Guidelines for Verification of Ensemble Forecasts

0. Aim of the document

The aim of this document is to provide some guidelines and a theoretical background for the common methods used to verify probabilistic and ensemble forecast systems. These guidelines will be used as a starting point for incorporating the relevant probabilistic scores and accompanying graphics in the VERSUS software package. The information contained within this document will also form part of the VERSUS User's Manual.

1. Introduction

The verification of a deterministic weather prediction system consists of the comparison of gridded model output, which can be interpolated or not, with point observations. A number of statistical scores evaluate different aspects of model performance while the forecast "error" is defined simply as the difference between the forecast value and the observation. The uncertainty associated with the forecast value is, however, not estimated. An Ensemble Prediction System (EPS), which is a probabilistic forecast system, aims to quantify this uncertainty using a set of perturbed Initial Conditions (ICs) and/or perturbed model formulations. Verification methods applied to ensemble forecasts have two main objectives: 1) to assess the characteristics of the ensemble distribution and 2) to verify the probability forecasts. EPS forecasts represent only one category of probabilistic forecasts; others types are associated with a dichotomous forecasted parameter or with a consistent set of probability values assigned to several categories of the predicted parameter.

In general, four aspects must be verified in order to properly measure the quality of an ensemble system: 1) equal likelihood of each ensemble member, 2) superiority of ensemble mean to single control forecast, 3) high spread-skill relation and 4) reliable probability. It should be noted that these four aspects are interrelated. Since all perturbed ICs could possibly be true and all perturbed physics or varying physics schemes or alternative models are also equally plausible, the performance of all ensemble members should, in principle, be equivalent to one another on average. If this is not the case, it is indicative of problems with the choice of ensembling technique employed. For example, either the IC perturbations are too large or alternative models, physics schemes or perturbations are not equally plausible.

A number of statistical scores have been developed and are applied in order to evaluate the usefulness of an EPS forecast system with respect to each of the four aforementioned aspects. In the following sections, the main attributes of an EPS forecast system will be presented and a set of the most common statistical scores, which will be implemented in the VERSUS software package, and their significance will be discussed. In addition, examples of the graphical representation of these statistical scores, which must be developed will also be provided.

2. Basic concepts

2.1 Statistical Framework

In the case of a dichotomous predictant, the same statistical framework for verification such as that described by Wilson (2002) can be applied with some simplifications. The joint distribution of forecasts and observations can be represented as p(x,f), where f represents the forecasts and x represents the observations, and p(f,x) is the joint probability of f and x. This joint distribution can be factored in two different ways:

(a) as the calibration-refinement (CR) factorization,

$$p(f, x) = p(x|f)\, p(f) \qquad (1)$$

where $p(x|f)$ is the conditional distribution of observations given the forecast, and $p(f)$ is the marginal (i.e., unconditional) distribution of the forecasts;

and (b) using the likelihood-base rate (LBR) factorization,

$$p(f, x) = p(f|x)\, p(x) \qquad (2)$$

where $p(f|x)$ is the conditional distribution of forecasts given the observation, and $p(x)$ is the marginal distribution of the observations. Each factorization involves the combination of a conditional distribution and a marginal distribution. The CR factorization (1) involves the conditional distribution of *observations given forecasts* (called *"calibration"*) and the marginal distribution of *forecasts* (called *"refinement"*). The likelihood-base rate factorization (2) involves the conditional distribution of *forecasts given observations* (called the "*likelihood*") and the marginal distribution of *observations* (called the "*base rate*"). In the case of probabilistic forecasts of a dichotomous event, the verification framework is greatly simplified because there are only two possible observations.

Differences between $p(x)$ *and* $p(f)$ describe the unconditional biases in the forecast probabilities. The conditional distribution $p(x|f)$ describes the conditional reliability of the forecast probabilities when compared to $p(f)$ and "resolution" when only its sensitivity to $p(f)$ is being considered. For a given level of reliability, forecasts that contain less uncertainty, i.e. ''sharp forecasts'', may be preferred over ''unsharp'' ones since they contribute less uncertainty to decision making. In contrast, $p(f|x)$ measures the ability of the forecasts to ''discriminate'' between different observed outcomes. An ensemble forecasting system is discriminatory with respect to an event if it consistently forecasts the event's (observed) occurrence with a probability higher than chance (i.e. climatology) and consistently forecasts its (observed) non-occurrence with a probability lower than chance. In general, the utility of a forecasting system will depend on several attributes of forecast quality (Jolliffe and Stephenson, 2003). In the following section, a short description of each one of these attributes is provided.

2.2 Attributes

Reliability is a measure of how closely the forecast probabilities correspond to the conditional frequency of occurrence of the event. For example, when a probability forecast of 0.20 is issued, we would expect the event to occur 20% of the time. Reliability is a measure of how well this holds up in reality. It should be noted that a forecasting system that simply forecasts the climatological probabilities of events may be reliable, but is not useful. This aspect can be improved by calibration, essentially relabeling the forecast probability values.

