VERIFICATION OF WEATHER ELEMENT FORECASTS FROM AN ENSEMBLE PREDICTION SYSTEM

L. J. Wilson
Meteorological Research Branch
Dorval, Quebec, Canada

Abstract: Weather element forecasts obtained from an EPS are essentially "posteriori" estimates of the weather element probability distribution, given the model forecasts. These can be compared to various "prior" or unskilled estimates taken from climatology (the climatological distribution for the valid time) or persistence (the climatological distribution given yesterday's observation). The verifying observation can be regarded as a single outcome of a stochastic process. Existing verification methods for both deterministic (e.g. temperature) and categorical forecasts (e.g. POP) will be examined in the context of these concepts. Also, these concepts point to some additional verification measures which are described.

1. INTRODUCTION

Until the advent of ensemble prediction systems (EPS), verification of NWP model results usually consisted either of comparisons of point output from the model on a grid with analyses, or interpolated model output with point observations. Many different measures have been defined and used, such as mean absolute and root mean square error, anomaly correlations, S1 scores to name a few. These scores are computationally different and express different aspects of the model performance, but in all cases the elements of the verification sample are produced by space and time matching of single point deterministic forecasts from the model with the corresponding observation. The "forecast error" is determined by simple difference between the forecast value and the observed value of the weather element. The forecast value is presented as a "best estimate" and any associated uncertainty is not estimated. In terms of the verification, uncertainty is simply an unspecified component of the error.

The advent of EPS's recognizes the uncertainty in the forecast due to the uncertainty in the initial conditions of the model. The ensemble represents an attempt to quantify the uncertainty, indeed to estimate the full distribution of possible outcomes given what is hoped is a realistic range of possible initial conditions. Verification methods heretofore applied to deterministic forecasts from models are inadequate to the task of verification of the output of an EPS, and new measures must be sought. Since some of the output products from the ensemble are presented as probabilities estimated from the ensemble distribution, many of the traditional verification measures used for probabiliy forecasts from statistical interpretation systems are directly applicable. However, these verification measures are inadequate for verification of the ensemble output with specific observations, such as temperatures, pressure, height or wind.

In this paper I propose some verification methods for ensemble forecasts which are consistent with the stochastic nature of the output of an EPS and which follow standard theory and practice of statistics. It is convenient to consider the subject in terms of "deterministic" and "categorical" or probabilistic forecasts.

Deterministic forecasts are specific forecasts of the value of a surface weather element at a particular location. They carry the units of the weather element. Examples are temperature, pressure and wind. Categorical forecasts are expressed as probabilities of occurrence of one or more mutually exclusive and exhaustive categories of a weather element. Such forecasts must also state the category boundary. Examples are "probability of temperature anomaly greater than +4 degrees", probability of precipitation, etc. I will restrict the discussion to the verification of surface weather elements, but it can be easily extended to upper air weather elements.

2. VERIFICATION OF DETERMINISTIC FORECASTS FROM THE ENSEMBLE.

The output of an EPS is a collection of deterministic forecasts which comprise an estimate of the distribution of each weather element at each point in the model domain. This is in fact a conditional distribution: The forecast is the probability distribution of the weather element given the model run and the ensemble of initial conditions. In statistical (Bayesian) terms, this can be considered an "a posteriori" distribution. That is, it is the estimated (altered) distribution following the addition of new knowledge about the predictand. The corresponding "prior" distribution represents knowledge about the weather element before the model is run. This can be either the climatological distribution for the day of the year, or, it could be a persistence distribution. The latter is the conditional distribution of today's values of the weather element given yesterday's value.

In this context, the verifying observation is a single realization, an individual extracted from the climatological distribution. For verification purposes, one is interested in the probability that the observed value of the weather element will occur, given the model's estimate of the posterior distribution. Logically, if the model has skill, the probability of obtaining a particular observation given the ensemble should be larger than the climatological or persistence probability. It is this conditional probability that can be used to verify the ensemble prediction, $P(X_{obs}| X_{eps})$, for weather element "X".

Figure 1 shows schematically the verification model described above, using temperature as an example. The desired probability is $P(T_{obs}| T_{eps})$. To compute this requires an estimate of the parameters of the EPS distribution, and assumptions about the distribution shape. For a first approximation, it can be assumed the EPS distribution will be of the same family as the underlying climatological distribution for the weather element in question. One would expect the EPS to predict the same distribution as in nature, and distribution-fitting, with appropriate goodness of fit tests would itself provide useful verification information about the EPS.

Although the ensemble is an attempt to estimate the full posterior distribution, it is not possible to do so with the small ensembles that are produced operationally. To delineate the shape of the distribution would require many hundreds or thousands of ensemble members. The parameters of the distribution can be estimated with more stability and confidence on a small sample. Such a parametric method also permits prior knowledge about the expected distribution shape to be incorporated into the assessment of the forecast.

For temperature, the underlying distribution is normal, as shown on Figure 1, and so estimation of the parameters of this distribution simply means computing the mean and standard deviation from the ensemble. Once the parameters have been estimated, the probability can be computed directly from the formula for the distribution. Since the distribution is continuous, one must define a reasonable finite range of values around the observed temperature which are considered "correct". Strictly speaking, +/- 0.5 degree could be chosen since it is the resolution limit for temperature, but larger ranges could also be used.

