On the Weighting of Multimodel Ensembles in Seasonal and Short-Range Weather Forecasting

SOPHIE CASANOVA AND BODO AHRENS

Institute for Atmospheric and Environmental Sciences, Goethe-University Frankfurt, Frankfurt, Germany

(Manuscript received 1 December 2008, in final form 4 May 2009)

ABSTRACT

The performance of multimodel ensemble forecasting depends on the weights given to the different models of the ensemble in the postprocessing of the direct model forecasts. This paper compares the following different weighting methods with or without taking into account the single-model performance: equal weighting of models (EW), simple skill-based weighting (SW), using a simple model performance indicator, and weighting by Bayesian model averaging (BMA). These methods are tested for both short-range weather and seasonal temperature forecasts. The prototype seasonal multimodel ensemble is the Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER) system, with four different models and nine forecasts per model. The short-range multimodel prototype system is the European Meteorological Services (EUMETNET) Poor-Man's Ensemble Prediction System (PEPS), with 14 models and one forecast per model. It is shown that despite the different forecast ranges and spatial scales, the impact of weighting is comparable for both forecast systems and is related to the same ensemble characteristics. In both cases the added value of ensemble forecasting over single-model forecasting increases considerably with the decreasing correlation of the models' forecast errors, with a relation depending only on the number of models. Also, in both cases a larger spread in model performance increases the added value of combining model forecasts using the performance-based SW or BMA weighting instead of EW. Finally, the more complex BMA weighting adds value over SW only if the best model performs better than the ensemble with EW weighting.

1. Introduction

The purpose of an ensemble prediction system (EPS) is to predict forecast probabilities of weather and climate events by the integration of an ensemble of numerical predictions (Lorenz 1965; Molteni et al. 1996; Ehrendorfer 1997; Palmer 2000). The ensemble members differ because of slightly different initial states, different model setups, or the application of different model systems. In the latter case the ensemble forecast is performed by a multimodel EPS (Atger 1999; Ebert 2001). The spread of the ensemble forecast provides a measure of the trustworthiness of the forecast. Additionally, the average forecast of multiple ensemble members often performs better than a forecast by any single member.

If a global EPS consists of multiple forecast members provided by a single-model, then the EPS predicts forecast probabilities of weather events by the integration of an ensemble of numerical weather predictions, which are initialized with slightly different states. These initial states are usually assumed to be equally realistic, and therefore all forecast members get the same weight in the determination of the forecast probability of an event (see the discussion in Katz and Ehrendorfer 2006). In the case of a limited-area EPS this can change. For example, the Consortium for Small-Scale Modeling (COSMO) Limited-Area EPS (LEPS; Marsigli et al. 2005) weights its members differently. These members are driven by a subset of the members of a global-scale EPS only. The subset of global-scale members is selected by a cluster analysis, and these members have different importance depending on the cluster sizes. This importance is inherited by the members of the limited-area EPS as weights in either the determination of forecast probabilities or the averaging of the members to a single deterministic mean forecast (Ahrens and Walser 2008; Jaun et al. 2008).

Corresponding author address: Bodo Ahrens, Altenhoefer Allee 1, Institute for Atmospheric and Environmental Sciences, Goethe-University Frankfurt, D-60438 Frankfurt/Main, Germany.
E-mail: bodo.ahrens@iau.uni-frankfurt.de

DOI: 10.1175/2009MWR2893.1

If an EPS consists of a set of forecasts by several different models, then the members are well discriminable because the various models have, in general, different forecast performance. It is then useful to apply different weights to the multimodel EPS members depending on the performance of the models in the preceding forecasts. For example, Krishnamurti et al. (1999) combine the deterministic forecasts of several models using different weights, which are obtained by means of a multilinear regression of the forecast anomalies during a training period. This yields a mean deterministic forecast. The purpose of their application was seasonal ensemble forecasting with global models. A recent paper by Marrocu and Chessa (2008) compares different weighting methods of deterministic short-range forecasts by three limited-area models. For example, they show that Bayesian model averaging (BMA; Raftery et al. 2005; Sloughter et al. 2007) can improve the EPS forecasts of the raw equally weighted ensemble. Performing numerical experiments with toy models, Weigel et al. (2008) have shown that both equally and unequally weighted multimodel ensembles can perform better than the best model within the ensemble of models.

Here, we investigate different multimodel ensemble weighting methods at two different forecast scales. These scales are (i) the global seasonal forecast scale, and (ii) the limited-area short-range forecast scale. For these two scales, the different sources of forecast errors are most important. The forecast performance at the seasonal scale is largely dependent on the description of the physics in the model system. At the limited-area short-range scale the performance depends mainly on the forecast initialization of the limited-area models, and the quality of the driving global models and their initialization. It is the major goal of this paper to compare the impact of weighting in the postprocessing of multimodel ensembles with different forecast scales, and to understand the similarities and/or differences between these impacts. For this goal we implemented three methods of weighting in ensemble forecast postprocessing: equal weighting of models (EW), simple skill-based weighting (SW), using a simple model performance indicator, and weighting by BMA. These methods are described in more detail below.

As a prototype of a global seasonal forecast system, we use the hindcasts from the Development of a European Multimodel Ensemble System for Seasonal to Interannual Prediction (DEMETER) project (Palmer et al. 2004). We can use 33 yr of hindcasts with this product; favoring it over the newer product of the Ensemble Based Predictions of Climate Changes and their Impacts (ENSEMBLES) project (information online at http://www.ensembles-eu.org), which has at present 10 yr of available seasonal hindcasts. As the short-range product we use multimodel forecasts that are collected and delivered by the Short-range Numerical Weather Prediction (SRNWP)–Poor Man's EPS (PEPS; see, e.g., Heizenreder et al. 2006) project. It collects the operational forecasts by limited-area models from more than 20 European national meteorological services.

Often, the ensemble of forecasts delivered by an EPS is interpreted and evaluated probabilistically (Katz and Ehrendorfer 2006; Ahrens and Walser 2008), or the members are applied as input for impact models (Jaun et al. 2008). The ensemble mean deterministic forecast is easier for interpretation. Here we will focus on the impact of weighting on the performance of the ensemble mean forecast. This simplifies the discussions and allows us to use deterministic scores in the evaluation of the ensemble forecasts. Consequently, this paper does not investigate the impact of weighting on the probabilistic calibration of the ensembles.