Table 1: Attributes of forecast quality for probabilistic forecasts (adapted from Murphy 1997 and Murphy and Winkler 1992)

Attribute	Definition	Basic distribution(s)	Graphs and measures
Sharpness (refinement)	Degree to which probability forecasts approach zero and 1; "spread" of distribution of forecasts	$p(f)$	• Histogram of $p(f)$ • Variance of forecasts, σ_f^2
Resolution	Difference between $\mu_{x\|f}$ and μ_x, considered over all values of f	$p(x\|f)$, $p(x)$	• Resolution component of Brier Score • Attributes diagram
Discrimination	Degree to which forecasts discriminate between occasions when x=1 and occasions when x=0	$p(f\|x)$	• Discrimination diagram (plot of likelihood functions) • Difference in conditional means: $\mu_{f\|x=1} - \mu_{f\|x=0}$
Bias	Difference between mean forecast and mean observation	$p(f)$, $p(x)$	Mean Error (ME): $ME = \mu_f - \mu_x$
Reliability (Calibration)	Degree of correspondence between conditional relative frequencies, $p(x\|f)$ and f, considered for all values of f	$p(x\|f)$	• Reliability diagram • Attributes diagram • Reliability measure from Brier score decomposition
Accuracy	Average degree of correspondence between f and x	$p(f,x)$	• Brier score = MSE • Other scores
Skill	Accuracy of forecasts relative to accuracy of forecasts based on a standard of comparison (e.g., climatology)	$p(f,x)$	• Brier skill score, BSS • Correlation, $\rho_{f,x}$, measures potential skill • ROC Area

Resolution provides a measure of how well the observations are "sorted" among the different forecasts. We would expect that the mean observation varies between forecasts and also differs from the overall mean observation.

Sharpness indicates the degree of "spread" or variability in the forecasts. While probability forecasts vary between 0 and 1, perfect forecasts only include the two end points, 0 and 1. Therefore, sharper forecasts will tend toward values close to 0 and 1, and sharpness measures the degree to which the forecasts approach these extreme values. Forecasts with greater variability (e.g., measured by the standard

deviation) are sharper forecasts. The basic shape of the distribution of forecasts can also provide feedback concerning the degree of sharpness of a set of forecasts. If the histogram of forecast relative frequencies is either “bell-shaped” or flat, then the forecasts cannot be considered very sharp. On the other hand, a U-shaped histogram, with most or all of the frequency at 0 or 1, indicates that the forecasts are relatively sharp. Finally, the histogram corresponding to perfect forecasts has two spikes, one at 0 and one at 1. It should be noted that sharpness is just one attribute describing forecast quality; sharp forecasts are not necessarily accurate.

Discrimination is a measure of how well the forecasts discriminate between events and non-events. Ideally, the distribution of forecasts in situations when the forecast event occurs should differ from the corresponding distribution in situations when the event does not occur.

Bias refers to the overall (average) error in the forecasts. It is simply a calculation of the difference between the mean forecast and the mean observation.

Accuracy is a measure of the overall correspondence between the forecasts and observations. A number of different scores are appropriate for probability forecasts, including the Brier score, the correlation coefficient and the ROC area.

Skill evaluates the relative accuracy by comparing the accuracy of the forecasts against the accuracy of some standard of comparison, such as climatological values or persistence.

Based on the above, it is evident that a variety of different attributes may be of interest when evaluating the quality of a set of forecasts. No single attribute or measure alone can provide a complete picture of the characteristics of the forecasts. It is therefore suggested that a range of the aforementioned attributes are calculated in order to deliver an overall view of the quality of the forecasts. Of course, for very specific forecast applications, some attributes may be more useful than others.

3. Statistical scores

3.1 Deterministic metrics for the ensemble mean forecast

There is a debate whether or not is a good idea to examine separately the ensemble mean as a normal deterministic output and what the benefits are. The verification of the ensemble mean can provide a general outlook of the skill of the model, filtering out smaller unpredictable scales and is needed in evaluating the spread-skill relation, something essential for an EPS system.

3.1.1 Continuous parameters

Mean error

The mean error (ME) measures the average difference between a set of forecasts and corresponding observations. Here, it measures the average difference between

the ensemble mean forecasts and observations. The ME of the ensemble mean forecast $\overline{Y}$ given the observation, x, is given by:

$$ME = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{Y}_i\right)$$

Versus computations: Mean value of all ensemble members for each parameter

Root mean square error

The mean square error (MSE) measures the average square error of the forecasts. The Root Mean Square Error (RMSE) provides the square root of this value, which has the same units as the forecasts and observations. Here, the forecast corresponds to the ensemble mean value and an 'error' represents the difference between the ensemble mean $\overline{Y}$ and the observation, x. The equation for the RMSE is:

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{Y}_i\right)^2\right]^{0.5}$$

Versus computations: Mean value of all ensemble members for each parameter

Correlation coefficient

The correlation coefficient measures the strength of linear association between two variables. Here, it measures the linear relationship between n pairs of ensemble mean forecasts and corresponding observations. A correlation coefficient of 1.0 denotes a perfect linear relationship between the forecasts and observations. A correlation coefficient of -1.0 denotes a perfect inverse linear relationship (i.e. the observed values increase when the forecasts values decline and vice versa). The ensemble mean forecast may be perfectly correlated with the observations and still contain biases, because the correlation coefficient is normalized by the overall mean of each variable. A correlation coefficient of 0.0 denotes the absence of any linear association between the forecasts and observations. However, a low correlation coefficient may occur in the presence of a strong non-linear relationship, because the correlation coefficient measures linear association only. The Pearson product-moment correlation coefficient, r, which is given by:

$$r = \frac{Cov(x, \overline{Y})}{Std(x) \cdot Std(\overline{Y})}$$

where Cov(x, $\overline{Y}$) is the sample covariance between the ensemble mean forecasts and their corresponding observations. The sample standard deviations of the forecasts and observations are denoted Std($\overline{Y}$) and Std(x) , respectively. The sample covariance between the n pairs of forecasts and observations is

$$Cov(x, \overline{Y}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(\overline{Y}_i - \mu_{\overline{Y}})$$

where μ_Y and μ_x are the overall sample means of the (ensemble mean) forecasts and observations, respectively.

