Met Office

Meteorology Research & Development

Medium-range multi-model ensemble combination and calibration

Technical Report No. 517

Christine Johnson and Richard Swinbank

email:nwp_publications@metoffice.gov.uk

Medium-range multi-model ensemble combination and calibration

Christine Johnson

Abstract

As part of its contribution to THORPEX, the Met Office has developed a global, 15-day multi-model ensemble. The multi-model ensemble combines ensembles from ECMWF, Met Office and NCEP and is calibrated to give further improvements. The ensemble-post processing includes bias correction, model-dependent weights and variance adjustment, and is based on a moving-average over past observation-forecast pairs. The post-processing parameters are calculated separately for each grid-point and forecast lead-time, and we show that the optimal size of the training data set is dependent on the forecast lead-time.

Verification shows that the multi-model ensemble gives an improvement in comparison to a calibrated single-model ensemble, particularly for surface temperature. However, the benefits are smaller for mean-sea-level-pressure (mslp) and 500hPa height. The reason for this is attributed to the higher degree of similarity between forecast-errors for mslp and 500hPa height than for temperature. The results also show only small improvements from the use of the model-dependent weights and the variance-adjustment. This is because the models have similar levels of skill, and the multi-model ensemble variance is already generally well calibrated.

In conclusion, we demonstrate that the multi-model ensemble does give benefit over a single-model ensemble. However, as expected, the benefits are small if the models are similar to each other and further post-processing gives only relatively small improvements.

Contents

1 Introduction

Ensemble forecasts, used in probabilistic weather prediction, aim to represent the uncertainty that arises from errors both in the initial conditions, or analysis, and in the forecast model. Typically, an ensemble prediction system generates a range of initial conditions by adding small perturbations to the analysis, and then evolving each initial condition with the numerical forecast model (e.g. Toth and Kalnay, 1993). It is also common to add stochastic perturbations throughout the model integration, to account for errors from the model parameterizations (e.g. Houtekamer et al., 1996). However, it is difficult to represent all the errors arising from the forecast model.

The aim of a multi-model ensemble is to account for the errors in both the initial conditions and the forecast model, by combining together ensembles from different centres, and hence combining different analyses, perturbation generation methods, and forecast models. It has been shown in the context of seasonal forecasting (Palmer et al., 2004; Hagedorn et al., 2005; Weigel et al., 2008) that the combination of ensembles from different models results in more skill than the single ensembles considered separately, this improvement being not just due to the increased ensemble size, but from the information provided by the different models. Benefits in the medium-range (15-day) have also been shown (Harrison et al., 1999; Evans et al., 2000; Mylne et al., 2002), with the improvement being attributed to the models exploring different regions of phase space. Further studies have shown that although sometimes the multi-model ensemble is not always the most skilful, it is better than the worst single-model ensemble, as emphasised by Matsueda et al. (2007), and as we can not predict which the worst single-model ensemble will be at a particular time and location, this gives the multi-model ensemble an advantage.

Although a multi-model ensemble combines the strengths from different models, some models might be better than others at different times, and hence further improvements might be made by giving the models different weights. Previous research has shown that model-dependent weights can give improvements, but care needs to be taken in how they are calculated and used. For example, Raftery et al. (2005) concluded that the Bayesian Model Averaging (BMA) model-dependent weights created a better deterministic forecast, and Stefanova and Krishnamurti (2002) also showed an improvement from model-dependent weights, when considering probability forecasts from seasonal multi-model forecasts. However, Doblas-Reyes et al. (2005) concluded that model-dependent weights gave no significant improvement in the DEMETER (Development of a European Multi-model Ensemble for Seasonal to Interannual Prediction) ensemble.

Further improvements can be made to ensemble forecasts through calibration. In calibration, the forecasts are adjusted so that the average statistical properties are similar to those of a reference data set. For example, if forecast temperatures are consistently too high, these can be reduced through bias correction. Thus, it may be the case that a calibrated multi-model ensemble has little benefit over a calibrated single-model ensemble. The results in Doblas-Reyes et al. (2005) suggested that both calibration and combination improve the ensemble predictions, although the results are sensitive to the reference data sets. Hagedorn et al. (2008) also found that a calibrated multi-model ensemble was slightly more skillful than a calibrated single model ensemble.

	ECMWF	Met Office	NCEP
Number of members	51	24	21
Perturbation method	Singular vectors	ETKF	Bred Vectors
Horizontal Resolution	50km (TL399) day 0-10, 80km (TL255) day 10-15	90km (1.25/0.833)	105km (T126)
Vertical Levels	62	38	28

Table 1: *Comparison of ensembles.*

The THORPEX Interactive Grand Global Ensemble (TIGGE) is designed to allow the enhanced collaboration of operational centres, and to investigate new techniques to combine and calibrate ensembles. Preliminary results on the ensembles archived in the TIGGE database have been compiled by Park et al. (2008). The results showed that for 500hPa height, there was only a small benefit from a simple multi-model ensemble in comparison to the ECMWF ensemble. The purpose of this report is to investigate further the benefits of multi-model ensemble combination, particularly looking at other variables, and also from more sophisticated combination methods.

The structure of the paper is as follows. In section 2 we describe the component model ensembles, and also the general framework in which the combination and calibration is addressed. We then go on to describe the procedure by which the calibration parameters are estimated, and how these parameters are used to correct the bias of the single model ensembles, define the model-dependent weights and adjust the variance of the multi-model ensemble. The results are presented in section 3, beginning with the impact of bias-correction, and a simple combination of models, and then examining the impact of the weights and variance adjustment. The ensembles are verified both using RMS errors of the ensemble mean, and Brier skill scores. In section 4 we present a general discussion of the benefits of multi-model ensembles and suggestions for future work. The final conclusions of the paper are given in section 5.

2 Procedure

Three ensembles have been chosen as the components of the multi-model ensemble. These are ECMWF, Met Office, and NCEP (GFS), which have been chosen because they are all accessible in real-time, and because they have similar levels of skill. The Met Office 15-day ensemble is an extension of the MOGREPS short-range ensemble (Bowler et al., 2008), which is currently run for research purposes as part of the Met Office contribution to THORPEX. The studies from Titley et al. (2008) showed that although the Met Office ensemble does not quite have the same skill as the ECMWF ensemble, many of the aspects are competitive. A separate study (Buizza et al., 2005) showed that the NCEP ensemble is also competitive, especially during the first 5-days. The details of the models are presented in table 1. The Met Office and NCEP ensembles have similar resolution grids and numbers of members, and the ECMWF has twice the number of members, and higher resolutions both in the horizontal and the vertical. The ensembles are post-processed on a $1^\circ/1^\circ$ grid, for three variables: mean sea level pressure (mslp), 2m temperature and 500hPa height. We concentrate on results for mslp and 2m temperature, as the results for 500hPa height are similar to those for mslp.

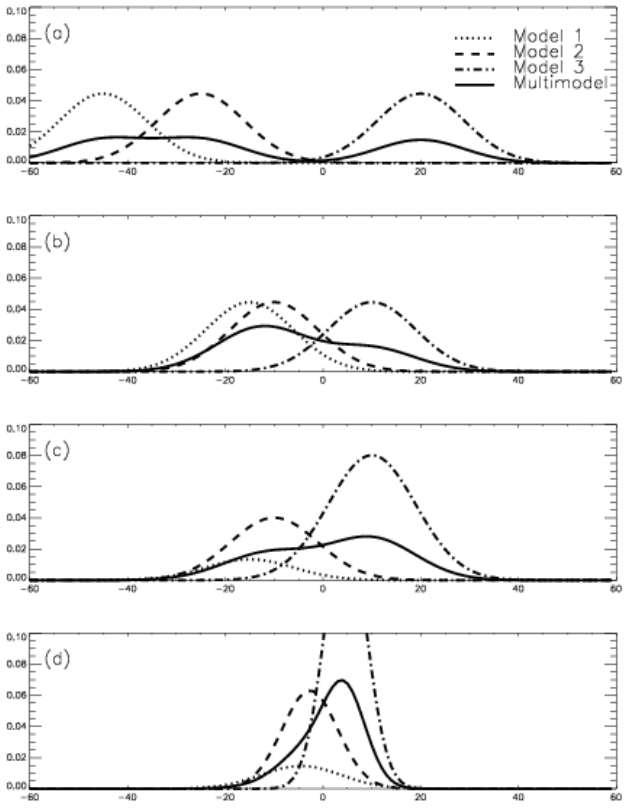


Figure 1: *Illustration of the steps in calibrating and combining the multi-model ensemble. Each panel shows the pdfs for each ensemble, with the x-axis representing the value of the forecast variable, e.g. temperature. The panels show the pdfs for (a) raw data, (b) after bias correction (c) after model-dependent weights and (d) after variance adjustment.*

2.1 Calibration and combination framework

We now describe a very general calibration and combination procedure in a similar way to Raftery et al. (2005) and, at this stage, without making any assumptions about how the biases, weights and variance adjustments should be estimated. We assume that there are M calibrated single-model ensembles and that the k^{th} single model ensemble has N_k ensemble members, given by x_k^i. The true or verifying state is given by y. From the law of total probability (Raftery et al., 2005) the multi-model probability density function (pdf) of the variable x is given by an average of the pdfs from the single-models,

$$p(x) = \sum_{k=1}^{M} p(x|\mathcal{M}_k)p(\mathcal{M}_k) \tag{1}$$

where $p(x|\mathcal{M}_k)$ is the pdf based on model $\mathcal{M}_k$ and $p(\mathcal{M}_k)$ is the probability of $\mathcal{M}_k$ being the best model. These probabilities can be viewed as model-dependent weights w_k such that $p(\mathcal{M}_k) = w_k/M$, where $0 \leq w_k \leq M$ and $\sum_{k=1}^{M} w_k = M$.

The calibration and combination procedure is split into three steps: bias (first moment) correction, weighting and variance (second moment) adjustment, as illustrated in Fig.1. The ensembles are combined and calibrated in a framework where each single-model pdf is represented by a normal (Gaussian) distribution with mean and variance given by the ensemble mean and ensemble variance respectively. The pdfs for the three single models are shown by the dashed curves, and the pdf for the multi-model,

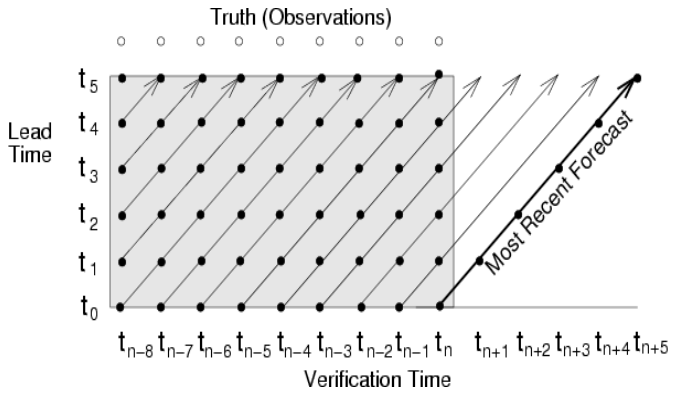


Figure 2: *Illustration of the available bias correction data. A single forecast is shown by a diagonal arrow, as a function of the lead time and verification time, and the most recent forecast is shown by the thick diagonal arrow. The observations, shown by the unshaded circles, are only available to the present time, t_n, meaning that only the forecast data in the shaded box can be used to estimate the bias in the most recent forecast.*

given by an average of the three pdfs, is shown by the solid curve.

