Verification of Precipitation Forecasts: A Survey of Methodology Part I: General Framework and Verification of Continuous Variables

Laurence J. Wilson
Environment Canada

1. Introduction

Precipitation forecasts, whether they are expressed in the form of amounts over a specific period of time, or as instantaneous rates, are difficult to verify because the distribution of precipitation is highly variable in space and time. Variations occur on all scales, but there is usually a significant contribution to the overall variability from smaller scale components. Forecasts and observations may be expressed with different implicit resolutions, which presents problems for the matching of forecasts to the verifying observations.

This paper describes a framework for verification in general, and discusses its application to precipitation verification in particular. Issues concerning the selection of a verification strategy are identified as a starting point for verification of the forecasts obtained from the Sydney Forecast Demonstration Project (FDP), which took place between September 1 and November 30, 2000. Following a discussion of the issues and principles of verification, some specific verification measures for continuous and categorical forecasts are described along with their characteristics.

2. Principles of Verification

2.1 Forecast "goodness"

Allan Murphy (1993) defines the "goodness" of a forecast in terms of three aspects: First, forecasts should be *consistent.* That is, the forecast should always agree with the forecaster's true belief about the future weather. Verification scores which encourage consistent forecasts are called *strictly proper*, which means that the forecaster cannot better his score by systematically issuing a forecast which does not agree with his judgement about the future weather.

The second aspect of goodness identified by Murphy is *quality*, which may be defined as the correspondence between forecasts and the corresponding observations. It is the quality of forecasts that is measured by verification methods; this is the subject of the rest of this paper.

The third aspect of "goodness" is *value*, defined as the increase or decrease in economic or other kind of value to someone as a result of using the forecast. The assessment of value always requires additional information from the user of the forecast, information that describes as objectively as possible the nature of the user's sensitivity to weather events. This additional information is combined with verification information to assess value. Since forecast users are sensitive to weather events in different ways, normally an assessment of forecast value is specific to a particular user and must be recalculated for each different user. The assessment of forecast value involves the branch of statistics known as *decision theory.*

2.2 Basic Principles of Verification

There are some basic principles that apply generally to verification activity. The first of these is that verification of the quality of forecasts has value only if the information generated leads to a

decision about the forecast or system being verified. That is, there should be a target user of the verification information who is identified before the start. The purpose of the verification should also be known in advance because that is the only way the verification methodology can be tuned to meet the needs of the user of the verification information. Decisions resulting from verification activity need not involve a change in the forecast product, a decision that a product is “good enough as is” is also a perfectly valid use of verification information. If verification is done without the knowledge and participation of a user, then the results are most likely to be filed in a drawer somewhere and never used.

A second principle governing verification is that no single measure exists that provides complete information about the quality of a forecast product. All scoring systems are deficient in one or more ways, which means that it is usually necessary to use a variety of measures to obtain reasonably complete verification information. This also means that it is important to be aware of the limitations of the various scores so that they are not used incorrectly.

A third principle that applies to the forecast rather than to the verification is that forecasts must be stated so that they are verifiable. Forecasts that are stated using vague terminology such as “chance of rain” are not verifiable unless further information is provided on the meaning of “chance”. A corollary to this principle is that the predicted quantity must be stated clearly and completely. For example, a precipitation forecast should include statements about the temporal and spatial resolution. Is it valid for a point or a specific area, and is it an instantaneous value, an average over a specific period of time, or integrated over a period of time? In all these examples, the spatial and temporal range of validity should be specified along with the forecast. Verification results are more likely to be fair and unbiased when the forecast quantity is stated completely in advance of the occurrence of the event, so that no a posteriori reinterpretation of the forecast event is possible.

2.3 Factors to consider in verification strategy

All verification activity begins with a matched set of forecasts and observations, a *joint distribution*. Before preparing the data, it is important to consider several factors which have an impact on the data processing.

2.3.1 Goals of the verification

Verification is carried out for a wide variety of reasons, but they can be generally divided into three types: administrative, scientific, and value. Examples of administrative verification goals include justifying the cost of the provision of weather services, justifying the purchase of new equipment, or monitoring the quality of forecasts over periods of years to track improvements in the forecast system. Such purposes normally require considerable summarizing of the verification information into as few values as possible. A single measure may be used, averaged over a large number of forecasts for a large number of locations, over perhaps a full year. Sometimes, administrators may even request the summarizing of verification output over several weather elements, in an attempt to arrive at a single value that characterizes the quality of all weather forecasts from a national weather service. In general, summarizing by averaging or other means, involves the loss of verification information, and can obscure differences in characteristics of the quality of forecasts. However, it is a legitimate procedure for many administrative purposes, as long as one is careful not to “read too much” into the verification results.

Scientific goals of verification concentrate more on learning about the different aspects of the quality of the forecast, to identify its strengths and weaknesses in sufficient detail to indicate possible improvements in the forecast product. In short, the information is sought to help direct R&D. Scientific verification usually focusses on more specific questions about the quality of the forecast, such as how well extreme values are predicted, or whether there are biases in the forecast.

When the goal of the verification is to determine the value of the forecast to a user or a group of users, then the verification information which is needed is specific to that user, and depends on the nature of his (economic) sensitivity to weather. As mentioned above, additional objective information is required from the user to combine with the forecast verification information.

