Analysis of features in a sliding threshold of observation for numeric evaluation (STONE) curve

Michael W. Liemohn,[1] Joshua G. Adam,[1] Natalia Yu. Ganushkina[1,2]

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

[2] Finnish Meteorological Institute, Helsinki, Finland

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

In preparation for *Space Weather*

Key Points:

- The STONE curve, an event detection sweeping-threshold data-model comparison metric, reveals thresholds where the model matches the data
- STONE curves can be nonmonotonic, revealing the location and size of clusters of model under- or over-estimations of the observations
- STONE curve features are analyzed, quantifying the shape of nonmonotonicities relative to distribution characteristics and other metrics

AGU Index Terms:

1984	Statistical methods: Descriptive (4318)
4318	Statistical analysis (1984, 1986)
7924	Forecasting (1922, 2722, 4315)
0550	Model verification and validation
2764	Plasma sheet

Keywords:

ROC curve, STONE curve, data-model comparison, model validation, forecasting, plasma sheet electrons

Abstract

We apply idealized scatter-plot distributions to the sliding threshold of observation for numeric evaluation (STONE) curve, a new model assessment metric, to examine the relationship between the STONE curve and the underlying point-spread distribution. The STONE curve is based on the relative operating characteristic (ROC) curve but is developed to work with a continuous-valued set of observations, sweeping both the observed and modeled event identification threshold simultaneously. This is particularly useful for model predictions of time series data, as is the case for much of terrestrial weather and space weather. The identical sweep of both the model and observational thresholds results in changes to both the modeled and observed event states as the quadrant boundaries shift. The changes in a data-model pair's event status result in nonmonotonic features to appear in the STONE curve when compared to a ROC curve for the same observational and model data sets. Such features reveal characteristics in the underlying distributions of the data and model values. Many idealized datasets were created with known distributions, connecting certain scatter-plot features to distinct STONE curve signatures. A comprehensive suite of feature-signature combinations is presented, including their relationship to several other metrics. It is shown that nonmonotonic features appear if a local spread is more than 0.2 of the full domain, or if a local bias is more than half of the local spread. The example of real-time plasma sheet electron modeling is used to show the usefulness of this technique, especially in combination with other metrics.

Plain Language Summary

Many statistical tools have been developed to aid in the assessment of a numerical model's quality at reproducing observations. Some of these techniques focus on the identification of events within the data set, times when the observed value is beyond some threshold value that defines it as a value of keen interest. An example of this is whether it will rain, in which events are defined as any precipitation above some defined amount. A method called the sliding threshold of observation for numeric evaluation (STONE) curve sweeps the event definition threshold of both the model output and the observations, resulting in the identification of threshold intervals for which the model does well at sorting the observations into events and nonevents. An excellent data-model comparison will have a smooth STONE curve, but the STONE curve can have wiggles and ripples in it. These features reveal clusters when the model systematically overestimates or underestimates the observations. This study establishes the connection between features in the STONE curve and attributes of the data-model relationship.

1. Introduction

Given a data set of continuous values and model output that is trying to reproduce that set of observations, there are many ways to conduct a quantitative comparison between these two number sets. Metrics, equations or techniques for comparing model output with a corresponding data set, come in many forms, but all are statistical analysis tools that help numerically specify what can usually be seen qualitatively from a scatterplot of the number sets against each other. Many well-known and useful metrics exist, as summarized by research studies such as Murphy (1991), Kubo et al. (2017), and Morley et al. (2018), or as reviewed in books, such as those by Joliffe & Stephenson (2012) and Wilks (2019). Each metric distills some aspect of the data-model relationship down to a single number or curve, which can then be interpreted with respect to the particular assessment being conducted. It is important to choose metrics that focus on the facet of the data-model relationship that matters, and combinations of metrics can often lead to additional insights (e.g., Potts, 2012, Liemohn et al., 2021). The decisions resulting from metrics usage could range anywhere along the Application Usability Level process (Halford et al., 2019), from a scientific conclusion at level 1 to a validation assessment at level 3 or 6 to an operational task at level 9.

One style of data-model comparison is event detection, in which the otherwise continuous number sets are reduced to yes-no binary designations depending on the number's value relative to some threshold value defining "events" (see, e.g., the review by Hogan and Mason, 2012). Because of the transformation from real values into yes-no labels, this technique is sometimes called categorical evaluation. Given event identification thresholds for the two number sets, the scatterplot is converted into a 2x2 matrix, called a contingency table or confusion matrix, counting the points within each quadrant of the scatterplot above and below each threshold. That is, the exact values no longer matter, only the event status matters, and values just barely beyond the threshold are counted as events equally with those that are far beyond the threshold. This is useful if the assessment being conducted is not concerned with matching the exact values but rather cares more about the model's ability to sort the observations according to event status. Many metrics have been created from these four count values to assess the quality of the model at achieving a good separation of observed events and nonevents.

An extension of event detection methods that more fully utilizes the continuous aspects of the two original number sets is the technique of sliding the thresholds of event identification. These two thresholds, one for the observations and one for the model output, do not have to be the same number. When the observed event identification threshold is held constant and the model threshold is swept, this yields a new contingency table at each modeled event identification threshold setting, from which metrics as a function of threshold setting can be calculated (e.g., Mason, 1982). These curves of metrics reveal the threshold settings where certain metrics are optimized, allowing users to choose the model threshold that best suits their needs.

The usefulness of sweeping the threshold extends beyond these metrics curves, though, with the technique of plotting the metrics against each other. A technique that has found particular usefulness across Earth and space science is the relative operating characteristic (ROC) curve (see, e.g., Hogan & Mason, 2012). Originally known as the receiver-operator characteristic curve because of its development by the radar community, the ROC curve plots two metrics against each other: probability of detection (POD) and probability of false detection (POFD). By holding the observed threshold fixed and sweeping the modeled threshold, the resulting POD and POFD curves monotonically vary from one to zero (from low to high threshold setting, respectively), resulting in a ROC curve that monotonically progresses from (1,1) to (0,0) in POFD-POD space. The area under the curve (AUC), sometimes converted into the ROC skill score (Swets, 1986), is then used as an overall measure of the quality of the model at correctly sorting the observations into events and nonevents.

The technique of holding the observed events fixed and sliding the model threshold through a continuous model output number set has been done for many Earth and space science applications. The study by Mathieu & Aires (2018) swept model thresholds in order to determine the best settings for certain climate-based predictors (e.g., rainfall, temperature, drought conditions) of corn yield, specifically assessing which predictors were best at determining corn yield losses. Another example of the usage of sliding threshold technique is the study by Manzato (2005), who swept the model threshold to optimize weather forecast model usage. They conclude that the odds ratio metric is particularly useful for maximizing another metric, the Heidke skill score. A planetary science example is that of Azari et al. (2018), who swept thresholds to determine the optimal settings for classifying hot plasma injection events in Saturn's magnetosphere. Their follow-up study (Azari et al., 2020) assessed their injection event determination model against several machine learning approaches, showing that the ROC curves for their model are as good or better than "black box" approaches (that is, including physics often helps with event classification). Sliding thresholds are used in earth science studies, too, for example when Meade et al. (2017) swept model event settings to determine which stress metrics are most effective at predicting aftershocks following major earthquakes.

All of the example usages mentioned above held the observed events fixed and varied only the model event threshold setting. This is very useful when the observed events are known; e.g., either an earthquake was recorded or one wasn't. In addition, this technique is powerful when the "model" is actually a driver parameter and has a different value range and perhaps even different units than the observations that it is trying to sort. In these cases, sliding only the model event identification threshold is possible.

Sometimes, however, the data are real numbers; to use a space weather example, magnetic perturbation values as a function of time at a particular ground station. Furthermore, you might have a model that is attempting to exactly reproduce this number set. In this particular case, there is no need to keep the observed event identification threshold constant; it can be swept along with the model event identification threshold. Such a technique was conducted by Liemohn et al. (2020) to introduce the analysis method they called the sliding threshold of observations for numeric evaluation (STONE) curve. The STONE curve is like the ROC curve in that it is a plot of POD versus POFD, but the underlying contingency tables for each point on the curve are created by sliding both event identification thresholds simultaneously. They showed that this can result in a STONE curve that varies like the ROC curve from (1,1) to (0,0) but is less restricted in its path between these endpoints. Specifically, the STONE curve does not have to be monotonic but might double back on itself in either the *x* or *y* axis direction. This is because all of the data-model paired points in the scatterplot begin in the "hits" quadrant of the contingency table when both thresholds are set very low but end in the "correct negatives" quadrant when the sweep is done and both thresholds are set very high. In between, the points usually pass through the "misses" or "false alarms" quadrants along the way as the thresholds are changed. This leads to the misses and false alarms cell counts increasing and decreasing throughout the threshold sweep, possibly resulting in times where the POD or POFD metrics temporarily increase.

Liemohn et al. (2020) showed two space weather examples of the usage of the STONE curve. The resulting nonmonotonicities were qualitatively interpreted as intervals when clusters of points were quite far from the "ideal fit" diagonal line through the data-model scatterplot. That study hinted that the size of the nonmonotonic feature in the STONE curve could be related to the size or location of the cluster of overestimated or underestimated values.

In this study, a systematic quantification is conducted of the relationship between features of the STONE curve and features of the data-model scatterplot. This is done by imposing known features into the scatterplot, varying the magnitude of the nonideal aspects of the distribution and assessing the impact on the resulting STONE curve. The methodology for creating these distributions and analyzing their features relative to the STONE curve is described in section 2. Section 3 presents these systematic variations and notes the relationships, while section 4 is a discussion of how to interpret these features. In section 5, the newfound quantitative relationships of the STONE curve to scatterplot features is applied to real-time space weather model results from the Inner Magnetosphere Particle Transport and Acceleration Model (IMPTAM), compared with satellite data, and used in conjunction with other data-model comparison metrics. A summary is provided in section 6.

2. Methodology

Our main method of analysis for this study is the creation of idealized distributions with randomly assigned points in both the *x* and *y* axis directions. The distributions will be known and therefore the appearance of features in the STONE curve can be systematically quantified against the imposed features of these distributions. All distributions are created using the skew norm distribution of Azzalini & Capitanio (1999), as implemented in the scipy.stats package within Python. Each distribution to be analyzed is constructed with 2000 paired data-model points per scatterplot, defined with a linear relationship confined to the zero-to-one range along the *x* axis. The full data set is created by concatenating 10 subsets of 200 points each, each with a uniform width in the *x* axis direction. The points along that axis are randomly distributed within each narrow range, while the values in the other axis are set with a random sampling from a Gaussian distribution with a specified mean and standard deviation relative to the unity-slope, zero-offset "perfect fit" line.

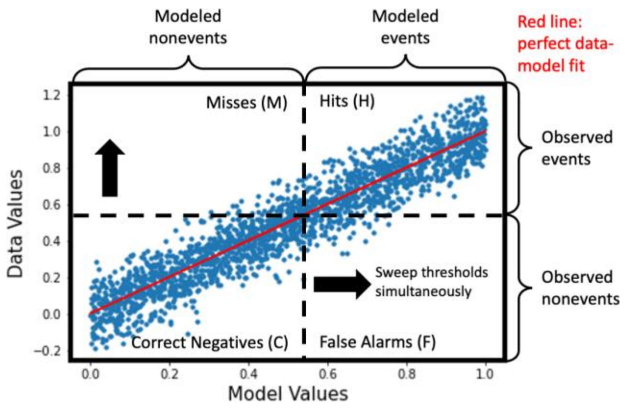

Figure 1. Data-versus-model scatterplot with event identification thresholds drawn. The arrows indicate the sweep conducted to generate a collection of contingency tables as a function of threshold setting, moving both thresholds simultaneously.

Figure 1 shows an example scatterplot, created with random values along the *x* axis and a Gaussian distribution in the *y* direction of spread 0.1 that follows the *y* = *x* perfect data-model fit (shown as the red diagonal line on the plot). All of the *x*-axis "model values" are contained within the [0,1] range; the Gaussian spread in the *y*-axis "data values" yields some points that are below zero, especially at low *x* values. Both number sets have a mean of 0.500 and the root-mean-square error (RMSE) between them is 0.099, a score very close to the imposed spread of 0.100.

Two event identification threshold lines are also drawn in Figure 1 (as black dashed lines), one for the model values and the other for the data values. These two thresholds divide the scatterplot into quadrants, labeled in Figure 1 as hits (*H*), misses (*M*), false alarms (*F*), and correct negatives (*C*). The contingency table is created by simply counting the points within each quadrant. In this example, most of the points are in the two correct cells (*H* and *C*), with very few points in the two error cells (*M* and *F*).

In the creation of the STONE curve, the two thresholds are swept simultaneously from very low to very high values. As the sweep continues, the cross-over point of the two thresholds will always occur at a value along the red perfect fit line. With each increment of the threshold setting, some points will move from *H* into the other quadrants. A few points that are very close to the red perfect fit line will jump directly from *H* into *C*, but most will pass first through an error cell of *H* or *F* and then on to *C* at some higher threshold setting.

