The Benefits of Multianalysis and Poor Man’s Ensembles

Neill E. Bowler, Alberto Arribas, and Kenneth R. Mylne

Met Office, Exeter, United Kingdom

(Manuscript received 18 September 2007, in final form 18 February 2008)

ABSTRACT

A new approach to probabilistic forecasting is proposed, based on the generation of an ensemble of equally likely analyses of the current state of the atmosphere. The rationale behind this approach is to mimic a poor man’s ensemble, which combines the deterministic forecasts from national meteorological services around the world. The multianalysis ensemble aims to generate a series of forecasts that are both as skillful as each other and the control forecast. This produces an ensemble mean forecast that is superior not only to the ensemble members, but to the control forecast in the short range even for slowly varying parameters, such as 500-hPa height. This is something that it is not possible with traditional ensemble methods, which perturb a central analysis.

The results herein show that the multianalysis ensemble is more skillful than the Met Office’s high-resolution forecast by 4.5% over the first 3 days (on average as measured for RMSE). Similar results are found for different verification scores and various regions of the globe. In contrast, the ensemble mean for the ensemble currently run by the Met Office performs 1.5% worse than the high-resolution forecast (similar results are found for the ECMWF ensemble). It is argued that the multianalysis approach is therefore superior to current ensemble methods. The multianalysis results were achieved with a two-member ensemble: the forecast from a high-resolution model plus a low-resolution perturbed model. It may be possible to achieve greater improvements with a larger ensemble.

1. Introduction

Ensemble forecasting has its roots in attempts to understand the limits of deterministic prediction of the atmospheric state (Lewis 2005). By running a number of forecasts from a set of initial conditions, which are consistent with our knowledge of the current state of the atmosphere, we hope to gain an insight into the uncertainty in the forecast. Generally, this has been performed by creating a set of perturbations to add to a given best guess (or analysis) of the current state of the atmosphere (Toth and Kalnay 1993; Buizza and Palmer 1995).

An additional benefit of ensemble forecasting is that the ensemble mean forecast typically outperforms a forecast based on a single run of a numerical model. The latter forecasts are often described as “deterministic” forecasts. Because each ensemble member has a different realization of certain less-predictable small-scale features, the ensemble mean forecast will not contain such features, because these have been averaged out. This averaging is a curse as well as a blessing, because it means that the ensemble mean forecast will become increasingly smooth as the forecast progresses and the uncertainty increases. Thus, one needs to be very careful how the ensemble mean forecast is used (Smith 2003). This means that the ensemble mean is of little use on its own, and it is often supplemented by the probability of various events occurring derived from the whole ensemble. Nonetheless, any improvement to the ensemble mean forecast has a large effect on the quality of the ensemble forecast (Buizza et al. 2005).

The size of an ensemble is typically much smaller than the number of degrees of freedom in a numerical model [the number of grid points of an operational numerical model is currently $O(10^8)$]. This means that the focus in ensemble forecasting has been to choose perturbations to the deterministic analysis, which grow very rapidly. The two schemes initially used for medium-range forecasting are error breeding (Toth and Kalnay 1993) and singular vectors (Buizza and Palmer...
Results from a poor man’s ensemble are shown in Richardson (2001; Mylne et al. 2002) have tried to ascertain the relative merits of running an ensemble from a set of differing analyses using either a single model or multiple models (known as the multianalysis and multicenter approaches, respectively). These experiments were different from a poor man’s ensemble because initial condition perturbations generated by singular vectors were added to the analyses generated from individual centers. Thus, the singular vector perturbations could attempt to counter any lack of spread in a poor man’s ensemble. The results showed that both the multicenter and multianalysis ensembles performed very well, generally giving better performance than the ECMWF ensemble (which is based on singular vectors alone). At this time the ECMWF ensemble was heavily underspread in the short range, and Richardson (2001) performed a further test that included “evolved” singular vector perturbations to the initial condition. This ensemble was nearly as skillful as the multianalysis ensemble.

There are currently a number of regional EPSs that are similar in spirit to either a poor man’s or multianalysis ensemble (Tracton et al. 1998; Stensrud et al. 1999; Garcia-Moya et al. 2007; Eckel and Mass 2005). Notable among these is the ensemble run by the Instituto Nacional de Meteorología (INM; Garcia-Moya et al. 2007). Initial conditions for this ensemble are given by analyses from the Met Office, ECMWF, Deutscher Wetterdienst (DWD), and National Centers for Environmental Prediction (NCEP). Five different forecast models are run from each of these starting conditions, with these models being provided by the Met Office, the University Corporation for Atmospheric Research (UCAR), DWD, the HIRLAM consortium, and the Consortium for Small-Scale Modeling (COSMO). This results in an ensemble containing 20 ensemble members. Despite the obvious difficulties in maintaining five separate numerical models and receiving data from four different centers, the results have been impressive, with good ensemble spread and low error in the ensemble mean forecast. Stensrud et al. (1999) developed an ensemble based on a combination of differing analyses, error-breeding perturbations, and two different forecast models. They reported a comparison between a high-resolution forecast and the ensemble mean of an ensemble of forecasts formed from one of the models. For many variables the ensemble mean had a lower mean absolute error (MAE) than the high-resolution forecast, such as temperature and wind speed at 700 hPa and above. However, for other variables (notably 500-hPa height) the high-resolution model had a lower MAE.

Results from a poor man’s ensemble are shown in
These results are taken from data calculated by Arribas et al. (2005), but not shown in that paper. This shows the RMSE of the ensemble mean forecast for both the ECMWF and a poor man’s ensemble (consisting of the six forecast models that were most readily available at the time). Also shown is the RMS spread for each ensemble. The forecasts are verified against ECMWF analyses over Europe. Verifying against ECMWF analyses, rather than independent observations, artificially reduces the RMSE of the ECMWF ensemble, particularly in the short range. At all lead times the ensemble mean of the poor man’s ensemble is superior to the ensemble mean from the ECMWF ensemble, with a reduction in the RMSE typically in excess of 10%. This result does not necessarily imply that the poor man’s ensemble is better than the ECMWF ensemble (Buizza et al. 2003). From Fig. 1 it is clear that the poor man’s ensemble is underspread (the spread of the ensemble is less than the error of the ensemble mean).

There are a number of practical problems that hinder the widespread use of poor man’s ensembles: There are issues related to the difficulty of transferring large amounts of data between NMSs, there is reluctance for NMSs to base their operational forecast output on the output of other NMSs over which they have no control, and there is also a feeling that with a poor man’s ensemble “we don’t know what we’re doing.” However, probably most important is that poor man’s ensembles are severely limited in the number of ensemble members that are available.

