A new parametric model for the assessment and calibration of medium-range ensemble temperature forecasts

Stephen Jewson,* Anders Brix and Christine Ziehmann
RMS, London, UK

Abstract

We present a new method for the assessment and calibration of medium-range ensemble temperature forecasts. The method is based on maximizing the likelihood of a simple parametric model for the temperature distribution, and leads to some new insights into the predictability of uncertainty.

Keywords: ensemble forecast; probabilistic forecast; medium-range forecast; ECMWF; forecast calibration; forecast assessment; spread regression; likelihood

1. Introduction

A number of different methods have been used for the assessment and calibration of ensemble forecasts (for example, see Atger, 1999; Coelho et al. 2004; Richardson, 2000; Roulston and Smith, 2002; Wilson et al., 1999). In many applications of ensemble forecasts the forecast is used to derive the probability of a certain outcome, such as temperature dropping below zero, conditional on all the information available at the time the forecast is made. In this context, the reliability diagram and the relative operating characteristic (ROC) are both useful tools (see, for example, Anderson, 1996; Eckel and Walters, 1998; Talagrand et al., 1997; Hamill, 1997; Mason, 1982; Swets, 1988; Mason and Graham, 1999).

In other applications of ensemble forecasts, however, the forecast is interpreted as providing a mean and a distribution of future values of temperature, again conditional on all the available information at the forecast time. For example, in the field of weather derivatives the calculation of the fair strike for a certain class of weather swap contract\(^1\) needs an estimate of the conditional mean of the future temperatures, while the calculation of the fair premium for weather option contracts needs an estimate of the whole conditional distribution of future temperatures (see Jewson and Caballero, 2003, for details). Additionally, the assumption is often made that the conditional distribution of temperature is normal since this allows the temperature forecast to be summarized succinctly using just the conditional mean and standard deviation. For such mean-and-distribution or mean-and-standard deviation-based applications of ensemble forecasts the reliability diagram and the ROC are not particularly appropriate.

In this paper we present a new parametric model for the assessment and calibration of ensemble temperature forecasts based on analysis of the mean and standard deviation of the conditional distribution of temperatures. We call this model ‘spread regression’. Spread regression has been developed to respond to the need for a simple and practical method for assessment and calibration that can be used by companies that make use of ensemble forecasts in the weather derivative market. We postulate a parametric model for the mean and standard deviation and fit the parameters of the model using the maximum likelihood method. This approach has a number of advantages relative to the assessment and calibration methods mentioned above. The model is simple, easy to interpret, and the entire ensemble distribution can be calibrated in one simple step. Also the model gives a clear indication of how many days of useful information there are in a forecast.

In Section 2 we describe the data sets we use for this study. In Section 3 we describe the statistical model that forms the basis for the method we propose. In Section 4 we describe the results from fitting the model. In Section 5 we discuss extensions to other distributions and in Section 6 we summarize our results and draw some conclusions.

2. Data

We will base our analyses on 1 year of ensemble forecast data for the weather station at London’s Heathrow airport, WMO number 03 772. The forecasts are predictions of the daily average temperature, and the target days of the forecasts run from 1 January 2002

---

\(1\) Linear swaps on a linear temperature index.
to 31 December 2002. The forecast was produced from the European Centre for Medium Range Weather Forecasting (ECMWF) model (Molteni et al., 1996) and downscaled to the airport location using a simple interpolation routine prior to our analysis. There are 51 members in the ensemble. We will compare these forecasts to the quality-controlled climate values of observed daily average temperature for the same location as reported by the United Kingdom Meteorological Office (UKMO).