Versus computations: Mean value of all ensemble members, mean value of observations, standard deviation of mean ensemble forecasts, standard deviation of observations.

3.1.2 Dichotomic parameters

If only the ensemble mean is examined, EPS forecasts can be treated through the measures used for completely "confident" forecasts of dichotomic parameters, namely all the already inserted scores in VERSUS. Many of the attributes discussed in previous paragraphs, can be evaluated using these commonly used measures. For example, POD and POFD are related to discrimination, and FAR is related to reliability. Unfortunately, because only three numbers are required to specify the joint distribution of forecasts and observations in this case (i.e., the dimensionality of the completely confident dichotomous forecast verification situation is three), the measures are also strongly related, in sometimes complex ways. Improvements in one measure (e.g., POD) generally are associated with degradations in another measure (e.g., POFD, FAR). Thus, it is critical to consider a variety of measures when evaluating these types of forecasts, despite their apparent simplicity. One particularly important dependency is the strong relationship of FAR, CSI, and other measures to the climatological probability, $p(x=1)$ (Brown and Young 2000; Mason, 1989). This relationship makes it inappropriate to compare forecasts for situations with different climatological probabilities, and also limits use of these measures for certain types of observations (Brown and Young 2000).

3.2 Probabilistic Scores

Brier Score

Answers the question: What is the magnitude of the probability forecast errors?

The Brier Score (BS) measures the average square error of a probability forecast. It is analogous to the mean square error of a deterministic forecast, but the forecasts, and hence error units, are given in probabilities. The Brier Score measures the error with which a discrete event, such as 'flooding', is predicted. It is given from:

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

- N = number of points in the "domain" (spatio-temporal)
- o_i = 1 if the event occurs
  = 0 if the event does not occur
- f_i = is the probability of occurrence according to the forecast system (e.g. the fraction of ensemble members forecasting the event)

It is sensitive to climatological frequency of the event. In the absence of any forecasting skill, the best strategy to optimise the Brier Score is to forecast the climatological frequency. The more rare an event, the easier it is to get a good *BS* without having any real skill. For this reason, the Brier Skill Score (see below) is

preferred because it references the score to climatology (sample or long-term). The perfect score is 0 and is possible for perfect deterministic forecast.

Versus computations: f_i *fraction of ensemble members forecasting an event*

Brier Score Decomposition

The Brier Score can be decomposed into several components that are relevant for interpretation of the sources of errors in the forecasts (Murphy 1973) which is useful for exploring dependence of probability forecasts on ensemble characteristics:

$$BS = \frac{1}{N}\sum_{k=0}^{M} N_k (f_k - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=0}^{M} N_k (\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o})$$

reliability *resolution* *uncertainty*

ō total frequency of the event (sample climatology)

For this decomposition, it is assumed that there is a discrete number of forecast possibilities, *M*, and the forecasts and observations have been sorted by the forecast value. Each of the terms in can be interpreted in the context of attributes of forecast quality as it was discussed above.

The first term is a **reliability** measure: for forecasts that are perfectly reliable, the sub-sample relative frequency is exactly equal to the forecast probability in each sub-sample. It measures the difference between the forecast and the mean observation associated with that forecast value, over all of the forecasts.

The second term is a **resolution** measure: if the forecasts sort the observations into sub-samples having substantially different relative frequencies than the overall sample climatology, the resolution term will be large. This is a desirable situation, since the resolution term is subtracted. It is large if there is resolution enough to produce very high and very low probability forecasts.

The **uncertainty term** ranges from 0 to 0.25. If the event was either so common, or so rare, that it either always occur or never occur, then b_{unc}=0. When the climatological probability is near 0.5, there is more uncertainty inherent in the forecasting situation (b_{unc}=0.25).

Versus computations: f_i fraction of ensemble members forecasting an event
ō total frequency of the event (sample climatology)
sample climatology= obs. occurrences/ num. forecasts

Brier Skill Score

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?

The Brier Skill Score (BSS) measures the performance of one forecasting system relative to another in terms of the Brier Score (BS). The BS measures the average square error of a probability forecast of a dichotomous event. The BSS comprises a ratio of the BS for the forecasting system to be evaluated (the "main forecasting system"), over the BS for the reference forecasting system BS_{REF}. Commonly, the reference forecast is the sample climatology.

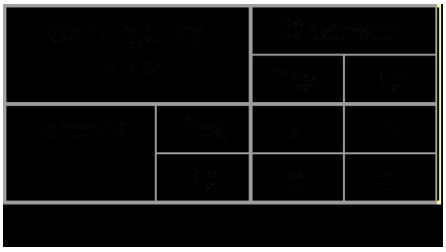
$$BSS = 1 - \frac{BS}{BS_{cli}} \qquad BS_{cli} = \bar{o}(1 - \bar{o})$$

As a measure of average square error in probability, values for the BS approaching zero are preferred. It follows that a BSS closer to 1 is preferred, as this indicates a low BS of the main forecasting system relative to the BS of the reference forecasting system. This score should always be applied to a sufficiently large sample, one for which the sample climatology of the event is representative of the long term climatology. The rarer the event, the larger the number of samples needed to stabilise the score. For best results the Brier skill score should be computed on the whole sample, i.e., the skill should be computed for an aggregated sample, not averaged for several samples.