The paper is organized as follows: Section 2 describes the ensembles used, the weighting methods, the method of performance evaluation, and the observational data applied for training of the weighting methods and for evaluation. Sections 3 and 4 give a summary of global and local performance, respectively, of the differently weighted ensembles. Section 5 discusses the impact of the different weighting methods in detail. Section 6 gives the conclusions.

2. Data and methods

This section introduces the two ensemble forecast systems, the data for forecast postprocessing and evaluation, and the methods applied in postprocessing and evaluation.

a. Ensemble forecast products

The first ensemble forecast product considered is from the DEMETER project (Palmer et al. 2004), founded under the European Union Fifth Framework Environment Programme. It yields a seasonal superensemble (i.e., a multimodel with multiple ensemble members for each model prediction system) of seven global coupled ocean–atmosphere models integrated for different periods (all ending in 2001). The hindcasts are delivered on a spatial grid with 2.5° of horizontal grid spacing. Each model is integrated for 6 months and 9 times with different initial conditions (yielding nine forecasts per model). A complete description of the project and its main results can be found on the DEMETER Web site (online at http://www.ecmwf.int/research/demeter). In this paper, we show only the results for the 2-m temperature hindcasts for the spring

seasons (March–May; hindcasts are initialized in the beginning of February). Forecasts for the other seasons show similar results. To have a long enough period (1969–2001) for bias correction of the forecasts and training of the SW and BMA weighting methods we had to restrict ourselves to four models from the following: the Met Office (UKMO; United Kingdom), the Centre National de Recherches Météorologiques (CNRM; France), the Max Planck Institute for Meteorology (MPI; Germany), and the European Centre for Medium-Range Weather Forecasts (ECMWF; United Kingdom). As discussed in the introduction, the nine forecasts using the same model cannot be discriminated, and thus no unequal weighting can be applied. Additionally, this paper does not consider probabilistic information potentially delivered by ensemble forecasts. Therefore, we consider four deterministic forecasts, each of which are constructed by the averaging of nine single-model ensemble members. We call this ensemble system DEMETER-36.

The second forecast product is delivered by the SRNWP–PEPS (Heizenreder et al. 2006) project. It collects the operational forecasts from more than 20 European national meteorological services in order to construct a multimodel ensemble prediction system of high-resolution short-range regional models [i.e., different implementations of the regional models (see Heizenreder et al. 2006) of the four regional model consortia, Aire Limitée Adaptation Dynamique Développement International (ALADIN), COSMO, High-Resolution Limited-Area Model (HIRLAM), and UKMO organized in the Network of European Meteorological Services (EUMETNET; see information online at http://srnwp.met.hu)]. The model setups have grid lengths between 7 and 22 km and apply different model domains, initialization, and coupling models. All forecasts are interpolated onto a forecast grid with a horizontal grid spacing of 7 km. The possible ensemble size of the PEPS depends on the region of interest. For this study we have restricted ourselves to the daily maximum temperature forecasts initialized at 0000 UTC with a maximum lead time of 30 h from 14 model implementations, which cover Germany very well, and are available for this paper in the period from November 2006 to February 2008. We have removed from the dataset all of the days with one or more missing model, and we end up with about 300 forecast days. This allows us to deal with a homogenous ensemble; We call this ensemble PEPS-14.

The different ensemble sizes and types of DEMETER-36 and PEPS-14 complicate the goal of comparing the impact of weighting methods in postprocessing. Therefore, in the following discussions the focus will be on the comparison of two derived ensemble products. The first one, named DEMETER-4, is constructed by random selection of one forecast member out of nine per model for each forecast event. Therefore, a forecast with DEMETER-4 consists of one realization of each of the four models. The second derived product, named PEPS-4, is created by subsampling of PEPS-14 without replacement, yielding forecast ensembles of four models. In the following we show forecast experiments where 50 random subsamples (out of 1001 possible) are selected and considered in the discussion of weighting effects.

b. Observational reference

For postprocessing and evaluation of the forecasts it is important to have adequate observational data. In the case of DEMETER, the seasonal mean 2-m temperature of the 40-yr ECMWF Re-Analysis (ERA-40; Uppala et al. 2005) is taken as the reference at the spatial grid of $2.5° \times 2.5°$ of the DEMETER product. In total, evaluation data are available for 144×71 grid points and 33 yr, and thus the total evaluation sample contains about 3×10^5 events.

In the case of PEPS postprocessing and evaluation, we use daily maximum temperature observations of 357 operational weather stations of the German meteorological service. The observational data are compared with the forecast value in the closest grid cell of PEPS. Because of missing observations and models, the number of days available for evaluation is about 300 (which is slightly dependent on the station). Thus, the number of events available for the evaluation is about 10^5.

c. Postprocessing

The postprocessing applied here consists of two steps: first, a bias correction of the single-model forecasts is applied; second, the model forecasts are weighted before deterministic mean ensemble forecasts are considered in the evaluation. Bias correction and weighting are based on empirical information generated from comparisons of forecasts and observational references in training periods.

For DEMETER, there are 33 seasonal forecasts of consecutive spring seasons available, which can be used in the forecast experiments discussed here. Leave-one-out cross-validation is applied: the current season is applied for validation, and the remaining 32 seasons are used as the training data in postprocessing. Therefore, the validation sample is 33 postprocessed forecasts per grid point. For PEPS, the training periods for the current forecast are composed of the previous 35 forecast days, which may not be consecutive because of the gaps in the data. Therefore, about 265 post processed PEPS forecasts are available per station.

1) BIAS CORRECTION

The most important step in postprocessing the direct model outputs is the correction for model mean bias. For DEMETER, the bias is estimated per grid point for each model in the multimodel ensemble. This is the mean difference of the seasonal forecasts and the ERA-40 data in the training period. The biases are spatially variable, but per grid point the standard error of the bias is typically about 3 times smaller than the bias estimate despite the small sample of seasonal forecasts in the training periods of DEMETER.

For PEPS, the bias is calculated stationwise for the current forecast day using the forecast and observational data from the training dataset based on the 35 previous days. Here, the bias varies in space and time. However, using a local bias correction calculated on the last 35 days is easily implemented, and takes into account the current season and weather situation. There are uncertainties in bias estimation, but they do not limit the following discussion on the relative impact of different weighting schemes.

2) WEIGHTING OF MULTIMODEL ENSEMBLES

For the multimodel ensembles PEPS-14, PEPS-4, DEMETER-36, and DEMETER-4, we tested three methods of ensemble weighting. The first method is EW. In this method each model gets the same weight in the multimodel forecast; the model with index m gets the weight $w_m = 1/M$, with M the number of models. This is the most common method in single-model ensemble weighting (Katz and Ehrendorfer 2006). Within EW any knowledge about the performance of the models is neglected.