The design of the verification and the measures that are chosen depend on the goal of the verification and on the user of the results. It is therefore important to state the question to be answered by the verification as completely and in as much detail as possible, before the work starts. Examples are: “How accurate are the model’s grid-box average precipitation forecasts?” or, “Is there any skill in point forecasts of 6-h precipitation accumulation?”, or, “Have station temperature forecasts improved over the last 5 years?”

2.3.2 Type of forecast

The design and selection of a verification methodology depends also on the type of forecast, or the type of predictand. Meteorological variables can be divided into three distinct types, continuous, categorical and probabilistic. Continuous variables are those for which the forecast is expressed as a specific value or range of values of the variable, for example, a temperature of 10 degrees or a 6 hour precipitation accumulation of 5.5 mm. Parameters that are usually forecast as continuous variables are temperature, wind direction and speed and upper air variables such as geopotential height and temperature.

WIth a categorical variable, the forecast is for the occurrence or non-occurrence of a particular predictand category, such as weather types snow or rain. All the weather elements are naturally categorized by their occurrence or non-occurrence at a particular place and time, but it is also common to express continuous variables as categorical variables by selecting a set of threshold values to distinguish the categories. For instance, precipitation forecasts are sometimes expressed categorically by setting thresholds at 1mm or 5mm or 10mm or 20mm, then examining the performance of a forecast of the occurrence of precipitation amounts above each of these thresholds. Variables may be categorized into two categories by setting one threshold, or into n categories by setting n-1 thresholds simultaneously. In all cases, the continuous forecasts are “binned” into one of the categories, and the categories are mutually exclusive and exhaustive. That is, each forecast value of the continuous variable is mapped into one and only one category. Effectively, the categorization process means that a smooth distribution of forecast values is represented by a histogram of frequencies of occurrence in each of the categories. It should be noted that the categorization of a continuous variable represents a transformation of the forecast, and forecast information may be lost in the process. Categorical forecasts are represented mathematically by assigning a value of 1 to the category which is forecast and 0 to all the other categories. For verification, these forecasts are matched with observations expressed the same way.

Probabilistic forecasts consist of probabilities of occurrence of the categories of a categorical variable. In a sense, categorical forecasts are a special case of probabilistic forecasts where only two probabilities, 1 (100%) and 0 are allowed to be predicted. Probabilistic forecasts allow for the full range of probabilities to be assigned to each of the categories, with the only restriction that

the probabilities must sum to 1 for a set of mutually exclusive and exhaustive categories. (One and only one of the categories must occur). Meteorological variables that are normally treated categorically include those which are inherently categorical such as precipitation occurrence, the occurrence of other weather and obstructions to vision, and precipitation type. Variables that have continuous distributions but are usually predicted categorically include cloud amount, and precipitation amount.

Another type of forecast used with continuous variables is called a *credible interval* forecast. These are forecasts of ranges of the predictand, where the range is defined according to the forecaster's probability estimate that the event will occur in the range. For example, a 50% percent credible interval forecast of temperature between 12 and 16 degrees might mean that the forecaster believes there is a 50% chance the temperature will lie in this range, with an expected value of 14 degrees. In terms of the three types of forecasts above, the credible interval forecast is a probabilistic forecast where the category thresholds are set by the forecaster on each occasion, rather than being selected a priori and remaining constant for all forecasts. When the forecaster is less confident, his 50% range will be wider. Credible interval forecasts share some characteristics with continuous forecasts also in the sense that any value of the variable can be forecast as the expected value. Credible interval forecasts are not widely used, but they do convey more information than either a forecast of a specific value or a probability forecast of a specific interval.

2.3.3. Attributes of the forecast

Once one has assembled the joint distribution of forecasts and observations, It is possible to examine its characteristics in three main ways: One can look at the overall correspondence or association between the pairs of observations and forecasts, that is, to calculate statistics based on the full joint distribution. Then, alternatively, one can set conditions on the distribution, for instance fixing the forecast values within a narrow range and examining the observations only for cases where forecasts were within that particular range. Verification results obtained this way are said to be *conditional on the forecast*, because it is the forecast that has been restricted to a particular set of values. The distribution of observations obtained for such a subset of cases is said to be a *conditional distribution*. In practice, verification measures that are conditional on the forecast are computed by first stratifying or "binning" the data sample by forecast value, which is equivalent to setting the condition. Similarly, if the observation is constrained to a specific set of values, then by "binning" the data sample according to observation value, statistics can also be computed which are *conditional on the observation*. These three types of analysis of the joint distribution allow one to examine different aspects of the quality of forecasts, called *attributes*.

Table 1: The nine attributes of forecasts, with definition and some related verification measures.

ATTRIBUTE	DEFINITION	RELATED MEASURES
1. Bias	Correspondence between mean forecast and mean observation	bias (mean forecast probability-sample observed frequency)
2. Association	Strength of linear relationship between pairs of forecasts and observations	covariance, correlation
3. Accuracy	Average correspondence between individual pairs of observations and forecasts	mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), Brier score (BS)
4. Skill	Accuracy of forecasts relative to accuracy of forecasts produced by a standard method	Brier skill score, others in the skill score format.

Table 1: The nine attributes of forecasts, with definition and some related verification measures.

