



Numerical Weather Prediction

The benefits of multi-analysis and poor-man's ensembles



NWP Technical Report No. 505

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

email:nwp_publicationsmetoffice.gov.uk

The benefits of multi-analysis and poor-man's ensembles

Neill E. Bowler* , Alberto Arribas & Kenneth R. Mylne

Met Office, Fitzroy Road, Exeter, EX1 3PB, UK.

Abstract:

We propose a new approach to probabilistic forecasting, 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 met services around the world. The multi-analysis ensemble aims to generate a series of forecasts that are as skilful as each other *and* the control forecast. This produces an ensemble mean forecast which is superior not only to the ensemble members, but to the control forecast in the short range. This is something that it is not possible with traditional ensemble methods, which perturb a central analysis.

Our results show that the multi-analysis ensemble is more skilful than the Met Office's high-resolution forecast by 4% over the first three days (on average as measured for RMSE). In contrast, the ensemble mean for the ensemble currently run by the Met Office performs 2% worse than the high-resolution forecast (similar results are found for the ECMWF ensemble). We argue that the multi-analysis approach is therefore superior to current ensemble methods.

Crown Copyright © 2007

KEY WORDS ensemble forecasting short-range NWP

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. Since each ensemble member has a different realisation of certain less-predictable small-scale features, the ensemble mean forecast will not contain such features, these having been averaged-out. This averaging is a curse as well as a blessing, since 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 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, 1995). Error breeding repeatedly re-scales the differences between two runs of a numerical model. The repeated re-scaling ensures that any differences which are not rapidly growing are rapidly damped. Thus, the perturbations project on those structures which have been growing rapidly in the recent past. The singular vectors method uses a linearised version of the numerical model to calculate structures which are expected to grow rapidly in the near future.

The ensemble run by the Meteorological Service of Canada has recently upgraded to using an ensemble Kalman filter (EnKF) (Houtekamer and Mitchell, 2001). This system attempts to perturb all the inputs to the data assimilation to provide an ensemble of analyses which are consistent with the uncertainty in the best-guess analysis. Inputs such as the observations, the background forecast, the sea-surface temperature and the version of the forecast model

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

used are all perturbed. This system provides analyses which are of comparable quality to the 3D-Var data assimilation method, as measured by the root mean square error (RMSE) of 6-h forecasts against radio-sonde observations (Houtekamer *et al.*, 2005).

1.1 Poor man's ensembles

A poor-man's ensemble is one which is created by collecting the output from the deterministic forecasts from National Meteorological Services (NMSs) around the world. It is described as a poor-man's ensemble since it does not require any additional model runs to produce it. A number of studies have investigated their properties (Atger, 1999; Ziehmann, 2000; Ebert, 2001; Buizza *et al.*, 2003; Arribas *et al.*, 2005). Many of these have concluded that a poor-man's ensemble is an effective method for ensemble forecasting. For example, Atger (1999) fitted a Gaussian distribution to a very small poor-man's ensemble and found that this provided more skilful forecasts (in terms of Brier skill score for the first five days) of 500 hPa height than the ECMWF ensemble prediction system (EPS). Buizza *et al.* (2003), with a very limited poor-man's ensemble, found that it provided better forecasts than the ECMWF EPS, but that a planned upgrade to the ECMWF EPS was even more skilful than the poor-man's ensemble. Arribas *et al.* (2005) tested a hybrid ensemble with some ensemble members from the poor-man's ensemble along with some from the ECMWF EPS, as well as testing a standard poor-man's ensemble. They found that the hybrid ensemble performed best, with the poor-man's ensemble being nearly as skilful.

Other studies (Harrison *et al.*, 1995; Evans *et al.*, 2000; Richardson, 2001; Mylne *et al.*, 2002) have tried to ascertain the relative merits of running an ensemble from a set of differing analyses using a single model or multiple models (known as the multi-analysis and multi-centre approaches, respectively). These experiments were different to a poor-man's ensemble since initial condition perturbations generated by singular vectors were added to the analyses generated from individual centres. 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 multi-centre and multi-analysis 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 under-spread in the short-range, and Richardson (2001) performed a further test which included 'evolved' singular vector perturbations to the initial condition. This ensemble was nearly as skilful as the multi-analysis ensemble.

There are currently a number of regional EPSs which are similar in spirit to a poor-man's or multi-analysis ensemble (Tracton *et al.*, 1998; Stensrud *et al.*, 1999; Garcia-Moya *et al.*, 2007; Eckel and Mass,

2005). Notable amongst these is the ensemble run by INM (Garcia-Moya *et al.*, 2007). Initial conditions for this ensemble are given by analyses from the Met Office, ECMWF, DWD and NCEP. Five different forecast models are run from each of these starting conditions, these models being provided by the Met Office, UCAR, DWD, the HIRLAM consortium and the COSMO consortium. 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 centres, the results have been impressive, with good ensemble spread and low error in the ensemble mean forecast.

Results from a poor-man's ensemble are shown in figure 1. These results are taken from data calculated by Arribas *et al.* (2005) but not shown in that paper. This shows the root mean square error (RMSE) of the ensemble mean forecast for the ECMWF ensemble and a poor-man's ensemble (consisting of the 6 forecast models which 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 figure 1 it is clear that the poor-man's ensemble is under-spread (the spread of the ensemble is less than the error of the ensemble mean).