Relative Operating Characteristic

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

The Relative Operating Characteristic (ROC; also known as the Receiver Operating Characteristic) measures the quality of a binary prediction or “decision” based on the forecast probability. A binary prediction is generated from the forecast by defining a probability threshold above which the discrete event is considered to occur. For example, a decision maker might issue a flood warning when the forecast probability of a flood exceeds 0.9. The ROC curve plots the forecast quality for several probability thresholds. For example, given a decision on whether to issue a flood warning, a probability threshold of 0.7 corresponds to a higher level of risk aversion (i.e. a lower threshold for warning) than a probability of 0.9. As the threshold declines, the probability of correctly detecting an event (the Probability of Detection or POD) will increase, but the probability of False Detection will also increase.
- X-axis: False Alarm Rate or probability of an incorrect forecast of an event
- Y-axis: the POD or probability with which an event is correctly forecast to occur.



$$POD = \frac{a}{a+c} = \frac{\text{number of correct forecasts of the event}}{\text{total number of occurrences of the event}}$$

$$F = \frac{b}{b+d} = \frac{\text{number of non correct forecasts of the event}}{\text{total number of non-occurrences of the event}}$$

A contingency table can be built for each probability class (a probability class can be defined as the % of ensemble elements which actually forecast a given event)

Versus computations: Contingency table for each probability class

Ranked Probability Score

Answers the question: How well did the probability forecast predict the category that the observation fell into?

The most common measure used to evaluate probability forecasts of multiple categories is the *Ranked Probability Score* (RPS). This measure is analogous to the BS, and has the form:

$$RPS = \frac{1}{J-1}\sum_{m=1}^{J}\left[\left(\sum_{j=1}^{m} f_j\right)-\left(\sum_{j=1}^{m} o_j\right)\right]^2$$

- *J* = number of forecast categories
- *oj* = 1 if the event occurs in category j
  = 0 if the event does not occur in category j
- *fj* = is the probability of occurrence in category j

This score is used to assess multi-category forecasts, where *J* is the number of forecast categories (for example, rainfall bins: 0-1 mm, 1-5 mm, 5-10 mm, etc.),The *RPS* penalizes forecasts less severely when their probabilities are close to the true outcome, and more severely when their probabilities are further from the actual outcome. For two forecast categories the *RPS* is the same as the Brier Score.

Ranked Probability Skill Score

Answers the question: What is the relative improvement of the probability forecast over climatology in predicting the category that the observations fell into?

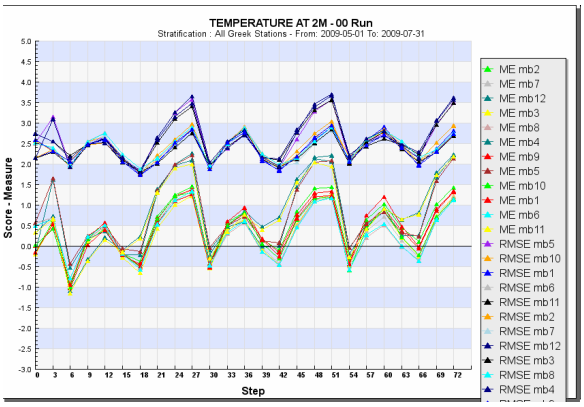
The RPSS measures the improvement of the multi-category probabilistic forecast relative to a reference forecast (usually the long-term or sample climatology). It is similar to the 2-category Brier skill score, in that it takes climatological frequency into account. Because the denominator approaches 0 for a perfect forecast, this score can be unstable when applied to small data sets. This score should always be applied to a sufficiently large sample, one for which the sample climatology of the event is representative of the long term climatology. The rarer the event, the larger the number of samples needed to stabilise the score. For best results the ranked probability skill score should be computed on the whole sample, i.e., the skill should be computed for an aggregated sample, not averaged for several samples.

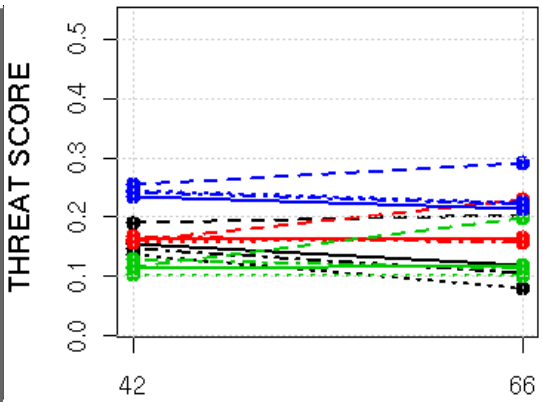
$$RPSS = \frac{\overline{RPS} - \overline{RPS}_{reference}}{0 - \overline{RPS}_{reference}} = 1 - \frac{\overline{RPS}}{\overline{RPS}_{reference}}$$

4. Graphical representation of scores and probabilities

4.1 Deterministic scores graphs

A composite graph with the values of RMSE values for each member for each parameter is a first necessary approach to present the statistical value of an ensemble system. Such capability has already been added to the VERSUS system. It is however essential to add as a separate line the calculated score for the ensemble mean. In this way the error spread around the mean forecast is more visible.