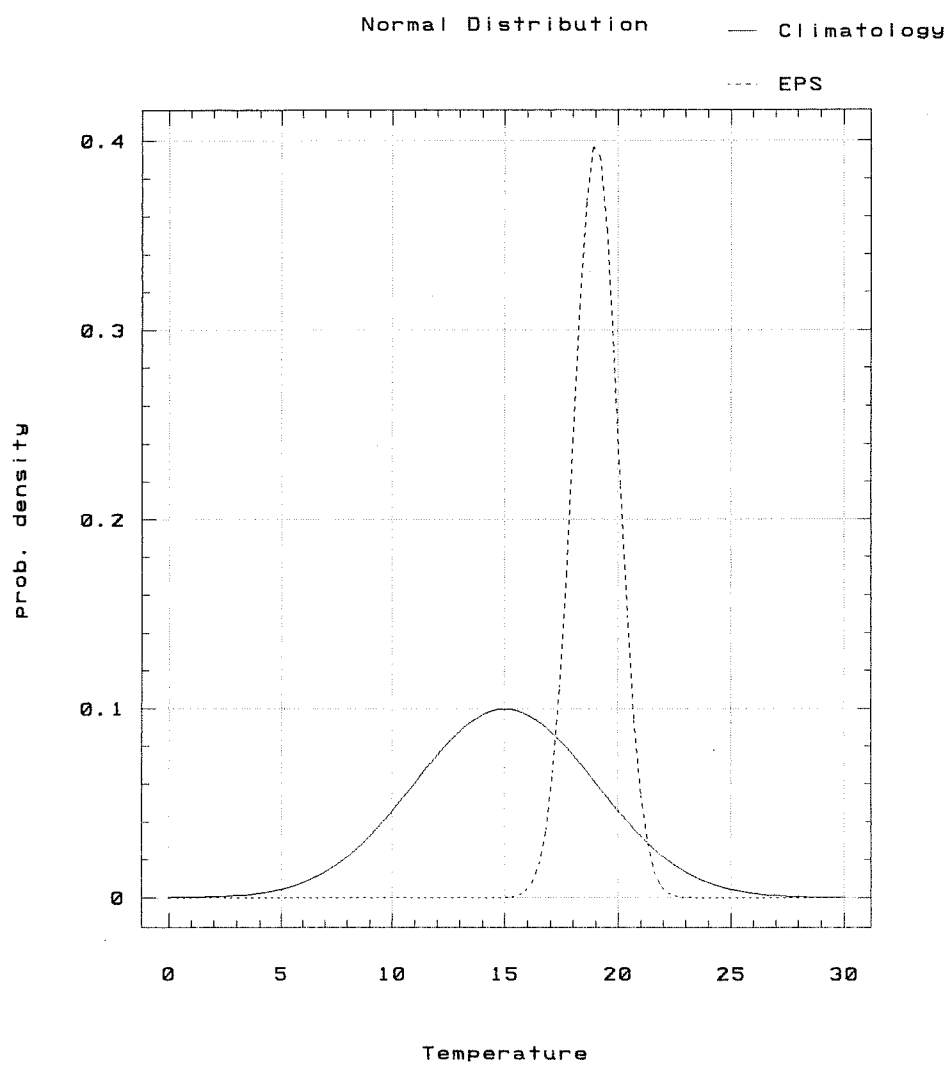

Figure 1. Comparison of hypothetical climatological and EPS distributions for a normally distributed weather element such as temperature. In this case, the EPS is predicting a temperature above normal, and has reduced the variance with respect to climatology.

This single probability $P(T_{obs}| T_{eps})$ has many of the characteristics of a good verification measure: First, it has a range of 0 to 1 and is positively oriented. A perfect score is obtained if all the ensemble members predict the verifying temperature. Second, this verification scheme encourages "precise" (sharp) forecasting by the ensemble, with a narrow distribution, *provided the ensemble is located in the right place*. Hedged forecasts (broader distribution) and sharp forecasts that "miss" the event are both penalized in this system. The ideal forecast is both accurately positioned (reliable) and sharp.

The above idea can be simply extended to skill measures by comparing with the corresponding a priori probability, that is, the probability $P(T_{obs}|T_{std})$. A skill score can be defined in the usual form,

$$\text{Skill} = [P(T_{obs}|T_{eps}) - P(T_{obs}|T_{std})]/[1.0 - P(T_{obs}|T_{std})]$$

where the numerator gives the difference (improvement) in the forecast probability compared to the standard forecast (T_{std}, usually climatology or persistence) and the denominator is the normalizing difference between the perfect score and the standard probability. The skill is thus a percentage improvement over the standard (unskilled) forecast. This kind of score has the advantage that the "difficulty" of forecasting is taken into account: For a given forecast probability, a "difficult" climatology will have a broader distribution, with lower climatological probabilities and correspondingly higher skill. For the special case where the observed temperature is near normal (near the climatological mean) the EPS shows skill by predicting a narrower distribution than the climatological distribution. Again, higher skill is obtained in more difficult situations for a given EPS probability.

