Numerical Weather Prediction

Explicitly accounting for observation error in categorical verification of forecasts

Forecasting Research Technical Report No. 456

Neill E. Bowler

email:nwp_publications@metoffice.gov.uk

© Crown copyright

Explicitly accounting for observation error in categorical verification of forecasts

Neill E. Bowler*
Met Office, Fitzroy Road, Exeter, EX1 3PB, UK.

March 8, 2005

Abstract

Given an accurate representation of errors in observations it is possible to remove the effect of those errors from categorical verification scores. The errors in the observations are treated as additive white noise which is statistically independent of the true value of the quantity being observed. This method can be applied to both probabilistic and deterministic verification where the verification method uses a categorical approach. In general this improves the *apparent* performance of a forecasting system, indicating that forecasting systems are generally performing better than they might first appear.

A major problem in the area of weather forecasting is caused by deficiencies in the observing network, either through the imperfect coverage of the observational network or through errors in the observations themselves. These deficiencies contribute to the initial condition uncertainties that have been the subject of great study [5]. Comparatively little has been written about the effect of observation errors on the verification of forecasts. Ciach & Krajewski [3] introduced an error separation technique for decomposing the mean square error of a forecast into terms involving the error in the observations and the error in the forecast. Anderson & later authors [1, 4, 6] used the rank histogram for verifying ensemble forecasts and showed how to remove the effect of observation errors from the verification of these forecasts. In this paper the categorical verification of forecasts is addressed, and it is shown how one may attempt to remove the effect of observation errors from such verifications.

1 Sources of errors

The source of observation errors changes dramatically depending on the type of observation. For example, radar-based estimates of surface rainfall rate are affected by the height of the radar beam above the ground and the size distribution of the raindrops, amongst other things. On the other hand the principle source of

*Electronic address: Neill.Bowler@metoffice.gov.uk

error in a rain-gauge measurement is that the measurement at that point is not necessarily representative of the rainfall rate at other points - so called representativity errors. Since Numerical Weather Prediction (NWP) models are formulated to forecast area-averaged quantities, representativity errors are ascribed as observational errors, rather than being seen as an inability of NWP models to represent sub-grid-scale variability.

In this study the imagined scenario is the verification of a forecast of the average temperature over London against a measurement of temperature at a single location. The reason for this choice is that the errors in surface observations are better characterised than the errors in remote-sensed observations. Furthermore, since the main error in a surface temperature measurement is due to the representation of small-scale processes the errors in the observations are unlikely to be correlated between one time and the next and for observations at different locations. Biases in the observations may introduce a correlation between observation errors at different times or locations, but provided these biases are known they may be removed. A serious problem in this and other methods may occur if biases are treated as random errors.

2 A desirable property of verification scores

An important property of verification scores is that the same score is achieved for a forecast, whatever the quality of the observational network. The importance of this property is seen when one considers a perfect deterministic forecast. Given a perfect observation this forecast would be seen to be perfect. However, any error in the observational network may obscure this fact, leading to the perfect forecast being believed to be in error!

One proposal for dealing with observational error would be to treat the observation as defining a probability density function for the truth and using probabilistic measures (such as the Brier score) to define the forecast quality [2]. However, such an approach would penalise a perfect forecast and lead to difficulties when comparing forecasts verified against different observations.

In general the problem of observational errors leads to the same problem as is faced in the verification of probabilistic forecasts - it becomes impossible to state whether a single forecast was "correct". However, by aggregating a number of forecasts, as in categorical verification, it becomes possible once again to recognise the quality of a perfect forecast.

3 Categorical verification

A categorical forecast is defined as one which forecasts whether a particular event will occur, for example will it rain in London tomorrow? Thus a whole wealth of forecast information is reduced to a forecast probability for an event to occur. In the case of a deterministic forecast a contingency table with four entries may be constructed as is shown in table 1. For a perfect forecast (and error-free observations) only hits

	Event Forecast	Event not forecast
Event observed	a (hit)	b (miss)
Event not observed	c (false alarm)	d (correct rejection)

Table 1: Contingency table for a categorical forecast. A perfect forecast would have zeroes in the off-diagonal elements.

	Event Forecast	Event not forecast
Event observed	$e = (1 - p_a)a + p_c c$	$g = (1 - p_b)b + p_d d$
Event not observed	$f = (1 - p_c)c + p_a a$	$h = (1 - p_d)d + p_b b$

Table 2: The contingency table for forecasts verified against observations, using the mis-categorisation approach.

and correct rejections would be seen. Innumerable skill scores may be derived from a categorical forecast [7] and their popularity lies in the simplicity of the verification method.

Many methods for the verification of probabilistic forecasts also use the categorical approach. The relative operating characteristic (ROC) uses the same contingency tables as for deterministic forecasts. It is common to calculate the Brier skill score by partitioning the forecast into a series of probability bins - this partitioning is necessary if the decomposition of this score into reliability and resolution is required. When forecasts are verified in this manner the same mis-categorisation and deconvolution approach detailed below may be used for these forecasts.

