *Journal of Geophysical Research:*
596      *Space Physics, 118* (1), 82-98. https://doi.org/10.1029/2012JA017923

597  Ganushkina, N. Y., Liemohn, M. W., Amariutei, O. A., & Pitchford, D. (2014). Low-energy
598      electrons (550 keV) in the inner magnetosphere. *Journal of Geophysical Research: Space*
599      *Physics, 119* (1), 246-259. https://doi.org/10.1002/2013JA019304

600  Ganushkina, N. Y., Amariutei, O. A., Welling, D., & Heynderickx, D. (2015). Nowcast model
601      for low-energy electrons in the inner magnetosphere. *Space Weather , 13* (1), 16-34.
602      https://doi.org/10.1002/2014SW001098

603  Ganushkina, N. Yu., Liemohn, M. W., & Dubyagin, S. (2018), Current systems in the Earth's
604      magnetosphere, *Reviews of Geophysics*, *56*(2), 309-332. https://doi.org/
605      10.1002/2017RG000590

606  Ganushkina, N. Y., Sillanpaa, I., Welling, D. T., Haiducek, J., Liemohn, M. W., Dubyagin, S., &
607      Rodriguez, J., (2019). Validation of Inner Magnetosphere Particle Transport and
608      Acceleration Model (IMPTAM) on the long-term GOES MAGED measurements of keV
609      electron fluxes at geostationary orbit. *Space Weather*, 17, 687-708.
610      https://doi/org/10.1029/2018SW002028

611  Hogan, R. J. & Mason, I. B. (2012). Deterministic Forecasts of Binary Events. In *Forecast*
612      *Verification* (eds I. T. Jolliffe and D. B. Stephenson). doi:10.1002/9781119960003.ch3

613  Ippolito, A., Scotto, C., Sabbagh, D., Sgrigna, V., & Maher, P. (2016). A procedure for the
614      reliability improvement of the oblique ionograms automatic scaling algorithm. *Radio*
615      *Science*, 51, 454– 460. https://doi.org/10.1002/2015RS005919

616  Jia, X., Hansen, K. C., Gombosi, T. I., Kivelson, M. G., Toth, G., DeZeeuw, D. L., & Ridley, A.
617      J., (2012). Magnetospheric configuration and dynamics of Saturn's magnetosphere: A
618      global MHD simulation. *Journal of Geophysical Research Space Physics, 117,* A05225.
619      https://doi.org/10.1029/2012JA017575

620  Jolliffe, I.T., & Stephenson, D.B. (2012). *Forecast verification: A practitioner's guide*
621      *in atmospheric science*. Wiley-Blackwell, Hoboken, NJ.

622  Katus, R. M., & Liemohn, M. W. (2013). Similarities and differences in low-to-mid-latitude
623      geomagnetic indices. *Journal of Geophysical Research, 118*, 5149-5156.
624      https://doi.org/10.1002/jgra.50501.

625  Li, X. (2004). Variations of 0.7–6.0 MeV electrons at geosynchronous orbit as a function of solar
626      wind. *Space Weather*, 2, S03006. https://doi.org./10.1029/2003SW000017

627   Liemohn, M. W., & Jazowski, M. (2008). Ring current simulations of the 90 intense storms
628         during solar cycle 23. *Journal of Geophysical Research, 113*, A00A17.
629         https://doi.org/10.1029/2008JA013466

630   Liemohn, M. W., Xu, S., Dong, C., Bougher, S. W, Johnson, B. C., Ilie, R., & De Zeeuw, D. L.,
631         (2017). Ionospheric control of the dawn-dusk asymmetry of the Mars magnetotail current
632         sheet. *Journal of Geophysical Research Space Physics, 122*, 6397-6414.
633         https://doi.org/10.1002/2016JA023707

634   Liemohn, M. W., Ganushkina, N. Y., De Zeeuw, D. L., Rastaetter, L., Kuznetsova, M., Welling,
635         D. T., Toth, G., Ilie, R., Gombosi, T. I., & van der Holst, B. (2018). Real-time SWMF
636         and CCMC: assessing the Dst output from continuous operational simulations. *Space*
637         *Weather, 16*, 1583-1603. https://doi.org/10.1029/2018SW001953.