Figure 2 shows the quadrant counts for *H*, *M*, *F*, and *C* as a function of the threshold setting. The sweep is conducted with a step size of 0.01. The threshold sweep starts well below the range for either the data or model number sets, so at the very low settings, all points are in the *H* cell. The sweep extends beyond the top of both ranges, so at the very high settings, all of the 2000 points in the scatterplot are in the *C* quadrant. In between, *H* decreases monotonically and *C* increases monotonically, but *M* and *F* rise and fall as points enter from the *H* cell and leave to join the *C* cell. Because this example scatterplot has a rather tight spread around the perfect fit line, the counts for the *F* and *M* quadrants are never a large fraction of the total. Near zero, *F* is larger than *C*, and near one, *M* is larger than *H*.

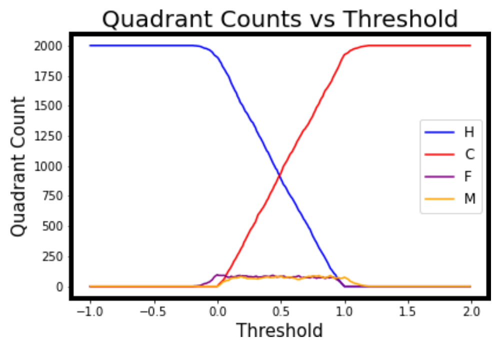

Figure 2. Quadrant counts for H, C, F, and M as a function of sweeping the data and model event identification thresholds simultaneously through the example scatterplot shown in Figure 1.

Metrics can be calculated from the resulting quadrant counts. For the STONE curve, the two metrics to be plotted against each other are POD, defined as hits over observed events:

$$POD = \frac{H}{H+M} \tag{1}$$

and POFD, defined as false alarms over observed nonevents:

$$POFD = \frac{F}{F+C} \tag{2}$$

The resulting STONE curve is shown in Figure 3a. As a reference to help the interpretation of this plot relative to the two above it, red dots are included every 0.1 along the threshold sweep. The POD and POFD values as a function of threshold are shown in Figure 3b. Because the scatterplot is fairly tight along the perfect fit line, these two curves are mostly monotonic, but not entirely. There are small intervals where one or the other of these two metrics increase during the upward sweep of the thresholds. The small increases in POD and POFD seen in Figure 3b are barely visible in the STONE curve in Figure 3a. With 2000 points in the number set and a few points in the F quadrant at a threshold setting of zero, on average there are roughly 19 points moving out of the *H* quadrant at each of the threshold increments between zero and one. About half of these move to *M* and the other half moving to *F*, with perhaps one or two converting directly to the *C* quadrant. A similar number is being converted out of *M* and *F* each threshold step. Poisson counting uncertainty dictate that there could be small fluctuations, on the order of 3, in the exact number of points moving between the quadrants. It happens occasionally, then, that this Poisson noise results a larger number of points converted out of *M* than into *M*, which would cause an increase in POD. A similar situation could arise for *F*, resulting in a very small increase in POFD. For this particular number set, the maximum increase in POD is 0.0094 and the maximum increase in POFD is 0.0038, with the majority of the increases below 0.002. Increases on the order of Poisson noise fluctuations are not significant and should not be interpreted as a meaningful feature of the STONE curve.

Figure 3a shows that the STONE curve comes very close to the upper left (0,1) corner of POFD-POD space. This indicates that these imposed *x*-axis “model” values are very good at sorting the *y*-axis “observations” into events and nonevents, regardless of the event threshold setting. It is well above the pink-dashed unity-slope line, drawn for reference (here and on all of the STONE curve plots below) to provide a comparison against the case when the model is equivalent to random chance.

Several additional metrics will be included in the analysis below. Because any single metric is designed to assess a specific aspect of the data-model relationship, several metrics are needed to fully quantify the goodness of the fit between two number sets. Categories for metrics have been defined by Murphy (1991), and a mapping of many event detection metrics to these categories has been provided by Liemohn et al. (2021). An accuracy metric is useful for determining the overall goodness of the fit between the two number sets. The F_1 score will be used in this study:

$$F_1 = \frac{2H}{2H + M + F} \tag{3}$$

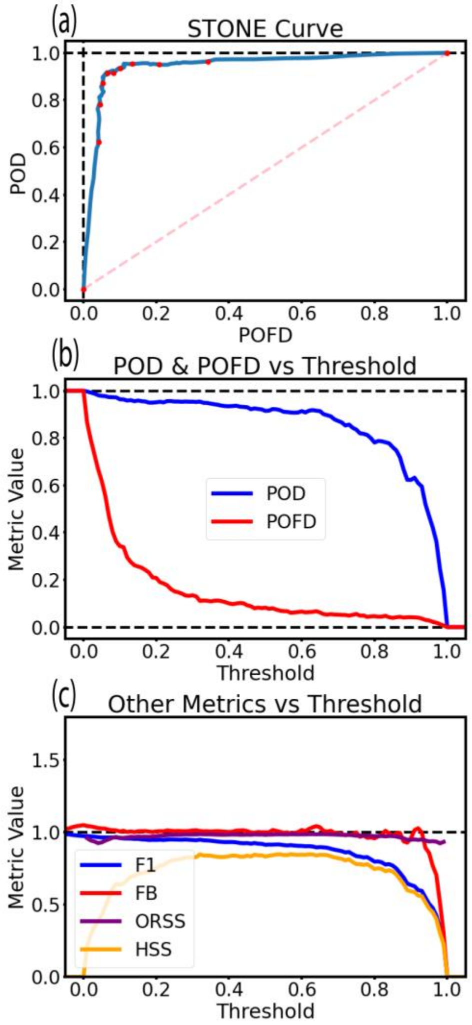

Figure 3. For the example scatterplot shown in Figure 1 with simultaneously threshold sweeps, (a) the STONE curve, with a red dot shown every 0.1 along the threshold sweep; (b) the POD and POFD scores as a function of threshold setting; and (c) four other metric scores, also as a function of threshold setting.

At the lowest threshold settings, everything is a hit, so F_1 will be one, its perfect score. As the thresholds sweep to higher values, hits are converted to either misses or false alarms, and F_1 will drop. This decrease does not have to be monotonic, however; it could increase if there is a cluster of points that leave the *M* or *F* quadrants for the *C* quadrant. At the highest threshold setting, it usually drops to zero when *H* = 0 and then becomes undefined when all points are in the *C* quadrant. For example, an F_1 of 0.5 could be achieved with H equal to the average of M and F while a score of 0.67 could be attained with H equal to the sum of M and F.

Accuracy metrics are nearly always symmetric, comparing the point count in H (perhaps also with C) against the combined value of $M + F$, all points in the error cells. To understand the asymmetry of the contingency table, a metric from the bias category is needed. For this study, frequency bias, FB, will be adopted:

$$FB = \frac{H + F}{H + M} \tag{4}$$

This metric compares the points with the model value in its event state to the points with the observed value in its event state. The ideal value for FB is one, with larger value indicating that the model overpredicts events and smaller values indicating an underprediction of events. The H is both the numerator and denominator acts to mitigate the influence of small but different F and M counts; if H is much larger than both error cell counts, then FB will be close to one regardless of the imbalance between F and M. A value of FB of 0.75 can be arrived at if $H = F = 2M$, while a score of 1.33 could be from $H = M = 2F$.

Another useful category to include in the analysis is association, which in the case of event detection metrics is assessing the balance of the contingency table and how well that balance favors the two good quadrants. We will use the odds ratio skill score, ORSS, which is typically written in this form:

$$ORSS = \frac{(H \cdot C) - (F \cdot M)}{(H \cdot C) + (F \cdot M)} \tag{5}$$

ORSS varies from a perfect score of +1 to a worst-case score of -1, with scores above zero indicating that the model is better than random chance. If the H times C product is double the value of the F times M product, then ORSS will be 0.33. If H and C are equal and double the values of F and M (also equal), then this combination yields ORSS = 0.6.

The final metric to be considered in this analysis is the Heidke skill score, HSS. Skill scores compare a metric score of the data-model comparison against that same metric for a reference model. In the case of HSS, the metric is “proportion correct” and the reference model is random chance, as given by the expected values for the contingency table cells given the same column and row totals. The formula for HSS is:

$$HSS = \frac{2[(H \cdot C) - (F \cdot M)]}{(H + M)(M + C) + (H + F)(F + C)} \tag{6}$$

If $F = M = 0$, then HSS will be one, its perfect score. If $H = C = 0$, then HSS reverts to $-2FM/(F^2+M^2)$, which is either zero if one or the other of F or M is zero and drops to its lowest value of -1 if F and M are equal. Any HSS score greater than zero indicates that the model is better than random chance. While this is sometimes taken as the threshold for a good HSS value, it is a relatively low bar to satisfy. If $H = C = 2F = 2M$, the case of a well-balanced contingency table with hits equal to the sum of the error cell counts, then HSS = 0.33.

These four additional metrics will be reported along with POD, POFD, and the STONE curve to assess the connection between known features in the scatterplot and calculated signatures in the metric values. They are shown in Figure 3c for the example distribution being considered in this section. At a threshold setting of zero, nearly all of the points are in the *H* quadrant, a few (those with negative *y* values) are in the *F* quadrant, and *M* = *C* = 0. This results in F_1 very close to one, FB slightly larger than one, an undefined ORSS, and HSS = 0. The values for these 4 metrics are close to one for most of the threshold sweep, until the threshold approaches a setting of one, in which case three of the four metrics plunge to zero. At a threshold setting of one, nearly all of the points are in the *C* quadrant, a few are in the *M* quadrant (those with *y* values above one), and *H* = *F* = 0. For these values, F_1, FB, and HSS are all zero and ORSS is undefined.

To conduct an assessment of how the STONE curve relates to features in the underlying scatterplot, two parameters are adjusted to this baseline data-model number set collection. The first is the spread of the distribution around the perfect-fit line, which will be systematically increased in either all or part of the *x* domain. The second parameter is the deviation of the local mean of the data minus model error distribution away from the perfect fit line. This change will be made for specific intervals of the *x* domain.

3. Results

Here we present the resulting STONE curves from the systematic variation of the data-model scatterplots. In all of the plots below, 2000 data-model pairs are used, with a threshold step size of 0.01. For each threshold setting, the points in each quadrant are counted, a contingency table is created, and the metrics listed above are calculated.

3.1. Variations in spread

Figure 4 shows a set of distributions with different settings for RMSE between the *y*-axis “data” and *x*-axis “model” number sets. The same RMSE is applied across the full (0,1) *x* domain. The top row presents the scatterplot, the second row the resulting STONE curve, the third row shows the underlying POD and POFD curves used to make the STONE curve, and the fourth row presents several other data-model comparison metrics.

It is seen that the STONE curves are very close to a perfect value in the upper-left corner (see Figure 4b), but pull away from this ideal as RMSE is increased. None of the STONE curves, however, include significant nonmonotonic features. This is revealed by the nearly monotonic curves of POD and POFD; while some very small increases are seen in every curve due to Poisson counting noise, the POD and POFD curves steadily decrease throughout the threshold sweep from low to high values. For the largest RMSE case, the POD and POFD curves (see in Figure 4o) lack the steep slopes seen for the other RMSE settings, indicating that this spread is seriously degrading the quality of the data-model comparison. The POD values are still larger than the POFD values for all threshold settings, though, so the STONE curve in Figure 4n is above the unity-slope “random chance” reference line.

The additional metrics in the lower row are shown for context. When the STONE curve is very close to the upper-left corner, all four of the chosen metrics are close to one for most of the zero-to-one threshold setting range. As the RMSE increases, these metrics worsen in some or all of the threshold setting range. For example, for the smallest RMSE setting used in Figure 4, HSS peaks at 0.93 (seen in Figure 4d), while for largest RMSE, HSS only reaches a maximum value

