NDFD Verification Score Definitions

Continuous (CNT) Scores

Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is a measure of forecast accuracy. A small value indicates a better score, a perfect score is zero. MAE is defined as:

$$MAE = \sum_{i=1}^{N} \frac{(|f_i - o_i|)}{N}$$

where N = the total number of observations, f = the forecasts, and o = the observation.

On our page, gridded and point verification scores include MAE scores for the following weather elements: Max Temp, Min Temp, Sfc Temp, Dew Point, Wind Speed, Wind Direction, Wind Gust, QPF06, QPF01, Snow, and Wave Height.

Bias

The mean algebraic error (bias) indicates whether a forecast is too high or too low in predicting a certain parameter. For example, a positively biased temperature forecast indicates that forecasts were, on average, too warm. Similarly, a negatively biased temperature forecast indicates that forecasts were, on average, too cool. Using another example, a positively biased wind speed forecast indicates that forecasts were, on average, predicting wind speeds that were too high. A bias of zero is possible if a forecaster's over-forecasting and under-forecasting cancel each other or if the forecast is perfect. Bias should be looked at in conjunction with MAE to determine forecasting error. The bias is defined as:

$$Bias = \sum_{i=1}^{N} \frac{(f_i - o_i)}{N}$$

where N = the total number of observations, f = the forecasts, and o = the observation.

On our page, gridded verification scores include bias scores. Point verification scores include bias scores for the following weather elements: Max Temp, Min Temp, Sfc Temp, Dew Point, Wind Speed, Wind Gust, and Relative Humidity.

Root Mean Square Error (RMSE)

The square root of the mean of the square of the differences of the forecasts and observations.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (f_i - o_i)^2}{N}}$$

where N = the total number of observations, f = the forecast, and o = the observation.

The root mean square error is a measure of forecast accuracy. The lower the mean square error, the more accurate the forecasts.

Mean Absolute Difference (MAD)

The Mean Absolute Difference (MAD) is a comparison of two sets of forecasts. On our page, NDFD is used as the base forecast system and different versions of the Blend and WPC forecasts are used for comparison. A smaller value indicates a smaller change between the two systems, with zero indicating no difference between the systems. MAD is defined as:

$$MAD = \sum_{i=1}^{N} \frac{(|f_{ai} - f_{bi}|)}{N}$$

where N = the total number of observations, f_a = the first forecast system, and f_b = the base forecast system for comparison (NDFD)

On our page, gridded verification scores include MAD scores for the following weather elements: Max Temp, Min Temp, Sfc Temp, Dew Point, Wind Speed, Wind Direction, Wind Gust, QPF06, Snow, Sky Cover, and Wave Height.

Contingency Table Scores (CTS)

Heidke Skill Score (HSS)

The Heidke Skill Score (HSS) is a measure of skill in forecasts. It is defined as follows:

$$HSS = \frac{NC - E}{T - E}$$

where NC equals the number of correct forecasts (in other words, the number of times the forecast and the observations match), T equals the total number of forecasts, and E equals the number of forecasts expected to verify based on chance.

This can be calculated using a contingency table:

Heidke Skill Score Table

	Forecast Category					
Observed Category		1	2	…	m	Total
	1	X_{11}	X_{12}	…	X_{1m}	X_{1p}
	2	X_{21}	X_{22}	…	X_{2m}	X_{2p}
	…	…	…	…	…	…
	m	X_{m1}	X_{m2}	…	X_{mm}	X_{mp}
	Total	X_{p1}	X_{p2}	…	X_{pm}	X_{pp}

where m is the number of categories, the element Xij indicates the number of times the forecast was in the jth category and the observation was in the ith category. The row and column totals are shown by the subscript (and category) p.

$$NC = \sum_{i=1}^{m} X_{ii}$$

$$T = X_{pp}$$

$$E = \sum_{i=1}^{m} \frac{X_{ip} X_{pi}}{T}$$

A negative HSS indicates that a forecast is worse than a randomly based/generated forecast. On our page, we indicate a perfect score in two ways. First, for a sample whose forecasts and observations fall into more than one category (i.e., the matched forecast and observation totals occupy more than one cell of the contingency table), the computed HSS=1.0. Second, for a sample whose forecast/observation total occupies only one cell of the contingency table (there was no reason to forecast anything but the most commonly observed condition), we set the HSS equal to 9997. This difference helps to highlight the stations that achieved a perfect score under more difficult forecast conditions.

