METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **15**: 31–40 (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/met.51

RMetS
Royal Meteorological Society

Understanding forecast verification statistics

S. J. Mason*
International Research Institute for Climate and Society, Columbia University, Palisades, USA

ABSTRACT: Although there are numerous reasons for performing a verification analysis, there are usually two general questions that are of interest: are the forecasts good, and can we be confident that the estimate of forecast quality is not misleading? When calculating a verification score, it is not usually obvious how the score can answer either of these questions. Some procedures for attempting to answer the questions are reviewed, with particular focus on p-values and confidence intervals. P-values are shown to be rather unhelpful in answering either question, especially when applied to probabilistic verification scores, and confidence intervals are to be preferred. However, confidence intervals cannot reveal biases in the value of a score that arises from an inadequate experimental design for testing on truly out-of-sample observations. Some specific problems with cross validation are highlighted. Finally, in the interests of increasing the insight into forecast strengths and weaknesses and in pointing towards methods for improving forecast quality, a plea is made for a more discriminating selection of verification procedures than has been adopted to date. Copyright © 2008 Royal Meteorological Society

KEY WORDS bootstrap; cross validation; effectiveness; equitability; locality; p-values; permutation; propriety

Received 18 September 2007; Revised 6 December 2007; Accepted 2 January 2008

1. Introduction

There are numerous reasons for verifying forecasts (Jolliffe and Stephenson, 2003), but having calculated any verification score, or performed any verification procedure, two related questions almost invariably arise: do the results mean that the forecasts are good (is there a strong correspondence between the forecasts and the observations (Murphy, 1993)); and do the results provide an accurate indication of how good (or bad) subsequent forecasts will be (is the score misleadingly high or low)? For example, given a set of paired forecasts and observations of precipitation occurrences for the last 30 days that give a Brier score of 0.1, the following question arises: is this score of 0.1 a good score? If the score is good then it may be concluded that the forecasts are good, but not if the score is bad. However, neither conclusion can be drawn if it is suspected that the score is not an accurate indication of the quality of the forecasts. Specifically, even if an excellent score was achieved, some kind of guarantee that these forecasts will continue to score about as well in the future would be desired. For example, if a new forecast procedure has been tested, it would be desirable to know whether the test results are believable rather than being taken by surprise at the quality of the forecasts once they are issued in an operational setting.

In this review article, procedures for addressing these two questions about the goodness of, and about the uncertainty in, the verification score are discussed. In addition to answering these questions, the procedures provide a means of comparing forecasts, which may indicate whether one forecaster is better than another, or whether the forecasts have improved since they were last evaluated, for example.

2. Is the score good?

For many verification scores the numerical value of the score is essentially an abstract number, and so the inexperienced practitioner may have no idea as to what the value indicates. Two questions arise when trying to interpret a score: does the score indicate that the forecasts are in fact good; and what does the score itself mean? The second question relates to identifying which of the many attributes of forecast quality (Wilks, 2006a) are communicated by the score, and is discussed only briefly here (further discussion is provided in Section 4.5 and in Mason and Stephenson (2008)). Instead the focus is on the first question: is, for example, a Brier score of 0.1 a good score? Unfortunately, there are too many unknowns to give a simple answer to this question. Firstly, a score of 0.1 would be a better score if the events occurred about half the time than if they occurred almost all the time (or almost never). If the events occur most of the time (or very infrequently), the forecaster would presumably learn this fact rather quickly, and would learn to issue a probability close to 1 (or 0) most of the time. The forecaster would then frequently be issuing probabilities very close to perfect forecasts, and so would easily achieve a low Brier score. The inherent uncertainty, or the

* Correspondence to: S. J. Mason, International Research Institute for Climate and Society, The Earth Institute of Columbia University, Palisades, NY 10964-8000, USA. E-mail: simon@iri.columbia.edu

'base rate' (Murphy and Winkler, 1987) of the events has to be considered when interpreting and comparing Brier (and many other) scores. Thus, the decomposition of the Brier score by Murphy (1973) includes an uncertainty term that is independent of the forecasts, indicating that a simple comparison of Brier scores for winter compared to summer daily precipitation occurrence, for example, is not straightforward if there is some seasonality in the frequency of precipitation.

2.1. Calculating *p*-values

Another difficulty in interpreting the value of a score is that a low score (assuming that it is negatively oriented) is more convincing if it was calculated using a large number of forecasts compared to from just a handful of cases. It is much easier for a good score to be achieved accidentally given only a few forecasts, but a large number of forecasts are only likely to score well if the forecasts are in fact good. This problem of sensitivity of the score to the sample size of the forecasts is closely related to the uncertainty in the score (if a larger sample were available, would the forecasts score similarly?). It is considered in greater detail in Section 3, but for the present purposes a suggested solution might be to calculate the probability that the value that was achieved could have been bettered by accident. What are the chances that a Brier score of 0.1 or better (i.e. less) could have been obtained by accident? It seems that this probability, known as a *p*-value could address the problem of the base rate, as discussed above, as well as the problem of the sample size, for example if the base rate is 0.5 achieving a Brier score of 0.1 by accident for a fixed sample size would be harder than if the base rate were 0.2. (The fact that this assertion is incorrect is demonstrated in Section 2.2).

There are numerous ways of calculating *p*-values, all of which are based upon determining the distribution of possible scores under the null hypothesis that there is no relationship between the forecasts and the observations. In other words, what scores would be achieved given lots of sets of completely useless forecasts? For some verification scores, the distribution of these values is known exactly, and the probability of scoring at least as well as the actual forecasts can be calculated from the left tail-area (or right tail-area for positively oriented scores) of this distribution. Examples include the hit rate, which follows a hyper-geometric distribution (Agresti, 2002, 2007), the correct number, which follows a binomial distribution (Mason, 2003), and the trapezoidal area beneath the relative operating characteristics (ROC) curve, which can be transformed to follow a *U*-distribution (Mason and Graham, 2002; Mason, 2003). For other scores this distribution can be approximated; for example, Pearson's product moment correlation coefficient can be transformed to a Student's *t*-statistic (Sheskin, 2007), and the ROC area can be transformed to follow an approximate Gaussian distribution, which is easier than using the *U*-distribution if the sample size is large (Mason and Graham, 2002).

