Tellus (2005), 57A, 234–252
Printed in UK. All rights reserved

TELLUS

The rationale behind the success of multi-model ensembles in seasonal forecasting – II. Calibration and combination

By FRANCISCO J. DOBLAS-REYES*, RENATE HAGEDORN and T. N. PALMER, *ECMWF, Shinfield Park, Reading RG2 9AX, UK*

(Manuscript received 6 April 2004; in final form 24 September 2004)

ABSTRACT

The DEMETER multi-model ensemble system is used to investigate the enhancement in seasonal predictability that can be achieved by calibrating single-model ensembles and combining them to issue multi-model predictions. The forecast quality of both deterministic and probabilistic predictions is assessed and compared to the skill of a simple multi-model ensemble where all the single models are equally weighted. Both calibration and combination are carried out using cross-validation. Single-model seasonal ensembles are calibrated using canonical correlation analysis for model adjustment and variance inflation for reliability enhancement. Results indicate that both model adjustment and inflation increase the skill of tropical predictions for single-model ensembles, provided that the training time series are long enough. Some improvements are also found for extratropical areas, although mostly due to an increase of reliability associated with the inflation. The beneficial impact of calibration is smaller for the simple multi-model than for the single-model ensembles due to the relatively high reliability of the former. The raw single-model predictions are also linearly combined using grid-point multiple linear regression to create an optimized multi-model system. Results indicate that the forecast quality of the simple multi-model ensemble is generally difficult to improve using multiple linear regression due to the lack of robustness of the regression coefficients. As in the case of the calibration, longer time series would be preferred to achieve a significant forecast quality improvement. Over the tropics, a multiple linear regression, that uses the principal components of the model anomalies for the target area as predictors indicates a substantial gain in skill even with the available sample size. The implications of these results in an operational context are discussed.

1. Introduction

Ensemble weather and climate forecasting (Epstein, 1969; Molteni et al., 1996; Palmer, 2002) has been principally developed so as to estimate forecast uncertainty due to unavoidable errors in the initial conditions. However, initial conditions are not the sole source of uncertainty in weather and climate forecasting. Additional contributions come from model uncertainty. Estimation of the part of forecast uncertainty linked to model error is still an open problem that suffers from a lack of analytical solutions (Palmer, 2001). A pragmatic method to deal with this problem is based on the multi-model approach (Tracton and Kalnay, 1993; Harrison et al., 1995; Vislocky and Fritsch 1995; Doblas-Reyes et al., 2000; Evans et al., 2000; Fritsch et al., 2000; Palmer et al. 2000, 2004). The multi-model approach consists of combining predictions issued with different forecast systems to produce a more skilful and reliable estimate of the forecast probability density function (PDF). In the case of dynamical forecasting, the use of several models as members of an ensemble is a way of taking into account model uncertainty in the simulation of the climate system. If, in addition, each model can produce an ensemble of simulations with slightly different initial conditions, the multi-model ensemble system will sample the two sources of forecast uncertainty mentioned above, i.e. model error and uncertainty linked to initial conditions.

As described in Barnston et al. (2003), Hagedorn et al. (2005) and Palmer et al. (2004), a simple way of issuing predictions with a multi-model system consists in pooling together all the ensemble members from all the single-model ensembles and assigning an equal weight to each model. This system, denoted hereafter as the 'simple multi-model ensemble', has shown a clear superiority in terms of forecast quality over every single-model ensemble (Hagedorn et al., 2005). A substantial component of this superiority is due to the higher reliability of the simple multi-model. Because single-model predictions are not perfectly reliable, there is the possibility of improving the skill of the single-model ensembles by calibrating the probabilities of the single-model

*Corresponding author.
e-mail: francisco.doblas-reyes@ecmwf.int

ensembles through a simple inflation of the model variance (von Storch, 1999). In addition, model anomalies tend to be systematically misplaced when contrasted with reference data sets. Thus, an additional calibration of the single-model ensembles may be achieved by adjusting the model simulations, taking into account the linear correlation between the model and observed patterns of variability (Feddersen et al., 1999). Both techniques are used here to create calibrated single-model ensembles for which the forecast quality can be compared to (1) the uncalibrated single-model ensembles, (2) the simple multi-model ensemble that is used as a benchmark and (3) a calibrated simple multi-model ensemble constructed from adjusted single models.

There is substantial evidence (e.g. Shukla et al., 2000) of the fact that different single models have very different performances, the ranking of model skills changing as a function of lead time, variable, region, etc. Given that the underlying hypothesis in the simple multi-model method is that each model is independent and equally skilful whereas this condition is far from being satisfied, the combination of the single models through optimal weighting may increase further the forecast quality of multi-model predictions. Model combination has already been applied in the development of better standards of reference (Murphy, 1992) and in the improvement of predictions by independent forecast combination (Thompson, 1977). The predictions combined can be either statistical or dynamical, generally the combination methods not making any distinction between the forecast system used. Examples of this can be found in Fraedrich and Leslie (1987) while a comprehensive overview of prediction merging and combination can be found in Clemen (1989). The combination of single-model predictions to create a multi-model forecast by fitting different weights has already been addressed in long-range forecasting (Fraedrich and Smith, 1989; Sarda et al., 1996; Krishnamurti et al. 1999, 2000; Pavan and Doblas-Reyes 2000; Derome et al., 2001; Kharin and Zwiers 2002; Peng et al., 2002; Coelho et al., 2004; Metzger et al., 2004). Although some of the examples found a slight increase of forecast skill, there is a great range of results, and some of the experiments proved that it is possible to degrade the quality of the predictions if the underlying assumptions of the statistical combination models are not satisfied. Most of these examples are focused in formulating ensemble-mean or deterministic predictions. However, the superiority of the multi-model concept is more clearly evidenced in a probabilistic framework (Doblas-Reyes et al., 2000; Palmer and Shukla 2000; Palmer et al., 2004). Therefore, the advantages of model combination may be more clearly illustrated when dealing with probabilistic predictions. Some efforts in this direction have been described in Rajagopalan et al. (2002) and Stefanova and Krishnamurti (2002), while the method described in Coelho et al. (2004) provides an estimate of the uncertainty around ensemble-mean predictions that can be used to estimated probabilities.

In order to make a comprehensive assessment of the possible benefits of dynamical single-model ensemble combination to construct multi-model predictions, a set of multiple linear regression models is tested and compared with the simple multi-model. The comparison is carried out on the basis of probabilistic forecast quality properties to take into account the inherent probabilistic nature of long-range predictions.

The paper is organized as follows. The DEMETER experiment and the data used are described in Section 2. The methods employed to calibrate and combine the forecasts, along with the strategy to assess the forecast quality, are introduced in Section 3. In Section 4 we present the results from the single-model calibration and the model combination, illustrating the advantages and disadvantages of these post-processing techniques. Section 5 contains a summary and discussion of the results, along with a description of their operational consequences.

2. Data and experiment description

2.1. Data

The DEMETER project[1] has been funded under the European Union (EU) Framework-V Environment Programme to assess the skill and potential economic value of multi-model ensemble seasonal forecasts. One of the principal aims of DEMETER has been to advance the concept of multi-model ensemble prediction by installing a number of state-of-the-art global coupled ocean–atmosphere models on a single supercomputer and producing a series of multi-model ensemble hindcasts with common archiving and common diagnostic software.

The DEMETER prediction system comprises the global coupled ocean–atmosphere models of the following institutions: the European Centre for Research and Advanced Training in Scientific Computation, France (CERFACS); Centre National de Recherches Météorologiques, France (CNRM); the European Centre for Medium-Range Weather Forecasts (ECMWF); Istituto Nazionale de Geofisica e Vulcanologia, Italy (INGV); Laboratoire d'Océanographie Dynamique et de Climatologie, France (LODYC); Max-Planck Institut für Meteorologie, Germany (MPI); the UK Meteorological Office (UKMO). To assess seasonal dependence in forecast skill, the DEMETER hindcasts have been started four times a year from 1 February, 1 May, 1 August and 1 November at 00 GMT. The atmospheric and land-surface initial conditions are taken from the ECMWF Reanalysis[2] (ERA-40) data set. The ocean initial conditions are obtained from ocean-only runs forced by ERA-40 fluxes, except in the case of MPI that used a coupled initialization method. Ocean observations have been assimilated only in the UKMO ocean-only run after 1987. Each hindcast has been integrated for six months and comprises an ensemble of nine members. Hindcasts have been produced over the period 1958–2001, although the common period to the seven models is 1980–2001. A significant

[1] A complete description of the project and its main results can be found on the DEMETER website, http://www.ecmwf.int/research/demeter.
[2] See http://www.ecmwf.int/research/era.

Table 1. Number of models and sample length of the sets of hindcasts used in the construction of different multi-model predictions

Set of hindcasts	Number of models	Length of the sample
LSSP	7	1980–2001
SSLP	3	1959–2001
SSSP	3	1980–2001

part of the DEMETER data set is freely available for research purposes through a public on-line data retrieval system installed at the ECMWF.[3] The data can be retrieved in both GRIB and NetCDF format.

A key result is that probability scores for the simple multi-model ensemble are, on average, better than those from any of the single-model ensembles, this improvement not being only a consequence of the increase in ensemble size. Results indicate that the multi-model ensemble is a viable pragmatic approach to the problem of representing model uncertainty in seasonal-to-interannual prediction, leading to a more reliable and skilful forecasting system than that based on any one single model.

2.2. *Experimental design*

A range of experiments has been carried out to test the performance of different methods of single-model and multi-model calibration as well as single-model combination. To assess the benefits of the different methods, the calibration, combination and score calculations are carried out over three regions before the corresponding average skill score is computed. These regions are the tropical band (20°S–20°N), the Pacific–North America (PNA; 30°N–75°N, 160°W–40°W) and the North Atlantic–Europe (NAE; 35°N–75°N, 60°W–45°E) regions.

Three different sets of single-model ensembles are used for calibration and combination (Table 1). A first set of predictions for seven models and the period 1980–2001 will be referred to as 'large set–short period' (LSSP). Another set of predictions from the three models available for the period 1959–2001 is denoted as 'small set–long period' (SSLP). For comparison purposes, a 'small set–short period' (SSSP) has also been assessed, using the same three models as in SSLP.

The performance of the DEMETER system has been evaluated from this comprehensive set of hindcasts (Palmer et al., 2004; Hagedorn et al., 2005). In addition, a complete set of diagnostics can be accessed on-line at http://www.ecmwf.int/research/demeter/verification. The predictions considered in this paper correspond to the seasonal averages for the simulation period going from months 2 to 4 and months 4 to 6. That is, the results are described for the following seasons: December to February (DJF), March to May (MAM), June to August (JJA) and September to November (SON) in the case of the two–four month predictions, and for February to April (FMA), May to July (MJJ), August to October (ASO) and November to January (NDJ) for the four–six month predictions. For brevity, only predictions for the hindcasts starting in May and November will be discussed here. Due to the model biases, the raw model output in long-range dynamical forecasting is not generally useful (Déqué, 1991). Thus, anomalies have to be computed as departures from the corresponding long-term climatology. Calculation of model and reference anomalies, and the adjustment of the coefficients for the inflation and combination have been carried out in one-year out cross-validation mode (Michaelsen, 1987; Wilks, 1995) to obtain unbiased estimators. This implies eliminating the target year from the computation and using the rest of the sample to perform the statistical analyses.

[3] See http://www.ecmwf.int/research/demeter/data.

