Numerical Weather Prediction

Accounting for the effect of observation errors on verification of MOGREPS

NWP Technical Report No. 506

Neill E. Bowler

email:nwp_publicationsmetoffice.gov.uk

Accounting for the effect of observation errors on verification of MOGREPS

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

Abstract:
A number of papers have looked at the effect of observation errors on verification statistics. In this paper, those methods are brought together in order to assess the importance of observation errors on verification statistics for the Met Office short-range ensemble prediction system.
The results indicate that the effect of observation errors is substantial - reducing the apparent skill of the forecast system by around 1-day in forecast lead time. The effect of observation errors are typically largest at short lead-times when forecast errors are smallest.

KEY WORDS ensemble forecasting observation verification error

1 Introduction

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. Comparatively little has been written about the effect of observation errors on the verification of forecasts. Recently a few papers have addressed this issue. These methods are briefly reviewed here, and then applied to forecasts from the Met Office Global and Regional Ensemble Prediction System - MOGREPS, (Bowler *et al.*, 2007).

Ciach and Krajewski (1999) 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 (1996) & later authors (Hamill, 2001; Saetra *et al.*, 2004) used the rank histogram for verifying ensemble forecasts and showed how to remove the effect of observation errors from the verification of these forecasts. Bowler (2006) developed a method for accounting for the effect of observation errors in categorical verification of forecasts. Recently, Candille & Talagrand (personal commun.) have introduced a method which treats the observation defining a probability distribution, and using methods for the verification of probability forecasts to assess the impact of observation errors. That method will not be used in this study.

*Correspondence to: N. E. Bowler, Met Office, FitzRoy Road, Exeter, EX1 3PB, UK. E-mail: Neill.Bowler@metoffice.gov.uk

In this paper a set of forecasts will be verified using a number of verification methods. In section 2 the root-mean-square error is considered. The rank-histogram is used in section 3, and categorical verification of forecasts is examined in section 4. In section 5 verification using analyses is considered and how this can lead to some difficulties in the interpretation of results. Methods for estimating the observation error are discussed in section 6. Final discussions are presented in section 7.

The forecasts verified here are from the global component of the Met Office short-range ensemble prediction system MOGREPS (Bowler *et al.*, 2007). This system has 23 perturbed forecasts, plus a control forecast, and initial condition perturbations are derived from the Ensemble Transform Kalman Filter (ETKF). The forecasts (except in section 5) are verified against radio-sonde observations across the globe. The observation errors are taken from the standard estimates used by the data assimilation system. In this paper we choose to focus on forecasts of wind speed at 850 hPa, for which radiosonde observations are expected to have an error of 1.6 m/s. This is the standard value used by the Met Office data assimilation system and includes the fact that the observation may not be representative of the average conditions over a model grid-length. For categorical verification, the event is chosen to be the wind speed being at least 10 m/s. Similar results have been seen for verification of other quantities. Ideally account should be taken of the effect of varying climatology (Hamill & Juras, 2007), and confidence intervals should be used, but this is beyond the scope of the current study. A simple bias correction has been applied to all the forecasts before verification is performed.

2 Root-mean-square error

The root-mean-square error of a forecast is given by the RMS difference between the forecast and the verification as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i - v_i)^2}. \quad (1)$$

Here f_i is the forecast value for some quantity, and v_i is the value against which the verification is performed. Typically, the verification will be performed against either observations or an analysis of the atmospheric state. Both these sources of information will be corrupted with errors, which has a consequent effect on the score achieved by the forecast. As shown by Ciach and Krajewski (1999), if the errors in the observations can be treated as additive noise, then the RMSE of the forecast as measured against observations is given by

$$RMSE_o = \sqrt{RMSE_t^2 + RMSE_e^2} \quad (2)$$

where $RMSE_t$ is the value of the RMSE if the true state of the atmosphere is used in the verification, and $RMSE_e$ is the RMSE of the observations, as measured against the truth. Of course, the truth is not a quantity which is known, but it is the quantity against which we would wish to verify. Provided estimates of the observation errors can be obtained, then equation 2 may be inverted to give an estimate of $RMSE_t$.