The above discussion applies to cases where the EPS distribution is unimodal. If there is sufficient evidence that it is multimodal, as indicated for example by the cluster analysis, then the EPS distribution can be treated as separate distributions of the same form, and the parameters estimated separately for each distinct cluster. Figure 2 illustrates this situation, for two clusters. The accuracy score then becomes the sum of the products of the prior probability of occurrence of each cluster and the likelihood of obtaining the observation given the occurrence of the cluster:

$$P(T_{obs}|T_{eps}) = P(T_{obs}|T_{eps1})P(T_{eps1}) + P(T_{obs}|T_{eps2})P(T_{eps2}) + \ldots + P(T_{obs}|T_{epsn})P(T_{epsn}),$$

for n clusters. Since the estimates of the distribution parameters become unreliable for small samples, this refinement will be of benefit only when the distribution is clearly multimodal, for example when there is a bifurcation of the ensemble "plume". The corresponding skill score is,

$$\text{Skill} = [P(T_{obs}|T_{eps1})P(T_{eps1}) + \ldots + P(T_{obs}|T_{epsn})P(T_{epsn}) - P(T_{obs}|T_{std})] / [1.0 - P(T_{obs}|T_{std})]$$

For temperature and other relatively smooth and continuous elements such as msl pressure, geopotential height and upper air temperatures, the underlying distribution is usually nearly normal. For many other surface weather elements, however, it is not so. Precipitation amount during a given period (QPF) requires a highly skewed distribution, which can model the high probability density near 0. One distribution which has this characteristic is the two-parameter Gamma distribution, which, depending on the values of the parameters, may have an exponential shape with high density near 0, or a slightly skewed shape for cases with significant probability density away from the lower boundary of 0. The less skewed forms might be appropriate in cases where the ensemble suggests precipitation is likely, where the ensemble-predicted mean amount is considerably higher than the climatological mean amount. Figure 3 shows two examples of the Gamma distribution with different parameters, as might be suggested by an ensemble forecasts for no precipitation and some precipitation. Mielke (1973) has suggested that a two-parameter family of distributions called the Kappa distribution might fit daily precipitation data better than the Gamma distribution, but the Gamma distribution is more widely used and programs are available to fit its parameters.

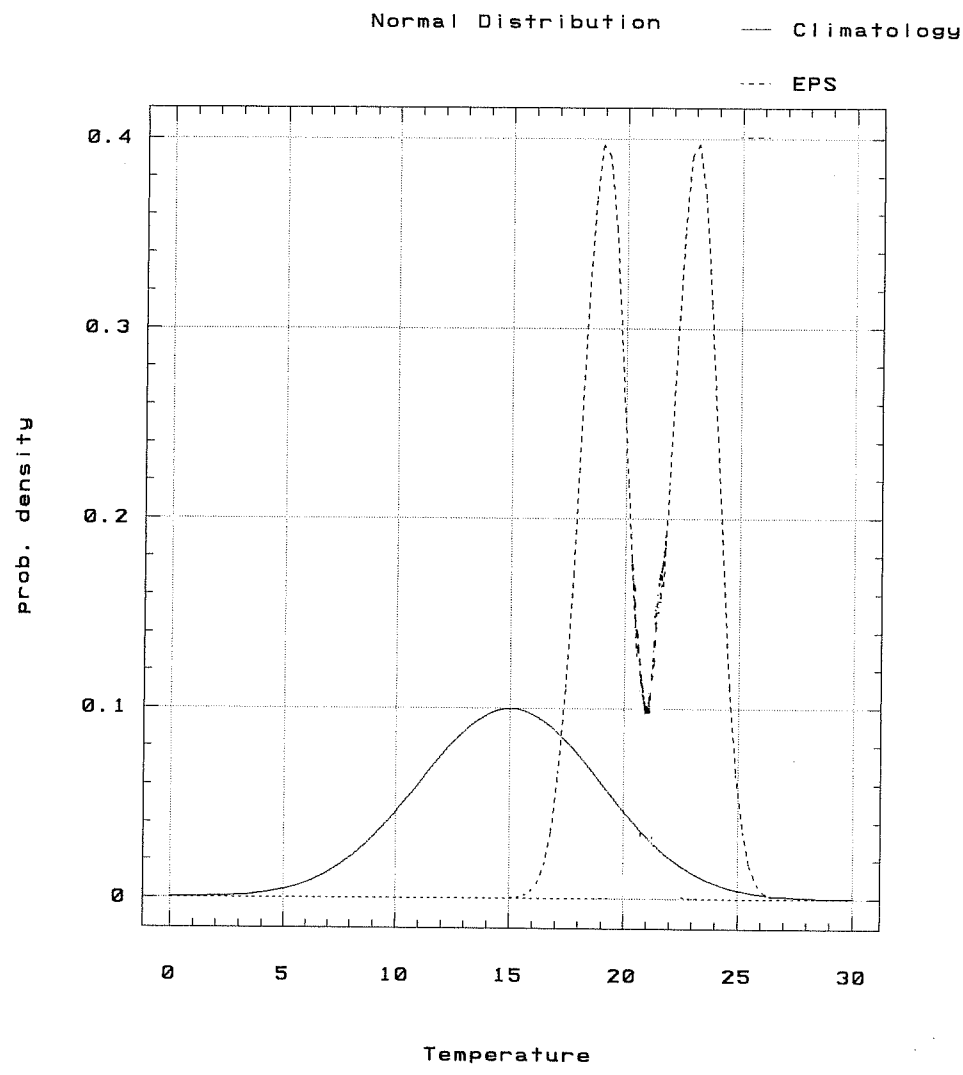

Figure 2. Comparison of hypothetical climatological and EPS distributions for a normally distributed weather element where the EPS has predicted a bimodal distribution.