4 Accounting for observation errors

In the following a method for performing verification of categorical forecasts is sought which will produce the same results independent of observational errors. If an observation is in error, the event may have been *observed* to have occurred when it did not happen, or vice versa. This means that an event may have been categorised as a hit when it should have been categorised as a false alarm (see table 1). It is therefore natural to define a probability that a false alarm is mis-categorised as a hit (p_c, and similarly for the other three probabilities). If table 1 is taken as the contingency table when the forecast is verified against the truth, then table 2 gives the contingency table which would result when verification is performed against the (noisy) observations. The probabilities of mis-categorisation $p_{a,b,c,d}$ are related to the magnitude of the observational error, and e, f, g and h are the observed contingency table values.

From a corrupted contingency table, such as table 2, it is possible to re-construct the true contingency table values by solving the set of four equations for a, b, c and d, provided the probabilities of mis-categorisation are known. Table 3 gives the reconstructed values, and it is clear that tables 1, 2 and 3 all give the same values if

	Event Forecast	Event not forecast
Event observed	$\frac{(1-p_c)e-p_cf}{(1-p_a)(1-p_c)-p_ap_c}$	$\frac{(1-p_d)g-p_dh}{(1-p_b)(1-p_d)-p_bp_d}$
Event not observed	$\frac{(1-p_a)f-p_ae}{(1-p_a)(1-p_c)-p_ap_c}$	$\frac{(1-p_b)h-p_bg}{(1-p_b)(1-p_d)-p_bp_d}$

Table 3: Reconstructed contingency table in terms of the measured values and the probabilities of mis-categorisation.

the mis-categorisation probabilities are zero. So, provided that the probability of an event being mis-categorised is known, then the contingency table which describes the verification against truth can be re-constructed. However, the estimation of these probabilities is not a trivial matter.

5 Estimation of probabilities

In order to estimate the probability that an event is mis-categorised it is convenient to consider the observation of a continuous variable, such as the temperature. The reason for considering continuous variables is two-fold; observational errors are often defined in terms of continuous variables, and they permit the definition of a probability density function for the observational errors. The error in an observation is defined as the difference between the observed value and the true value (which is generally unknown). It is assumed that the error in one observation is independent of the error in another observation and that they are both independent of the value of the truth. If this is the case then it is reasonable to define some probability density of the observational errors, P_e.

The true value of the quantity which is being observed can, in general, be described as being drawn from some frequency distribution, P_t. If, as already assumed, the observation errors can be treated as additive white noise then the distribution of the observations, P_o may be written as the convolution of the distribution of the truth with the pdf of the observation errors

$$P_o(x|F) = \eta \int_{-\infty}^{\infty} P_t(y|F)P_e(x-y|F)dy \qquad (1)$$

where x is the observed value, y is the true value and the distributions have been conditioned on the event being forecast to occur (F) and η is a normalisation constant. No assumptions have been made about the shape of the distributions and these may be different for forecasts of the event to not occur. Once the distribution of the truth has been estimated via the deconvolution, the correct contingency table values may be estimated by calculating what fraction of this distribution lies above the event threshold.

	Event Forecast	Event not forecast
Event observed	2114	401
Event not observed	386	2099

Table 4: Contingency table for the example forecast when verified against the truth. Approximately 84% of the forecasts lie in the diagonal elements, which is close to the figure that would be expected.

6 An example

To demonstrate the deconvolution process an example is considered where all the data have been simulated and therefore the truth is known. Consider forecasting the average temperature over London, and comparing this forecast with the climatological temperature. An event is said to have occurred when the average temperature over London is above the climatological average value. The forecast is verified against a single observation located in central London, which provides an unbiased estimate of the average temperature for London.

In the imagined case 5000 forecasts of the average temperature over London are made, of which $\frac{1}{2}$ forecast the event to occur. The distribution of the truth is chosen to be taken from a Gaussian distribution with standard deviation of $2°$C. When the event is forecast to occur (to not occur) the mean of the distribution of the truth is chosen to be $2°$C above (below) the climatological average. The choice of $2°$C for the mean and standard deviation is arbitrary and has been chosen to illustrate this method. When forecasts are verified against the truth the contingency table is given by table 4. The expected value for the diagonal elements is $\frac{2500}{2}\left(1+erf\left(\frac{1}{\sqrt{2}}\right)\right) \simeq 2103$, and the values seen in the contingency table are close to this. erf is the normal error function, defined as $erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$.

The error in the observation is chosen to be taken from a Gaussian distribution with mean zero and standard deviation of $2°$C (this is a typical value for the estimated error in a surface temperature observation used in the global data assimilation system of the Met Office). When the forecasts are verified against the noisy observations the contingency table is given by table 5. Clearly the forecast appears to perform worse, simply due to the errors in the observations. The expected value for the diagonal elements of the contingency table in this case is $\frac{2500}{2}\left(1+erf\left(\frac{1}{2}\right)\right) \simeq 1901$, and the observed contingency table values are close to this.