The skill of the ensemble mean forecast, and its comparison with a forecast from a high-resolution deterministic system, has been considered in the context of traditional (single analysis) ensembles. Rodwell (2006) looked at the anomaly correlation coefficient (ACC) for the ECMWF high-resolution and ensemble mean forecasts. The high-resolution forecast was significantly more skillful than the ensemble mean for forecasts of 500-hPa height for the first 5 days and for the forecasts of potential temperature on the surface with PV = 2 for the first 3 days. For forecasts of 24-h-accumulated precipitation the high resolution was not more skillful, and the ensemble mean forecast was significantly more...
skillful at day 2 and beyond. The variation in the results for different forecast variables relates to the validity of the linearity assumption, which is discussed below.

b. Why are poor man’s ensembles better?

An important issue that has been discussed recently by Palmer et al. (2006) is the relationship between the mean square error (MSE) of an ensemble member forecast and the ensemble mean forecast. This relationship is

\[ \text{MSE}(\text{member}) = \text{MSE}(\text{mean}) + \text{MSS}, \]

where MSS denotes the spread of the ensemble. For a well-calibrated ensemble the spread will equal the error of the ensemble mean forecast (on average), giving

\[ \text{MSE}(\text{member}) = 2\text{MSE}(\text{mean}). \]

For traditional ensembles, which are generated by adding perturbations centered around a high-resolution analysis, the ensemble mean and control forecast are very similar on large scales in the short range (0–3 days). This is related to the validity of the linearity assumption used in four-dimensional variational data assimilation (4DVAR) and in calculating singular vectors (Gilmour et al. 2001); that is, the perturbations are sufficiently small that their evolution can be treated as linear. Note that the validity of the linearity assumption is not directly dependent on the method chosen to generate the ensemble perturbations. Provided that the perturbation magnitude is sufficiently small and the time scales are sufficiently short, then this assumption will hold; we would expect this to be true for most ensemble forecasts of broad-scale variables (such as 500-hPa height). For ensembles that are initially centered around the control analysis, they will remain centered around the control forecast for as long as this assumption holds, which is typically in the short range for large scales. Thus, a good approximation for these ensembles is

\[ \text{MSE}(\text{member}) \approx 2\text{MSE}(\text{control}). \]

The implication of this is that on average the ensemble members are always much less skillful than the control forecast. However, it would be wrong to assume that this is an unavoidable characteristic of ensemble forecasting. It is a consequence of degrading the best-guess analysis by adding perturbations around it.

In many ways the traditional approach of adding perturbations to a best-guess analysis is an appropriate strategy if only a single data assimilation cycle is available. However, the degradation of the perturbed forecasts hinders the interpretation of traditional ensemble forecasts. A common method for presenting ensemble information is via “postage stamp” charts. These charts display the forecasts from each ensemble member side-by-side for a particular area. The interpretation has been that any of the scenarios presented could occur, and are equally likely to occur. However, because the ensemble members are degraded forecasts relative to the control, the control forecast is more likely to be close to the truth than any one of the other ensemble members. This effect is largest for forecasts covering a large area at short range; for point forecasts at long range the chance that the control forecast is better than any other ensemble members is reduced (Palmer et al. 2006). This makes the job of a forecaster difficult, because they are often required to combine a set of forecasts that are of unequal skill.

In contrast to traditional ensembles, a poor man’s ensemble does not generate perturbations that are degraded relative to a control forecast. We can understand why a poor man’s ensemble performs well by looking at the skill of the ensemble mean with respect to the control forecast. Equation (1) may be rearranged to give

\[ \text{MSE}(\text{mean}) = \text{MSE}(\text{member}) - \text{MSS}. \]

In a poor man’s ensemble each member is approximately as skillful as a control forecast from a traditional ensemble. This means that the ensemble mean forecast from a poor man’s ensemble will have lower RMSEs than any of the forecasts from which the ensemble is composed. Each of the forecasts derives from an analysis produced independently by different NMSs; none of them is degraded with respect to the control forecast. Model and observation uncertainties are represented by the diversity of approaches used at different NMSs. The differences in the forecasts serve to create the spread that ensures that the ensemble mean forecast is better than any of the contributing forecasts individually. The attribute of having different forecasts of similar skill is central to the success of poor man’s ensembles. Because none of the forecasts have been substantially degraded relative to each other we may say that the poor man’s ensemble has generated nearly the correct spread “for free” (see Fig. 1). Further explanation of the relationship between the error of the ensemble mean and control forecasts is given in the appendix.

The above analysis has focused on the performance of the ensemble mean in terms of MSE (or equivalently RMSE). However, as discussed in section 4, the use of RMSE as a verification score can cause problems, because smoothing the forecast fields can result in a reduction of the forecast error. However, the RMSE of the ensemble mean is the natural quantity to consider because the spread of an ensemble is normally tuned.
to match the RMSE of the ensemble mean. A reduction in the RMSE of the ensemble mean should be accompanied by a reduction in the ensemble spread, and would combine to give a reduction (improvement) in the Brier score of the ensemble forecast, which is the main measure of ensemble quality. Additionally results for the MAE and anomaly correlation coefficient are presented in section 4.

2. Proposal for a new approach

Inspired by the performance of poor man’s ensembles, we propose a multianalysis approach for generating ensemble forecasts. The aim is to mimic the performance of a poor man’s ensemble and calculate an ensemble mean forecast that is more skillful than the deterministic, high-resolution forecast even in the short range. To improve the ensemble mean the approach illustrated in Eq. (4) is followed. The aim is to produce a set of analyses from which the forecasts are as different as possible from each other, while minimizing the degradation in quality of each forecast. This may result in a reduction in spread of the ensemble relative to a traditional ensemble, but at this stage the ensemble spread is treated as a secondary concern, and will be discussed later. The differences in the analyses will therefore contribute to generating an ensemble mean that is more skillful than any of the ensemble members, including the control forecast.

This proposal is different from other methods for generating an ensemble of analyses (Houtekamer et al. 1996; R. Buizza and M. Fisher 2007, personal communication), which include perturbations to the observations. At this stage the sole aim is to improve the ensemble mean performance, not to create a reliable ensemble.