ATTRIBUTE	DEFINITION	RELATED MEASURES
5. Reliability	Correspondence of conditional mean observation and conditioning forecast, averaged over all forecasts	Reliability component of BS, MAE, MSE of binned data from reliability table
6. Resolution	Difference between conditional mean observation and unconditional mean observation, averaged over all forecasts	Resolution component of BS
7. Sharpness	Variability of forecasts as described by distribution of forecasts	Variance of forecasts
8. Discrimination	Difference between conditional mean forecast and unconditional mean forecast, averaged over all observations	Area under ROC, measures of separation of conditional distributions; MAE, MSE of scatter plot, binned by observation value
9. Uncertainty	Variability of observations as described by the distribution of observations	Variance of observations

Murphy (1993) identifies nine attributes of forecasts. In addition to the general goal of the verification and the type of forecast variable, the selection of a verification methodology depends also on which attributes of the forecast are to be studied. Table 1 shows the nine attributes and their definition in terms of unconditional and conditional distributions. The first 4 of these, bias, association, accuracy and skill, relate to the analysis of the unconditional joint distribution of forecasts and observations; the next three relate to stratification by (conditioning on) the forecast, while the last two are associated with stratification of the dataset by observation. The attributes can be considered to be different dimensions of the verification problem; if one wishes to obtain reasonably complete verification information about a set of forecasts, it is necessary to examine several of these attributes using the appropriate measures. Attributes are not dependent on the type of forecast. Though they might be given different names, corresponding attributes exist for continuous forecasts and categorical or probabilistic forecasts.

Allan Murphy in his landmark article with Winkler (Murphy and Winkler 1987) states that all verification information is contained in the joint distribution of forecasts and corresponding observations. Murphy and Winkler describe two types of factorization of the joint distribution, which correspond to the two types of stratification described above. The *calibration-refinement* factorization is equivalent to stratification by forecast value. It is expressed more formally by writing the joint distribution

$$p(f, x) = p(x|f)p(f)$$

where $p(x|f)$ is the conditional distribution of the observations given the forecast and $p(f)$ is the unconditional (marginal) distribution of forecasts. Stratification by observations leads to the other factorization, called *likelihood-base rate* by Murphy and Winkler (1987),

$$p(f, x) = p(f|x)p(x)$$

where $p(f|x)$ is the conditional distribution of the forecast given the observations and $p(x)$ is the unconditional (marginal) distribution of the observations, which is the sample climatological distribution. The joint distribution $p(f, x)$ contains all the information needed to compute verification measures and to assess the attributes of the forecast; once the forecasts are matched with their

corresponding observations, the different verification measures represent different ways of processing the component events of the joint distribution.

2.3.4 Data issues

The matching of forecast and observation is not always a trivial process, especially when forecast and observation have different characteristics. For variables such as precipitation rate or amount, which often exhibit considerable small scale variability in space and time, the determination of the joint distribution may be especially difficult and choices must be made. It is convenient to characterize observation datasets by three characteristics, *sampling*, *resolution*, and *proxy data*.

It should always be the intent of verification activity to obtain the best possible observation dataset for the purpose. However, observations are never complete in space and time. The *resolution* of the observation can be considered to be the inherent spatial averaging area and/or the temporal averaging period of the actual measurement. For example, for a satellite observation, the spatial resolution is the “footprint” or the pixel size of the measurement; for a radar observation, it is the horizontal projection of the sampling volume, which varies with range. For rain gauge observations, the resolution may be determined in terms of representativeness, which will vary according to the siting, but would be expected to be not more than 100 to 200 m radius in the vicinity of the gauge. The temporal resolution can be defined in terms of the inherent averaging period of the observation, which is essentially instantaneous for satellite and radar observations, but may be much longer for gauges, depending on the instrument. Gauges normally report accumulated precipitation, which is averaged over the accumulation period.

The issue of observation *sampling* frequency is particularly important for elements with small scale variability because of the need for high spatial and temporal sampling frequency to adequately describe the small scale variations. A gauge is taking observations continuously if it is a recording gauge (high sampling frequency) but may have a long averaging period (low resolution), depending on how observations are reported. Satellite observations are typically relatively high resolution in space, but the sampling frequency may be very low in time, perhaps with a return period of several days to a particular location. Radar data is high resolution in space and continuous spatial sampling (all locations within the radar range are sampled), but the temporal sampling is discrete, depending on the return period of the radar to a particular sampling location.

It is tempting to try to combine the continuous temporal sampling of gauges with the continuous spatial sampling of radar to obtain a complete spatial and temporal representation of precipitation occurrence. However, radar data is *proxy* data, which means that the physical parameter is not observed directly. Rather, an equation (the Z-R relationship) must be used to convert the signal received by the radar into an estimate of the physical parameter, rainfall. By means of ground-truthing studies, precipitation observed by gauges can be related to corresponding radar signatures. However there are several potential or real sources of error in such relationships: shadowing, anomalous propagation and ground clutter in the radar echo and uncertainty due to differences in the location of the sampled radar volume vs. the corresponding ground location. Gauge observations are subject to representativeness errors related to their siting and location, which means that the true relationship between radar-sampled precipitation and gauge precipitation observations will vary from station to station. Despite all these potential errors, ground-truthing is certainly worth doing, to optimally extract information about the precipitation field from both sources of data.