Where such distributions can be used there is usually considerable computational advantage and will give exact *p*-values in the cases where the distribution is not approximated. However, for many scores, the distributions cannot be modelled exactly or approximated by theoretical distributions and even for those that can, assumptions can be restrictive. The most common assumption is that each forecast–observation pair has to be independent of every other pair. This assumption is often violated if the verification score is calculated using forecasts for different locations and when there is spatial correlation among the locations, or if both the forecasts and observations are not independent temporally. For some distributions, such as the Student's *t*-statistic for the correlation, it is assumed that the forecasts and observations are both normally distributed. Violations of this assumption can be quite severe. As an example, consider a set of forecasts and observations of January–March 1971–2000 seasonal rainfall totals for Brisbane, Australia (Figure 1). The observed rainfall has a skewness of 1.9, whereas the forecasts are only slightly positively skewed. The forecasts were obtained from a simple linear regression with the preceding December value of the NIÑO 3.4 index, and have a correlation with the observed rainfall of about 0.35. Using the Student's *t*-approximation, the probability of achieving a correlation this strong by accident is about 2.8%, which is about half the probability obtained using re-sampling procedures that do address the distributional assumptions (described below). Spearman's correlation, which adjusts for distributional violations, is about 0.16 and is much weaker than Pearson's.

An alternative to using a theoretical distribution is to use re-sampling procedures in order to generate an empirical distribution for the values of the verification score. Since it is the probability of exceeding the observed score by accident that is needed, the procedure is to generate a large number of scores against forecast–observation

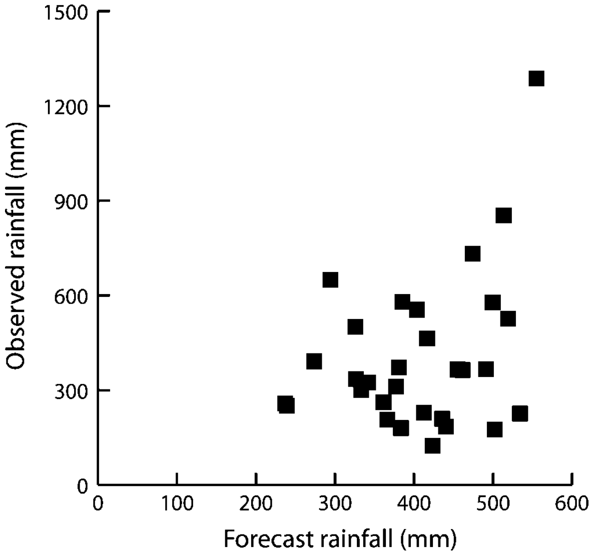

Figure 1. Forecasts and observations of January–March 1971–2000 seasonal rainfall totals for Brisbane, Australia (27°27′04″S, 153°01′55″E).

pairs in which there is no prior reason for expecting the forecasts to provide any useful indication of the observations. The simplest way to generate these scores is to reorder the observations randomly (or the forecasts – it makes no difference) so that the original forecasts are no longer paired with the original observations (except by chance). There is now no reason to expect the forecasts to score well, but some random arrangements of the forecasts and observations will make the forecasts look good by accident. The *p*-value can be estimated by calculating the scores for a large number of these random pairings (typically a few thousand), and calculating the proportion of the scores that exceed the score for the correct forecast–observation pairing. Because the same data are used in each permutation, the respective distributions of the forecasts and of the observations remain unchanged, and so the calculated scores will account for the effects of skewness, for example. An example is shown in Figure 2, where 10 000 correlations have been calculated from permutations of the observed rainfall for Brisbane. Note that the distribution is symmetric about zero, which is to be expected given that random forecasts are just as likely to be accidentally 'good' as accidentally 'bad'. The *p*-value for the actual forecasts is calculated as the right tail-area, as indicated.

Unfortunately, the permutation procedure just described is not assumption free. Specifically, it is assumed that each forecast–observation pair is independent of every other pair. This same assumption was made when using a theoretical distribution. One solution is to use 'block re-sampling' (Elmore *et al*., 2006; Wilks, 2006a). If there is serial correlation in the forecasts and/or observations (techniques for spatially correlated data are discussed later), a random set of consecutive data could be selected, and the temporal autocorrelation would be retained within each block. The discontinuities between blocks may be small enough so that the degree of dependence in the re-sample is approximately the same as in the original data. A more flexible approach is to generate a synthetic set of observations or forecasts using random numbers with the same distributional properties and dependencies as the original data. For example, if the observations are serially correlated, a random autoregressive series could be used to generate a random set of observations. More generally, the forecasts or observations can be replaced with a set of random numbers with the same distributional and dependency properties as the original data.

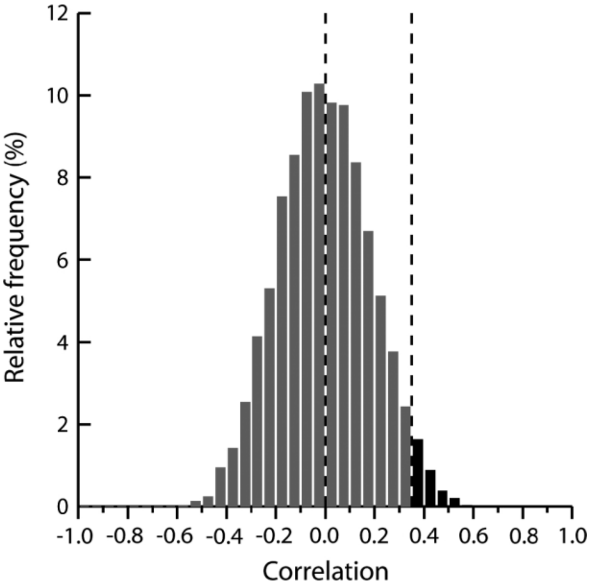

Figure 2. Relative frequencies of 10 000 Pearson's correlations between forecasts and observations of January–March 1971–2000 seasonal rainfall totals for Brisbane after permuting the observations. A dashed line is drawn at 0.0, which is the expected value of the correlations. A second dashed line is drawn at 0.35, which is the correlation value for the correct forecast–observation pairings (from Figure 1), and correlations in the right tail that are larger than this value are shaded black.