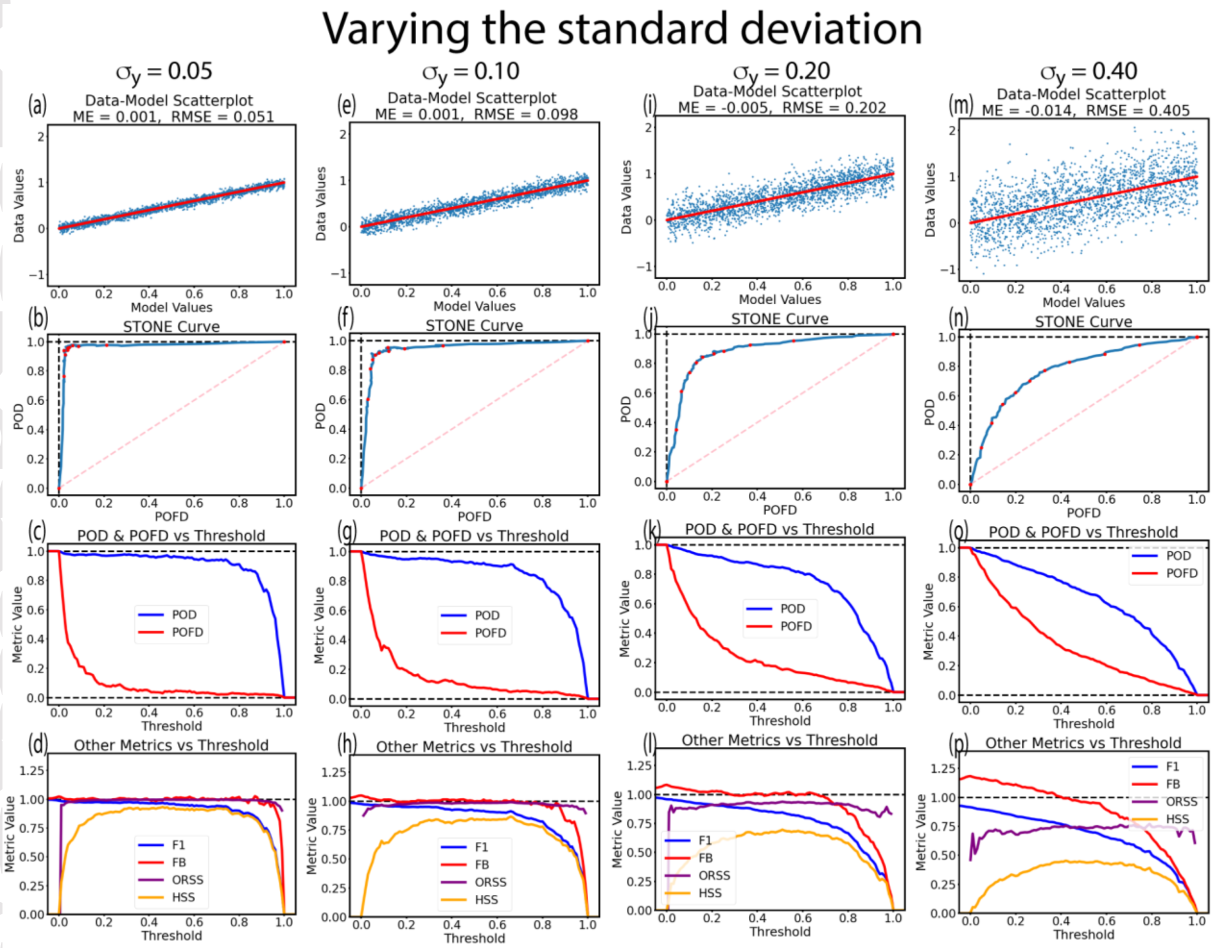

Figure 4. Scatterplots (top row), the corresponding STONE curves (second row), its underlying POD and POFD values as a function of threshold (third row), as well as several other metrics versus threshold (fourth row). The formats of the plots are identical to those in Figures 1 and 3. The four columns have data-model errors with Gaussian distributions with no offset but different spreads, as indicated.

of 0.45 (Figure 4p). This is still a number indicating substantial skill relative to random chance, but the interpretation of such a value for HSS depends on the specific data-model comparison being conducted.

Figure 5 shows a slightly different case, in which RMSE is only increased at the high end of the *x* domain. To create these distributions, the *x* domain was segmented into 10 equal intervals, each with 200 randomly distributed values. The corresponding *y* values are a Gaussian spread around the *x* axis value, with an imposed RMSE of 0.1 for the first 6 bins and then incrementally increasing the RMSE in the remaining 4 bins. For the left column, the increase increment is 0.01, so the final *x*-axis bin has an imposed RMSE of 0.14. The second column has an increment of 0.02 (maximum RMSE in the last bin of 0.18), the third column has an increment of 0.03 (maximum RMSE of 0.22), and the fourth column has an increment of 0.04 (maximum RMSE of 0.26). The panels of Figure 5 are in the same format as those of Figure 4.

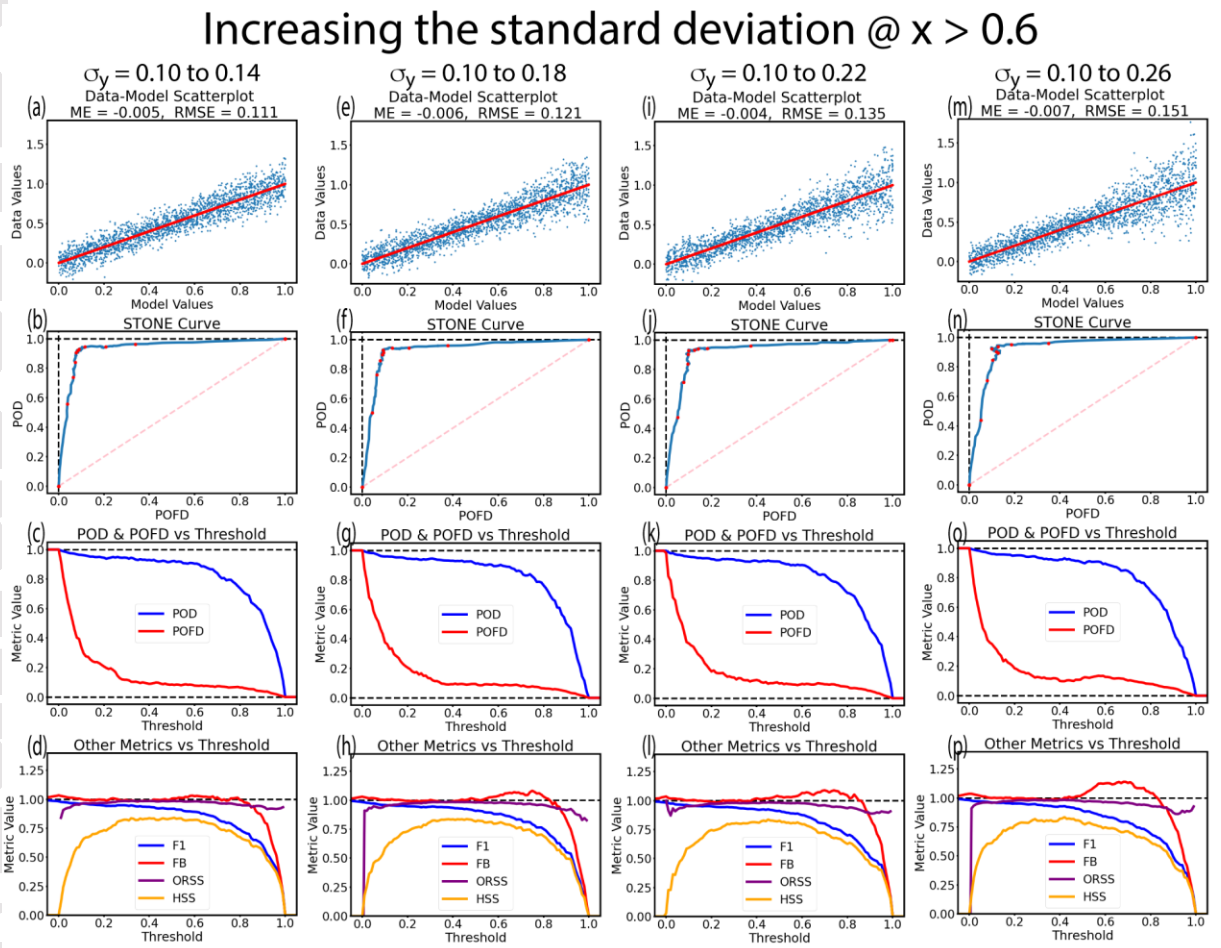

Figure 5. Like Figure 4, but systematically varying the width of the distribution but only for *x* > 0.6. All distributions have σ = 0.10 for x < 0.6, then increase within each 0.1 interval by a given increment: left column by 0.01, second column by 0.02, third column by 0.03, and right column by 0.04.

For this group of distributions, the STONE curves in the second row show the progression from monotonicity to a curve containing a nonmonotonic wiggle. Here, a "wiggle" is defined as a statistically significant increase in the *x*-axis value, POFD, while the *y*-axis value, POD, continues to decrease. That is, a wiggle is a left-to-right oscillation in the STONE curve. This is the case for the two column on the right, seen in Figures 5j and 5n. These increases are seen in the POFD values displayed in Figures 5k and 5o. The wiggle is very subtle in Figure 5j, but it exists for a relatively large number of threshold steps. For a threshold setting of 0.58, the POFD value in Figure 5j is 0.089; at a threshold of 0.67 (9 steps later), POFD has risen to 0.102. This increase is larger than the Poisson noise fluctuations and indicates a response to a real feature in the relationship between the two number sets. The increase in POFD is even more dramatic in Figure 5o, rising from a relative minimum of 0.097 at a threshold of 0.41 up to a relative maximum of 0.135 at a threshold of 0.57. This results in a small but noticeable wiggle in the STONE curve, just at the moment of its closest approach to the upper-left corner.

The wiggle can be related to features in other metric values as a function of threshold setting. It is particularly seen in FB, which increases slightly above unity in the vicinity of the wiggle. As seen in equation (4), FB includes *F* in the numerator, indicating that at these threshold settings, there is an imbalance in the contingency table between the two error cells, specifically an excess of *F* counts relative to *M*. As points preferentially move from *H* to *F* (instead of equally to *M*), FB and POFD systematically increase. As the threshold continues to increase, eventually these points will move from *F* to *C*, and both FB and POFD will be reduced. The other metrics, in particular ORSS and HSS, show slight downward kinks beginning at the same threshold as the increase in FB and POFD.

3.2. Variations in both spread and offset

Another test is to not only vary the spread but also impose a slight shift of the bias between the *y* values relative to the *x* values. The plots from this experiment are shown in Figure 6. The distributions are constructed in the same manner as those in Figure 5, but in addition to an incremental increase in RMSE for the last 4 *x*-axis bins, an increment shift in the mean value between the 200 *x* and *y* values in that bin is also imposed. Because Figure 5 reveals a slight wiggle for an incremental RMSE change of 0.03, this RMSE increment is imposed for all of the distributions in Figure 6. The bias increment, always downward for this set, is varied from -0.1 in the first column (for a maximum offset of -0.4 in the final *x*-axis bin) up to an increment of -0.4 (for a maximum offset of -0.16).

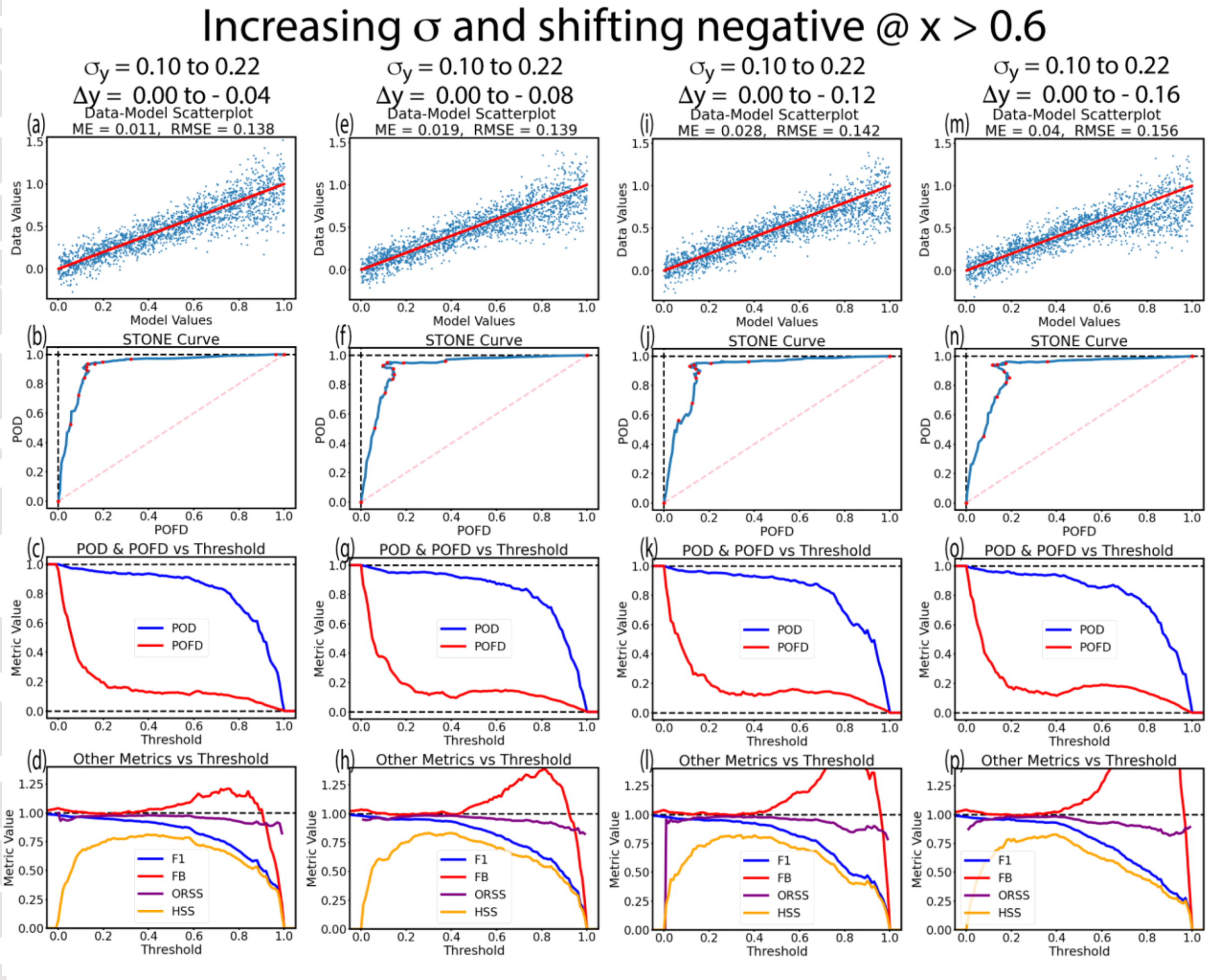

Figure 6. Like Figure 5, but also varying the offset of the Gaussian error distribution away from the unity-slope line. All distributions have σ = 0.10 and Δy = 0 for x < 0.6, then increase the spread within each 0.1 interval by 0.03. Each column has a different negative Δy shift at high *x*: left column by 0.01 bin , second column by 0.02, third column by 0.03, and right column by 0.04.