Verification of the RMSE of MOGREPS ensemble mean forecasts for wind speed at 850 hPa is shown in figure 1. The RMSE of the ensemble mean forecast is clearly less when the effect of observation errors is accounted for. Also note that since it is the square of the RMSE that is subtracted, the effect of observation error is largest at short lead-times, when the forecast error is smallest. Also shown is the RMS spread of the ensemble forecasts. An ideal ensemble forecast would have a spread equal to the RMSE of the ensemble mean. The MOGREPS ensemble appears to be under-spread, even after accounting for observation error.

Evaluations of ensemble prediction systems have sometimes focused on the ensemble spread, and whether the spread is growing sufficiently fast. Since observation errors affect the apparent rate of growth of the RMSE of the ensemble mean forecast, the conclusion about whether the ensemble spread is growing fast enough may be affected, although in figure 1 the difference in growth rates appears small.

Another issue is the quantification of improvements in the forecasting system. If the RMSE of a forecast (as measured against observations) reduces by 1%, how much has the forecast really improved? Using the data from figure 1, the forecast will have improved by between 1.25% and 1.5% to give a 1% improvement in the RMSE measured against observations. This means that forecast improvements may be considerably larger than they appear to be. Conversely, an improvement to the observing network could improve the apparent quality of a forecast, even if the system has not been changed.

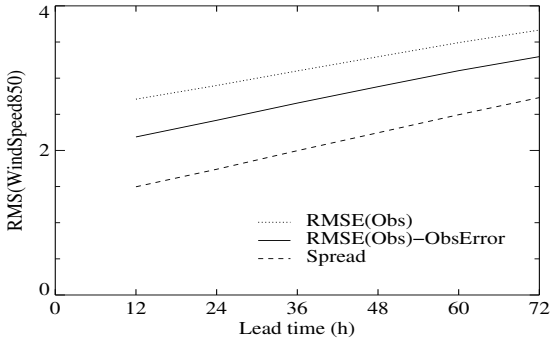

Figure 1. The root-mean-square error, and RMS spread, of forecasts of wind speed at 850 hPa, verified against radio-sonde observations.

3 Rank-histograms

The rank histogram was developed independently by a number of authors (Anderson, 1996; Hamill and Colucci, 1997; Talagrand *et al.*, 1997) in order to assess whether an ensemble forecast is reliable. If the ensemble forecast is reliable, then the set of ensemble member forecast values (at a given point) will be drawn from the same distribution as the true state. If this is the case, then it implies that if an n-member ensemble and the verification are pooled into a vector and sorted from lowest to highest, then the verification is equally likely to occur in each of the $n+1$ possible ranks. By repeated calculation of the rank for many independent sample points, the pooled result should be a uniform histogram. This is a general result that is true irrespective of the shape of the distribution which defines the truth.

In order to account for the effect of observation error, it is appropriate to randomly perturb the forecast values for each ensemble member by the observation error, as was suggested by Anderson (1996); Hamill (2001); Saetra *et al.* (2004). The effect of this approach is to increase the spread of the ensemble to account for the effect of observation error. A lot of focus on rank histograms is on the frequency with which the verification lies outside the range of the ensemble (the outlier frequency). Since the verification can often lie close to, but outside the forecast values then observation errors can have a very large effect on rank histograms (as argued by Saetra *et al.* (2004)).

Results for the MOGREPS ensemble, for forecasts of 850 hPa wind speed made at a lead time

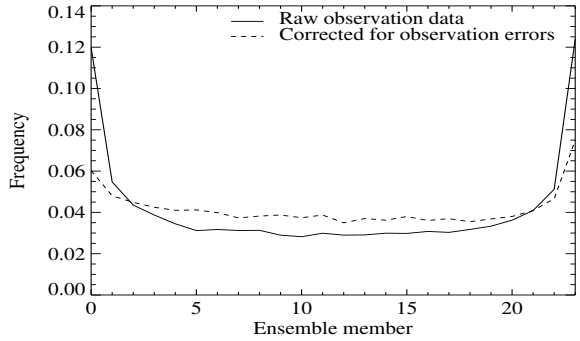

Figure 2. Rank histogram of 72h forecasts of wind speed at 850 hPa, verified against radio-sonde observations.