For spatially correlated data, similar techniques to those used for temporal dependency can be applied. A random surface could be generated, for example, to provide a synthetic set of observations, and is likely to be preferable to some form of block sampling in which the map of observations is re-shuffled, just like in the classic 'fifteen puzzle', and which becomes complicated to apply in the context of irregularly spaced station data (Lahiri and Zhu, 2006). However, there are often additional considerations when working with spatial data. It is first helpful to distinguish various ways in which forecasts that have a spatial component may be verified. The simplest case is when there is one forecast for each location, and a score is desired for the entire map; the second case is when there are numerous forecasts for each location, and a score is desired for all maps and all locations together; the third case is similar to the second, but a score is desired for each location. These cases are discussed in turn.

For a single forecast at numerous locations, perhaps the primary consideration is the need to consider the skill obtained from predicting the climatological distribution of the forecast parameter (e.g. colder temperatures towards the poles and warmer temperatures in lower latitudes). If the re-sampling or synthetic data do not recognize the climatological distribution of the data, the verification scores will be unrealistically bad, and so the *p*-value for the original data will be artificially low (Hamill and Juras, 2006). For predictands measured on continuous scales (e.g. 2 m temperature), the spatial effects often can be removed by converting the data to anomalies by subtracting the respective mean values for each location (the so-called anomaly correlation performs exactly this operation (Wilks, 2006a)), or by dividing by the mean if the predictand has an absolute lower limit of zero (e.g. precipitation). For some predictands it may be necessary to account for spatial differences in variance – an issue that is not often considered. After removing, or accounting for the spatial features of the climatology of the data, it is then necessary to account for the remaining spatial correlation, otherwise the spatial degrees of freedom will be artificially high in the synthetic data. Near surface temperature anomalies, for example, tend to be spatially coherent, and so if a forecast at one location is accurate, the forecast at a nearby location is likely to be accurate also. If the synthetic data are not spatially correlated even after considering the climatology, it will be much harder to obtain a good score by accident, and so p-value will be artificially low. Some stochastic multi-site weather generator procedures (e.g. Wilks, 1998) explicitly consider all these issues, and so are effective ways of generating synthetic surfaces for the calculation of p-values.

If forecasts are to be verified by pooling sets of forecasts for different locations and calculating one overall score, the procedures for calculating p-values are essentially identical to those described in the previous paragraph, although it may be necessary to account for any temporal correlation from forecast to forecast for the same reasons as it is necessary to account for the spatial correlation (Zwiers, 1987). Again, stochastic multi-site weather generators can be used for the synthetic forecasts, although this procedure can be cumbersome (Wilks, 1997).

If verification scores are to be calculated for each location to obtain a map of forecast quality, some of the locations are likely to score well simply by accident. This problem of 'multiplicity' (Katz, 1988; Brown and Katz, 1991; Katz and Brown, 1991) is often addressed by calculating a p-value for the entire map. This so-called 'field significance' (Livezey and Chen, 1983) is an attempt to estimate the probability of obtaining a map with the same proportional coverage of locations with good forecasts, and is distinct from the p-value for a score calculated by pooling all forecasts as described in the previous paragraph. The standard procedure is to calculate p-values for each location by randomly reordering the forecasts or the observations for all locations simultaneously, and then for each permutation to go back and count the number of locations at which the p-value was less than a threshold amount. From these counts a probability mass function of frequencies of different numbers of locations with statistically significant scores is obtained (commonly $p/ < 0.05$). The field significance is estimated by counting the proportion of times that the number of locations with statistically significant verification scores equalled or exceeded the number for the correct forecast–observation pairings. As before, the procedure assumes temporally independent forecasts and observations. If this assumption is invalid, block sampling can be used, or synthetic sets of forecasts generated using the weather generator technology described earlier.

This procedure for calculating field significance has a few limitations, primarily relating to the fact that it ignores the strength of the results, and to some conservative effects resulting from the discreteness of the counting of the number of significant results. Effectively, the forecasts are considered to be either 'good' or 'bad' depending on whether the score is statistically significant, but no consideration is given to exactly how good' the forecasts are. Consequently, if the forecasts are very good over only a small part of the map, and are useless elsewhere, the forecasts as a whole will probably be rejected as statistically insignificant. Recently, simpler and more powerful procedures have been identified that are insensitive to the effects of spatial correlation and that can be applied when serial correlation is minimal. These tests are based on the smallest p-value in the field, and on the false detection rate (Ventura *et al.*, 2004; Wilks, 2006b). Note that these tests answer a question different from that of Livezey and Chen (1983). At the risk of over-simplifying, the latter addresses the question: are the forecasts good over a large proportion of the field? The alternative tests address the question: are the forecasts good anywhere in the field?

2.2. Problems with p-values

Up to now it has been left implicit that p-values *are* a useful means of converting an uninterpretable skill score to a value that helps address the question of whether the scores are good. Indeed, this assumption is widespread, and the calculation of p-values is common in the atmospheric sciences. However, in attempting to answer the question of whether the score is a good score, p-values are not necessarily very helpful (Nicholls, 2001; Jolliffe, 2004, 2007). What p-values tell us is how confident we can be that the forecasts have some skill, which is not the same as telling us how much skill the forecasts have. For example, given a very large sample of marginally good forecasts, the p-value will be very small, telling us that we can be very confident that the forecasts are marginally good. To illustrate, Finley's well-known tornado forecasts (Murphy, 1996) are frequently cited as examples of bad forecasts because against a strategy of perpetual forecasts of no tornadoes they perform very poorly. Modifying the numbers slightly for the sake of simplification, over a series of about 2800 forecasts about 50 tornadoes were observed, 100 were forecast, and 30 were forecast correctly. Against a strategy of random guessing, Finley's forecasts score well on most verification scores (Mason, 2003), and given the large sample size the probability of scoring 30 or more hits by accident is about 1.0×10^{-32} (i.e. virtually impossible). From this p-value alone it might be concluded that Finley's forecasts were miraculously good. However, if the size of the sample is reduced by a factor of 10, then even with a perfect hit rate the p-value still increases by a factor of about 10^{24}. The point of this example is that p-values are not very helpful in addressing the difficulty of comparing scores for differently sized forecast sets, as had been posited at the beginning of the previous Section.