For windspeed, the Weibull distribution has been found to fit well (Somerville and Bean, 1978). This distribution takes a highly skewed shape also with high density near 0 for some values of its parameters, but also takes a near-normal shape if the probability density is concentrated away from 0. Like the Gamma distribution, it is designed for positive-definite variables. It is negatively-skewed (higher density below the mode than above), which makes it useful for windspeed where underprediction is more likely than overprediction.

For variables defined on the range 0 to 1 such as percent of cloud cover, the Beta distribution can be used (Somerville and Bean, 1978). It is also a two-parameter distribution, and includes not only the U-shape typical of cloud amount, but also the uniform distribution and near-normal forms. It can be positively skewed or negatively skewed or symmetric, depending on the parameters.

Sometimes a transformation of the variable makes it easier to fit a distribution. Examples are the "cube-root normal" distribution for QPF, where the cube root of the precipitation amount is found to have a normal distribution, and the "Lognormal" distribution, by which the logarithm of the visibility is approximately normal.

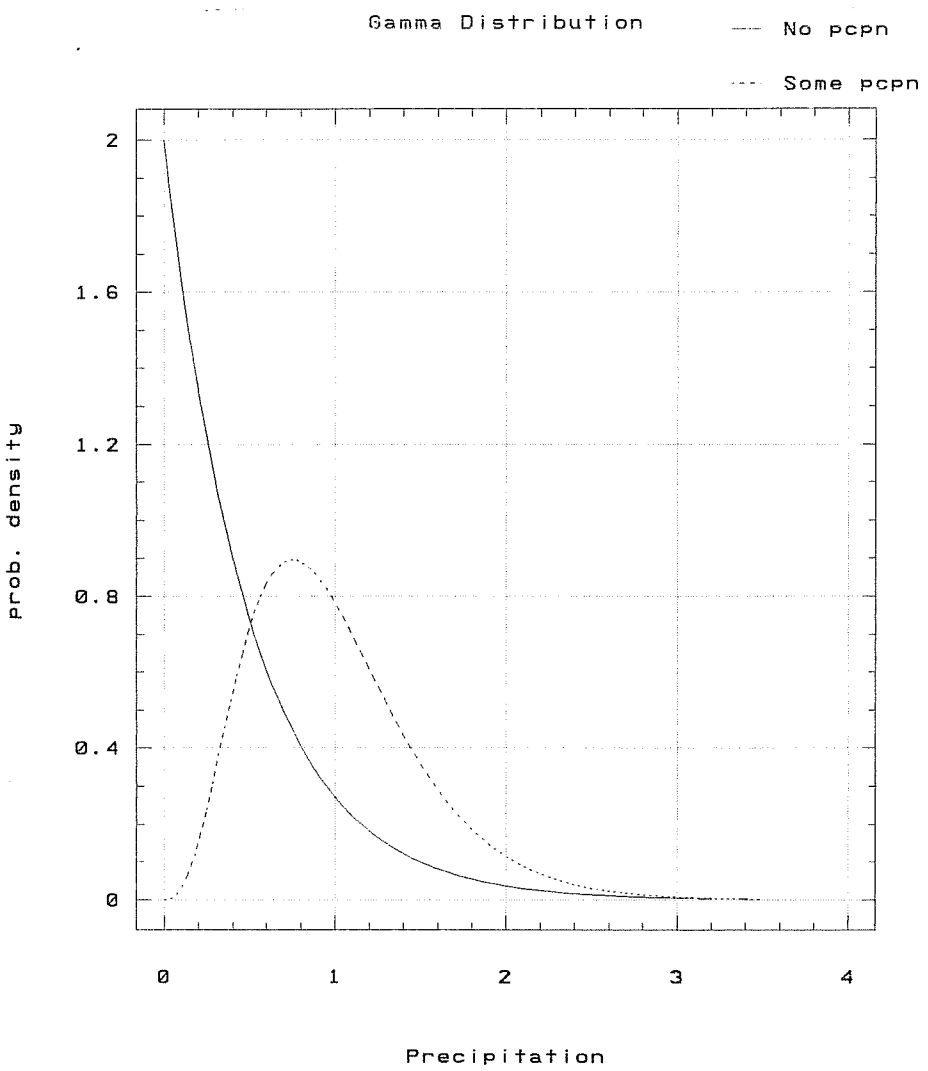

Figure 3. An example of the two-parameter Gamma distribution, which might be fitted to QPF forecasts. The two curves are possible shapes of predicted distributions for cases where precipitation is unlikely and where some precipitation is likely.