As the second method, a simple skill-based weighting method that considers model performance is applied grid point–wise for DEMETER and stationwise for PEPS. This SW method uses the inverses of the mean-square errors of the model forecasts in the training period as weighting factors (after normalization by the sum of the weights). This method neglects any interdependence of the performances of the models in the ensemble. Application of the mean absolute error as an alternative simple performance measure has been tried and does not change the following discussion and conclusions.

The third method applied is BMA. BMA has been applied in probabilistic weather forecasts by Raftery et al. (2005), Sloughter et al. (2007), Wilson et al. (2007), and Marrocu and Chessa (2008). To the full extent of the knowledge of the authors, the BMA method is applied here for the first time to seasonal temperature forecasts. BMA is a statistical way of postprocessing forecast ensembles to create predictive probability density functions for weather-related quantities. The predictive probability density function is estimated as a weighted average of individual density functions centered on the individual bias-corrected forecasts. Here, it is assumed that the temperature forecasts are well approximated by normal distributions. It is further assumed that the error variances of all models are approximately equal. Then, the weights and the common variance can be determined by a maximum likelihood procedure [using the expectation-maximization algorithm as in Raftery et al. (2005)] that is applied over each forecast's training period. The BMA is applied per grid point for DEMETER and per station for PEPS. Here, only the BMA weights are necessary for calculation of the deterministic mean ensemble forecast. The BMA method weights better models relatively more than is done by SW. However, it additionally respects intermodel relationships. If two models show relatively high forecast error correlation, then one model is downweighted to reduce redundant information in the ensemble. This yields a larger risk of overfitting the training data than that for SW [discussed in Hamill (2007) as a response on an application of BMA in Wilson et al. (2007)].

d. Evaluation scores

In this paper, to keep things as simple as possible, only the means of weighted ensemble forecasts are evaluated and discussed. Therefore, no probabilistic evaluation with scores, like the Brier score, is performed. We have chosen instead the well-known and easily interpretable mean-square error (MSE) score

$$\mathrm{MSE} = \frac{1}{IT}\sum_{i}^{I}\sum_{t}^{T} a_i (y_{it} - o_{it})^2, \qquad (1)$$

with o_{it} the value of the observational reference, and y_{it} the forecast at time t and location i. The factor a_i considers the gridcell area at location i: in the case of PEPS $a_i = 1$ for all locations, and in the case of DEMETER $a_i = \cos(\mathrm{latitude}_i)$.

Here, T is the number of forecast experiments evaluated for each station or grid point; that is, T is about 260 for PEPS depending slightly on station data availability, and T is 33 for DEMETER. In case of total evaluations over all forecast events in space and time, I is 357 (the total number of stations) for PEPS and 10 244 (the total number of grid points over the globe) for DEMETER. Therefore, the number of evaluated forecast events $I \times T$, and thus the evaluation uncertainty, is comparable for the total DEMETER and PEPS evaluations. For local evaluations $I = 1$ for PEPS, yielding about 260 local evaluations. In order to enhance the comparability of the local evaluation, the forecast events of the DEMETER

TABLE 1. Performance of PEPS-14 and the average performance of 50 random samples using PEPS-4. Different postprocessing methods are applied (EW, SW, and BMA) and these are compared to non-ensemble methods (AVER and BEST). The performance is given by the RMSE (K) of the forecasts in the evaluation period. The performance of the persistence forecast REF is 3.49 K. Bootstrapping gives 90% uncertainty estimations below ±0.02 K in all cases.

Method	EW	SW	BMA	AVER	BEST
PEPS-14	1.51	1.46	1.40	1.86	1.45
PEPS-4	1.58	1.54	1.53	1.85	1.61

product are pooled in small regions of 3×3 pixels ($7.5° \times 7.5°$), yielding $I = 9$. The impact of the different temporal and spatial dependence of the events on the comparison of the PEPS and DEMETER results is small and therefore neglected.

Below we want to compare the performance of EPSs with different space and time scales, and at different locations. The relative performance of forecasts by the evaluated forecast system (Fcst1) in comparison with forecasts by the reference forecast system (Fcst2) is measured by a normalized skill score (Wilks 2006). Here, we use the skill score (SS) defined by

$$\mathrm{SS}(\mathrm{Fcst1}, \mathrm{Fcst2}) = 1 - \frac{\mathrm{MSE}(\mathrm{Fcst1})}{\mathrm{MSE}(\mathrm{Fcst2})}, \qquad (2)$$

The skill score value is zero if the performance of Fcst1 equals the performance of Fcst2. The skill score is unity if Fcst1 is perfect in terms of MSE.

Often a forecast system is compared against a simple reference forecast (REF). In the case of DEMETER, a good, simple REF is the climatological value (i.e., the mean seasonal temperature in the training period). In the case of PEPS, the REF chosen here is the persistence forecast (i.e., the observed temperature value on the day before the forecast day).

A main goal of this paper is to compare different weighting methods in forecasting. This is done by calculating the skill scores of, for example, BMA relatively to EW, which we denote as SS(BMA, EW). Additionally, it is useful to compare a multimodel ensemble forecast system with the MSE of the best model forecast system (BEST) in the evaluation period, or with the average MSE of the single-model forecasts (AVER). BEST can be interpreted as the extreme weighting example, where all of the weight is put on the model, which is the best on average in the evaluation period. Skill score values with AVER apply the MSE of the single-model forecasts in the evaluation of SS, and thus give an average skill of the models neglecting the forecast skill that is added by the ensemble approach.

TABLE 2. Same as Table 1, but for DEMETER-36 and DEMETER-4. The performance of the climatological forecast REF is 0.75 K. Bootstrapping gives 90% uncertainty estimations below ±0.01 K in all cases.

Method	EW	SW	BMA	AVER	BEST
DEMETER-36	0.71	0.70	0.70	0.81	0.70
DEMETER-4	0.76	0.71	0.71	1.01	0.76

3. Total performance of the ensemble forecasts

In this section we discuss the total performances of the PEPS and DEMETER forecast systems in different postprocessing configurations. We show the square root of the MSE score (RMSE) averaged over all of the events in space and time. Therefore, the spatial sum in the score calculation is globally over all of the grid points in the case of the DEMETER systems, and over all 357 German weather station locations in the case of PEPS.

