V. Developmental Testbed Center Objective Evaluation Background

New Objective Verification Approaches

Subjective verification of model forecasts has been a cornerstone to HWT activities in previous years. This approach has provided valuable insights into how forecasters use numerical models, and facilitates the gathering of information about the value of new guidance tools from the perspective of a forecaster. In addition, traditional verification measures (e.g., Equitable Threat Score or ETS) used for synoptic scale and mesoscale model forecasts of discontinuous variables such as precipitation typically provide less useful information (and even misleading information) about forecast accuracy as the scale of the phenomena being evaluated decreases. This is because the ETS is proportional to the degree of grid scale overlap in space and time between the forecasts and observations, and there is typically low predictability on convective scales. Despite these limits, operational severe weather forecasters have often found value in WRF forecasts of thunderstorms and convective systems, since they can provide unique information about convective mode, coverage, and evolution that is not resolved by mesoscale models using parameterized convection. In recent years, we have found that subjective evaluation has great potential to serve as a comparative benchmark for assessing new objective verification techniques designed for high resolution NWP, and has had a significant positive impact on model development strategies.

In order to better utilize subjective and objective verification techniques in a complementary manner, simulated composite reflectivity and 1-hr QPF output from several model runs will be evaluated using subjective visual comparisons and objective statistical measures produced by the Developmental Testbed Center's (DTC) Meteorological Evaluation Tool (MET). The focus this year will be on probabilistic predictions, particularly of extreme precipitation events and severe weather as it relates to aviation weather. All members of the Center for Analysis and Prediction of Storms (CAPS) Storm Scale Ensemble Forecast (SSEF) system will be evaluated for select variables. Ensemble products from the fifteen members selected by the NOAA Storm Prediction Center (SPC) will also be evaluated. Operational (or near-operational) models will be used as a baseline for comparison. These include the North American Model (NAM, the High Resolution Rapid Refresh (HRRR, and ensemble products from the Short Range Ensemble Forecast SREF. Other contributing models will be brought in and archived for retrospective studies.

MET is designed to be a highly-configurable, state-of-the-art suite of verification tools. We will focus on the use of the object-based verification called Method for Object-based Diagnostic Evaluation (MODE) that compares gridded model data to gridded observations for the QPF and simulated reflectivity forecasts. MODE output will be tested to evaluate its ability to diagnose different types of convective modes considered important in forecasts and observations of convective weather, such as linear systems, discrete cells, and MCS's. Traditional verification statistics will also be computed. Details about the DTC MET system is at http://www.dtcenter.org/met/users/ .

Verification "truth" will be provided by NSSL National Mosaic and Multi-Sensor QPE (NMQ) multi-sensor Quantitative Precipitation Estimates (QPE) and three-dimensional radar reflectivity data bases. See http://www.nssl.noaa.gov/projects/q2/ for more information about the NMQ.

Models and Fields to be Evaluated

FCST Field	Observation	Grid-Stat	MODE	Models
Prob of Exceed (0.25", 0.5", 1", 2" over 3 and 6 hrs)	0.25", 0.5", 1", 2" QPE over 3 and 6 hrs	Brier Score, Decomp of Briar score, Area under ROC, Reliability Diagram	None	Ensemble products from CAPS and SREF
50% Prob of Exceed (0.25", 0.5", 1", 2" over 3 and 6 hrs)	0.25", 0.5", 1", 2" QPE over 3 and 6 hrs	None	MMI, Intersection Area, Area Ratio, Centroid Distance, Angle Difference, % Objects and Area Matched, 50th and 90th Percentile of Variable	Ensemble products from CAPS and SREF
0.25", 0.5", 1.0", 2" QPF over 3 and 6 hrs	0.25", 0.5", 1.0", 2" QPE over 3 and 6 hrs	GSS, CSI, FAR, PODY, FBIAS	Same as above	CAPS members, CAPS ens mean, SREF ens mean, HRRR, NAM
Sim. CompositeRefl (20,30,40,50 dBZ)	Q2 Composite refl (20,30,40,50 dBZ)	GSS, CSI, FAR, PODY, FBIAS	Same as above	CAPS members, CAPS ens mean, SREF ens mean, HRRR, NAM
18 dBZ Echo Top (18, 25, 30, 35, 40, 45 kft)	Q2 18dBZ Echo Top (18, 25, 30, 35, 40, 45 kft)	GSS, CSI, FAR, PODY, FBIAS	Same as above	CAPS members, CAPS ens mean, SREF ens mean, HRRR, NAM
Prob of 40dBZ echos	Q2 Composite reflectivity (40dBZ)	GSS, CSI, FAR, PODY, FBIAS	None	Ensemble products from CAPS and SREF
50% Prob of 40dBZ echos	Q2 Composite reflectivity (40dBZ)	None	Same as above	Ensemble products from CAPS