There are a number of practical problems that hinder the wide-spread use of poor-man's ensembles. There are issues related to the difficulty of transferring large amounts of data between NMSs. There is also a reluctance for NMSs to base their operational forecast output on the output of other NMSs - over which they have no control. 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.

1.2 Why are Poor man's ensembles better?

An important issue which has been discussed recently 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

$$MSE(member) = MSE(mean) + MSS \quad (1)$$

where MSS denotes the spread of the ensemble. For a well-calibrated ensemble the spread will equal the

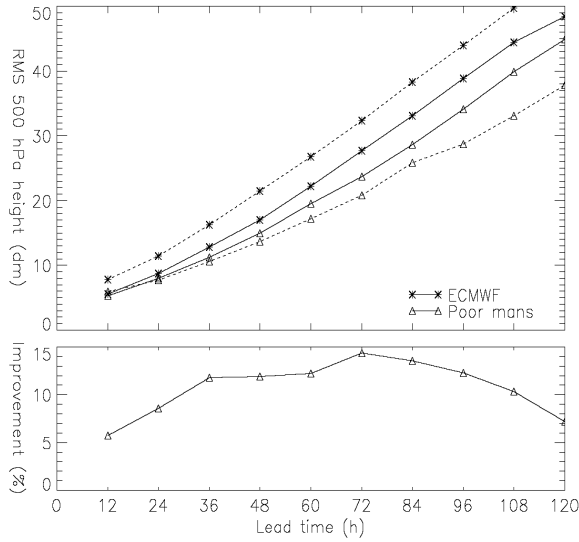


Figure 1. Root-mean square error (solid) of ensemble-mean forecasts of 500hPa height, and RMS spread (dashed) of ensemble, for ECMWF ensembles (stars) and a poor-man's ensemble (triangles) verified over Europe. Below is shown the relative improvement in RMS error for the poor-man's ensemble mean over the ECMWF ensemble mean.

error of the ensemble mean forecast (on average), giving

$$MSE(member) = 2MSE(mean) \quad (2)$$

For traditional ensembles, which are generated by adding perturbations centred 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 validity of the linearity assumption used in 4D-Var and in calculating singular vectors (Gilmour *et al.*, 2001). For ensembles which are initially centred around the control analysis, they will remain centred 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

$$MSE(member) \simeq 2MSE(control). \quad (3)$$

The implication of this is that on average the ensemble members are always much less skilful 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, since 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 and at short-range; for point forecasts and 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, since they are often required to combine a set of forecasts which are of unequal skill.

In contrast to traditional ensembles, a poor-man's ensemble does not generate perturbations which 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. Eq. 1 may be re-arranged to give

$$MSE(mean) = MSE(member) - MSS. \quad (4)$$

In a poor-man's ensemble each member is approximately as skilful 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 derive 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 serves 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. Since 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 figure 1). Further explanation of the relationship between the error of the ensemble mean and control forecasts is given in the appendix.

As is discussed in section 4, the use of RMSE as a verification score can cause problems, since smoothing to 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 since 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 in the Brier score of the ensemble forecast, which is the main measure of ensemble quality.

2 Proposal for a new approach

Inspired by the performance of poor-man's ensembles we propose to use a multi-analysis approach for generating ensemble forecasts. The aim is to mimic the performance of a poor-man's ensemble and calculate an ensemble mean forecast which is more skilful than the deterministic, high-resolution forecast even in the short-range. In order to improve the ensemble mean the approach illustrated in eq. 4 is followed. Each analysis will be a best-guess, approximately as skilful as each other. The differences in the analyses will therefore contribute to generating an ensemble mean which is more skilful 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, M. Fisher, 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. This issue of how to generate a reliable ensemble will be discussed later.

2.1 Data Assimilation framework

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

$$\bar{x}_a = \bar{x}_f + \mathbf{P}^f \mathbf{H}^T (\mathbf{H} \mathbf{P}^f \mathbf{H}^T + \mathbf{R})^{-1} (y - \mathbf{H} \bar{x}_f) \quad (5)$$

where \bar{x}_f is the ensemble mean forecast from the previous cycle, \mathbf{P}^f is the forecast error covariance matrix, \mathbf{H} is the observation operator (here considered linear), \mathbf{R} is the observation error covariance matrix, T denotes the matrix transpose, and y are the current observations. When updating each ensemble member it is necessary to perturb the observations to maintain sufficient spread in the ensemble. For a poor-man's ensemble a different, static, \mathbf{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. Due to this myriad of differences between the forecasts, and the chaotic nature of the atmospheric system, the analyses of each centre do not converge to the same solution, even though they are all attempting to solve the same problem.

3 Description of tests

3.1 Set up of multi-analysis system

In the tests that have been run, an N216 forecast (approximately 60km resolution in the mid-latitudes) has been used in conjunction with the 4D-Var system run at N108 resolution. This is compared with the forecasts from the operational suite, which are run at N320 resolution (approximately 40km in the mid-latitudes). All the forecasts use 50 vertical levels, and are performed for data times between 0Z on 10th May and 12Z on 19th May 2006. The forecasts are all initialised from the analysis of the operational suite valid at 0Z on 6th May allowing 4 days for each system to "spin-up".

The set-up of the system (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 4D-Var trials suite developed by Mike Thurlow.