4.2 Reliability diagrams

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

The reliability diagram measures the accuracy with which a discrete event is forecast by an ensemble or probabilistic forecasting system. According to the reliability diagram, an event should be observed to occur with the same relative frequency as its forecast probability of occurrence over a large number of such forecast-observation pairs. The Reliability diagram plots the average forecast probability within each bin on the x-axis. The y-axis shows the corresponding fraction of observations that fall in each bin. If the forecast is perfectly reliable, the observed fraction within each bin will equal the average of the associated forecast probabilities, forming a diagonal line on the reliability diagram. Deviation from the diagonal line represents bias in the forecast probabilities, notwithstanding sampling uncertainty. The reliability diagram may be computed for several discrete events.

Each event is represented by a separate reliability curve. The number of forecasts that fall in each bin is referred to as the 'sharpness' of the forecasts and is displayed as a histogram for each of the forecast bins. Ideally, the forecast probabilities will be sharp, i.e. issued with little uncertainty, but also reliable. The reliability diagram is conditioned on the forecasts,p(x|f) and it is a good partner to the ROC

To construct a reliability diagram, do the following:
1. For each forecast probability category count the number of observed occurrences
2. Compute the observed relative frequency in each category k

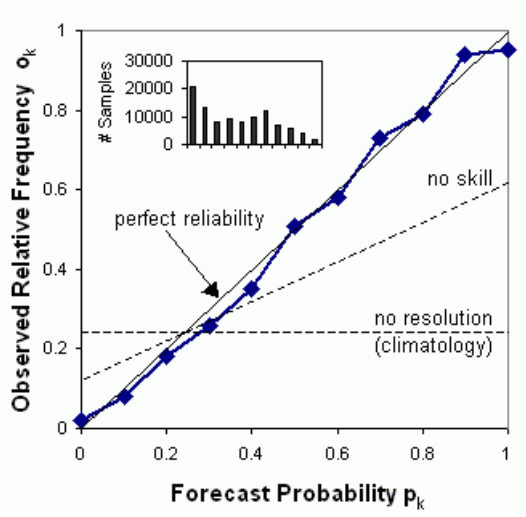
obs. relative frequency$_k$= obs. occurrences$_k$/ num.forecasts$_k$

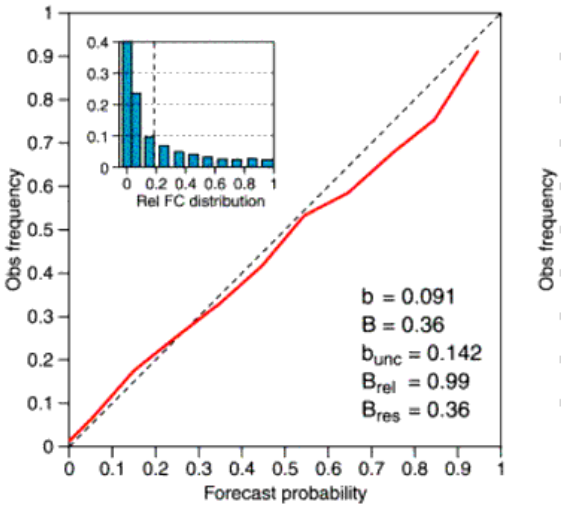
3. Plot observed relative frequency vs forecast probability
4. Plot sample climatology ("no resolution" line)

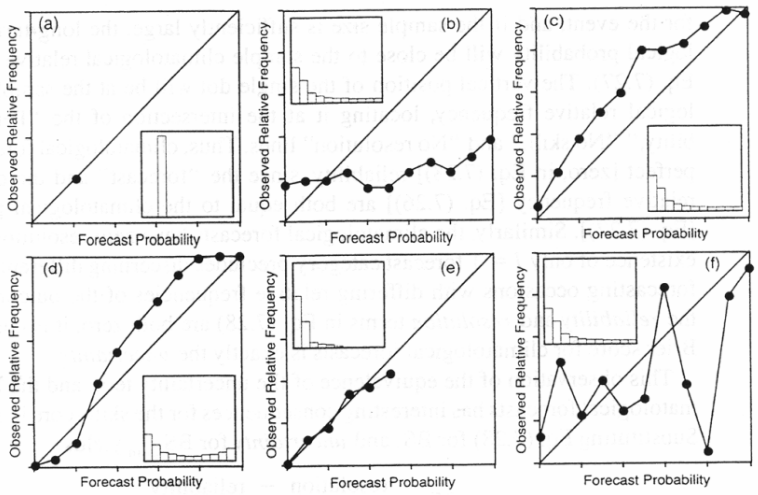
   sample climatology= obs. occurrences/ num. forecasts

5. Plot "no-skill" line halfway between climatology and perfect reliability (diagonal) lines
6. Plot forecast frequency separately to show forecast sharpness

In a perfect reliable system the forecast probability is equal to the observed frequency so the graph is a straight line oriented at 45^o to the axes. If the curve lies below the 45° line, the probabilities are overestimated whether if the curve lies above the 45° line, the probabilities are underestimated. The more flat is the curve the lower resolution the probabilities have. The frequency of the forecasts in each probability bin of the histogram shows the sharpness of the forecast. Below are given examples of graphs created to represent the reliability diagram and ways to interpret a graph that was created.







a.climatoligical forecast, **b.**minimal resolution, **c.**underforecasting bias, **d.**good resolution at the expense of reliability, **e.**reliable of rare event, **f.**small sample size and small ensemble

The reliability term measures the mean square distance of the graph line to the diagonal line (see top left diagram). Points between the "no skill" line and the diagonal contribute positively to the Brier skill score (resolution > reliability). The resolution term measures the mean square distance of the graph line to the sample climate horizontal dotted line.