Table 1 gives the RMSE for PEPS in different configurations. As expected, the multimodel ensemble PEPS-14, with 14 models, performs better than the average of randomly selected ensembles of four models in PEPS-4. Both ensembles perform independently of the weighting method, which is obviously better than the average single-model forecast AVER. It is more interesting to note that BMA weighting performs best, but the added value over SW weighting is only small on average [the skill of BMA over SW given by SS(BMA, SW) is between −0.05 and 0.05 for PEPS-4 and 0.05 for PEPS-14]. Using EW, and thus neglecting model performance in weighting, gives the worst performance in either ensemble forecasting case. For PEPS the model forecasts are very good in comparison to the chosen simple reference by persistence forecasting REF: SS(AVER, REF) = 0.47.

For DEMETER-4, the RMSE scores given in Table 2 show behavior that is similar to that of PEPS. The SW and BMA weightings are comparable, but equal weighting is worse [SS(BMA, EW) = 0.06]. The AVER is worst of all, and this time it is even worse than the reference forecast REF (i.e., the climatological forecast, which has a RMSE of 0.75 K). This shows that it is on average worst to use a single forecast of a single-model only. Also, the DEMETER-4 forecast with EW is worse than the climatology. The results are different for DEMETER-36. DEMETER-36 uses four models, as does DEMETER-4, but has nine forecast members per model. With nine forecast members per model the equal weighting is already almost as efficient as applying weighting that considers the model performance [SS(BMA, EW) = 0.02 only]. Obviously, the better performance of DEMETER-36 with EW in comparison to DEMETER-4 with EW is

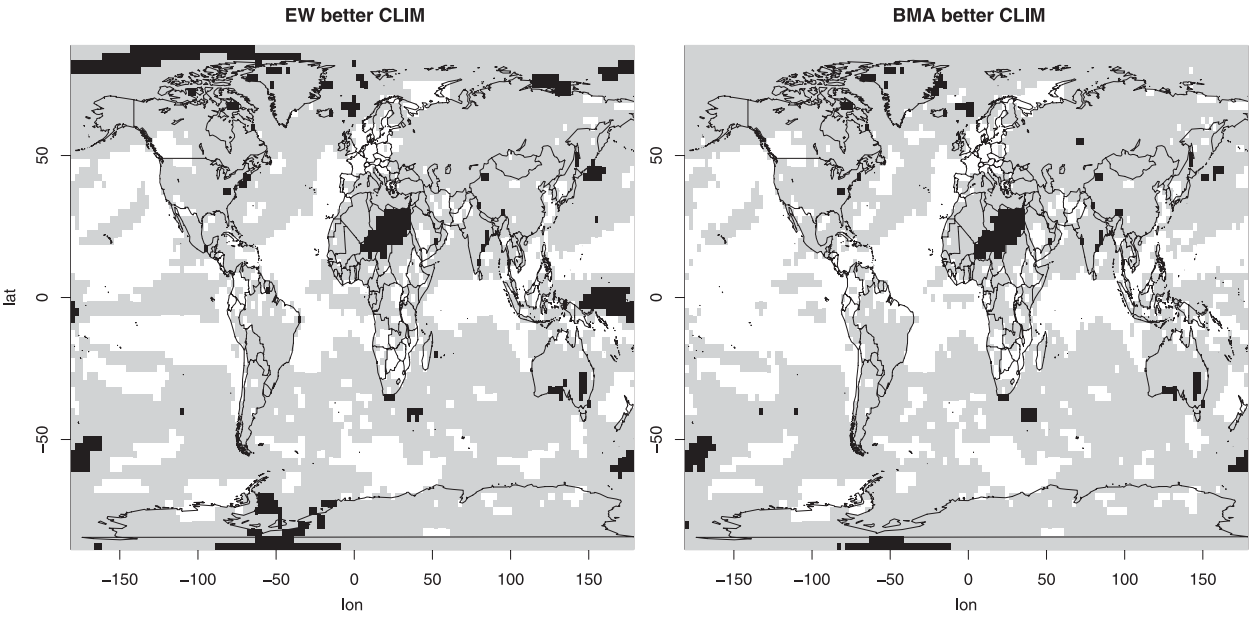

FIG. 1. Skill of the DEMETER-36 forecast system using (left) EW and (right) BMA. Marked regions indicate where DEMETER forecasts are significantly better than climatology (white), DEMETER forecasts are significantly worse (black), and there is no significant difference in performance at the 10% significance level (gray).

because of the multiple forecast members per model. When DEMETER-36 is applied with performance weighting it only slightly adds value to the forecasts. However, it is important to note that DEMETER-4, with four SW- or BMA-weighted forecasts, is as good as DEMETER-36, which has 36 forecasts in the applied deterministic evaluation.

4. Local performance of the ensemble forecasts

The performances of the forecast systems are spatially heterogeneous. Figure 1 compares the performance of the DEMETER-36 ensemble forecast with the climatological reference forecast. This is accomplished by means of a paired Wilcoxon test for differences between the forecasts and ERA-40 at the 10% significance level. It shows that there are distinct regions with DEMETER-36 forecasts that are better than the climatological reference forecast. However, there are also regions within DEMETER forecasts that are significantly worse than climatology (e.g., in parts of the El Niño region, in the Arctic, and in northern Africa). DEMETER-36 with EW is significantly better than climatology in 29% and worse in 4% of the grid points.

Figure 1 also shows that using BMA instead of EW in postprocessing can improve the forecasts, with 32% of the grid points being significantly better than climatology, and 2% being significantly worse than climatology. These improvements over EW are mainly in the Arctic (delimited with latitudes greater than 70°N) and in the traditional Niño-3 and Niño-4 regions [delimited with latitudes between 5°S and 5°N and longitudes between 160°E and 90°W (Trenberth 1997)]. This is consistent with the results from Doblas-Reyes et al. (2005). The mean skill scores of the different weighting methods over climatology for the Arctic region are, respectively, −0.07, 0.06, and 0.06 for the EW, SW, and BMA methods. For the Niño-3 and -4 regions they are 0.35, 0.62, and 0.66, respectively. The results in these two regions will be further discussed and compared to the results in the other regions later in the text. It is important to note that the relatively small skill score values in the Arctic are not necessarily a consequence of either model or weighting difficulties in the Arctic region, but mainly result because of smaller potential predictability of the climate system in the Arctic than in the tropics at the seasonal scale (Kumar et al. 2007). Nevertheless, the SW and BMA weighting is able to slightly increase the skill score values.