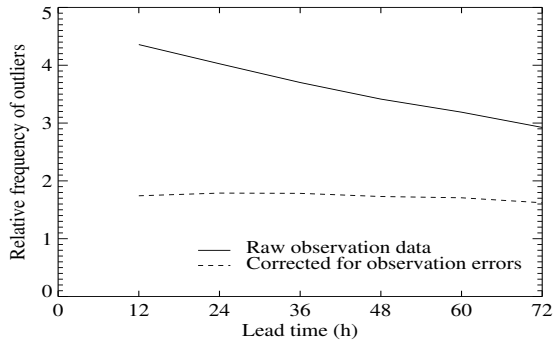

Figure 3. The frequency with which the verification lies outside the range of the ensemble, relative to the number expected for a flat rank histogram.

of T+72 hours, verified against radio-sonde observations is shown in figure 2. For the ideal flat rank histogram the verification would lie outside the range of the ensemble $2/24 \sim 8.33\%$ of the time (the control forecast is not included in this assessment). In reality, the observation lies outside the range of the ensemble approximately 24% of the time. This reduces to around 13% of the time once the effect of observation errors has been accounted for.

The frequency of outliers as a function of lead time is shown in figure 3. This has been normalised by the expected number of outliers (for a flat rank-histogram) so a reliable ensemble would have a value of one, and a value of two indicates the verification lying outside the range of the ensemble twice as often. As the lead time increases the frequency of outliers, when verification is performed against observations, decreases rapidly. When account is taken of observation errors, however, the frequency of outliers is approximately constant with lead time. This indicates a persistent feature of the effect of observation errors - they are larger at shorter lead time.

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

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

4 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 I. For a perfect forecast (and error-free observations) only hits and correct rejections would be seen. Innumerable skill scores may be derived from a categorical forecast (Stephenson, 2000) and their popularity lies in the simplicity of the verification method.

Accounting for observation errors by perturbing the forecast value by the observation error would not be acceptable, since this would reduce the measured skill of the forecast. Instead, the effect of observation errors must be accounted for by treating the error as a perturbation to the observations, and trying to remove the effect of this perturbation. Bowler (2006) performed this using a deconvolution method.

Consider the instance when some event has been forecast to occur. For a large number of such cases, one may consider the true value of the quantity being forecast to have been drawn from some distribution, P_t. Similarly, the observations may be taken as being drawn from another distribution, P_o. Finally, we define the observation errors to be taken from a third distribution, P_e. If the errors in the observation are independent of the true value of the quantity being observed, then the distribution of the observations may be written as the convolution of the distribution of the truth with the pdf of the observation errors

$$P_o(x|F=1) = \int_{-\infty}^{\infty} P_t(y|F=1)P_e(x-y|F=1)dy \tag{3}$$

where x is the observed value, y is the true value and the distributions have been conditioned on the event being forecast to occur ($F=1$). This formula can be equally applied for forecasts of the event not occurring ($F=0$). 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