Data assimilation framework

To shed light on the relationship between the multianalysis approach and other ensemble methods, we consider the framework provided by the EnKF; (Evensen 1994). The update equation of the Kalman filter is the same as the equation that is solved by variational methods (such as 3DVAR and 4DVAR; see Lorenc 2003), although these use an approximate solution. For the EnKF the ensemble mean analysis is calculated according to

\[
\tilde{x}_a = \tilde{x}_f + P^f H^T (H P^f H^T + R)^{-1} (y - H \tilde{x}_f),
\]

where \( \tilde{x}_f \) is the ensemble mean forecast from the previous cycle, \( P^f \) is the forecast error covariance matrix, \( H \) is the observation operator (here considered linear), \( R \) is the observation error covariance matrix, \( T \) denotes the matrix transpose, and \( y \) is the current observations. When updating each ensemble member it is necessary to perturb the observations to maintain sufficient spread in the ensemble.

The analogy with a poor man’s ensemble is that a different, static \( P^f \) is used for each member. Errors in the forecast model are accounted for by using a different forecast model for each ensemble member. The observations are not perturbed, but observation errors are accounted for by the fact that each model will use a slightly different set of observations and different observation operators. Because of the myriad differences between the forecasts and the chaotic nature of the atmospheric system, the analyses of each center do not converge to the same solution, even though they are all attempting to solve the same problem.

3. Description of tests

a. Setup of multianalysis system

In the tests that have been run, an N216 forecast (0.83° × 0.55° resolution, approximately 60 km in the midlatitudes) has been used in conjunction with the 4DVAR system run at N108 (1.67° × 1.11° resolution). This is compared with the forecasts from the operational suite, which are run at N320 (0.56° × 0.37° resolution, approximately 40 km in the midlatitudes). All of the forecasts use 50 vertical levels and are performed for data times between 0000 UTC 10 May and 1200 UTC 19 May 2006. The forecasts are all initialized from the analysis of the operational suite that is valid at 0000 UTC 6 May, allowing 4 days for each system to spin up.

The setup of the forecast and analysis experiments (apart from resolution) is the same as that for the high-resolution deterministic global model, which became operational on 14 March 2006. The system is based on the standard Met Office 4DVAR trials suite developed by M. Thurlow (2006, personal communication). This includes an observation processing system that performs quality control on the observations and calculates an estimate of the observation error. Atmospheric data assimilation is performed using the 4DVAR method (Rawlins et al. 2007). The linear model used in this scheme has a simplified parameterization of surface friction, boundary layer processes, and large-scale condensation. The minimization uses an extra term in the cost function to penalize the production of gravity waves, known as the Jc term (Gauthier and Thepaut 2001). Soil properties (including soil moisture content) are reset to climatological values on a weekly basis and are allowed to run freely between.
surface temperature is based on an analysis correction scheme (Lorenc et al. 1991) and is performed on a daily basis. The analysis of sea ice extent is performed by NCEP, using the method of Grumbine (1996), on a daily basis.

In these experiments only very small ensembles are considered, because of practical considerations. The question of the optimal ensemble size has been addressed by a number of authors (Leith 1974; Du et al. 1997; Toth and Kalnay 1997). Generally ensemble sizes in excess of eight appear to be necessary for the best performance of ensemble mean forecasts, although Toth and Kalnay (1997) concluded that a number greater than 40 may be beneficial. Thus, it may be possible to achieve improved results to the ones reported here by using a greater number of ensemble members.

b. Perturbation strategies for the data assimilation cycle

A number of small differences in the data assimilation and forecast cycles were used to perturb the analyses produced. These perturbations are designed to produce differences in the analyses, but without degrading their quality. The control forecast was based on the standard suite as used by the high-resolution forecast with no perturbations. This provides a baseline measure of the effect of changing the resolution of the forecast model; note, however, that the atmospheric data assimilation is still calculated at N108 resolution for all runs. Zhu and Thorpe (2007) found that model resolution was an important source of model uncertainty, so the differences between the control and the high-resolution forecasts may be substantial.

Ensemble member 1 was created by using the same suite as the control forecast, but introducing a small perturbation to the analysis produced at 0600 UTC 6 May 2006. This perturbation was based on the differences between the analysis at 0000 and 0600 UTC. This test was designed to demonstrate that two identical cycles with slightly different starting conditions do not converge, but do produce very similar forecasts.

Ensemble member 2 is the same as the control cycle, but a random component was applied to the thinning of satellite observations. Because satellite observations are more dense than can be assimilated, they are routinely thinned to a grid with spacing of approximately 154 km. Normally the observations that are closest to the points of the 154-km grid are assimilated, but we are free to choose an observation that is not the closest to the grid points. Member 2 uses a random choice of observation within each 154-km grid box, rather than closest. If the errors of the observations within a grid box are independent, then one would expect this forecast to have substantial differences from the control forecast.

The background error covariance matrix used by the 4DVAR scheme is determined using the “National Meteorological Center” (NMC) method of Parrish and Derber (1992). The covariance matrix used operationally is based on the difference between two forecasts that are valid at the same time, but for forecast lead times of $T + 6$ and $T + 30$; these are referred to as the early covariances. The late covariances are based on the differences between $T + 24$ and $T + 48$ forecasts. These late covariances were used operationally before June 2005, at which point the assimilation switched to using the early covariances. Ensemble member 3 uses the late covariances and random thinning of satellite observations.

Ensemble member 4 is the same as the control forecast, but a 3DVAR analysis scheme (Lorenc et al. 2000) is used instead of a 4DVAR scheme. This scheme uses a time interpolation between model states within the assimilation window to calculate the model forecast of the observation at the observation time [this is the so-called 3D first guess at appropriate time (3D-FGAT) method]. This is in contrast to the basic 3DVAR, which assumes that all observations are made at the middle of the assimilation window. The differences between the model forecast and observed values are then used in calculating increments to apply to the model state in the middle of the assimilation window. In 4DVAR a linear model is used to propagate each increment back to the start of the assimilation window. The Met Office used the 3DVAR scheme operationally before 5 October 2004. The forecasts from the 3DVAR analysis are expected to produce worse forecasts than the control, but they are much cheaper computationally.