In situ observations from gauges are normally represented with certainty in the dataset of observations, that is, by means of a specific precipitation amount or categorically as 1 or 0 for the occurrence or non-occurrence of precipitation respectively. The use of proxy data with its associated errors and uncertainties raises the prospect of a probabilistic representation of observation data, for example as a credible interval around a specific precipitation amount, or a probability that the precipitation actually occurred over a preset threshold, given the evidence from the radar signal. Such probability estimates should be calibrated using ground-truthing studies.

Given all these limitations of observations, the question arises: Is it appropriate to process the observation data, by analysis or other means, to match it to the forecast? In general, the answer to this is “no” because this leaves the verification process open to biases in favour of the forecasts being verified. There are, however, exceptions to this guideline, as described below. How to match the forecast and observation depends on the characteristics of each. Given that model precipitation forecasts are often claimed to be “grid-box” averages, then the matching problem is one of resolution differences: how to match a forecast that has spatial resolution equal to the size of the grid box to irregularly spaced gauge observations which have high spatial resolution. The simplest and fairest is to treat the observations as estimates of the grid box average and match them with the forecasts. In the absence of information on the systematic (climatological) distribution of precipitation in the grid box area, a single point estimate of precipitation will tend to underestimate slightly the areal mean since precipitation amounts are distributed according to the gamma distribution (Mielke 1973). The gamma distribution is skewed and thus has a mode (maximum probability density) lower than the mean. If spatial climatology information is available for an area, the observation could be adjusted in light of this information to obtain a better estimate of the area mean precipitation. Analysis of scattered observations onto the grid of the model is not recommended because this systematically processes the data distribution so that it has resolution characteristics of the model being verified. This is especially true of analysis routines that use trial fields, but is true to some extent of all spatial analysis routines. It is questionable whether point observations should be analyzed at all for verification purposes because the drastic undersampling of the true precipitation field translates into uncertainties in the analysis at all scales. In other words, if a specific precipitation field were sampled at different points with the same spatial sampling frequency, then a different analysis would result, whatever the resolution of the analysis. Again, this is particularly true when the field contains large amounts of small scale variability. In general, an attempt to filter the observations so that they contain only the scales that can be resolved by the model is really an attempt to enable a model to obtain a perfect verification score without producing a perfect forecast. The inability of a model to resolve all scales is simply one source of error which should not be eliminated from the verification.

If one has the luxury of having more than one point observation within a grid box, then an average of the observations within the grid box will provide a better estimate of the grid box average observed precipitation. It is fair and perhaps preferable to carry out verification analysis of model forecasts at several different resolutions, by averaging both grid point forecasts and observations over larger areas. The greater the number of samples in an area, the more accurate the estimate of the area average precipitation, again assuming a climatologically homogeneous distribution.

When the forecast is expressed as point values at locations different from the observation points, then it is preferable to go from forecast to observation, that is, to interpolate the forecasts to the observation sites rather than the other way around. This recognizes that the forecast is limited to estimates of the forecast quantity at a subset of sites; failure to forecast at all the sites is a limitation of the forecast which is expressed through the interpolation.

Under some circumstances, processing of observations is fair. For example, when the purpose of the verification is to assess the quality of changes to the physics of a model, it may be preferable for research purposes to process the observations to remove small scale components of the predictand which have nothing to do with the changes in the model, and which would be seen as noise that would obscure any differences due to differences in model formulation. The essence here is that the purpose of the verification is the comparison of two versions of the model at the same resolution; both versions would be expected to be subject to the same limitations of resolution.

A second type of instance when processing of observations is acceptable are those situations when the expression of the forecast is placed under specific limitations which are known in advance (a priori). An example of this is Harold Brooks' application of a Gaussian kernel distribution around severe weather reports to determine a "practically perfect" forecast (Brooks, in this volume). A practically perfect forecast is the forecast that would have been made under the known constraints, if the forecaster knew the outcome (observations) in advance. That is, it is the best forecast that could be made. One should still distinguish this from a true perfect forecast, however, which in the case of a precipitation forecast, would predict the precipitation field perfectly at all scales. Restrictions are often placed on the expression of forecasts to prevent overconfidence when it is not justified by the state of the art. Practically perfect forecasts are therefore with reference to the state of the art rather than perfect in any absolute terms.

In summary, statistical estimation of the observed counterparts to forecast quantities is preferable to analysis of observations onto a specific grid, and it is desirable to take advantage of all sources of observation information to obtain optimal estimates.

3. Verification methods for continuous forecasts

3.1 Scatter plots

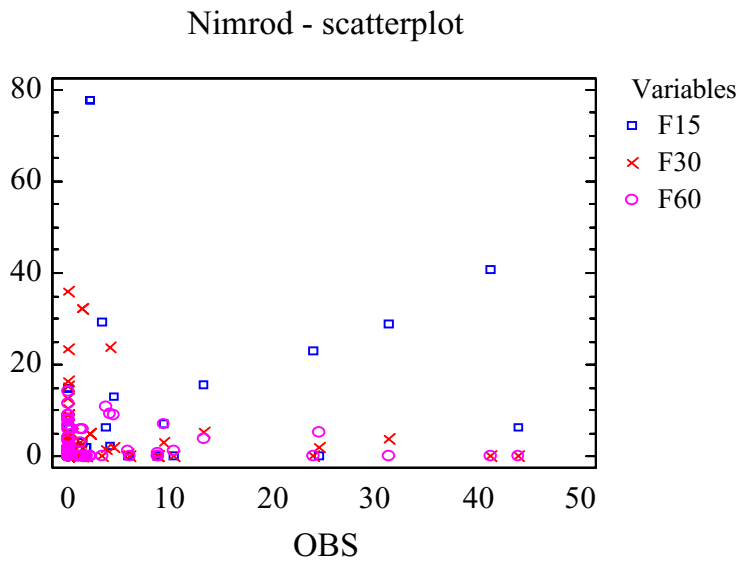

Figure 1. An example of a scatter plot, showing forecasts vs. observations for 15, 30 and 45 minute forecasts of instantaneous precipitation rate in mm/h from the Nimrod system.