Probability of Detection (POD)

The Probability of Detection is a measure of categorical forecast performance equal to the total number of correct forecasts (hits) divided by the total number of events observed (hits + misses). POD values range from 0 to 1, with 0 being no skill and 1 being a perfect forecast. It is defined as follows:

$$POD = \frac{A}{A + C}$$

This can be computed using a 2X2 contingency table:

POD Contingency Table		**Event Observed**	
		Yes	**No**
Event Forecast	**Yes**	A	B
	No	C	D

False Alarm Ratio (FAR)

The False Alarm Ratio is a measure of categorical forecast performance equal to the number of false alarms divided by the total number of event forecasts. FAR values

range from 0 to 1, with 1 being no skill and 0 being a perfect forecast. It is defined as follows:

$$FAR = \frac{B}{A+B}$$

This can be computed using a 2X2 contingency table:

FAR Contingency Table		Event Observed	
		Yes	No
Event Forecast	Yes	A	B
	No	C	D

Success Ratio

The Success Ratio (SR) is a measure of categorical forecast performance equal to the number of hits divided by the total number of event forecasts (hits + false alarms). We only display the success ratio in the Performance Diagrams. A perfect score is SR=1.

$$SR = \frac{A}{(A+B)} = 1 - FAR$$

This can be computed using a 2X2 contingency table:

SR Contingency Table		Event Observed	
		Yes	No
Event Forecast	Yes	A	B
	No	C	D

Probability of False Detection

The Probability of False Detection, also known as the False Alarm Rate (not to be confused with the False Alarm Ratio), is a verification measure of categorical forecast performance equal to the number of false alarms divided by the total number of

non-events. We only display the Probability of False Detection in the ROC Curve. A perfect score is POFD=0.

$$POFD = \frac{B}{(B+D)}$$

This can be computed using a 2X2 contingency table:

SR Contingency Table		Event Observed	
		Yes	**No**
Event Forecast	**Yes**	A	B
	No	C	D

Critical Success Index (CSI)

The Critical Success Index, also known as the Threat Score, is a measure of categorical forecast performance equal to the total number of correct event forecasts (hits) divided by the total number of yes-forecasts plus the number of misses (hits + false alarms + misses). The CSI is not affected by the number of non-event forecasts that verify. However, the CSI is a biased score that is dependent on the frequency of the event. For an unbiased version of the CSI, see the Gilbert Skill Score. CSI values range from 0 to 1, with 0 being no skill and 1 being a perfect forecast. The CSI is defined as follows:

$$CSI = \frac{A}{A + B + C}$$

This can be computed using a 2X2 contingency table:

CSI Contingency Table		Event Observed	
		Yes	**No**
Event Forecast	**Yes**	A	B
	No	C	D

Gilbert Skill Score (GSS)

The Gilbert Skill Score, also known as the Equitable Threat Score, is a measure of categorical forecast performance which takes into account the number of hits due to chance. Hits due to chance is the event frequency multiplied by the number of event forecasts. The GSS is the total number of correct event forecasts minus the hits due to chance (hits - chance hits) divided by the total number of forecasts plus the number of misses minus the hits due to chance (hits + false alarms + misses - chance hits). GSS values range from -⅓ to 1. A no-skill forecast would have GSS=0. A perfect forecast would have GSS=1. The GSS is defined as follows:

$$GSS = \frac{A - CH}{A + B + C - CH}$$

where

$$CH = \frac{(A + B)(A + C)}{A + B + C + D}$$

This can be computed using a 2X2 contingency table:

GSS Contingency Table		**Event Observed**	
		Yes	**No**
Event Forecast	**Yes**	A	B
	No	C	D

Frequency Bias

The Frequency Bias (FB) is a measure of bias for categorical forecasts, equal to the total number of event forecasts (hits + false alarms) divided by the total number of observed events (hits + misses). This type of bias is also known as overall bias, systematic bias, or unconditional bias. A perfect score is FB=1. Scores greater than [less than] 1 indicate over-forecasting [under-forecasting].

$$FB = \frac{(A+B)}{(A+C)}$$

This can be computed using a 2X2 contingency table:

FB Contingency Table		Event Observed	
		Yes	No
Event Forecast	Yes	A	B
	No	C	D

Percent Correct

The Percent Correct is a measure of forecast accuracy. It is defined as follows:

$$PC = \frac{NC}{T}$$

where NC equals the number of correct forecasts (in other words, the number of times the forecast and the observations match), and T equals the total number of forecasts. The percent correct is calculated from the contingency table counts (CTC) produced by MET.