Throughout this paper all equations and all values have had the seasonal cycles in the mean and the standard deviation removed using the method of Caballero et al. (2002). The seasonal cycle in the mean was removed by transforming the data to Fourier space, setting the first three harmonics of the annual cycle to zero, and transforming back. The result of this step we call ‘single anomalies’. The single anomalies were then squared to create a variance process and the same procedure was applied to remove the seasonal cycle in this variance process. The final result, which has almost no residual seasonal cycle in either mean or variance, we call ‘double anomalies’ or ‘standardized anomalies’. Removing the seasonal variance removes most of the seasonality in the forecast error statistics, and justifies the use of non-seasonal parameters in the statistical models for temperature that we propose.

### 3. The moment-based ensemble assessment and calibration model

For forecasts of temperature anomalies it has long been recognized (see, for example, Leith, 1974) that the use of a final linear regression step between ensemble mean and observations can eliminate bias and minimize the mean square error (MSE). For our purposes we will write this regression step as

$$ T_i \sim N(\alpha + \beta m_i, \sigma) $$

where $T_i$ is the observed temperature on day $i$, $N(\mu, \sigma)$ represents a normal distribution with mean $\mu$ and standard deviation $\sigma$, $m_i$ is the forecast of the temperature (in our case, the ensemble mean) and $\alpha$, $\beta$ and $\sigma$ are free parameters. The symbol $\sim$ means ‘has the distribution’ and Equation (1) should thus be read ‘the distribution of $T$ on day $i$, conditioned on $m_i$, is modelled as a normal distribution with mean $\alpha + \beta m_i$ and standard deviation $\sigma$’.

The values for $\alpha$, $\beta$ and $\sigma$ come from fitting the model, and this is usually done using least squares linear regression. One justification for the use of least squares linear regression is that for this particular model least squares estimates are also maximum likelihood estimates as long as we assume that the forecast errors are uncorrelated in time. We note that although the model in Equation (1) postulates that the data comes from a normal distribution, it can be applied in situations in which the data is not strictly normal. There are many limitations of the linear regression model, and there are many methods that one could try and use to improve on it. These include:

- use of the ensemble spread as an additional predictor;
- use of the individual ensemble members as predictors;
- use of lagged predictors;
- use of variables other than temperature as predictors;
- use of a parametric non-linear function of $m$ instead of a linear function;
- use of a stochastic parameter framework;
- use of neural networks;
- use of the Bayesian methods of Coelho et al. (2004) that can allow the incorporation of prior climatological information.

Many of these methods could also be used in conjunction: for instance, one could consider using the ensemble spread as a predictor in the Bayesian framework of Coelho et al. (2004), and so on. A long-term goal of research in this area is to work out which of these methods can really improve on the linear regression model, and by how much. In this paper our contribution is to study one of these methods in particular: the idea of using the ensemble spread as an extra predictor.

One of the assumptions in the linear regression model is that the standard deviation of the forecast errors $\sigma$ is constant. However, it is well documented that the size of forecast errors varies in time (Palmer and Tibaldi, 1988) and that there is a relationship between the ensemble spread and the size of forecast errors (Toth et al., 2000). It thus makes sense to attempt to generalize the model in Equation (1) to a model that takes these temporal variations in $\sigma$ into account. We will do this using the model

$$ T_i \sim N(\alpha + \beta m_i, \gamma + \delta s_i) $$

where the free parameter $\sigma$ has been replaced by a linear function of the ensemble spread $s_i$, and two new parameters $\gamma$ and $\delta$ have been introduced. Modelling the standard deviation as a linear function of the ensemble spread in this way allows for both time variation and the correction of biases in the predicted uncertainty.\(^2\)

The optimum parameters for this model can no longer be fitted using least squares linear regression. However, they can be fitted if we can identify a cost function that can be minimized or maximized by varying the parameters. There are various possibilities for such a cost function, but one of the most natural is the likelihood, which, in the classical (i.e. non-Bayesian) statistical modelling framework that we

---

\(^2\) We note that one could alternatively model the variance as a linear function of the spread squared.
will use, is defined as the probability density of the observations as a function of the parameters. Rather than maximize the likelihood, we actually maximize the log-likelihood. This gives the same values for the parameters since logarithm is a monotonically increasing function.