Since verification data comes in the form of bivariate joint distributions, it is a simple matter to plot the data as a forecast vs. observed scatterplot. Scatter plots are fundamental, as they provide an instant visual comparison of forecasts and corresponding observations, and all the data are visible - there is no loss of information. For large samples, scatter plots may become cluttered, in which case the point density can be reduced in various ways, for example, by binning the data and plotting a single point at the mean value of each bin.

Figure 1 shows an example of a scattershot for the Sydney FDP data. There are 152 points plotted, representing forecasts and obser-

vations at 152 station locations for a specific time. The verifying observations are radar rain rates, and they are matched to the nearest grid point of the forecast. All forecasts and observations are valid at 0330 UTC on November 3, 2000. It is evident from the scatter plot that there is a relationship between forecast and observation at only a few of the locations, and then only for the fifteen minute forecast projection. The outliers in the sample are also immediately evident, for example the forecast of 78 when 3 was observed and a forecast of 7 when 44 was observed.

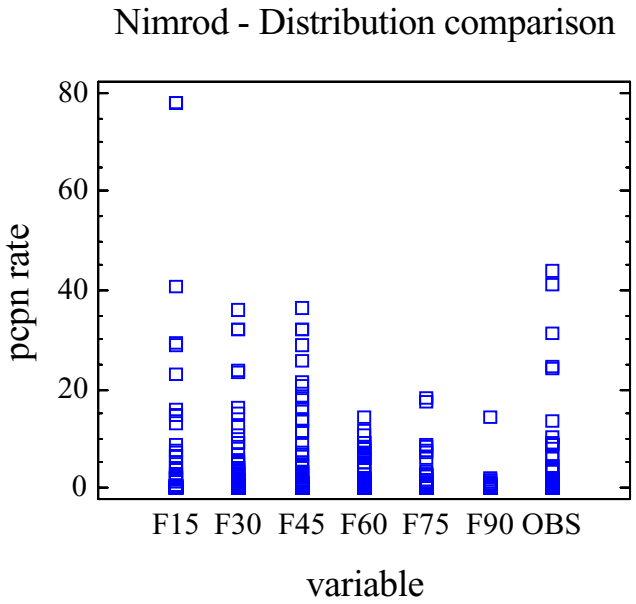

Figure 2. Another graphical way to compare distributions of several variables. Example from Nimrod, for observations compared to 15 to 90 minute forecasts. Units are mm/h.

.Figure 2, also for the Sydney FDP data, shows another way of comparing the distribution of several variables. Here it can be immediately seen that the shortest range forecasts up to 45 minutes have approximately the same dispersion as the observations, but the longer range forecasts from 60 to 90 minutes, have a much smaller variability than the observations. This form of display does not evaluate the joint distribution, but rather allows comparison of the characteristics of the distributions of forecasts and observations (the two marginal distributions).

Scatter plots are essentially graphical depictions of the joint distribution of forecasts and observations. They are simple and extremely valuable diagnostic tools. If used early in a verification analysis, they help identify suspicious data values, as well as helping identify specific events in the sample that warrant further attention.

3.2. Scores for continuous forecasts.

The most frequently used scores for continuous forecasts are bias, mean absolute error (MAE), root mean square error (RMSE), variance explained or reduction of variance, and correlation. Each of these evaluates a specific attribute of the forecast, summarizing it by means of a single value for a given dataset. These are summarized in Table 2 along with their characteristics and ranges.

Table 2: Summary of the equations and characteristics for the most frequently used verification measures for continuous forecasts.

Measure	Equation	Range	Best score	Characteristics
bias	$\sum_{i=1}^{N} \frac{(f_i - x_i)}{N}$	-∞ to ∞	0	mean error over a sample

Table 2: Summary of the equations and characteristics for the most frequently used verification measures for continuous forecasts.

Measure	Equation	Range	Best score	Characteristics
MAE	$\sum_{i=1}^{N} \frac{\lvert f_i - x_i \rvert}{N}$	0 to ∞	0	average magnitude of errors, linear
RMSE	$\sqrt{\sum_{i=1}^{N} \frac{(f_i - x_i)^2}{N}}$	0 to ∞	0	quadratic scoring rule: larger errors carry higher weights; comparison RMSE - MAE indicate the error variance; $rmse \geq mae$; CAUTION: sometimes mean error is subtracted when computing these scores
RV	$1 - \frac{\sum_{i=1}^{N} (f_i - x_i)^2}{\sum_{i=1}^{N} (c_i - x_i)^2}$	-∞ to 1	1	a skill score in the usual format with respect to climatology. Is interpreted as the % of predictand variance explained, or the % improvement over climatology. Can be referenced either to long term or sample climatology. CAUTION: can become unstable if climatology is accurate, and for small samples.
correlation (r)	$\frac{\sum_{i=1}^{N} (f_i - \bar{f})(x_i - \bar{x})}{N s_f s_x} = \sqrt{RV}$	-1 to +1	1	The degree of linear association; insensitive to bias; covariance normalized by the product of the standard deviations; square root of the RV when RV positive.