The next three ensemble members use a new scheme that perturbs the calculated departure points in the semi-Lagrangian advection scheme. For each time step in the forecast model the value of a variable at a given grid point is calculated by estimating the location from which the fluid at that point would have come. This is illustrated for a 2D grid in Fig. 2a. For a 3D grid, at each time step the value of a variable at a given point is interpolated from the values at the eight grid points nearest to the estimated departure point. The interpolation of each time step creates a repeated filtering of the fields being forecast. Removing the effect of this interpolation from ensemble and climate forecasts has been the subject of study recently (Shutts 2005; Jung et al. 2005). The schemes developed thus far have relied on adding vorticity perturbations to undo the effect of the interpolation. However, in this study we have per-
(a) The semi-Lagrangian advection scheme interpolates the value at a given point from the four nearest points upstream of the point being considered. (b) In the perturbed scheme one of the four points is chosen with a probability related to the area weighting of the boxes shown, and the departure point is moved a factor $\alpha$ closer to this point (plus sign). The bottom-left point would be selected with a probability $(1 - L_z)(1 - L_x)$ so the points nearest the standard departure point are most likely to be selected.

In this scheme the departure point perturbation is restricted to the 2D case; this means considering interpolation from the four grid points that are nearest to the calculated departure point. The semi-Lagrangian scheme still interpolates in three dimensions, but the perturbation is restricted to the horizontal dimension. Rather than interpolating the field at a given location from the four grid points nearest to the calculated departure point, it is possible to take the value from one of the four grid points directly. This would be equivalent to moving the calculated departure point to rest exactly at one of the four possible points. However, such a scheme would likely cause the model forecast to fail because it would introduce large amounts of noise into the advection scheme. An alternative is to move the calculated departure point closer to one of the four grid points. This allows the introduction of a tunable parameter $\alpha$ (see Fig. 2b), which can control the strength of the perturbation. Here, $\alpha = 0$ corresponds to no perturbation, and $\alpha = 1$ corresponds to moving the departure point to rest exactly on one of the four grid points.

The scheme implemented in this paper randomly chooses one of the four grid points in order to reduce the effect more directly.

To summarize the performance of the ensemble mean forecast over a number of forecast variables and lead times, an index of scores has been devised. The weights given to the forecast variables are inspired by the global NWP index that is used at the Met Office to summarize deterministic model performance. The index is compiled by dividing the forecast error from each system by the forecast error of the high-resolution forecast. These values are averaged over all lead times to find an average, normalized forecast error for each forecast variable. This average forecast error is computed between the original departure point and the chosen grid point. The random numbers used to choose the grid point toward which to move the departure point have no spatial or temporal correlation. Thus, the perturbed departure point scheme will introduce noise at the grid scale of the model. This is different from the approach taken by others (Shutts 2005) who have used a spatial and temporal correlation to a vorticity perturbation. The perturbed departure point scheme is used in the forecast model only, and is not applied to the linear model used in 4DVAR.

Ensemble members 5 and 6 use the same setup as ensemble members 2 and 3, respectively, but in addition they use the perturbed departure points scheme. These members are expected to produce worse forecasts than the members without the perturbed departure points scheme. However, it may be possible that the extra spread created by using these members means that they can make a positive contribution to the quality of the ensemble mean forecast. Ensemble member 7 uses the same setup as ensemble member 6, but uses a different (lower) weight given to the $J_c$ filtering term (Gauthier and Thepaut 2001). The standard value of these weights is $600$ and $10^{10}$, where the penalty is applied to the full fields or tendency fields. Ensemble member 7 uses the values of $1/10$th this size. This filtering term is used to suppress undesirable gravity waves in the data assimilation process, so ensemble member 7 may have more evidence of gravity waves in its analysis.

A summary of the configuration used by each ensemble member is shown in Table 1.

c. Score calculation

To summarize the performance of the ensemble mean forecast over a number of forecast variables and lead times, an index of scores has been devised. The weights given to the forecast variables are inspired by the global NWP index that is used at the Met Office to summarize deterministic model performance. The index is compiled by dividing the forecast error from each system by the forecast error of the high-resolution forecast. These values are averaged over all lead times to find an average, normalized forecast error for each forecast variable. This average forecast error is computed...
combined with the sum calculated for the other forecast variables, using the weights shown in Table 2. Thus, the value for the index is unity for the high-resolution forecast and values greater than (less than) one indicate that the forecast is less skillful (more skillful) than the high-resolution forecast.

The rationale behind the weights used in the index is as follows: 500-hPa height represents the overall circulation forecast by the model, and so it should receive high weight. Other variables that are near the surface should also receive high weights, because surface weather is important to many Met Office customers. Additionally, wind at high levels should receive a high weight, because the aviation industry is one of the key customers of the Met Office. Thus, the index is designed to give an overall picture of forecast quality, weighted toward the main customers of the Met Office.

All of the forecasts presented in this paper are verified against radiosonde observations from across the globe that are valid at 0000 and 1200 UTC between 10 and 22 May 2006. Because there are more radiosondes in the Northern Hemisphere than elsewhere, the statistics will be biased toward the ensemble performance in the Northern Hemisphere.

For the graphs summarizing the relative improvement of a forecast system over the high-resolution forecast, confidence intervals have been calculated. These error bars are 1.96 times standard deviation estimates of each value, estimated from the standard deviation of the component RMS errors, and using an estimate of the number of spatial and temporal degrees of freedom in a similar set of forecasts. These error bars represent 95% confidence intervals on the improvement in the ensemble mean forecast. The number of degrees of freedom is derived using the Z method, as described by Wang and Shen (1999).

To transform the anomaly correlation into a score similar to the RMSE, the calculation involves the square root of one minus the anomaly correlation coefficient. Using \( \sqrt{1 - \text{ACC}} \) changes the ACC from a skill score (positively oriented) to a penalty function (negatively oriented). The square root is necessary because the ACC involves the covariance between the forecast and observed anomaly, and the RMSE is the square root of a variance-like term. For the climatology, an in-sample climatology has been used. The observed values for each variable were plotted against latitude and a fourth-order polynomial fitted to these points. This polynomial fit was then used to give the climatology. This calculation of the climatology is rather crude, and the results for ACC may vary if a more sophisticated climatology is used.