The first step is to estimate the biases of the individual ensembles so that the average ensemble mean is closer to the average true value. Let us imagine that models 1 and 2 have negative bias, and that model 3 has a positive bias. Then bias correction applied to each individual model will shift the models towards zero, as shown in Fig.1(b). Notice that the bias correction has no effect on the spread of the individual ensembles.

The second step is to define the model-dependent weights. It is likely that some models exhibit more skill than other models in certain situations. Therefore, it is sensible to estimate model-dependent weights, w_k, to apply to each pdf. In Fig.1(c), model 1 has been assigned the smallest weight, and model 3 has been assigned the largest weight.

The final step is to adjust the variance of the multi-model ensemble so that it is, on average, representative of the uncertainty in the multi-model ensemble mean. Let us imagine that we have a relatively accurate multi-model ensemble mean. Then the calibration will reduce both the within-model variance (the variances of the single model ensembles) and the between-model variance (the variance of the single model means around the multi-model mean), as shown in 1. Note that although the variance adjustment does not alter the multi-model ensemble mean, it does affect the means of the component models.

2.2 Moving-average estimates

If we assume that the model errors are stationary over time, then it would be possible to generate estimates for the calibration parameters by averaging over a sample of forecast errors over a long period of time. This approach, known as Model Output Statistics (MOS) has been successfully used in the statistical post-processing of many weather and climate forecasts (e.g. Wilks, 2006; Doblas-Reyes et al., 2005; Stephenson et al., 2005; Kharin and Zwiers, 2002). However, MOS requires a large time period of calibration data, with an identical forecast model. If the forecast model changes, then the MOS statistics need to be recalculated, making the process unfeasible for the realistic situation with many model upgrades.

An alternative method is to perform an *on-line* estimation, as used by for example Cui et al. (2004) and Woodcock and Engel (2005). In this technique, the parameter is updated over time, removing the need for a large set of calibration data and making the calculation more efficient. Further, and perhaps

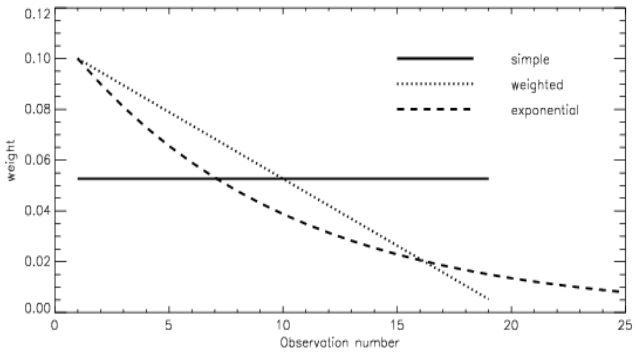


Figure 3: *Illustration of the size of the data window associated with a particular value of μ. In this case, the data window is of length $Q = 19$, and for the exponential moving average, $\mu = 0.1$.*

more importantly, the estimate easily accommodates for changes in the bias due to model upgrades and also allowing for a better estimate of errors that depend on the particular weather conditions (flow-dependent errors).

The moving-average estimate, also known as an exponential moving average, is a simple method where all the previous forecast errors are averaged together, but using an exponentially increasing weighting so that the most recent data has the largest weight. The method is currently a popular way to correct the biases for atmospheric forecasts, (e.g. Cui et al., 2004; Woodcock and Engel, 2005). Although the method averages together all the previous data, it can still be implemented as a so-called on-line or sequential update. The moving-average can also be considered as a type of Kalman filter (Homleid, 1995), but where we assume that there are no observational errors so that no filtering is necessary.

The aim is to find an estimate of the calibration parameter p_s^n for a forecast at lead time s, and starting at verification time n. The calibration parameter could be for example, the average forecast error (i.e. the bias), or the average variance. The estimate $\hat{p}_s^n$ is given by

$$\hat{p}_s^n = (1 - \mu)\hat{p}_s^{n-1} + \mu p_s^{n-s} \tag{2}$$

which is an average of the previous estimate $\hat{p}_s^{n-1}$ and the most recent parameter p_s^{n-s}. Some of the parameter estimates, such as the bias, require forecast-observation pairs as the input data. This means that the most recent parameter is actually at the most recent observation time, which is from the forecast at verification time $n - s$ and with lead time s.

As the input data is forecast-observation pairs, all the data to be used in the estimate is from the past, as illustrated in Fig. 2. The most recent forecast starts from an analysis at verification time t_n, with a lead time of t_0. The forecast is integrated over 5 steps so that the forecast at a lead time of t_5 should be close to the observations at a verification time of t_{n+5}. The estimate of the bias of the most recent forecast is based on the forecast errors of the past. These forecast errors are given by the differences between the forecasts and the verifying observations (the unshaded circles). As we only have observations of the past available, we can only use the forecast values in the shaded box to estimate the bias of the most recent forecast. For example, to calibrate the forecast at a lead time of t_s, the most recent forecast error that is available is the difference between the forecast that starts at verification time t_{n-s} with a lead time of t_s and the observation at verification time t_0. Thus, there is a time-lag of s steps between the forecast to be corrected and the most recent forecast error. The lag between these forecasts increases with increasing lead time, and we will see that this time-lag plays an important role in determining an optimal estimate of the calibration parameters.

The parameter μ determines the smoothness of the estimate. If μ is close to zero, then the recent data will have very little influence, and the resulting bias will be close to the average climatological estimate. If μ is close to one, then the bias will be dominated by the most recent data. However, as the forecast statistics are likely to change quickly over time, it is likely that the most recent data no longer provides a good estimate. Thus, the optimal value for μ lies somewhere in between these two extremes.

It is useful to know the equivalent size of the calibration data set associated with a particular value of μ, and this is possible by comparing the exponential weighted average (the running mean) with a weighted moving average. In all cases, the sum of the weights over time is equal to one. In a weighted moving average of Q observations, the weights are given by $w_i = 2(Q-i+1)/q(q+1)$, so that they are arithmetically decreasing with the most recent observation having the largest weight $(2/(Q+1))$. In an exponential moving average, there are an infinite number of observations. However, by comparing an exponential moving average with a weighted moving average, so that the exponential moving average has the same weight as the weighted moving average for the most recent data $(\mu = 2/(Q+1))$, as shown in Fig. 3, then we can say that a value of μ is similar to a calibration data set of length $(2-\mu)/\mu$. Thus, an update parameter of $\mu = 0.1$ is equivalent to a window length of $Q = 19$, and an update parameter of $\mu = 0.01$ is equivalent to a window length of $Q = 199$.

2.3 Bias Correction

In bias correction or first moment calibration, we adjust the mean of the ensemble, so that the average value of the mean over time is the same as the average value of the observation over time. That is, we find the value b such that

$$E(\bar{x} - b) = E(y) \tag{3}$$

Thus, the bias is the average of the forecast errors over time - dominated by flow dependent errors when averaged over a short time period, and dominated by seasonally varying errors when averaged over a long time period. In reality, we can only use past errors to estimate the bias, therefore we can only correct the forecast errors that persist over the time period. Thus, the forecast errors can be considered to have two components - a systematic or predictable component, which has little variability over time, and a random component which has large variability over time making it highly unpredictable. In performing bias correction, we are aiming to estimate the systematic, or predictable component of the forecast error, based on the past forecast errors.

The ensemble-mean biases for each single model are estimated using the moving-average estimate, and applied to every ensemble member from the same model, so that the single-model ensemble-mean is bias corrected. The bias correction is applied to the individual single-model ensembles before the ensembles are combined; this allows for the fact that the models might have different biases, and ensures that the multi-model variance is not artificially inflated due to different biases.

2.4 Combination

The multi-model ensemble is given by the union of the ensemble members from the single model ensembles. From this multi-model ensemble, we can then derive probabilities and ensemble mean and variance, which can be written in terms of the single-model ensemble values.

The multi-model ensemble probability

$$p_{MM} = \frac{1}{M} \sum_k w_k p_k \tag{4}$$

is given by a weighted average of the single-model probabilities, $p_k = \frac{1}{N_k}\sum_k o_k^i$, where o_k^i is the binary forecast of whether or not the event will occur, based on ensemble member i from model k. The multi-model ensemble mean

$$\bar{x}_{MM} = \frac{1}{M}\sum_{k=1}^{M} w_k \bar{x}_k \tag{5}$$

is given by a weighted average of the single-model means, $\bar{x}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} x_k^i$, and the multi-model ensemble variance

$$\sigma_{MM}^2 = \frac{1}{M}\sum_{k=1}^{M} w_k \sigma_k^2 \tag{6}$$

is given by a weighted average of the single model variances - around the multi-model ensemble mean. These variances can be also be written in terms of the between-model and within-model variances,

$$\sigma_k^2 = \frac{1}{N_k}\sum_{i=1}^{N_k}(x_k^i - \bar{x}_{MM})^2 \tag{7a}$$

$$= \zeta_k^2 + \nu_k^2 \tag{7b}$$

where

$$\zeta_k^2 = (\bar{x}_k - \bar{x}_{MM})^2 \tag{7c}$$

is the between-model variance for model k, which is the squared distance between the single model mean and the multi-model mean, and

$$\nu_k^2 = \frac{1}{N_k}\sum_{i=1}^{N_k}(x_k^i - \bar{x}_k)^2 \tag{7d}$$

is the within-model variance for model k, which is the variance of the single model members around the single model mean.

2.5 Measure of similarity

To define the weights and variance adjustment, it is useful to have an objective measure of the similarity between models. Here, we define a measure that is computed for each grid point using moving-average estimates. If we assume that the single-model ensemble means have similar level of skill, then the average between-model variance gives an indication of the similarity between the forecast errors. If the forecast errors are exactly the same for the two models, then the between-model variance would be zero. To obtain an absolute measure of the similarity, then we compare the between-model variance with the mean-square-error of the multi-model ensemble mean. That is, we are aiming to assess the proportion of the phase-space of uncertainty that is spanned by the contributing models.

The measure of similarity, S, is defined as

$$S = \frac{E(\zeta^2)}{E(\zeta^2) + MSE} \tag{8}$$

where MSE is the mean-square-error of the multi-model ensemble mean and $E(\zeta^2)$ is the average between-model variance.

$$E(\zeta^2) = \frac{1}{M}\sum_k E(\zeta_k^2) \tag{9}$$

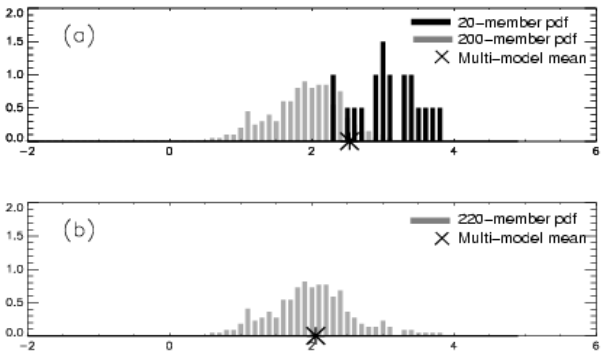


Figure 4: *Illustration of two alternatives for combining ensembles where model 1 has 200 members and model 2 has 20 members. In (a) we consider the members to be sampled from two distinct distributions and the multi-model mean is given by the average of the single model means, and in (b) we consider the members to be sampled from a single distribution and the multi-model mean is given by the average of the members from both models.*

S ranges from zero to one, with a smaller value indicating a larger similarity. When the models are identical, then the average between model variance is zero, so that $S = 0$. When the models are different to one another and are spanning a relatively large part of the true uncertainty, then the between-model variance will be large in comparison to the MSE so that $S = 1$.