This shows the dependence of the different methods' efficiency on the region of interest. For PEPS, forecasts also a high interstation variability in skill scores, which can be observed with the SS(EW, REF) varying from −0.51 to 0.93. In this paper we will use this wide range of efficiency for both the DEMETER and PEPS ensembles to gain insight into which parameters affect the performance of the different weighting methods.

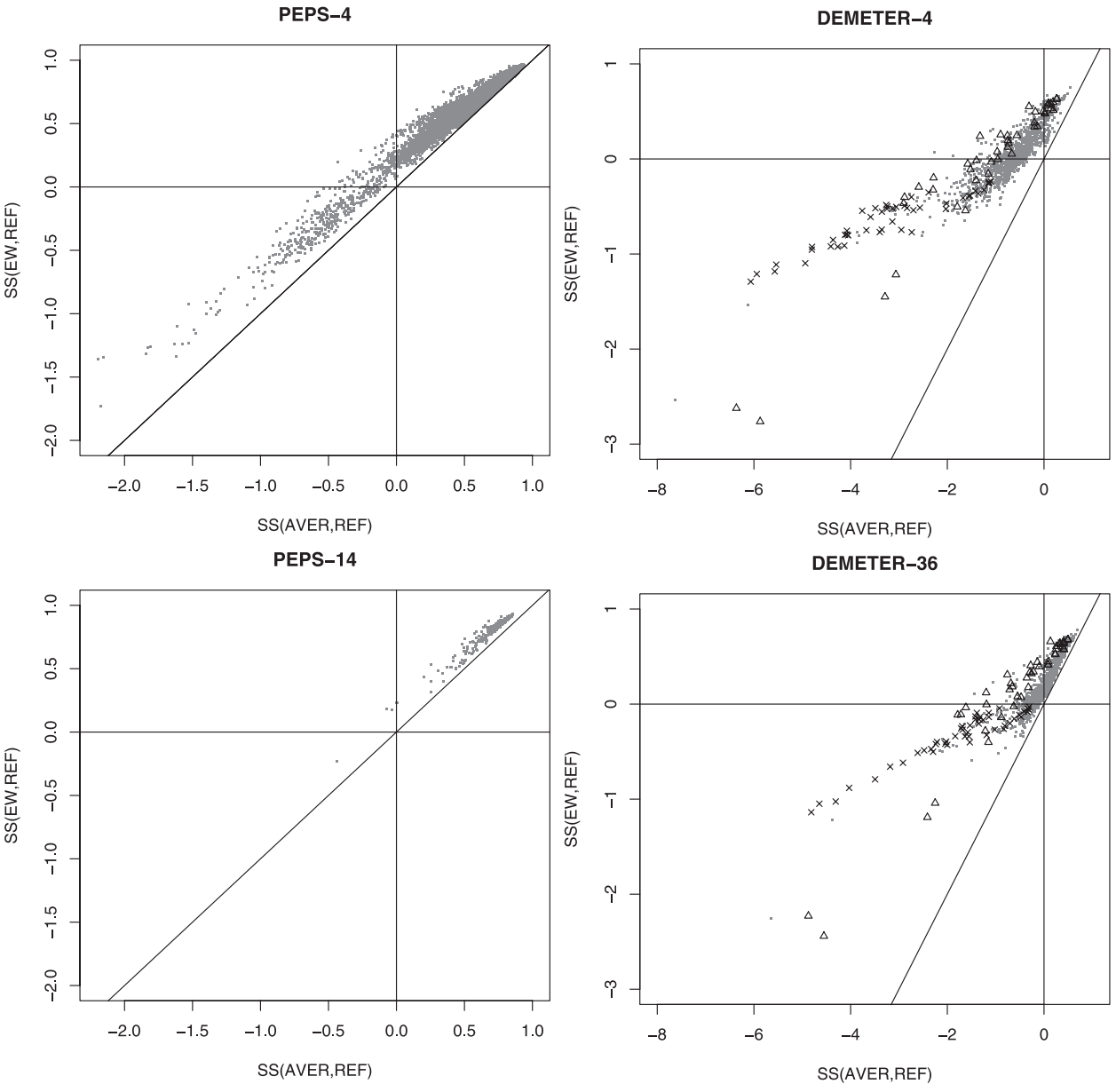

FIG. 2. Scattergrams showing SS(EW, REF) vs SS(AVER, REF): (top) (left) PEPS-4 and (right) DEMETER-4; (bottom) (left) PEPS-14 and (right) DEMETER-36. In the DEMETER scattergrams the pairs of values in the Arctic region are marked with "x" and in the El Niño region with "Δ." These regions are defined in the text in section 4. The diagonal (the line where the skill of an equal weighting combination is equal to the average single-model skill) is drawn.

5. Discussion of the different ensemble weighting methods

a. Equal weighting

First of all, for both PEPS and DEMETER, the multimodel ensemble performance depends on the average performance of the individual models. Figure 2 shows skill score values of SS(EW, REF) versus SS(AVER, REF) for all PEPS stations, and averaged values for regions with $I = 3 \times 3$ grid points for DEMETER. This figure indicates a strong linear correlation between the performance of the multimodel ensemble forecasts with equal weighting and the average performance of the models. The linear correlation between SS(EW, REF) and SS(AVER, REF) is 0.98 and 0.89 for PEPS-4 and DEMETER-4, respectively. It is noteworthy that the averaged ensemble forecasts always perform better than the averaged performance of the single-model forecasts