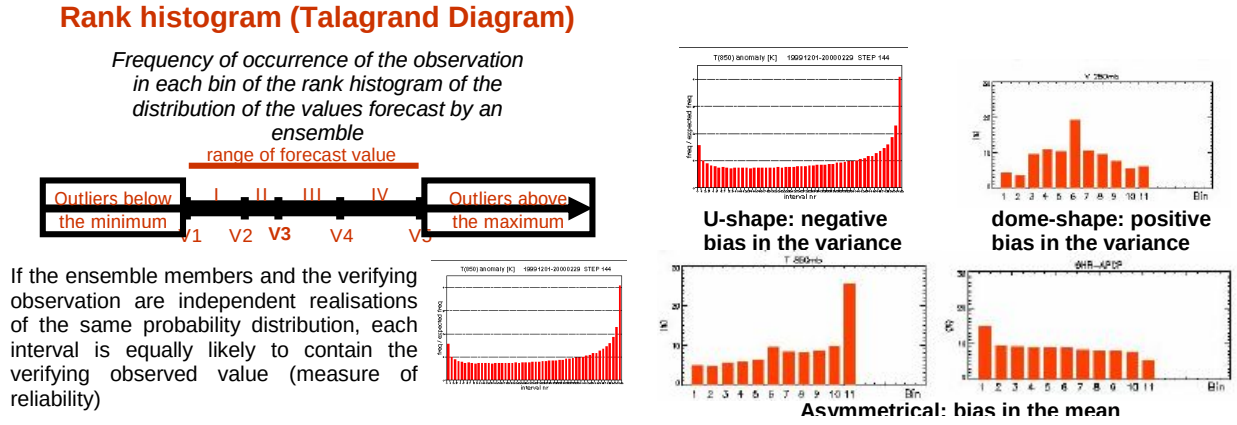
4.3 Rank Histogram

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

Also known as a "Talagrand diagram", this method checks where the verifying observation usually falls with respect to the ensemble forecast data, which is arranged in increasing order at each grid point. In an ensemble with perfect spread, each member represents an equally likely scenario, so the observation is equally likely to fall between any two members. It measures how well the ensemble spread of the forecast represents the true variability (uncertainty) of the observations.

To construct a rank histogram, do the following:
1. At every observation point rank the N ensemble members from lowest to highest. This represents N+1 possible bins that the observation could fit into, including the two extremes
2. Identify which bin the observation falls into at each point
3. Tally over many observations to create a histogram of rank.

Interpretation:
Flat: ensemble spread correctly represents forecast uncertainty
U-shaped: ensemble spread too small, many observations falling outside the extremes of the ensemble
Dome-shaped: ensemble spread too large, most observations falling near the center of the ensemble
Asymmetric: ensemble contains bias



Explanation of Talagrand diagram construction and interpretation of its shapes

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

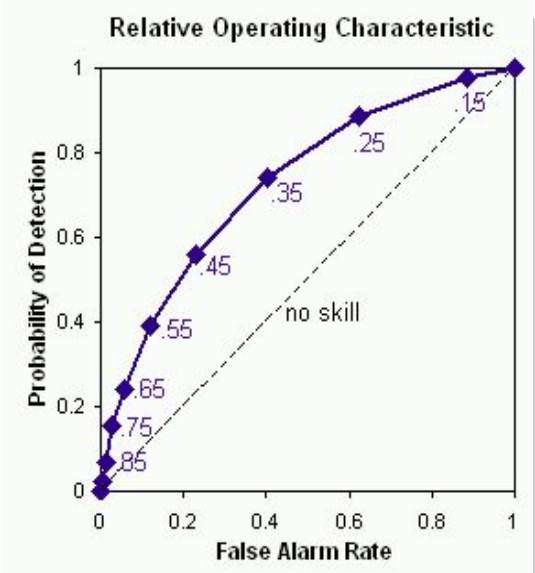
4.4 ROC curves

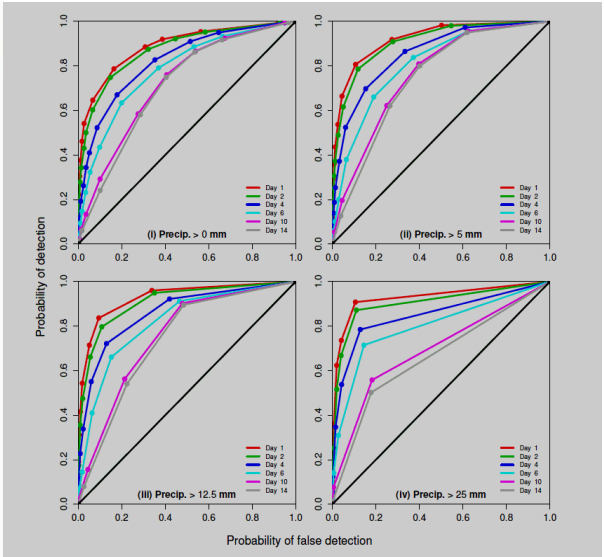
ROC measures the ability of the forecast to discriminate between two alternative outcomes, thus measuring resolution. It is not sensitive to bias in the forecast, so says nothing about reliability. A biased forecast may still have good resolution and produce a good ROC curve, which means that it may 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 corresponding forecast?) It is therefore a good companion to the reliability diagram, which is conditioned on the forecasts.