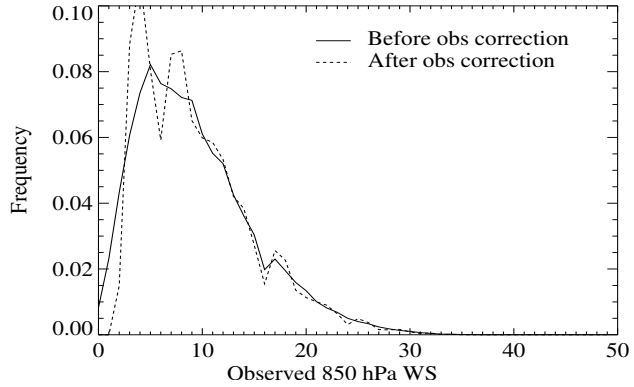

Figure 4. The distribution of observed values, not conditioned on the forecast

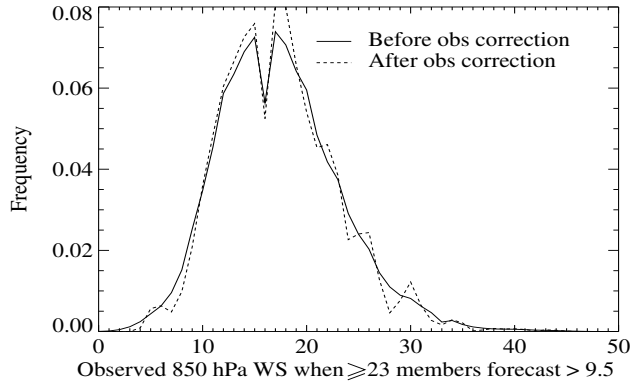

Figure 5. The distribution of observed values, given that the wind speed was forecast at T+72h to be at least 10 m/s by all ensemble members.

values may be estimated by calculating what fraction of this distribution lies above the event threshold.

Results for the MOGREPS ensemble are shown in figures 4 and 5. Each figure shows the distribution of observed values, before and after the deconvolution has been applied. Figure 4 shows the distribution of observed values without any conditioning on the forecast value. As can be seen, most of the observed values lie below the 10 m/s threshold, indicating that the wind speed is often observed to be below this threshold. Similarly, figure 5 shows the observed distribution, given that all the ensemble members forecast the wind speed to be greater than 10 m/s. In this case, most of the observations are above the threshold.

In both figures 4 and 5, the dashed line shows the distribution of observations after the deconvolution has been applied. These distributions are close to the original distribution of the observations, indicating that observation error has a relatively small effect. However, the removing the effect of observation errors narrows the distribution, decreasing the number of events that were observed to occur when the forecast said they would not, and vice-versa. Thus, the system is judged to be more skilful when account is taken of the effect of observation errors. The deconvolved distribution in figure 4 is noisy for low wind speeds. This is because the observed distribution drops-off sharply at low wind speeds, and it is difficult for the deconvolved distribution to replicate this behaviour. This suggests that the error model of additive noise is inappropriate at these wind speeds.

The kink in the distribution of observations (see figures 4 and 5) at wind speeds of around 16 m/s is due to problems with fitting the data into discrete bins. Many of the radio-sondes report the wind speed in whole numbers of metres per second. However, many of the radio-sondes report the wind speed in whole numbers of knots. These two conventions mean that there is no binning method which will be well suited to both types of data, and hence the kink observed. This rounding has been reproduced in the code used in this study, so should not affect the results presented. Given the observations are believed to have errors of around 1.6 m/s, it is strange that some reports only contain information to the nearest 1 m/s!

Plots of the relative operating characteristic (Mason & Graham, 1999) show the hit rate (H) against the false alarm rate (F) for different confidence levels (such as probability of precipitation greater than 50%). The hit rate and false alarm rate are defined as follows

$$H = \frac{a}{a+b} \tag{4}$$

$$F = \frac{c}{c+d} \tag{5}$$

where a, b, c and d are the standard contingency table values shown in table I. Values at different probability thresholds define a series of points, which are often joined by straight lines. Observation errors can be accounted for by calculating the distribution of observations for each probability threshold, then performing the deconvolution, giving results as shown in figures 4 and 5. From these the contingency table values may be calculated, and therefore H and F.

Figure 6 shows the relative operating characteristic (ROC) for MOGREPS forecasts, before and after account is taken of observation errors. The system is noticeably more skilful when observation errors are accounted for. This may appear surprising, after having noted that observation errors have a small effect on the distribution of observed values. However, the quality of a forecast system is typically judged on the small number of misses and false alarms that it is verified to have, and these numbers are much more sensitive to the effect of observation errors.