Another problem with p-values arises when trying to calculate them for probabilistic scores. Imagine two sets of forecasts: forecaster A issues a probability of 10% for case 1, and 90% for case 2; forecaster B issues probabilities of 20 and 80%, respectively. An event occurs in case 2, but not in case 1. Forecaster A scores 0.01 on the Brier score, and forecaster B scores 0.04. Forecaster A issues better forecasts than B, but using a permutation procedure both forecasters have identical p-values. The problem with the permutation procedure

DOI: 10.1002/met

is that the sampling distribution of the Brier score (and other probabilistic scores) under the null hypothesis is a function of the forecast probabilities (Mason, 2004). As a result, a simple monotonic transformation of the forecast probabilities will adjust the Brier score and its sampling distribution without affecting the p-value. The effect is to ignore the sharpness of the forecasts in exactly the same way as the area beneath the ROC curve does (Mason and Graham, 2002). Effectively, therefore, when estimating p-values for Brier scores the reliability of the forecasts is ignored, and instead the p-value for the ROC area is returned. The situation is somewhat complicated in the case of multi-category verification scores that are non-local such as the ranked probability score (Epstein, 1969; Murphy, 1969, 1970, 1971), although the general principle that the p-values do not adequately consider differences in reliability still applies.

Given these problems with p-values, their calculation is perhaps to be discouraged; they tell us only how confident we can be that the forecasts do not have zero skill, and they do not adequately account for the reliability of probabilistic forecasts. Instead, it is more informative to consider the question of how confident we can be that the score gives a realistic estimate of the quality of subsequent forecasts. This question is addressed in the following Section.

3. Is the score misleading?

In asking whether the verification analysis provides an accurate indication of the quality of subsequent forecasts, there are two possible reasons why the results may be misleading: only a limited sample of forecast–observation pairs is available, and the forecasts may have been uncharacteristically good (or bad) over this period or the experiment may have been inherently biased.

3.1. Are the results accurate?

Because only a limited number of forecast–observation pairs is available, the following question arises: would similar results be obtained if another set of forecast–observation pairs from the same forecast system were available? For example, it would be desirable to have some kind of guarantee that when a new forecast procedure is implemented into operations it is not suddenly going to start performing worse than anticipated. As discussed in Section 2.1 on the calculation of p-values, it is much easier to get a good score by accident if only a few forecasts are available than if there is access to a large sample. This problem of sample size is a much bigger issue for seasonal climate forecasts than for weather forecasts, since the number of weather forecasts made *per* year is at least 30 times larger than the number of seasonal forecasts. However, even if a large number of forecast–observation pairs were available, there may still be an interest in sampling uncertainty because of a concern about how much forecast quality can be expected to vary 'naturally'. For example, there is some debate as to whether the predictability of ENSO varies inter-decadally (Ji *et al*., 1996; Kirtman and Schopf, 1998), but how can it be decided whether this variability is real or simply a reflection of sampling error?

Knowing the sampling uncertainty in verification statistics not only provides an indication as to whether the results may be misleading but can also help to address the question of whether the forecasts are good: We can only be confident that the forecasts are good if they score well *and* if the uncertainty in the score is small. It is quite easy for forecasts to score well for a short period purely by accident, and so a good score is much less likely to be adequate evidence of good forecasts if the score was calculated using only a small sample compared to using a large sample. In the previous Section the p-value was suggested as a way of assessing whether the sampling uncertainty is large enough for us to suspect that the forecasts may have scored well by accident, but some problems were raised in the interpretation of the p-values.

A more satisfactory approach is to attempt to answer the question: what range of scores would be obtained given different sets of forecasts from the same forecast system? Typically all the forecast–observation pairs would be used to obtain as accurate an estimate of the verification score as possible, and thus taking subsets to recalculate the score will overestimate the sampling uncertainty. Instead, a bootstrapping procedure is often used. Bootstrapping is a re-sampling procedure that is distinct from the permutation procedure used to calculate the p-values. In the permutation procedure, the objective is to generate a new set of forecast–observation pairs in which the observations are unrelated to the forecasts except by accident. In the bootstrap procedure, however, the objective is to generate a new set of forecast–observation pairs in which the quality of the forecasts is consistent with the quality one would expect from the forecast system – i.e. the new forecasts should be as good as for the original set, subject only to sampling differences (Wilks, 2006a). Although there are a number of bootstrapping designs, the most commonly used procedure is to sample forecast–observation pairs randomly, keeping the forecast and observation together (unlike with the permutation procedure). The new sample, which for the method described here should be of the same size as the original sample, differs from the original because it is sampled with replacement. It is thus likely to include some forecast–observation pairs more than once, and is likely to omit some pairings altogether. If there are only a few forecasts that contribute strongly to the score being good, for example, then bootstrap samples that omit one or more of these forecasts will score very poorly, but samples that include these forecasts multiple times will score very well, and so the sampling uncertainty in the score will be very large.

In the case of the observations being defined as categories, some care should be taken to ensure that the definition of the categories remains consistent between the original data and the bootstrap sample. In some cases

DOI: 10.1002/met

the observations in the bootstrap sample may need to be re-classified. For example, if the observations are defined as a value exceeding the median, observations that were above median in the original sample may no longer be above median in the bootstrap sample unless the median was defined using an independent climatology. (The same is true of observations below the median.) In this event the observations would need to be re-classified. However, if the categories are defined by independent criteria such as temperatures exceeding 0 °C or precipitation exceeding a trace amount, the observations should probably not be re-classified. The simple rule is that if the categories are defined in a way to ensure that there are at least an approximately set number of observations in each category, then the observations need to be re-classified in each bootstrap sample to ensure that the relative frequencies remain approximately constant. Similar considerations may need to be taken for the forecasts as well. Forecast probabilities may need to be re-estimated to account for changes in model climatology; the forecasts may have to be re-calibrated because of differences in systematic errors from bootstrap sample to bootstrap sample. These problems can be simplified considerably if independent data are available for model training and calibration.

