# Ensemble Verification User's Manual

## Weather Forecast Research Team, University of British Columbia

Katelyn Wells

Last Updated: November 8th, 2011

## Table of Contents

## Quick Reference Summary

| Metric | Range | Desirable | Meaning |
|---|---|---|---|
| **Rank Histogram** | x: 0 to *n* ensemble members<br>y: 0 to 100 (% of forecasts falling in the corresponding bin) | Equal percent of all members in each bin (flat histogram) | Illustrates dispersion of ensemble members relative to observations |
| **Reliability Diagram** | x: 0 to 1 (Observed Frequency)<br>y: 0 to 1 (Forecast Probability) | Forecast probability = Observed frequency (Line $y = x$) | Indicates how well the probability forecast of an event corresponds to the observed frequency |
| **Sharpness** | x: 0 to *K* probability bins used in the Reliability Diagram<br>y: 0 to 1 (Portion Forecasts falling into each bin) | Want narrow peak at either extreme of *x*-axis<br>(a forecast system that produces a lot of 0 and 100 % forecast probabilitys) | Property of the forecast system only. Shows the distribution of the forecast probabilities<br>*Note:* Sharp forecasts will only be accurate if they also exhibit good reliability |
| **ROC Diagram:** | x: 0 to 1 (False Alarm Rate)<br>y: 0 to 1 (Hit Rate) | More hits, less false alarm rates (all points in the upper left corner) | Shows frequency of hit rates compared to false alarm rates |
| **Brier Score** | 0 to 1 | 0 | The magnitude of the forecast errors. See page 15 for variations in the calculation |
| **Brier Skill Score** | -∞ to 1 | 1 = perfect<br>0 = no skill relative to reference forecast (or uncertainty)<br>< 0 = worse than reference forecast but still may have skill | Brier Score converted into a positively oriented skill score |
| **Reliability** | ≥ 0 | Positive small number (Ideally 0) | Indicates how well the probability forecast of an event corresponds to the observed frequency |
| **Relative Reliability** | 0 to 1 | 0 | Reliability÷Uncertainty<br>The reliability (as defined above) given |

| | | | |
|---|---|---|---|
| | | | the inherit Uncertainty |
| **Resolution** | ≥ 0 | Close to 1 | The ability of the probability forecast system to differentiate between possible observed oucomes |
| **Relative Resolution** | 0 to 1 | 1 | Resolution ÷ Uncertainty<br>The resolution (as defined above) given the inherit Uncertainty |
| **Uncertainty (Degree of Difficulty)** | 0 to 1 | N/A<br>(Property of the observation only, For BS = 0, Uncertainty = 1) | Difficulty in forecasting the event<br>In a perfect forecast Uncertainty = Resolution, therefore BS = 0 |
| **CRPS** | ≥ 0 | 0<br>(Negatively oriented, smaller values are preferable) | Integral of all possible Brier Scores |
| **CRPS potential** | ≥ 0 | 0<br>(Negatively oriented, smaller values are preferable) | CRPS given a perfectly reliable forecast (Reliability = 0) |
| **CRPSS** | -∞ to 1 | 1 = perfect<br>0 = no skill relative to reference forecast (or uncertainty)<br>< 0 = worse than reference forecast but still may have skill | The CRPS converted into a positively oriented skill score (Calculated in the same way as BSS) |
| **Mean Absolute Error Skill Score** | -∞ to 1 | 1 = perfect<br>0 = no skill relative to reference forecast (or uncertainty)<br>< 0 = worse than reference forecast but still may have skill | MAE converted into a positively oriented skill score based on climatology |

## Web User's Interface

There are a variety of graphical products and tables for each station and forecast day.
Steps:

1. Select the station or region name (only one can be chosen at a time)
2. Select the forecast day (only one can be chosen at a time)
3. Select the variable (only one can be chosen at a time)
4. Select the metrics (one up to all can be chosen)
5. Select the threshold (the metrics requiring thresholds will have a choice of viewing 'All', two, or one threshold)

## Rank Histogram

*Answers the question*: how well does the ensemble spread of the forecast represent the true variability (uncertainty) of the observations?

### *Description:*

The rank histogram is a common approach to evaluating whether a collection of ensemble forecasts satisfy a consistency condition. This means the histogram is used to evaluate the spread of the ensemble. It is a metric that displays where and how often the verifying observation usually falls with respect to the ensemble forecast data.

To create a rank histogram for each of the *n* forecasts of a given variable the $n_{ens}$ ensemble member forecasts are ranked from lowest to highest along the *x*-axis, creating $n_{ens} + 1$ bins (including the bins higher and lower than the two most extreme members). The observation corresponding to that $n^{th}$ forecast is then placed in the appropriate bin (between whatever two ensemble members it falls between). This is repeated *n* times (for each forecast) in the evaluation period, creating a histogram. In an ensemble with perfect spread, the observations verify equally between each of the members, and the histogram is flat, meaning each member represents an equally likely scenario (fig. 1b). This is a good means for detecting systematic flaws in an ensemble system.