2.6 Weights

The weights can be viewed as the probability of a particular model being the best model, and aim to give more weight to the most skillful model. There are various methods to define these weights. One method is to use multiple-regression based weights, as used by Krishnamurti et al. (1999); Kharin and Zwiers (2002); Doblas-Reyes et al. (2005); Stephenson et al. (2005). Although this is a well-defined method for computing a weighted mean, it is not straightforward to extend this to weighted probabilities. In particular, care needs to be taken in using the regression coefficients to define the weights (Stefanova and Krishnamurti, 2002).

A competing method to compute the weights is Bayesian model averaging (BMA) (Raftery et al., 2005; Wilson et al., 2007), where the weights and variances are computed simultaneously through a maximum likelihood estimation which aims to fit the pdfs to the calibration data.

A much more simple method is to use a skill-based method where the weights are dependent on a measure of forecast skill. This method is commonly used in the generation of consensus forecasts (Woodcock and Engel, 2005). Despite the simplicity of the method, the derived weights are surprisingly similar to those derived from multiple-regression and BMA, as seen in the studies by Raftery et al. (2005); Johnson (2006). Therefore we choose to use the skill-based method here.

The definition of the optimal weights is based on both the skill of the models and also on the similarity between the models. The reason why we need to consider the similarity between the models is because the component ensembles have different numbers of ensemble members. If we consider the definition of the multi-model ensemble mean, then one way to create the mean would be to take an average of the three single model ensembles, so that the multi-model ensemble mean is an average of the single model ensemble means. An alternative would be to take an average of all the ensemble members, so that the model with more members will get more weight. If the models have the same

number of members, then both methods will give the same result.

Figure 4 illustrates the difference between these two alternatives. In both (a) and (b), ensemble members have been drawn from two Gaussian distributions with different means, but the same variance (0.5). Ensemble 1 has 200 members drawn from a distribution with mean 2 and ensemble 2 has 20 members drawn from a distribution with mean 3. In reality we don't know the underlying parent distributions of the ensemble members, as we only have a small sample of members. If we assume that the members are sampling from different parent distributions, so that the models are providing different samples, as in (a), then the members from ensemble 2 provide more information than those from ensemble 1, so that the multi-model-mean should be formed by the average of the means of the single-model means. If we assume that the members from both ensembles are actually from the same parent distribution, as in (b), then we can assume that all members should be treated equally, so that the multi-model mean should be formed by the average of the members.

Another possibility is that the multi-model mean lies in between these two extremes. This can be achieved using the measure of similarity to define an effective number of members for each single-model ensemble.

In Eq(5), we formulated the multi-model ensemble mean as a weighted combination of the single model ensemble means. Here, we define the weight given to each ensemble mean as

$$w_k = \frac{N_k}{\tilde{T}_k} \frac{\gamma}{MSE_k} \tag{10}$$

where MSE_k is the mean-square-error of the bias-corrected ensemble mean for model k, and γ is a normalization factor to ensure that $\frac{1}{M}\sum w_k = 1$. N_k is the number of members in model k and we define $\tilde{T}_k$, below, as an effective average number of members.

In the case where the models are dissimilar, and hence providing members from different samples, then we will have $\tilde{T}_k = N_k$, the effective average number of members is the same as the number of members in that model, so that the weight is only based on MSE_k. Together with (5) this means that the multi-model ensemble mean is given by a weighted combination of the single-model ensemble means.

In the case where the models are identical, and hence providing members from the same sample, then we will have $\tilde{T}_k = T$ where

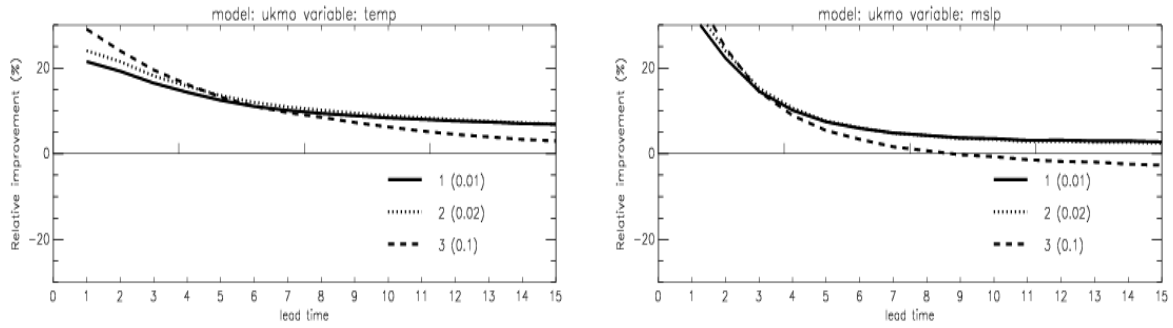
$$T = \frac{1}{M}\sum_k N_k \tag{11}$$

is the actual average number of members in each model. If the number of ensemble members in that model N_k is larger than the average number of members, then more weight is given to that single-model ensemble mean. This means that the multi-model ensemble mean is given by a weighted combination of all the ensemble members.

The measure of similarity is used to define the effective average number of members. Based on experience, we have found that better results are obtained if the similarity values are skewed more towards 0 and 1, which can be achieved by transforming the measure of similarity. There are a variety of ways in which to do this; we have chosen to use a formula similar to the tanh function. The transformed similarity measure is

$$\tilde{S} = \frac{1}{2}\left(\frac{\theta^{\lambda(S-1/2)} - 1}{\theta^{\lambda(S-1/2)} + 1} + 1\right) \tag{12}$$

and we have chosen to use the values $\theta = 4$ and $\lambda = 10$. This formula gives values of $\tilde{S}$ are similar to S in that when the models are similar, then $S = \tilde{S} = 0$ but when the models are dissimilar then $S = \tilde{S} = 1$.



(a) 2m temperature

(b) mslp

Figure 5: *Percentage improvement in globally averaged RMS errors for (a) 2m temperature and (b) mslp, predicted by the Met Office ensemble mean, as a function of lead time. The upper plot shows the RMS errors and the lower plot shows the difference in RMS errors from that of the raw ensemble mean. The bias correction uses $\mu = 0.01$ (solid), $\mu = 0.02$ (dotted) and $\mu = 0.1$ (dashed). The data are verified against multi-model analyses over a 250 day period ending on 29 April 2008.*

The effective number of members for each model is then defined as

$$\tilde{T}_k = (1 - \tilde{S})T + \tilde{S}N_k \tag{13}$$

If the models are similar then $S = 0$ so that $\tilde{T}_k = T$. i.e the effective average number of members is identical to the actual average number of members. If the models are dissimilar, then $S = 1$ so that $\tilde{T}_k = N_k$, the effective number is equal to the number of members for that model.

2.7 Variance Adjustment

The aim of variance adjustment is to correct the second moment of the pdf. In a similar way to equation (3) for bias correction, We require,

$$E(x^i - \bar{x})^2 = E(y - \bar{x})^2 \tag{14}$$

where x^i is a member of the multi-model ensemble and $\bar{x}$ is the multi-model ensemble mean. This is equivalent to requiring that the multi-model ensemble variance is, on average, equal to the mean square error of the multi-model ensemble mean. This is achieved by adjusting both the between-model variance and the within-model variance simultaneously,

$$x_k^i = \bar{x}_{MM} + \beta(\bar{x}_k - \bar{x}_{MM}) + \alpha_k(x_k^i - \bar{x}_k) \tag{15}$$

where β adjusts the between-model variance and α_k adjusts the within-model variance for model k. Note that we now assume that the single model ensemble members and means, x_k^i and $\bar{x}_k$, are bias-corrected. Although the variance adjustment alters the means of the component ensembles, it does not alter the mean of the multi-model ensemble.

Based on Eq(6) this gives a new calibrated multi-model variance

$$\sigma_{MM}^2 = \frac{1}{M}\sum_k w_k(\beta^2\zeta_k^2 + \alpha_k^2\nu_k^2) \tag{16}$$

and hence we require

$$MSE_{MM} = \frac{1}{M}\sum_k w_k(\beta^2 E(\zeta_k^2) + \alpha_k^2 E(\nu_k^2)) \tag{17}$$

An extra constraint is required to determine the ratio of the between-model variance to the within-model variance. We have chosen to apply the following constraint.

$$E(\zeta_k^2) = S\alpha_k^2 E(\nu_k^2) \tag{18}$$

where $E(\nu_k^2) = E(x_k^i - \bar{x}_k)^2$ is the within-model variance and $E(\zeta_k^2) = E(\bar{x}_k - \bar{x}_{MM})^2$ is the between-model variance.

This constraint means that, for models that are dissimilar, the between-model variance should be equal to the within-model variance. This constraint means that the within-model variance of a single model is representative of the distance from the multi-model ensemble mean. Thus, if one model has a relatively high between-model variance - i.e. it is very different from the other models, then the corresponding ensemble members will have a large spread. This constraint should also ensure that the pdfs from the different models overlap within the multi-model ensemble, whilst retaining the individual identity of each model.

However, if the models are similar to each other, then we will expect that the between-model variance would be small, and so in this case, we weaken the constraint so that that the between-model variance is smaller than the within-model variance. In the extreme case where the models are identical, then $S = 0$ and the between-model variance is zero.

Using (18) to substitute for $\alpha_k^2 E(\nu_k^2)$ in (17) then we obtain

$$\kappa = \frac{MSE_{MM}}{\frac{1}{M}\sum_j w_j E(\zeta_j^2)} \tag{19a}$$

$$\beta^2 = \frac{\kappa S}{1+S} \tag{19b}$$

$$\alpha_k^2 = \frac{\kappa}{1+S}\frac{E(\zeta_k^2)}{E(\nu_k^2)} \tag{19c}$$

where κ is the ratio between the mean-square error of the multi-model ensemble mean and the overall between-model variance.

β^2 is large if the expected between-model variance is too small relative to the mean-square-error, and α^2 is large if the within-model variance is too small compared to the adjusted between-model variance.

2.8 Summary

In summary, we have used moving-average estimates of the ensemble statistics to define the biases of the single model ensembles, the model-dependent weights, and the parameters to simultaneously adjust the between and within model variances, and hence the overall multi-model ensemble variance. Both the weights and the variance adjustment parameters have made use of a measure of the similarity of the ensembles, again defined from moving-average estimates.

The moving-average estimates use past data to define the parameters. This means that the calibration data ages more when calibrating the forecast at longer lead times. We will see in the results section that this ageing means that it is harder to calibrate the ensemble at longer lead times.