As with the permutation procedure, a large number of bootstrap samples (typically a few thousand) is generated and the sampling distribution of the score is built up. This distribution is often summarized by quoting the scores towards each tail of the distribution, defining confidence limits for the score. Typically the 0.025 and the 0.975 quantiles are quoted, defining the 95% confidence limits for the score. An example of bootstrapped correlations is shown in Figure 3, using the Brisbane data from Figure 1. The lower and upper confidence limits are −0.13 and 0.63 respectively, and thus contain zero, implying that the uncertainty in the quality of the forecasts is sufficiently large for it to be quite possible that there is no skill at all. Note, however, that the interpretation of confidence limits is a little complicated (Jolliffe, 2007). It is tempting to assume that there is a 95% probability that the 'true' score lies within the confidence interval, but all that can be said is that 95% of confidence intervals will contain the 'true' score.

Bootstrapping is not the only way of obtaining confidence intervals. Just as the distribution of scores under the null hypothesis (that the forecasts are useless) could be represented by an exact distribution, an approximate distribution, or empirically (Section 2.1), the distribution of the sample score too can be represented using theoretical or empirical distributions. Jolliffe (2007) provides a detailed description of these and other alternatives together with examples.

Regardless of the exact method used, such procedures for assessing the sampling uncertainty in the results of verification analyses are useful not only for indicating whether the results may be misleading but can also be used to assess whether differences in results (perhaps between a model with an established and one with a new prototypical convective parameterization scheme, or for measuring the changes in forecast quality over time, for example) are meaningful. Given sampling distributions for the scores of two forecast systems, the probability of obtaining a score from one of these systems that is better than another could then be calculated; if this probability is not substantially different from 50% the quality of the forecasts from the two systems is indistinguishable. However, a more powerful procedure is to obtain a distribution for the difference in the scores of the two systems, and then to calculate a p-value, or preferably a confidence interval, for the difference (Jolliffe, 2007). Remembering the problems in interpreting the p-value that were discussed in Section 2.2, the p-value will indicate how confident we can be that the scores differ by any amount; given a large sample, the p-value may indicate that it is reasonable to reject the null hypothesis that the quality of the forecasts from the two systems is the same, but the difference in the scores may in fact be very small.

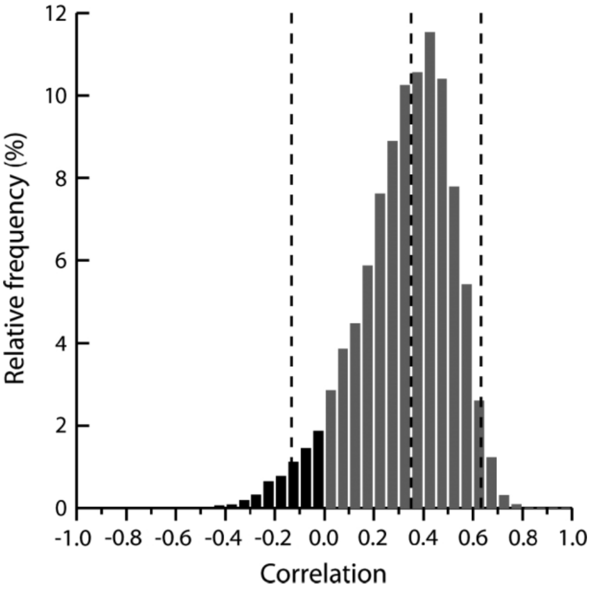

Figure 3. Relative frequencies of 10 000 bootstrapped Pearson's correlations between forecasts and observations of January–March 1971–2000 seasonal rainfall totals for Brisbane. A dashed line is drawn at 0.35, which is the correlation value for the correct forecast–observation pairings (from Figure 1), and at −1.14 and 0.64, which are the 95% confidence limits. The interval contains zero, and correlations in the left tail less than zero are shaded black.

3.2. Are the results biased?

It is well recognized that forecasts should be evaluated using a set of data that is independent of the one used to train the forecast system (Davis, 1976; Chelton, 1983), and that skill levels can be considerably overestimated when search procedures are used to select the predictors (Rencher and Pun, 1980; Wilkinson and Dallal, 1981). Ideally, a forecast system should be evaluated by considering those forecasts issued in real-time, but this requirement can be impractical if forecasts are issued

DOI: 10.1002/met

infrequently (such as with seasonal forecasts, for example). Instead, various procedures have been designed for generating a series of out-of-sample hindcasts; all such procedures involve using a subset of historical observational data to train the forecast model and in generating a set of hindcasts for the period for which the data have been withheld.

The most widely used out-of-sample testing procedure is cross validation (Stone, 1974; Geisser, 1975). In the atmospheric sciences, cross validation typically involves leaving out each forecast–observation pair in turn (possibly with a few additional pairs if the data are autocorrelated) reconstructing the forecast model and predicting the omitted observation so that if there are n forecasts, n cross-validated forecasts are generated (Michaelsen, 1987; Elsner and Schmertmann, 1994). Unfortunately, while this procedure slightly underestimates the quality of the forecasts when the correct predictors are used (Barnston and van den Dool, 1993), a less widely recognized problem is that under certain circumstances it can substantially overestimate the quality of the forecasts. Specifically, if there is a candidate pool of predictors, and one or only a few forecasts are withheld at each step the skill will be overestimated even if the predictors are reselected at each step in an attempt to avoid leakage (Shao, 1993; Rivals and Personnaz, 1999). If the number of candidate predictors is large, it may be necessary to withhold up to 60% of the forecasts (Xu and Liang, 2001).

An alternative to cross validation is to use a retroactive forecast procedure in which the model initially is trained to use only the first few forecast–observation pairs (e.g. Mason and Mimmack, 2002). The first omitted observation is then forecast (or the first few observations), and the model is updated using an expanded training period. Although this procedure gives a realistic indication of the quality of the forecasts if they had been issued operationally (as long as there has been no application of posterior knowledge about which predictors to use), the quality of subsequent forecasts may be underestimated because the forecast model should improve as the training periods lengthen (Unger personal communication; and see Wilks, 2006a).