### *Interpretation:*

Interpretation is based on the comparison of the shape of the histogram to perfect rank uniformity line (the straight line that represents perfect spread). There are five main deviations that can illustrate the characteristic ensemble dispersion and bias errors of the forecast.

1. *Overforecasting Bias*—Fig. 1a—The histogram is skewed right, meaning the bars are higher on the left than the right. This results from average forecast probabilities that are larger than the observed relative frequencies. More observations fall in the lower bins thus the forecasted values are too high.
2. *Rank Uniformity*—Fig. 1b—A flat histogram indicating the ensemble spread is about right to represent forecast uncertainty. It results for them observation being equally distributed between the ranked ensemble members over time.
3. *Underforecasting Bias*—Fig. 1c—The histogram is skewed to the left, meaning the bars are higher on the right than on the left. This results from forecast probabilities that are smaller than the corresponding observed relative frequencies. More observations fall into higher bins thus the forecasted values are too small.
4. *Underdispersion (overconfident)*—Fig. 1d—This produces a U-shaped histogram and occurs when the ensemble spread is too small, with many observations falling outside the extremes of the ensemble. This happens when the ensemble members tend to be too much like each other, and different from the observation.
5. *Overdispersion (underconfident)*—Fig. 1e—The ensemble produces histograms with relative frequencies that are highest in the middle ranks. This means that the ensemble spread is too large with most observations falling near the center of the ensemble. The range of the ensemble is excessive.

Note:
A flat histogram does not necessarily indicate a good forecast; it only measures whether the observed probability distribution is well represented by the ensemble.

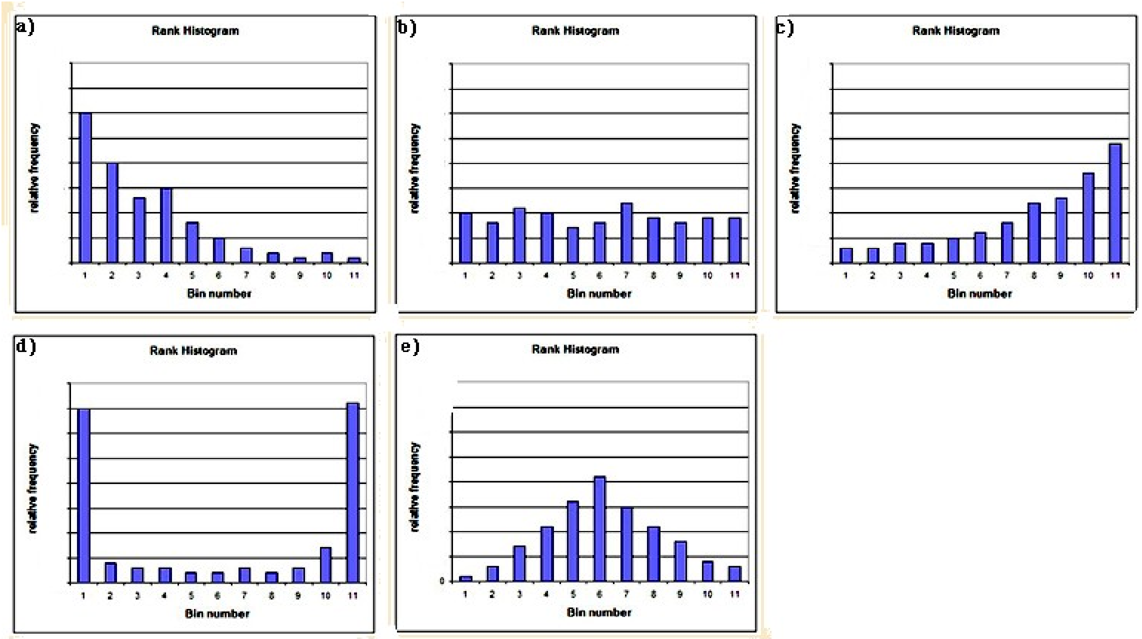


**Figure 1—Rank Histogram**

*http://www.eumetcal.org/resources/ukmeteocal/verification/www/english/msg/ver_prob_forec/uos4b/uos4b_ko1.htm*

## Reliability Diagram and Sharpness

*Answers the question:* how well do the predicted probabilities of an event correspond to their observed frequencies?

*Description:*

A reliability diagram is used for assessing probability forecasts for binary events such as the probability of measurable precipitation. It measures the accuracy with which a discrete event is forecast by an ensemble or probabilistic forecasting system.