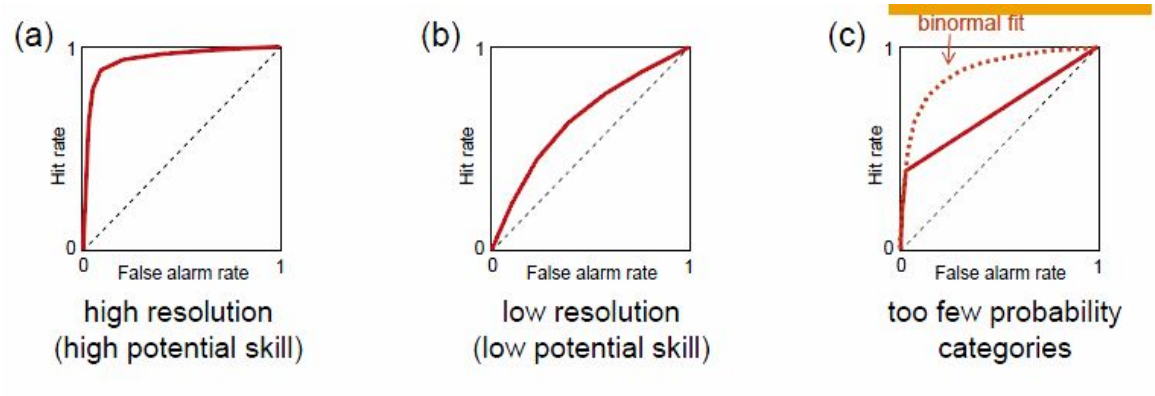
To construct a ROC curve, do the following:

1. From original dataset, determine bins
   - You can use binned data as for Reliability diagram but there must be enough occurrences of the event to determine the conditional distribution given occurrences – may be difficult for rare events.
   - Generally you need at least 5 bins.
2. For each probability threshold, determine HR and FA
3. Plot HR vs FAR to give empirical ROC curve.
4. Use binormal model to obtain ROC area; recommended whenever there is sufficient data >100 cases or so.





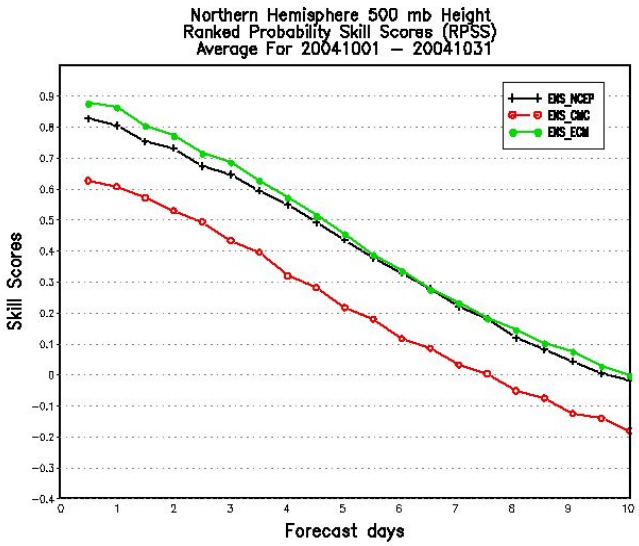
The area under the ROC curve ("ROC area") is a useful summary measure of forecast skill. Perfect: ROC area = 1, No skill: ROC area = 0.5, ROC skill score ROCS= 2 (ROC area -0.5).



Interpretation of ROC curves

4.5 Ranked Probability Skill Score Graph

Measures the improvement of the multi-category probabilistic forecast relative to a reference forecast (usually the long-term or sample climatology). It is rather unstable when applied to small data sets. An example graph is given below.



5. Spread-Skill Relationship

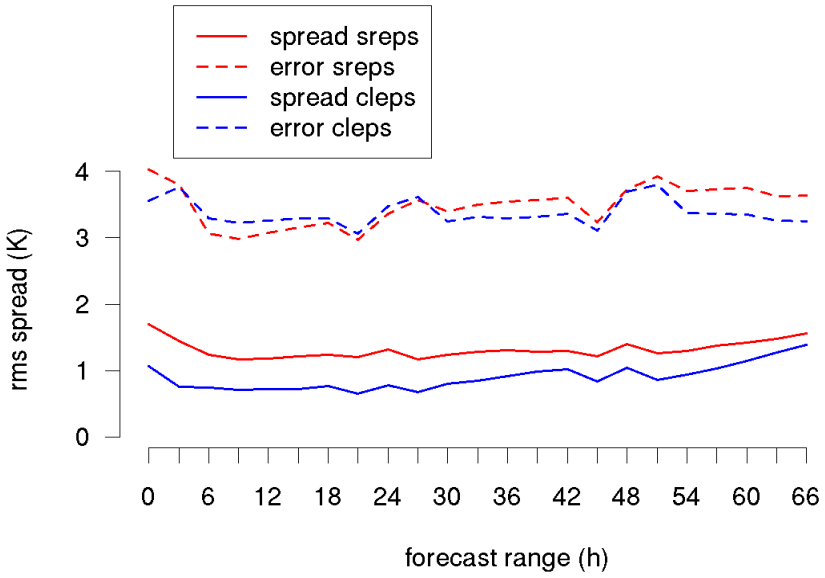
EPS are designed to represent the full uncertainty range as realistic as possible. To assess if this is the case, the spread-skill relationship (SSR) is often investigated. Palmer et al. [2005] showed that in a 'perfect ensemble' the mean of the spread should be equal to the root mean square error over the same period. The unbiased estimator of the standard deviation is given by:

$$spread = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(f_m - \overline{f})^2}$$

where M is the ensemble size, f_m is the forecast value of the m^{th} member and f represents

$$\overline{f} = \frac{1}{M}\sum_{m=1}^{M} f_m$$

When the spread line is below and far from the RMSE line then the ensemble is considered underdispersive. Example graph of this spread/skill correlation is given below.