### 4. Results

#### a. Overall results for each member

Relative values for the index of forecast variables for all the experiments described in section 3b are shown in Table 3. The value of the index for each combination has been normalized by the value of the index for the high-resolution forecast. Each of the forecasts are approximately as skillful as each other (to within 1.5%). Because the premise of improving the ensemble mean depends on having forecasts that are approximately equally skillful, this is a reassuring result. The table also shows the root-mean-square differences between the forecast members. All of the low-resolution forecasts have large differences with the high-resolution forecast, indicating that the change of horizontal resolution is an important component in generating spread in a multi-

<table>
<thead>
<tr>
<th>Variable</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height at 500 hPa</td>
<td>25%</td>
</tr>
<tr>
<td>Temperature at 850 hPa</td>
<td>15%</td>
</tr>
<tr>
<td>Temperature at 500 hPa</td>
<td>10%</td>
</tr>
<tr>
<td>Temperature at 250 hPa</td>
<td>5%</td>
</tr>
<tr>
<td>Wind speed at 850 hPa</td>
<td>15%</td>
</tr>
<tr>
<td>Wind speed at 500 hPa</td>
<td>10%</td>
</tr>
<tr>
<td>Wind speed at 250 hPa</td>
<td>20%</td>
</tr>
</tbody>
</table>
The RMSE, MAE, and the square root of one minus the ACC are shown.

Index of forecast variables. The numbers in this table have been normalized by the error of the forecasts from the high-resolution model. The numbers in this table have been normalized by the error of the forecasts from the high-resolution model.

The next most important differences are created by the perturbed departure points scheme. Although this scheme degrades the forecast, the extra difference that it creates with the reference forecasts means that the ensemble mean of the high-resolution forecast and member 7 is the most skillful forecast combination. Surprisingly, the control forecast appears to be more skillful than the high-resolution forecast. Although this result is not statistically significant, it appears to be largely due to the better forecasts of the low-resolution model of temperature and wind speed at 500 hPa, and wind speed at 250 hPa.

The member that uses a 3DVAR analysis (member 4) gives a noticeably worse forecast than the control, demonstrating the benefits of the more sophisticated analysis scheme. However, because there are substantial differences between this forecast and the control forecast (around 31% of the RMSE of the high-resolution forecast) this degradation is not reflected in a degradation of the ensemble formed by member 4 and the high-resolution forecast. The differences between member 1 and the control forecast are small, indicating that two parallel assimilation cycles on average produce very similar forecasts. However, the feedback of a different forecast model onto the analysis is significant. The difference between the member using the control analysis and the high-resolution forecast model (member 8) and the high-resolution forecast is much larger than the difference between member 8 and the control forecast. Therefore, it is the differences in the analyses of the high-resolution and control forecasts that are most important, not the differences in the forecast model.

Aside from the results for the RMSE, Table 3 also shows results for the MAE and ACC. For all three scores the combination of the high-resolution forecast with member 7 is judged to be the most skillful, with between 3.8% and 4.7% improvement in skill relative to the high-resolution forecast alone. Because the MAE and ACC are less susceptible to being improved by the smoothing of forecasts, these results give confidence that the improvements in the ensemble mean forecast are not simply due to smoothing caused by the averaging.

A further test of the robustness of the forecast improvement, relative to smoothing, was performed. A simple smoothing (box averaging with the length of the sides between 3 and 15 grid points) was applied to the deterministic forecast. This gave a maximum reduction of the RMSE for the index of forecast variables of 2.6%. This suggests that the averaging of the ensemble mean is more beneficial than can be achieved by a simple smoothing, which supports the results found by Toth and Kalnay (1997).

In addition to these results, which are based on the global network of radiosondes, results have been calculated for various regions of the globe. The subregions that have been used are the Northern Hemisphere extratropics (20°–90°N), the tropics (20°S–20°N), and the Southern Hemisphere extratropics (90°–20°S). The results for the Northern Hemisphere are very similar to the global results, because most of the global radiosondes are in this region. For the Southern Hemisphere the low-resolution model is more skillful than the high-resolution model, and the difference between the two forecasts is greater than for forecasts in the Northern Hemisphere. Both of these facts combine to indicate that the improvement in skill from using a two-member ensemble-mean over the high-resolution forecast is around 5%–6% (as opposed to 3.5%–4.5% for the globe). The reverse is true in the tropics, where the low-resolution forecasts are less skillful than the high-resolution forecasts and the differences between the forecasts at the two resolutions are less than those for the global results. The two-member ensemble mean forecasts are still more skillful than the high-resolution forecasts, but only by 1%–2%. Generally, the distribution of changes in skill between the different forecast

---

**Table 3. Summary of the results for each ensemble member configuration and the differences between each member in terms of the index of forecast variables.** The numbers in this table have been normalized by the error of the forecasts from the high-resolution model. The RMSE, MAE, and the square root of one minus the ACC are shown.

<table>
<thead>
<tr>
<th>Ensemble member</th>
<th>High resolution</th>
<th>Control</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>1.000</td>
<td>0.991</td>
<td>0.989</td>
<td>0.991</td>
<td>0.987</td>
<td>1.004</td>
<td>1.015</td>
<td>1.013</td>
<td>1.017</td>
<td>1.009</td>
</tr>
<tr>
<td>RMSE (average with high resolution)</td>
<td>—</td>
<td>0.966</td>
<td>0.964</td>
<td>0.964</td>
<td>0.959</td>
<td>0.968</td>
<td>0.957</td>
<td>0.956</td>
<td>0.953</td>
<td>0.970</td>
</tr>
<tr>
<td>RMS difference high resolution</td>
<td>—</td>
<td>0.331</td>
<td>0.334</td>
<td>0.344</td>
<td>0.358</td>
<td>0.359</td>
<td>0.434</td>
<td>0.436</td>
<td>0.433</td>
<td>0.363</td>
</tr>
<tr>
<td>RMS difference control</td>
<td>0.331</td>
<td>—</td>
<td>0.104</td>
<td>0.131</td>
<td>0.175</td>
<td>0.308</td>
<td>0.297</td>
<td>0.303</td>
<td>0.299</td>
<td>0.232</td>
</tr>
<tr>
<td>MAE</td>
<td>1.000</td>
<td>0.994</td>
<td>0.993</td>
<td>0.995</td>
<td>0.992</td>
<td>1.005</td>
<td>1.025</td>
<td>1.024</td>
<td>1.018</td>
<td>1.015</td>
</tr>
<tr>
<td>MAE (average with high resolution)</td>
<td>—</td>
<td>0.966</td>
<td>0.965</td>
<td>0.964</td>
<td>0.960</td>
<td>0.968</td>
<td>0.960</td>
<td>0.959</td>
<td>0.956</td>
<td>0.971</td>
</tr>
<tr>
<td>ACC score</td>
<td>1.000</td>
<td>0.998</td>
<td>0.997</td>
<td>0.999</td>
<td>0.995</td>
<td>1.011</td>
<td>1.028</td>
<td>1.025</td>
<td>1.019</td>
<td>1.015</td>
</tr>
<tr>
<td>ACC score (average with high resolution)</td>
<td>—</td>
<td>0.971</td>
<td>0.970</td>
<td>0.970</td>
<td>0.965</td>
<td>0.974</td>
<td>0.967</td>
<td>0.965</td>
<td>0.962</td>
<td>0.975</td>
</tr>
</tbody>
</table>

---
variables is similar across the regions, except for wind speed at 850 hPa. In the northern and southern extratropics the ensemble mean forecasts are clearly more skillful than the high-resolution forecasts, which is not the case in the tropics. The results for the southern extratropics and the tropics are not statistically significant. The improvement of the ensemble mean forecasts over the high-resolution forecasts for the globe is statistically significant at the 2.5% level.