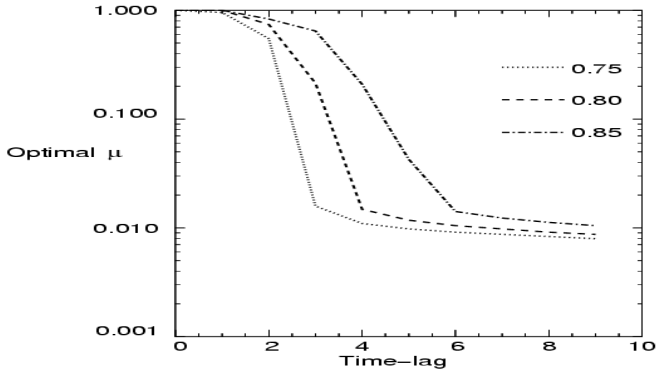


Figure 6: *The optimal value of μ as a function of the time-lag for three timeseries with different autocorrelation values: 0.75,0.8 and 0.85.*

3 Results

We now present the results from the multi-model combination and calibration applied to the three global model ensembles. We begin by examining the impact of the bias correction that is applied to the single model means. This is followed by the results from the simple and weighted combination methods, and finally by the results from the variance-adjustment.

3.1 Bias correction

The results for the bias correction of Met Office ensemble mean are first presented. The bias correction is also applied to the NCEP and ECMWF ensembles means, and as the overall conclusions are similar, the results are not shown here. The bias correction is repeated with three different value of μ: 0.01, 0.02 and 0.1, and a estimate of the bias is found for each grid point and each lead time, for the three variables under consideration.

The differences between globally averaged RMS errors for the raw and bias corrected 2m temperature are shown in Fig. 5a. All three bias correction methods give a reduction in the RMS error at all lead times. At short lead times (less than 5 days), the lowest RMS errors are achieved with $\mu = 0.1$, with a 30% improvement and at longer lead times (greater than 5 days), the lowest RMS errors are achieved with $\mu = 0.02$, giving a 5% improvement.

Similar results are found for mslp (Fig. 5b). At short lead times (less than 3 days), the lowest RMS errors are achieved with $\mu = 0.1$; at medium lead times (between 3 and 7 days) the lowest RMS errors are achieved with $\mu = 0.02$; and at longer lead times (greater than 7 days) the lowest RMS errors are achieved with $\mu = 0.01$. An important difference here to the 2m temperature results is that at longer lead times, the bias correction is actually detrimental to the forecast if the value of μ is too large.

In summary, the optimal value for μ depends on both the lead time, and on the variable: a smaller value of μ should be used for longer lead times, and for mslp; a large value of μ should be used for shorter lead times and for 2m temperature.

3.1.1 Interpretation of bias correction results

To give further insight into these results, we consider some idealized experiments based on autoregressive timeseries. We assume that the true bias is given by a stationary, first-order autoregressive AR(1)

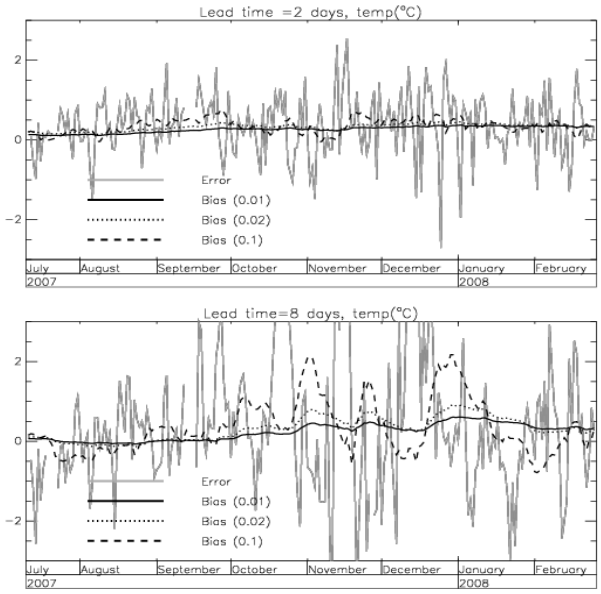


Figure 7: *2m temperature forecast error timeseries from the Met Office ensemble at point* $(50°N, 0°E)$*, at lead times of 2 and 8 days over the period July 2007 to February 2008.*

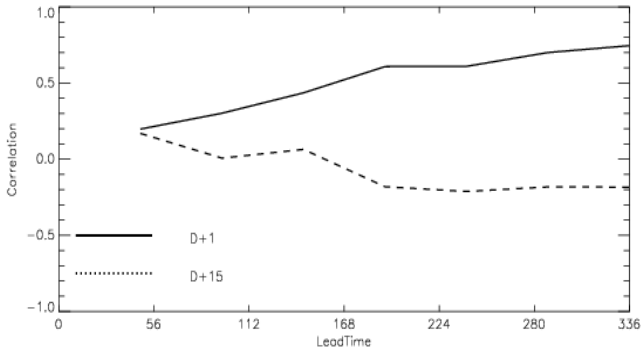


Figure 8: *Autocorrelation values for 1 and 15 day time lags, as a function of lead time. The data is the 2m temperature forecast error timeseries from the Met Office ensemble at point* $(50°N, 0°E)$

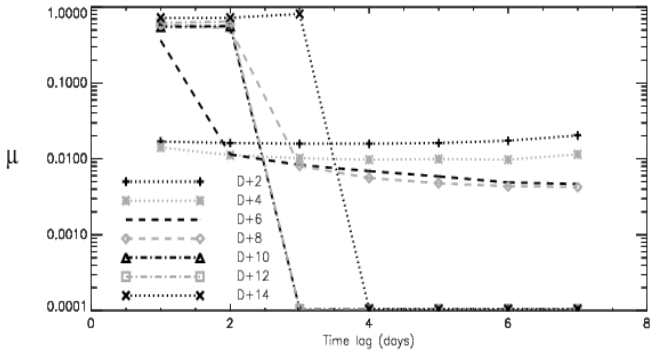


Figure 9: *The optimal values of μ computed for timeseries from various lead times (or forecast ranges, D+2 to D+14) and for bias correction using a range of time-lags for the most recent observation. The data is the 2m temperature forecast error timeseries from the Met Office ensemble at point $(50°N, 0°E)$ over 250 days of data ending on 2008/02/25 with the optimal value for μ based on verification over the last 120 days of data.*

process (e.g. Papoulis and Pillai, 2002; Smith and Krajewski, 1991), about mean $\bar{b} = 0$,

$$b^n = \rho b^{n-1} + W^n \tag{20}$$

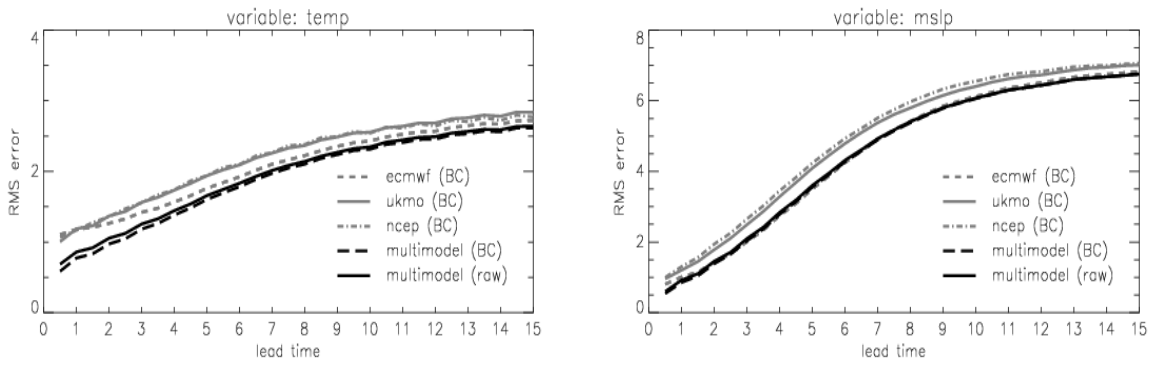
where $\rho = E(a^n a^{n-1})$, known as the autocorrelation or temporal correlation, is the expected value of the bias with a time-shifted version of itself, and the Wiener process W^n is a random number drawn from a Gaussian distribution with variance, $V = \sigma^2(1 - \rho^2)$, where σ^2 is a specified scalar equal to the variance of b^n.

Five randomly-generated timeseries are created with 2000 data points in each timeseries. The optimal value of μ is selected based on the RMS errors over the last 120 data points of these bias estimates for these timeseries. The results are shown for three different autocorrelation values in Fig. 6. As expected, the optimal value for μ decreases with increasing time-lag. But also, the optimal value for μ increases with increasing autocorrelation. Thus, if there is a high temporal correlation in the data, then the estimate will draw close to the most recent data; if there is a low temporal correlation, then the estimate will draw close to the long term average.

We then consider a forecast-error timeseries of the Met Office ensemble mean at a single grid-point. Timeseries of the 2m temperature forecast error at $(50°N, 0°E)$ are shown in Fig. 7. The errors are highly variable over time, and have a larger magnitude with increasing lead time. Also shown are the estimated biases using the three different values of μ. With $\mu = 0.01$, the bias varies very slowly over time, whereas for $\mu = 0.1$, the bias has larger amplitudes and is more variable. The time lag between large errors and large biases are noticeable for $\mu = 0.1$. This time lag means that sometimes the estimated bias has the opposite sign to the actual error.

Figure 8 shows the correlation values of the timeseries with time-shifted versions of itself, where the time-shift is either 1-day ($E(b_s^n - \bar{b}_s)(b_s^{n-1} - \bar{b}_s)$) or 15-days ($E(b_s^n - \bar{b}_s)(b_s^{n-15} - \bar{b}_s)$). For a 1-day time-shift, we see that the correlation increases with increasing lead time. For a 15-day time-shift we see that the correlation decreases with increasing lead time, such that at long lead times, there is actually a negative correlation. This contrast between the 1-day and 15-day shifted correlations means that the temperature timeseries is more complicated than a simple AR process. However, we can still apply the general principles derived from the AR process experiments.

From the AR process experiments we found that if the timeseries has a large autocorrelation then a relatively large value of μ should be used, whereas if the timeseries has a small autocorrelation, then



(a) 2m temperature

(b) mslp

Figure 10: *RMS errors of the (a) 2m temperature and (b) mslp, ensemble means for the three single models and for the raw and bias corrected multi-model ensemble, both using a simple combination. The data is verified globally over 250 days ending on 29 April 2008.*

a relatively small value of μ should be used. For the real data, the 1-day autocorrelation at short lead times is relatively small so that a small value of μ is required. However, the 15-day autocorrelation at long lead times, is actually negative, so that an even smaller value of μ is required.

As the 1-day autocorrelation for the long lead time timeseries is relatively large, we would expect to see that if we applied bias correction using more recent data (of course this is not possible in practise), that we would use a larger value of μ than for the short lead time data. Similarly we would expect that if we applied bias correction to the short lead time timeseries using older data, that a smaller value of μ would be required, because the autocorrelation for a 15-day time lag is smaller (perhaps negative) than for a 1-day time lag.

The results from these experiment are shown in Fig. 9. The dotted line with crosses shows the optimal values for μ for the 15-day timeseries computed with a range of recent data times. If the recent data is only from the previous day then the optimal value for μ is close to one, whereas if the recent data is from longer lead times, then the optimal value for μ is much smaller (here, taking the lowest value tried, 0.0001 - in fact there may be no value of μ that actually gives an improvement over the raw data). For the 6-day timeseries, the optimal value for μ when the most recent data is from the previous day is about 0.4 (lower than for the 15-day timeseries) because the data has a smaller autocorrelation). This value then reduces as the most recent data increases to 7 days. This is because the autocorrelation for say a 4-day time lag is smaller. For the 2-day timeseries data, there is little change in the optimal value for μ. There is perhaps a slight increase, but this may be due to the small sample size.