Table 1 summarizes the distributions which have been used to fit meteorological data. Once an appropriate distribution is chosen, its parameters can be estimated from the ensemble as described above. Statistical programs are readily available for distribution fitting. One of these is STATGRAPHICS, a PC program that will estimate the parameters of 14 different distributions given a dataset.

The verification method described in this section is a "parametric" method, where the 32 members of the ensemble are used to estimate the parameters of an assumed distribution. It has not been demonstrated that the assumption that the posterior distribution is the same shape as the prior distribution is valid. The ensemble is designed in part to find the extremes of the distribution with relatively few members; the perturbed analyses are by no means random in a statistical sense. What impact this might have on the output distribution is not known. One way of checking the distribution shape would be to compile a large sample of ensembles for a particular weather element, compiling them and plotting the frequency distribution. This can be compared with the frequency distribution of the corresponding observations.

Table 1. Summary of theoretical distributions that have been found to fit observations of weather elements. Algorithms to fit a dataset to most of these distributions are generally available.

Weather Element	Distribution	Characteristics
Temperature Geopotential height Upper air temperature	Normal	Two-parameter; mean and standard deviation Symmetric, bell-shaped
Precipitation (QPF)	Gamma Kappa (cube-root normal)	**Gamma:** Two-parameter, "shape and spread" Positively skewed (median higher than mode) Applies to variables bounded below; approaches normal when well away from lower bound **Kappa:** similar to Gamma in form, but not as well known. Not likely to be included in distribution-fitting programs **Cube root normal:** The cube root of precipitation amount has been found to be approximately normally distributed.
Wind speed	Weibull	Two-parameter Negatively skewed (median lower than mode) Applies to variables bounded below
Cloud amount	Beta	Two-parameter; a family of distributions including the uniform and U-shaped as special cases Intended for variables which are bounded above and below such as probability estimates, and cloud amount Negatively or positively skewed, depending on parameters
Visibility	Lognormal	Normal distribution with logarithmic x-axis; applies to positive-definite variables.

3. VERIFICATION OF THE ENSEMBLE

One measure that has been used to assess the performance of the ensemble as a whole is the Talagrand diagram. A Talagrand diagram is designed to verify the estimated distribution via the random variable "P", the probability that the verifying observation will lie in the interval between adjacent ensemble members. For ensembles of 32 members, the range of the verifying element is divided into 33 segments, with the end segments open-ended, and the expected value of P is about .03 for each of the ranges. If the EPS has truly divided the range into equal probability segments, on average, a plot of the frequency of occurrence of the verifying observation over the 33 ordered intervals for a large sample should display a uniform distribution. The Talagrand diagram can be applied to any variable, but it usually has been applied to 500 mb heights and upper air temperatures.

For a sample of ensembles, the total sample variance is composed of two components, the variance about ensemble mean, and the variance of the ensemble mean. The former can be calculated by subtracting the ensemble mean from each member of the ensemble, within each ensemble, squaring and summing over all the data. The latter is simply the variance of the ensemble means taken together. The total sample variance can be compared with the variance of the verifying observations. The two should agree.

In this context, care should be taken with the interpretation of the Talagrand diagram. Preparation of the diagram for an ensemble verification dataset usually results in a U-shaped distribution, rather than the

expected uniform distribution. The open-ended extreme categories usually show higher-than-expected-frequencies with a modest under-estimation of the frequencies in the middle of the range. Thus the verifying observation lies outside the predicted range of the distribution more often than expected. This can be due either to underprediction of the spread of the distribution, or to an error in the location of the ensemble mean (the placement of the distribution) or both. Likely, the verifying observation lies too often outside the ensemble for both reasons. Therefore, it can be misleading to conclude just from the Talagrand diagram that the ensemble spread should be increased. The ensemble attempts to estimate the frequency distribution of weather elements arising from the uncertainty in the initial conditions. Errors in the model simulation will not be accounted for, but will contribute to the total error. Results shown at this workshop indicating that improvements to the model resolution lead to a greater reduction in the tendency for verifying observations to lie outside the ensemble than is achieved by increasing the ensemble size tend to support this point.

The ensemble mean should be treated with caution in summary verification of the ensemble. First, it does not represent a specific trajectory of the model. It may not even be close, for example when the ensemble bifurcates with near equal numbers of members on both sides of the bifurcation. Second, the mean is statistically different from individual outcomes. Its estimated variance is $S^2/32$ for an ensemble of 32 members, where S^2 is the variance estimated from the ensemble. The mean is thus a very conservative estimator of a forecast value of the weather element, which will generally score well on a quadratic scoring rule. However, the ensemble mean will never predict extreme events. For all these reasons, it is not useful to devise verification measures that involve matching the ensemble mean with individual outcomes or individual ensemble members. It is, however, useful to compute summary linear scores such as the overall bias between the ensemble mean and a set of observations. This does not require matching of individual ensemble means with specific verifying observations.