3. Methodology

The methods employed to calibrate the predictions and to combine the single-model ensembles are described in this section. While the combination needs to be undertaken using several models, the calibration has been carried out separately for each single-model and for the multi-model ensembles. The dependence of the results on the data used to train the statistical models is a drawback of any post-processing method. The quality of the data sets that are employed as a reference has usually been considered as a limitation to estimate the actual forecast quality of the predictions (e.g. Mason et al., 1999). The same data sets are used either to calibrate predictions or to construct weighted multi-models, being an additional source of forecast error in post-processed seasonal predictions. Given that assessing the impact of such a source of uncertainty requires the use of independent data sets, which may not be available for some of the variables dealt with here, this problem will not be addressed in this paper.

3.1. *Calibration*

A sample of climate model output is not equivalent to a sample of observed climate. Calibration can be considered as a way of obtaining predictions with average statistical properties similar to those of a reference data set. This implies that dynamical predictions need to be calibrated against a reference data set. Calibration has been often addressed using ad hoc methods given the large variety of statistical models available. As an illustration, variance inflation (von Storch, 1999) and model adjustment (Feddersen et al., 1999; Feddersen, 2000) have been used in this paper as calibration strategies. Model adjustment has also been used to perform a downscaling of seasonal predictions (Feddersen, 2003; Feddersen and Andersen, 2005).

Single-model probabilistic predictions may be unreliable due to symmetric, conditional biases linked to either model error or weaknesses of the method used for generating the ensemble

(Atger, 2003), or both. Therefore, an inflation of the ensemble spread is required in order to obtain reliable probabilities (Hamill and Colucci, 1998). The inflation modifies the predictions to have the same interannual variance as the reference data set at every grid point. Note that the simple method applied here is equivalent to the additive randomization described in von Storch (1999). If z_i is an ensemble-mean prediction for any grid point in time i and z_{ij} is the difference of ensemble member j with the ensemble mean, the inflated estimate of the ensemble member j can be expressed as

$$w_{ij} = \alpha z_i + \beta z_{ij}. \tag{1}$$

The coefficients α and β are found under the constraints that (1) the standard deviation of the inflated prediction is the same as that of the reference s_r and (2) the correlation ρ between the inflated prediction and the reference does not change with the inflation. A solution to this problem is

$$\alpha = \mathrm{abs}(\rho)\,\frac{s_\mathrm{r}}{s_\mathrm{em}} \tag{2}$$

$$\beta = \sqrt{1-\rho^2}\,\frac{s_\mathrm{r}}{s_\mathrm{e}}, \tag{3}$$

where s_em is the standard deviation of the ensemble mean and s_e is the square root of the mean variance of the ensemble. This method corrects the underestimation or overestimation of the ensemble spread with no modification of the ensemble-mean correlation. The experiment that performed the inflation of the different ensembles is denoted as INF hereafter (Table 2). When applied in cross-validation, the inflation is equivalent to the improvement of the prediction reliability with minor prejudice to the resolution or relative operating characteristic (ROC) area (see Appendix A for their definition) as described in Kharin and Zwiers (2003a,b). The method is applied separately for each grid point. However, Atger (2003) showed that the inflation with area-averaged parameters could provide better results because the parameter estimates are more robust. The inflation is always carried out in cross-validation mode.

To correct for systematic spatial shifts of the simulations, a canonical correlation analysis (CCA; von Storch and Zwiers 1999) between time series of ensemble-mean predictions and the reference data set is performed for each single-model ensemble. CCA searches for sequences of linearly related modes that can be used to form an expansion of the time series of the matrices **A** and **B**, the model and reference data sets, respectively. **A(B)** is an $N_\mathrm{T} \times N_\mathrm{p}$ matrix where each column is the ensemble-mean (reference) time series for a grid point. There are N_p grid points in the target region and N_T cases, where the number of cases is the amount of time-steps. The reference and model data sets are standardized to have zero mean and standard deviation equal to one. As the computations are carried out in cross-validation mode, the CCA is computed as many times as the number of years available. This implies that after removing the target year, the number of cases available in the training data set is $N_\mathrm{T} - 1$. A principal component analysis is performed with this sample in order to reduce the spatial dimension to a number smaller than the number of grid points. Only the components explaining at least 90% of the variance have been retained, which results in a different number of principal components depending on the target year, region and variable. The CCA decomposition of the original fields can be written as

$$\mathbf{A} = \mathbf{D}(\mathbf{C_A F})^\mathbf{T} \tag{4}$$

$$\mathbf{B} = \mathbf{E}(\mathbf{C_B G})^\mathbf{T}. \tag{5}$$

Matrices **D** and **E** contain the expansion coefficients (time series) for the simulated and reference data sets, respectively, while the columns of **F** and **G** contain the corresponding patterns and $\mathbf{C_A}$ and $\mathbf{C_B}$ are the covariance matrices of the fields **A** and **B**. **T** denotes transposed matrix. As CCA maximizes the linear correlation between pairs of patterns, the relative importance of the modes is given by the correlation between the expansion coefficients, which is stored in a diagonal matrix **H**.

The prediction field has been reconstructed separately for each target year using a certain number of patterns ranging from

Table 2. Set of calibration and combination experiments

Post-processing	Acronym	Description
None	SMM	Simple multi-model ensemble (with equal weighting) is constructed with original single-model ensembles.
Calibration	INF	Inflated single-model ensembles. A calibrated simple multi-model ensemble (with equal weighting) is obtained from the original single-model ensembles and then inflated.
Calibration	ADI	Inflated and CCA-based adjusted single-model ensembles. The calibrated simple multi-model (with equal weighting) is obtained from the CCA-based adjusted single models previous to inflation and then inflated.
Combination	REG	Multi-model with the coefficients computed at each grid point using multiple linear regression.
Combination	RPC	Multi-model with the coefficients computed at each grid point using principal component linear regression.
Combination	RAP	Multi-model with the coefficients computed at each grid point using multiple linear regression with the model principal components for the area as predictors.

$J = 1$ to 4. The underlying hypothesis is that the most predictable patterns in a linear sense will be captured by the leading CCA modes, although in theory this only holds for statistically stationary fields. In order to correct the spatial biases of the model variability, the reconstruction of the predictions should express the predictable time series within the phase space of the reference data. The adjustment of the predictions is carried out in cross-validation. Therefore, the estimated canonical correlation patterns for the training period are used to reconstruct an adjusted single-model ensemble for the target year τ, $\mathbf{a}_\tau^{\mathrm{rec}}$

$$\mathbf{a}_\tau^{\mathrm{rec}} = \mathbf{a}_\tau \mathbf{D}' \mathbf{H}' (\mathbf{C}_\mathbf{A} \mathbf{G})'^{\mathbf{T}}, \tag{6}$$

where $\mathbf{a}_\tau$ is the vector of ensemble members for time τ. The prime indicates that the matrices are truncated for the J leading canonical correlation patterns. The number of patterns included in the reconstruction is arbitrary, so that an additional adjustment has been performed using a selection of the most robust canonical patterns. The selection is based on the double cross-validation concept (Feddersen et al., 1999). A set of CCA analyses is made using the training data set and excluding one of the $N_\mathrm{T} - 1$ cases each time. The expansion coefficients are reconstructed from the CCA decompositions. The canonical patterns corresponding to the reconstructed expansion coefficients with a correlation above a certain threshold were then used to adjust the target year.

Given that CCA maximizes the correlation between pairs of patterns of the two fields, it is preferred to maximum covariance analysis because large variability differences are generally found between the models and the reference. The drawback of the method is that the most correlated patterns are also those that reduce the ensemble spread (Feddersen et al., 1999), so that an inflation of the ensemble predictions is required. The previously described inflation method has been applied and the results of the adjustment and posterior inflation will be denoted as experiment ADI (Table 2).

3.2. Combination

The linear combination of N_M predictions for a given grid point can be written as

$$\mathbf{y} = \mathbf{X}\mathbf{b} \tag{7}$$

where $\mathbf{b}$ is the vector of coefficients, $\mathbf{X}$ is an $N_\mathrm{T} \times N_\mathrm{M}$ matrix of predictors, N_T is the sample size or number of cases and $\mathbf{y}$ the vector of predictands. Both predictors and predictands are assumed to have zero mean. In this multiple linear regression problem, the least-squares solution for $\mathbf{b}$ is obtained from

$$\mathbf{C}_{xy} = \mathbf{C}_{xx}\mathbf{b}, \tag{8}$$

where $\mathbf{C}_{xy}$ and $\mathbf{C}_{xx}$ are the covariance matrices of the predictand with the predictors and of the predictors (von Storch and Zwiers, 1999), respectively. In our case, the predictand is the reference value and the predictors the ensemble-mean values of each single model for the specific grid point. This regression is also known as inverse regression (Coelho et al., 2004) because the observations are regressed against the predictions. The prediction for time τ is obtained by substituting the corresponding predictors for that time in the regression equation. An estimate of the prediction uncertainty (Anderson, 1972) assumes a Gaussian error distribution with the mean given by the predicted value $\mathbf{y}_\tau$ and the standard deviation by

$$s_\tau = s_0 \sqrt{\left(1 + \frac{1}{N_\mathrm{T}} \mathbf{x}_\tau \mathbf{C}_{xx}^{-1} \mathbf{x}_\tau\right)}. \tag{9}$$

$\mathbf{x}_\tau$ is the vector of predictor values for time τ and s_0 is the standard deviation of the regression residuals. The method applied to single grid points will be referred to in the following as experiment REG (Table 2).

As pointed out by Kharin and Zwiers (2002), the superiority of this linear combination is more easily evidenced when the sample size is large. This not being generally the case in seasonal forecasting, a way to avoid biased estimates of the coefficients is to compute them with an independent sample. However, the short length of the time series available only allows for the use of cross-validation as an alternative method. As in the rest of the paper, a leave-one-out scheme has been used to estimate the vector of coefficients. This implies that the parameter estimation is carried out separately for each target year using the $N_\mathrm{T} - 1$ remaining values.

Some authors (e.g. Thompson, 1977; Vislocky and Young, 1989) have pointed out that an optimal blend of predictions will perform better than any of the individual systems, provided that they have some skill and some independence. In the case of combination with multiple linear regression, the independence criterion is also important because co-linearity of the predictors (linear correlation between the ensemble mean of the different single models) introduces a large uncertainty in the estimated coefficients that is particularly evidenced in cross-validation mode. In this sense, Kharin and Zwiers (2002) suggest that the poorer performance of the combined multi-model predictions through multiple linear regression is due to overfitting or, in other words, biased estimates of the coefficients. Because co-linearity of the predictors may explain part of the failure, principal component regression (von Storch and Zwiers, 1999) offers an alternative way of performing the regression with linearly uncorrelated variables. The matrix of predictors for a grid point can be decomposed using a principal component analysis

$$\mathbf{X} = \mathbf{PQR}, \tag{10}$$

where $\mathbf{P}$ and $\mathbf{R}$ are the matrices whose columns contain the principal components and the loadings, respectively. $\mathbf{Q}$ is the matrix of eigenvalues. This method intends to build up a reduced set of uncorrelated variables that explain most of the variability of the N_M predictions. The principal components have been computed using both the covariance and the correlation matrix of the single-model ensemble-mean predictions. An interesting feature of the principal component regression is that a number $K < N_\mathrm{M}$

of principal components can be retained to construct a matrix $\mathbf{P}'$. The criterion used was based on retaining a certain amount of the variance between the models. A multiple linear regression between $\mathbf{P}'$ and $\mathbf{y}$ is then carried out to estimate a new set of regression coefficients. Finally, the predicted value for time τ is obtained from the regression equation with the vector of predictors given by

$$\mathbf{p}_\tau = \mathbf{x}_\tau \mathbf{R}'^{\mathbf{T}} \mathbf{Q}'^{-\mathbf{1}}, \tag{11}$$

where the prime indicates that the matrices are truncated to the K leading components. It is important to note that, as a consequence of the cross-validation process, the reconstructed predictors are not uncorrelated. However, some tests have indicated that the correlation between the reconstructed PCs is not statistically significant.