These results demonstrate that a two-member ensemble (the high-resolution forecast plus a low-resolution forecast of similar quality) can produce an ensemble mean forecast that is around 4% more skillful than either forecast individually. To examine the performance of the forecasts more closely, such as looking at the performance against different forecast variables, we choose the forecasts from member 7 as the combination of this forecast with the high-resolution forecast that performs well for the above verification.

b. Small ensemble results

The following results are based on the ensemble containing the high-resolution forecast member 7. Similar results are seen combining the high-resolution forecast with forecasts from other members. It is possible to achieve slightly better results by adding another member to this ensemble (a combination of the high-resolution member 3 and 7 forecasts gives best results), but these will not be reported here. Adding even more ensemble members does not improve the forecast, which is different from the results of other studies (Du et al. 1997; Toth and Kalnay 1997). In this study a major source of forecast differences is the resolution of the forecast model, and only one run at the high-resolution is available. Therefore, having a larger ensemble, with more low-resolution forecast members, would reduce the ensemble spread, leading to worse results. This suggests that even better results may be achieved if a larger ensemble with a number of high- and low-resolution forecasts was available.

Figure 3 shows the RMSE as a function of lead time for forecasts of 500-hPa height from the high-resolution model, member 7, and the ensemble mean. Also shown is the RMS spread of this two-member ensemble. It can be seen that the forecast from the high-resolution model is more skillful than that of member 7. This improvement is not due to the resolution change, because the control forecast performs as well as the high-resolution forecast in the extratropics (not shown) but it is due to the perturbed departure points scheme. The ensemble mean forecast is more skillful than either of the individual ensemble members. The improvement of the ensemble mean forecast over the high-resolution forecast varies between none (at $T + 12$) and 3.7% (at $T + 72$).

The ability to produce a forecast that has a lower RMSE for 500-hPa height than the high-resolution forecast in the early part of the forecast period is not trivial. Rodwell (2006) found that the ECMWF high-resolution forecast had a significantly higher anomaly correlation coefficient than their ensemble mean forecast for the first 5 days. We know of no other system, aside from a poor man’s ensemble, that can produce forecasts of 500-hPa height that are more skillful than the high-resolution model.

Similar results to Fig. 3, but for wind speed and temperature at 850 hPa, are shown in Figs. 4 and 5, respectively. At these lower levels of the atmosphere the high-resolution model would be expected to perform better than the lower-resolution models, because the effect of variations in surface height will be more noticeable at this level. This is reflected in both figures, for which the
low-resolution forecast performs worse. However, in both cases the ensemble mean forecast has lower errors than the high-resolution forecast.

For all of the variables shown in Figs. 3–5 the spread of the multianalysis ensemble is much less than the error in the ensemble mean forecast. This indicates that the multianalysis ensemble is unlikely to be useful without further modifications, which will be discussed later. The spread is also less than that seen with the poor man’s ensemble (cf. Figs. 1 and 3), indicating that there are some uncertainties sampled by that ensemble that are not accounted for here.

The improvements in the ensemble mean forecast for the multianalysis ensemble over a number of variables, relative to the high-resolution forecast, are shown in Fig. 6. For all variables the ensemble mean forecast is more skillful than the high-resolution forecast. The benefit of the ensemble mean varies with variable considered, but the index is over 4% higher for the ensemble mean than the high-resolution forecast. This benefit comes from the averaging in the ensemble mean forecast, and the ensemble members perform with similar skill to the high-resolution forecast.

c. Comparison with MOGREPS

How do the results of Fig. 3 compare with what can be achieved using a traditional ensemble where ensemble members are created by perturbing a central analysis? Figure 7 shows comparable results for the Met Office Global and Regional Ensemble Prediction System (MOGREPS; Bowler et al. 2008). The perturbations to the initial conditions are created using the ensemble transform Kalman filter (ETKF; Bishop et al. 2001) and the forecasts are run at N144 resolution (1.25° × 0.83° resolution, around 90 km in the mid-latitudes). The control forecast is run at N144 resolution and is started from the high-resolution analysis, but without any perturbation. From this figure a number of things are apparent. The perturbed ensemble members are much less skillful than any of the control forecast, the ensemble mean forecast, or the high-resolution forecast. One may note that the ensemble mean forecast is slightly
less skillful than the control forecast. This is believed to be due to the ensemble spread being slightly too large.

Figure 8 shows similar summary results for MOGREPS to the results shown in Fig. 6 for the multianalysis ensemble. As was seen for 500-hPa height, the ensemble mean forecast is less skillful than the high-resolution forecast. For some variables, such as 500-hPa temperature, MOGREPS performs better than the high-resolution forecast. Overall, the results are clearly worse than for the multianalysis ensemble.

5. Practical considerations

a. Dealing with the ensemble spread

In this paper, we have focused entirely on the performance of the ensemble mean forecast in terms of RMSE. Another important aspect of an ensemble prediction system is the spread of the ensemble. From Figs. 3–5 it can be seen that the multianalysis ensemble has around half the spread that would be desired for a perfect ensemble. Originally, when embarking on this study, we had hoped to produce a well-calibrated ensemble for which each forecast is almost as skillful as the high-resolution forecast. A poor man’s ensemble (see Fig. 1) is close to this goal. This, however, may be out of reach for a single-model system because of the difficulties in modeling errors in the forecast model.

Because it has now been demonstrated that it is possible to generate some ensemble spread without degrading individual ensemble members, it is reasonable to work to avoid degrading each ensemble member more than is absolutely necessary. One possible way to gain the benefits of a multianalysis ensemble, while maintaining a large ensemble spread, is to use a system similar to that tested by Evans et al. (2000). This involves generating ensemble perturbations, such as singular vectors or ETKF perturbations, which degrade the ensemble member forecast performance, and adding these to the ensemble of analyses. Because the multianalysis ensemble already has some spread, these perturbations may be smaller in amplitude than if they were added to a single analysis, meaning that the ensemble member forecasts will be more skillful when
considered individually. Another approach would be to perturb the observations used by each of the analysis schemes (Houtekamer et al. 1996). However, one would have to be confident that the perturbations to the observations did not introduce undesirable features in the assimilation.