The diagram is created by plotting *f(q)* vs. *q* where *f(q)* is the conditional probability of the event to occur given that the forecast probability was equal to *q*. *f(q)* can easily be estimated by counting the relative frequency of the observed event over cases when event (lets call it A) was forecast to occur with probability *q*. This is called the observed relative frequency. 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 sampling uncertainty will decline as the width of the bin increases but the precision of the diagram will also decline. The proportion of forecast probabilities in each bin is often included as a histogram, called a sharpness diagram (see below).

A forecast is defined as reliable if *f(q) = q* (i.e. perfect reliability). Perfect reliability is also plotted on the same diagram, which is the dashed line *f(q) = q* ( i.e., the line *y* = *x*). If the forecast is perfectly reliable the observed fraction within each bin will equal the average of the associated forecast probabilities. The deviations from this line are what determine the level of resolution of the binary forecast, (i.e., it represents bias in the forecast probabilities, notwithstanding sampling uncertainty).

When assessing a diagram it must be noted that deviations from the diagonal are not necessarily indicative of true deviations from reliability but can also be due to sampling variations (i.e. variations in data collection or station readings). Thus when the statistics are based on a finite sample the reliability curve for even a perfectly reliable forecast system is expected to exhibit sampling variations around the diagonal.

### *Sharpness Diagram*

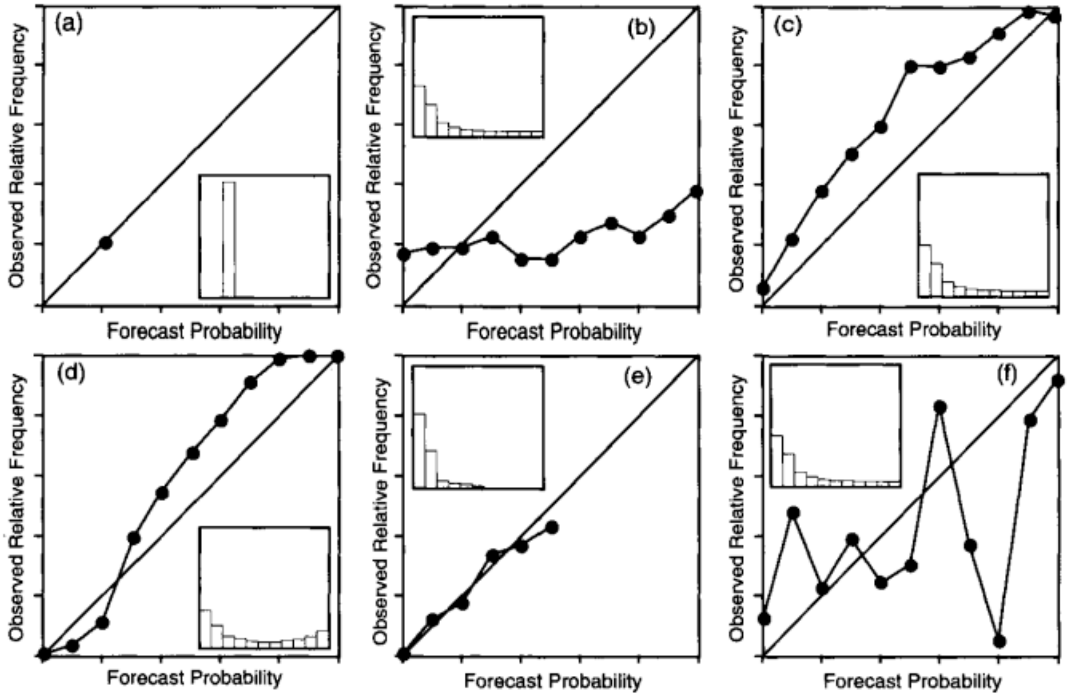
This histogram in the lower corner of the reliability diagram is a sharpness diagram. It is a property of the forecast only and does not take into account how the probability forecast verifies against the observations. It shows portion of forecasts that fall into each probability bin. This means it shows the tendency of the ensemble forecast system to predict probabilities near *0* or *1*, as opposed to more non-committal probabilities. A sharper probabilistic forecast tends to forecast the more extreme probabilities (0 or 1 near the extremes of the *x*-axis).

### *Interpretation:*

Here we compare the slope of the reliability diagram (called a calibration function) with perfect reliability. There are five main deviations that can illustrate the resolution or calibration level of the forecast (Fig. 2).