One of the main disadvantages of the methods described above is that the combination is carried out separately for each grid point. This implies that the spatial correlation of the fields is not taken into account. Using a reduced set of predictors that represents most of the variability of the region is a way of introducing the spatial information in the combination. A natural and simple set of variables to carry out that task consists in the leading principal components over a region. $\mathbf{O}$ is an $N_\mathrm{T} \times N_\mathrm{R}$ matrix, with the columns being the grid-point ensemble-mean time series for the area and for every single model. This matrix is used to compute $\mathbf{U}$, which contains the principal components. The predicted value and its uncertainty are computed using eqs. (7) and (9), with the vector of predictors for time τ given by

$$\mathbf{u}_\tau = \mathbf{o}_\tau \mathbf{V}'^{\mathbf{T}} \mathbf{T}'^{-\mathbf{1}}, \tag{12}$$

where $\mathbf{o}_\tau$ is the vector of ensemble-mean predictions and $\mathbf{V}$ and $\mathbf{T}$ are the loading and eigenvalue matrices, respectively. The prime indicates that the matrices are restricted to the L leading components. The regression may be performed either on the grid-point values of the reference data set or on the leading principal components of the reference data set over the region. In the second case, the predicted principal components and their uncertainty are used to obtain a prediction for the full field. As before, cross-validation is used for both the calculation of the principal components and the regression. As discussed above, the cross-validation applied to the principal component computation prevents the reconstructed predictors from being uncorrelated. This experiment is denoted hereafter as RAP (Table 2). Yun et al. (2005) describe results of a similar method to produce multi-model seasonal predictions.

The probabilistic predictions obtained from the regression are not reliable, so that an inflation of the predictions may still improve their reliability. The inflation applied is similar to that discussed previously. The coefficients α and β are computed as in eqs. (2) and (3) with s_e being the square root of the average of the variances estimated with eq. (9). The ensemble-mean prediction is multiplied by α, while the prediction uncertainty is scaled with β.

3.3. Forecast quality assessment

Forecast quality assessment is an essential component of the forecast formulation process (Jolliffe and Stephenson, 2003). The aim is to summarize and assess the overall forecast quality using a statistical description of how well the predictions match the verification data. Both deterministic (ensemble mean) and probabilistic predictions have been considered. The skill of the former has been estimated using the anomaly correlation coefficient (*ACC*). Three sources of uncertainty in common scoring metrics of probabilistic forecasts should be considered: improper estimates of probabilities from small-sized ensembles, insufficient number of forecast cases, and imperfect reference values due to observation errors. A way to alleviate this problem is to use several scoring measures to offer a comprehensive picture of the forecast quality of the system. Forecast quality is a multidimensional concept described by several scalar attributes that provides useful information about the performance of a forecasting system. Therefore, no single measure is sufficient for judging and comparing forecast quality. Forecast quality has been evaluated in this paper using measures of bias, reliability and accuracy. The set of verification measures are the frequency bias, the ROC area under the curve, and the Brier score (*BS*), including its decomposition in resolution and reliability terms. Although all these measures are thoroughly described in Thornes and Stephenson (2001), Jolliffe and Stephenson (2003), Doblas-Reyes et al. (2003) and Mason (2004), a brief description can be found in Appendix A.

Three types of events have been considered for verification: predictions above the upper and below the lower tercile, and predictions above the median. The thresholds (median or terciles) have been defined separately for the hindcast and the verification data sets. Tercile boundaries were computed using two different methods. A simple way of estimating the quantiles of a sample is to rank the values and find the boundaries that create equally populated bins. A more sophisticated method that allows for a high-resolution estimate has been used instead. It relies upon a Gaussian-kernel estimator of the population PDF (Silverman 1986). As mentioned above, there is an inherent uncertainty in these estimates that translates into an increased uncertainty in the forecast quality measures.

4. RESULTS

4.1. Calibration

As shown in Hagedorn et al. (2005), the simple multi-model (SMM) ensemble performs better than any single-model ensemble for most of the regions, parameters and lead times. An example of forecast quality measures for late winter (November start date, 4–6 month hindcast, FMA) predictions of mean sea level pressure above the median over the tropical band is shown in Fig. 1. The open bars represent the scores of the single-model and

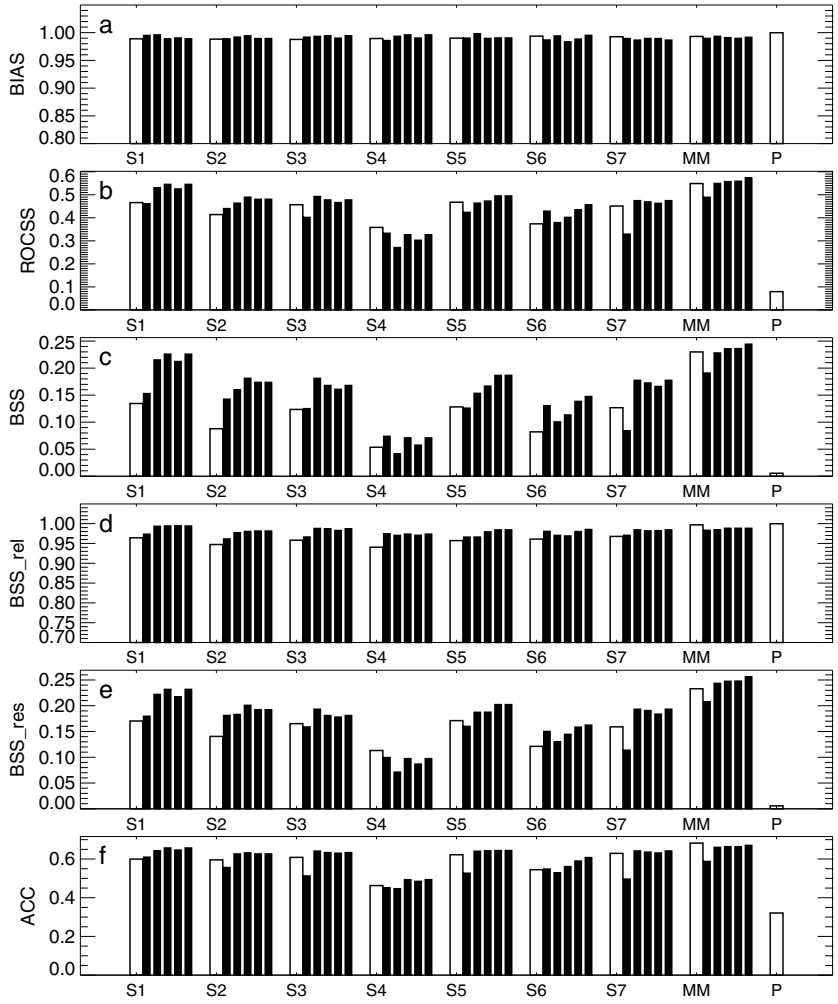

Fig 1. (a) Frequency bias, (b) *ROCSS*, (c) *BSS*, (d) BSS_{rel}, (e) BSS_{res} and (f) *ACC* for seven single-model ensembles (S1–S7), the simple multi-model (MM) ensemble and persistence (P) predictions of late winter (November start date, 4–6 month hindcast, FMA) mean sea level pressure above the median over the tropical band. The open bars represent the raw model output. The solid bars correspond to the results obtained with the adjusted and inflated single-model ensembles when retaining 1–4 canonical patterns (from left to right). The fifth solid bar corresponds to the adjusted single model constructed with only the most robust canonical patterns and then inflated. The calibrated simple multi-model ensemble is obtained from the ADI method of Table 2. The hindcasts are for the period 1980–2001.

SMM ensembles. Seven single models have been used to construct the SMM ensemble, with hindcasts for the period 1980–2001. For comparison, a prediction based on persistence is also displayed (open bar on the right end of each plot). This probabilistic prediction has been obtained from the seasonal anomaly of the three months previous to the start date of the hindcasts. The probability of persistence is then linearly relaxed towards a climatological probability forecast with a time-scale of six months.

As expected, the SMM ensemble performs better than any single-model ensemble. The SMM ensemble shows a high relative operating characteristic skill score (*ROCSS*) that denotes a greater ability to distinguish between events and non-events. A strong agreement can be found between the results of the ROCSS and the resolution component of the Brier skill score (BSS_{res}), due mainly to the fact that both are measures of potential predictability. In that sense, their results are comparable to the *ACC*, although the latter is defined for the ensemble-mean prediction while the former scores are probabilistic and depend on the event that is verified. The SMM also displays a higher reliability than any single-model ensemble. These results indicate that, as in Hagedorn et al. (2005), the improvement in the BSS of the SMM ensemble with respect to single-model ensembles is due to an increase in both reliability and resolution. Additionally, it can be observed in Fig. 1 that every single model and the SMM ensembles are better than persistence except in terms of reliability, which is a result common to other areas, variables and lead times.

The solid bars in Fig. 1 correspond to the results obtained with the adjusted and inflated (ADI) ensembles when retaining one to four leading canonical patterns (from left to right). The fifth bar corresponds to the forecast quality measures for the adjusted single-model ensembles obtained with only the most robust canonical patterns and then inflated. Double cross-validation as in Feddersen et al. (1999) has been used in such a way that, when no robust canonical pattern could be found, a climatological prediction was taken. The correlation threshold used to retain a canonical pattern is 0.15. The ADI simple multi-model ensemble was constructed by pooling together the ensemble members from the adjusted single models previous to

inflation. The adjusted simple multi-model ensemble was then inflated. The frequency bias of the ADI ensembles as well as that of the original output is overall close to 1. This is a desirable feature in order to avoid any hedging of the predictions to optimize specific scores. The reliability is improved for the ADI single-model ensembles with low BSS_{rel}, which is mainly due to the inflation. A similar increase in reliability is found for the INF ensembles but with a slight decrease of *ROCSS* and BSS_{res}, as in Kharin and Zwiers (2003a,b). The gain in reliability with respect to the original single-model ensembles is similar to that obtained with the SMM ensemble. This result is also valid for most of the regions and variables and illustrates one of the advantages of the simple multi-model.

The ADI single-model ensembles show an overall increase of *ROCSS* and BSS_{res}, especially with a large number of patterns retained. Better results, on average, can be found for the best set of canonical patterns, as in Feddersen et al. (1999). While the increase in reliability can be considered a consequence of the inflation, the improvement in terms of resolution is due to the CCA-based adjustment. The *ACC* results are close to those of *ROCSS* and BSS_{res}, although the improvement over the original single-model skill scores is not always found for the three skill scores. As a consequence of the gain in reliability and resolution for the ADI single-model ensembles, the *BSS* also improves. An interesting feature of the model adjustment is that in some specific cases, as for model 1 in Fig. 1, the calibrated scores are almost as good as those of the SMM ensemble. However, when this occurs, the specific single model depends on the event, region, variable, etc.