The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE; Richardson et al. 2005) is very similar to the multicenter ensemble considered by Evans et al. (2000). However, each forecast center will calibrate its own ensemble to have approximately the same spread as the error in its ensemble mean forecast. Because the spread of a poor man’s ensemble is substantial (see Fig. 1), these perturbations may be larger than necessary. Ideally, one would reduce the size of the perturbations applied to each contributing center’s ensemble to ensure that the resulting multicenter ensemble is well calibrated.

b. Computer time comparison

One difficulty with a multianalysis ensemble is that the analysis system is computationally expensive, and running a number of analyses would require compromises in other parts of the forecasting system. Table 4 gives typical computational costs (in CPU seconds on the NEC SX6 at the Met Office) for the various forecasting system components. A standard forecast cycle consists of eight 6-h 4DVAR analyses per day (each analysis is repeated once the full set of observation data has been received). Additionally, two forecasts are run to $T + 171 \ h$, two forecasts are run to $T + 70 \ h$, and four forecasts are run to $T + 12 \ h$.

Rather than run a full forecast cycle, one might run an extra data assimilation cycle without forecasting to provide extra initial condition perturbations for an ensemble forecast. Removing the forecasting burden substantially reduces the computational cost of running such a system. However, the cost is still equivalent to over 20 ensemble members (10 members per cycle for two cycles per day). This would entail a large increase in the computational cost of the ensemble suite. Because the 3DVAR method is much less expensive than 4DVAR, it is more realistic to run such a scheme to provide extra initial conditions for an ensemble. The cost of such a system is less than four ensemble members, thus representing less than a 10% increase in the cost of ensemble forecasting at the Met Office.

Also included are tentative estimates for running an N512L90 forecast model, which would give a horizontal resolution of around 25 km in the midlatitudes (with an assumed 10-min time step, which compares with the 15-min time step used by the N320 model). This is included because ECMWF recently upgraded their high-resolution model from T511 (~40 km) to T799 (~25 km). This upgrade only improved the performance of the 500-hPa height forecast by 1% in the Northern Hemisphere (M. Miller 2007, personal communication), which is less than the improvement we report here. It should be noted that since the resolution upgrade some problems with the data assimilation system have been identified and addressed. Nonetheless, a 4% improvement in forecast quality typically represents between 1.5- and 2-yr development of the Met Office forecast system.

One extra comparison is between this system and the ensemble Kalman filter, which is the subject of much study (e.g., Houtekamer and Mitchell 2001; Ott et al. 2004). The EnKF requires a very large number of ensemble members (~100) in order to provide forecasts of comparable quality to 3DVAR systems. This is a prohibitively high cost, and is likely to remain high for the foreseeable future. It may be possible to reduce the number of ensemble members, but this would require supplementing the background error covariance matrix with one generated from a long time series of statistics (Hamill and Snyder 2000). Although four-dimensional versions of the EnKF are possible, if the EnKF uses a static background error covariance matrix, then it becomes the equivalent of a 3DVAR analysis scheme. This implies that if an EnKF is to be competitive with 4DVAR, then it must avoid the use of static background error covariances through very large ensembles. Therefore, investigating methods that can provide improved forecasts at the cost of performing a small number of additional analyses may provide a better way forward.

6. Conclusions

In this paper, results from a multianalysis ensemble have been shown. The aim of these tests has been to mimic some of the properties of a poor man’s ensemble,
in particular the improved performance of the ensemble mean forecast in the short range. This has been achieved, with the RMSE of the ensemble mean forecast less than that of the high-resolution forecast for all quantities. Most notable is that the ensemble mean forecasts of 500-hPa height are better for the multi-analysis ensemble than for the high-resolution forecast. It is not possible to achieve this improvement for short-range forecasts using traditional ensembles. Therefore, this is one respect in which an ensemble of analyses is intrinsically superior to all of the ensemble methods that depend on a single analysis. These results were achieved with a two-member ensemble: the forecast from a high-resolution model plus a low-resolution perturbed model. It may be possible to achieve greater improvements with a larger ensemble. Tests were also performed with a three-member ensemble, and it was possible to achieve a further, small increase in skill. Comparisons with a simple smoothing method applied to the deterministic forecast have shown that the multi-analysis ensemble improves the ensemble mean forecast by more than is possible by using simple smoothing alone.

By focusing on the ensemble mean performance it has been possible to generate a certain amount of spread in the ensemble without degrading the quality of individual ensemble members. For such an ensemble the interpretation of the products is made simpler, because each forecast from the ensemble is equally likely to occur.

These tests have given an insight into why a poor man’s ensemble is an effective forecasting tool. Experiments with the random thinning of observations have shown that this source of uncertainty has a small effect on the forecast. Much more significant are the choice of model resolution and the background error covariance matrix. These will affect the determination of the current analysis, but their effect will also feed back to the initial conditions through the repeated iteration of analysis cycles. The chaotic nature of the atmosphere, amplifying small differences between forecasts, serves to magnify the importance of this feedback. Therefore, we assess that the initial condition spread in a poor man’s ensemble derives from differences in model formulation, as described above, which amplify through a repeated analysis cycle. Differences in the analysis caused by differences in the resolution of the forecast model have a larger impact than differences in the resolution of the forecast model alone.

To gain the benefit from a multi-analysis ensemble, while generating an ensemble forecast with appropriate spread, it may be necessary to include initial condition perturbations that degrade the forecast. Such a setup was tested by Evans et al. (2000) and Mylne et al. (2002). It is our intention to test this approach within MOGREPS. It is envisaged to run a low-resolution (N216) 3DVAR analysis alongside the high-resolution (N320) 4DVAR analysis. Some of the MOGREPS perturbations will be added to the 3DVAR analysis, and the rest will be added to the 4DVAR analysis. Provided the spread of the perturbations is chosen appropriately, this should lead to an ensemble forecast that has a similar spread in the short range to the current system, but an improved ensemble mean forecast. This kind of system is also very similar to that of the TIGGE multi-model ensemble (Richardson et al. 2005). Given that the spread of a poor man’s ensemble is often quite close to the RMSE of the ensemble mean forecast (see Fig. 1), the initial condition perturbations in this setup will need to be quite small, which is not the case with TIGGE. Therefore, postprocessing of TIGGE forecasts may be necessary to reduce the spread of the contributing ensembles in order to achieve a well-calibrated ensemble.