Maximizing the likelihood (or log-likelihood) is the standard way to fit parameters in classical statistics (see, for instance, textbooks such as Casella and Berger, 2002, or Lehmann and Casella, 1998) and, asymptotically, gives the most accurate possible estimates of the parameters for most statistical models. For our particular model the likelihood is

$$l = -\frac{1}{2} \ln(2\pi \text{det}) - \frac{1}{2} (T - \hat{\mu})^T \hat{\Sigma}^{-1} (T - \hat{\mu})$$  (3)

where \( l \) is the log-likelihood, \( T \) is a vector of all the observed temperatures, \( \hat{\mu} \) is a vector of the predicted conditional mean temperature with \( \hat{\mu}_i = \alpha + \beta m_i \), \( \hat{\Sigma} \) is the covariance matrix between forecast errors, and \( \text{det} \) is the determinant of \( \hat{\Sigma} \).

We will make the approximation that the forecast errors are uncorrelated in time. This means that \( \hat{\Sigma} \) is diagonal, and simplifies the likelihood to

$$l = l(\alpha, \beta, \gamma, \delta) = -\frac{1}{2} \sum_{i=1}^{n} \ln(2\pi \hat{\sigma}_i)$$

$$- \frac{1}{2} \sum_{i=1}^{n} \frac{(T_i - \hat{\mu}_i)^2}{\hat{\sigma}_i^2}$$  (4)

where \( \hat{\sigma}_i = \gamma + \delta s_i \). As we will see later when we consider residuals from the model, the assumption that the forecast errors are uncorrelated in time is not entirely correct. However, maximizing the full log-likelihood given by Equation (3) is somewhat challenging: developing numerical methods for that purpose (or applying methods that have already been developed in other fields) is one of the most pressing areas for future work arising from this study, and is a work in progress for the authors.

To avoid potential confusion we note that the definition of the likelihood that we are using in this study comes from classical statistics. A slightly different definition has been used in meteorology, following Murphy and Winker (1987). We discuss the differences between these two definitions in Jewson (2003d).

As with the linear regression model, this model is also not restricted to cases in which temperature is exactly normally distributed: the assumption of the normal distribution merely provides a metric in which the likelihood can be calculated and the parameters fitted. This metric is most appropriate when the data is at least close to normally distributed. For cases when the data is not close to normal other distributions can be used, or the data can be transformed to normal.

There are a number of useful features of the model we present. These include the following:

- Once the parameters have been fitted to past historical data, calibration of future ensemble forecasts is easy since it just involves applying linear transformations to the ensemble mean and standard deviation. The calibrated values for the mean and the standard deviation can be used to define the whole forecast distribution, or can be used to shift and stretch the individual ensemble members, if individual ensemble members need to be preserved. In the latter case non-normality in the distribution of the original ensemble members will not be destroyed.

- The optimum values of the parameters in Equation 2 have clear interpretation and give us useful information about the performance of the ensemble. For instance, \( \alpha \) identifies a bias in the mean, and \( \beta \) represents a scaling of the forecast towards climatological values. In a perfect forecast, \( \alpha \) would be zero and \( \beta \) would be one. The spread parameters \( \gamma \) and \( \delta \) combine to optimize the prediction of uncertainty about the mean. The value of the ensemble spread \( s \) varies in time because of the dependence of the growth rate of differences between ensemble members on the actual model state. The calibrated standard deviation value \( \sigma_i = \gamma + \delta s_i \) additionally includes uncertainty due to model error. If the spread of the ensemble contains very little real information, \( \delta \) will tend to be small, and \( \gamma \) will tend to be large to compensate.