638   Liemohn, M. W., McCollough, J. P., Jordanova, V. K., Ngwira, C. M., Morley, S. K., Cid, C.,
639         Tobiska, W. K., Wintoft, P., Ganushkina, N. Y., Welling, D. T., Bingham, S., Balikhin,
640         M. A., Opgenoorth, H. J., Engel, M. A., Weigel, R. S., Singer, H. J., Buresova, D.,
641         Bruinsma, S., Zhelavskaya, I., Shprits, Y. Y., & Vasile, R. (2018). Model evaluation
642         guidelines for geomagnetic index predictions. *Space Weather*, 16, 2079– 2102.
643         https://doi.org/10.1029/2018SW002067

644   Ma, Y. J., Nagy, A. F., Russell, C. T., Strangeway, R. J., Wei, H. Y., & Toth, G. (2013). A
645         global multispecies single-fluid MHD study of the plasma interaction around Venus.
646         *Journal of Geophysical Research Space Physics, 118*, 321–330.
647         https://doi.org/10.1029/2012JA018265

648   Manzato, A. (2005). An odds ratio parameterization for ROC diagram and skill score indices.
649         *Weather and Forecasting, 20,* 918-930. https://doi.org/10.1175/WAF899.1

650   Manzato, A. (2007). A note on the maximum Peirce skill score. *Weather and Forecasting, 22,*
651         1148-1154. https://doi.org/10.1175/WAF1041.1

652   Mathieu, J. A., & Aires, F. (2018). Using neural network classifier approach for statistically
653         forecasting extreme corn yield losses in Eastern United States. *Earth and Space Science*,
654         5, 622– 639. https://doi.org/10.1029/2017EA000343

655   Meade, B. J., DeVries, P. M. R., Faller, J., Viegas, F., & Wattenberg, M. (2017). What is better
656         than Coulomb failure stress? A ranking of scalar static stress triggering mechanisms from
657         $10^5$ mainshock-aftershock pairs. *Geophysical Research Letters*, *44*, 11,409– 11,416.
658         https://doi.org/10.1002/2017GL075875

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

662   Muller, R. H. (1944). Verification of short-range weather forecasts (a survey of the literature).
663         *Bulletin of the American Meteorological Society, 25*, 18-27.

664   Murphy, A. H. (1996). The Finley Affair: a signal event in the history of forecast verification.
665         *Weather and Forecasting, 11*, 3-20.

666 Murphy, A.H. & Winkler, R.L. (1987). A general framework for forecast verification. *Monthly*
667         *Weather Review, 115*, 1330-1338. https://doi.org/10.1175/1520-
668         0493(1987)115%3C1330:AGFFFV%3E2.0.CO;2

669 Reiff, P. H. (1990). The use and misuse of statistics in space physics. *Journal of Geomagnetism*
670         *and Geoelectricity, 42,* 1145–1174. https://doi.org/10.5636/jgg.42.1145.

671 Reuter, B., & Schweizer, J. (2018). Describing snow instability by failure initiation, crack
672         propagation, and slab tensile support. *Geophysical Research Letters*, 45, 7019– 7027.
673         https://doi.org/10.1029/2018GL078069

674 Rostoker, G. (1972). Geomagnetic indices. *Reviews of Geophysics and Space Physics*, *10*, 935–
675         950.

676 Rowland, W., & Weigel, R. S. (2012). Intracalibration of particle detectors on a three-axis
677         stabilized geostationary platform. *Space Weather, 10* (11).
678         https://doi.org/10.1029/2012SW000816

679 Pulkkinen, A., L. Rastätter, M. Kuznetsova, H. Singer, C. Balch, D. Weimer, et al. (2013).
680         Community-wide validation of geospace model ground magnetic field perturbation
681         predictions to support model transition to operations. *Space Weather, 11*, 369–385.
682         https://doi.org/10.1002/swe.20056

683 Sillanpaa, I., Ganushkina, N. Y., Dubyagin, S., & Rodriguez, J. V. (2017). Electron Fluxes at
684         Geostationary Orbit From GOES MAGED Data. *Space Weather, 15* (12), 1602-1614.
685         doi: 10.1002/2017SW001698

686 Stefanescu, E. R., et al. (2014). Temporal, probabilistic mapping of ash clouds using wind field
687         stochastic variability and uncertain eruption source parameters: Example of the 14 April
688         2010 Eyjafjallajökull eruption. *Journal of Advances in Modeling Earth Systems*, 6, 1173–
689         1184. https://doi.org/10.1002/2014MS000332

690 Stephenson, D.B., Casati, B., Ferro, C.A.T. & Wilson, C.A. (2008), The extreme dependency
691         score: a non-vanishing measure for forecasts of rare events. *Meteorological Applications*,
692         *15*, 41-50. https://doi.org/10.1002/met.53

693 Toth, G., B. van der Holst, I. V. Sokolov, D. L. De Zeeuw, T. I. Gombosi, F. Fang, et al. (2012).
694         Adaptive numerical algorithms in space weather modeling. *Journal of Computational*
695         *Physics, 231*, 870-903. https://doi.org/10.1016/j.jcp.2011.02.006

696 Wilks, D. S. (2012). *Statistical methods in the atmospheric sciences* (3rd ed.). Oxford: Academic
697         Press.