The wiggle in the STONE curves is visible in every panel of the second row of Figure 6. For the smallest offset increment, the STONE curve (seen in Figure 6b) wiggle is small and the increase in POFD occurs near a threshold setting of 0.6 (seen in Figure 6c). For the other three bias increment settings, the STONE curve wiggle is clear, with the POFD increase beginning at a threshold setting around a value of 0.4. This is because the points in the final two x-axis bins have a spread and bias setting that allows some points to be at y values at low as 0.4. This begins the imbalance of the conversion of points out of the *H* quadrant, now favoring *F* over *M*.

The additional metrics shown in the lower row reflect this imbalance of *F* over *M*. It is most clearly seen in the FB metric, peaking at a value of 1.87 for the largest imposed offset increment (Figure 6p). As seen in Figure 5, the other metrics have a downward change in slope at the same threshold setting as the initial increase in FB and POFD. Before this downward trend, though, the metrics have very good scores because the imposed spread is small for the left section of the distribution.

Figure 7 shows a very similar experiment as that shown in Figure 6 but this time imposing a positive bias between the *y* and *x* values in the four highest *x*-axis bins. Exactly the same settings are used for this set of distributions, with a 0.1 spread for *x* below 0.6, then a 0.3 RMSE setting for *x* greater than 0.6. The offsets are incremented in these bins of increased spread, with imposed increments of +0.01, +0.02, +0.03, and +0.04 for the four distributions, respectively.

In this case, only the first distribution has a STONE curve with a very subtle but statistically significant wiggle. From Figure 7c, at a threshold setting of 0.48, POFD is 0.092; it then rises to 0.107 at a threshold of 0.58. This is a similar feature to what was seen in the third column of Figure 5. The other three STONE curves in the second row of Figure 7 have no significant features beyond Poisson noise fluctuations. The metrics in the last row of Figure 7 reflect this subtle or nonexistent feature set in the STONE curves. In Figure 4d, the first distribution with the smallest imposed offset has a slight increase in FB. All of the distributions, though, have an FB curve that drops below unity at lower *x* values than previously seen in Figures 4 – 6. The other three metrics (F1, ORSS, and HSS) have nearly identical curves for the four distributions.

The distributions used in Figure 7 are included to illustrate the point that not all offsets result in nonmonotonic features in the STONE curve. This set has an offset that is positive, so the increased spread at large *x* does not result in additional points in the *F* quadrant. They remain in the *H* quadrant until the final threshold steps of the sweep. In fact, the F quadrant has a reduced count for high *x* values for these distributions, causing the early downward shift in FB. The upward shift of the distributions does not, however, result in an increased count in M until the very last threshold steps of the sweep. So POD never undergoes an increase for these distributions. In short, this upward offset at high *x* values is not revealed by the STONE curve.

As another test case to further reveal features of the STONE curve, Figure 8 shows a set of distributions that are essentially the inverse of those used in Figure 6. That is, an increased spread (RMSE of 0.3) and upward offset are applied at low *x* values (below 0.4), while the distribution above *x* = 0.4 has no offset and an imposed RMSE of 0.1 (the nominal case). The offset values are incremented the same as in Figure 6, but now upward so the distributions remain more within the (0,1) range in the *y* values. The offset is largest in the lowest *x*-axis bin.

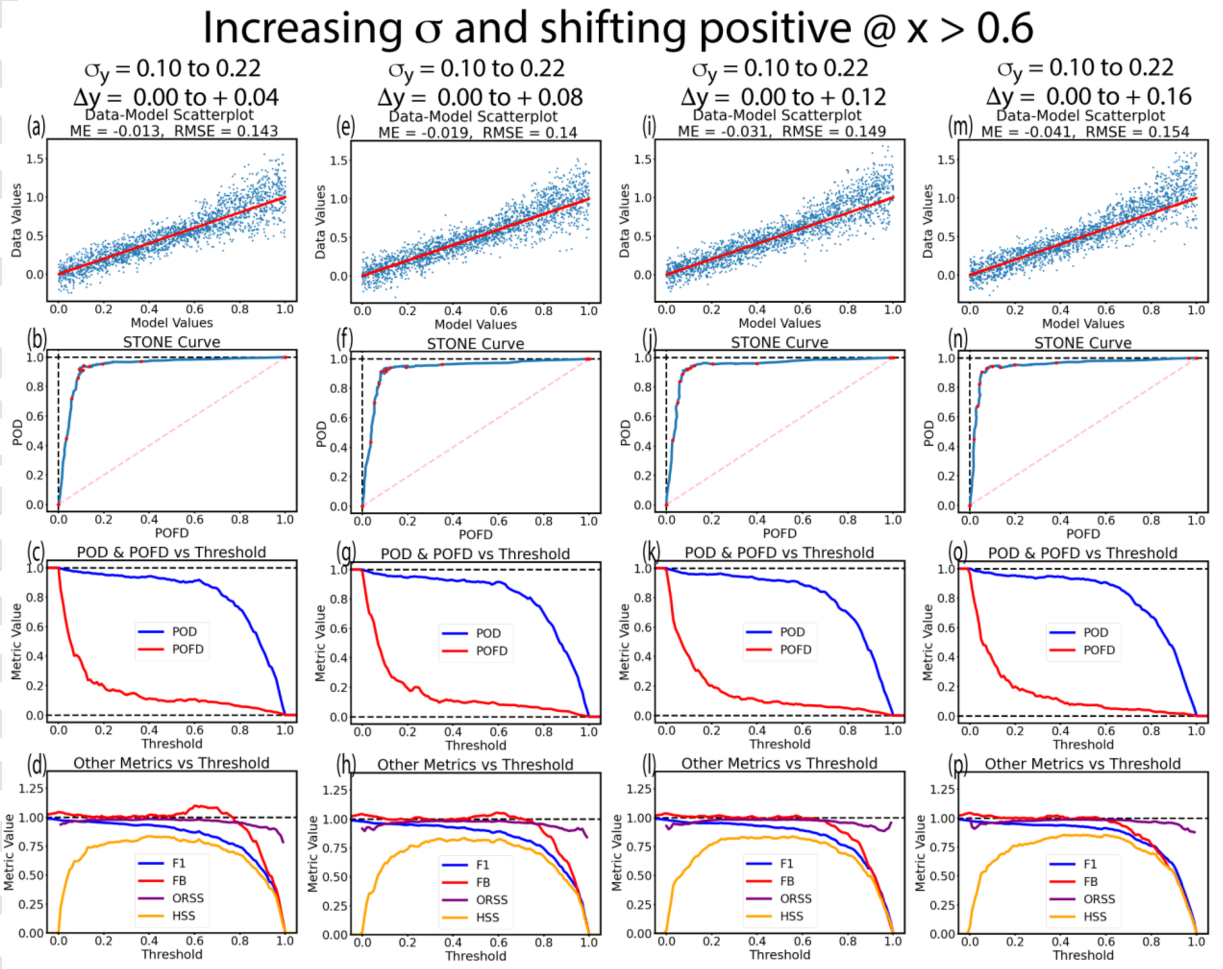

Figure 7. Like Figure 6 except with positive shifts in the Gaussian peak relative to the unity slope line.

The STONE curves in the second row of Figure 8 all include a ripple feature. A "ripple" is defined here as a statistically significant increase in the *y*-axis value, POD, while the *x*-axis value, POFD, continues to decrease. More plainly, a ripple is an up-and-down oscillation in the STONE curve. The POFD curves in the third row of Figure 8 include small fluctuations due to Poisson noise in the distributions but the increases in POD that is seen in these plots exist over a larger span of thresholds and reveal an important feature of the underlying data-model comparison. In particular, they show that there is a cluster of points in the *M* quadrant that are being quickly converted into the *C* quadrant, faster than new points are entering the *M* quadrant. Remembering equation (1), this causes a systematic decrease in the denominator of POD and therefore an increase in this metric. The larger the spread and upward offset in the points at low *x* values, the larger the cluster in the *M* quadrant that is not removed until high threshold settings, resulting in a clear increase in POD and therefore a ripple in the STONE curve.

For the first distribution with the smallest imposed offset, the ripple is subtle. The POD curve in Figure 8c reaches a relative minimum value of 0.871 at a threshold of 0.41 and a relative maximum of 0.899 11 bins later at a threshold of 0.52. The change in POD for the next-largest offset increment (Figure 4f) is already more clearly seen, with a relative minimum of 0.867 at a threshold of 0.34, rising to a POD value of 0.901 at a threshold of 0.58. Because it occurs over a longer interval of threshold settings, the ripple is more apparent in the STONE curve of Figure 8f than the one in Figure 8b. The largest setting for the imposed offset has a very clear ripple, with a POD change of 0.051 from relative minimum to maximum, but the span of *x* values over which this occurs is nearly identical to the other cases in this set.

The additional metrics shown in the last row of Figure 8 are somewhat different than their counterparts in earlier figures. The F_1 metric scores decrease sooner (at lower threshold settings) than earlier cases, although they remain quite good through most of the threshold sweep. The FB scores drop below unity at low thresholds, then increase back to unity for the second half of the sweep before plummeting to zero as all points are converted to *C*. ORSS appears to be a mirror image of its shape in previous plots, being reduced at the low threshold settings and better at the higher settings. Finally, HSS also appears to be a mirror image of earlier plots, with lower values at low thresholds and higher values at higher thresholds. All of these features are caused by the imposed difference between the *y* and *x* values at low *x*; once the

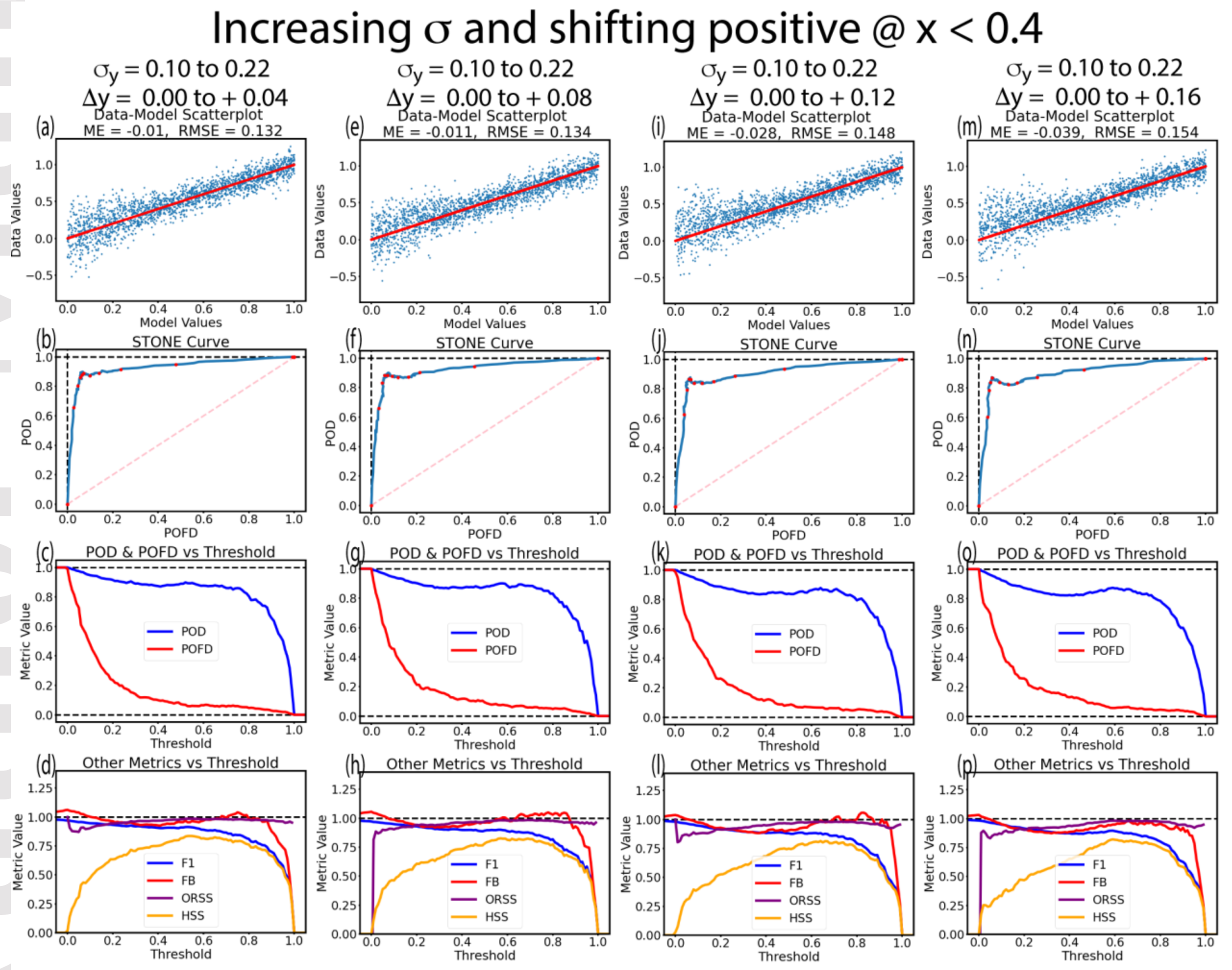

Figure 8. Like Figure 6 except with the increasing spread and positive shift of the error distribution at low *x* value, below 0.4.

thresholds increase to a point where these points with large differences are all contained within the *C* quadrant, they no longer influence the metrics, which return to their values from the nominal distribution case presented in Figure 3.