- It is very easy to calculate approximate uncertainty levels on the values of the parameters as part of the fitting procedure. This is done using the curvature of the log-likelihood at the maximum (see the above references on likelihood methods). These uncertainty levels give us a clear answer to the question of whether the ensemble forecast has useful skill at different lead times. For instance, once \( \beta \) is not significantly different from zero we can say that the ensemble mean no longer contains useful information (at least not within this framework) and once \( \delta \) is not significantly different from zero then we can say that the ensemble spread no longer contains useful information. This raises the interesting possibility that we might identify situations in which the mean may contain more days of useful information than the spread. One caveat is associated with the assumption that the forecast errors are uncorrelated, which we will see is slightly incorrect. It is not immediately clear what effect more accurate modelling of the forecast errors would have on the parameter and uncertainty estimates, but in similar modelling situations it is often the case that the estimated parameters are the same but the uncertainty levels are wider, and it would seem possible that this would be true here. Nevertheless even if the forecast errors were perfectly modelled, then we would still expect the uncertainty estimates to be too narrow since they rely on the assumption that the model is correct (which is wrong because all models are wrong) and that the data is stationary (which is wrong because of seasonality and model
Assessment and calibration of medium-range ensemble temperature forecasts

changes). These are generic problems with model-based uncertainty estimators.

- It is often necessary to decide which of two forecasts is the more accurate. If two forecasts are both calibrated using Equation (2) then the log-likelihood provides a natural way to compare the forecasts. Log-likelihood measures the ability of the forecast to represent the whole distribution of observed temperatures, and can be thought of a generalization of mean square error. It can be presented in a number of ways such as log-likelihood or log-likelihood skill score. A further discussion of the use of likelihood-based scores for evaluating ensemble forecasts is given in Jewson (2003e).

Forecasts calibrated using Equation (2) will not necessarily minimize MSE. Users interested purely in a single forecast that minimizes MSE should thus calibrate using Equation (1). However, users interested in predictions of uncertainty, or, equivalently, in the whole conditional distribution of possible temperatures, should calibrate using Equation 2. In practice we have found that the mean temperature prediction produced by Equation (2) is close to that produced by Equation (1), presumably because the fluctuations in uncertainty are not large.

4. Results

The first set of results we show are the residuals from the spread-regression model. These give an indication as to how well the model is able to represent the data. The upper left-hand corner of Figure 1 shows the autocorrelation function for the residuals for three different leads. We see that the residuals are weakly autocorrelated in time. This contradicts our modelling assumption that they are uncorrelated in time, and is probably the most important flaw in the spread regression model, as used in this paper. We do not think that modelling this aspect of the data more accurately will have a major impact on the results, but we cannot be sure, and so developing models that can cope with this autocorrelated forecast errors correctly is a high priority for future work. The other panels in Figure 1 show QQ plots of the residuals against a normal distribution for three leads (QQ plots are a standard way to compare distributions). We see that the residuals are reasonably close to being normally distributed, which validates our assumption to that effect.

The optimum values for the parameters in Equation (2) for our 1 year of forecast data and observations are shown in Figure 2. In each case we show the approximate 95% sampling error confidence intervals around the optimum parameters. In some cases they are so narrow that they are hard to see in the graphs.

Looking at $\alpha$ we see that there is a small and roughly constant bias in the temperatures produced by the ensemble. Correction of the ensemble mean (or each ensemble member) using $\alpha$ would eliminate this bias, as long as the ensemble remains stationary.

The parameter $\beta$ is slightly less than 1 at all leads. This shows that the ensemble mean varies too much: either the ensemble mean, or each ensemble member, should be reduced by the factor $\beta$ towards the

Figure 1. Analysis of the residuals from the spread-regression model. The upper left panel shows the ACF of the residuals for three different lead times. The other panels show QQ plots of the residuals for leads 2, 5 and 9.
climatology. Such a damping factor is presumably required because the ensemble members are more correlated with each other than they are with the observations and because the ensemble is finite in size. Even at lead 10 $\beta$ is highly significantly different from zero, implying that the ensemble mean still contains useful predictive ability at that lead. If we allow ourselves to extrapolate the $\beta$ curve to longer leads by eye, it would seem likely that the ensemble mean would still contain useful predictive information even beyond that.