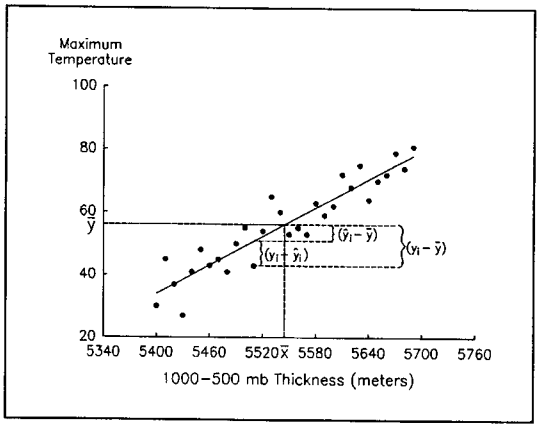

Figure 3. Diagram showing partitioning of the variance of predictand y into explained and unexplained components. In this example, the predictor (forecast) is "thickness" and the predictand (observations) is "maximum temperature".

In the table, the sample size is N, the forecast and observed variables are (f_i, x_i) respectively, c_i is the corresponding climatological forecast value, the overbar indicates the sample mean, and (s_f, s_x) are the standard deviations of the forecast and observations. The reduction of variance or variance explained is related to the partitioning of the variance of the predictand (observations) into two components, the portion explained and the portion unexplained by the forecast (the errors). This is illustrated graphically in Figure 3. Figure 3 is actually an illustration of a bivariate regression; regression, correlation and reduction of variance are all closely related.

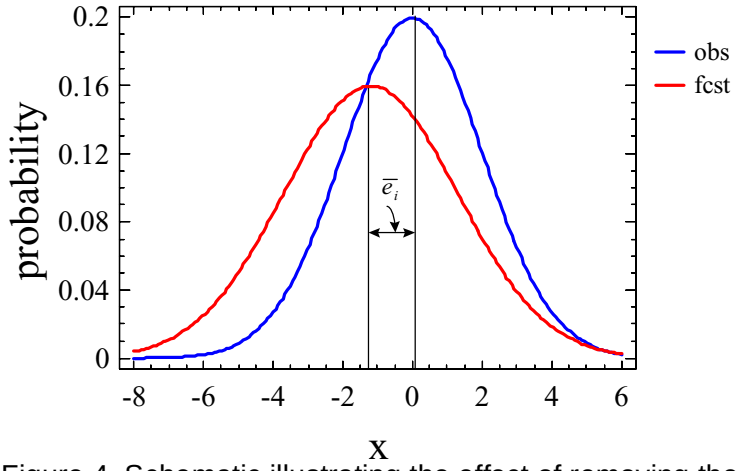

Figure 4. Schematic illustrating the effect of removing the average error before computing the MAE or RMSE. This is equivalent to repositioning the forecast distribution so that the means are coincident.

Figure 4 illustrates the effects of removing the bias before computing the MAE or the RMSE. The bias is represented by the difference between the forecast and observed mean, the difference in the means of the two distributions. This distance is effectively removed from each error value, which is equivalent to adjusting the forecast distribution to be centered on the observed distribution before computing the MAE or RMSE. There are occasions when this is desirable, for example when two forecasts are being compared against the same set of observations. Separation of errors in this way allows for direct comparison of the variable component of the forecast error from the two forecast systems.

In summary, we can define the following statistics of the error, where the error of the i^{th} event is defined by $e_i = f_i - x_i$:

$$\bar{e} = bias = \bar{f} - \bar{x} = ME$$

$$MSE = \frac{\sum e_i^2}{N}$$

$$S_e^2 = error\ variance = \frac{\sum (e_i - \bar{e})^2}{N} = MSE - (ME)^2$$

$$Stderror = \sqrt{S_e^2}$$

The bias and the error variance are the first two moments of the distribution of errors, which is formed by pairwise subtraction of the observations from the forecasts. The RMSE calculated by first removing the sample bias is equal to the standard error, however, this is NOT the same as subtracting the bias from the RMSE computed from the original data. The mean squared error is reduced by the square of the mean error when the mean error is removed first. When the bias is large, this can make quite a difference. Sometimes the bias is associated with the accuracy and the (R)MSE is associated with the confidence on can place in the forecast. It is often useful to separate bias and error variance when comparing different forecast systems so that one may compare the components of the error which can be easily corrected (bias) vs. those which are harder to correct (error variance, or *unexplained variance*)

One verification measure for continuous forecasts that is less often used is called Linear Error in Probability Space (LEPS). First introduced by Ward and Folland (1991), this method evaluates forecasts in terms of the cumulative distribution function (cdf) of the observations. The formula is:

$$LEPS = \frac{\sum_{i=1}^{N} |F_o(f_i) - F_o(x_i)|}{N}$$

where F_o is the climatological cdf, or the cdf of the observations (sample climatology). It is easiest to use the sample cdf rather than long-term climatology, but the latter would be useful when it one wishes to compare the accuracy of forecasts of common events vs. forecasts of extreme events. The range of LEPS is 0 for a perfect forecast to 1 for the worst possible forecast. Since the score is sensitive to differences in the cumulative probability of forecast and observed, forecast errors in the higher probability density part of the distribution are weighted more strongly. This could be quite appealing in the case of precipitation verification. An error of, say, 1 mm in a quantitative precipitation forecast would be assigned a higher penalty if the observed amount were small, and a much lower penalty if the observed amount were high. However, both missed events and false alarms would be given relatively high penalties.