4. Is the verification score a good score to use?

In the previous Sections the question of how to interpret a specific verification score has been considered, with a focus on attempting to answer the questions of whether the forecasts are good, and whether the score has provided a misleading indication of the quality of the forecasts. A more elementary question is whether the verification score itself is a good score to use. Unfortunately, this question is not asked sufficiently frequently, and a subset of scores receives wide use apparently simply because they are widely used. Pearson's correlation coefficient, for example, is almost invariably used regardless of distributional assumptions when Spearman's correlation may be more appropriate, and almost as powerful, while Kendall's τ barely receives any notice despite it having some intuitive properties (Sheskin, 2007). Given the wide selection of scores available, it is helpful to have some criteria for selecting an appropriate score for the immediate context. The following properties are discussed, namely, propriety, equitability, effectiveness, locality, and understandability. For much more detailed discussions of propriety and equitability, see the article by Jolliffe (2008) in this issue.

4.1. Propriety

One of the most important properties of a verification score for probabilistic forecasts is whether it encourages the forecaster to hedge. If the forecaster is concerned to achieve the best possible score, then it may be in his/her interests to issue a forecast that is inconsistent with his/her beliefs about what is likely to happen. A strictly proper score is one for which the forecaster uniquely optimizes the expected score by forecasting his/her true beliefs (Bröcker and Smith, 2007; Jolliffe, 2008). In most situations a strictly proper score would be desirable so that the forecaster does not issue a misleading forecast. Scores that are proper also have an advantage of making comparisons between forecasts easier. Unfortunately, many of the skill scores used in the atmospheric sciences are not strictly proper (Murphy, 1973; Gneiting and Raftery, 2007; Jolliffe, 2008).

An additionally interesting problem is that when a proper score is used in a reward function (e.g. if a forecaster receives a bonus if his/her Brier score is less than a threshold, or if the forecaster with the smallest Brier score receives a bonus) then the score is rendered improper (Roulston, 2007). Consider the case of the forecaster who is offered a bonus if his/her weather forecasts over a 10-day period achieve a Brier score of less than 0.1. After 9 days the forecaster's score is 0.11, and thinks that the probability of an event on the 10th day is 20%. The Brier score is strictly proper, thus the forecaster's expected score is minimized if (s)he states that the probability of the event is 20%, i.e. the expected score is 0.115 if the forecast is consistent with the forecaster's beliefs, and is 0.119 if the forecaster hedges by issuing a probability of 0%. However, if the forecaster does issue a probability of 20%, the bonus will not be awarded whether the event occurs (score is 0.163) or does not occur (score is 0.103). Conversely, if the forecaster issues a probability of 0% for the event (against his/her true belief) then while (s)he will not get the bonus if the event does occur (score is 0.199) (s)he will get the bonus in the more likely event that it does not (score is 0.099).

4.2. Equitability

Equitable verification scores score all naïve forecast strategies equally (Gandin and Murphy, 1992). Equitability is primarily of interest for scores for so-called deterministic forecasts (i.e. forecasts of specific values without an indication of the uncertainty in the forecasts). The linear error in probability space (LEPS) (Ward and Folland, 1991; Potts *et al*., 1996) and Gerrity scores (Gerrity, 1992), for example, score random guesses and perpetual forecasts of any one outcome equally. The same is true of correlation if the correlation is defined as zero for forecasts with no variance, but is not true of the mean squared error. For probabilistic forecasts there is a variety of forecast strategies that can be viewed as equally naïve (e.g. random forecasts and perpetual forecasts of constant probabilities, including those of climatological probabilities), but it is not possible for a probabilistic verification score to be equitable and proper (Jolliffe and Stephenson, 2008). These strategies for probabilistic scores are recognized as being equally naïve because they all have zero resolution (the expected outcome is the same regardless of the forecast), but they do not have equal reliability (only the climatological forecasts will have perfect reliability), and since reliability is an important attribute of probabilistic forecasts, it is appropriate that these different strategies would receive different scores. Equitability is therefore not considered an important property for probabilistic forecasts.

4.3. Effectiveness

An effective score is one which monotonically improves as the distance (however it is measured) between the forecast and the observation decreases (Friedman, 1983; Nau, 1985). The most often quoted example of a score that is ineffective is the original version of the LEPS score, which for large numbers of categories could score a forecast with an extremely large error less severely than one with only a large error (Potts *et al*., 1996). For probabilistic forecasts, because of the way effectiveness is defined, it is closely related to the property of locality, which is discussed in the following Section.

4.4. Locality

Locality is a property that applies to verification scores for probabilistic forecasts. A skill score is local if it depends only on the probability assigned to the outcome (Bröcker and Smith, 2007, Mason *et al*., 2008, Mason and Stephenson, 2008). Examples of local skill scores include the quadratic score, which measures the squared probability error only for the category that verifies, and the ignorance score (Roulston and Smith, 2002). It has become widely accepted that locality is not a desirable property of verification scores for two reasons: The idea of crediting forecasts that issue high probabilities to outcomes that are close to the verification (i.e. for 'near-misses') seems intuitively appealing, and non-local scores can be less sensitive to the categorization of the observed values than local scores – the more the categories that are used the lower the score tends to be (Daan, 1985). If a local score is used, the score typically drops with an increase in the number of categories because the probability assigned to the verifying category is divided between neighbouring categories. Although this latter argument may be appealing in some contexts (e.g. for comparing forecasts that are presented with differing degrees of precision), forecast systems that provide probabilities for large numbers of categories are attempting to communicate more information than systems with only a few, and unless this increase in precision can be matched by an increase in sharpness the penalty may be warranted.

The intuitive argument for the consideration of 'distance' is perhaps more appealing than the argument about the sensitivity of the local scores. However, the consideration of distance does not always produce intuitive results. For example, the ranked probability skill score (Epstein, 1969; Murphy, 1969, 1970, 1971) was explicitly designed to account for the ranking of the categories in a multi-category system and to credit those forecasts that have high probabilities close to the verifying category. This consideration of distance means that a forecast with a higher probability assigned to the verifying category will not necessarily achieve a better score than a forecast with a lower probability. Imagine one forecast with probabilities of 52, 33, and 15% to the categories one to three, respectively, and a second forecast with probabilities of 50, 45, and 5%. If the first category verifies, the first forecast with the highest probability on the verifying category scores worse than the second forecast (0.2529 compared to 0.2525). Perhaps more fundamentally, probabilistic forecasts provide indications of the likelihoods of different outcomes, and it seems consistent with the interpretation of these probabilities to verify the forecast only on the basis of the probability assigned to the verification. The desirable properties of probabilistic forecasts are that these probabilities be high, subject to reliability. Local scores will measure these two properties of sharpness and reliability, whereas it is unclear exactly what non-local scores will measure (Mason *et al*., 2008). However, as with the other properties of verification scores, it is important to consider what properties are relevant for the specific context since locality may not always be a desirable property.