In spite of the positive impact of the single-model calibration, the ADI simple multi-model displays a very small increase of skill. None the less, a positive impact of the calibration for the simple multi-model can be found when longer time series are available. Figure 2 depicts the *ROCSS*, *BSS* and BSS_{rel} for the three single-model ensembles and their corresponding simple multi-model for the period 1959–2001. These three models are the same as the first three in Fig. 1. Skill scores for the unadjusted ensembles are lower than in the previous figure, although persistence is still beaten in terms of *ROCSS*. The ADI single-model ensembles show an important increase in *ROCSS* (and also in BSS_{res}) with respect to the original ensembles. Along with an improvement of the reliability due to the inflation, the skill improvement of the calibration explains the change of sign from negative to positive *BSS*, which is then better than for persistence. In addition, these improvements are at the origin of a general increase of the multi-model ensemble skill. Both reliability and resolution increase and the *BSS* increases up to three times the SMM value.

Part of the decrease in skill observed when using longer time series (compare open bars in Figs. 1 and 2) could be linked to the relative inability of the single models to correctly reproduce the observed multi-annual variability, as suggested by Pavan et al. (2005). It has been found that single models reproduce such variability in the first month of the simulation. However, the hindcasts for longer lead times display a clear disagreement with the low-frequency variability of the observations. This issue is currently under investigation and additional experimentation is being carried out.

Predictability is higher over the tropical regions than in the extratropics. Therefore, an improvement of predictability for the latter region may be more relevant than for the former. Figure 3 displays the *ROCSS*, *BSS* and BSS_{rel} for the same models of Fig. 1, but over the NAE region. Predictions are for winter (November start date, 2–4 month hindcast, DJF) 2-m temperature of the upper tercile. This is another example of the superiority of the SMM ensemble over the original single-model ensembles. As before, the ADI skill scores tend to be higher for the adjustment that uses the set of robust canonical patterns. There is a general improvement of the ADI single-model reliability, the solid bars improving every time with respect to the value

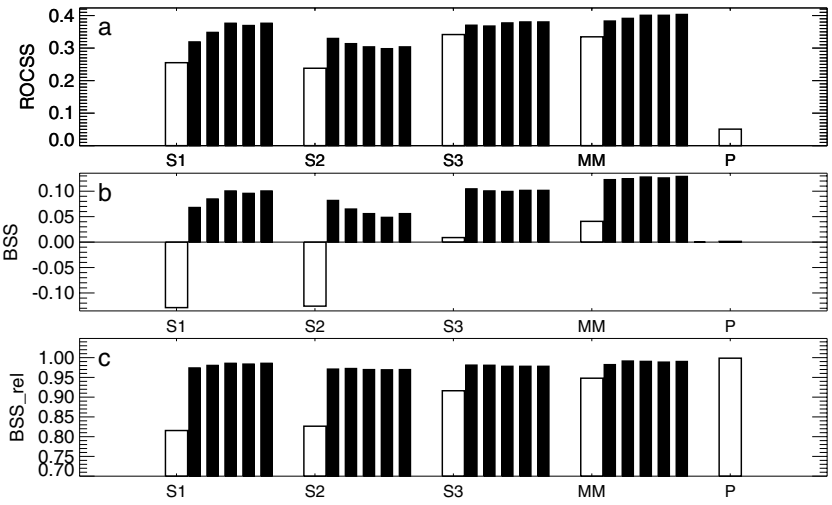

Fig 2. (a) *ROCSS*, (b) *BSS* and (c) BSS_{rel} for three single-model ensembles (S1–S3), the corresponding simple multi-model (MM) and persistence (P) predictions of late winter (November start date, 4–6 month hindcast, FMA) mean sea level pressure above the median over the tropical band. The reading of open and solid bars is as in Fig. 1. The hindcasts are for the period 1959–2001.

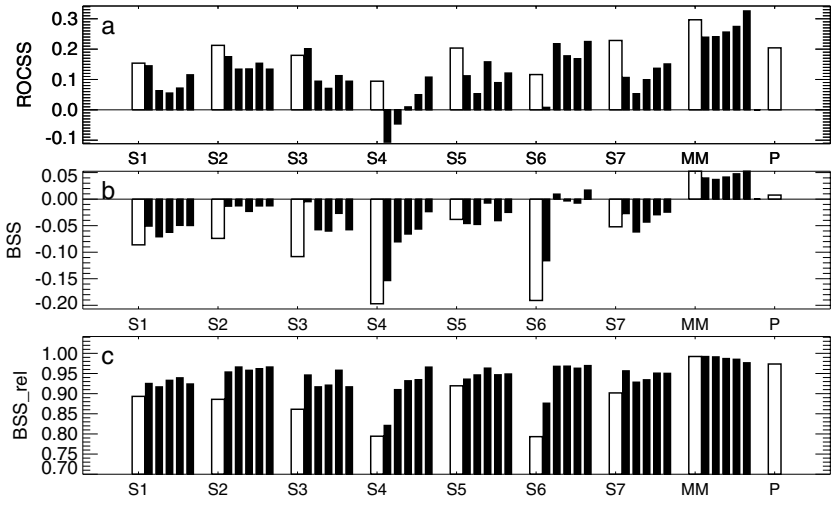

Fig 3. As in Fig. 2, but for seven single-model ensembles and predictions of winter (November start date, 2–4 month hindcast, DJF) 2-m temperature in the upper tercile over the NAE region. The hindcasts are for the period 1980–2001.

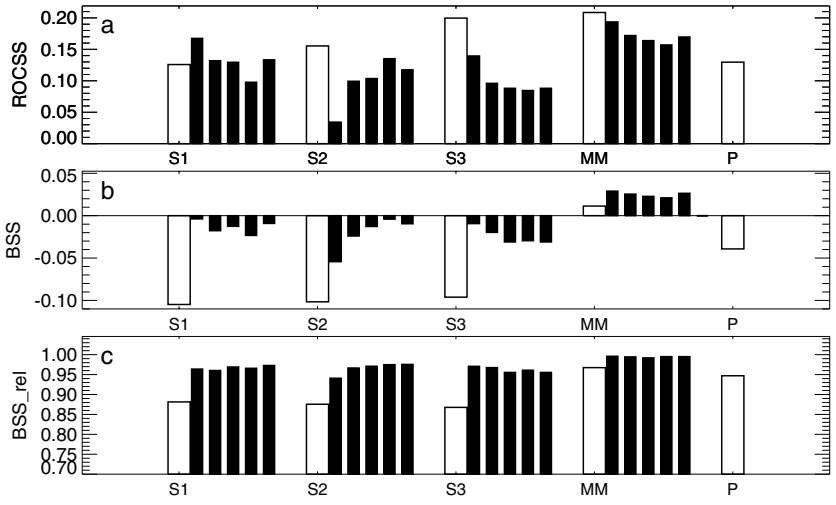

Fig 4. As in Fig. 3, but for three single-model ensembles and hindcasts for the period 1959–2001.

of the unadjusted ensemble. This increases the *BSS*, although a general decrease of the resolution, which has a behaviour similar to the *ROCSS*, prevents a higher rise. The skill scores of the ADI simple multi-model ensemble do not generally increase with the calibration, the reliability skill score of the simple multi-model being already close to 1, and therefore difficult to improve. Similar results have been found for the other two events verified over the same region.

Given that a large improvement in scores is found when no cross-validation is used (not shown), one of the reasons for the poor performance of the adjustment method may be the small sample size that gives less robust estimates. Figure 4 shows the *ROCSS*, *BSS* and BSS_{rel} for the set of single-model ensembles available for the period 1959–2001. The same event, variable and lead time as in Fig. 3 are depicted. As in the case of the tropical region, the skill scores of the original single-model ensembles are slightly lower with this longer set of hindcasts. A comparison with the persistence scores indicates that the single-model ensembles have low reliability. The SMM skill scores are also lower than in Fig. 3, part of this decrease being due to the smaller number of models considered. As in Fig. 3, the *ROCSS* is not improved for the ADI single-model (except for model 1) and multi-model ensembles. However, the ADI single-model ensemble *BSS* is greatly improved, mainly due to an increase in reliability, which allows them to have a less negative skill score. The ADI simple multi-model marginally improves the *BSS* with respect to that of the SMM ensemble.

The example described in Fig. 4 can be used to illustrate how the CCA-based adjustment can increase the skill of the single-model ensembles. Figure 5a depicts the grid-point correlation between the ensemble-mean prediction of model 1 and ERA-40. Areas with positive correlation appear over the North Atlantic, the European continent showing mainly zero values. Figure 5b shows the correlation of the ADI ensemble-mean using the first, second and fourth canonical patterns. The third pattern was found in the double cross-validation process to be not robust enough to be retained in the reconstruction. The correlation decreases over the western subtropical North Atlantic. However, there is a

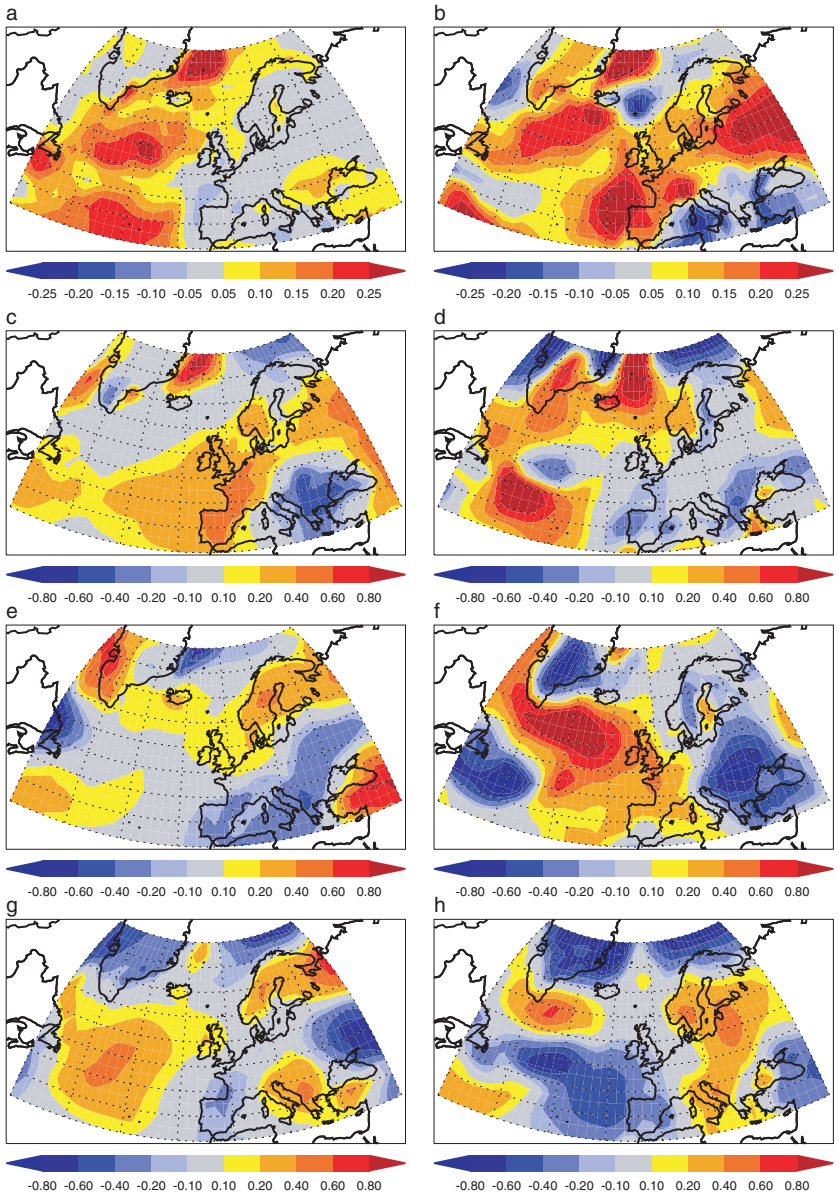

Fig 5. Grid-point ensemble-mean correlation with ERA-40 of (a) the winter (November start date, 2–4 month hindcast, DJF) 2-m temperature issued with model 1 and (b) the corresponding ADI prediction. Three canonical patterns (patterns 1, 2 and 4) have been retained for the reconstruction. The patterns are reproduced in (c), (e) and (g) for ERA-40 and in (d), (f) and (h) for the model. The hindcasts are for the period 1959–2001.

substantial gain over continental Europe (with two spots of negative correlation over the Mediterranean), the regional *ACC* between ensemble-mean predictions and ERA-40 increasing from 0.08 to 0.14. The shift of the adjusted model anomalies is the consequence of the three pairs of canonical patterns shown in Fig. 5. This result indicates that calibration for this area, season and variable can be beneficial as skill is improved in areas where the predictions might be most useful.