The contingency table for verification against the observations (table 5) provides the basis for estimating the values seen in the true contingency table (table 4). In order to do this the mean and variance of the observed values are measured, from which the mean and variance of the true values will be estimated. For forecasts of the event to occur, the observations are estimated to be taken from a Gaussian distribution with mean $2.054°$C above the event threshold, and standard deviation of $2.898°$C. Given that the pdf of the observation error is known to have zero mean and standard deviation $2°$C, the distribution of the truth for forecasts of the event to

	Event Forecast	Event not forecast
Event observed	1938	599
Event not observed	562	1901

Table 5: Contingency table for the example forecast when verified against the observations. Approximately 77% of the occasions lie in the diagonal elements, which is close to the expected frequency of 76%.

	Event Forecast	Event not forecast
Event observed	2091	418
Event not observed	409	2082

Table 6: Contingency table for the example forecast when verified against the observations, and corrected for the effect of observation error. The statistics are close to those obtained for the verification against the truth (table 4).

occur, is estimated to have standard deviation $\sqrt{2.898^2 - 4} \simeq 2.097°$C. Therefore, the estimated number of true hits is given by

$$N_{\text{true hits}} = \frac{2500}{2}\left(1 + erf\left(\frac{2.054}{2.097\sqrt{2}}\right)\right) \simeq 2091 \qquad (2)$$

which is very close to the expected true value. A similar process is followed for occasions where the event is forecast to not occur, and the full reconstructed contingency table for this set of forecasts is given in table 6. The values in this table are much closer to the values in table 4 than those in table 5, indicating that the correction procedure has a positive effect.

7 Discussion

The sensitivity of the reconstruction to the specification of the observation errors will be common to all the approaches for removing the effect of observation errors. If the observation error is estimated to be too small, then some effect of the observation errors will remain in the verification statistics. However, if the error in the observations is estimated as being larger than reality then the method here may lead to an impossible deconvolution. In the same situation the Ciach & Krajewski [3] method may indicate a negative mean square error. The approach of Anderson [1] would suggest that the spread of the ensemble is too great. Any error separation is likely to be particularly difficult for forecasts where the observation error is much larger than the true forecast error since the statistics will be dominated by the observation error. A useful estimate of the observation error may be that used in the data assimilation scheme of the NWP model providing the forecasts, although this may only provide reliable estimates for "traditional" observations such as radiosondes or surface observations.

Calculating the probability of mis-categorisation allows an interesting insight into the effect of observation errors on (uncorrected) contingency table values. If the distribution of the truth is uni-modal and $a > c$, then the true values in category c will regularly be close to the event threshold. For true values in category a the peak of the frequency distribution will be away from the event threshold. Therefore, if the observation errors follow a symmetric uni-modal distribution centred around zero, then the probability of mis-categorisation for true values in category c will be greater than for those in category a ($p_c > p_a$). However, observation errors will ensure that the distribution of the observations is wider than the distribution of the true values, and hence the effect will be to increase the number of false alarms at the expense of decreasing the number of hits. This means that observation errors tend to equalise the values of a and c (and similarly for b and d). So, although in this case $p_c > p_a$ it is also true that $ap_a > cp_c$. This equalisation will often lead to an *apparent* decrease in skill, even though the forecast has not changed.

8 Conclusion

The use of categorical verification of forecasts is extremely widespread in weather prediction. However, little attempt has been made thus far to quantify the effect of errors in the observations on this verification. In this paper a deconvolution approach has been introduced which can be used to restore the contingency table scores which would have been achieved had the observations not been corrupted by noise, under the assumption that this noise is white, additive and independent of the true value of the quantity being observed. This approach can be applied to the verification of both deterministic and probabilistic forecasts.

It is crucial to the success of this technique that the errors in the observations are well known. If this is not the case, then spurious results can be obtained, and potentially even negative entries in the contingency table. This is the same issue that affects the error separation method of Ciach & Krajewski [3]. If an NWP forecast is being produced, then the data assimilation system will require estimates of the errors in the observations, and this may be a useful source for the verification system.

Acknowledgements

I would like to thank Beth Ebert for her stimulating discussions that helped the development of this paper, and Sarah John for her clear suggestions for revisions.

References

[1] J. L. Anderson. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9:1518–1530, 1996.

[2] G. Candille and O. Talagrand. Impact of observational errors on the validation of ensemble prediction systems. Poster at Ensembles Workshop, Exeter, abstract available at http://cccma-meetings.seos.uvic.ca/cgi-bin/ensemble/abstracts/view_particular_abstract.exe?a=10095, 2004.

[3] G. J. Ciach and W. F. Krajewski. On the estimation of radar rainfall error variance. *Advances in Water Resources*, 22 No. 6:585–595, 1999.

[4] T. M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129:550–560, 2001.

[5] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20 No. 2:130–148, 1963.

[6] O. Saetra, H. Hersbach, J-R. Bidlot, and D. S. Richardson. Effects of observation errors on the statistics for ensemble spread and reliability. *Monthly Weather Review*, 132 No. 6:1487–1501, 2004.

[7] D. B. Stephenson. Use of the "odds ratio" for diagnosing forecast skill. *Weather and Forecasting*, 17 No. 2:221–232, 2000.