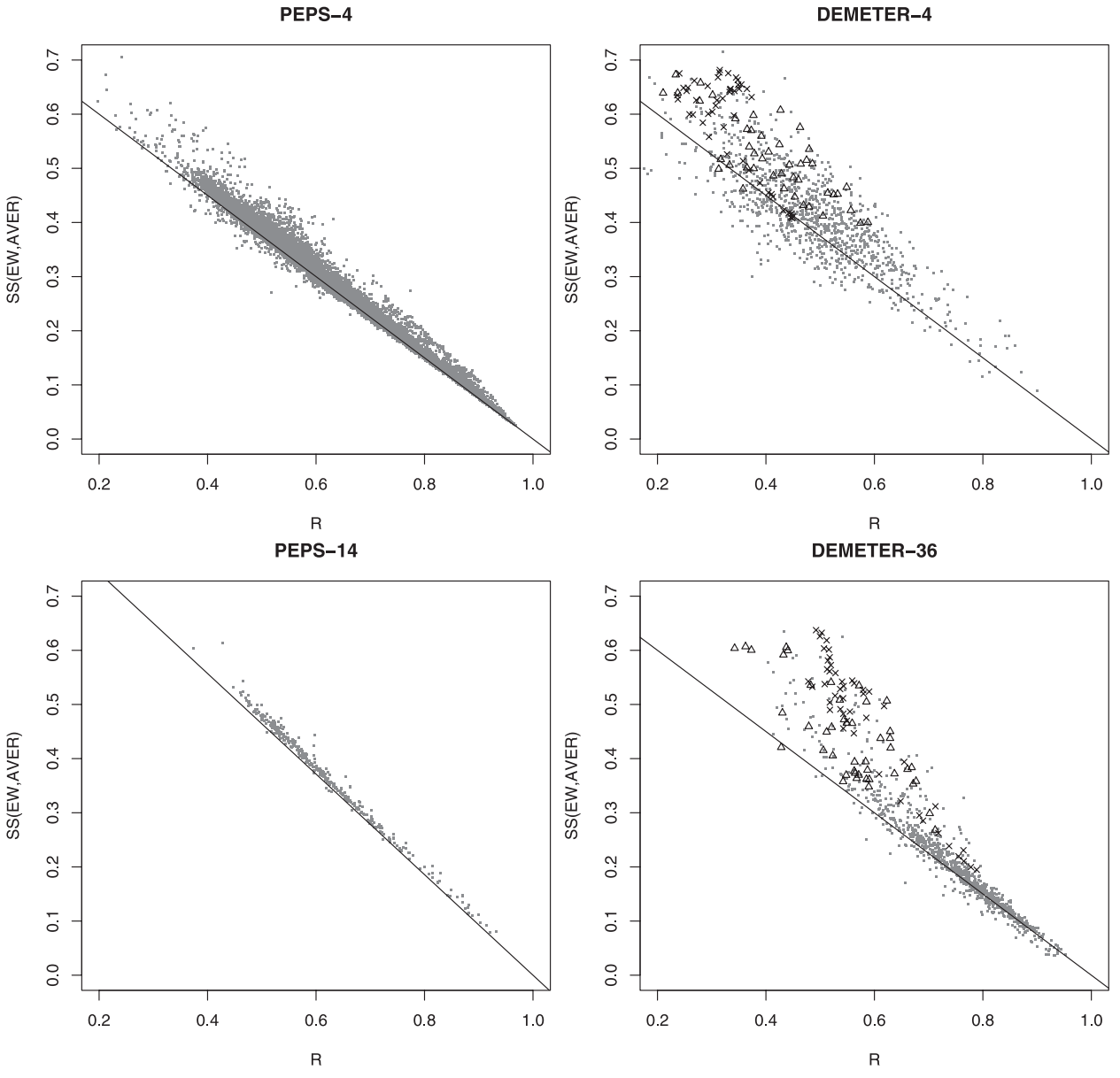

FIG. 3. SS(EW, AVER) of the ensemble forecasts with equal weighting normalized with AVER vs the mean-error correlation of the single-model forecasts R: (top left) PEPS-4 and (top right) DEMETER-4; (bottom left) PEPS-14 and (bottom right) DEMETER-36. The ideal regression line $(N - 1)/N(1 - R)$ with N the number of models assuming model-independent error variances and covariances is drawn. The symbols are the same as in Fig. 2.

given with AVER (as expected, because averaging minimizes the MSE for normal random variables). However, the added value of multimodel ensembles over the individual models is quite variable. Figure 2 shows that for DEMETER the added value is greater in the Arctic and El Niño regions (defined in section 4) than in many other regions, and it is especially pronounced if the model performance is poor on average.

Why is the positive impact of model combination so variable? The added value of ensemble forecasting is related to the independence of the different models. On average, an ensemble mean forecast is better than a single-model forecast if the errors of the forecasts in the ensemble compensate each other to some extent. This is illustrated in Fig. 3. In these scattergrams, the gain through ensemble forecasting is shown over a correlation parameter R. This parameter R quantifies the temporal correlation of the forecast errors of the models by averaging the pairwise correlation coefficients of the forecast errors. In case of PEPS-4, DEMETER-4, and

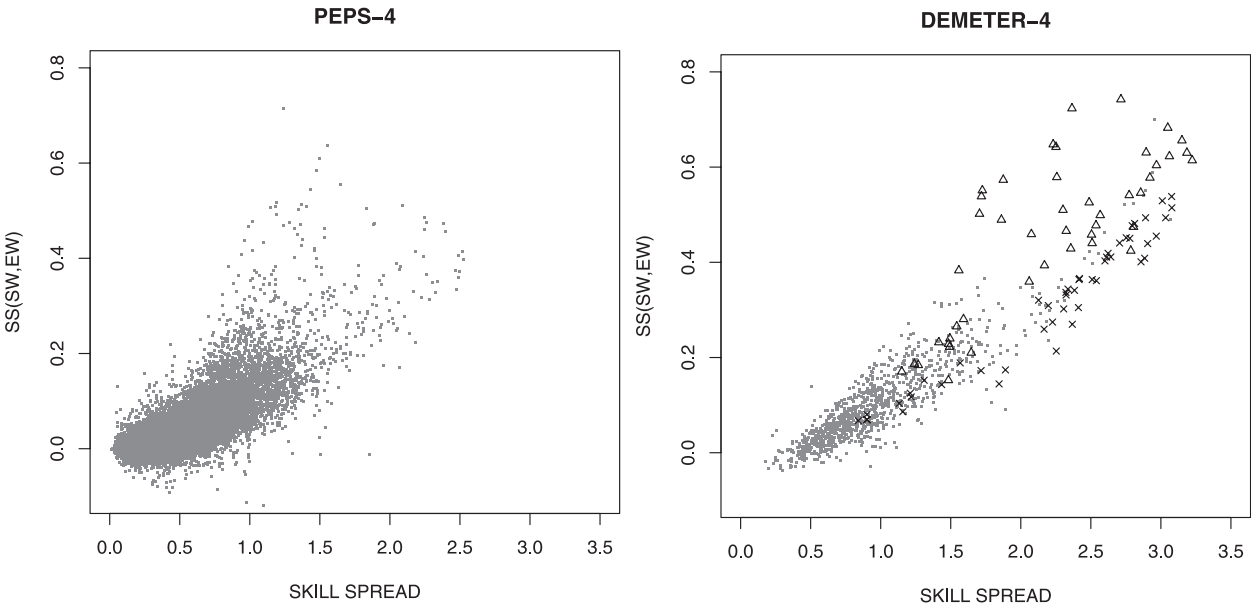

FIG. 4. SS(SW, EW) of the ensemble forecasts with simple skill-based weighting relative to the forecasts with equal weighting vs the relative spread of model skills (SKILL SPREAD); (left) PEPS-4 and (right) DEMETER-4. The symbols are the same as in Fig. 2.