Precipitation is a special case in itself. Unfortunately, homogeneous global precipitation data sets are not available before 1979 (Adler et al., 2003). As a consequence, precipitation hindcasts have been calibrated only for the set of predictions for the period 1980–2001. ADI single-model ensembles show a much higher reliability for all areas and seasons (not shown). Over the tropics, the ADI single-model ensembles do not show improvements in *ROCSS* and BSS_{res}. This reduces the positive effect of the gain in reliability when the *BSS* is computed. As for the rest of the variables, the verification of probabilistic predictions offers a more complex picture of predictability than the *ACC*, which provides a single estimate regardless of the threshold the user is interested in. As extratropical skill is very low, the interpretation of the results in terms of improvements is difficult. As in previous cases, the most relevant feature is a gain in reliability that increases the *BSS*.

The calibration process reduces the skill score differences between the single-model and the multi-model ensembles. The previous examples indicate that single-model ensembles generally improve their reliability with the calibration and, sometimes, can

perform as well as the simple multi-model ensemble. However, the single-model ensemble that achieves such an improvement changes with the region, variable and lead time considered. As in Hagedorn et al. (2005), the main advantage of the simple multi-model is that, in spite of not being the best forecast system for every possible prediction, the likelihood of having a better score is higher than for any single model.

Figure 6 depicts the *BSS* of the SMM ensemble versus the *BSS* of the ADI simple multi-model for the LSSP and SSLP sets of hindcasts. Predictions for different regions, lead times, events and variables have been included, although the detail of the results is sacrificed to obtain an overall picture of the performance. The calibration of the single-model ensembles, based on CCA and posterior inflation for the 22-yr (1980–2001) long data sets, improves the reliability of the calibrated simple multi-model predictions in most cases, but only increases their resolution in some specific cases, most of them in extratropical areas. In those cases, the ADI simple multi-model performs better than the SMM ensemble, as indicated in Fig. 6a for the *BSS* (seven models are used), while barely any improvement is found for the tropical area. Different results are obtained with the longer time series available (43 yr of hindcasts, 1959–2001) for three of the models. This larger sample size allows more robust estimates of the canonical patterns and the posterior multi-model inflation. As a consequence, the *BSS* for the ADI simple multi-model is increased with respect to the SMM *BSS* (Fig. 6b) in a larger proportion of cases than for the shorter time series. This improvement is achieved as a result of a consistent increase in BSS_{rel}, BSS_{res} only increasing for the tropical region. It has been found that in both sets of time series the improvement is more important in the 4–6 month seasonal hindcasts than in the 2–4 month ones.

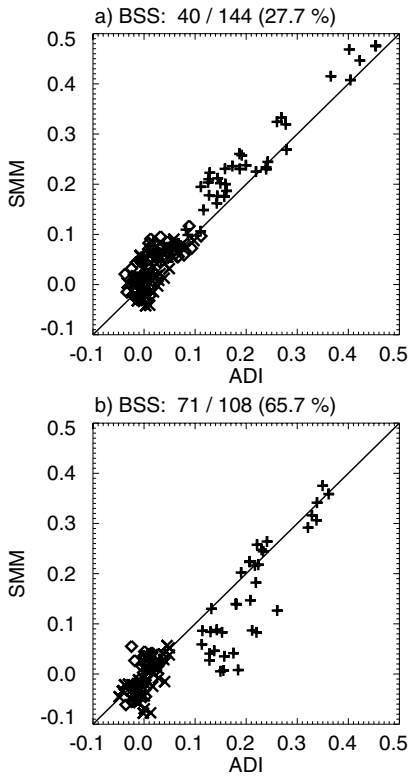

Fig 6. BSS for the SMM ensemble versus the ADI simple multi-model constructed from adjusted single models using the most robust canonical patterns for (a) the LSSP (seven models for the period 1980–2001) and (b) the SSLP (three models for 1959–2001) sets of hindcasts. Scores for three areas have been included: tropical band (plus symbols), PNA (diamonds) and NAE (crosses). Results for three events (anomalies within the upper and lower tercile and anomalies above the median), three variables (2-m temperature, 500-hPa geopotential height and sea level pressure), two lead times (one- and three-month lead) and two start dates (May and November). Precipitation results are only included in (a). The numbers on top of each panel correspond to the number of times the score of the system in the abscissa has a larger value than the one in the ordinates axis, the total number of cases and, in brackets, the percentage.

4.2. Combination

Previous results have shown that the ADI simple multi-model might be, in some cases, as skilful as (or even more skilful than) the simple multi-model obtained from the original single-model ensembles, depending on the sample size. In this section we intend to explore whether linear model combination can achieve a similar degree of improvement with regard to the SMM ensemble.

Multiple linear regression at each grid point has been employed by Krishnamurti et al. (1999, 2000) and Kharin and Zwiers (2002) as a method to formulate multi-model predictions, finding slightly contradictory results. Figure 7 shows the frequency bias, *ROCSS*, *BSS* and BSS_{rel} for the SMM ensemble versus the values for the REG multi-model. The LSSP set of hindcasts (seven models, 22 yr) has been used. As in Fig. 6, predictions for different regions, lead times, events and variables for the period 1980–2001 have been included. In this example the predictors have been standardized before computing the coefficients. This approach estimates the best linear unbiased estimator (BLUE) as in Sarda et al. (1996), Pavan and Doblas-Reyes (2000) and Derome et al. (2001). The combined predictions tend to be more skilful with this method than in a multiple linear regression where each single model keeps its own variability, although differences are minor (not shown). In order to compensate for the loss of variability in the regression process, the inflation method has been applied in cross-validation mode to the predictions. As in the INF experiment, the inflation increases the reliability of the predictions introducing a small decrease in *ACC*, *ROCSS* and BSS_{res} (not shown).

The SMM predictions have a frequency bias lower than 1 and worse than for the REG multi-model predictions, indicating that the SMM ensemble tends to underpredict the events. In contrast,

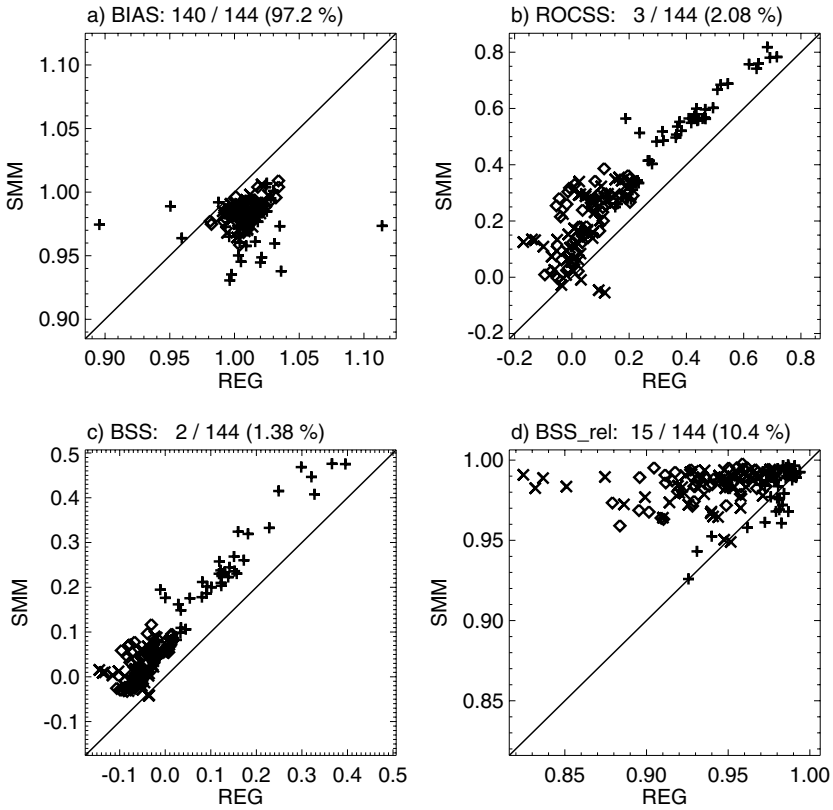

Fig 7. (a) Frequency bias, (b) *ROCSS*, (c) *BSS* and (d) BSS_{rel} for the SMM ensemble (ordinates) versus those for the REG multi-model predictions (abscissa). The BLUE method has been used for the multiple linear regression. The LSSP set of hindcasts (seven models, 22 yr, period 1980–2001) was used. Symbols are as in Fig. 6.

there is a decrease of BSS_{rel} for the REG predictions with respect to the SMM ensemble. The lack of robustness of the inflation parameters estimated in cross-validation mode may play a role in the decrease of reliability. *BSS*, *ROCSS* and *ACC* are worse for the REG multi-model, except for a few cases over the extratropics, where the skill is already very low. To illustrate the actual potential of this method to combine single-model predictions, Fig. 8 depicts the same skill scores as Fig. 7, but for the longer time series of the SSLP set of hindcasts (three models, 43 yr). The REG multi-model forecast quality values are closer to those of the SMM in this case, indicating that the length of the time series used to estimate the coefficients is an important factor in the construction of REG predictions. The REG multi-model *BSS* improves with respect to the SMM ensemble for almost all the tropical cases, mainly due to an increase in reliability linked to a higher-than-one frequency bias. *ROCSS* also increases for some tropical events and so does BSS_{res}. Nevertheless, an overall decrease of the forecast quality is found for the extratropical events.

The previous results suggest that there is a relationship between the improvement in the skill of the REG predictions and the sample size used for the estimation of the coefficients. The beneficial impact of the sample size can be evidenced more clearly by comparing Fig. 9, the *BSS* of the SSSP multi-model (three models, 22 yr), with the *BSS* panel of Fig. 8 (three models, 43 yr). The clear improvement for some areas appearing in the former are totally absent in the latter. The improvement of the REG multi-model *BSS* with longer time series is mainly due to a decrease of the difference in BSS_{res} of the REG multi-model with the SMM ensemble. Stefanova and Krishnamurti (2002) found that the resolution term is the most important contributor to the improvement of regression-based multi-model predictions with regards to a simple multi-model probabilistic hindcast system.