The area under the ROC curve gives a single-figure summary of the information contained in the ROC plot, and can be taken as a skill score of the forecast. For simplicity the area is calculated using a series of straight-line segments, rather than the superior method advocated by Wilson (2000). The

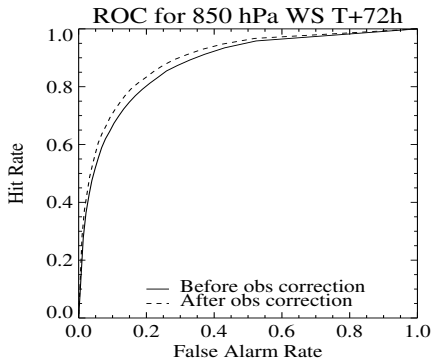

Figure 6. Relative operating characteristic (ROC) for 72h ensemble forecasts of the event of the wind speed being greater than 10 m/s.

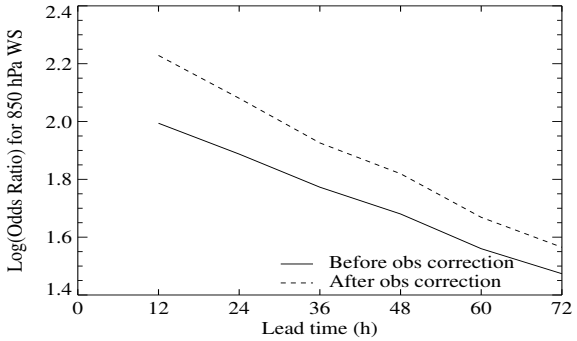

Figure 8. Log of the odds ratio of the control forecast of the ensemble, for forecasting the event of the wind speed being greater than 10 m/s.

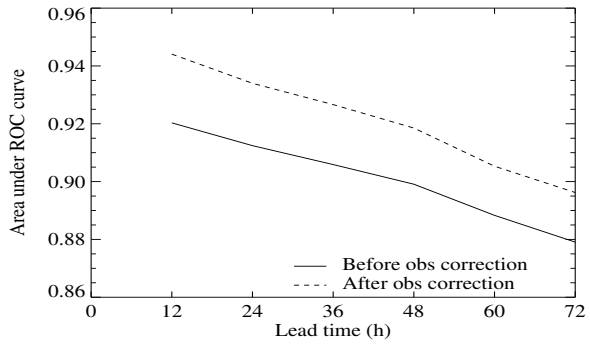

Figure 7. Area under ROC curve as a function of lead time for forecasting the event of the wind speed being greater than 10 m/s.

ROC area for MOGREPS forecasts of wind speed of 10 m/s is plotted as a function of lead time in figure 7. This indicates that the effect of observation errors is equivalent to a degradation of forecast quality of around one day.

Although much of the focus of this paper has been on the performance of probabilistic forecasts, the deconvolution method outlined above can be used for deterministic forecasts. This allows the calculation of the contingency table (see table I) before and after the correction for observation errors has been applied. A popular skill-score for the deterministic forecast is the odds ratio (Stephenson, 2000), and is plotted in figure 8 for the MOGREPS control forecast. The effect of observation errors is clear, implying a degradation in forecast quality equivalent to a lead time between 12 and 24 hours.

The deconvolution algorithm is based on a simple optimisation routine. An initial deconvolved distribution is considered, which is convolved with the distribution for the observation error, and this distribution is compared with the distribution of the observations. A small change to the initial deconvolved distribution is then tested. The modified distribution is convolved with the observation error distribution, and this is compared with the distribution of the observations. If the small change brings the convolved distribution closer to the observed distribution, then the change is accepted, and the process repeats.

5 Verification against analysis

It is quite common in numerical weather prediction to perform verification against an analysis of the atmospheric state. This can pose problems, since the errors in the analysis are not easily found, and those errors may be related to the errors in the forecast being verified. The analysis is typically calculated by combining information from a forecast from the previous analysis cycle (background forecast) and the latest observations. Clearly, errors in the background forecast will propagate into the new analysis, although their impact will be reduced by the use of the latest observations. Simmons & Hollingsworth (2002) estimated the the correlation between analysis and forecast error is between 0.37 and 0.5 for 1-day forecasts of geopotential height at 500 hPa. Any persistent model errors will, by definition, be present in the forecast, but therefore also in the analysis. This makes the use of the analysis for verification very troublesome indeed!