4.5. Understandability

A neglected property of verification scores is their understandability. While, for example, the Gerrity and LEPS scores have a number of desirable and elegant mathematical properties (Section 4.2), they are not easily understandable by non-specialists. Similarly, in Section 4.4, the ranked probability skill score was criticized for failing to represent a simple measure of the reliability and sharpness of probability forecasts. Conversely, for probabilistic forecasts, the simple linear score (Wilson *et al*., 1999) may not be proper, but it is intuitively appealing, and so may be an appropriate score for communicating the quality of forecasts to non-specialists. In general, it is important to consider what questions the practitioner wants answered when performing a verification analysis, and to understand the strengths and weaknesses of each verification score. Is the score sensitive to distributional assumptions? Does it matter that biases in the mean and variance are ignored by correlation measures? To whom is the score to be communicated, and what is it meant to convey? Is the score to be used to compare forecast systems? If so, what properties of the forecasts are most important? For example, how important is reliability? More generally, Jolliffe and Stephenson (2003) ask the key question: is the score informative?

The question about the desirability of different properties of the forecasts deserves further attention, since too often the generic question is asked: which is the better forecast system? Given the multi-faceted nature of forecast quality (Murphy, 1991) comparing forecasts using a single score is likely to be unfruitful because it is unlikely to indicate in what respect a set of forecasts is better or worse than the other. It needs to be realized more widely that quality may differ between two forecast systems without one being better in every respect than the other. Even if there is a genuine difference, which specific attribute of the forecasts has improved?

5. Summary

This article has presented a review of some procedures for interpreting the results of a verification analysis of forecasts. The focus has been on interpreting verification scores, with the specific objective of trying to answer the questions: Are the forecasts good, and can we be confident that the estimate of forecast quality is not misleading? A widely used procedure for deciding whether forecasts are good is to calculate the probability that useless forecasts could have scored at least as well simply by accident. These so-called p-values are estimated from the left or right tail of a distribution for the values of the verification score under the null hypothesis that there is no relationship between the forecasts and the observations. Four approaches were described for defining the distribution of the scores, namely, using an exact theoretical distribution, an approximate theoretical distribution, or an empirical distribution using either permuted or synthetic data. Regardless of the type of distribution used, however, p-values are not very informative since they leave the question of whether the forecasts have ‘good’ skill unanswered. Further problems with p-values apply when calculating the statistical significance of probabilistic verification scores. Specifically, standard permutation techniques do not address the reliability component of forecast quality.

Instead of using p-values, confidence intervals for verification scores are to be preferred. Confidence intervals may provide not only an indication of how misleading the verification results might be but also some insights to the question of how good the forecasts are. However, they will not point to problems in experimental design that may have introduced some biases into skill estimates. While cross validation is widely used in an attempt to minimize biases in skill estimates, the commonly used leave-one-out procedure can be ineffective.

A more general problem relates to the selection of the verification score *per se* rather than the interpretation of the values of the score. Too often, scores seem to be selected without due consideration for the specific information that is actually desired from a verification analysis. Scientists should consider questions of the desirable attributes of verification scores in the context of the specific objectives of the analysis. A more discriminating selection of verification procedures than has been adopted to date in the atmospheric sciences is encouraged, and should facilitate a more insightful assessment of forecast quality.

Acknowledgements

This note was funded by Cooperative Agreement AN07 GP0213 from the National Oceanic and Atmospheric Administration (NOAA). The views expressed herein are those of the author, and do not necessarily reflect the views of NOAA or any of its sub-agencies. Helpful discussions with DS Wilks are gratefully acknowledged, as are the comments of anonymous referees. The rainfall data for Brisbane were supplied by A Watkins, Bureau of Meteorology.

References

Agresti A. 2002. *Categorical Data Analysis*, 2nd edn. Wiley-Interscience: Hoboken; 734.

Agresti A. 2007. *An Introduction to Categorical Data Analysis*, 2nd edn. Wiley-Interscience: Hoboken; 372.

Barnston AG, van den Dool HM. 1993. A degeneracy in cross-validated skill in regression-based forecasts. *Journal of Climate* **6**: 963–977.

Bröcker J, Smith LA. 2007. Scoring probabilistic forecasts: the importance of being proper. *Weather and Forecasting* **22**: 382–388.

Brown BG, Katz RW. 1991. Use of statistical methods in the search for teleconnections. *Teleconnections Linking Worldwide Climate Anomalies: Scientific Basis and Societal Impact*. Cambridge University Press: Cambridge; 371–400.

Chelton DB. 1983. Effects of sampling errors in statistical estimation. *Deep-Sea Research* **30**: 1083–1103.

Daan H. 1985. Sensitivity of verification scores to the classification of the predictand. *Monthly Weather Review* **113**: 1384–1392.

Davis R. 1976. Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *Journal of Physical Oceanography* **6**: 249–266.

Elmore KL, Baldwin ME, Schultz DM. 2006. Field significance revisited: spatial bias errors in forecasts as applied to the Eta model. *Monthly Weather Review* **134**: 519–534.

Elsner JB, Schmertmann CP. 1994. Assessing forecast skill through cross validation. *Weather and Forecasting* **9**: 619–624.

Epstein ES. 1969. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology* **8**: 985–987.

Friedman D. 1983. Effective scoring rules for probabilistic forecasts. *Management Science* **29**: 447–454.

Gandin LS, Murphy AH. 1992. Equitable skill scores for categorical forecasts. *Monthly Weather Review* **120**: 361–370.

Geisser S. 1975. The predictive sample reuse method with applications. *Journal of the American Statistical Association* **70**: 320–328.