3.3. Variations in offset

The two sections above show that an offset is more effective than an increased spread at creating a feature of interest in the STONE curve. To further examine this relationship, the RMSE of the *y*-value Gaussian distribution will held constant at 0.1 and systematic offsets at different *x* ranges will be imposed.

The first of these assessments is shown in Figure 9. In the creation of these distributions, three of the *x*-axis bins are given negative *y*-value offsets. The 200 points in the *x* = [0.7, 0.8] interval are defined with the maximum offset (-0.025, -0.05, -0.1, and -0.2 for the four columns in Figure 9, respectively), and the two *x*-axis intervals on either side of this (the [0.6, 0.7] and [0.8, 0.9] intervals) are given half of that bias. The fourth distribution, with an offset of twice the spread in the *x* = [0.7, 0.8] interval, has very few points above the unity-slope line in this *x* range.

The STONE curves for this set of distributions are shown in the second row of Figure 9. Three of the four have a feature that appears as a downward-to-the-right diagonal excursion of the STONE curve that returns nearly along itself. This feature is especially evident in the STONE curves from the larger offsets but is present even in the curve for the smallest offset. As seen in the POD and POFD curves plotted in the third row of Figure 9, the two features of wiggle and ripple are both occurring in these STONE curves.

Consider the most obvious example, the largest offset setting, and the POD and POFD curves in Figure 9o. At lower thresholds, POFD quickly drops to a relative minimum of 0.120 at a threshold setting of 0.36. As the threshold rises from this level, a few points from the downward-shifted distribution at high *x* values move from the *H* to *F* quadrant. POFD then steadily rises to a relative maximum of 0.209 at a threshold of 0.56, after which it then decreases towards zero. This rise in POFD creates a wiggle in the STONE curve. The POD curve appears to be making a steady decline from one towards zero, but stops its descent at a threshold of 0.62 (at which point POD = 0.850). It then increases to a new maximum of 0.996 at a threshold of 0.82, after which it quickly drops to zero. This rise in POD is caused by the lack of points above the unity-slope line in the *x* = [0.7 0.8] interval; very few points shifted from *H* to *M*, yet many left *M* for *C*. This yields a smaller denominator for POD and the metric increases, creating a ripple in the STONE curve. The POD and POFD curve increases are less pronounced in Figures 9g and 9k, but still exist and are large enough to cause a noticeable feature in the corresponding STONE curves of Figures 9f and 9j. The combination of a wiggle immediately followed by a ripple forms the nearly diagonal line excursion in the STONE curve.

Examining Figure 9b, the STONE curve from the distribution with the smallest imposed shift, it is seen that it has a small feature of this same type, occurring between threshold settings of 0.6 to 0.8. The change in POD and POFD are both 0.010, just a bit more than the maximum swing caused by Poisson noise (as evaluated above in Figure 3). This is, therefore, a significant feature of the same type and origin as seen in the other STONE curves of Figure 9, but the increases are difficult to distinguish in the POD and POFD curves of Figure 9c because they are so close to the noise level.

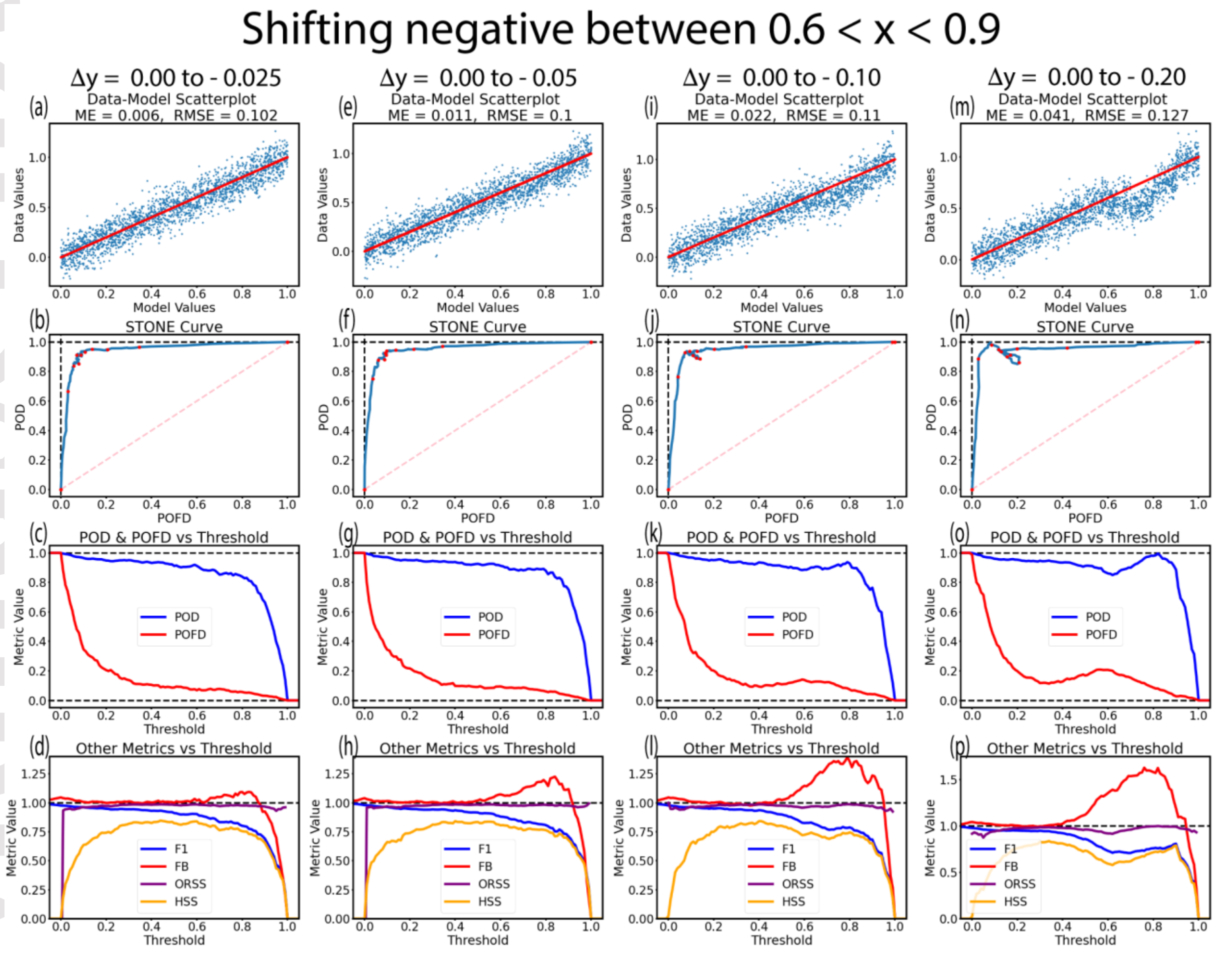

Figure 9. Like Figure 4 but with a constant spread of 0.1 and a negatively shifting bias to the error distribution in three *x*-axis bins (between 0.6 and 0.9), two halfway and the middle the full shift. From left to right, the maximum shifts are -0.025, -0.05, -0.10, and -0.20.

The last row of Figure 9 shows how this diagonal-line feature in the STONE curve relates to other metrics. At the start of the diagonal excursion, where POFD begins its increase, all of the chosen metrics worsen, with F_1, ORSS, and HSS starting downward trends and FB increasing away from unity. When the STONE curve starts its return along the diagonal – when POFD starts to decline and POD starts to increase – it is seen that $F_{1,}$ ORSS, and HSS reverse their trend and recover somewhat. FB continues to rise away from unity, though, because the decrease in the *F* count is slower than the decrease in the *M* count.

An analogous test can be conducted for a systematic positive shift in the distribution. This is shown in Figure 10. For optimal effect on the STONE curve, the shift is imposed at low *x* values, so the *y* values stay mostly within the (0,1) range. The largest offset is applied in the *x* = [0.2, 0.3] interval, with the two neighboring intervals given half of the full bias. The amplitudes of the shifts are the same size as those imposed for the distributions in Figure 9.

The resulting effect of a positive shift on the STONE curves presented in the second row of Figure 10 is to create the same diagonal excursion as seen in Figure 9. The main difference with the previous set is that the excursion occurs at an lower threshold setting. For the smallest offset , the increase is only significant for POFD, not for POD, so the STONE curve in Figure 10b only exhibits a small wiggle near a threshold of 0.3. For the next column, the STONE curve in Figure 10f contains a diagonal excursion but, like Figure 9b, it is barely above the Poisson noise level. The diagonal excursion is more visible in the STONE curves of the next two examples, Figure 10j and 10n. For these distributions, the POFD curves reach a relative minimum near a threshold of 0.2, rise to a peak near a threshold of 0.4, at which POD reaches a relative minimum and begins its ascent to a maximum near a threshold of 0.6.

The other metrics plotted in the final row of Figure 10 only have substantial features for the largest of the imposed shifts (Figure 10p). Because the shift is positive, the *M* quadrant has excess points compared to *F*, so FB drops to values below unity. Even for the large setting, though, FB never attains a poor score; it is always above 0.8 in the threshold region of the diagonal excursion of the STONE curve. The HSS curve in Figure 10p is a nearly perfect mirror image of the HSS curve in Figure 9p.

The final assessment to conduct with shifting the distribution is shown in Figure 11, in which two shifts are applied. The procedure is a superposition of the two techniques used for the

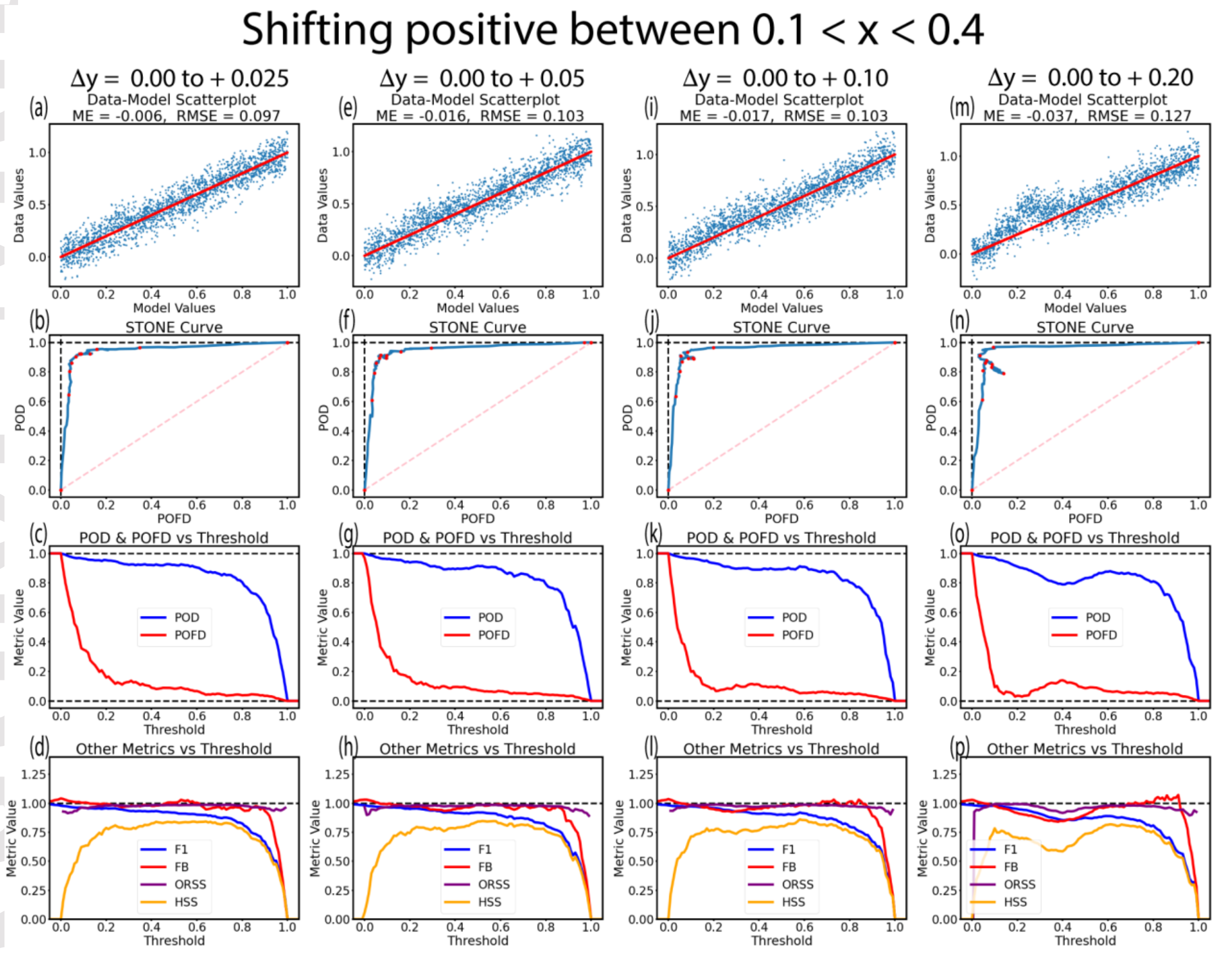

Figure 10. Like Figure 9 but with a positive shift in 3 *x*-axis bins (between 0.1 and 0.4).

distributions in Figure 9 and 10, offsetting the distributions by a maximum of 0.2 in the *y* direction in two bins (both *x* = [0.2, 0.3] and *x* = [0.7, 0.8]), in either the upward or downward direction.