The computational cost of the multi-analysis ensemble has been examined. The cost of running a separate N216L50 forecast cycle is similar to the cost increase in upgrading from a 50- to 70-level version of the N320 forecast. A 3DVAR analysis is much cheaper than this, although the forecasts produced from this system are inferior to those produced using 4DVAR. However, the forecast improvement from combining the low-resolution 3DVAR forecast with the high-resolution 4DVAR forecast is still around 3%.

In the case of a traditional ensemble it is a simple matter of choosing a priori the best member; provided that the verification region is large enough, the control will be the best forecast. In the case of a multi-analysis ensemble each member is approximately equally likely to the best member. This raises questions about the use of deterministic forecasts: how does one choose between a set of different forecasts of approximately equal skill? One may consider running a forecast from the ensemble mean analysis, but this would be a poor proxy for the ensemble mean forecast itself, because much of the benefit in the ensemble mean forecast comes from the smoothing of uncertain features. On the other hand, using the ensemble mean forecast may be seen as unacceptable because it does not contain small-scale information. One solution could be to choose the individual member closest to the ensemble mean over the area of interest, or, if identifiable, the member closest to the mode. The problem of choosing between forecasts of approximately equal skill has been present for many years because of the existence of poor man’s ensembles, though political and operational con-
straints have hindered NMSs basing their output on data from external sources. Now these constraints may be mitigated by the use of a multianalysis ensemble.

Further avenues of study include running such a system for a second case, for a longer series of forecasts. This would allow a greater level of confidence in the improvements gained from a multianalysis system. Because resolution plays a very important part in the benefits from a multianalysis ensemble, a logical step would be to test differences in the inner-loop resolution of the analysis system.

At this stage, one returns to the quote of Box (1987) “All models are wrong, but some are useful.” From the work presented above we have a better understanding of how to define useful; that is, a model is useful if it can improve the ensemble mean forecast. However, we are still a long way from understanding why an ensemble of configurations of the same model should be more useful than just one. Why can a forecast from a low-resolution version of a model add useful information to forecasts from a high-resolution version of the same model?

Acknowledgments. The authors thank the many people who have contributed to this work through discussion and supply of some supplementary code. In particular, we thank Adam Maycock, Nigel Wood, and David Thompson.

APPENDIX

A Vector Explanation of the Ensemble Mean Error

As an extra explanation of why an ensemble approach works, we consider the conditions under which an ensemble of two members has a lower RMSE than either member. Figure A1 shows a vector illustration of the errors of two forecasts. The truth is taken as the origin of the diagram; \( e_1 \) is the error of the forecast from member 1, and \( e_2 \) is the error of the forecast from member 2. From this representation it is clear that the error of the ensemble mean forecast is

\[
\left\| \frac{e_1 + e_2}{2} \right\|^2 = \frac{1}{4} \| e_1 \|^2 + \frac{1}{4} \| e_2 \|^2 + \frac{1}{2} e_1 \cdot e_2. \tag{A1}
\]

where \( \| \cdot \|^2 \) denotes the square of the length of the vector (mean square error) and \( e_1 \cdot e_2 \) is the scalar product of the two errors. Similarly, if we consider the difference between \( e_2 \) and the error of this ensemble mean (in effect the mean square spread of the two-member ensemble) then its length is given by

\[
\left\| \frac{e_2 - e_1}{2} \right\|^2 = \frac{1}{4} \| e_1 \|^2 + \frac{1}{4} \| e_2 \|^2 - \frac{1}{2} e_1 \cdot e_2. \tag{A2}
\]

Rearranging Eq. (A2) for \( e_1 \cdot e_2 \) and substituting into Eq. (A1) gives

\[
\left\| \frac{e_1 + e_2}{2} \right\|^2 = \frac{1}{2} \| e_1 \|^2 + \frac{1}{2} \| e_2 \|^2 - \left\| \frac{e_2 - e_1}{2} \right\|^2. \tag{A3}
\]

This is the expression for the error of the ensemble mean in terms of the error of the two forecasts and the difference between them. Writing this in the notation of Eq. (1) we have

\[
\text{MSE(ensemble)} = \frac{1}{2} \text{MSE(member 1)} + \frac{1}{2} \text{MSE(member 2)} - \text{MSE}. \tag{A4}
\]

![Fig. A1. The improvement in an ensemble mean forecast. The length of each vector corresponds to an RMS distance: (a) a case where the two forecasts are of similar skill, so the ensemble mean is more skillful than either; and (b) a case where member 2 is less skillful than member 1, and the ensemble mean is not more skillful than member 1.](image-url)
Now, if we require the ensemble mean forecast to have smaller errors than the errors of ensemble member 1, we must have

\[
\text{MSS} > \frac{1}{2} \text{MSE}(\text{member 2}) - \frac{1}{2} \text{MSE}(\text{member 1}).
\]  

(A5)

Thus, for the two-member ensemble to have smaller errors than either forecast individually, then the second forecast needs to be more different from the first forecast than it is worse. It is for this reason that the mean-square differences between the high- and low-resolution forecasts are reported in Table 3. This also illustrates that although the perturbed departure point scheme degrades the forecasts, it introduces even greater differences with the reference forecast, and hence improves the ensemble mean. Figure A1 shows two different cases for the possible errors and differences of two forecasts. In Fig. A1a the forecasts are of equal skill, and there is a considerable spread between the forecasts. Consequently the RMS error of the ensemble mean (represented by the length of the dashed vector) is smaller than the RMS errors of either forecast. In Fig. A1b member 2 is less skillful than member 1, and the difference between the forecasts is smaller than that for Fig. A1a. In this case the ensemble mean is not more skillful than member 1.

The condition described in Eq. (A5) explains why adding a poor-quality forecast to a poor man’s ensemble can improve the ensemble mean. Provided that the forecast is different from the other forecasts more than it is worse, then the ensemble mean will improve. For a traditional ensemble the ensemble mean is equal to the control in the short range. In this case each ensemble member is exactly as much worse compared to the ensemble mean as it is different.

The derivation reported here can be extended to cover many forecasts, where each are given unequal weights, and to include the anomaly correlation as well as the MSE. Such results may be made use of in the context of combining forecasts from the TIGGE multimodel ensemble.

REFERENCES