The fact that our values of $\delta$ are significantly different from zero out to the end of the forecast (just) shows that there is significant information in the ensemble spread too. However, in this case if we extrapolate to higher lead times by eye it seems unlikely that there would be any more skill in $\delta$. Since $\delta$ is below one and $\gamma$ is non-zero we see that the standard deviation of the ensemble is not an optimal estimate of the uncertainty of the prediction.

The $\gamma + \delta s$ transformation can change both the mean spread (the time mean of the standard deviation across the ensemble) and the variability of that spread (the standard deviation in time of the standard deviation across the ensemble). To measure the effect on the mean spread, Figure 3 shows values of $(\gamma + \delta X)/\overline{s}$ (where the overbar indicates the mean in time over the year of data), which shows the factor by which the transformation increases the mean spread. We see that at short lead times the ensemble spread $s$ is far too small on average and the calibration increases the spread by factors of around 4 (at lead 0) and 2 (at lead 1). At longer lead times the ensemble spread is still too small on average by a factor of around 1.2. This underestimation of the spread from ensemble forecasts has been noted by a number of authors such as Ziehmann (2000) and Mullen and Buizza (2000). It is likely to be due to model error in the prediction model and due to the fact that the forecast is a prediction of a large-scale flow while the observation is site-specific and hence affected by small-scale variability not represented in the model.

The size of the effect of the calibration on the variability (in time) of the spread is given by the value of $\delta$. Since $\delta$ is significantly different from one at all lead times beyond the first, we conclude that the variability of the spread from the ensemble needs to be reduced to be optimal at those lead times. This could be because the variability of the ensemble spread is too large, or because the variability of the ensemble spread is not highly correlated with the real variability of skill.
Assessment and calibration of medium-range ensemble temperature forecasts

Figure 4. Both lines show the ratio of the standard deviation in time of the standard deviation across the ensemble to the mean in time of the standard deviation of the ensemble. This ratio is given the name coefficient of variation of spread (COVS) in the text. The solid line was estimated using the uncalibrated ensemble, and the dashed line using the calibrated ensemble.

We can see from the values of δ that the variability of the ensemble spread alone will overestimate the state-dependent predictability of this model by a large factor at long leads. A better estimate for the level of state-dependent predictability is given by the variability of the calibrated spread, which is smaller by the δ factor.

Figure 4 shows the ratio of the standard deviation of the ensemble spread to the mean ensemble spread at different lead times. We call this ratio the coefficient of variation of the spread (COVS):

\[
\text{COVS} = \frac{\text{sd}(s)}{\bar{s}}
\]

Figure 4 shows the COVS estimated from both the uncalibrated and the calibrated ensemble data. These values give an indication of how much extra information we get about the forecast uncertainty by using the (uncalibrated or calibrated) spread of the ensemble rather than using a level of uncertainty which is constant with time. The uncalibrated data suggests that variations in uncertainty that are 20–55% of the mean uncertainty are predictable using the ensemble spread. However, because the uncalibrated data both underestimates the total spread (the numerator in the COVS) and overestimates the predictable part of the variability of the spread (the denominator in the COVS) these values seem to be overestimates. The calibrated data suggests that variations in the uncertainty that are only 5–20% of the mean uncertainty are predictable using the ensemble spread.

5. Other distributions

In cases where the forecast errors are not close to normally distributed, one can use other distributions. For example, in the case where the forecast errors show skew, the skew-normal distribution \( SN \) can be used. The skew-normal distribution is a generalization of the normal distribution which has a third parameter, and includes the possibility of modelling skew (Azzalini, 1985). Suppressing the index \( i \) for clarity we then have

\[
T \sim SN(\alpha + \beta m, \gamma + \delta s, \zeta + \eta k)
\]

where we have introduced the ensemble skew \( k \) and two new parameters \( \zeta \) and \( \eta \).