The first column of Figure 11 has a downward shift at both low and high *x* values. At low threshold settings, a few points pass through the *M* quadrant but, due to the downward shift in the x = [0.1, 0.4] interval, the points in the M quadrant are evacuated and not replaced. This yields a relative minimum in POD near a threshold of 0.1 (Figure 11c) and results in a shallow ripple in the STONE curve (Figure 11b). As the threshold passing a setting of 0.4, it is now completely past the low-x-value downward shifted part of the distribution (all of those points are not in the C quadrant). From here, the STONE curve mimics that in Figure 9n. The downward shift at in the high x range then contributes a large number of points into the F quadrant, causing a rise in POFD and therefore a wiggle in the STONE curve. This is immediately followed by an increase in POD and the return from the wiggle has a superimposed ripple, making the STONE curve retrace its diagonal excursion. The other metrics for this distribution, shown in Figure 11d, resemble those of Figure 9p with the added features at low threshold values of a reduced ORSS and an elevated FB.

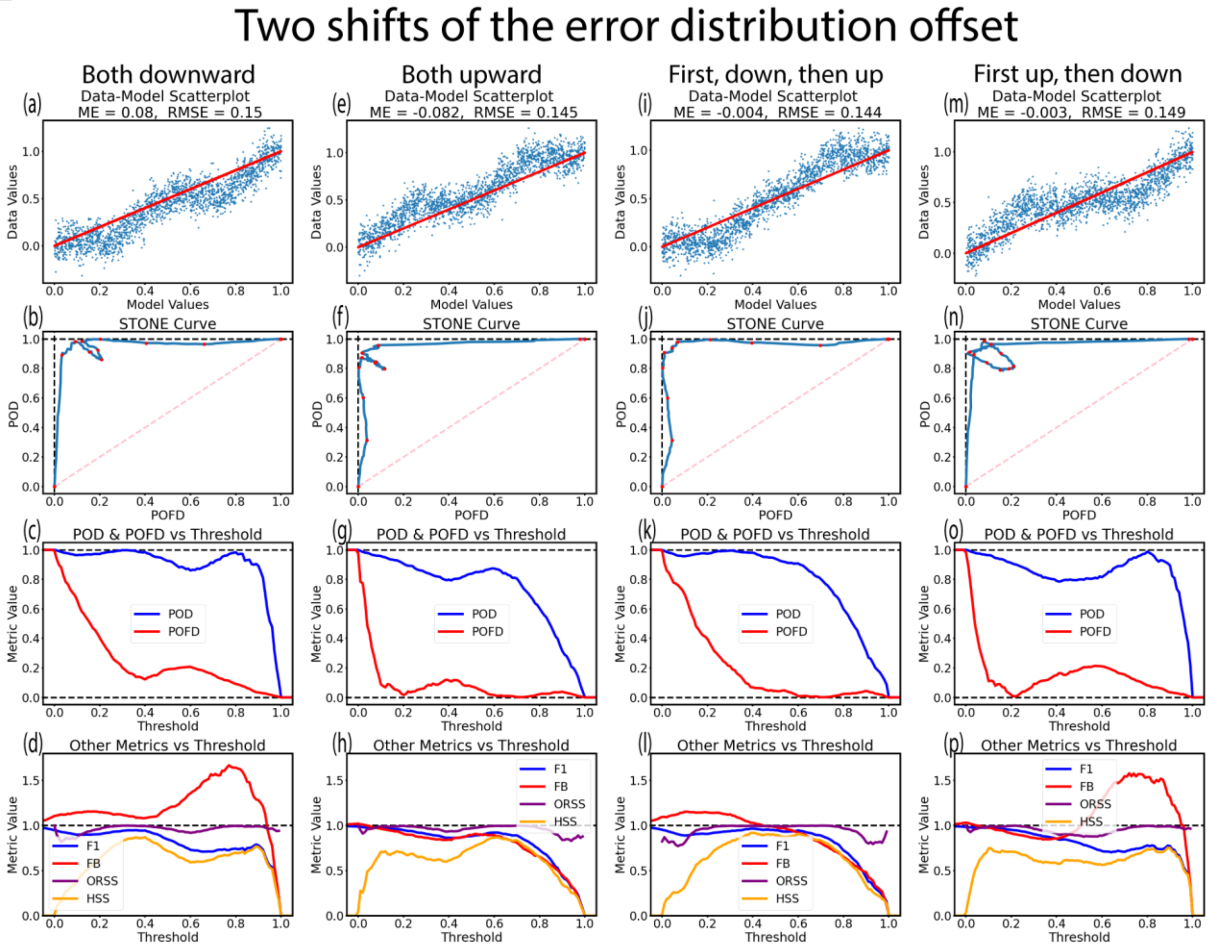

Figure 11. Like Figure 9 but with two shifts in 3 x-axis bins (one between 0.1 and 0.4, and another between 0.6 and 0.9).

The second column of Figure 11 presents the case of two upward shifts in the distribution relative to the unity-slope line. Because of the upward shift at low *x* values, the first half of the STONE curve in Figure 11f resembles that of Figure 10n. At higher thresholds, however, the second upward shift causes a dearth of *F* quadrant counts, with a nearly perfect POFD (equal to 0.002) at a threshold setting of 0.73. At the end of the threshold sweep, the return of the distribution to being centered on the unity slope line causes an increase in *F* quadrant counts and therefore a slight increase in POFD to 0.038 at a threshold setting of 0.90. This appears as a wiggle in the STONE curve as it descends nearly parallel to the *y* axis towards its high-threshold (0,0) location. The other metrics, shown in Figure 11h, resemble those in Figure 10p, except that all of the metrics are a bit worse at the high-end of the threshold sweep.

The third column of Figure 11 is the scenario with a downward shift in the low-*x* range and an upward shift in the high-*x* range. These options were not explored in the distributions of Figures 9 and 10, but were part of the distributions considered in the first two columns of Figure 11. This distribution's STONE curve, shown in Figure 11j, is the nearly perfect case of the *x*-axis "model" values correctly sorting the *y*-axis "data" values into events for all threshold settings. It is not quite perfect, though. The few points above the unity-slope line at very-low *x* values results in a ripple in the STONE curve as it runs nearly parallel to the top axis, while the few points below the unity-slope line at very-high *x* values produces a wiggle in the STONE curve while it descends nearly parallel to the left axis. In Figure 11(l), it is seen that both the ORSS and HSS curves are nearly symmetric about a threshold of 0.5, with both attaining excellent values in the middle of the threshold sweep. Because F_1 and FB are heavily dependent on *H*, which monotonically decreases as the threshold is swept, these two metrics do not exhibit symmetry.

Finally, the fourth column of Figure 11 shows the case of an upward shift at low-*x* values and a downward shift at high-*x* values. This is the combination of the fourth case from Figures 9 and 10. The scatterplot in Figure 11m shows a broad interval in the center from roughly $x = 0.25$ to $x = 0.75$ in which the *y* axis distribution has essentially no slope. This imposed configuration has profound implications for the STONE curve and other metrics. Figure 11n shows that the STONE curve twice comes very close to the upper-left (0,1) corner of POFD-POD space; at a threshold of 0.21 it attains (0.007, 0.904) and then at a threshold setting of 0.80 it reaches (0.082, 0.988). In between these two thresholds, though, the STONE curve undergoes a wiggle and then a ripple. First, POFD increases as points from the "zero slope" portion of the distribution enter the *F* quadrant. While POFD is still increasing, the increasing threshold makes the points at low-*x* values above the unity-slope line leave the *M* quadrant for the *C* quadrant, increasing POD. This offset of the maximum in POFD and the minimum in POD is clearly seen in Figure 11o. The effect is that the ripple begins before the wiggle reaches its maximum excursion, so instead of the STONE curve retracing its original path, it now swings in an more circular pattern. After reaching its peak POD value and second closest approach to the (0,1) corner, the STONE curve then plummets along the left axis. The metrics in the final panel (Figure 11p) show that the ORSS and HSS curves are again symmetric about a threshold of 0.5 with a relative minimum score at the middle of the sweep. F_1 and FB are not symmetric and have essentially the worst elements of the corresponding curves from Figures 9p and 10p.

3.4. Variations in event identification threshold sweep stepsize

One more assessment to conduct is with respect to the threshold step size. This is shown in Figure 12, for which four different threshold increment settings were used four of the distributions above (one each from Figures 5 and 10, plus two from Figure 11). Figure 12 is shown in a different format than Figures 4 – 11, showing only the STONE curves. The number of steps for the top row is very coarse, then it was increased by a factor of three to four between each adjacent row in Figure 12 (with the 100 steps used above as the version between the third and fourth row).

It is seen in the STONE curves in each column of Figure 12 are essentially the same. Nearly all significant features are still visible in the top row with a threshold step size of 0.2. The only feature that is missing is the subtle wiggle and ripple in the third column. With a step size of 0.1, all main features are captured by the sweep. The only addition with more threshold steps is

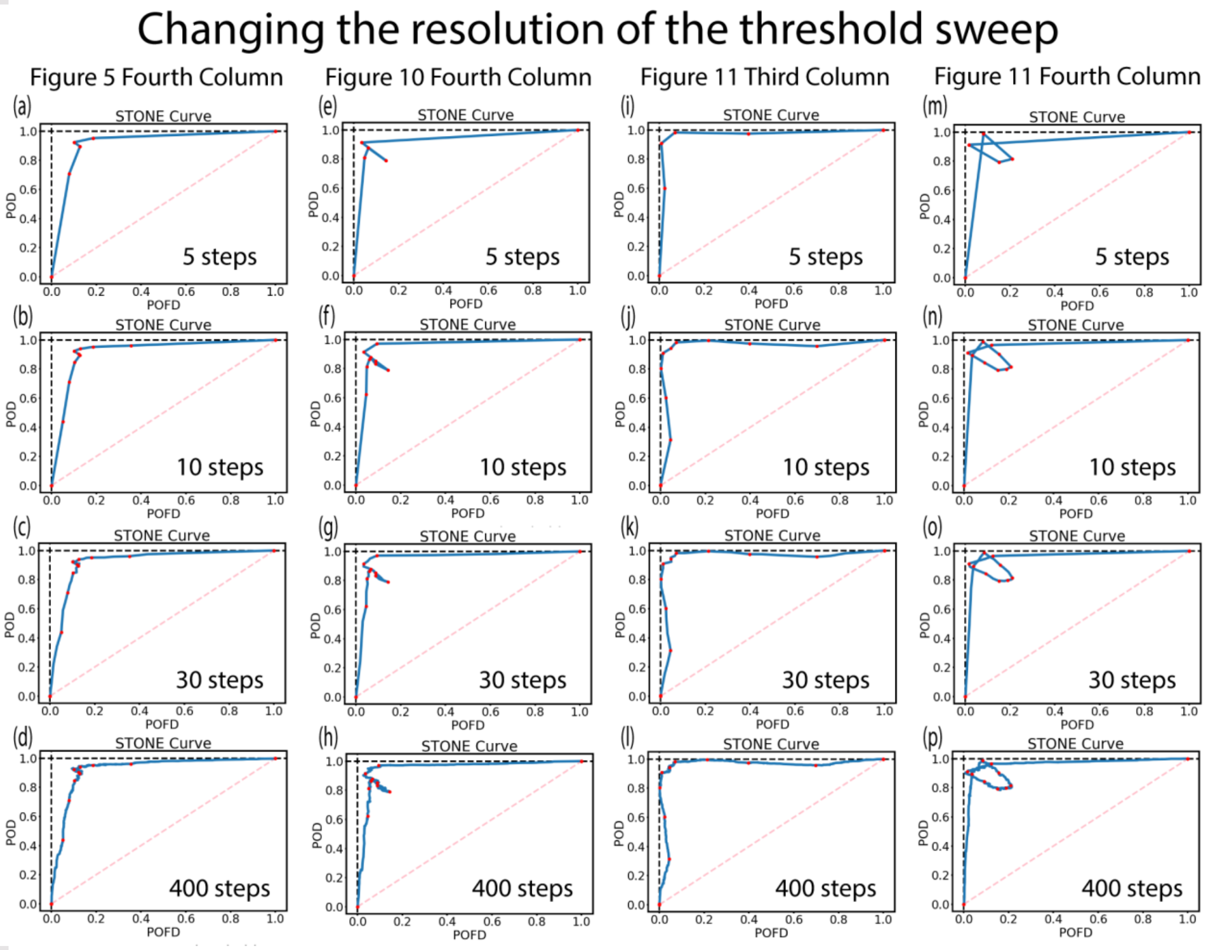

Figure 12. Changing the resolution of the threshold sweep when calculating the STONE curve and other metrics. Four of the distributions from previous figures were used, as indicated above each column. The original STONE curve above was conducted with 100 steps between 0 and 1; these STONE curves were created with a different number of steps in that interval: 5 for the top row, 10 for the second row, 30 for the third row, and 400 for the fourth row.