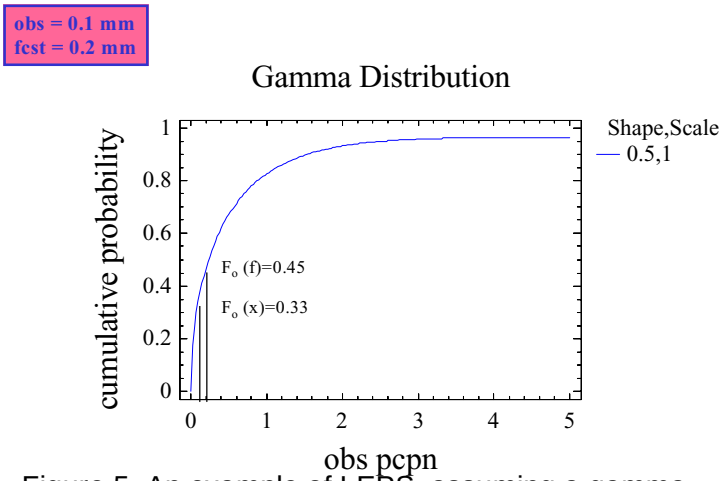

Figure 5. An example of LEPS, assuming a gamma cdf, for an observation of 0.1 mm and forecast of 0.2 mm precipitation. In this case the score value would be 0.12.

Figure 5 shows an example of the computation of the LEPS score for an observation of 0.1 mm and forecast of 0.2 mm precipitation. The observations are assumed to be distributed according to the gamma distribution which has the cdf as shown. One can visualize from the graph that the penalty assigned to errors in QPF would be relatively small when both observation and forecast are for extremes.

One can formulate a skill score for LEPS, in the standard format, $SS = \frac{score - standard\ score}{perfect\ score - standard\ score}$ The easiest standard to use would be the median of the distribution, which has a LEPS value of 0.5. Note that the median, or other statistic of the distribution is not known in advance if the sample distribution is used to evaluate LEPS, which means that the "standard forecast" is not available to the forecaster when he makes his forecast. For this reason, it might be preferable to calculate LEPS with the long term climatological distribution if a reasonable estimate of skill is wanted. The LEPS skill score with respect to the median is:

$$SS_{LEPS} = \frac{\sum |F_o(f) - F_o(x)| - \sum |0.5 - F_o(x)|}{0 - \sum |0.5 - F_o(x)|}$$
$$= 1 - \frac{\sum |F_o(f) - F_o(x)|}{\sum |0.5 - F_o(x)|}$$

It would be worthwhile to assess LEPS as a verification method for precipitation.

4. Verification of Categorical Forecasts

Categorical forecasts are forecasts of the occurrence or non-occurrence of specific categories of the predictand variable. If, for example, one wishes to predict precipitation accumulation over 25 mm in 6 h, the 25 mm in 6 h amount becomes a threshold for this category, and a categorical forecast would be a “yes-or-no” forecast of the occurrence of this event. Categorical forecasts are restricted in the information they contain since the forecaster has no means of conveying his judgement about the level of uncertainty in the forecast value.

Categorical forecasts are usually verified using contingency tables and various scores associated with them. Contingency tables are formed by making a table of all the possible combinations of forecast and observed categories, and tallying up the number of forecast-observation pairs that fit each combination. To this is added marginal totals formed by summing the rows and columns of the table, and the sample size, which appears in the lower right corner as a sum of the marginal totals. Though the entries of the table are usually total numbers of events for each forecast-observation category combination, it is possible to divide through by the sample size and express each total as a percentage of the sample. Contingency tables are equivalent to scatter plots; all the information of the joint distribution of forecasts and observations is contained in the table. To satisfy oneself of this equivalence, it is necessary only to imagine category thresholds imposed on a scatter plot, and represented graphically by vertical and horizontal lines drawn at the threshold value. These effectively divide the scatter plot into a table, and the points on the plot can be simply summed within each of the boxes so formed, in order to turn a scatter plot into a contingency table for categorized variables.

The dimensions of the contingency table are *K* x *K* where *K* is the number of categories. The simplest and probably the most studied table is for *K=2*, the 2 x 2 table for verification of two-state categorized variables (binary variables). In the following discussion, contingency table scores are presented for the 2 X 2 version, but all can be easily generalized to more than two categories. Two by two tables are often used in situations where one of the two categories is infrequent and/or important, to verify specific events of importance, for example the occurrence of extreme or severe weather. It is common to present the table with the rarer category as the one of interest, associated with the “yes” forecast, and in the upper left of the table.