The skew-normal model can be fitted using maximum likelihood methods exactly as for the normal distribution. One of the results from such a fitting process would be a clear indication as to whether the forecast being calibrated does or does not contain statistically significant information about the skew of observed temperatures (this question has been discussed by, among others, Denholm-Price and Mylne, 2003, and Stephenson and Doblas-Reyes, 2000).

For extremely non-normal distributions for which even the skew-normal is not non-normal enough, non-parametric distributions may be more appropriate. A simple non-parametric method would be to use a kernel density, with a single free parameter for the width of each kernel (see Bowman and Azzalini, 1997, for a description of kernel densities). Such a method would look a little like the method of Roulston and Smith (2003) even though it is justified in a completely different way.

6. Conclusions

We have described a simple parametric method for the assessment and calibration of ensemble temperature forecasts. The method consists of applying linear transformations to the mean and standard deviations from the ensemble. The parameters of the model can be fitted easily using the maximum likelihood method. The model has various advantages and disadvantages relative to other calibration models currently in use. The main disadvantage is that the model only works for forecast errors that are reasonably close to normally distributed, although extensions have been described that should overcome that limitation. The advantages of the model are that:

- the calibration of forecasts using the model is extremely simple;
- the model is transparent and easy to understand;
- the model separates skill in predictions of the mean and the spread;
- calculating approximate confidence intervals on parameters is easy;
- the model gives a clear indication of how many days of useful skill there are in a forecast.

We have applied the model to 1 year of site-specific ECMWF ensemble forecasts. We find that the forecasts have highly significant skill for predicting both the
mean and the standard deviation out to 10 days. The forecasts underestimate the mean uncertainty, as has been reported in other studies. They also overestimate the variability of the uncertainty. For these forecasts we estimate that the predictable part of the uncertainty is only between 5% and 20% of the mean uncertainty, depending on lead time. For some applications this variability in the uncertainty may be small enough that it can be ignored and one could make the simplifying assumption that the uncertainty is constant in time.

Further work includes:

- developing algorithms that avoid having to make the assumption that the forecast error is uncorrelated in time;
- out-of-sample testing of the calibrated forecasts, using both measures from within the framework (i.e. likelihood) and also other measures such as rank histograms, reliability diagrams and ROCs.

Finally we note that during the time take to review this article, a number of areas of further research have already been completed and published as non-peer reviewed technical reports. These include:

- a comparison of the linear regression and spread regression models using in- and out-of-sample tests (Jewson, 2003c);
- an assessment of whether generalizing the distribution of the forecast errors in the spread regression model to a general non-normal distribution brings any benefit (Jewson, 2003b);
- an assessment of the relative contribution of the ensemble mean and the ensemble spread to the skill of probabilistic forecasts generated by the spread regression model (Jewson, 2003a).

Acknowledgements

We would like to thank K. Mylne for providing us with the forecast data, Risk Management Solutions for providing us with the observational data, and S. Mason, F. Doblas-Reyes and D. Anderson for useful discussions on the subject of ensemble validation. This research was funded by the authors as part of an on-going project to try and develop more useful weather forecasts for the energy and weather derivatives industries.