DEMETER-36, the number of error pairs is six, and in the case of PEPS-14 the number of pairs is 91. If the error variances and covariances are assumed to be model independent, it is possible to show that SS(EW, AVER) = $(N - 1)/N(1 - R)$ with N the number of models. Figure 3 shows very clearly that the added value of the ensemble increases when the models are less correlated. Despite the difference in scales of the seasonal and short-range forecasts, this increase is extremely similar for the PEPS-4 and DEMETER-4 ensembles. We can see that for grid points in the El Niño and Arctic regions the correlation of the forecast errors is very small, so that the extra skill added by the ensemble is particularly high.

For DEMETER-36, each of the four model forecasts already has an average of nine members, and because of the smoothing effect of averaging the correlation of the forecast errors tends to be larger. This explains why the added value of the multimodel ensemble is smaller than that for DEMETER-4 (Table 2). It is also interesting to note that for DEMETER-36 the grid points in the Arctic and El Niño regions show a tendency to have a larger dependence on error correlation. This is related to the observation that in these regions the spread in model skill is larger than elsewhere, as discussed in more detail in the next paragraph (see Fig. 4).

As expected, the skill score versus the error correlation dependence is steeper for PEPS-14 (based on $N = 14$ models) than for the three ensemble products based on four models. This is because more models with the same mean error correlation yield more independent information in the ensemble and this adds value to the ensemble forecast. Thus, the multimodel ensembles are able to compensate for errors (in initialization of the forecasts and model physics) as long as there are enough debiased forecasts with independent errors.

b. Simple skill-based weighting

Figure 4 shows that the ensemble forecasts with SW perform better than those with EW on average. This is quantified by the skill score SS(SW, EW) with values above zero most of the time. In Fig. 4 the skill scores are shown only for PEPS-4 and DEMETER-4, but the same is true for PEPS-14 and DEMETER-36. This is affirmed by the percentages of grid points/stations with SW better than EW, given in Table 3. The figure also depicts the positive dependence of this skill score on the relative spread of the skill of the models in the ensembles. This relative spread is calculated as the difference between

TABLE 3. Percentages of grid points/stations with better forecasts on average using one of the performance-based weighting methods (SW, BMA, or BEST) than using EW.

Method	SW	BMA	BEST
PEPS-4	89	62	33
PEPS-14	98	86	57
DEMETER-4	95	88	49
DEMETER-36	66	59	56

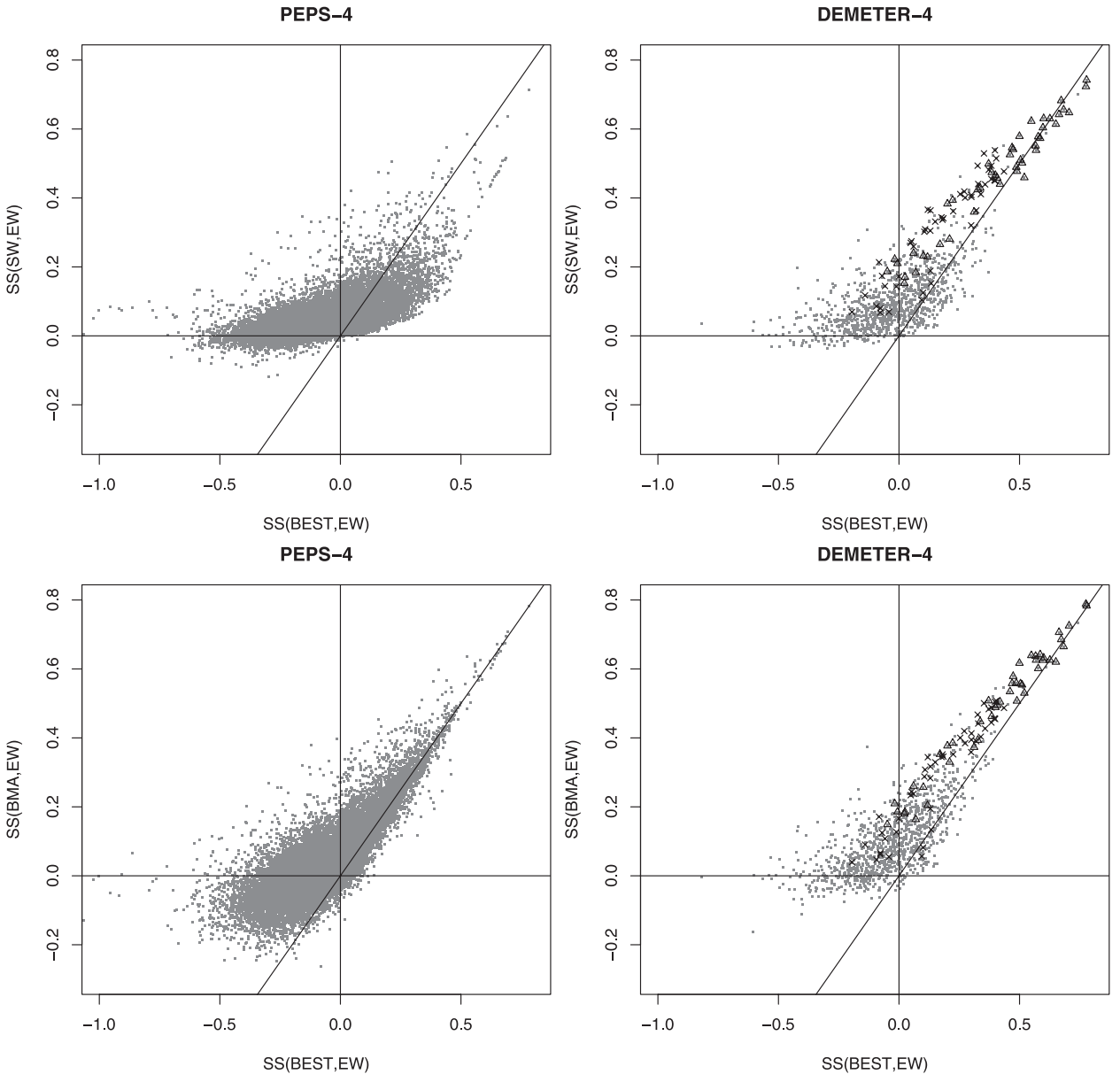

FIG. 5. Scattergrams of the (top) SS(SW, W) and (bottom) SS(BMA, EW) vs SS(BEST, EW): (top left) PEPS-4 and (top right) DEMETER-4; (bottom left) PEPS-14 and (bottom right) DEMETER-36. The first diagonal is drawn. The symbols are the same as in Fig. 2.

the MSE of the worst model minus the MSE of the best model divided by the averaged MSE of all models, with the MSEs evaluated for all of the forecast events. The scattergrams in Fig. 4 show that SW is able to put more weight on the better models when the relative spread is larger, and this improves the forecasts. It is interesting to note that the dependence of the added value of SW over EW on spread in model skill is similar for the short-range PEPS forecasts and the long-range DEMETER forecasts, except that the relative spread can reach much greater values for DEMETER.