Table 1. List of variables (and thresholds) to be evaluated during SE 2010. Many will be available in real-time and others will be generated retrospectively. Traditional and Spatial metrics for which models are also listed.

DTC Verification Metrics Summary

1.Traditional Verification Metrics – *excerpted from the WWRP/WGNE Joint Group on Forecast Verification Research website on Forecast Verificaiton: Issues, Methods and FAQ (http://www.cawcr.gov.au/projects/verification/)*

1a. Statistics for dichotomous (2-category) variables

For dichotomous variables (e.g., precipitation amount above or below a threshold) on a grid, typically the forecasts are evaluated using a diagram like the one shown in Fig. 1. In this diagram, the area "**H**" represents the intersection between the forecast and observed areas, or the area of **Hits**; "**M**" represents the observed area that was missed by the forecast area, or the "**Misses**"; and "**F**" represents the part of the forecast that did not overlap an area of observed precipitation, or the "**False Alarm**" area. A fourth area is the area outside both the forecast and observed regions, which is often called the area of "**Correct Nulls**" or "**Correct Rejections**".

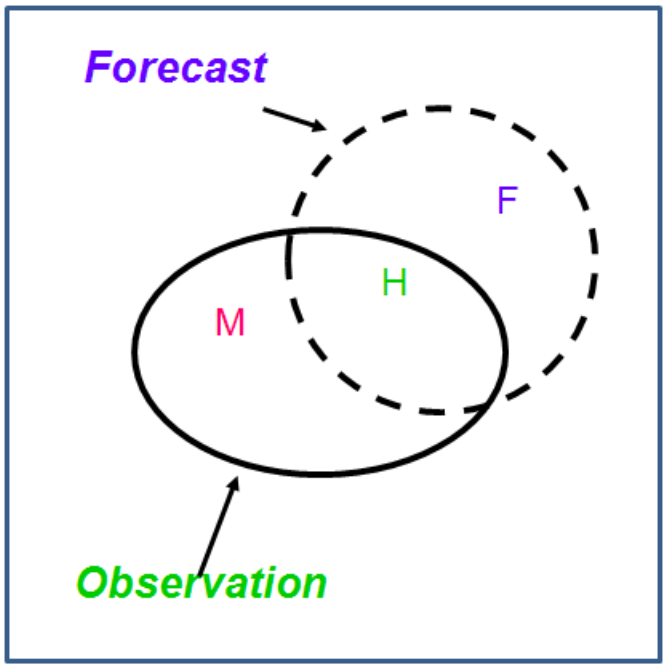


Figure 1. Diagram showing hits, misses, and false alarms for dichotomous forecast/observations.

This situation can also be represented in a "contingency table" like the one shown in Table 1. In this table the entries in each "cell" represent the counts of hit, misses, false alarms, and correct rejections. The counts in this table can be used to compute a variety of traditional verification measures, described in the following sub-sections.

Table 1. Contingency table illustrating the counts used in verification statistics for dichotomous (e.g., Yes/No) forecasts and observations. The values in parentheses illustrate the combination of forecast value (first digit) and observed value. For example, YN signifies a Yes forecast and and a No observation.