1. *Underforecasting (low bias)*—Fig. 2c— The calibration curve is entirely to the left of perfect reliability. This indicates that the forecast probabilities are consistently too low relative the observed frequencies. This is unconditional bias and shows that the forecasts are miscalibrated or not reliable.
2. *Overforecasting (high bias)*— not shown—The calibration function is entirely to the right of the perfect reliability line. This indicates that the forecast probabilities are consistently too high compared to the observed frequency (opposite to what is shown in Fig. 2c).
3. *Good resolution (underconfident)*—Fig. 2d— More subtle and indicates conditional bias. The calibration curve straddles the perfect reliability line starting on the right side then moving to the left side. This shows overforecasting biases associated with lower forecast probabilities and underforecasting biases associated with high forecast probabilities (see above for definitions of these kinds of biases).

4. *Poor resolution (overconfident)*—not shown— is the reverse of *good resolution*, where the calibration curve starts on the left side of the perfect line, then moves to the right side of the line, indicating underforecasting biases associated with lower forecast probabilities and overforecasting biases associated with higher probabilities. (Opposite to Fig. 2d)
5. *Good Calibration*—not shown— is when the calibration curve lines up with the perfect line (not exactly but with slight variations as already discussed).



## Figure 2—Reliability Diagrams

*Wilks (1995)*

*Figure 2 (a):* Climatological Forecasts
*(b):* Forecasts exhibiting minimal resolution
*(c):* Forecasts showing an underforecasting bias
*(d):* Forecasts showing good resolution
*(e):* Reliable forecasts of a rare event
*(f):* Verification data set limited by small sample size

*Notes specific to our products:*

This metric is calculated by using observed and forecasted departures from climatology. Climatology is defined as a 10-day moving average of 30 years of observations (approximately 1980-2010). Thresholds are based on this climatology.
Thresholds used:

- 24-hour accumulated precipitation: Absolute departure of 5mm, 10mm, 25mm, and 50mm

- Maximum daily temperature and minimum daily temperature: Absolute departure of 2 ˚C, 5 ˚C, and 10 ˚C

For the first run of this metric there is only one station per graph. The length of the data set is confined to the number of usable days in the water year (those with all 42 ensemble members and a corresponding observation). This means that a full year will be a few days short of 365 days because every day with missing data is skipped.
The number of probability bins used in the metric is 11 (i.e., K = 11). Since the data set is relatively small using more bins creates graphs that are 'jumpy' and unrealistic (Fig. 1f).

The second run of this metric now includes many stations per graph, lengthening the data set. Below is a list of stations, by code, that are sorted into different regions. Some stations may be missing for two different reasons: one, there are no forecasts for this station for the water year, or there is no climatological data available at this time.

Water year: 2009-2010:

Lower Mainland
YXX WZA ALU COQ CQM GOC YHE WAH STA STV YVR

Vancouver Island
BCK YBL YQQ CMX CRU ELK ASH ERC GLD HEB SCA YYJ WOL *Missing: GLD, CRU, HEB*

Columbia Region
BAR YCP YCG YXC CRS DCD EAC FQR WGE RAD MCD MOL MOR MTR PKM YRV RGR *Missing: EAC, MOL, MOR, MTR, PKM*

Peace River
FIN YXJ YGS YQU HFF ING KWA YZY PAR ALC DNV PHH PCP TPR PYN YXS SUN YOJ YPE *Missing: HFF, KWA, PAR, ALC, DVN, PHH, PCP, TPR, PYN, SUN*

Pacific Northwest
GEG SEA PDX

California
BFL LAX SFO *Missing: LAX, SFO*

Cheakamus/Clowhom
CMU CLO CMS YPW WAE

Bridge
BLN BRI LJU MIS NTY SON

North/Central Coast
YBD YPR YZT YYD TAH YXT *Missing: YBD, YPR, YZT, YYD, TAH, YXT (all)*

# ROC Diagram

*Answers the question:* How well can a forecast discriminate between events and non-events?

*Description:*

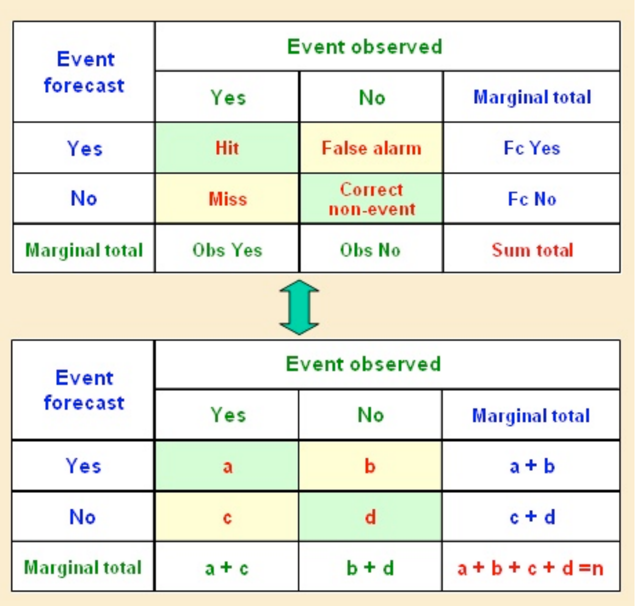
The Relative Operating Characteristic, or Receiver Operating Characteristic (ROC) diagram is a graph of Probability of Detection (hit rate) against Probability of False Detection (false alarm rate) as a predetermined threshold changes. It is discrimination-based forecast verification but unlike the reliability diagram the ROC diagram does not provide a full depiction of the joint distribution of forecasts and observations. It is only conditioned on the observation. Hit rate (H): The total number of correct event forecasts (hits) divided by the total number of events observed. False alarm rate (F): The instances where the event was predicted but did not occur divided by the total number of non-events observed. The whole curve is plotted in a unit square (*1 x 1*).