3.2 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. Ensemble member 1 was created by using the same suite as the control forecast, but introducing a small perturbation to the analysis produced at 6Z on 6th May 2006. This perturbation was based on the differences between the analysis at 0Z and 6Z. 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. Since satellite observations are more dense than can be assimilated, they are routinely thinned to a grid with spacing of approximately 154km. Normally the observations which are closest to the points of the 154km grid are assimilated, but we are free to choose an observation which is not the closest to the grid points. Member 2 uses a random choice of observation within each 154km grid-box, rather than closest.

The background error covariance matrix used by the 4D-Var scheme is determined using the 'NMC' method of Parrish and Derber (1992). The covariance matrix used operationally is based on the difference between two forecasts, 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. 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 3D-Var analysis scheme is used instead of a 4D-Var scheme. This is expected to produce worse forecasts than the control, but is much cheaper computationally.

The next three ensemble members use a new scheme which 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 figure 2(a). Thus for a 3D-grid, at each time-step the value of a variable at a given point is interpolated from the values at the 8 nearest grid-points to the estimated departure point. The interpolation 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 perturbed the calculated departure points in order to reduce the effect more directly.

In this scheme the departure point perturbation is restricted to the 2-D case, this means considering interpolation from the 4 grid-points that are nearest to the calculated departure point. The semi-Lagrangian scheme still interpolates in 3 dimensions, but the perturbation is restricted to the horizontal. Rather than interpolating the field at a given location from the 4 grid-points nearest to the calculated departure point, it is possible take the value from one of the 4 grid-points directly. This would be equivalent to moving the calculated departure point to rest exactly at one of the 4 possible points. However, such a scheme would likely cause the model forecast to fail since 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 4 grid-points. This allows the introduction of a tunable parameter α (see figure 2(b)) which can control the strength of the perturbation. $\alpha = 0$ corresponds to no perturbation, and $\alpha = 1$ corresponds to moving the departure point to rest exactly on one of the 4 grid-points.

The scheme implemented in this paper randomly chooses one of the 4 grid-points in the horizontal. The probability of choosing one of the grid-points is based on an area-weighted scheme for the proximity of the unperturbed departure point to each of the 4 grid-points to ensure highest probability of moving towards the nearest points. This also means that the average departure point calculated is the same as when no perturbation is used. In this study

we used $\alpha = 0.5$, which means that the perturbation to the calculated departure point moves the point to be half-way between the original departure point and the chosen grid-point. The random numbers used to choose the grid-point towards which to move the departure point have no spatial correlation.

Ensemble members 5 and 6 are the same as ensemble members 2 and 3, respectively, but they use the perturbed departure points scheme. Ensemble member 7 is the same as ensemble member 6, but uses a different (lower) value of the 'Jc' filtering term (Gauthier & Thépaut, 2001). This filtering term is used to suppress undesirable gravity waves in the data assimilation process, so ensemble member 7 uses less filtering than the normal assimilation.

Ensemble member 8 uses the analysis produced by the system of the control forecast. However, the forecast is run using the high-resolution (N320) forecast model. The aim of this member is to determine if the differences in the analyses produced by the two systems, or the differences in the forecast model are more important.

4 Results

The following results are based on the ensemble containing the high-resolution forecast, and members 3 and 7. This set of forecasts was chosen since it minimises the root-mean-square error (RMSE) of the ensemble mean forecast for the index of forecast variables discussed later. A different set of forecasts could be chosen with similar results, although here the best combination only is presented. Adding more ensemble members does not improve the forecast, since this would mean the ensemble would have many more low-resolution forecasts than high-resolution forecasts (of which only one is available).

Figure 3 shows the RMSE as a function of lead time for forecasts of 500 hPa height from the high-resolution model, and members 3 and 7. Also shown are the RMS spread of this three member ensemble, and the RMSE of the ensemble mean forecast. These forecasts (as all presented in this paper) are verified against radio-sonde observations from across the globe valid and 0Z and 12Z between the 10 and 22 May 2006. Since there are more radio-sondes in the northern hemisphere than elsewhere, the statistics will be biased towards the ensemble performance in the northern hemisphere. It can be seen that the forecast from the high-resolution model is approximately as skilful as member 3, and more skilful than member 7. This illustrates for 500 hPa height that the enhanced horizontal resolution only weakly influences the quality of the forecast. The ensemble mean forecast is more skilful than any of the individual ensemble members. The improvement of the ensemble mean forecast over the high-resolution forecast varies between none (at T+0) and 4% (at T+72).