Forecast	**Observed**		
	Yes	**No**	
Yes	Hits (YY)	False alarms (YN)	YY + YN
No	Misses (NY)	Correct rejections (NN)	NY + NN
	YY + NY	YN + NN	Total = YY + YN + NY + NN

Base rate

$$\text{Base rate} = \frac{\text{Hits} + \text{Misses}}{\text{Total}} = \frac{\text{YY+NY}}{\text{Total}}$$

Also known as **sample climatology** or **observed relative frequency of the event.**

Answers the question: *What is the relative frequency of occurrence of the Yes event?*

Range: 0 to 1.

Characteristics: Only depends on the observations. For convective weather can give an indication of how "active" a day is.

Probability of detection (POD)

$$\text{POD} = \frac{\text{Hits}}{\text{Hits} + \text{Misses}} = \frac{\text{YY}}{\text{YY} + \text{NY}}$$

Also known as **Hit Rate.**

Answers the question: *What fraction of the observed Yes events was correctly forecasted?*

Range: 0 to 1. **Perfect score:** 1.

Characteristics: Sensitive to hits, but ignores false alarms. Good for rare events. Can be artificially improved by issuing more Yes forecasts to increase the number of hits. Should be used in conjunction with the false alarm ratio (below) or at least one other dichotomous verification measure. POD also is an important component of the Relative Operating Characteristic (ROC) used widely for evaluation of probabilistic forecasts.

False alarm ratio (FAR)

$$\text{FAR} = \frac{\text{False alarms}}{\text{Hits} + \text{False alarms}} = \frac{\text{YN}}{\text{YY} + \text{YN}}$$

Answers the question: *What fraction of the predicted "yes" events did* ***not*** *occur (i.e., were false alarms)?*

Range: 0 to 1. **Perfect score:** 0.

Characteristics: Sensitive to false alarms, but ignores misses. Very sensitive to the climatological frequency of the event. Should be used in conjunction with the probability of detection (above). Relative Operating Characteristic (ROC) used widely for evaluation of probabilistic forecasts.

Bias

$$\text{Bias} = \frac{\text{Hits} + \text{False alarms}}{\text{Hits} + \text{Misses}} = \frac{\text{YY} + \text{YN}}{\text{YY} + \text{NY}}$$

Also known as **Frequency Bias.**

Answers the question: *How similar were the frequencies of Yes forecasts and Yes observations?*

Range: 0 to infinity. **Perfect score:** 1.

Characteristics: Measures the ratio of the frequency of forecast events to the frequency of observed events. Indicates whether the forecast system has a tendency to underforecast (Bias < 1) or overforecast (Bias > 1) events. Does not measure how well the forecast gridpoints correspond to the observed gridpoints, only measures overall relative frequencies. Can be difficult to interpret when number of Yes forecasts is much larger than number of Yes observations.

Critical Success Index (CSI)

Also known as **Threat Score (TS).**

$$\text{CSI} = \text{TS} = \frac{\text{Hits}}{\text{Hits} + \text{Misses} + \text{False alarms}} = \frac{\text{YY}}{\text{YY} + \text{NY} + \text{YN}}$$

Answers the question: *How well did the forecast "yes" events correspond to the observed "yes" events?*

Range: 0 to 1, 0 indicates no skill. **Perfect score:** 1.

Characteristics: Measures the fraction of observed and/or forecast events that were correctly predicted. It can be thought of as the *accuracy* when correct negatives have been removed from consideration. That is, CSI is only concerned with forecasts that are important (i.e., assuming that the correct rejections are not important). Sensitive to hits, penalizes both misses and false alarms. Does not distinguish the source of forecast error. Depends on climatological frequency of events (poorer scores for rarer events) since some hits can occur purely due to random chance.

Non-linear function of POD and FAR. Should be used in combination with other contingency table statistics (e.g., Bias, POD, FAR).