In conclusion, we see that a smaller value of μ is required at longer lead times. In fact, at longer lead times the 1-day autocorrelation is larger, so we might expect a larger value of μ to give better results. Indeed, if we could use observations of the future, then we would be able to use a large value of μ and subsequently obtain a better bias correction at longer lead times. However, as there is a larger time lag from the most recent observation, and there are smaller autocorrelation values for larger time-shifts, a smaller value of μ gives better results.

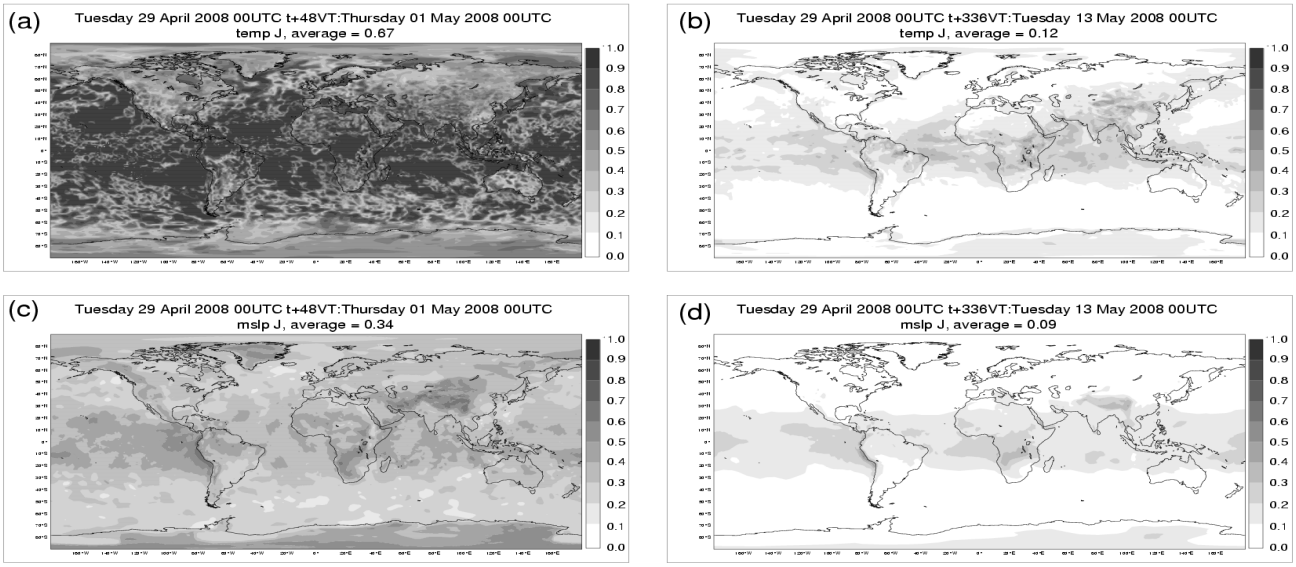


Figure 11: *Plots of similarity, S, on 29 April 2008, for 2m temperature (top: a,b) and mslp (bottom: c,d), at T+48 (left: a,c) and at T+336 (right: b,d). S is defined by equation (8). High values (shaded grey) indicate that the models are dissimilar, low values (shaded white) indicate that the models are similar.*

3.2 Simple combination

The aim of a multi-model ensemble is to give a better representation of the uncertainty, arising from both the initial conditions and from the forecast model. We might expect that the different ensembles might span different parts of phase space, according to their different biases, and hence that these biases would cancel each other out in the multi-model ensemble mean. Therefore, we address the question of whether a raw multi-model ensemble-mean is better than a bias-corrected single-model ensemble-mean. And further, whether a bias-corrected multi-model ensemble-mean is better than a raw multi-model ensemble-mean.

For all the experiments in this subsection we assume that the weights given to each model are equal, $w_k = 1$. We also apply bias correction using $\mu = 0.01$. Figure 10a shows the RMS errors of the ensemble means for the three bias corrected ensembles, and also for the raw and bias-corrected multi-model ensemble, for 2m temperature. We see that the three component-model ensembles have similar errors, with the ECMWF ensemble having a slightly better performance.

The raw multi-model ensemble is the combination of the raw ensembles, and has even lower RMS errors than any of the bias-corrected component model ensembles. Thus the benefits from the multi-model combination outweigh the benefits from bias correction. Figure 10a also shows a bias-corrected multi-model ensemble, which is the combination of the bias-corrected ensembles. This has slightly smaller RMS errors, particularly at the shorter lead times. However, the impact of the bias correction on the multi-model ensemble is far less than on the single model ensembles.

The effect of the multi-model ensemble combination for mslp is shown in Fig. 10b. Both the raw and bias-corrected multi-model ensembles have very similar RMS errors to the ECMWF ensemble.

These results show that the impact of multi-model combination is equal to or better than the impact of bias correction. This means that the impact of bias correction is not only to cancel the biases between the two models. There are a number of other reasons that the multi-model ensemble is giving a better performance over the bias-corrected single-model ensemble. The first is that the multi-

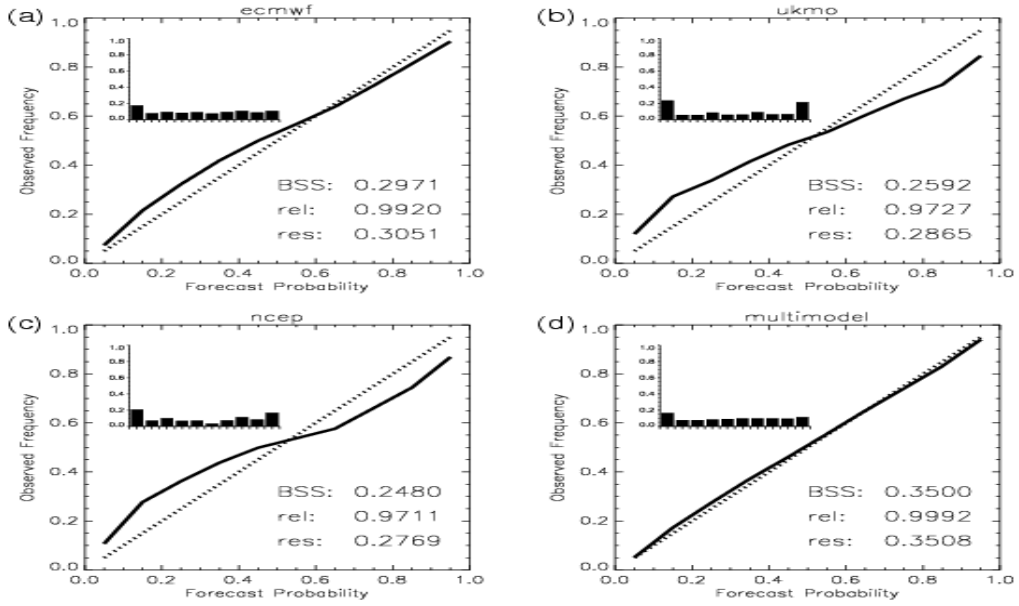


Figure 12: *Reliability and sharpness diagrams for 2m temperature greater than the climatological mean, at T+240. The bias-corrected single model ensembles for (a) ECMWF, (b) Met Office and (c) NCEP and the bias-corrected multi-model ensemble (d) are shown. The data is globally averaged over 120 days until 29 April 2008.*

model ensemble is cancelling errors in the model that can not be predicted using bias correction. That is, the multi-model ensemble is cancelling out the random, unpredictable components of the errors. The second reason is that the multi-model ensemble always performs better than the worst model. As the three ensembles have similar levels of skill, then if assume that the identity of the worst model changes for different situations, then the multi-model ensemble will have better skill.

These reason show that the multi-model ensemble gives more benefit for 2m temperature than for mslp. For both variables, the three single models have similar levels of skill although, again for both variables, the ECMWF model can be identified as having slightly better skill. Thus, the reason can not be solely attributed to the fact that a best model can be identified and we must concluded that it is also associated with the similarity of the models. If the models are identical, then there can be no error cancellation. However, if the models are dissimilar, then the multi-model ensemble give benefits from the cancelling of errors between the models. It is likely that there is less similarity between the 2m temperature forecasts than the mslp forecasts, over both time and space. This is because there is a larger impact on 2m temperature from the model parameterizations such as land-surface schemes, of which there is a large variety between different forecast models (e.g. Pitman et al., 1999).

It is important to emphasise the difference between the similarity of the forecasts, and the similarity of the forecast skill. In fact, for the multi-model combination to give an improvement over the single model, it is important for the models to have similar levels of forecast skill. So, in conclusion, it is necessary for the models to have dissimilar forecasts, but similar levels of forecast skill. This conclusion has also been reached by Hagedorn et al. (2005), who summarised that "the key to the success of the multi-model concept lies in combining independent and skillful models, each with its own strengths and weaknesses".

3.2.1 Similarity

The results from the simple combination showed that there is a greater benefit from the multi-model ensemble for 2m temperature than there is for mslp. This difference was partly attributed to the differences in the similarities of the forecast errors, so that in comparison to mslp, 2m temperature has less similarity between the forecast errors, so that there is more cancelling of errors.

The plots in Fig. 11 show the similarity values for mslp and 2m temperature, at two times T+48 and T+336. The dark shading indicates regions that have a large value of S (models are not similar) and the light shading indicates regions that have a small value of S (models are similar).

We see that the forecasts for mslp are more similar than for 2m temperature at both T+48 and T+336. The dissimilarity between the temperature forecasts can perhaps be attributed to the differences in the land surface schemes in the two models, whereas mslp will be governed more by the dynamical cores, which are perhaps more similar.

We also see that the models at T+336 are more similar than at T+48. This shows that there is a good spread in analyses at the initial time, but as the forecasts evolve, they develop the same errors so that the forecasts become similar to each other and do not span the full phase space.

The plots also show that the forecasts exhibit more similarity in the extratropics, than in the tropics. This is perhaps because the errors in the extratropics are dominated by dynamical errors, and there are very little differences in the dynamics of the three models, whereas errors in the tropics are more likely to be a result of differences in convection schemes, which have more differences between the three models.

3.3 Brier Skill Scores

We now look at the effect of the combination on the probabilistic prediction of the ensemble. We compute reliability tables based on probabilities above climatological thresholds such as the mean, or the 90^{th} percentile. This climatology data is based on the ERA-40 reanalysis database, and provided by M. Leutbecher, ECMWF. If we were to use fixed thresholds then we would expect to see false skill in the Brier skill scores, because for example the temperatures over the tropics will always be greater than $0^{\circ}C$, and hence the forecasts are always correct (Hamill and Juras, 2006). By using climatological thresholds and assuming that the model climatology is the same as the ERA-40 climatology, we can neglect this effect, and average the verification data globally, giving a large sample size.

The reliability and sharpness diagrams are shown in Fig. 12. The three single models are showing an over forecasting bias for low frequencies, and an under-forecasting bias for high frequencies, as a result of poor resolution. The ensembles also have a relatively high sharpness, with large frequencies at the two extremes. The multi-model ensemble has less sharpness, as indicated by a flatter sharpness graph, but has a higher resolution and reliability, as indicated by the solid line (actual) being closer to the dotted line (ideal). These improvements in both the reliability and the resolution lead to a higher overall Brier skill score.