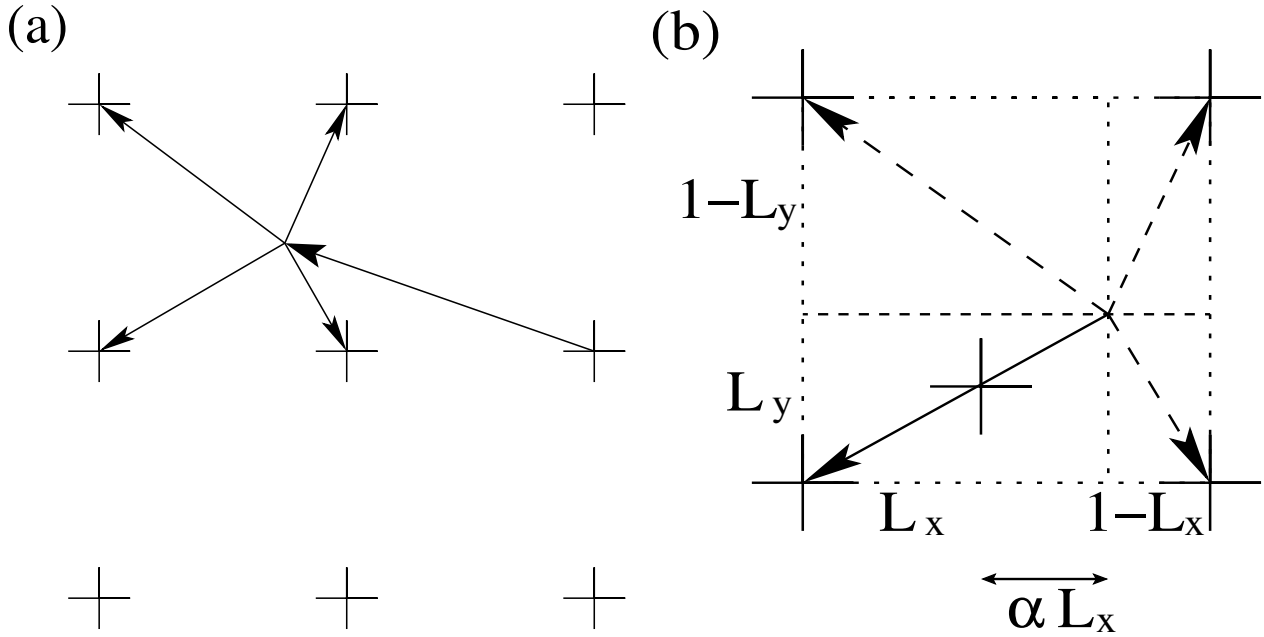


Figure 2. (a) The semi-Lagrangian advection scheme interpolates the value at a given point, from the four nearest points upstream of the point being considered. (b) In the perturbed scheme one of the four points is chosen with a probability related to the area-weighting of the boxes shown, and the departure point is moved a factor α closer to this point (+). The bottom left point would be selected with a probability $(1 - L_x)(1 - L_y)$ so the points nearest the standard departure point are most likely to be selected.

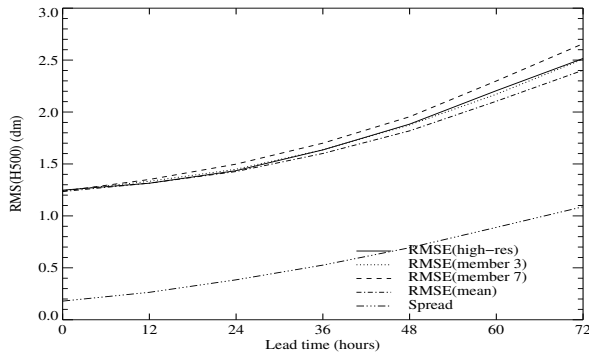


Figure 3. Root-mean square error of forecasts of 500hPa height, and RMS spread of ensemble

The ability to produce a forecast which has a lower RMSE 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 (and therefore lower RMSE) than their ensemble mean forecast for the first 5 days. We know of no other system, aside from a poor-man's ensemble, which can produce forecasts of 500 hPa height which are more skilful than the high-resolution model.

Similar results to figure 3, but for wind speed and temperature at 850 hPa, are shown in figures 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,

since the effect of variations in surface height will be more noticeable at this level. This is reflected in figure 4, for which the lower-resolution forecasts perform clearly worse. Member 7 is the least skilful of all the forecasts, indicating that the perturbed departure points scheme is not performing well for this variable. Consequently the ensemble mean forecast performs worse than the high-resolution forecast. However, for 850 hPa temperature, the difference in performance between the three forecasts is much less. Members 3 and 7 perform worse than the high-resolution forecast, but the degradation is less than for wind speed. Additionally, the spread of the ensemble is greater (relative to the RMSE of the high-resolution forecast). Both these facts result in the RMSE of the ensemble mean forecast for this variable being much less than the RMSE of any of the members individually.

For all the variables shown in figures 3-4 the spread of the multi-analysis ensemble is much less than the error in the ensemble mean forecast. This indicates that the multi-analysis ensemble is unlikely to be useful without further modifications, which will be discussed later. The spread is also less than is seen with the poor-man's ensemble (c.f. figures 1 and 3) indicating that there are some uncertainties sampled by that ensemble which are not accounted for here.

How do the results of figure 3 compare with what can be achieved using a traditional ensemble where ensemble members are created by perturbing a central analysis? Figure 6 shows comparable results for the Met Office ensemble (MOGREPS) (Bowler *et al.*, 2007). The perturbations to the initial conditions are

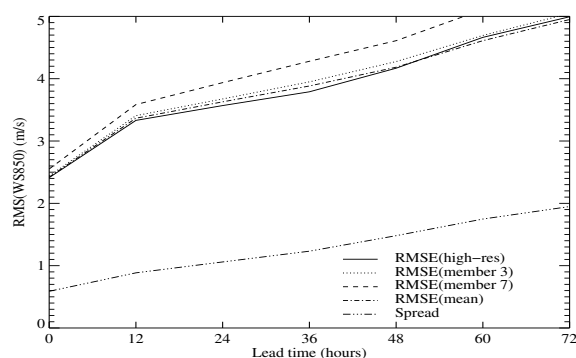


Figure 4. Root-mean square error of forecasts of 850hPa wind speed, and RMS spread of ensemble