A comparison of Figs. 7c and 9 reveals a small improvement of the REG prediction constructed with three single models with respect to the prediction made using seven single models. This indicates that an excess of models may have a negative impact on the multiple linear regression if the samples are not long enough, as also found by Kharin and Zwiers (2002) and Krishnamurti et al. (2000). Part of the lack of robustness of the estimates is due to the co-linearity of the single-model predictions. The single models used tend to show similar skill over certain regions. Some of the coupled models may present a similar behaviour because they share the same atmospheric or ocean component. This lack of independence appears as co-linearity between the predictors of the regression and prevents the estimates of the model weights from being robust enough, especially when a large number of models are used. To reduce the impact of this burden, principal component regression has been used, where the principal component analysis has been computed in cross-validation separately for each grid point (RPC experiment). Figure 10 depicts the *ROCSS* for the RPC predictions versus that of the REG

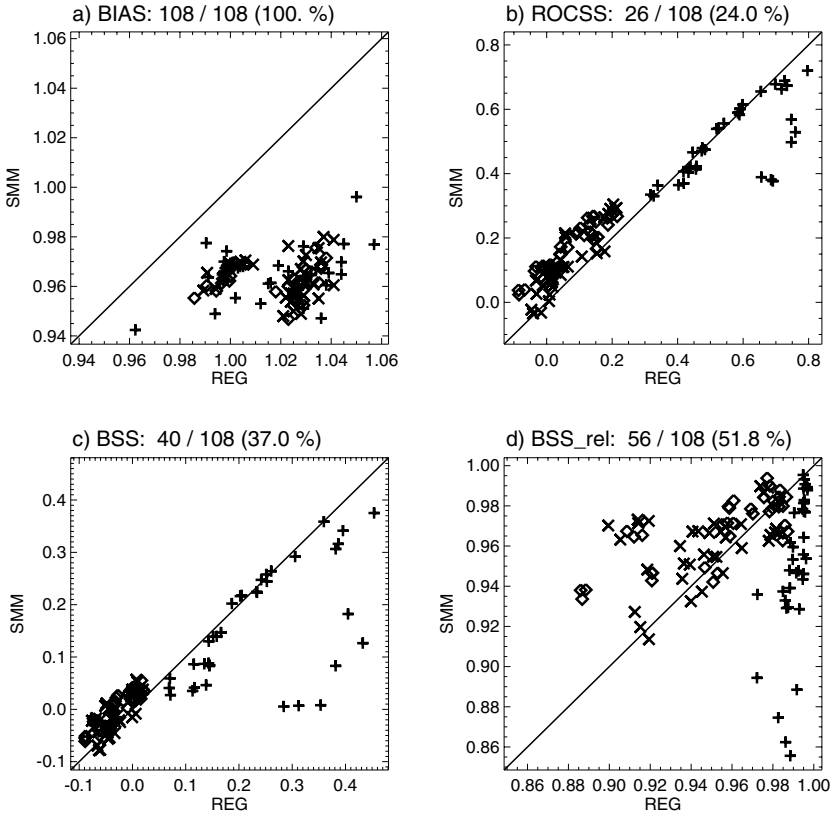

Fig 8. As Fig. 7, but for the SSLP set of hindcasts (three models, 43 yr, period 1959–2001). Precipitation results are not included.

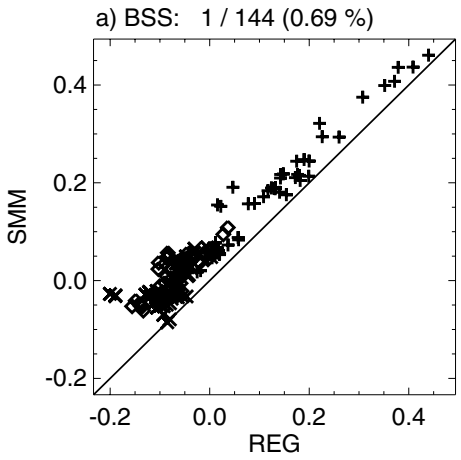

Fig 9. BSS for the SMM ensemble (ordinates) versus the REG multi-model predictions (abscissa). The BLUE method has been used, as well as the SSSP set of hindcasts (three models, 22 yr, period 1980–2001). Symbols are as in Fig. 6.

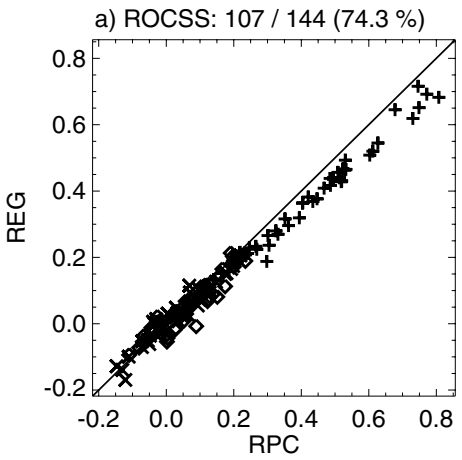

Fig 10. ROCSS for the REG (ordinates) versus the RPC predictions (abscissa). Symbols are as in Fig. 6. The LSSP set of hindcasts (seven models, 22 yr, period 1980–2001) and the BLUE method have been used in both cases. The principal components used as predictors in the RPC experiment explain at least 95% of the variance.

predictions, both computed with the LSSP set of hindcasts. The BLUE approach has been used. The input for the regression is the set of uncorrelated principal components that explain at least 95% of the covariance (three to five principal components depending on the variable, season and region). Although there is a gain in skill with respect to the REG experiment that is also found with other skill scores, it is not yet enough to beat the SMM ensemble (not shown). These results do not show any sensitivity to the computation of the principal components using the covariance or the correlation matrix. Tests where fewer principal components have been retained (variance explained in the

range of 50–75%) have shown that the introduction of a reduced number of predictors reduces the skill with regard to the standard multiple linear regression version. It has also been found that the skill of the SSLP and SSSP multi-model versions obtained through multiple linear regression in each grid point is not improved by using uncorrelated variables. In other words, the beneficial impact of principal component regression is only observed when a large number of models are used.

The multiple linear regression method in each grid point allows the estimation of an upper limit of predictability when the method is applied without cross-validation. This setting is not feasible in a realistic framework so that the results should not be interpreted as a measure of useful skill. In a realistic setting, only a cross-validated forecast quality assessment provides some information about useful skill. Instead, these results illustrate the importance of cross-validation. The *ROCSS* of the REG predictions with the LSSP set of hindcasts computed in-sample is shown in Fig. 11a. As expected, the skill is well above that of the simple multi-model and that of the cross-validated prediction, especially for the extratropics. The *ROCSS* is high, not only for the tropical events, but also for the extratropical ones. The same conclusion applies to the other scores. This result can be considered as the maximum skill that could be achieved with such a multi-model if the coefficients of the linear combination could be perfectly estimated, which would only occur with infinitely long time series and a stationary climate. Not surprisingly, this upper limit of the skill depends strongly on the number of models, as shown in Fig. 11b, where the *ROCSS* is lower for the SSSP multi-model than for the LSSP. An increase in the number of predictors without cross-validation necessarily provides a higher skill than a smaller multi-model. A certain improvement may also be observed in actual predictions, provided that the coefficients are robust estimates. However, given that climate is non-stationary, even if the time series are long enough, the potential skill should be reduced because the statistical model does not take this feature into account. This can be noticed in Fig. 11c, where the *ROCSS* for the SSLP set of hindcasts (three models, 43 yr) appears to be lower than that of the SSSP (three models, 22 yr).

The use of grid-point predictors does not take into account the spatial distribution of the anomalies. A way to include the spatial information is to use as predictors the leading principal components obtained from all the single models over the region of interest. Figure 12a shows the *BSS* for the SMM ensemble versus that of the RAP experiment for the SSLP set of hindcasts. Up to four model principal components have been used. The regression has been performed on the leading principal components of the reference data set that explain at least 75% of the variance. The predicted reference principal components have then been used to reconstruct the predicted spatial field and its uncertainty before applying the inflation. The *BSS* increases with respect to that of the SMM ensemble for the tropical region due to a simultaneous increase in both resolution and reliability. This increase is for some cases not as large as for the REG experiment (Fig. 8), although the results tend to be similar. Some improvement is also found for extratropical areas, especially due to an increase in reliability. In this case, the forecast quality is higher than that obtained in the REG experiment. Figure 12b depicts the corresponding *ACC*. In contrast with the *BSS*, the *ACC* is

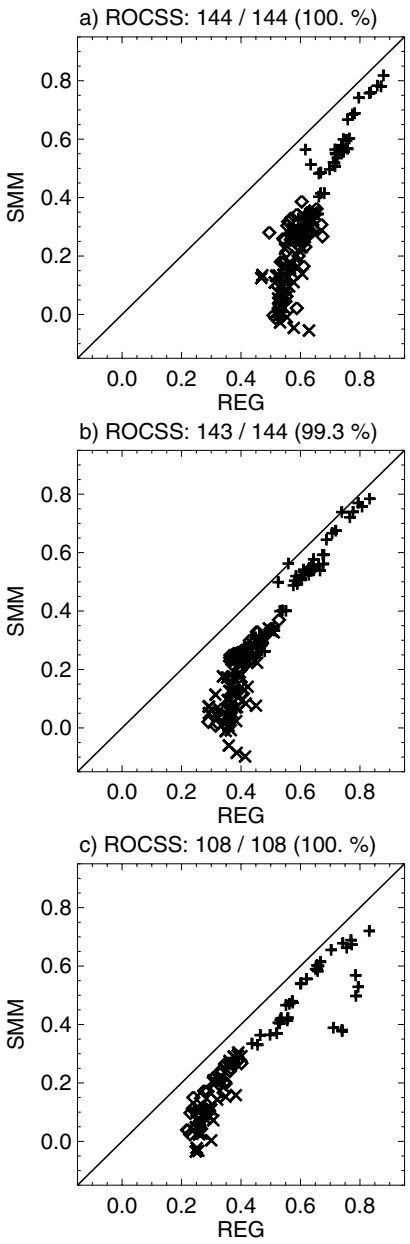

Fig 11. ROCSS for the SMM ensemble (ordinates) versus that of the REG multi-model predictions (abscissa, BLUE method) without cross-validation. The sets of hindcasts used are: (a) LSSP (seven models, 22 yr, period 1980–2001); (b) SSSP (three models, 22 yr, period 1980–2001); (c) SSLP (three models, 43 yr, period 1959–2001). Symbols are as in Fig. 6. Precipitation results are not included in (b).