4. VERIFICATION OF PROBABILITY FORECASTS FROM THE ENSEMBLE

A primary goal of ensemble prediction is the estimation of probabilities of occurrence of specific weather events based on the ensemble. Therefore, verification of probability forecasts from the ensemble is highly relevent to evaluation of the EPS. Probability forecasts are normally made by defining two or more categories of the predictand element, separated by one or more thresholds, then estimating the probability of occurrence of each category by a simple percentage of occurrence of the ensemble members in the category. In the contest of the above discussion, this is a "non-parametric" estimate of the probability of category occurrence because it makes no reference to the parameters of the distribution of the ensemble forecast.

If one has reason to believe the ensemble forecast distribution should take a specific known shape, it is possible to estimate probabilities from the ensemble using a parametric method. First, the parameters of the expected or assumed distribution are estimated and then probabilities can be obtained from the formula of the distribution. The advantages of this approach are that additional information about the distirbution can be incorporated, and the estimates may be less noisy than would be obtained directly from a 32 member

ensemble. A parametric approach could be compared and tested against the non-parametric approach first for a normally distributed variable such as surface temperature or 500 mb height.

Verification measures for probability forecasts are of two types, summary scores and graphical or tabular methods based on stratified data. The former attempt to summarize one or more attributes of the forecast quality in as few numbers as possible, usually a single value. Examples are the Brier Score and the Brier Skill Score. Such scores must be carefully interpreted and used. Graphical or tabular methods usually give more insight into the performance being measured because they do not attempt to summarize into a single value. The stratification of the data can be specified to describe specific attributes of the forecast. Examples of these methods include contingency tables, reliability tables and signal detection theory measures such as the relative operating characteristic (ROC). The following are some comments on the use and interpretation of these measures. They are discussed more fully with examples in Stanski et. al., 1990.

The Brier Score summarizes the overall accuracy of a probability forecast of a dichotomous (2-category) event. As it is usually presented, it is negatively oriented (0 is a perfect score) and has a range of 0 to 1. The Brier Score is highly sensitive to the sample climatological probability (The mean frequency of occurrence of the event in the verification sample), which means that it cannot be used to compare the accuracy of forecasts computed on different verification samples. The same is true of the Brier Skill Score (BSS), which is in the standard skill score format (score of the forecast - score of the standard forecast)/(perfect score - score of the standard forecast). A further problem with the BSS is that it can become unstable when the sample is too small and/or the forecast event rarely occurs. For example, a series of monthly BSS values for a station in a dry area is likely to be rather noisy because of variations in the sample climatology from month to month. It is often advisable to use a longer averaging period for this score, especially when the event is relatively rare.

A verification sample consisting of matched observations of a categorical event and the corresponding forecast probabilities can be stratified in two distinct ways. If the data are stratified according to the forecast probabilities (all the 10% +/- 5% forecasts put in one bin etc), then this leads to a reliability table and associated measures. Stratification by observation, that is, separating the occurrences of the event from the non-occurrences, leads to the likelihood diagram, measures of discrimination, and the relative operating characteristic (ROC) curve. These are the principle graphical (tabular) verification measures for probability forecasts, and are discussed briefly below. Contingency tables also are produced from partitioned data, but they are primarily designed for categorical forecasts where the uncertainty has been removed, and will not be discussed here. It should be noted that the two methods of stratification correspond to Murphy and Winkler's (1987) "calibration-refinement" and "likelihood-base rate" factorizations, respectively.

Reliability tables are primarily a way of presenting the degree of reliability of a probability forecast system in graphical form. "Reliability" is the degree to which the probability forecast agrees with the actual frequency of occurrence of the event. On the diagram, perfect reliability is represented by the 45 degree line. For points on that line, the actual frequency of occurrence of the event in the verification sample is exactly equal

to the forecast probability of the event. For perfect reliability, the forecast is "accurate" as stated: 10% means 10%, 70% means 70% etc, and systematic departures from the accurate probability forecast (biases) are easy to see on the diagram. However, there is another important attribute of a good probability forecast, it should have some degree of sharpness, the tendency to predict probabilities that are significantly different from the climatological frequency. A system which forecasts the climatological probability all the time is a reliable system, but the forecasts are not very useful. Sharpness can also be represented graphically on the reliability diagram by means of a histogram of the frequencies with which each of the probability categories has been forecast. A perfectly sharp forecast is in fact a categorical forecast, represented by bars only at either end of the histogram. If the perfectly sharp forecasts are also accurate (the event was forecast only when it occurred and was not forecast only when it did not occur), this is a perfect forecast. On the reliability table, that is represented by two points, at the lower left and upper right of the table. Real forecasts will lie somewhere in between these extremes. A relatively unsharp (smooth or conservative) forecast might produce a bell-shaped frequency histogram around the climatological frequency of occurrence. Sharper systems may give a uniform distribution of probabilities, while particularly sharp systems will tend towards a U-shaped distribution of probabilities. Sharpness can be expressed numerically by the variance of the forecast probabilities in the sample either with respect to the sample climatological frequency (the sample mean) or with respect to the long term climatological frequency. It should be noted that sharpness is NOT the same as resolution. Resolution is another attribute obtainable from the reliability table which measures the system's ability to separate the forecasts into categories which are associated with different relative frequencies of occurrence. Compared to resolution, sharpness is a simpler attribute to measure and control since it is a function only of the forecasts.