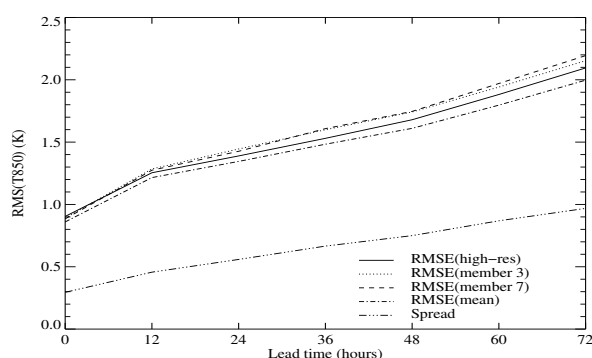


Figure 5. Root-mean square error of forecasts of 850hPa temperature, and RMS spread of ensemble

created using the Ensemble Transform Kalman Filter (ETKF) (Bishop *et al.*, 2001) and the forecasts are run at N144 resolution (around 90km 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 skilful than any of the control forecast, the ensemble mean forecast or the high-resolution forecast. The control forecast and the ensemble mean are less skilful than the high-resolution forecast. One may note that the ensemble mean forecast is slightly less skilful than the control forecast. This is believed to be due to the ensemble spread being slightly too large.

The improvements in the ensemble mean forecast for the multi-analysis ensemble over a number of variables, relative to the high-resolution forecast are shown in figure 7. The results have been compiled by finding the sum of the RMSE values for each forecast lead time, out to T+72, and weighting the sum linearly so that T+0 values receive weight 1, and T+72 values receive weight 0.5. The percentage improvement of this sum relative to the high-resolution forecast is reported in figure 7. An index of these values has been created to mimic the NWP index used by the Met Office, using weights shown

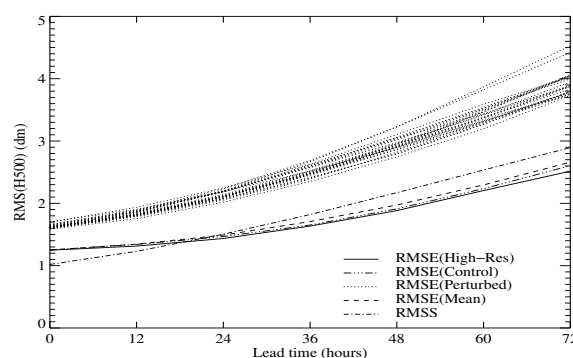


Figure 6. Root-mean square error of forecasts of 500hPa height, and RMS spread of ensemble for MOGREPS. Each of the ensemble member forecasts have a much higher RMSE than the high-resolution forecast, and the ensemble mean forecast is worse than the high-resolution forecast at all lead times.

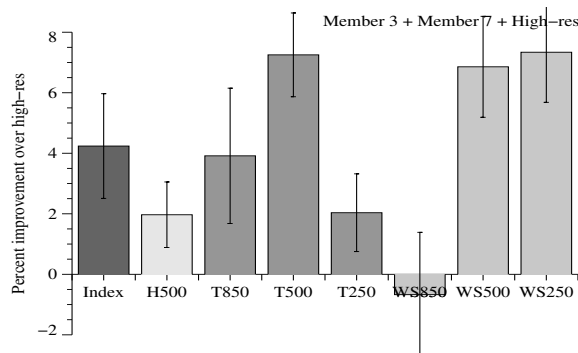


Figure 7. Index of forecast quality for an ensemble consisting of the deterministic forecast, and two lower-resolution forecasts

in table I. The error bars are the two 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. Since these are two-sigma error bars, they represent 95% confidence intervals on the improvement in the ensemble mean forecast. The number of degrees of freedom are derived using the Z method as described by Wang and Shen (1999). For all but the 850 hPa wind speed, the ensemble mean forecast is more skilful than the high-resolution forecast. The benefit of the ensemble mean varies with variable considered, but the index is 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.

Figure 8 shows similar summary results for the MOGREPS ensemble. As was seen for 500 hPa height the ensemble mean forecast is less skilful than the high-resolution forecast. For some variables, such as 500 hPa temperature, MOGREPS performs better than the high-resolution forecast.

Variable	Weight
Height at 500 hPa	0.25
Temperature at 850 hPa	2
Temperature at 500 hPa	1.6
Temperature at 250 hPa	1.2
Wind speed at 850 hPa	1
Wind speed at 500 hPa	0.5
Wind speed at 250 hPa	0.85

Table I. Weights given to forecast variables in order to calculate an index of forecast quality.

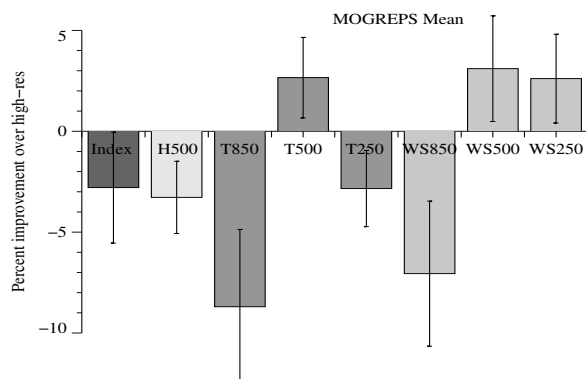


Figure 8. Index of forecast quality for the MOGREPS ensemble mean.

Overall, the results are clearly worse than for the multi-analysis ensemble.