It must be noted that the ROC diagram is not a measure of the bias in the forecast, so reliability is not described.  A good ROC curve may still be produced even if the forecast has bias. *The ROC diagram is then a good companion to the reliability diagram, which is conditioned on the forecasts.*

*F* and *H* are calculated based on a contingency table (see Fig. 3),

$$F = \frac{b}{b+d} \quad , \; H = \frac{a}{a+c}$$

Where *a, b, c,* and *d* are the various parts of the contingency table. *a* is number of times a "yes" forecast was followed by a "yes" occurrence i.e. a "hit". *b* is the number of times a "yes" forecast was followed by a "no" occurrence i.e. a  "false alarm". *c* is the number of times a "no" forecast was followed by a "yes" occurrence i.e. a "miss". *d* is the number of times a "no" forecast was followed by a "no" occurrence i.e. a "correct non-event".



| Event forecast | Event observed | | |
|---|---|---|---|
| | Yes | No | Marginal total |
| Yes | Hit | False alarm | Fc Yes |
| No | Miss | Correct non-event | Fc No |
| Marginal total | Obs Yes | Obs No | Sum total |

| Event forecast | Event observed | | |
|---|---|---|---|
| | Yes | No | Marginal total |
| Yes | a | b | a + b |
| No | c | d | c + d |
| Marginal total | a + c | b + d | a + b + c + d =n |

## Figure 3—ROC Contingency Table

*http://satreponline.org/vesa/verif/www/english/msg/ver_categ_forec/uos1/uos1_ko1.htm*

*Interpretation:*

Forecasts with better discrimination exhibit ROC curves approaching the upper left corner of the diagram more closely (i.e., high Probability of Detection and relatively low Probability of False Alarm indicates good performance [Fig. 3a]). Forecasts with very little ability to discriminate between events and non-events exhibit ROC curves very close to the line $y = x$, i.e., when hit rate = false alarm rate (or $H = F$ [Fig. 3b]). The area under the curves is used as a score, measuring discrimination. The area for a perfect discrimination is the unit square, $A_{perf} = 1$. The area of no skill is $A_{rand} = 0.5$. The discrimination is the ability of the forecast to correctly differentiate between that actual observed value and other potential outcomes. The area is the percentage of forecasts that discriminate correctly. The area $A$ under the ROC curve of interest can be expressed in standard skill score form: $SS_{ROC} = 2A - 1$.

*Notes specific to our products:*

This metric is calculated by using observed and forecasted departures from climatology. Climatology is defined as a 10-day moving average of 30 years of observations (approximately 1980-2010). Thresholds are based on this climatology.
Thresholds used:

- 24-hour accumulated precipitation: Absolute departure of 5mm, 10mm, 25mm, and 50mm
- Maximum daily temperature and minimum daily temperature: Absolute departure of 2 ˚C, 5 ˚C, and 10 ˚C

Often for higher thresholds there will be only one or two forecasts above this threshold, or there may not be any. In these cases a ROC diagram cannot be produced and 'No Figure Found!' will appear.

The second run of this metric now includes many stations per graph, lengthening the data set. Below is a list of stations, by code, that are sorted into different regions. Some stations may be missing for two different reasons: one, there are no forecasts for this station for the water year, or there is no climatological data available at this time.