Gilbert Skill Score (GSS)

Also commonly known as **Equitable Threat Score (ETS).**

$$\mathrm{GSS} = \mathrm{ETS} = \frac{\mathrm{Hits} - \mathrm{Hits}_{\mathrm{random}}}{\mathrm{Hits} + \mathrm{Misses} + \mathrm{False\ alarms} - \mathrm{Hits}_{\mathrm{random}}} = \frac{\mathrm{YY} - \mathrm{YY}_{\mathrm{random}}}{\mathrm{YY} + \mathrm{NY} + \mathrm{YN} - \mathrm{YY}_{\mathrm{random}}}$$

where

$$\mathrm{Hits}_{\mathrm{random}} = \mathrm{YY}_{\mathrm{random}} = \frac{(\mathrm{Hits} + \mathrm{False\ alarms})(\mathrm{Hits} + \mathrm{Misses})}{\mathrm{Total}} = \frac{(\mathrm{YY} + \mathrm{YN})(\mathrm{YY} + \mathrm{NY})}{\mathrm{Total}}$$

Answers the question: *How well did the forecast "yes" events correspond to the observed "yes" events (accounting for hits that would be expected by chance)?*

Range: -1/3 to 1; 0 indicates no skill. **Perfect score:** 1.

Characteristics: Measures the fraction of observed and/or forecast events that were correctly predicted, adjusted for the frequency of hits that would be expected to occur simply by random chance (for example, it is easier to correctly forecast rain occurrence in a wet climate than in a dry climate). The GSS (ETS) is often used in the verification of rainfall in NWP models because its "equitability" allows scores to be compared more fairly across different regimes; however it is not truly equitable. Sensitive to hits. Because it penalizes both misses and false alarms in the same way, it does not distinguish the source of forecast error. Should be used in combination with at least one other contingency table statistic (e.g., Bias).

1b. Statistics for continuous forecasts and observations – *excerpted from the WWRP/WGNE Joint Group on Forecast Verification Research website on Forecast Verificaiton: Issues, Methods and FAQ (http://www.cawcr.gov.au/projects/verification/)*

For this category of statistical measures, the grids of forecast and observed values – such as precipitation or reflectivity – are overlain on each other, and error values are computed. The grid of error values is summarized by accumulating values at all of the grid points and used to compute measures such as mean error and root mean squared error. This section is included for completeness but DTC will not be providing any continuous stats for this Spring Experiment.

These statistics are defined in the sub-sections below. In the equations in these sections, $\boldsymbol{f_i}$ signifies the **forecast value** at gridpoint *i*, $\boldsymbol{o_i}$ represents the observed value at gridpoint *i*, and $\boldsymbol{N}$ is the **total number** of gridpoints.

Mean error (ME)

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

Also called the **(additive) Bias.**

Answers the question: *What is the average forecast error?*

Range: minus infinity to infinity. **Perfect score:** 0.

Characteristics: Simple, familiar. Measures *systematic* error. Does not measure the magnitude of the errors. Does not measure the correspondence between forecasts and observations; it is possible to get a **perfect ME score for a bad** forecast if there are compensating errors.

Pearson Correlation Coefficient (r)

$$r = \frac{\sum_{i=1}^{n}(f_i - \overline{f})(o_i - \overline{o})}{\sqrt{\sum_{i=1}^{n}(f_i - \overline{f})^2}\sqrt{\sum_{i=1}^{n}(o_i - \overline{o})^2}}$$

where $\overline{f}$ is the average forecast value and $\overline{o}$ is the average observed value.

Also called the **linear correlation coefficient.**

Answers the question: *What is the linear association between the forecasts and observations?*

Range: -1 to 1. **Perfect score**: 1

Characteristics: *r* can range between -1 and 1; a value of 1 indicates perfect correlation and a value of -1 indicates perfect negative correlation. A value of 0 indicates that the forecasts and observations are not correlated. The correlation does not take into account the mean error, or additive bias; it only considers linear association.