The corresponding Brier skill scores for the range of ensembles are shown as a function of lead time in Fig. 13. The Brier skill scores for the three single models, shown in grey, decrease with increasing lead time, as expected. The scores are similar for all three ensembles, although the ECMWF is slightly better. The multi-model, shown by the dotted black line and lying underneath the solid black line, has a much better performance, at all lead times, with a 1-day increase in predictability at day 7 (the skill of the multi-model at day 8 is the same as the skill of the best single model at day 7). This improvement is also seen in both the reliability and the resolution components.

The Brier skill scores for mslp are shown in Fig. 14. In this case the ECMWF model can be clearly

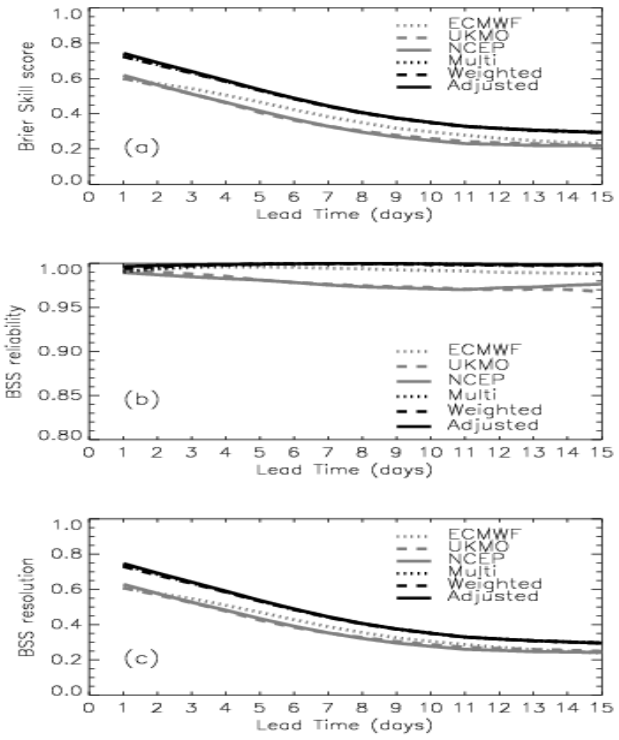


Figure 13: *(a) Brier skill score, (b) reliability component and (c) resolution component, for 2m temperature greater than the climatological mean as a function of lead time. The grey lines show the bias-corrected single-model ensembles (ECMWF, Met Office and NCEP) and the black lines show three different multi-model ensembles: simple combination (dotted), weighted (dashed), weighted and variance-adjusted (solid). The data is globally averaged over 120 days until 29 April 2008.*

identified as the best model, and the multi-model Brier skill scores give only a slight improvement on the ECMWF model.

The Brier skill scores for 2m temperature, but this time using the 90^{th} percentile as the threshold rather than the mean, are shown in Fig. 15. In a similar way to the scores for temperature using the mean threshold, the multi-model gives a considerable improvement at all lead times, with a 2-day increase in predictability at day 7.

3.4 Weighted combination

We now examine the impact of model-dependent weights. The mean-square-error, used to calculate the weights, is estimated using a running mean with update parameter $\mu = 0.01$.

The γ/MSE_k component of the weights for 2m temperature are shown in Fig. 16, for two different lead times. The values of γ are chosen so that these components sum to 3. To make a clear assessment of the individual model skills, we have not included the $\frac{N_k}{\tilde{T}_k}$ component, that accounts for the similarity and the number of members, in the plots. However, it is included in the weighted multi-model ensemble verification results. The grey and black shaded regions show the regions in which the most weight is given, whereas the white and dotted regions show the regions in which less weight is given. In general, we see that more weight is given to the ECMWF model in the extratropics, whereas more weight is given to the NCEP model in the tropics. We also see ocean/land contrasts so that more weight is given to the ECMWF model over the Amazon, in comparison to the other two models, and flow-dependent

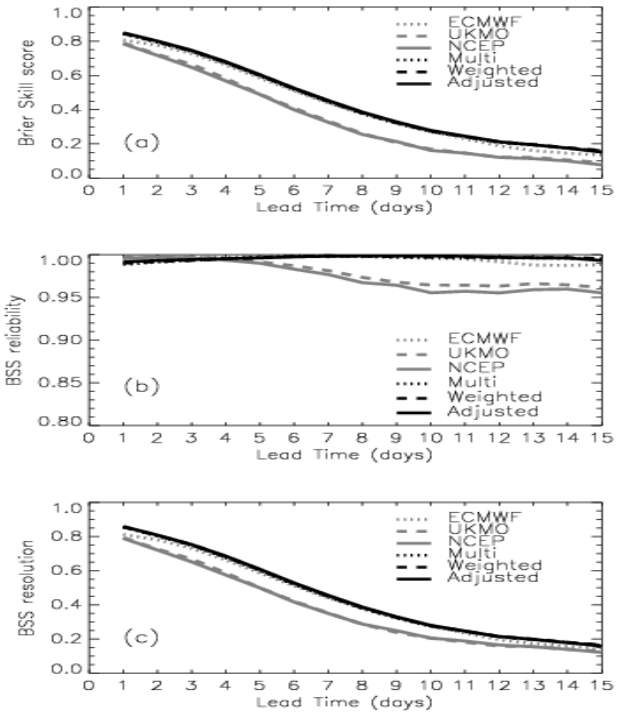


Figure 14: *As Fig. 13 but for mslp greater than the climatological mean.*

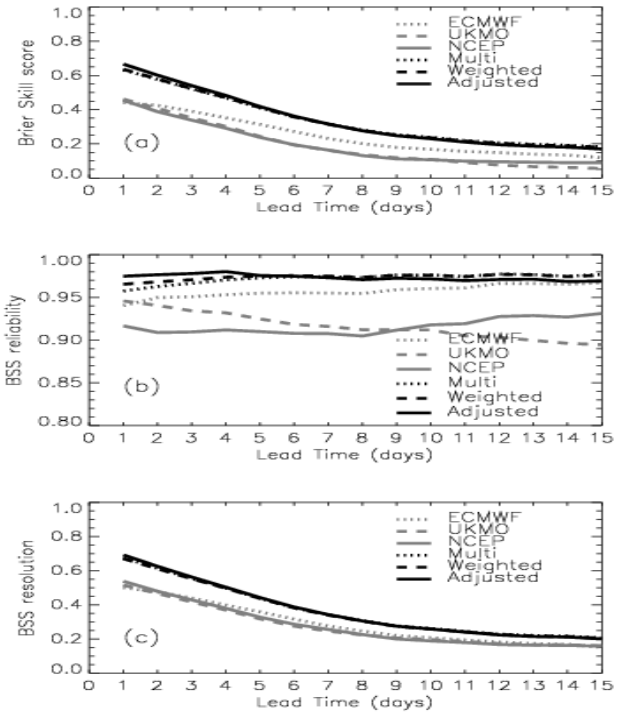


Figure 15: *As Fig. 13 but for 2m temperature greater than the* 90^{th} *percentile.*

Exp.	Description	Weight given to single model means
MM1	combine means with equal weights	$w_k = 1$
MM2	combine means with unequal weights (based on RMS error)	$w_k = \frac{\gamma}{MSE_k}$
MM3	combine members with equal weights	$w_k = \frac{N_k}{T}$
MM4	combine members with unequal weights (based on RMS error)	$w_k = \frac{N_k}{T}\frac{\gamma}{MSE_k}$
MM5	combination based on similarity and RMS error.	$w_k = \frac{N_k}{\tilde{T}_k}\frac{\gamma}{MSE_k}$

Table 2: *Different combination methods. w_k is the weight given to the ensemble mean for model k, MSE_k is the mean-square-error of the bias-corrected single model ensemble mean, γ is a normalization factor to ensure that the weights sum to the number of models, N_k is the number of members for model k, T is the average number of members (eq(11)), and $\tilde{T}_k$ is the effective average number of members, for model k (eq(13)).*

differences such as the in the ITCZ region over the eastern Pacific, where more weight is given to the ECMWF model than the Met Office model.

Generally, the weights are relatively close to one and the probabilities given by the component models are also relatively similar to one another, so the application of the weights make only a slight modification to the probability values, and hence only a slight modification to the overall verification scores. The impacts of the model-dependent weights on the Brier skill scores are shown in Figs. 13 to 15, by the dashed line. For the mean thresholds, the model-dependent weights have very little impact. However, more of an impact is made for the 90^{th} percentile threshold (Fig. 15). In particular, at short lead times, the model-dependent weights make an improvement in the ensemble reliability, whereas at longer lead times, the model-dependent weights are detrimental to the ensemble reliability. This negative impact at longer lead times is most likely due to a similar effect that was seen in the the bias correction results, in that there is a longer time-lag between the most recent data and the forecast time.

3.4.1 Comparison of different weights

The weights that have been specified in the multi-model ensemble are based on both the skill of the model and a measure of the similarity between the models. We now investigate the impact if a different set of weights are used. The five different weighting methods are listed in table 2. The results are assessed using the RMS errors of the multi-model ensemble means. The difference in the RMS errors of the multi-model ensemble means from using different model weightings to those from a simple combination of the single model means (MM1), are shown in Fig. 17.

For mslp (Fig.17a), we see that the RMSE difference for MM2 is negative at all lead times, with larger differences at short lead times, showing that the weighting using the RMS errors is giving an improvement over giving equal weight to each model. Similarly, we see that MM4 has better skill than MM3, showing that when combining the ensemble members, the MSE based weights also give an improvement. At all lead times, and for both equal and MSE-based weights, the combination of the ensemble members gives lower RMS errors than the combination of ensemble means. For temperature (Fig.17b), we see again that MM2 has better skill than MM1,and that MM4 has better skill than MM3, showing that the MSE-based weights are giving an improvement. However, in contrast to mslp, we see that at short lead times, the combination of members gives worse skill than the combination of the