Figure 9 shows the RMSE of the MOGREPS forecasts, as displayed in figure 1, but also with verification against analysis, where the analysis has been interpolated to the location of the radio-sonde observations. The RMSE of the forecast, when measured against analyses, is less than the RMSE measured against observations, even after observation error has been accounted for. This indicates that either errors in the analysis are highly correlated with forecast errors or that the observation errors are larger than the value used here - probably both are true. Another point to note is that the rate of growth of the RMSE measured against analyses is larger than the

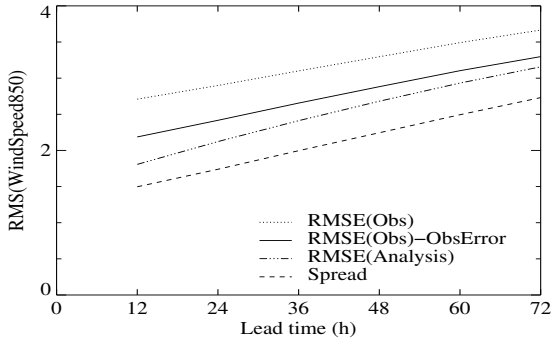

Figure 9. The root-mean-square error, and RMS spread, of forecasts of wind speed at 850 hPa, verified against radio-sonde observations and analyses.

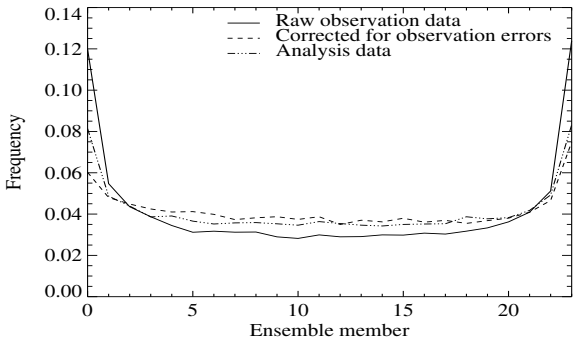

Figure 10. Rank histogram of 72h forecasts of wind speed at 850 hPa, verified against radio-sonde observations and analyses

rate of growth of RMSE measured against observations. As mentioned earlier, this may lead to erroneous conclusions about the rate of growth of the ensemble spread.

It is also instructive to look at the rank histogram verification of the MOGREPS ensemble using analyses. Figure 10 shows the same as figure 2, but with verification against analyses. A lower RMSE is normally consistent with seeing a lower outlier frequency in the ensemble verification. However, the verification against analysis gives a lower RMSE, but higher outlier frequency than the corrected verification against observations. This may be explained by the results of Candille & Talagrand (personal communication) who showed that accounting for the effect of observation error, by perturbing the ensemble forecasts, can compensate for a lack of spread in the ensemble. This suggests that the number of outliers seen after correcting for observation errors may be an under-estimate.

6 Methods to estimate observation error

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 principal source of 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.

Verification of short-range forecasts provides an immediate estimate of the value of the observation error. The observation error cannot be larger than the RMSE of the forecast measured against observations (assuming forecast and observation errors are uncorrelated) as is dictated by equation 1. A commonly used method to estimate the observation error is provided by Daley (1993) plotting the covariance between forecast errors as a function of the distance between observations. This, extrapolated to zero distance provides an estimate of the observation error, under the assumption that observation errors are uncorrelated in space. Another method for diagnosing observation error is provided by the data assimilation system (Desroziers et al., 2005). The differences between the observed values, the background forecast and the analysed values can be used as a consistency check on the values of the supplied observation and background errors.

7 Conclusion

In this paper a number of methods for accounting for observation error in verification statistics have been examined. In each case the effect of observation error has had a clear effect on the verification statistics, and may change the conclusions one draws about how the forecast system should be improved.