Mean squared error (MSE) and root-mean squared error (RMSE)

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

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

MSE can be re-written as

$\mathrm{MSE} = (\overline{f} - \overline{o})^2 + s_f^2 + s_o^2 - 2s_f s_o r_{fo}$,

where $\overline{f}$ is the average forecast value, $\overline{o}$ is the average observed value, s_f is the standard deviation of the forecast values, s_o is the standard deviation of the observed values, and r_{fo} is the correlation between the forecast and observed values. Note that $\overline{f} - \overline{o} = \mathrm{ME}$ and $s_f^2 + s_o^2 - 2s_f s_o r_{fo}$ is the estimated variance of the error, s_{f-o}^2. Thus, $\mathrm{MSE} = \mathrm{ME}^2 + s_{f-o}^2$. To understand the behavior of **MSE**, it is important to examine *both* of these terms of **MSE**, rather than examining **MSE** alone. Moreover, **MSE** can be strongly influenced by **ME**, as shown by this decomposition.

The standard deviation of the error, s_{f-o}, is simply $s_{f-o} = \sqrt{s_{f-o}^2} = \sqrt{s_f^2 + s_o^2 - 2s_f s_o r_{fo}}$.

Note that the standard deviation of the error (ESTDEV) is sometimes called the "Bias-corrected MSE" (BCMSE) because it removes the effect of overall bias from the forecast-observation squared differences.

Answers the question: *What is the average magnitude of the forecast errors?*

Range: 0 to infinity. **Perfect score:** 0.

Characteristics: Simple, familiar. Measures "average" error, weighted according to the square of the error. Does not indicate the direction of the deviations. The RMSE puts greater influence on large errors than smaller errors, which may be a good thing if large errors are especially undesirable, but may also encourage conservative forecasting.

2. MODE Summary Metrics

The Method for Object-based Diagnostic Evaluation (MODE) identifies and matches spatial objects in the forecast and observed fields. A convolution radius (*r*) and a precipitation/reflectivity threshold (*t*) are used to identify objects; different combinations of these parameters lead to objects with different characteristics, and can be used to evaluate forecasts as a function of threshold and scale.

In the object matching and merging[1] process, all possible pairs of forecast and observed objects are assigned a total "interest" value. This value is formulated from the weighted sum of specific interest values that are associated with differences in particular attributes between the forecast and observed objects. According to the current weighting scheme, the total interest value is large when objects are located close to each other and are about the same size, and is smaller for pairs of objects that are further apart and have different sizes. Note that users can specify other components of interest, and their relative weights, in the configuration file for running MODE, according to what is most relevant for their particular application.