Figure 4 shows two reliability tables for two different forecast methods, but based on the same data. The system on the right is reliable, but relatively conservative, since it does not attempt to forecast greater than 50% probability of this realatively rare event, 24 hour precipitation of more than 10mm. The system represented by the verification on the left is sharper, attempting to forecast all probabilities, but loses reliability in doing so. In terms of a quadratic scoring rule, the conservative system obtains a better score. Sharpness is penalized in quadratic scoring systems unless it is fully supported by reliability.

Stratification of the data sample according to the observation results in two separate distributions of forecast prabability, one for the occurrence of the event and the other for the non-occurrence. When plotted, these two distributions constitute a likelihood diagram. They provide graphically the information needed to assess the ability of the system to discriminate between situations leading to occurrence and situations leading to non-occurrence of the event. A perfect system would once again consist of two spikes, at 0% for the non-occurrences and at 100% for the occurrences. A perfectly non-discriminating system would be indicated by complete overlap of the two distributions, with identical means. In the general case, discrimination is most effective if the means of the two distributions are separated AND the dispersions of the two distributions are relatively small, to minimize the overlap area. Statistics which measure this are the separation of means expressed in terms of the standard deviation of one of the distributions and the Mahalanobis distance statistic

used in discriminant analysis, which is the squared distance between the distribution means divided by the pooled variance within the categories.

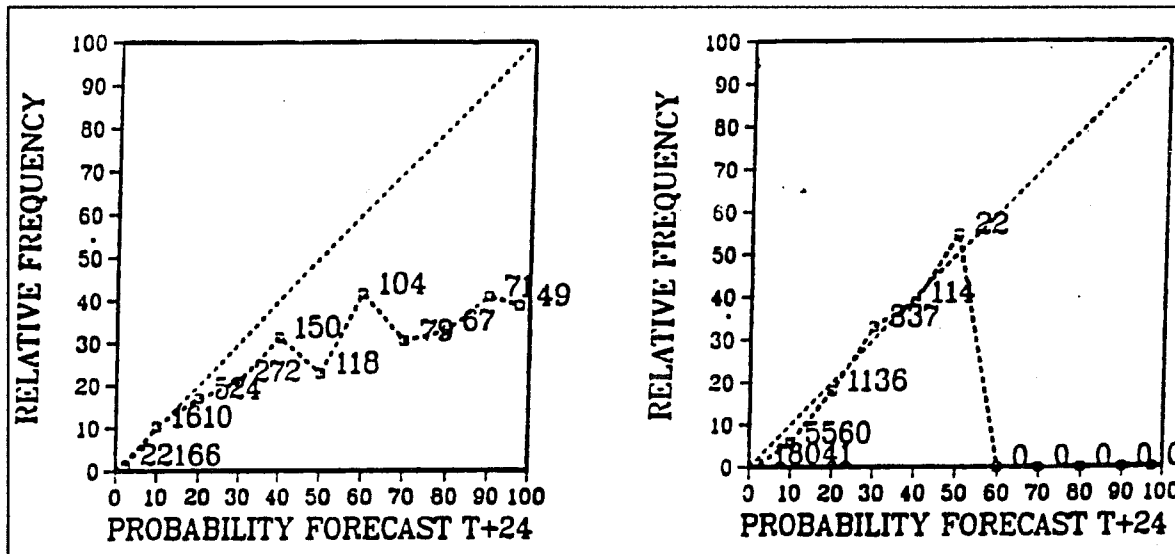

Figure 4. Reliability tables for 24h forecasts of probability of precipitation greater than 10mm in 24h averaged over a large sample for 72 Canadian stations. The forecasts were made by statistical processing of model output using discriminant analysis (left) and regression estimation of event probabilities (REEP) (right). Sample sizes in each 10% probability bin are plotted next to the points of the table.

The relative operating characteristic (ROC) curve, taken from signal detection theory, offers a way of assessing the probability forecasts as a basis for decisions about which category will occur. The concept is similar to the discriminant idea: Taking each 10% probability as a threshold for decision about whether or not the event will occur, the percentage of hits and percentage of false alarms are computed and plotted against each other. A full description of how to prepare the data for the curve is given in Stanski et. al., (1990), and Figure 5 is an example for probability of precipitation forecasts made in Canada over a one year period.

Ideally, if the forecast has detected the signal, if it has discriminated between occurrences and non-occurrences with some degree of skill, the percentage of hits will always exceed the percentage of false alarms, and the curve will lie in the upper half of the diagram. On this diagram, 0 skill is the 45 degree line. Figure 5 clearly indicates the difference in discriminating power between the 12h and 72h forecasts, but both show skill in this context. Quantitative measures associated with the ROC are the area under the curve, with a range of 0 to 1, positively oriented, and 0 skill level at 0.5, and the distance between the two conditional probability distributions. This is exactly the same measure described above, the distance between the distributions expressed in terms of the standard deviation (usually) of the non-occurrence distribution.