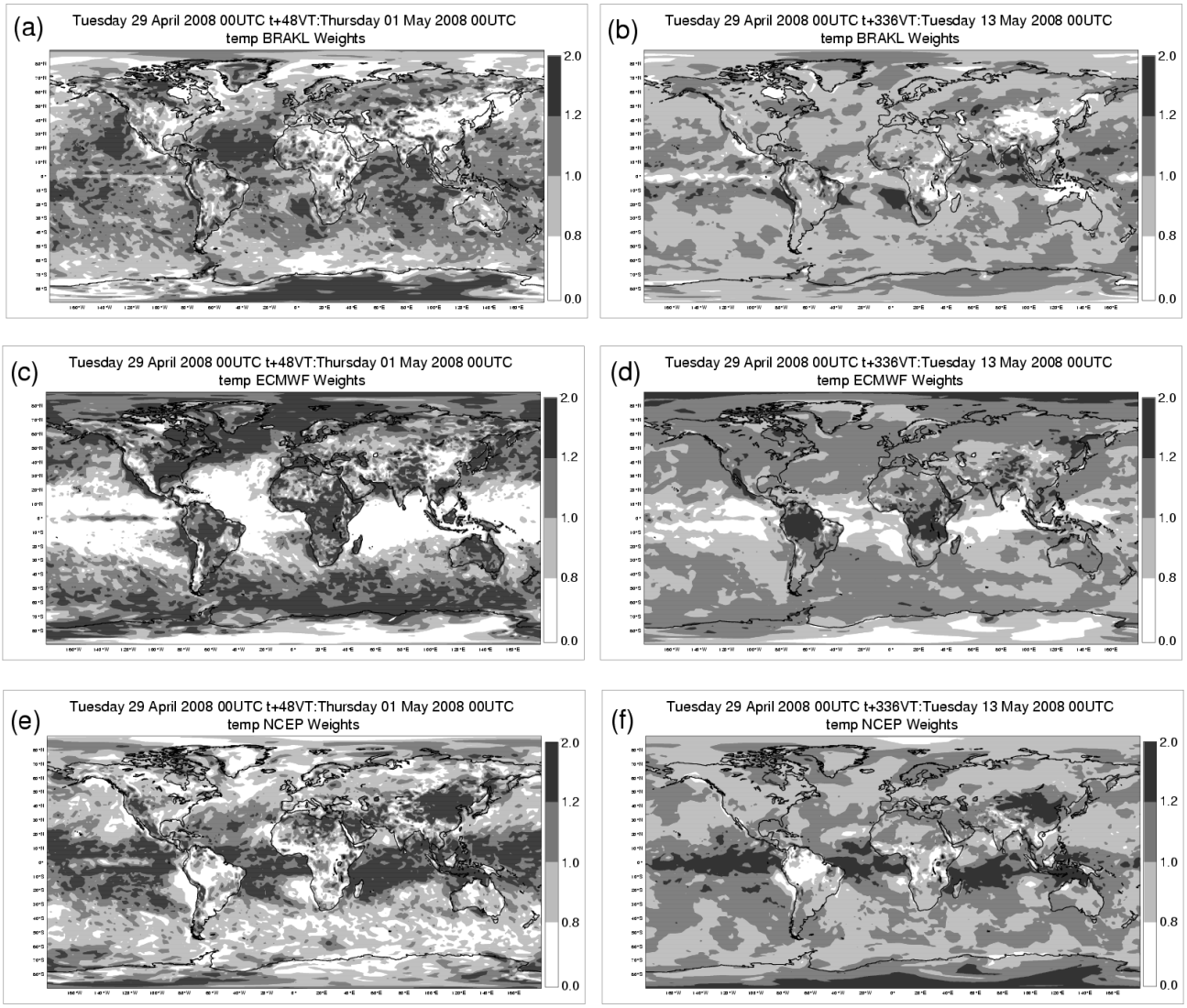
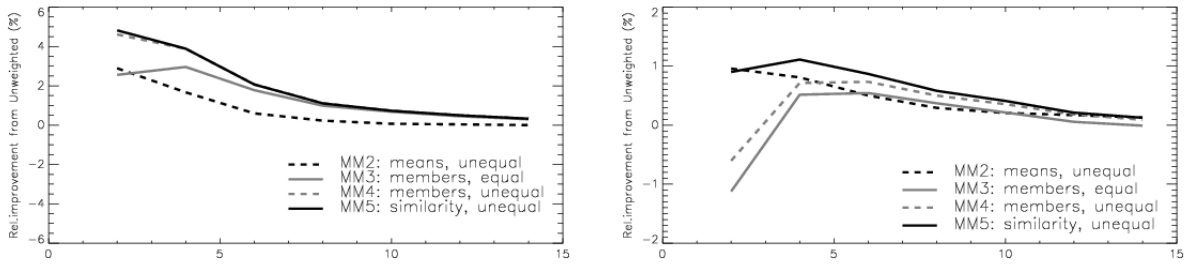


Figure 16: *Weights plots on 29 April 2008, for 2m temperature, at T+48 (left a,c,e) and at T+336 (right b,d,f) and for the three different models: Met Office (top a,b), ECMWF (centre c,d) and NCEP (bottom e.f). These weights are computed as γ/MSE_k where γ is a normalization factor.*



(a) mslp

(b) 2m temperature

Figure 17: *Percentage relative difference in RMS errors from that of the simple multi-model average (MM1) for 4 different methods for (a) mslp and (b) 2m temperature, as a function of lead time (days). Verification is globally averaged over 160 days of data ending on 29 April 2008. The grey dash line for MM4 is very close to the solid line for MM5.*

ensemble means. At long lead times, the RMS errors are similar to each other.

Thus, for mslp we see that it is better to combine the members, whereas for temperature, it is better to combine the means. This difference is due to the similarities between the forecasts. As shown in Fig.11, we saw that, in general, there is more similarity between the mslp forecasts than there is between the temperature forecasts. The combination to produce multi-model MM5 uses weights that are also based on the similarity index, as used in the multi-model that was verified with the Brier skill scores. This actually gives the best overall RMS errors for both variables. For mslp, MM5 has a similar level of skill to MM4, which was the optimal out of the 4 different combinations. For temperature, MM5 has a similar level of skill to MM2 at short lead times, and is higher than all the other combinations at longer lead times.

In conclusion, we have seen that the optimal weighting for the multi-model ensemble is based on both the skill and similarities of the component model ensembles. However, the impact of the weights is very small, with improvements in the RMS errors of up to only 4 %.

3.5 Variance Adjustment

We finally discuss the impact of the variance adjustment on the multi-model ensemble. An example of the inflation factors are shown in Fig. 18. We see that the inflation factor for the between model variance, β lies between 0.0 and 1.0 over the whole globe, showing that the between-model variance needs to be reduced, with larger reductions over the tropics. There is a different inflation factor α for the within-model variance for each model. In all three models, there is a large variance inflation over the ITCZ region over the Eastern Pacific. However, there are also regions where there are differences between the models. For example, for the Met Office model, the within-model variance is reduced in the extra tropics and at the poles, whereas for the NCEP model, the variance is reduced in the tropics. The most striking difference is that for the ECMWF model, there is a greater area over which the within-model variance is reduced. This region is mainly over the tropical and mid-latitude oceans. There are two reasons that the inflation factor for the ECMWF is smaller in these regions. The first is that the raw ECMWF ensemble generally has a larger within-model variance than the raw Met Office and NCEP ensembles. The second is that the ECMWF ensemble mean generally has a smaller between model-variance. That is, the ECMWF ensemble mean is closer to the multi-model ensemble mean than

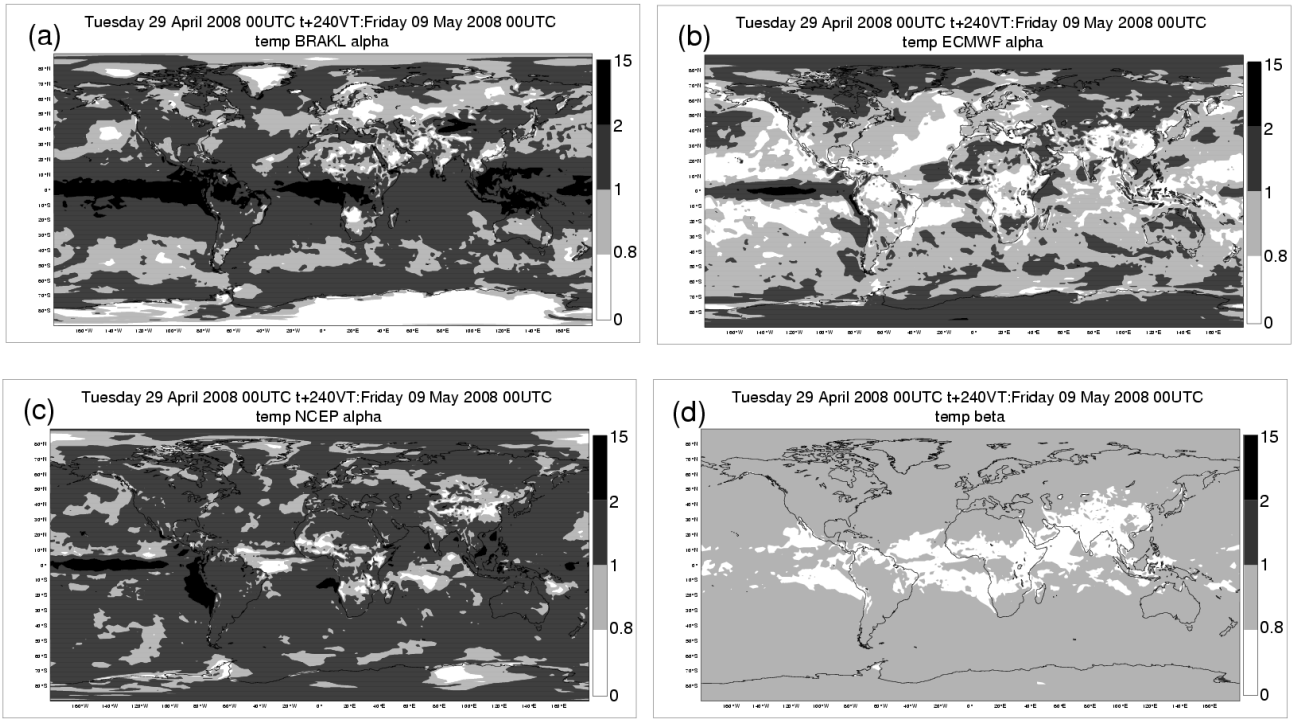


Figure 18: *Variance adjustment factors on 29 April 2008, for 2m temperature, at T+240. The within-model inflation factors α are shown for (a) Met Office (b) ECMWF and (c) NCEP and the between-model variance β is shown in (d). Values below 1 (white and light grey) indicate that the variance should be decreased, whilst values above 1, (dark grey and black) indicate that the variance should be increased.*

the other component models. From constraint 18, this results in a smaller within-model variance.

These results highlight that it is important to allow for different inflation factors for different models when adjusting the variance of the multi-model ensemble. Some ensembles may already have a much better calibrated spread than other models, and hence need to be treated differently. If every model was treated identically, then the multi-model ensemble spread might be dominated by the members from a particular model. 5

The impact of the variance adjustment is shown by the Brier skill scores in Figs. 13 to 15. In fact there is again generally little impact on the Brier skill scores. For 2m temperature greater than the 90^{th} percentile, we see an improvement in the reliability at short lead times, but a degradation at longer lead times. The reason that we don't see a significant improvement in the Brier skill scores is that the raw multi-model ensemble is already generally well calibrated, so that the variance adjustment is only making small adjustments. At short lead times, the calibration data remains a good sample so that it is possible to use the calibration parameters to make an improvement. However, at long lead times, the calibration data is no longer a good sample, and as only small adjustments are required, these adjustments actually make the skill worse. Following the interpretation of the bias correction results, it is likely that we would no longer degrade the forecast if a smaller value of μ were used in the moving-average. However, we are still unlikely to make any significant improvement.

4 Discussion

4.1 Benefits of multi-model ensembles

We have seen that the multi-model ensemble does give benefits over a single-model ensemble, with larger improvements for 2m temperature than for mslp. Although not shown, the results for 500hPa height are similar to those for mslp, showing a small improvement from the multi-model combination. These small benefits for 500hPa height are in agreement with the studies by Park et al. (2008). However, we should emphasise here that the small benefits of the multi-model ensemble for 500hPa height does not necessarily mean that the multi-model ensemble gives small benefits for all variables. Indeed, we have found here that the benefits are much larger for 2m temperature.