Gerrity JP. 1992. A note on Gandin and Murphy’s equitable skill score. *Monthly Weather Review* **120**: 2707–2712.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**: 359–378.

Hamill TM, Juras J. 2006. Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society* **132**: 2905–2923.

DOI: 10.1002/met

Ji M, Leetmaa A, Kousky VE. 1996. Coupled model predictions of ENSO during the 1980s and the 1990s at the National Centers for Environmental Prediction. *Journal of Climate* **9**: 3105–3120.

Jolliffe IT. 2004. P stands for .... *Weather* **59**: 77–79.

Jolliffe IT. 2007. Uncertainty and inference for verification measures. *Weather and Forecasting* **22**: 637–650.

Jolliffe IT. 2008. The impenetrable hedge: a note on propriety, equitability and consistency. *Meteorological Applications* **15**: 25–29.

Jolliffe IT, Stephenson DB. 2003. Introduction. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley: Chichester; 1–12.

Jolliffe IT, Stephenson DB. 2008. Proper scores for probability forecasts can never be equitable. *Monthly Weather Review* in press.

Katz RW. 1988. Use of cross correlations in the search for teleconnections. *Journal of Climatology* **8**: 241–253.

Katz RW, Brown BG. 1991. The problem of multiplicity in research on teleconnections. *International Journal of Climatology* **11**: 505–513.

Kirtman BP, Schopf PS. 1998. Decadal variability in ENSO predictability and prediction. *Journal of Climate* **11**: 2804–2822.

Lahiri SN, Zhu J. 2006. Resampling methods for spatial region models under a class of stochastic designs. *Annals of Statistics* **34**: 1774–1813.

Livezey RE, Chen WY. 1983. Statistical field significance and its determination by Monte Carlo techniques. *Monthly Weather Review* **111**: 46–59.

Mason IT. 2003. Binary events. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley: Chichester; 37–76.

Mason SJ. 2004. On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. *Monthly Weather Review* **132**: 1891–1895.

Mason SJ, Graham NE. 2002. Areas beneath the relative operating characteristics (ROC) and levels (ROL) curves: statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society* **128**: 2145–2166.

Mason SJ, Mimmack GM. 2002. Comparison of some statistical methods of probabilistic forecasting of ENSO. *Journal of Climate* **15**: 8–29.

Mason SJ, Stephenson DB. 2008. How can we know whether the forecasts are any good? *Seasonal Climate Variability: Forecasting and Managing Risk*. Kluwer Academic Publishers: Dordrecht, in press.

Mason SJ, Smith LA, Clarke L, Bröcker J. 2008. Locality and the ranked probability skill score. *Monthly Weather Review* Submitted to.

Michaelsen J. 1987. Cross-validation in statistical climate forecast models. *Journal of Climate and Applied Meteorology* **26**: 1589–1600.

Murphy AH. 1969. On the "ranked probability score". *Journal of Applied Meteorology* **8**: 988–989.

Murphy AH. 1970. The ranked probability score and the probability score: a comparison. *Monthly Weather Review* **98**: 917–924.

Murphy AH. 1971. A note on the ranked probability score. *Journal of Applied Meteorology* **10**: 155–156.

Murphy AH. 1973. A new vector partition of the probability score. *Journal of Applied Meteorology* **12**: 595–600.

Murphy AH. 1991. Forecast verification: its complexity and dimensionality. *Monthly Weather Review* **119**: 1590–1601.

Murphy AH. 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting* **8**: 281–293.

Murphy AH. 1996. The Finley affair: a signal event in the history of forecast verification. *Weather and Forecasting* **11**: 3–20.

Murphy AH, Winkler RL. 1987. A general framework for forecast verification. *Monthly Weather Review* **115**: 1330–1338.

Nau RF. 1985. Should scoring rules be effective? *Management Science* **31**: 527–535.

Nicholls N. 2001. The insignificance of significance testing. *Bulletin of the American Meteorological Society* **82**: 981–986.

Potts JM, Folland CK, Jolliffe IT, Sexton D. 1996. Revised "LEPS" scores for assessing climate model simulations and long-range forecasts. *Journal of Climate* **9**: 34–53.

Rencher A, Pun FC. 1980. Inflation of R^2 in best subset regression. *Technometrics* **22**: 49–53.

Rivals I, Personnaz L. 1999. On cross validation for model selection. *Neural Computation* **11**: 863–870.

Roulston MS. 2007. Performance targets and the Brier score. *Meteorological Applications* **14**: 185–194.

Roulston MS, Smith LA. 2002. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review* **130**: 1653–1660.

Shao J. 1993. Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**: 486–494.

Sheskin DJ. 2007. *Handbook of Parametric and Nonparametric Statistical Procedures*, 4th edn. Chapman and Hall/CRC: Boca Raton; 1776.

Stone M. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society* **36B**: 111–147.

Ventura V, Paciorek CJ, Risbey JS. 2004. Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. *Journal of Climate* **17**: 4343–4356.

Ward MN, Folland CK. 1991. Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea-surface temperatures. *International Journal of Climatology* **11**: 711–743.

Wilkinson L, Dallal GE. 1981. Tests of significance in forward selection regression with an F-to-enter stopping rule. *Technometrics* **23**: 377–380.

Wilks DS. 1997. Resampling hypothesis tests for autocorrelated fields. *Journal of Climate* **10**: 65–82.

Wilks DS. 1998. Multisite generalizations of a daily stochastic precipitation generation model. *Journal of Hydrology* **210**: 178–191.

Wilks DS. 2006a. *Statistical Methods in the Atmospheric Sciences*, 2nd edn. Academic Press: San Diego; 627.

Wilks DS. 2006b. On "field significance" and the false discovery rate. *Journal of Applied Meteorology and Climatology* **45**: 1181–1189.

Wilson LJ, Burrows WR, Lanzinger A. 1999. A strategy for verification of weather element forecasts from an ensemble prediction system. *Monthly Weather Review* **127**: 956–970.

Xu QS, Liang YZ. 2001. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems* **56**: 1–11.

Zwiers FW. 1987. Statistical considerations for climate experiments. Part II: multivariate tests. *Journal of Climate and Applied Meteorology* **26**: 477–487.

DOI: 10.1002/met