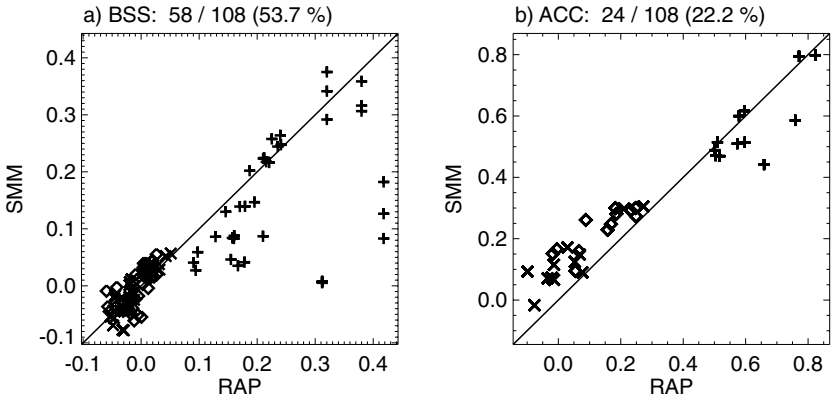

Fig 12. (a) *BSS* and (b) *ACC* for the SMM ensemble (ordinates) versus the corresponding skill scores of the RAP multi-model predictions (abscissa). Up to four model principal components have been used. The regression has been performed on the leading principal components of the reference data set that explain at least 75% of the variance. The predicted reference principal components have then been used to reconstruct the predicted spatial field and its uncertainty before applying the inflation. The SSLP set of hindcasts (three models, 43 yr, period 1959–2001) has been used. Symbols are as in Fig. 6.

clearly worse with respect to the SMM ensemble. This is another example of the benefits of considering forecast quality as a multidimensional magnitude. The reduction of the sample length has a strong impact in the combination. The combination of the SSSP set of hindcasts showed only a small improvement with regard to the SMM ensemble for the tropical areas and degradation elsewhere. Slightly worse results have been obtained when the regression is carried out separately for each reference grid point instead of for the reference leading principal components.

5. Discussion and concluding remarks

A comprehensive assessment of the impact of model calibration (using CCA and variance inflation) and combination (based on multiple linear regression) in the formulation of multi-model predictions has been described. The forecast quality of the systems has been discussed mainly for two different sets of hindcasts: (1) seven models and 22 yr; (2) three models and 43 yr. The results have been described taking into account previous published results of seasonal prediction calibration and combination.

The single-model seasonal predictions have been calibrated using CCA for model adjustment and variance inflation to increase the reliability. The gain in forecast quality has been compared with that of the original single-model ensembles and with the SMM ensemble. The results indicate that both model adjustment and inflation increase the skill of tropical predictions, provided that the sample size for training the statistical model is large enough. Some improvements are also found for extratropical areas, although it is mostly due to the increase in reliability. The larger improvements are found for longer-range predictions, a likely consequence of the drift of the coupled models beyond the initial month of the simulation. The skill increase is smaller for the calibrated simple multi-model than for the calibrated single-model ensembles.

The raw single-model predictions have also been linearly combined using grid-point multiple linear regression to create an optimized multi-model system. Results indicate that the overall forecast quality of probabilistic predictions is not significantly improved with respect to the SMM ensemble due to the lack of robustness of the coefficients. Much longer time series would be required to achieve a significant skill improvement as well as a reduction of the co-linearity between the single models. Vislocky and Fritsch (1995) showed that linear regression did not add any additional improvement over a simple multi-model in the case of temperature and precipitation short-term forecasts because of the correlation among the forecast systems. Similar conclusions were found by Kharin and Zwiers (2002) in the case of seasonal predictions. Principal component regression of the grid-point ensemble-mean hindcasts was used to solve the problem with very little success. The beneficial impact with regard to standard multiple linear regression is only observed when the number of single models is high.

A multiple linear regression experiment was carried out, for which the predictors are the leading principal components from a joint principal component analysis using the ensemble mean of every single model over a region. This method has the advantage of taking into account the spatial correlation of the anomalies and of estimating the coefficients from uncorrelated variables. It is shown that there is potential for improving the skill of the multi-model especially for the tropical band. Instead, the *ROCSS* and BSS_{res} of extratropical events do not increase when compared to the SMM ensemble. However, it is likely that a careful choice of the region used to compute the principal components may provide improved results for many areas and variables (Pavan et al., 2005).

Because of the difficulty in estimating the coefficients of the combination, an option to improve the SMM ensemble skill is to not use the model with the lowest skill for a given variable and region (the model selection is undertaken in cross-validation mode). A small increase of the skill scores is then found (not shown), especially when the number of single models used is small. In other words, the impact of an additional model is not large, even if it is the worst in the multi-model ensemble, provided that enough models are used. When the number of single models is small, the improvement of dropping a single-model ensemble off the multi-model is at best comparable to the

impact of the inflation. Still, all the models are required because the worst model in the multi-model ensemble is not the same for all regions and variables, in addition to the fact that a reliable choice of the worst model may be impossible most of the time.

A comparison of Figs. 6 and 12a suggests that both calibration and combination improve in a similar way the seasonal prediction forecast quality, the former being superior for extratropical predictions and the latter for tropical ones. This implies that calibration and combination complement each other. Stephenson et al. (2005) described a recently developed general framework to simultaneously achieve model calibration and combination in a Bayesian context. This framework will be used to explore those cases where post-processing seems to increase seasonal prediction skill.

Both deterministic and probabilistic predictions have been formulated and verified in this paper. The inherent probabilistic nature of seasonal forecasting implies that scores for probabilistic forecasts are more closely related to the kind of information most relevant to the users, and thus their results are more interesting and useful. Some examples have been found wherein either calibration or combination seemed to be beneficial in terms of ensemble-mean *ACC*. Nevertheless, the same conclusion did not hold in the case of the probabilistic predictions because either the reliability or the resolution (or the *ROCSS*) was degraded. The generally reduced information offered by the *ACC* should be taken as a warning when results of previous calibration and/or combination studies are considered. From an operational perspective, the benefits of post-processing methods need to be illustrated in terms of the gain in quality of either probabilistic forecasts as in Clark and Déqué (2003) or ensemble-mean forecasts with uncertainty estimates as in Coelho et al. (2003, 2004).

The forecast quality of the original single-model and SMM ensembles was found to be lower for the data set with longer time series. Additional analyses indicate that the single models have difficulties reproducing the multi-annual climate variability. This is harmful both for forecast quality assessment and for model post-processing (e.g. calibration and combination). The reasons for the reduction of forecast quality for the long sets of hindcasts have yet to be determined, although paucity of ocean observations before 1980 may be one of them. No ocean assimilation has been carried out in the hindcasts used in this paper (except for a short period in one of the models) because of the lack of ocean subsurface observations for the period before 1990. Similar obstacles are found for soil wetness initialization. This points towards the importance of properly initializing the coupled models, which should be a priority for future developments.

The results described in this paper have two relevant implications for an operational environment. First, the calibration of high-quality single-model ensembles and their combination require long hindcast data sets, which are difficult to create due to both computational constraints and uncertainties in the initialization of the soil and ocean components for early dates. Secondly, the wide range of results presented here implies that the areas and variables where calibration and/or combination are likely to introduce some skill increase should be carefully selected. Predictions for areas and variables where no or little improvement with respect to an SMM ensemble is found should be taken from the SMM ensemble. This implies that a thorough assessment of every multi-model forecast system is required. Both aspects, the availability of long data sets and the selection of the areas and variables where calibration and/or combination are beneficial, need to be considered in the context of the application to end-user models (e.g. Cantelaube and Terres 2005; Challinor et al., 2005; Marletto et al., 2005; Morse et al., 2005) and of the decision-making process of users (Sarewitz et al., 2000; Jochec et al., 2001; Hartmann et al., 2002; Murnane et al., 2002; Archer, 2003). This is the framework where the actual predictive potential of a multi-model seasonal forecast system can be determined.

6. Acknowledgments

This work was supported by the EU-funded DEMETER project (EVK2-1999-00024). The authors wish to thank David Anderson, Caio Coelho, Simon Mason, David Stephenson and Tim Stockdale for their constructive advice and support. Special thanks go to Magdalena Alonso Balmaseda for developing the inflation method and for her useful and stimulating remarks.

7. Appendix A

A set of verification measures has been used to assess the quality of the probabilistic hindcasts. These measures are the frequency bias, the ROC area under the curve and the Brier score (*BS*) along with the terms obtained from its decomposition.

The hit rate H, or the relative number of times an event was forecast when it occurred, and the false alarm rate F, or the relative number of times the event was forecast when it did not occur (Jolliffe and Stephenson, 2003), are two useful variables to characterize the performance of a prediction system. They are based on the likelihood-base rate factorization of the joint probability distribution of forecasts and verifications (Murphy and Winkler, 1987). The hit and false alarm rates take the form

$$H = \frac{a}{a+c}, \qquad F = \frac{b}{b+d}, \tag{A1}$$

where a, b, c and d are the elements of a contingency table containing the number of hits (a, number of cases when an event is forecast and is also observed), false alarms (b, number of cases the event is not observed but is forecast), misses (c, number of cases the event is observed but not forecast), and correct rejections (d, number of no-events correctly forecast) for all ensemble members.

The hit and false alarm rates allow the definition of a reliability measure, the frequency bias B (Thornes and Stephenson,

2001). The frequency bias is an attribute of forecast quality that corresponds to the ability of the forecast system to average probabilities equal to the frequency of the observed event. This measure indicates whether the forecasts of an event are being issued at a higher rate than the frequency of observed events and can be estimated from

$$B = \frac{a+b}{a+c}. \tag{A2}$$

A frequency bias greater than 1 indicates overforecasting, i.e. the model forecasts the event more often than it is observed. Consequently, a bias lower than 1 indicates underforecasting.

The ROC (Swets, 1973) is a signal-detection curve plotting the hit rate against the false alarm rate for a specific event over a range of probability decision thresholds (Mason and Graham, 1999; Kharin and Zwiers, 2003b) where a separate contingency table has been determined for every probability threshold. It indicates the performance in terms of hit and false alarm rates stratified by the verification. Ideally, the hit rate will always exceed the false alarm rate and the curve will lie in the upper left-hand portion of the diagram. The hit rate increases by reducing the probability threshold, but at the same time the false alarm rate is also increased. The standardized area enclosed beneath the curve, A, is a simple accuracy measure associated with the ROC, with a range from 0 to 1. A skill score can be constructed from this area as $ROCSS = 2 \times A - 1$ (Richardson, 2001). A system with no skill (made by either random or constant forecasts) will show a skill score of zero. As ROC is based upon the stratification by the verification, it provides no information about reliability of the forecasts, and hence the curves cannot be improved by improving the climatology of the system. In other words, this score is relatively independent of the forecast calibration and should be understood as a measure of potential predictability, being qualitatively similar to the resolution term of the *BS*.

The *BS* is an accuracy measure similar to the mean square error. It is defined as

$$BS = \sum_{i=1}^{N} (p_i - o_i)^2, \tag{A3}$$

where p_i represents the probability of predicting a dichotomous event and o_i is the corresponding observed probability, which usually is taken as 1 if the event occurs and 0 otherwise. The corresponding skill score is

$$BSS = 1 - \frac{BS}{p_c(1-p_c)}, \tag{A4}$$

where p_c is the climatological probability of the event. The range of *BSS* is the interval

$$\left[1 - \frac{1}{p_c(1-p_c)}, 1\right].$$

The *BS* can be decomposed into three non-negative terms:

$$BS = BS_{rel} - BS_{res} + BS_{unc}. \tag{A5}$$

The three terms are known, respectively, as: (1) reliability, which measures the conditional bias of the forecasts and is equal to the weighted average of the square differences between the forecasts and conditional observed probabilities; (2) resolution, which measures the ability of the forecast system to discriminate between event occurrences and non-occurrences; (3) uncertainty, which is the variance of the observation probabilities. The corresponding *BSS* can be written as

$$BSS = BSS_{res} + BSS_{rel}, \tag{A6}$$

where BSS_{rel} and BSS_{res} are the skill scores of the reliability and resolution components, respectively. Their values are found within the range [0, 1].

References

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P. and co-authors. 2003. Climatology project (GPCP) monthly precipitation analysis (1979—present). *J. Hydrometeor.* **4**, 1147–1167.