Poisson noise, the small-scale fluctuations along the curve. With 30 steps, there are with each threshold step, on average, about 30 points moving from *H* to *M* and another 30 moving to *F* (plus, perhaps, a few moving directly to *C*), and on average a similar number moving from *M* and *F* to *C*. Poisson noise on 30 counts rounds to 5, and the two counts (in and out of an error cell) will very rarely yield a fluctuation in POD or POFD from counting statistical uncertainty. With 400 steps, the 30 number drops to a little over 2, for which the Poisson noise over half of this value. Therefore, it will regularly experience counting uncertainty fluctuations. For the plots above with 100 steps, the 30 number is roughly 10. For this step size and average number of points moving from quadrant to quadrant, fluctuations are uncommon but expected. For these test distributions with 2000 points in the scatterplot, 100 steps across the domain of the *x*-axis "model" values is about the limit for a good STONE curve creation. As shown in Figure 12, fewer steps would still work to reveal most, if not all, of the main features of interest.

4. Discussion

The STONE curve is a data-model comparison tool that can be used when both the model and observed values are sets of real numbers and the model is trying to exactly reproduce the corresponding data. It has a calculation concept that is nearly identical to the ROC curve, plotting POD versus POFD, but with the threshold sweep occurring for both the model and data event identification thresholds (not just the model threshold, as is done for the ROC curve). This simultaneous sweep of both thresholds is only possible with the above-mentioned stipulations on the number sets.

As the threshold is swept from low to high values, the points in the data-model scatterplot are systematically shifting from one quadrant to another. At a very low threshold, all points are in the *H* quadrant. As the threshold increases, the points change from *H* to *C*, but usually not directly, most points visit the *F* or *M* quadrant along the way. The quadrant counts change in this way as the event identification threshold is increased: *H* always decreases; *C* always increases; and *F* and *M* increase then decrease (starting and ending at zero).

In general, the shape of a STONE curve resembles a ROC curve. For low thresholds, it has a value of (1,1) in POFD-POD space. It then moves "down and to the left" as the threshold is increased, eventually reaching the (0,0) corner for very high threshold settings. For well-behaved "good fit" data-model comparisons, the STONE curve is monotonic, like the ROC curve. The exception to this is if the threshold sweep step size is small and the number of points shifting from quadrant to quadrant with each step is, on average, less than 10. In this case, there will be small-scale fluctuations in the STONE curve due to the randomness of Poisson counting uncertainty. Such elements within a STONE curve are typically not important.

It has been shown above that there are two key nonmonotonic features of the STONE curve, a right-then-left wiggle and an up-then-down ripple. These are produced when there are clusters of points away from the unity-slope perfect fit line. A wiggle is produced when there is a small cluster of model over-predictions (points below the perfect fit line), producing an increase in *F* counts and therefore an increase in POFD over a small interval of the threshold sweep. A ripple is created in the STONE curve when there is a small cluster of model under-predictions (points above the perfect fit line), resulting in increased *M* for that part of the sweep. As these points leave the *M* quadrant, there is a corresponding rise in POD. If there are still many points spread around the perfect fit line, then the cluster will only influence *F* or *M*, producing either a wiggle or a ripple, respectively. If, however, the cluster is because of a shift in the distribution and there are few points spread around the perfect fit line, then the increase in one error quadrant corresponds to a decrease in the other, producing a wiggle-ripple combination. If this shift of the distribution exists over only a small, isolated portion of the domain, then the STONE curve will exhibit diagonal excursion and then retrace itself. If, however, there are upward and downward offsets close to each other in the domain, then the STONE curve will develop a circular pattern.

If there is large spread of the points relative to the perfect fit line but this spread is fairly uniform across the model value space, then the STONE curve will be monotonic. The bigger the spread, the farther the STONE curve will be from the ideal (0,1) value in the upper left corner of POFD-POD space, but it will not exhibit nonmonotonic features. A wiggle or ripple feature requires a cluster away from the perfect fit line; uniform spread will not cause these STONE curve properties. If, however, the spread around the perfect fit line is only in one part of the value domain space, then this will appear as a cluster and the STONE curve will include a wiggle or ripple.

It was shown above that these wiggle and ripple characteristics of the STONE curve will appear if the local RMSE of the distribution (spread of observed values within a very limited model value range) reaches a fractional value of 0.2 of the full model value domain. It will also appear if a local bias of the distribution (difference of the mean of the observed values and mean of the model values within a very limited model value range) is more than half of the local RMSE. Deviations larger than either of these thresholds result in clear wiggles and/or ripples in the STONE curve.

These wiggles and ripples in the STONE curve are useful for identifying the domain values where these clusters occur. These clusters will influence other metrics, as shown above for a few well-known formulas. Together, the feature in the STONE curve and the variation in the other metrics provide a robust description of the data-model relationship. It is highly encouraged to use the STONE curve with metrics from several categories (as discussed by Liemohn et al., 2021), including subsetting around the interval of the cluster.

The wiggle and ripple features of the STONE curve identify threshold settings for which a cluster of off-diagonal points moves into or out of the M or F quadrant. For a left-to-right wiggle, there was a rapid increase in F, revealing a cluster of model overestimations of the corresponding observed values. For an up-and-down ripple, there was a rapid decrease in M, revealing a cluster of model underestimations. These features indicate nothing specific about the physical process or numerical issue responsible for the model to be off from the observations, but rather simply identifies such clusters. A next step in the assessment would be to determine exactly where and/or when these over- or underestimations happened, a process that might lead to the discovery of some problem with physics or numerics the model. This, hopefully, results in a correction to improve the output.

This entire analysis has been conducted with idealized distributions over normalized value domains. Two issues should be stated regarding this. The first is that no derivations of statistical significance probabilities were conducted in this analysis (such as, for example, the probabilities associated with the significance of a t test). It is left as future work for a more theoretical investigation of STONE curve features relative to characteristics of the scatterplot distribution. The second is that, when applying the STONE curve to a specific data-model comparison, the scatterplot could be far more complicated than the simple variations of the imposed distributions above. It is hoped that the idealized nature of the distributions in this study provide clear connections between the scatterplot and the STONE curve. All of the imposed distributions used above were random values from Gaussian functions; real distributions might not have a Gaussian histogram, which could complicate the interpretation. We hope, however, that this work provides guidance to using the STONE curve with real data-model comparisons.

5. Application to Space Weather: IMPTAM-GOES comparisons

As an application of these relationships between scatterplot and STONE curve features, let's use the same example as in Liemohn et al. (2020), specifically the prediction of energetic electron observations (40 keV energy channel) from the geosynchronous orbiting environmental satellites (GOES) spacecraft (Rowland & Weigel, 2012) to real-time output from the inner magnetosphere particle transport and acceleration model (IMPTAM). IMPTAM has been running in real time for nearly a decade (Ganushkina et al., 2015), and these two particular number sets were originally compared by Ganushkina et al. (2019). The time period is from September 2013 through March 2015 and, as with Liemohn et al. (2020), the magnetic local time (MLT) of the comparison will be restricted to the dawn sector (specifically, the 03 to 09 MLT range), when the model performs the best of any MLT sector.

Figure 13 shows a set of plots for this data-model comparison. The first four plots are similar to what has been shown above, while the final two plots are new: histograms of subsets of one the number sets. Figure 13e shows discrimination subsetting, IMPTAM histograms using ranges of the GOES values, while Figure 13f is reliability subsetting, showing GOES histograms using ranges of the IMPTAM values.

To go along with these plots are statistics of the number sets, presented in Table 1. The first row of values are for the full number sets, while the two lower groupings of values are for the discrimination and reliability subsetting intervals, respectively. Listed are the mean, standard deviation, and skewness coefficient (using the definitions in Liemohn et al., 2021), for both the GOES observations and the IMPTAM output, the RMSE score between them, along with the number of data-model pairs in each interval.

Table 1. Statistics of the number sets for the IMPTAM-GOES comparison

	Full number set statistics							
Interval	**Data Mean**	**Data St. Dev.**	**Data Skew**	**Model Mean**	**Model St. Dev.**	**Model Skew**	**RMSE**	**Count in subset**
Full range	4.58	0.46	0.4	4.57	0.77	-0.2	0.7	28659
	Discrimination subsetting (based on GOES value ranges)							
Interval (log flux)	**Data Mean**	**Data St. Dev.**	**Data Skew**	**Model Mean**	**Model St. Dev.**	**Model Skew**	**RMSE**	**Count in subset**
Up to 4.0	3.83	0.18	-3.9	3.91	0.77	1.0	0.8	2367
4.0 to 4.5	4.30	0.14	-0.4	4.36	0.73	0.08	0.7	11533
4.5 to 5.0	4.67	0.13	0.8	4.76	0.68	-0.3	0.7	9241
5.0 to 5.5	5.25	0.14	0.02	4.94	0.63	-0.5	0.7	4347
5.5 to 6.0	5.63	0.10	1.1	5.10	0.68	-0.8	0.9	1165
6.0 and above	6.14	0.04	-0.2	5.7	0.08	0.5	0.5	6
	Reliability subsetting (based on IMPTAM value ranges)							
Interval (log flux)	**Data Mean**	**Data St. Dev.**	**Data Skew**	**Model Mean**	**Model St. Dev.**	**Model Skew**	**RMSE**	**Count in subset**
Up to 4.0	4.30	0.39	0.8	3.56	0.35	-1.0	0.9	7176
4.0 to 4.5	4.50	0.39	0.8	4.26	0.14	-0.9	0.5	5568

4.5 to 5.0	4.65	0.42	0.5	4.75	0.14	0.2	0.4	6943
5.0 to 5.5	4.78	0.43	0.3	5.23	0.14	0.1	0.6	5601
5.5 to 6.0	4.82	0.47	0.2	5.70	0.14	0.4	1.0	2867
6.0 and above	4.64	0.69	-0.6	6.19	0.19	2.1	1.7	504

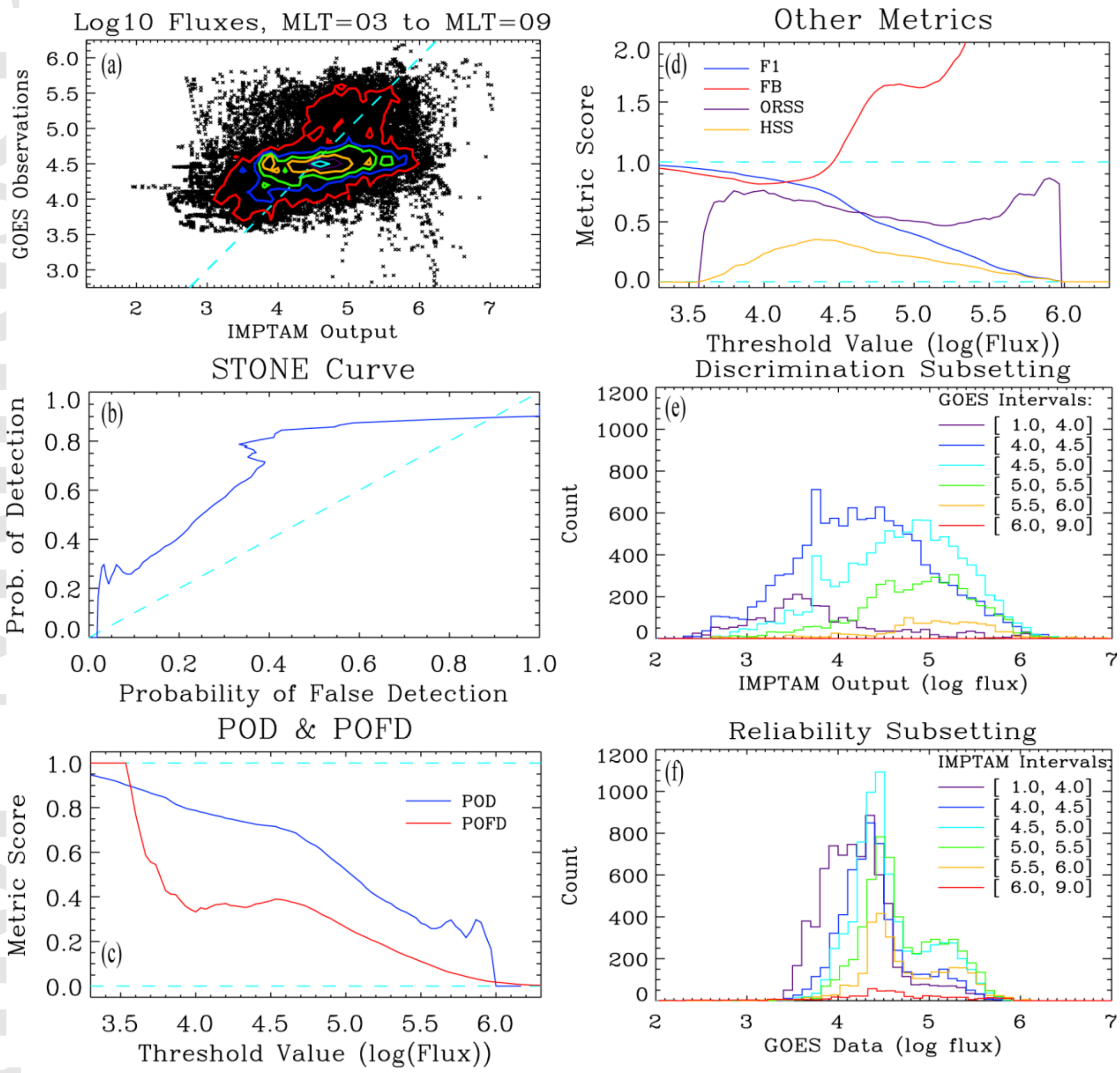

Figure 13. Comparison of 40 keV electron differential number fluxes (log base 10 of electrons cm^{-2} s^{-1} sr^{-1} keV^{-1}) from GOES and real-time IMPTAM. Shown are: (a) scatterplot of the data-model pairs, with 10 bins per decade and contours drawn every 40 counts/bin increment; (b) the STONE curve; (c) the POD and POFD curves as a function of threshold; (d) the four other metrics discussed in this study; (e) discrimination subsetting histograms of IMPTAM values for 6 ranges of the GOES values; and (f) reliability subsetting histograms of GOES values for 6 ranges of the IMPTAM values. Dashed lines in cyan are included for reference.