Relative values for this index of forecast variables for all the experiments described in section 3.2 are shown in table II. This 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 change of horizontal resolution is an important component in generating spread in a multi-analysis ensemble. 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 and member 7 is the most skilful forecast combination. Surprisingly, the control forecast appears more skilful 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 which uses a 3D-Var analysis (member 4) gives a noticeably worse forecasts than the high-resolution, demonstrating the benefits of the more sophisticated analysis scheme. However, because there are substantial differences between this forecast and the control forecast (around 43% 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 will 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 but 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.

Using root-mean-square error as a measure of the forecast quality is known to have deficiencies, since a forecast may be improved by a simple smoothing algorithm. RMSE heavily penalises any large deviations of the forecast from the verification, through the use of the square of the error. When a simple smoothing (box-averaging with length of side between 3 and 15 grid-points) is applied to the deterministic forecast, the maximum reduction of the RMSE for the index of forecast variables is around 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 (1993). The comparison between the multi-analysis and MOGREPS ensemble means is not prone to the issue of using RMSE, since the means of both ensembles are constructed similarly. Hence the conclusion that the multi-analysis ensemble is able to improve the ensemble mean in a way that traditional ensemble cannot is not susceptible to issues related to the choice of verification measure in the same way as comparison with a deterministic forecast.

5 Practical considerations

5.1 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 figures 3 to 5 it can be seen that the multi-analysis ensemble has slightly less than 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 skilful as the high-resolution forecast. A poor-man's ensemble (see figure 1) is close to this goal. This, however, may be out of reach for a single-model system as described here.

Since it has now been demonstrated that it is possible to generate some ensemble spread

Ensemble member	High-res	Control	1	2	3	4	5	6	7	8
RMSE	1.000	0.990	0.989	0.991	0.987	1.005	1.015	1.013	1.007	1.008
RMSE (average with hi-res)	-	0.970	0.969	0.968	0.964	0.973	0.963	0.961	0.959	0.973
RMS difference hi-res	-	0.471	0.476	0.489	0.509	0.510	0.616	0.619	0.615	0.514
RMS difference control	0.471	-	0.149	0.187	0.250	0.436	0.423	0.431	0.425	0.329

Table II. 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 normalised by the RMSE of the high-resolution forecast.

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 multi-analysis ensemble, whilst 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. Since the multi-analysis 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 skilful 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 TIGGE (TIGGE, 2005) multi-model ensemble is very similar to the multi-centre ensemble considered by Evans *et al.* (2000). However, each forecast centre will calibrate its own ensemble to have approximately the same spread as the error in its ensemble mean forecast. Since 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 centre's ensemble to ensure that the resulting multi-centre ensemble is well-calibrated.

5.2 Computer time comparison

One difficulty with a multi-analysis 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 III 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 8 6-hour 4D-Var analyses per day (each analysis is repeated once the full set of observation data has been received). Additionally, 2 forecasts are run to T+171, 2 forecasts are run to T+70, and 4 forecasts are run to T+12.

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 2 cycles per day). This would entail a large increase in the computational cost of the ensemble suite. Since the 3D-Var data assimilation method is much less expensive than 4D-Var 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 4 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 25km in the mid-latitudes (with an assumed 10 minute time-step which compares with the 15 minute time-step used by the N320 model). This is included because ECMWF recently upgraded their high-resolution model from T511 (~ 40km) to T799 (~ 25km). This upgrade only improved the performance of the 500 hPa height forecast by 1% in the northern hemisphere (Martin Miller, 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 years 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 (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 3D-Var data assimilation 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 4 dimensional versions of the EnKF are possible, if the EnKF uses a static background error covariance matrix,

then it becomes the equivalent of a 3D-Var analysis scheme. This implies that for an EnKF to be competitive with 4D-Var than it must avoid the use of static background error covariances through very large ensembles. Therefore, investigating methods which can provide improved analyses at the cost of performing a small number of additional analyses may provide a better way forward.

6 Conclusion

In this paper, results from a multi-analysis 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 the high-resolution forecast for all quantities, except wind speed at 850 hPa. 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 other ensemble methods which depend on a single analysis. 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, since 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, such as described above, which amplify through a repeated analysis cycle. Differences in the resolution of the forecast model have a larger impact on the analysis than differences in the resolution of the forecast model alone.

In order to gain the benefit from a multi-analysis ensemble, whilst generating an ensemble forecast with appropriate spread, it may be necessary to include initial condition perturbations which degrade the forecast. Such a set-up was tested by Evans *et al.* (2000). This kind of system is also very similar to that of the TIGGE multi-model ensemble (TIGGE, 2005). Given that the spread of a poor-man's ensemble is often quite close to the RMSE of the ensemble mean forecast (see figure 1) the initial condition perturbations in this set-up will need to be quite small, which is not the case with TIGGE. Therefore, post-processing 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 a 70 level version of the N320 forecast. A 3D-Var analysis is much cheaper than this, although the forecasts produced from this system are inferior to those produced using 4D-Var. However, the forecast improvement from combining the low-resolution 3D-Var forecast with the high-resolution 4D-Var forecast is still around 2.5%.

In the case of a "traditional" ensemble it is a simple matter to choose 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 be 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, since 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 since 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 due to the existence of poor-man's ensembles, though political and operational constraints have hindered NMSs basing their output on data from external sources. Now these constraints may be mitigated by the use of a multi-analysis ensemble.