Anderson, T. 1972. *The Statistical Analysis of Time Series*. Wiley, New York.

Archer, E. R. 2003. Identifying underserved end-user groups in the provision of climate information. *Bull. Am. Meteorol. Soc.* **84**, 1525–1532.

Atger, F. 2003. Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.* **131**, 1509–1523.

Barnston, A., Mason, S., Goddard, L., Dewitt, D. and Zebiak, S. 2003. Multi-model ensembling in seasonal climate forecasting at IRI. *Bull. Am. Meteorol. Soc.* **84**, 1783–1796.

Cantelaube, P. and Terres, J.-M. 2005. Seasonal weather forecasts for crop yield modelling in Europe. *Tellus* **57A**, 476–487.

Challinor, A. J., Slingo, J. M., Wheeler, T. R. and Doblas-Reyes, F. J. 2005. Probabilistic simulations of crop yield over western India using the DEMETER seasonal hindcast ensembles. *Tellus* **57A**, 498–512.

Clark, R. and Déqué, M. 2003. Conditional probability of seasonal predictions of precipitation. *Q. J. R. Meteorol. Soc.* **129**, 179–193.

Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *Int. J. Forecasting* **5**, 559–583.

Coelho, C. A., Pezzulli, S., Balmaseda, M., Doblas-Reyes, F. J. and Stephenson, D. B. 2003. Skill of coupled model seasonal forecasts: A bayesian assessment of ECMWF ENSO forecasts. ECMWF Technical Memorandum No. 426.

Coelho, C. A., Pezzulli, S., Balmaseda, M., Doblas-Reyes, F. J. and Stephenson, D. B. 2004. Forecast calibration and combination: a simple bayesian approach for ENSO. *J. Climate* **17**, 1504–1516.

Déqué, M. 1991. Removing the model systematic error in extended range. *Ann. Geophys.* **9**, 242–251.

Derome, J., Brunet, G., Plante, A., Gagnon, N., Boer, G. J. and co-authors 2001. Seasonal predictions based on two dynamical models. *Atmosphere–Ocean* **39**, 485–501.

Doblas-Reyes, F. J., Déqué, M. and Piedelièvre, J.-P. 2000. Multi-model spread and probabilistic seasonal forecasts in PROVOST. *Q. J. R. Meteorol. Soc.* **126**, 2069–2088.

Doblas-Reyes, F. J., Pavan, V. and Stephenson, D. B. 2003. Multi-model seasonal hindcasts of the NAO. *Climate Dyn.* **21**, 501–514.

Epstein, E. 1969. Stochastic dynamic prediction. *Tellus* **21**, 739–759.

Evans, R., Harrison, M. S. J., Graham, R. J. and Mylne, K. R. 2000. Joint medium-range ensembles from the Met Office and ECMWF systems. *Mon. Wea. Rev.* **128**, 3104–3127.

Feddersen, H. 2000. Impact of global sea surface temperature on summer and winter temperatures in Europe in a set of seasonal ensemble simulations. *Q. J. R. Meteorol. Soc.* **126**, 2089–2110.

Feddersen, H. 2003. Predictability of seasonal precipitation in the Nordic region. *Tellus* **55A**, 385–400.

Feddersen, H. and Andersen, U. 2005. A method for statistical downscaling of seasonal ensemble predictions. *Tellus* **57A**, 398–408.

Feddersen, H., Navarra, A. and Ward, M. 1999. Reduction of model systematic error by statistical correction for dynamical seasonal prediction. *J. Climate* **12**, 1974–1989.

Fraedrich, K. and Leslie, L. M. 1987. Combining predictive schemes in short-term forecasting. *Mon. Wea. Rev.* **115**, 1640–1644.

Fraedrich, K. and Smith, N. 1989. Combining predictive schemes in long-range forecasting. *J. Climate* **2**, 291–294.

Fritsch, J. M., Hilliker, J., Ross, J. and Vislocky, R. L. 2000. Model consensus. *Wea. Forecasting* **15**, 571–582.

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

Hamill, T. and Colucci, S. J. 1998. Verification of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.* **126**, 711–724.

Harrison, M. S. J., Palmer, T. N., Richardson, D. S., Buizza, R. and Petroliagis, T. 1995. Joint ensembles from the UKMO and ECMWF models. In: *ECMWF Seminar Procedings: Predictability*,Vol. 2, pp. 61–120. ECMWF, Reading, UK.

Hartmann, H. C., Pagano, T. C., Sorooshian, S. and Bales, R. 2002. Confidence builders: evaluating seasonal climate forecasts for user perspectives. *Bull. Am. Meteorol. Soc.* **83**, 683–698.

Jochec, K., Mjelde, J., Lee, A. and Conner, J. 2001. Use of seasonal climate forecasts in rangeland-based livestock operations in west Texas. *J. Appl. Meteorol.* **40**, 1629–1639.

Jolliffe, I. T. and Stephenson, D. B. 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, New York.

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

Kharin, V. V. and Zwiers, F. W. 2003a. Improved seasonal probability forecasts. *J. Climate* **16**, 1684–1701.

Kharin, V. V. and Zwiers, F. W. 2003b. On the roc score of probability forecasts. *J. Climate* **16**, 4145–4150.

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z. and co-authors. 1999. Improved weather and seasonal climate forecasts from multi-model superensemble. *Science* **285**, 1548–1550.

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., LaRow, T. E., Bachiochi, D. R. and co-authors. 2000. Multi-model ensemble forecasts for weather and seasonal climate. *J. Climate* **13**, 4196–4216.

Marletto, V., Zinoni, F., Criscuolo, L., Fontana, G., Marchesi, S. and co-authors. 2005. Evaluation of downscaled DEMETER multi-model ensemble seasonal hindcasts in northen Italy by means of a model of wheat growth and soil water balance. *Tellus* **57A**, this issue.

Mason, S. J. 2004. On using climatology as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.* **132**, 1891–1895.

Mason, S. J. and Graham, N. E. 1999. Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting* **14**, 713–725.

Mason, S. J., Goddard, L., Graham, N. E., Yulaeva, E., Sun, L. and co-authors. 1999. The IRI seasonal climate prediction system and the 1997/98 El Niño event. *Bull. Am. Meteorol. Soc.* **80**, 1853–1873.

Metzger, S., Latif, M. and Fraedrich, K. 2004. Combining ENSO forecasts: a feasibility study. *Mon. Wea. Rev.* **132**, 456–472.

Michaelsen, J. 1987. Cross-validation in statistical climate forecast models. *J. Climate Appl. Meteor.* **26**, 1589–1600.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. 1996. The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119.

Morse, A. P., Doblas-Reyes, F., Hoshen, M. B., Hagedorn, R., Thomson, M. C. and co-authors. 2005. A forecast quality assessment of an end-to-end probalistic multi-model seasonal forecast system using a malaria model. *Tellus* **57A**, 454–475.

Murnane, R. J., Crowe, M., Eustis, A., Howard, S., Koepsell, J. and co-authors. 2002. The weather risk management industry's climate forecast and data needs: a workshop report. *Bull. Am. Meteorol. Soc.* **83**, 1193–1198.

Murphy, A. H. 1992. Climatology, persistence, and their linear combination as standards of reference in skill scores. *Wea. Forecasting* **7**, 692–698.

Murphy, A. H. and Winkler, R. 1987. A general framework for forecast verification. *Mon. Wea. Rev.* **115**, 1330–1338.

Palmer, T. N. 2001. A non-linear dynamical perspective on model error: a proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Q. J. R. Meteorol. Soc.* **127**, 279–304.

Palmer, T. N. 2002. Predictability of weather and climate: from theory to practice – from days to decades. In: *ECMWF Seminar Proceedings: Predictability of Weather and Climate*. ECMWF, Reading, UK. pp. 1–14.

Palmer, T. N. and Shukla, J. 2000. Editorial to DSP/PROVOST special issue. *Q. J. R. Meteorol. Soc.* **126**, 1989–1990.

Palmer, T. N., Brankovic, C. and Richardson, D. S. 2000. A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Q. J. R. Meteorol. Soc.* **126**, 2013–2034.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M. and co-authors. 2004. Development of a European multi-model ensemble system for seasonal to interannual prediction (DEMETER). *Bull. Am. Meteorol. Soc.* **85**, 853–872.

Pavan, V. and Doblas-Reyes, F. J. 2000. Multi-model seasonal hindcasts over the Euro-Atlantic: skill scores and dynamical features. *Climate Dyn.* **16**, 611–625.

Pavan, V., Marchesi, S., Morgillo, A., Cacciamani, C. and Doblas-Reyes, F. J. 2005. Downscaling of DEMETER winter seasonal hindcasts over northern Italy. *Tellus* **57A**, 424–434.

Peng, P., Kumar, A., van den Dool, H. and Barnston, A. G. 2002. An analysis of multi-model ensemble predictions for seasonal climate anomalies. *J. Geophys. Res.* **107**, doi:10.1029/2002JD002712.

Rajagopalan, B., Lall, U. and Zebiak, S. E. 2002. Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Mon. Wea. Rev.* **130**, 1792–1811.

Richardson, D. S. 2001. Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Q. J. R. Meteorol. Soc.* **127**, 2473–2489.

Sarda, J., Plaut, G., Pires, C. and Vautard, R. 1996. Statistical and dynamical long-range atmospheric forecasts: experimental comparison and hybridization. *Tellus* **48A**, 518–537.

Sarewitz, D., Pielke R. A. Jr, and Byerly, R. Jr. 2000. *Prediction: Science, Decision Making and the Future of Nature*. Island Press, Washington, DC.

Shukla, J., Anderson, J., Baumhefner, D., Brankovic, C., Chang, Y. and co-authors. 2000. Dynamical seasonal prediction. *Bull. Am. Meteorol. Soc.* **81**, 2593–2606.

Silverman, B. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

Stefanova, L. and Krishnamurti, T. N. 2002. Interpretation of seasonal climate forecast using Brier skill score, the Florida State University superensemble, and the AMIP-I data set. *J. Climate* **15**, 537–544.

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

Swets, J. 1973. The relative operating characteristic in psychology. *Science* **182**, 990–1000.

Thompson, P. D. 1977. How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.* **105**, 228–229.

Thornes, J. and Stephenson, D. B. 2001. How to judge the quality and value of weather forecast products. *Meteorol. Appl.* **8**, 307–314.

Tracton, M. S. and Kalnay, E. 1993. Operational ensemble prediction at the National Meteorological Center: practical aspects. *Wea. Forecasting* **8**, 379–398.

Vislocky, R. L. and Fritsch, J. M. 1995. Improved model output statistics forecast through model consensus. *Bull. Am. Meteorol. Soc.* **76**, 1157–1164.

Vislocky, R. L. and Young, G. S. 1989. The use of perfect prog forecasts to improve model output statistics forecasts of precipitation probability. *Wea. Forecasting* **4**, 202–209.

von Storch, H. 1999. On the use of inflation in statistical downscaling. *J. Climate* **12**, 3505–3506.

von Storch, H. and Zwiers, F. W. 1999. *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge, 494 pp.

Wilks, D. S. 1995. *Statistical Methods in the Atmospheric Sciences*. Academic Press, New York.

Yun, W., Stefanova, L., Mitra, A., Vijayakumar, T., Dewar, W. and co-authors. 2005. A multi-model superensemble algorithm for seasonal climate prediction using DEMETER forecasts. *Tellus* **57A**, 280–289.