Water year: 2009-2010:

Lower Mainland
YXX WZA ALU COQ CQM GOC YHE WAH STA STV YVR

Vancouver Island
BCK YBL YQQ CMX CRU ELK ASH ERC GLD HEB SCA YYJ WOL *Missing: GLD, CRU, HEB*

Columbia Region

BAR YCP YCG YXC CRS DCD EAC FQR WGE RAD MCD MOL MOR MTR PKM YRV RGR *Missing: EAC, MOL, MOR, MTR, PKM*

Peace River
FIN YXJ YGS YQU HFF ING KWA YZY PAR ALC DNV PHH PCP TPR PYN YXS SUN YOJ YPE *Missing: HFF, KWA, PAR, ALC, DVN, PHH, PCP, TPR, PYN, SUN*

Pacific Northwest
GEG SEA PDX

California
BFL LAX SFO *Missing: LAX, SFO*

Cheakamus/Clowhom
CMU CLO CMS YPW WAE

Bridge
BLN BRI LJU MIS NTY SON

North/Central Coast
YBD YPR YZT YYD TAH YXT *Missing: YBD, YPR, YZT, YYD, TAH, YXT (all)*

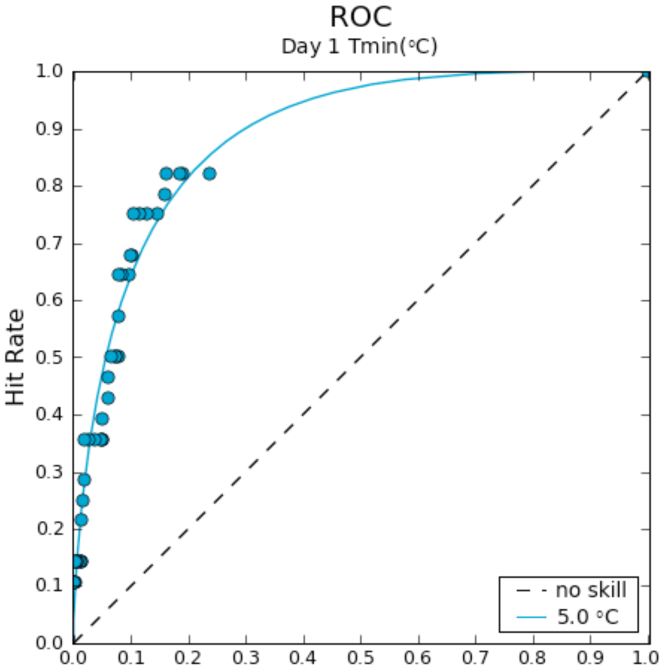


**Figure 3a—ROC Good Discrimination**

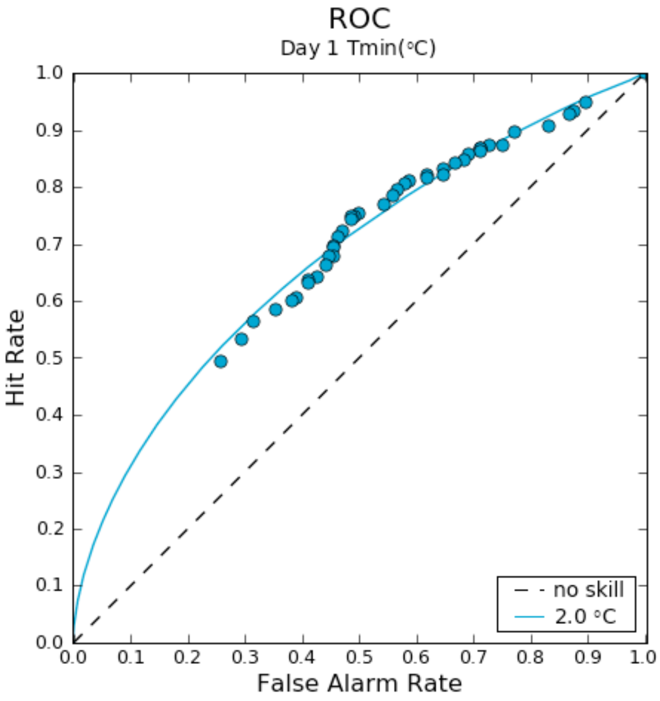


**Figure 3b—ROC Poor Discrimination**

## Brier Score and Brier Skill Score

*Brier Score answers the question*: what is the magnitude of the probability forecast errors?
*Brier Skill Score answers the question:* what is the relative skill of the probabilistic forecast over that of climatology or other reference forecasts, in terms of predicting whether or not an event occurred?

*Description/Interpretation:*
The Brier Score (BS) is the mean squared error of the probability forecasts. The error is (fcst prob – obs)$^2$ , where obs = 1 if the event occurs and obs = 0 if the event does not occur. The score then averages the squared differences between the pairs of forecast probabilities and matching binary observations. For perfect forecasts *BS = 0* and less accurate forecasts receive higher Brier Scores, and the score only takes on the values between *0* and *1*.

The Brier Skill Score (BSS) is also calculated: $BSS = 1 - BS/BS_{ref}$. $BS_{ref}$ is the Brier Score of the reference forecasts, which may be from relevant climatological data, or other forecasts. The range of the BSS is negative infinity to *1*, where *1* is a perfect skill score and where 0 indicates no skill when compared to the reference forecast.