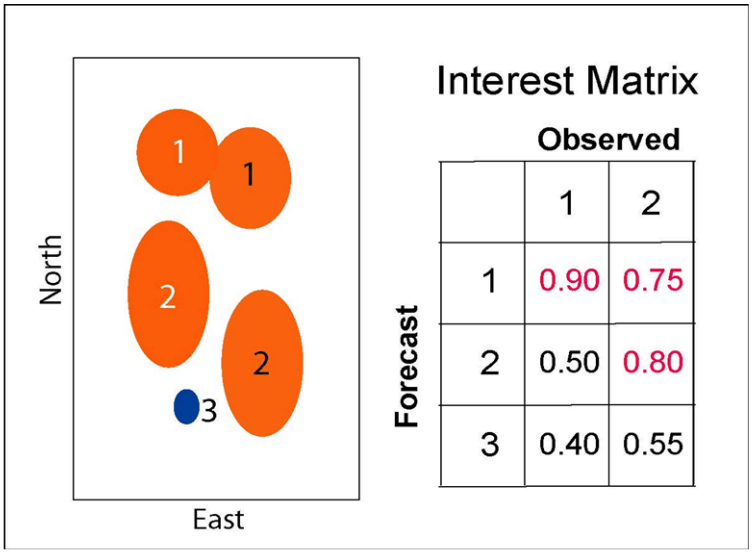


Figure 2 illustrates a scenario in which three forecast objects and two observed objects have been identified in the two fields. The total interest values for all of the pairs of forecast and observed objects are shown in the associated table. In previous work an interest threshold of 0.70 has been found to be a reasonable indicator of a good match. Thus, in this case, forecast object 1 is a good match with both observed objects 1 and 2, and forecast object 3 matches well with observed object 2. Forecast object 3 does not match well with either of the observed objects, mostly because of its small size. Because both forecast objects 1 and 2 match observed object 2, and forecast object 1 also matches observed object 1, these objects form a matched "cluster" in the forecast and observed fields.

Some of the forecast attributes that are (or can be considered) in determining matches between objects include object size, distribution of intensity values, orientation angle, and location. Comparisons of these attributes, along with the total interest values, also can be used to help measure the quality of the forecast performance.

[1] "Merging" refers to the connection of objects in the same field, while "matching" refers to the connection between objects in the forecast and observed field.

Median of Maximum Interest (MMI)

This measure is computed using the total interest values for all of the pairs of objects. It considers the maximum total interest values associated with each forecast object and each observed object. From this set, the median value is computed and is the MMI.

Example: Forecast and observed objects in Fig. 2
Maximum interest values for all of the forecast and observed objects are as follows:

For forecast object 1, the maximum total interest is 0.90.
For forecast object 2, the maximum total interest is 0.80.
For forecast object 3, the maximum total interest is 0.55.
For observed object 1, the maximum total interest is 0.90.
For observed object 2, the maximum total interest is 0.80.

The median of those 5 numbers is 0.80, so MMI = 0.80.

This number can be small because no objects match well, or because there are many extra objects that don't match well.

Larger MMI values imply a better match between forecast and observed objects.

Area-Weighted CSI (AWCSI)

Area Weighted Critical Success Index (AWCSI)

AWCSI = [(hit area weight) * #hits] / [(hit area weight * # hits) + (miss area weight * # misses) + (false alarm area weight * # false alarms)]

Where each area weight is the ratio of size of the (hit, miss, or false alarm) objects to the total area of all objects and # hits = number of matched objects; # misses = # unmatched observed objects; and # false alarms = # unmatched forecast objects.

Answers the question: How well did the forecast "yes" objects correspond to the observed "yes" objects?

Range: 0 to 1, 0 indicates no skill. Perfect score: 1.

Characteristics: Measures the area-weighted fraction of observed and/or forecast events that were correctly predicted. It can be thought of as the /accuracy/ when correct negatives have been removed from consideration, that is, /TS/ is only concerned with forecasts that count. Sensitive to hits, penalizes both misses and false alarms. Does not distinguish source of forecast error.
In a grid-based CSI each gridpoint that is counted in computing the CSI contributes represents an area with the same size but with MODE objects, the various objects can have a wide variety of sizes. Thus, area weighting makes sense. and observed objects.

Median Intersection over Area (MIA)

Ratio of intersection area to union area (unitless). Ranges from zero to one: One is perfect, smaller implies less overlap. This measure is the mean for all clusters of objects with interest values greater than 0.7.

Median Area Ratio (MAR)

Ratio of the areas of two objects defined as the lesser of the forecast area divided by the observation area or its reciprocal (unitless). The ideal value is 1, since this means that the forecast and observed objects are exactly the same size. Smaller implies that the forecast was either too small or too large. This measure is the mean for all clusters of objects with interest values greater than 0.7.

Median Centroid Distance (MCD)

Distance between two objects centroids (in grid units). Smaller is better, since this means the objects are closer. This measure is the mean for all clusters of objects with interest values greater than 0.7.