References

Anderson JL. 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. Journal of Climate 9: 1518–1530.
Atger E. 1999. The skill of ensemble prediction systems. Monthly Weather Review 127: 1312–1327.
Azzalini A. 1985. A class of distributions which includes the normal ones. Scandinavian Journal of Statistics 12: 171–178.
Bowman A, Azzalini A. 1997. Applied Smoothing Techniques for Data Analysis. Oxford Science: Oxford.
Caballero R, Jewson S, Brix A. 2002. Long memory in surface air temperature: detection, modelling and application to weather derivative valuation. Climate Research 21: 127–140.
Casella G, Berger RL. 2002. Statistical Inference. Duxbury: Pacific Grove, CA.
Coelho C, Pezzulli S, Balmaseda M, Doblas-Reyes F, Stephenson D. 2004. Forecast calibration and combination: a simple Bayesian approach from ENSO. Journal of Climate 17: 1504–1516.
Dennholm-Price J, Mylne K. 2003. Can an ensemble give anything more than gaussian probabilities? Nonlinear Processes in Geophysics 10: 469–475.
Eckel A, Walters M. 1998. Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. Weather Forecasting 13(December): 1132–1147.
Hamill T. 1997. Reliability diagrams for multcategory probabilistic forecasts. Weather Forecasting 12(December): 736–741.
Jewson S. 2003a. Comparing the ensemble mean and the ensemble standard deviation as inputs for probabilistic temperature forecasts. Working paper, arXiv: physics/0310059.
Jewson S. 2003b. Do probabilistic medium-range temperature forecasts need to allow for non-normality? Working paper, arXiv: physics/0310060.
Jewson S. 2003c. Moment based methods for ensemble assessment and calibration. Working paper, arXiv: physics/0309042.
Jewson S. 2003d. A note on the use of the word ‘likelihood’ in statistics and meteorology. Working paper, arXiv: physics/0310020.
Jewson S. 2003e. Use of the likelihood for measuring the skill of probabilistic forecasts. Working paper, arXiv: physics/0308046.
Jewson S, Caballero R. 2003. The use of weather forecasts in the pricing of weather derivatives. Journal of Applied Meteorology 40(4): 377–389.
Lehmann E, Casella G. 1998. Theory of Point Estimation. Springer-Verlag: Berlin.
Leith C. 1974. Theoretical skill of Monte Carlo forecasts. Monthly Weather Review 102: 409–418.
Mason I. 1982. A model for assessment of weather forecasts. Australian Meteorological Magazine 30: 291–303.
Mason S, Graham N. 1999. Conditional probabilities, relative operating characteristics, and relative operating levels. Weather Forecasting 14(5): 713–725.
Molteni F, Buizza R, Palmer T, Petralignis T. 1996. The ECMWF ensemble prediction system: methodology and validation. Quarterly Journal of the Royal Meteorological Society 122: 73–119.
Mullen S, Buizza R. 2000. QPF over the US by the ECMWF EPS. Monthly Weather Review 129: 638–663.
Murphy J, Winkler R. 1987. A general framework for forecast verification. Monthly Weather Review 115: 1330–1338.
Palmer T, Tibaldi S. 1988. On the prediction of forecast skill. Monthly Weather Review 116(December): 2453–2480.
Richardson D. 2000. Skill and relative economic value of the ECMWF ensemble prediction system. Quarterly Journal of the Royal Meteorological Society 649–668.
Roulston M, Smith L. 2002. Evaluating probabilistic forecasts using information theory. Monthly Weather Review 130: 1653–1660.
Roulston M, Smith L. 2003. Combining dynamical and statistical ensembles. Tellus A 55: 16–30.
Stephenson D, Doblas-Reyes P. 2000. Statistical methods for interpreting monte-carlo ensemble forecasts. Tellus A 52: 300–322.
Swets J. 1988. Measuring the accuracy of diagnostic system. Science 240(June): 1285–1293.
Talagrand O, Vautard R, Strauss B. 1997. Evaluation of probabilistic prediction systems. In Proceedings, ECMWF Workshop on Predictability, 20–22 October 1997. Available from ECMWF, Reading: 1–25.
Toth Z, Zhu Y, Marchok T. 2000. On the ability of ensembles to distinguish between forecasts with small and large uncertainty. Weather Forecasting 15: 436–447.
Wilson L, Burrows W, Lanzinger A. 1999. A strategy for verification of weather element forecasts from ensemble prediction systems. Monthly Weather Review 127: 956–970.
Ziehmann C. 2000. Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. Tellus A 52: 280–299.