This can be calculated using a contingency table:

	Forecast Category					
Observed Category		1	2	…	m	Total
	1	X_{11}	X_{12}	…	X_{1m}	X_{1p}
	2	X_{21}	X_{22}	…	X_{2m}	X_{2p}
	…	…	…	…	…	…
	m	X_{m1}	X_{m2}	…	X_{mm}	X_{mp}
	Total	X_{p1}	X_{p2}	…	X_{pm}	X_{pp}

where m is the number of categories, the element Xij indicates the number of times the forecast was in the jth category and the observation was in the ith category. The row and column totals are shown by the subscript (and category) p.

$$NC = \sum_{i=1}^{m} X_{ii}$$

$$T = X_{pp}$$

Contingency Table Counts for Probabilistic Forecasts

Relative Frequency

Relative Frequency (RF) is a measure of the number of occurrences a forecast falls within a certain bounded error. It is defined as follows:

$$RF_i = \sum_{i=1}^{N} \frac{n_i}{N}$$

where n_i equals the count of direction errors in a certain category and N equals the total number of forecasts/observations.

For example, if there are 10 forecasts with 8 of the forecasts having an error less than or equal to 5 degrees, the relative frequency of forecasts with error of 5 degrees or less is 0.8.

Contingency Table Statistics for Probabilistic Forecasts

Brier Score

The Brier Score is the mean square error applied to probability forecasts. A common form of this score, called the half-Brier score, used on these verification web pages, is defined as follows:

$$\sum_{i=1}^{N} \frac{(f_i - o_i)^2}{N}$$

where N = the total number of observations, f = the forecasts, and o = the observation.

- The probability forecasts used in these computations have 1% precision.

- The observation is set equal to one if precipitation greater than or equal to 0.01 inches occurred, or to zero if no precipitation (or a trace) occurred.
- The score (half-Brier score) has a range of 0 to 1, with lower scores indicating better forecasts. Example: If the PoP forecast is 100% in 10 cases, and it rains in all 10 cases, then the Brier score is 0. Similarly, if the PoP forecast is 50% in 10 cases, and it rains in only 5 of them, then the Brier score is 0.25.
- Generally, the rarer the event, the better the Brier score, regardless of the forecast skill. Therefore, care must be used when comparing the Brier scores for different locations or seasons.
- On our page, this score is used only for point verification for the weather element PoP.

Probabilistic Statistics

Continuous Ranked Probability Score

The CRPS is a measure of accuracy and reliability of a continuous forecast. It is a negatively oriented score with a perfect score of CRPS=0. It is the squared difference between the forecast probability and the observation, which collapses to the MAE for a deterministic case.

$$CRPS = \int_{-\infty}^{\infty} (P(f) - P(o))^2 dx$$

Spread-Skill Score

The spread-skill score is an experimental score that relates the accuracy and reliability of a probabilistic forecast. For a forecast with a near-normal error distribution, the ratio of the 50% spread to the MAE should be near 1 for a reliable system. Scores less than 1 indicate under-dispersion while scores greater than 1 indicate over-dispersion.

$$SSS = \frac{\frac{1}{N}\sum_{i=1}^{N}(P_{75i} - P_{25i})}{\frac{1}{N}\sum_{i=1}^{N}|P_{50i} - o_i|}$$

Probability Integral Transform Histogram

The Probability Integral Transform (PIT) is the percentile location on a Cumulative Density Function (CDF) of the observation. A PIT Histogram is constructed by putting the PITs into the appropriate percentile bins and calculating the height of the bin by dividing the count of the bin by the expected count for that bin. It is used to assess the reliability of a probabilistic forecast system. A reliable system will have a flat PIT Histogram with a height of 1, meaning that the actual number of cases in the percentile bin is near the expected count. A PIT Histogram with a U-shape [O-shape] shows under-dispersion [over-dispersion]. A PIT Histogram skewed to the left [right] shows over-forecasting [under-forecasting].

Receiver Operating Characteristic Curve

The ROC Curve is a scatter plot of the Probability of False Detection (False Alarm Rate) and the Probability of Detection for each user-specified probability threshold (5, 10, 25, 50, 75, 90, 95).