Further avenues of study include running such a system for a second case, for a longer series of forecasts. This will allow a greater level of confidence in the improvements gained from a multi-analysis system. Since resolution plays a very important part in the benefits from a multi-analysis ensemble, a

Forecast system component	CPU time (SX6 seconds)	Ensemble member equivalent
UM forecast (N144 T+72)	5,143	1
N320L50 forecast cycle	458,696	89.19
N216L50 forecast cycle	205,384	33.93
N216L50 analysis cycle	110,732	21.53
N216L50 (3D-Var) analysis cycle	18,344	3.57
N512L90 forecast cycle	2,264,647	440.34

Table III. Average computer time (equivalent on one processor of NEC SX6) for components of the operational suite

logical next step would be to test differences in the inner-loop resolution of the analysis system. In addition testing different estimates of the background error covariance matrix may yield further benefits.

At this stage, one returns to the quote of George Box “All models are wrong, but some are useful”. From the work presented above we have a better understanding of how to define useful - a model is certainly 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?

Acknowledgements

The authors would like to 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 9 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

$$\left\| \frac{e_1 + e_2}{2} \right\| = \frac{1}{4} \|e_1\| + \frac{1}{4} \|e_2\| + \frac{1}{2} e_1 \cdot e_2 \quad (6)$$

where $\|\cdot\|$ 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

$$\left\| \frac{e_2 - e_1}{2} \right\| = \frac{1}{4} \|e_1\| + \frac{1}{4} \|e_2\| - \frac{1}{2} e_1 \cdot e_2. \quad (7)$$

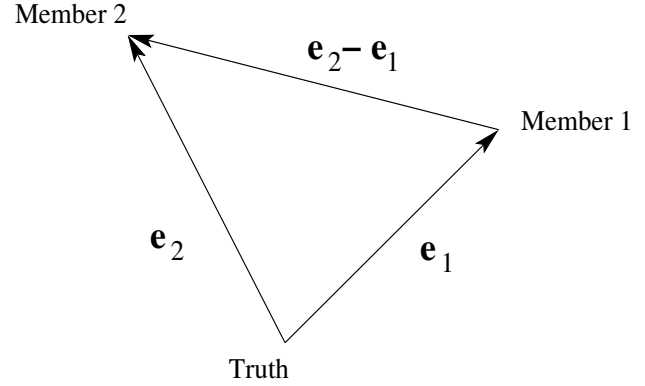


Figure 9. Index of forecast quality for the MOGREPS ensemble mean.

For the ensemble mean to be closer to the truth than ensemble member 1, we require

$$\frac{1}{2} \|e_1\| + \frac{1}{2} \|e_2\| - e_1 \cdot e_2 < \|e_1\|. \quad (8)$$

Re-arranging this and writing in terms of the notation of equation 1 then we find

$$MSS > \frac{1}{2} MSE(member2) - \frac{1}{2} MSE(member1). \quad (9)$$

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-resolution and low-resolution forecasts are reported in table II. 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.

The condition described in equation 9 explains why adding a poor-quality forecast to a poor-man's ensemble can improve the ensemble mean. Provided that the forecast is more different to the other forecasts than it is worse, then the ensemble mean will improve. For “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 than the ensemble mean as it is different.