For the RMSE, accounting for observation errors is a simple post-processing procedure. The reduction in RMSE of 850 hPa wind speed for forecasts between 0 and 3 days when accounting for observation errors was between 10% and 20%. The effect of observation errors was even more dramatic for rank histograms with a halving of the frequency with which the verification lies outside the range of the ensemble. The behaviour of the number of outliers as a function of lead time also appears sensitive to the effect of observation errors. The frequency of outliers reduces rapidly with forecast lead time when verification is performed against raw observations, but is approximately constant when account is taken of observation errors.

For categorical verification of forecasts, account was taken of observation errors using a deconvolution approach. The binning of observations posed problems and care was needed to replicate the

method used in the raw data. Observation errors have a clear effect, which is again most substantial in the short range. Verification against analysis was examined and it was shown that interpretation is difficult for this due to correlations between the errors in the analysis and forecast.

Central to this work is an accurate estimate of the observation errors. A number of methods exist for these estimates, and it is advisable to use a combination of approaches.

There are some verification statistics (such as continuous rank probability skill score Hersbach (2000)) for which the effect of observation errors is still unquantified. However, there are now a large range of scores for which the effects are understood to some degree.

References

Anderson, J.L., "A method for producing and evaluating probabilistic forecasts from ensemble model integrations" *Journal of Climate*, **9**, 1518-1530 (1996)

Bowler, N.E., "Explicitly Accounting for Observation Error in Categorical Verification of Forecasts" *Monthly Weather Review*, **134**, 1600-1606 (2006)

Bowler, N.E., Arribas, A., Mylne, K.R., Robertson, K.B. and Beare, S.E., "The MOGREPS short-range ensemble prediction system" *submitted to Quarterly Journal of the Royal Meteorological Society* (2007)

Ciach, G.J. and Krajewski, W.F., "On the estimation of radar rainfall error variance" *Advances in Water Resources*, **22**, 585-595 (1999)

Daley, R., "Estimating observation error statistics for atmospheric data assimilation" *Annales Geophysicae*, **11**, 634-647 (1993)

Desroziers, G., Berre, L., Chapnik, B., and Poli, P., "Diagnosis of observation, background and analysis-error statistics in observation space" *Quarterly Journal of the Royal Meteorological Society*, **131**, 3385-3386 (2005)

Hamill, T.M. and Colucci, S.J., "Verification of eta-RSM short-range ensemble forecasts", *Monthly Weather Review*, **125**, 1312-1327 (1997)

Hamill, T.M., "Interpretation of rank histograms for verifying ensemble forecasts" *Monthly Weather Review*, **129**, 550-560 (2001)

Hamill T.M. and Juras J., "Measuring forecast skill: is it real skill or is it the varying climatology?", *in press, Quarterly Journal of the Royal Met. Soc*

Hersbach H., "Decomposition of the continuous ranked probability score for ensemble prediction systems", *Weather and Forecasting*, **15**, 559-570 (2000)

Mason, S.J., and Graham N.E., "Conditional probabilities, relative operating characteristics, and relative operating levels", *Weather and Forecasting*, **14**, 713-725 (1999)

Saetra, O., Hersbach, H., Bidlot, J.R., and Richardson, D.S., "Effects of observation errors on the statistics for ensemble spread and reliability" *Monthly Weather Review*, **132**, 1487-1501 (2004)

Simmons A.J., and Hollingsworth, A., "Some aspects of the improvement in skill of numerical weather prediction" *Quarterly Journal of the Royal Meteorological Society*, **128**, 647-677 (2002)

Stephenson, D.B., "Use of the 'odds ratio' for diagnosing forecast skill" *Weather and Forecasting*, **17**, 221-232 (2000)

Talagrand, O., Vautard, R. and Strauss, B., "Evaluation of probabilistic prediction systems" in *Proceedings of the workshop on predictability*, ECMWF, Reading, Berkshire, UK, 1-26 (1997)

Wilson, L.J., Comments on "probabilistic predictions of precipitation using the ECMWF ensemble prediction system", *Weather and Forecasting*, **15**, 361-364, (2000)