Median Angle Difference (MAD)

Difference between the axis angles of two objects (in degrees). This is only meaningful if objects seem to be more linear than circular, e.g. lines of thunderstorms. When they are linear, this measure tells you how well the angle of the forecast line matches the angle of the observed line. Smaller differences are better. This measure is the mean for all clusters of objects with interest values greater than 0.7.

Intensity with confidence intervals

10th, 25th, 50th, 75th, and 90th percentiles of intensity of the filtered field within the object (various units). This tells you the distribution of values within an object (think of this as the numeric equivalent of a boxplot). There are no ideal values. However, if you compare the distribution of values within a forecast object and an observed object, you would like them to match up. We recommend checking to see how close the median and 90^{th} percentile values are. This will tell you if you forecast is too intense or not intense enough. This measure is the mean for all clusters of objects with interest values greater than 0.7.

Median P50 Difference

First, the difference between the forecast and observed 50^{th} percentile intensity (median) for matched objects is calculated. The median of the difference for given time is then calculated and plotted.

Median P90 Difference

First, the difference between the forecast and observed 90^{th} percentile intensity for matched objects is calculated. The median of the difference for given time is then calculated and plotted.

Areal Coverage (ACOV)

Proportion of observation grid points inside the object. Intended to be used similar to Base Rate for traditional statistics.

3. Probabilistic Evaluation – *excerpted from the WWRP/WGNE Joint Group on Forecast Verification Research website on Forecast Verificaiton: Issues, Methods and FAQ (http://www.cawcr.gov.au/projects/verification/).*

A probabilistic forecast gives a *probability* of an event occurring, with a value between 0 and 1 (or 0 and 100%). In general, it is difficult to verify a single probabilistic forecast. Instead, a set of probabilistic forecasts, p_i, is verified using observations that those events either occurred (o_i=1) or did not occur (o_i=0).

An accurate probability forecast system has:

* *reliability* - agreement between forecast probability and mean observed frequency
* *sharpness* - tendency to forecast probabilities near 0 or 1, as opposed to values clustered around the mean
* *resolution* - ability of the forecast to resolve the set of sample events into subsets with characteristically different outcomes

Brier score -

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 = \frac{1}{N}\sum_{k=1}^{K} n_k(p_k - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o})$$

Brier score provides the user with a measure of *the magnitude of the probability forecast errors.* Measures the mean squared probability error. Murphy (1973) showed that it could be partitioned into three terms: (1) *reliability*, (2) *resolution*, and (3) *uncertainty*. These variables will also be made available during this Spring Experiment.

Range: 0 to 1. **Perfect score:** 0.

Characteristics: Sensitive to climatological frequency of the event: the more rare an event, the easier it is to get a good *BS* without having any real skill. Negative orientation (smaller score better) - can "fix" by subtracting *BS* from 1.

Brier skill score -

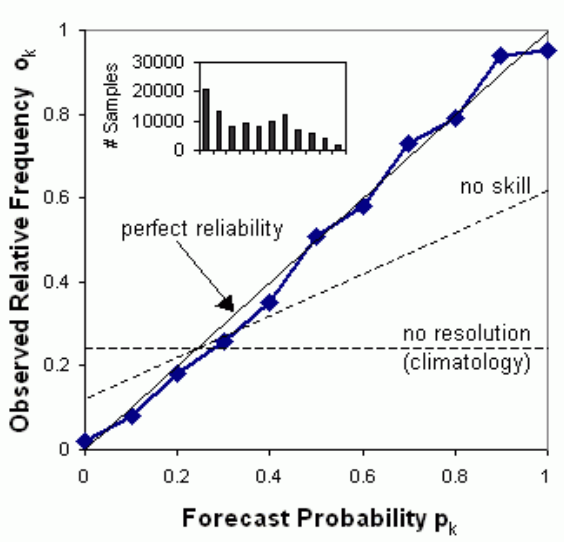
$$BSS = \frac{BS - BS_{reference}}{0 - BS_{reference}} = 1 - \frac{BS}{BS_{reference}}$$

Answers the question: *What is the relative skill of the probabilistic forecast over that of climatology, in terms of predicting whether or not an event occurred?*

Range: -∞ to 1, 0 indicates no skill when compared to the reference forecast. **Perfect score:** 1.