References

- Atger F., "The skill of ensemble prediction systems", *Monthly Weather Review*, **127**, 1941-1953 (1999)
- Arribas, A., Robertson, K.B., Mylne, K.R., "Test of a poor man's ensemble prediction system for short-range probability forecasting", *Monthly Weather Review*, **133**, 1825-1839 (2005)
- Bishop, C.H., Etherton, B.J. and Majumdar, S.J., "Adaptive sampling with the ensemble transform kalman filter. Part 1: theoretical aspects", *Monthly Weather Review*, **129**, 420-436 (2001)
- Bowler, N.E., Arribas, A., Mylne, K.R., and Robertson, K., "The MOGREPS short-range ensemble prediction system Part I: System description", *Met Office technical report number 497* (2007)
- Buizza, R. and Palmer, T.N., "The singular-vector structure of the atmospheric general circulation" *J. Atmos. Sci.*, **52**, 1434-1456 (1995)
- "Benefits of increased resolution in the ECMWF ensemble system and comparison with poor-man's ensembles" Buizza R., Richardson, D.S., Palmer, T.N., *Q. J. Royal Meteorological Society*, **129**, 1269-1288 (2003)
- Buizza, R., Houtekamer, P.L., Toth, Z., Pellerin, G., Wei, M., and Yuejian Z., "A comparison of the ECMWF, MSC and NCEP global ensemble prediction systems", *Monthly Weather Review*, **133**, 1076-1097 (2005)
- Ebert, E.E., "Ability of a poor man's ensemble to predict the probability and distribution of precipitation", *Monthly Weather Review*, **129**, 2461-2480 (2001)
- Eckel, F.A. and Mass, C.F., "Aspects of effective mesoscale, short-range ensemble forecasting" *Weather and Forecasting*, **20**, 328-350 (2005)
- Evans, R.E., Harrison, M.S.J., Graham, R.J., Mylne, K.R., "Joint medium-range ensembles from the Met. Office and ECMWF systems", *Monthly Weather Review*, **128**, 3104-3127 (2000)
- Evensen, G., "Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics", *J. Geophysical Research-Oceans*, **99**, 10143-10162 (1994)
- Garcia-Moya, J.A., Callado, A., Santos, C., Santos, D. and Simarro, J., "The role of mass transfer in describing the dynamics of mesoscale convective systems", *3rd International Verification Methods Workshop*, ECMWF, (2007) (available at: http://www.ecmwf.int/newsevents/meetings/workshops/2007/jwgv/workshop_presentations/index_poster.htm)
- Gauthier, P., Thépaut, J.N., "Impact of the digital filter as a weak constraint in the preoperational 4DVAR assimilation system of Meteo-France", *Monthly Weather Review*, **129**, 2089-2102 (2001)
- Gilmour, I., Smith, L.A., Buizza, R., "Linear regime duration: Is 24 hours a long time in synoptic weather forecasting?", *Journal of the Atmospheric Sciences*, **58**, 3525-3539 (2001)
- Hamill, T.M., Snyder, C., "A hybrid ensemble Kalman filter-variational analysis scheme", *Monthly Weather Review*, **128**, 2905-2919 (2000)
- Harrison, M.S.J., Palmer, T.N., Richardson, D.S., Buizza, R., Petrolia, T., "Joint ensembles from the UKMO and ECMWF models", *ECMWF workshop on predictability*, **II**, 61-120 (1995)
- Houtekamer, P.L., Lefavre, L., Derome, J., Ritchie, H., Mitchell, H.L., "A system simulation approach to ensemble prediction", *Monthly Weather Review*, **124**, 1225-1242 (1996)
- Houtekamer, P.L., Mitchell, H.L., "A sequential ensemble Kalman filter for atmospheric data assimilation", *Monthly Weather Review*, **129**, 123-137 (2001)
- Houtekamer, P.L., Mitchell, H.L., Pellerin, G., Buehner, M., Charron, M., Spacek, L., Hansen, B., "Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations", *Monthly Weather Review*, **133**, 604-620 (2005)
- Jung, T., Palmer, T.N., Shutts, G.J., "Influence of a stochastic parameterization on the frequency of occurrence of North Pacific weather regimes in the ECMWF model", *Geophysical Research Letter*, **32**, L23811 (2005)
- Lewis, J.M., "Roots of ensemble forecasting", *Monthly Weather Review*, **133**, 1865-1885 (2005)
- Lorenc, A.C., "The potential of the ensemble Kalman filter for NWP - a comparison with 4D-Var", *Q. J. Royal Meteorological Society*, **129**, 3183-3203 (2003)
- Mylne, K.R., Evans, R.E., Clark R.T., "Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting", *Q. J. Royal Meteorological Society*, **128**, 361-384 (2002)
- Ott, E., Hunt, B.R., Szunyogh, I., Zimin, A.V., Kostelich, E.J., Corazza, M., Kalnay, E., Patil, D.J., "A local ensemble Kalman filter for atmospheric data assimilation", *Tellus*, **56A**, 415-428 (2004)
- Palmer, T., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M., Smith, L., "Ensemble prediction: a pedagogical perspective", *ECMWF newsletter*, **106**, 10-17 (2006) (available at: <http://www.ecmwf.int/publications/newsletters/pdf/106.pdf>)

- Parrish, D.F., Derber, J.C., "The National Meteorological-Centers spectral statistical-interpolation analysis system", *Monthly Weather Review*, **120**, 1747-1763 (1992)
- Shutts, G., "A kinetic energy backscatter algorithm for use in ensemble prediction systems", *Q. J. Royal Meteorological Society*, **131**, 3079-3102 (2005)
- Richardson, D., "Ensembles using multiple models and analyses", *Q. J. Royal Meteorological Society*, **127**, 1847-1864 (2001)
- Rodwell, M.J., "Comparing and combining deterministic and ensemble forecasts: how to predict rainfall occurrence better", *ECMWF newsletter*, **106**, 17-23 (2006) (available at: <http://www.ecmwf.int/publications/newsletters/pdf/106.pdf>)
- Smith, L.A., "Predictability and chaos", in *Encyclopedia of the Atmos. Sci.*, **4**, 1777-1785, Academic Press (2003) Eds. J.R. Holton, J.A. Curry, J.A. Pyle
- Stensrud DJ, Brooks HE, Du J, Tracton MS, Rogers E, "Using ensembles for short-range forecasting", *Monthly Weather Review*, **127**, 433-446 (1999)
- Richardson, D.S., Buizza, R., Hagedorn, R., "First workshop on the THORPEX interactive grand global ensemble (TIGGE): Final report" (2005) (available at: http://www.wmo.int/pages/prog/arep/thorpex/pdf/tigge_first_workshop_report.pdf)
- Toth, Z. and Kalnay, E., "Ensemble Forecasting at the NMC: The generation of perturbations", *Bull. Amer. Meteor. Soc.*, **74**, 2317-2330 (1993)
- Toth, Z. and Kalnay, E., "Ensemble forecasting at NCEP and the breeding method", *Monthly Weather Review*, **125**, 3297-3319 (1997)
- Tracton M. S., Du J., Toth Z. and Juang H., "Short-range ensemble forecasting (SREF) at NCEP/EMC" *12th Conf. on Numerical Weather Prediction, Phoenix*, American Meteorological Society, 269-272 (1998)
- Wang, X., Shen, S.S., "Estimation of spatial degrees of freedom of a climate field", *J. Climate*, **12**, 1280-1291 (1999)
- Ziehmann, C., "Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models", *Tellus*, **52A**, 280-299 (2000)