The nice thing about the BS is that it can be decomposed into three components: reliability, resolution and uncertainty. *BS = "reliability" - "resolution" + "uncertainty"*:

- Reliability tests how well the forecast probability of an event corresponds to the observed frequency, and is represented in the Reliability diagram (see Fig. 2). It summarizes the calibration or conditional bias of the forecasts.

- Resolution is the degree to which the forecast probabilities can sort observations into subsamples of events with different frequencies of occurrence. If the forecasts can differentiate from climatology this term will be large.

- Uncertainty is the variability of the observations. It is only a function of the observations. The greater the uncertainty, the more difficult the forecast will be.

Since a small BS is ideal, the reliability term ideally should be small whereas the resolution term should be large (or the same magnitude as the uncertainty term).

*Notes specific to our products:*
This metric is calculated by using observed and forecasted departures from climatology. Climatology is defined as a 10-day moving average of 30 years of observations (approximately 1980-2010). Thresholds are based on this climatology.
Thresholds used:

- 24-hour accumulated precipitation: Absolute departure of 5mm, 10mm, 25mm, and 50mm
- Maximum daily temperature and minimum daily temperature: Absolute departure of 2 ˚C, 5 ˚C, and 10 ˚C

There are two different Brier Scores and Brier Skill Scores that are displayed. These are the 'Direct' and 'Decomposed'.

*BS Direct:* calculated by: $\frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2$, where *y* is the forecast probability and *o* is the matching observation, *n* is the number of forecasts within each probability category, $o_k$ is the number of observations within each probability category, and the number of probability category bins is equal to the number of ensemble members.

BS Decomposed: is calculated by $Reli\ = \frac{1}{\mathrm{n}}\sum_{i=1}^{I} N_i(y_i - \bar{o})^2$, $Resol = \frac{1}{n}\sum_{i=1}^{I} N_i(\bar{o}_i - \bar{o})^2$, and $Unc\ = \bar{o}(1 - \bar{o})$, Making the BS decomposition: $BS = Reli - Resol + Unc$, where *n, o, and y* are as above, and K is the number of probability bins used to calculate $N_i$ and $o_i$ (in our case K = 11)
Note: since the number of bins (K) is not the same as the number of ensemble members these Brier Scores for direct and decomposed have different values.

When there is a divide by zero error (because the threshold was so high there were no observed events), one of two values will appear

- *nan*: If no score is calculable
- -∞: If -∞ is within the calculable range(e.g. for BSS)

## CRPS Decomposition

*Answers the question:* what is the forecasts ability to produce sharp and appropriately centered probability distributions?

*Description/Interpretation:*

The continuous ranked probability score (CRPS) is a generalization of the Brier score. Instead of two classes (the event occurs or does not occur) like the Brier score, the range of the parameter of interest is divided into an infinite number of classes.   The CRPS is sensitive to the entire permissible range of the parameter of interest and does not require the introduction of a number of predefined classes. It can also be interpreted as an integral over all possible Brier scores (i.e., the intergrated difference between the forecasted and observed CDF's). For a deterministic forecast the CRPS is equal to the mean absolute error. This means ensemble forecasts can be compared to deterministic forecasts using the CRPS

The CRPS can also be decomposed into reliability, resolution and uncertainty:

- Reliability tests whether for each bin *i* on average the verifying analysis was found to be with a fraction of *i/N* below this bin, which relates very closely to the reliability diagram. $\text{CRPS}_{\text{potential}}$ is also calculated which is the CRPS that would occur if the system were perfectly reliable.

- Resolution expresses the improvement gained by issuing probability forecasts that are case dependant.  It is the difference between the $\text{CRPS}_{\text{potential}}$ and the climatological uncertainty.

- Uncertainty is equal to the potential reliability for a forecast system based on the sample climatology data.

The CRPS is negatively oriented, meaning that smaller values are better (0 being preferred).

## Mean Absolute Error Skill Score

*Description/Interpretation:*

The skill score is the measure of the relative quality of the forecasting system compared to some reference forecast.  Here the reference forecast is either the direct model output or the score for a different model, or most often climatology. In our case we have used climatology based on a 10 day running average of 30 years of observations (1980- 2010). The general form of a skill scores is: $SS = (S - S_o) / (S_1 - S_o) \; x \; 100$, where $S$ is the forecast score $S_o$ is the reference forecast score and $S_1$ is the best possible score. The mean absolute error skill $S$, the forecast score: $S = MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$, where $f$ is the forecast value and $y$ is the corresponding observation. The skill score ranges between *0* and *1* (*0* to *100%*).  A value of *1* is the maximum value and indicates a perfect forecast, and a value of *0* indicates a model forecast equivalent to the reference forecast.  A negative value implies that the model is worse than the reference, but does not necessarily mean that the model has no skill at all. Skill scores have the advantage over raw scores because they help take account of non-stationarities in the system to be forecast.