Table 3 shows a schematic of a 2 by 2 contingency table indicating the common nomenclature of the individual cells of the table. the letters a, b, c, d represent the total events from the sample which fit the indicated forecast-observed combination. The marginal totals correspond to the marginal (unconditional) distributions of forecast and observed values for the categorical variable.

Table 4 summarizes the common verification measures associated with contingency tables, along with their formulae and characteristics.

Table 3: Nomenclature for the cells of a 2 by 2 contingency table.

Observations \ Forecasts	Yes	No	Total
Yes	HITS a	MISSES b	Total events observed $a + b = O_1$
No	FALSE ALARMS c	CORRECT NEGATIVES d	Total non-events Observed $c + d = O_2$
Total	Total events Forecast $(a + c) = F_1$	Total non-events Forecast $b + d = F_2$	Sample Size T

Table 4: Summary of common measures used to evaluate categorical forecasts via contingency tables.

Measure	Formula	Range; best score	Characteristics
proportion correct Hit Rate (PC)	$PC = \frac{a+d}{T}$	0 to 1; 1	-Dominated by common categories -Can be maximized by forecasting the most common category all the time.
Probability of detection (PoD); Prefigurance	$PoD = \frac{a}{O_1}; \frac{d}{O_2}$	0 to 1; 1	-sensitive only to missed events, not false alarms: can always be increased by overforecasting rare events
False Alarm Ratio (FAR)	$FAR = \frac{c}{F_1}; \frac{b}{F_2}$	0 to 1; 0	-sensitive only to false alarms, not missed events; can always be improved by underforecasting rare events.
Post-agreement (PAG)	$PAG = \frac{a}{F_1}; \frac{d}{F_2}$ $= (1 - FAR)$	0 to 1; 1	-same as FAR
Threat score; Critical Success Index (CSI) or (TS)	$TS = \frac{a}{a+b+c}$ $; \frac{d}{b+c+d}$	0 to 1; 1	-sensitive to both false alarms and missed events; a more balanced measure than either PoD or FAR

Table 4: Summary of common measures used to evaluate categorical forecasts via contingency tables.

Measure	Formula	Range; best score	Characteristics
Equitable Threat Score (ETS)	$ETS = \frac{a - \frac{F_1 O_1}{T}}{a + b + c - \frac{F_1 O_1}{T}}$ $; \frac{d - \frac{F_2 O_2}{T}}{b + c + d - \frac{F_2 O_2}{T}}$	0 to 1; 1	-Equitable: correct forecasts of either category get same overall score -Constant forecasts of either category get 0 -Threat score adjusted for the number expected correct by chance given the observation distribution.
Frequency bias	$bias = \frac{F_1}{O_1}; \frac{F_2}{O_2}$	0 to ∞; 1	-ratio of frequency forecast to frequency observed without reference to correctness. -comparison of forecast and observed distributions.
Heidke Skill Score (HSS)	$HSS = \frac{(a + d) - \frac{F_1 O_1 + F_2 O_2}{T}}{T - \frac{F_1 O_1 + F_2 O_2}{T}}$	$-\infty$ to 1; 1	-Skill score in the usual format with chance as the standard as shown here -can be computed using other standards; number correct by standard substituted for RHS term. -chance relatively easy to beat in practice.
Hanssen-Kuipers Discriminant; True Skill Statistic (KSS or TSS)	$KSS = \frac{ad - bc}{(a + b)(c + d)}$	$-\infty$ to 1; 1	-an equitable skill score against chance -random forecasts and constant forecasts get 0 score -in practice, gives results a lot like PoD -relates to stratification by observation, i.e. a measure of discrimination
False Alarm Rate (FA)	$FA = \frac{c}{O_2}; \frac{b}{O_1}$	0 to 1; 0	-false alarms relative to the total observations of each category -usually used in combination with the Hit Rate in the calculation of the relative operating characteristic curve -not to be confused with FAR, FA is rarely used alone.

Further details on the definition and interpretation of verification measures for both continuous and categorical forecasts can be found in Stanski et. al. (1989).

5. Recommendations

The above discussion leads to several recommendations regarding the verification of precipitation forecasts from the Sydney FDP. These are:

1. Use the gauge data as estimates of the area average precipitation for areas in which the observation sites are located. This is preferable to analyzing high resolution datasets that under sample the precipitation field.
2. Carry out ground truth studies of the radar observations wherever possible. Climatological studies of spatial precipitation distribution, as revealed by radar, are encouraged to improve estimates of areal average precipitation, and, when calibrated by collocated station data, to provide estimates of precipitation at higher spatial resolution.
3. Investigate the LEPS score for precipitation verification.
4. Allow for the expression of observation data probabilistically, with probabilities of the predictands estimated using all available sources of data.
5. Carry out verification analyses at different scales by averaging both model grid estimates and observations over successively larger areas.

6. References

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

Murphy, A.H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. and Forecasting*, **8**, 281-293.

Stanski, H.R., L.J. Wilson and W.R. Burrows, 1989: *Survey of Common Verification Methods in Meteorology*, World Weather Watch Tech. Rep. No. 8, TD 358. World Meteorological Organization, Geneva, 114 pp.

Ward, M.N., and C.K. Folland, 1991: Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea surface temperature. *Int. J. Climatol.*, **11**, 711-743.