This returns us to the question of why a multi-model ensemble should give benefits over a single model ensemble. For a multi-model ensemble to have better skill than the component model ensembles then we require that the component model ensembles have similar levels of skill but dissimilar types of forecast errors, as noted by Hagedorn et al. (2005). This ensures that the errors cancel each other out when combined in the multi-model ensemble. We have argued in this paper that although the component models have similar levels of skill, they also often have similar errors.

The issue of similarity between the component-model forecast errors has also been discussed by Tebaldi and Knutti (2007) in the context of climate prediction, who concluded that although some errors in a model might be random, others are the result of our limited understanding of the process and our ability to parametrize them efficiently in coarse resolution models.

For some variables, we have seen that the benefits of the multi-model ensemble are small, which draws us to the question of whether resources should be put into improving the best ensemble, or developing multi-model ensembles. Buizza et al. (2003) found that a higher resolution single-model ensemble performed better than a low resolution poor mans ensemble, so it would seem that it is possible for a single model ensemble to outperform a multi-model ensemble containing poor models. However, even with sophisticated stochastic physics schemes, it is unlikely that we will ever be able to truly represent the model uncertainty. Therefore, we should continue to pursue research in multi-model ensembles.

4.2 Future Work

The issue of similarity between models guides the future development of multi-model ensembles. If multi-model ensembles are to truly capture the forecast model uncertainty, then we must ensure that we are combining forecasts that have significant differences from each other. For example, when determining which models/forecasts to use in a multi-model ensemble then we should examine both the skill and similarity of the forecasts. It may be of value to blend forecasts from different model grid resolutions together, even if the lower resolution forecasts have slightly less skill, if this gives a better representation of the model uncertainty.

The issue of similarity also affects the choice of variables to combine in a multi-model ensemble; it is better to combine variables for which there is less similarity between models. Our understanding of model 'dynamics' is generally consistent throughout all forecast models: the forecast models have a coded representation of the same underlying physical equations. This means that there is a high degree of similarity between variables such as 500hPa height and mslp. However, we have less understanding of how to represent the model 'physics', such as convection, precipitation, and land surface processes. The greater variety between the model physics in forecast models means that there is less similarity between fields such as 2m temperature. Although not studied here, it is likely that there is also less similarity

between fields such as precipitation and sunshine duration, and therefore that we would see larger benefit from a multi-model ensemble for these fields. Fortunately, these are the type of variable that are of particular use to end users, hence further increasing the potential of the multi-model technique.

Further improvements could be made to the calibration method presented. For example, the similarity measure presented here assumes that the ensembles have similar levels of skill. In fact, if the ensembles have different levels of skill then the similarity measure may also be measuring this difference. It would be better if the similarity measure could be modified so that it measures only the similarity in the forecast errors. Another improvement would be in the definition of the model-dependent weights. In the current implementation, the moving-average estimates produce weights that have a high spatial variability, which can lead to noisy multi-model fields. Future work should aim to take into account the spatial correlations of the variables, perhaps through increased temporal filtering (using a smaller value of μ) or spatial averaging. A further problem with the calibration parameters is that they are based on the most recent data, and hence if there is regime change, they may no longer be relevant. The forecasting of such regime changes is in fact one of the prime purposes of medium-range forecasting. Therefore it would be of benefit if regime-dependent calibration parameters could be estimated, for example using a reforecasting approach (Hamill et al., 2004).

When examining the benefits of the multi-model ensemble against a calibrated single model ensemble, we only applied bias correction to the single-model ensemble. The benefit of variance adjustment on the single-model ensemble was not examined. As variance adjustment will not affect the ensemble mean, the conclusions from the RMS error results would not be affected. However, variance adjustment might improve the Brier skill scores of the single-model ensembles, hence reducing the benefit of the multi-model ensemble. This aspect should be investigated in future work.

5 Conclusions

A calibrated, global, 15-day multi-model ensemble has been developed as part of the Met Office's contribution to THORPEX. The multi-model ensemble blends together three single-model ensembles: ECMWF, Met Office and NCEP, and for three different variables: 500hPa height, mslp and 2m temperature. Calibration parameters are estimated for every lead time, variable and grid point using a moving-average of past data. These parameters are used to correct the bias of the single model ensemble means, define the model-dependent weights and apply variance-adjustment to the multi-model ensemble variance.

Verification against multi-model analyses includes RMS errors, and Brier skill scores. The results show that:

- The bias-correction studies highlighted that the calibration data used in the moving-average ages with increasing lead time. This means that a smaller update-parameter should be used at longer lead times.

- The bias-correction of the single-model ensemble mean gives up to a 30% improvement at 24 hours reducing to 5% at 15 days. However, there is less impact from the bias-correction on the multi-model ensemble mean.

- The raw multi-model ensemble mean has better skill than any of the bias-corrected single-model ensemble means, showing that multi-model combination has more impact than calibration. This improvement in skill is larger for 2m temperature than for mslp and 500hPa height.

- The multi-model combination gives up to a 2-day increase in predictability at day 7, as illustrated by the Brier skill scores for predicting 2m temperature greater than the 90th percentile.
- Different flavours of model-dependent weights were tested. The RMS errors showed that the optimal weights are those that consider both the skill and the similarities of the component-model ensembles.
- The model-dependent weights and variance-adjustment only gave an improvement up to day 6. This improvement was relatively small, and was only exhibited in the verification of extreme values. The weights and variance adjustment was detrimental at longer lead times because the calibration data had aged so that it was no longer a representative sample.

In summary, the results show that a simple combination of ensembles from different operational centres can give a significant improvement in the predictive skill. Further calibration such as bias correction and variance adjustment have been found to give further slight improvements, but the greatest benefit is from the combination of the ensembles, and hence that the combination of ensembles is a promising approach for future development. As the greatest benefits have been found for 2m-temperature, future research is focusing on the development of the multi-model ensemble for further surface 'weather-related' variables.

Acknowledgements

We thank ECMWF and NCEP for making their ensemble data available. We also thank the ensemble forecasting team at the Met Office for the development of the Met Office 15-day ensemble.

References

Bowler, N. E., A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare, 2008: The MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.*, in press.

Buizza, R., P. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC and NCEP global ensemble prediction systems. *Mon. Weather Rev.*, **133**, 1076–1097.

Buizza, R., D. Richardson, and T. N. Palmer, 2003: Benefits of increased resolution in the ECMWF ensemble system and comparison with poor-mans ensembles. *Q. J. R. Meteorol. Soc.*, **129**, 1269–1288.

Cui, B., Z. Toth, Y. Zhu, D. Hou, D. Unger, and S. Beauregard, 2004: The trade-off in bias correction between using the latest analysis/modeling system with a short, vs an older system with a long archive. THORPEX, Montreal, Canada, Symposium Proceedings, 281–284.

Doblas-Reyes, F. J., R. Hagedorn, and R. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting - II. Calibration and combination. *Tellus*, **57A**, 234–252.

Evans, R., M. Harrison, R. J. Graham, and K. Mylne, 2000: Joint medium-range ensembles from the Met. Office and ECMWF systems. *Mon. Weather Rev.*, **128**, 3104–3126.

Hagedorn, R., F. Doblas-Reyes, and R. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. *Tellus*, **57A**, 219–233.

Hagedorn, R., T. Hamill, and J. Whitaker: 2008, Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: 2-meter temperatures, submitted.

Hamill, T., J. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Weather Rev.*, **132**, 1434–1447.

Hamill, T. M. and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology. *Q. J. R. Meteorol. Soc.*, **132**, 2905–2923.

Harrison, M. J., T. N. Palmer, D. S. Richardson, and R. Buizza, 1999: Analysis and model dependencies in medium-range ensembles: Two transplant case studies. *Q. J. R. Meteorol. Soc.*, **125**, 2487–2515.

Homleid, M., 1995: Diurnal corrections of short-term surface temperature forecasts using the Kalman filter. *Weather and Forecasting*, **10**, 689–707.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Weather Rev.*, **124**, 1225–1242.

Johnson, C., 2006: On the benefits of multi-model ensembles: idealized studies using the Lorenz 1963 model. Technical Report FRTR 492, Met Office.

Kharin, V. and F. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate*, **15**, 793–799.

Krishnamurti, T., C. Kishtawal, T. LaRow, D. R. Bachiiiochi, C. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **13**, 4196–4216.

Matsueda, M., M. Kyouda, H. L. Tanaka, and T. Tsuyuki, 2007: Daily forecast skill of multi-center grand ensemble. *SOLA*, **3**, 29–32.

Mylne, K., R. Evans, and R. Clark, 2002: Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. *Q. J. R. Meteorol. Soc.*, **128**, 361–384.

Palmer, T., A. Alessandri, U. Andersen, P. Cantelaube, M. Davey, P. Delecluse, M. Deque, E. Diez, F. J. Doblas-Reyes, H. F. nd R. Graham, S. Gualdi, J.-F. Gueremy, R. Hagedorn, M. Hoshen, N. Keenlyside, M. Latif, A. Lazar, E. Maisonnave, V. Marletto, A. P. Morse, B. Orfila, P. Rogel, J.-M. Terres, and M. C. Thomson, 2004: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bull. Amer. Meteor. Soc.*, 853–872.

Papoulis, A. and S. Pillai, 2002: *Probability, random variables and stochastic processes*. McGraw Hill, fourth edition, 852 pp.

Park, Y.-Y., R. Buizza, and M. Leutbecher, 2008: TIGGE: preliminary results on comparing and combining ensembles. Technical report, ECMWF, technical Memo. 548.

Pitman, A. J., A. Henderson-Sellers, C. E. Desborough, Z.-L. Yang, F. Abramopoulos, A. Boone, R. E. Dickinson, N. G. an dR. Koster, E. Kowalczyk, D. Lettenmaier, X. Liang, J.-F. Mahfouf, J. Noilhan, J. Polcher, W. Qu, A. Robock, C. Rosenzweig, C. A. Schlosser, A. B. Shmakin, J. Smith, M. Suarez, D. Verseghy, P. Wetzel, E. Wood, and Y. Xue, 1999: Key results and implications from phase 1(c) of the Project for Intercomparison of Land-surface Parametrization Schemes. *Climate Dynamics*, **15**, 673–684.

Raftery, A., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.*, **133**, 1155–1174.

Smith, J. and W. Krajewski, 1991: Estimation of the mean field bias of radar rainfall estimates. *J. Appl. Meteor.*, **30**, 397–412.

Stefanova, L. and T. Krishnamurti, 2002: Interpretation of seasonal climate forecast using Brier skill score, the Florida state university superensembles and the AMIP I dataset. *J. Climate*, **15**, 537–544.

Stephenson, D., C. Coelho, F. J. Doblas-Reyes, and M. Balmaseda, 2005: Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. *Tellus*, **57A**, 253–264.

Tebaldi, C. and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Phil. Trans. Royal Society, A*, **365**, 2053–2075.

Titley, H., N. Savage, and R. Swinbank, 2008: Comparison between Met Office and ECMWF medium-range ensemble forecast systems. Technical Report FRTR 512, Met Office.

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Q. J. R. Meteorol. Soc.*, **134**, 241–260.

Wilks, D., 2006: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorological Applications*, **13**, 243–256.

Wilson, L., S. Beauregard, A. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Mon. Weather Rev.*, **135**, 1364–1385.

Woodcock, F. and C. Engel, 2005: Operational consensus forecasts. *Weather and Forecasting*, **20**, 101–111.