## Relative Mean Absolute Error

*Description/Interpretation:*

RMAE is similar to the Z-score, and is useful for comparing stations in different regions with differing observed variability, which makes some stations innately more difficult to forecast for.

For a given forecast variable, the Mean Absolute Error (as described above) is divided by its standard deviation. The standard deviation we use is calculated by computing standard deviations for each day of the year using the 30-year climatology, then smoothing them using a 10-day running mean.

Note: Since 24-hour accumulated precipitation cannot be assumed to be a normal distribution the RMAE cannot be calculated for this variable.

## Cost/Loss Evaluation

*Answers the question:* Given a user with a certain cost/loss ratio, how high does forecast confidence have to be, such that it is economically beneficial for a user to take preventative action (with associated costs) to protect against a loss?

*Description/Interpretation:*

Cost/loss evaluation is helpful when assessing the economic risk associated with making a decision based on a forecast. In general, the user has two alternative courses of action: either do nothing and risk incurring a loss $L$, or take some sort of preventative action of cost $C$ to protect against loss $L$. Table 1 is a 2 x 2 matrix showing the possible combinations of action and occurrence. An example of an "event" might be exceeding a rainfall threshold, say, 100mm, over a watershed, which would force a reservoir operator to spill water. The preventative action would be to lower the reservoir ahead of the event, at a cost of running the generator with reduced water height, which reduces the amount of power generated. If the operator doesn't take action, and the event occurs, then they would suffer a loss $L$ related to the amount of water spilled. If the event doesn't occur, then they have no costs or losses.

| **Action taken** | **Event occurs** | |
|---|---|---|
| | Yes | No |
| Yes | *C* | *C* |
| No | *L* | *0* |

**Table 1**

The decision of whether or not to take preventative action is based on the skill of the model, the model's forecasted probability of occurrence of the event, the climatological rate of occurrence of the event, and the expenses (costs and losses) unique to the user. With a probabilistic forecast system, for a given user and event, one can calculate a probability threshold (e.g., 60% chance of occurrence), beyond which the user should take preventative action.

For our evaluation purposes on the webpage, we're interested in which forecast system provides the greatest value. The value $V$ of a forecast is the reduction in expense gained from the forecast, relative to that from a perfect forecast:

$$V = (E_{climate} - E_{forecast})/(E_{climate} - E_{perfect})$$

The expense is a function of the $C/L$ ratio unique to the user. The user should determine his or her $C/L$ ratio, and then use the plots to determine which, if any, forecast provides the most value. If none of the forecasts provide value for their $C/L$ ratio, then the user should either a) always take preventative action (expense = $C$), or b) never take preventative action, and accept loss when it happens (expense = $\bar{O}L$, where $\bar{O}$ is the climatological rate of occurrence), whichever entails a lower expected expense.

## Definitions:

*Cumulative Distribution Function (CDF):* A distribution function in probability theory and statistics that completely describes the probability distribution of a real-value random variable. Cumulative distribution functions are also used to specify the distribution of multivariate random variables. It is the probability for a realization of the variable to be less than $x$ for any $x$. This probability is denoted by $F(x) : F(x) = P\{X \leq x\}$ $F(x)$ for a variable X. The CDF of a random variable is a monotonically increasing function from 0 to 1.

*Probability Density Function (PDF):* A function that describes the relative likelihood for this random continuous variable to occur at a given point in the observation space. The probability of a random variable falling within a given set is given by the integral of its density over the set.

*Random Variable:* A variable whose value is a numerical outcome of a random phenomenon.

*Discrete random variable:* A variable X that has a finite number of possible values. The probability distribution of X lists the values and their probabilities. Each probability is a number between 0 and 1 and all probabilities added together equal 1. To find the probability of any event add the probabilities of the particular values that make up the event.

*Continuous random variable:* A continuous random variable Y takes all values in an interval of numbers. The probability distribution of Y is described by a density curve. The probability of any event is the area under the density curve and above all the values of Y that make up the event.

# References

Hersbach, H., 2000: *Decomposition of the Continuous Ranked Probability Score for Ensemble Preiction Systems*. Weather and Forecasting, 15, 559-70.

McCollor, D. and R. Stull, 2008: *Hydrometeorological Short-Range Ensemble Forecasts in Complex Terrain. Part I: Meteorological Evaluation*. Weather and Forecasting, 23, 533-556.

Jolliffe, I.T., B.D. Stephenson, 2003: *Forecast Verification: A practitioner's Guide in Atmospheric Science*. J. Wiley, 254 pp.

Wilks, D.S., 2006: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.