6. Cost - Loss model

A cost-loss analysis is a useful tool for all decision makers, especially while forecasting precipitation events and identifying the skill of the forecast system. The score is based on the fact that a certain event can cause an economic loss L, which can be avoided if the event is forecasted, by taking a protective action whose cost is C. depending on the ratio between C and L the forecast system can either be useful or not for a user. The scores FAR and HR can be used as indicators for the skill of the forecast system.

	Event occurs	Event does not occur
Action taken	*C*	*C*
Action not taken	*L*	*0*

With a deterministic forecast system, the mean expense for unit loss is:

$$ME = FAR\frac{C}{L}(1-\bar{o}) - HR\bar{o}\left(1-\frac{C}{L}\right) + \bar{o},$$

ō=a+c is the sample climatology (the observed frequency), while a,c are taken from the contingency table prepared for a specific threshold of the event.

Contingency table		Observed	
		Yes	No
Forecast	Yes	a	b
	No	c	d

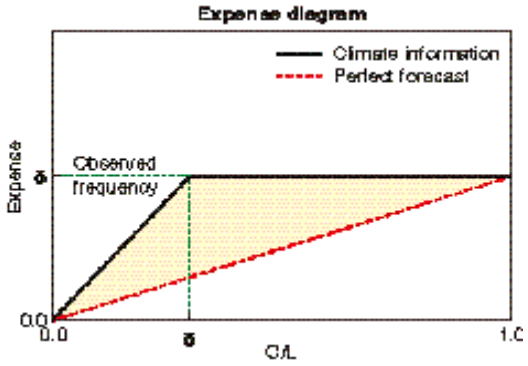
In the case of an EPS or probabilistic forecast system, the user has to fix a probability threshold k and when this threshold is exceeded a protective action needs to be taken. In this case, FAR and HR are calculated for the specific probability threshold. The value V of the forecast system is defined as a reduction in ME with respect to the ME sustained if only climatological information were available and it is expressed as a percentage of the value which would be achieved by a perfect forecast:

$$V = \frac{ME(climate) - ME(forecast)}{ME(climate) - ME(perfect)}.$$

The ME for a perfect forecast system is the expense taking a preventing action every time the event occurs. Both ME(climate) and ME(perfect formulas are given below together with a diagram.

$$ME(perfect) = \bar{o}\frac{C}{L},$$

$$ME(climate) = \min(\bar{o}, \frac{C}{L})$$



The action is taken always when C/L< ō and never in any other case.

7. Remarks

A number of measures can be added for a more spherical approach of an ensemble verification. Such methods can be those based on object-oriented verification, similar to the ones used for deterministic forecasts and described by Ebert and McBride (2000). Also a useful tool could be the ability to compare two or more ensemble systems and the statistically significant difference in their performance, using for example a bootstrap method (Wilks 1995, Hamill 1999). This can be the interest of a second phase of implementation of ensemble verification techniques inside VERSUS system.

References:

Brown, B. G. (2001). Verification of precipitation forecasts: A survey of methodology part ii: Verification of probability forecasts at points. In Proceedings of the WWRP/WMO Workshop on the Verification of Quantitative Precipitation Forecasts, Prague, 14-16 May 2001., NCAR, Boulder CO, USA.

Brown, B.G., and G.S. Young, 2000: Verification of icing and turbulence forecasts: Why some verification statistics can't be computed using PIREPs. *Preprints, 9th Conference on Aviation, Range, and Aerospace Meteorology*, 11-15 September, Orlando, FL, American
Meteorological Society (Boston), 393-398.

Ebert, E.E. and McBride, J.L., 2000. Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrology*, **239**, 179-202.

Hamill, T.M., 1999. Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155-167.

Jolliffe, I.T., Stephenson, D.B. (Eds.), 2003. Forecast Verification: A Practitioner's Guide in Atmospheric Science. John Wiley and Sons, Chichester, 240 pp.

Mason, I., 1989: Dependence of the critical success index on sample climate and threshold probability. *Australian Meteorological Magazine*, **37**, 75-81.

Murphy, A.H., 1973: A new vector partition of the probability score. *Journal of Applied Meteorology*,**12**, 595-600.

Murphy, A. H. and Winkler, R.L., 1987: A general framework for forecast verification. Monthly Weather Review, 115, 1330-1338.

Palmer, T., R. Buizza, R. Hagedorn, A. Lawrence, M. Leutbecher, and L. Smith, Ensemble prediction: A pedagogical perspective, ECMWF Newsletter, (106), 2005.

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Royal Meteorol. Soc.*, **126**, 649-667.

Stanski, H.R., Wilson L.J. and Burrows W.R., 1989. Survey of Common Verification Methods in Meteorology (WMO Research Report No. 89-5)

Talagrand., O., 1999 A posteriori verification of analysis and assimilation algorithms. In Proceedings of the ECMWF Workshop on Diagnosis of Data Assimilation Systems, 2-4 November pages 17--28, Reading.

Wilks D. S., 1995. Statistical methods in atmospheric sciences. Academic Press, New York, 467 pp

Wilson, L.J., Burrows, W.R. and Lanzinger, A., 1999: A strategy for verification of weather element forecasts from an ensemble prediction system. Monthly Weather Review, 127, 956-970.