First, let's consider the full number set comparison. Figure 13a is a scatterplot of the GOES flux values against the corresponding real-time IMPTAM output. In general, the points are spread fairly evenly on either side of the diagonal perfect fit line. Table 1 shows that the means are very close and both skewness coefficients are low. The IMPTAM values have a larger spread, which can been seen in the scatterplot. The GOES values are rather constricted in their range, with most measurements confined within half an order of magnitude and only small count values beyond this central peak. The IMPTAM values, however, are spread across a wider range, spanning two full orders of magnitude.

Figures 13b shows the STONE curve for this comparison. The STONE curve is nearly always above the unity slope diagonal line, indicating that IMPTAM has skill (better than random chance) at sorting the GOES dawnside 40 keV electron fluxes into high and low flux categories. The only threshold settings for which the STONE curve is below the diagonal line are at the very low and very high values.

As presented in Liemohn et al. (2020), this comparison yields both a left-right wiggle as well as an up-down ripple in the STONE curve. The wiggle occurs when POFD exhibits an increase with increasing threshold, seen in the red curve of Figure 13c as occurring between log-flux threshold settings of 4.0 and 4.7. This occurs when a large number of points enter the *F* quadrant. As seen in Figure 13a, the peak of the distribution is "below" the unity-slope line; as the threshold sweeps past this peak, they first enter the *F* quadrant before shifting to the *C* quadrant. When they leave *F* for *C*, POFD decreases and the STONE curve continues its trek towards the left-side axis. The ripple occurs when POD increases with increasing threshold. This occurs at rather high settings, between log-flux values of 5.5 and 6.0, as seen in the blue curve in Figure 13c. A ripple occurs when a cluster leaves the *M* quadrant. The cluster is seen in Figure 13a as the red contour extends around a group of points above the unity slope line. Because they are above the unity slope line, the model event identification threshold sweeps past them first, resulting in an increase in *M* (a reduction in POD). At a higher threshold setting, the observed event identification threshold sweeps past them, putting them in the *C* quadrant. Because so many leave at one time, POD increases.

The other metrics, shown in Figure 13d, also quantify features of the data-model relationship. The F_1 score (blue line) has a slight downward trend at low threshold settings as points are converted into the *M* quadrant. We know it is the *M* quadrant that dominates F_1 because the FB metric is below one and *M* is in the denominator of this equation. It then exhibits a downward kink around a log-flux value of 4.5. This is coincident with an upward turn of FB, indicating that it is due to many points entering the *F* quadrant. This is also the same threshold as when the STONE curve exhibited a left-to-right wiggle. Because the two number sets have every similar means, the ORSS metric has high values near each end of the threshold sweep, with a lower value in the middle because of the large spread of the IMPTAM model values relative to the GOES measured values. HSS is always at or above unity, confirming that IMPTAM has skill at organizing the GOES flux values into high-flux events and low-flux nonevents. Its peak score occurs just before the threshold is swept over the large cluster of points below the perfect fit line.

All of this can be further clarified with a subsetting analysis of the number sets. One method of subsetting, known as discrimination, is constructed using only the data-model pairs that lie within a specified range of observed values. Figure 13e shows histograms of IMPTAM flux values for 6 intervals of the GOES data range. As the observed interval is incremented upwards in flux range, it is seen that the modeled values are also shifting upward. This is what is expected.

Table 1 lists some key statistics of both the GOES and IMPTAM values in each of these GOES-value interval ranges. The means of the GOES values should increase, as they are limited within the intervals. There is no guarantee that the IMPTAM means should increase with increasing interval, though. For the lower 3 intervals, where most of the points are located, the log-flux means are within 0.1 of each other. Note that if a skew is moderate-to-large (say, above an absolute value of 0.7), then the distribution is most likely not close to a Gaussian distribution and the typical probabilities with statistical tests and inference cannot be applied. This is the case for several of the intervals in one number set or the other, so no mean-testing calculations are performed as the resulting *p*-value would be meaningless.

Figure 13f shows the opposite case of subsetting, known as reliability, in which value intervals of the model values define the subsets. This is exactly analogous to how the idealized distributions above were constructed and evaluated. Two features are immediately evident in these histograms of GOES flux values within each of the 6 IMPTAM value ranges: first, the modes of the 6 histograms are very close; and second, there is another peak in the distribution at high flux values. This lack of movement of the histogram modes is evident in the scatterplot – the GOES values do not have a large spread around a log-flux value of 4.5. The secondary peak in the histograms is also evident in the scatterplot – it is the extended region surrounded by the red contour above the perfect fit line.

These histograms are further quantified in the final section of numbers given in Table 1. Here, the IMPTAM means rise with increasing interval, as they should, but it is seen that the means of the GOES values are all within 0.5 of each other and all below 5.0. At the lowest interval, the GOES mean is well above the IMPTAM mean, it's then close for two of the intervals, and then it is lower than the IMPTAM mean in the top three intervals. The standard deviations of the GOES values are larger than those of the IMPTAM points within each interval, as expected, but these spread values are not as large as those of the IMPTAM spreads in the discrimination analysis.

The wiggle is due to the main grouping of points for which IMPTAM overestimates the GOES fluxes, while the ripple is due to high-flux observations for which IMPTAM underestimates the GOES fluxes. As seen in Figure 13f, the GOES measurements have a bimodal distribution, with a secondary peak at higher flux values than the primary peak. The ratio of the mode of the secondary-to-primary peak for the reliability intervals goes up 0.36 for the green, orange, and red curves (within 0.01). Table 1 shows that the calculated RMSE of 0.4 at the lower intervals but then much larger than this at the high log-flux intervals. Furthermore, the shift of the secondary peak from the primary one is a full order of magnitude, larger than half of this RMSE value. The combination of a relatively large secondary peak (as evidenced by the mode ratios) and a separation of the peaks much larger than the local spread results in an easily-discernible ripple in the STONE curve.

6. Conclusions

The STONE curve is a data-model comparison technique that is very similar to a ROC curve, but with a key difference: the event identification threshold is swept for both the model and data, not just the model threshold. The STONE curve is best used with a continuous-valued data set for which the model is trying to predict those exact values. The STONE curve answers the question: does the model predict events at each threshold setting? This is a question that cannot be answered by a ROC curve, for which the observed events and nonevents are fixed. The ROC curve is still very useful for what it does – optimizing a prediction of known observed events from a continuous-valued model – which is not something that the STONE curve can do. They are complementary but unique tools within the array of statistical methods available for comparing number sets.

The STONE curve identifies intervals of the event threshold identification setting range where the model performs well at sorting the observations into events and nonevents. When the STONE curve is close to the (0,1) upper left corner of POFD-POD space, this indicates that the observed values (classified as events and nonevents, whether above or below that particular threshold setting, respectively) are mostly classified correctly by the event-nonevent status of the corresponding model values. A perfect classification – one with no counts in the *M* and *F* error quadrants – could occur for multiple threshold settings, in which case the STONE curve would linger or return to the (0,1) corner. Furthermore, the STONE curve technique of sweeping both thresholds simultaneously identifies thresholds where metrics often surpass “goodness cutoffs.” This is especially seen in plotting a metric versus threshold, using the same technique of sweeping both thresholds together and then calculating other event detection metrics from the resulting collection of contingency tables.

An example was shown of how to use the STONE curve, in conjunction with other metrics, for a robust evaluation of a data-model comparison for magnetospheric real-time predictions. Note, however, that this technique is not limited to space physics. It is particularly useful for model predictions of time series data, as is the case for other fields, such as terrestrial weather. In fact, it can be used for any scientific discipline, any time there is a comparison of real-numbered observed values to a model output number set. It augments the standard set of metrics, providing a method to identify intervals for which the model is particularly good at reproducing the data, and other intervals for which there is a cluster of points far from a perfect match.

To summarize, the main findings of this study are as follows:

- A key feature of the STONE curve is that it can be nonmonotonic – exhibiting wiggle and ripple features. These have been quantified with idealized number set distributions.
- The left-right wiggle is produced when there is an influx of points into the *F* (false alarm) quadrant of the contingency table, that is when there is a cluster of model overpredictions.
- The up-down ripple is produced when there is a rapid outflow of points from the *M* (misses) quadrant, that is when there was a cluster of model underpredictions.
- These two features can occur independently or in combination.
- These extra characteristics of the STONE curve will appear if the local RMSE of the distribution (spread of observed values within a very limited model value range) reaches a fractional value of 0.2 of the full model value domain.

- The extra characteristics will also appear if a local bias of the distribution (difference of the mean of the observed values and mean of the model values within a very limited model value range) is more than 0.5 of the local RMSE.

Acknowledgments and Data

The authors would like to thank the US government for sponsoring this research, in particular research grants from NASA (NNX17AI48G, 80NSSC20K0353, 80NSSC17K0015, 80NSSC19K0077, 80NSSC21K1127, and NNX17AB87G) and NSF (AGS-1414517). The authors received funding from the European Union Horizon 2020 Research and Innovation programme under grant agreement 870452 (PAGER).

The specific number sets used in this study are available at the University of Michigan Deep Blue Data repository, https://deepblue.lib.umich.edu/data/ . The eventual permanent DOI will be here, which will be active after review: https://doi.org/10.7302/2mcx-5749 . Until then, the submitted files are available via anonymous download here: https://deepblue.lib.umich.edu/data/anonymous_link/download/a31c81d52f243a355e49b2098228f2fb8be20c5f838f7b09ace8f049e70012f1?locale=en . The Python and IDL code used to conduct this analysis is also available at this site. The number sets for the example IMPTAM-GOES comparison are available here, from a prior journal article publication (https://deepblue.lib.umich.edu/data/concern/data_sets/02870v99r?locale=en).

References

Azari, A. R., Liemohn, M. W., Jia, X., Thomsen, M. F., Mitchell, D. G., Sergis, N., et al. (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/10.1029/2018JA025391

Azari, A. R., J. Lockhart, M. W. Liemohn, & X. Jia (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

Azzalini, A. and Capitanio, A. (1999), Statistical applications of the multivariate skew normal distribution. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61: 579-602. https://doi.org/10.1111/1467-9868.00194

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

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

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

Hogan, R. J., & Mason, I. B. (2012). Deterministic forecasts of binary events. In I. T. Jolliffe & D. B. Stephenson (Eds.), *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (2nd ed., chap. 3, pp. 31–60). Chichester, UK: John Wiley, Ltd. https://doi.org/10.1002/9781119960003.ch3

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

Kubo, Yûki, Mitsue, Den, & Mamoru, Ishii (2017). Verification of operational solar flare forecast: Case of Regional Warning Center Japan. *Journal of Space Weather & Space Climate*, *7*, A20. DOI: 10.1051/swsc/2017018

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*, *6*, e2020EA001106. https://doi.org/10.1029/2020EA001106

Liemohn, M. W., Shane, A. D., Azari, A. R., Petersen, A. K., Swiger, B. M., & Mukhopadhyay, A. (2021). RMSE is not enough: guidelines to robust data-model comparisons for magnetospheric physics. *Journal of Atmospheric and Solar-Terrestrial Physics*, *218*, 105624. https://doi.org/10.1016/j.jastp.2021.105624

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

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

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

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

Murphy, A. H. (1991). Forecast verification: Its complexity and dimensionality. Monthly Weather Review, 119, 1590–1601.

Potts, J. M. (2021). Basic Concepts. In *Forecast Verification* (eds I. T. Jolliffe and D. B. Stephenson). https://onlinelibrary.wiley.com/doi/10.1002/9781119960003.ch2

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

Swets, J.A. (1986) Indices of discrimination or diagnostic accuracy: their ROCs and implied models. Psychological Bulletin, 99, 100–117.

Wilks, D. S. (2019). *Statistical methods in the atmospheric sciences* (4th ed.). Oxford: Academic Press.