Since the ROC curve uses cumulative frequencies in the computation of its variate, a stable graph can be obtained with smaller samples than is possible with the reliability table. However, the most important advantage of the ROC is that it is the only verification measure that I am aware of that permits the simultaneous evaluation of corresponding deterministic and probabilistic forecasts. In the context of the EPS, a probability forecast of, say, an 850 temperature anomaly greater than +4 can be computed and plotted. Then, the T213 model's prediction of that event (or the control model's prediction) for the same sample can be processed into a hit rate and false alarm rate and plotted on the same diagram. The result will be a single point. If the point lies above (below) the curve, the T213 forecast is better (worse).

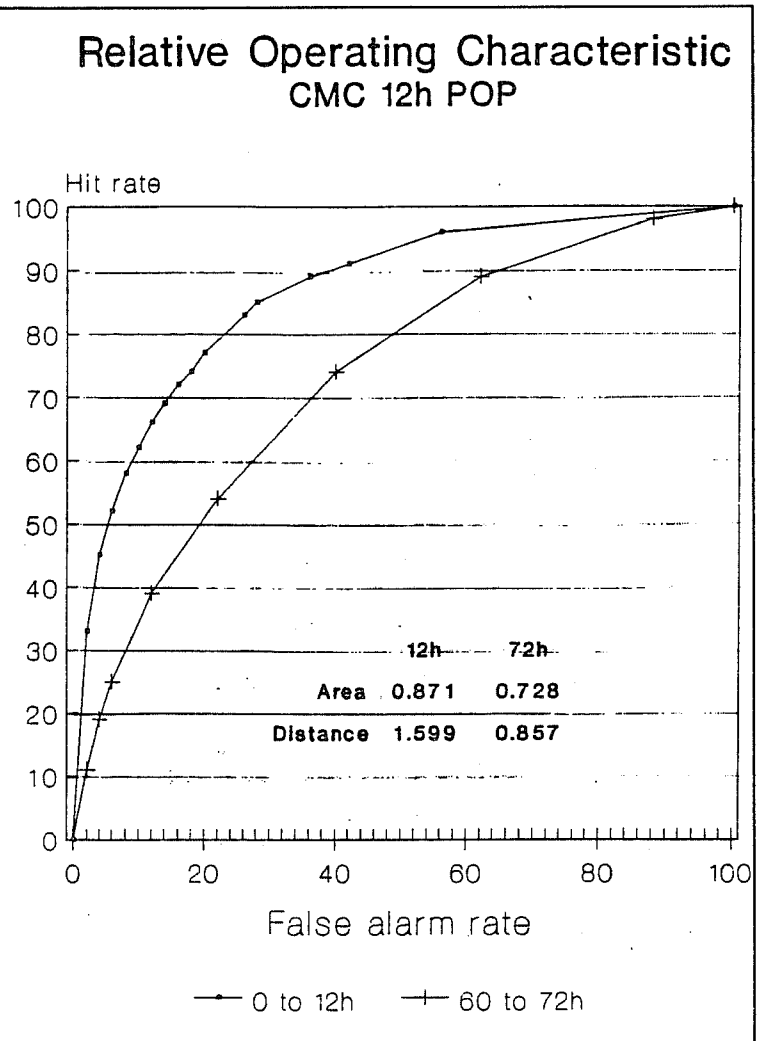

Figure 5. Example of ROC curves, for 12h (upper) and 72h (lower) operational probability of precipitation forecasts for Canadian stations. There are 7250 cases in the sample. The area under the curves and the separation of the two conditional distributions is tabulated on the figure.

5. CONCLUDING REMARKS

This paper describes a variety of verification measures appropriate for use in the evaluation of weather element forecasts from an ensemble prediction system. I have attempted to select measures that can be used together to provide a complete and consistent picture of the performance of the system, and which are consistent with statistical theory and practice. The measures that have been discussed are those that are most relevent to the output of the EPS and which are sensitive to important attributes of EPS products.

Since the EPS output represents an attempt to estimate the probability distribution of the model output arising from uncertainty in the initial conditions, some new formulations have been proposed in recognition of the stochastic nature of the EPS forecasts. These new methods should be validated on output from the EPS to assess their utility. On the other hand, verification rules that have been used to assess post-processed probability forecasts can be applied directly to the ensemble forecasts. The characteristics of these have been described with reference to

previous experience with other data sources. It is hoped this proposal will lead to a standard set of verification measures for EPS forecasts.

6. REFERENCES

Mielke, P., 1973: Another family of distributions for describing and analyzing precipitation data. J. Appl. Meteor., 12, 275-280.

Murphy, A.H., and R.L. Winkler, 1987: A general framework for forecast verification. Mon. Wea. Rev.,115, 1330-1338.

Somerville, P.N., and S.J. Bean, 1979: Probability modeling of weather elements. Preprints, Sixth Conference on Probability and Statistics in Atmospheric Sciences, Amer. Meteor. Soc., Boston, Mass., 173-175.

Stanski, H., L.J. Wilson, and W.R. Burrows, 1990: A Survey of Common Verification Methods in Meteorology, World Meteorological Organization, WWW Technical Report No. 8., 113 pp.