Characteristics: Measures the improvement of the probabilistic forecast relative to a reference forecast (usually the long-term or sample climatology), thus taking climatological frequency into account. Not strictly proper. Unstable when applied to small data sets; the rarer the event, the larger the number of samples needed.

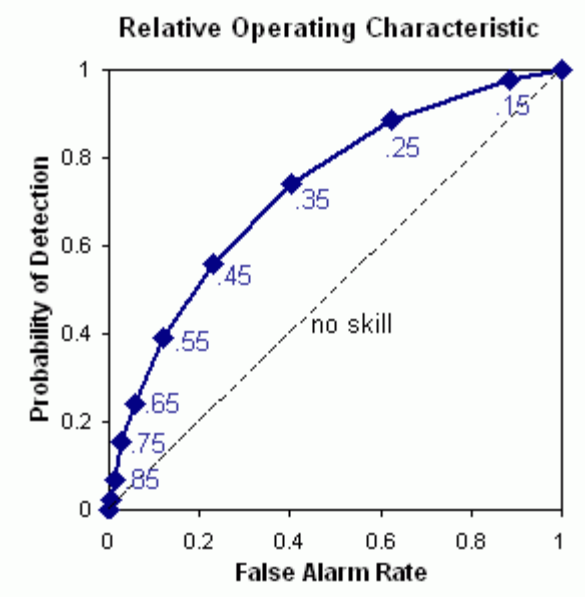
Reliability diagram - The reliability diagram plots the observed frequency against the forecast probability, where the range of forecast probabilities is divided into *K* bins (for example, 0-5%, 5-15%, 15-25%, etc.). The sample size in each bin is often included as a histogram or values beside the data points.



Reliability is indicated by the proximity of the plotted curve to the diagonal. The deviation from the diagonal gives the *conditional bias*. If the curve lies below the line, this indicates overforecasting (probabilities too high); points above the line indicate underforecasting (probabilities too low). The flatter the curve in the reliability diagram, the less resolution it has. A forecast of climatology does not discriminate at all between events and non-events, and thus has no resolution. Points between the "no skill" line and the diagonal contribute positively to the Brier skill score. The frequency of forecasts in each probability bin (shown in the histogram) shows the sharpness of the forecast. The reliability diagram is conditioned on the forecasts (i.e., given that X was predicted, what was the outcome?), and can be expected to give information on the real meaning of the forecast. It it a good partner to the ROC, which is conditioned on the observations.

Relative operating characteristic -Plot *hit rate (POD)* vs *false alarm rate* (*POFD*), using a set of increasing probability thresholds (for example, 0.05, 0.15, 0.25, etc.) to make the yes/no decision. The area under the ROC curve is frequently used as a score.

Answers the question: *What is the ability of the forecast to discriminate between events and non-events?*



ROC: **Perfect:** Curve travels from bottom left to top left of diagram, then across to top right of diagram. Diagonal line indicates no skill.
ROC area: **Range:** 0 to 1, 0.5 indicates no skill. **Perfect score:** 1

Characteristics: ROC measures the ability of the forecast to discriminate between two alternative outcomes, thus measuring resolution. It is not sensitive to bias in the forecast, so says nothing about reliability. A biased forecast may still have good resolution and produce a good ROC curve, which means that it may be possible to improve the forecast through calibration. The ROC can thus be considered as a measure of potential usefulness.
The ROC is conditioned on the observations (i.e., given that Y occurred, what was the correponding forecast?) It is therefore a good companion to the reliability diagram, which is conditioned on the forecasts.
More information on ROC can be found in Mason 1982, Jolliffe and Stephenson 2003 (ch.3), and the WISE site (http:/wise.cgu.edu/stdmod/measures6.asp).