Obviously, the spread in model skill is relatively large in the El Niño and Arctic regions, and in consequence the performance weighting is especially effective in these regions. In fact, in the Arctic region one model is much worse than the two others. In the El Niño region there is both a best and a worst model in terms of performance in the training periods. Elsewhere, the spread in model skill is small, and is even smaller for DEMETER-36 than for DEMETER-4. This explains why SW makes a smaller improvement over EW in case of the super-ensemble DEMETER-36 (see Tables 2 and 3).

TABLE 4. Percentages of grid points/stations where the method SW, BMA, or BEST performs best, taking into account only grid points/stations with SS(BEST, EW) > 0.

Method	SW	BMA	BEST
PEPS-4	11	60	29
PEPS-14	5	65	30
DEMETER-4	20	59	21
DEMETER-36	14	14	72

c. Bayesian model averaging

In Section 3 we have shown that BMA performs only slightly better than SW in the total averages. Table 3 shows that BMA is more effective than EW in fewer grid points/stations than SW is (perhaps because of the risk of misweighting with BMA). This is confirmed by Fig. 5, with more grid points/stations showing values of SS(BMA, EW) < 0 than SS(SW, EW) < 0. Like SW, the BMA method weights the ensemble members according to model skill, but less directly. Here, we investigate under what conditions the more complex BMA weighting can give better results than SW.

Figure 5 shows that BMA is the more effective weighting method if there is a locally best model. This can be seen very clearly for PEPS-4, where the locally best model is often better than the ensemble with SW (upper left panel). BMA is almost always able to improve on the locally best model (lower left panel).

This is quantified in Table 4. For the grid points/stations where the forecast with BEST is already better than that with EW, the forecast with BMA is the best one in about 60% of the cases. Thus, BMA is better than BEST and SW in 60% of the cases with relatively good singe model forecasts. Obviously, BMA is more capable of using the good information in consistently good single-model forecasts than SW. This is because the average standard deviation of the weighting coefficients of BMA is about 3 times larger than that of SW (not shown). However, as Table 4 shows, there is one exception: DEMETER-36. If the best model ensemble in the superensemble is better than the ensemble with EW, then in 72% of the cases BEST stays better than the ensemble with either BMA or SW. This is the positive impact of the single-model ensembles, which makes it difficult to add value through skill-based model weighting. Note that the locally best model is not necessarily the best model in all regions.

6. Conclusions

This paper has investigated the impact of weighting on multimodel ensemble forecasts of temperature with different spatial scales and forecast ranges. It has been shown that the impact is independent of the forecast scales. This is the case despite the different main sources of forecast errors for the seasonal-range forecasts by DEMETER (mainly errors in model description and boundary conditions), and the short-range forecasts by PEPS (mainly forecast initialization uncertainties).

For both forecast systems, DEMETER and PEPS, the ensemble forecasts are better than the average single-model forecasts [as extensively discussed for DEMETER in Hagedorn et al. (2005)]. Weighting of the ensemble members with model performance (determined in a training period) improves the forecasts (shown for short-range forecasts also by Marrocu and Chessa 2008). For performance-based weighting a simple skill-based weighting (SW) based on the bias-corrected model's MSE and Bayesian model averaging (BMA) has been applied.

In most of the cases the simple skill-based weighting method is as effective as, or even more effective than, the BMA. Considering the computational cost of the BMA method and its risk for overfitting (as discussed in Hamill 2007), SW seems to be sufficient in most cases. The BMA is worth its additional computational effort only if there are consistently good models in the ensemble, because BMA effectively puts more weight on them. However, this conclusion might be valid only for the deterministic evaluation performed here and the temperature forecasting evaluated here. Additionally, we have not considered possible multigridpoint or multistation extensions of the weighting methods.

Nevertheless, the study indicates: (i) that the ensemble forecasts improve at the short-range as well as at the seasonal-range scale with an increasing number of models in the ensemble with as low a forecast error correlation as possible, and (ii) that some simple skill-based weighting with model performance in a training period improves the forecasts efficiently and independent of the forecast scale. Application of a simple performance-based weighting method on the four models' DEMETER product with one forecast member per model performs as well as the full four models' DEMETER product with nine forecasts per model without performance weighting in the given forecast and evaluation setup. The efficiency of the different weighting methods at different forecast scales in a probabilistic forecasting setup has to be investigated in further studies.

Acknowledgments. The authors thank all of the groups involved in the DEMETER and PEPS projects, and especially M. Denhard and S. Trepte of DWD and the many people involved from ECMWF. Support from the LOEWE Biodiversity and Climate Centre is also acknowledged. We are grateful to two anonymous reviewers and T. Hamill for their fruitful comments.

REFERENCES

Ahrens, B., and A. Walser, 2008: Information-based skill scores for probabilistic forecasts. *Mon. Wea. Rev.,* **136,** 352–363.

Atger, F., 1999: The skill of ensemble prediction systems. *Mon. Wea. Rev.,* **127,** 1941–1953.

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

Ebert, E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.,* **129,** 2461–2480.

Ehrendorfer, M., 1997: Predicting the uncertainty of numerical weather forecasts: A review. *Meteor. Z.,* **6,** 147–183.

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

Hamill, T. M., 2007: Comments on "Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging." *Mon. Wea. Rev.,* **135,** 4226–4230.

Heizenreder, D., S. Trepte, and M. Denhard, 2006: SRNWP-PEPS: A regional multimodel ensemble in Europe. *The European Forecaster: Newsletter of the WGCEF,* No. 11, WGCEF, 29–35.

Jaun, S., B. Ahrens, A. Walser, T. Ewen, and C. Schär, 2008: A probabilistic view on the August 2005 floods in the upper Rhine catchment. *Nat. Hazards Earth Syst. Sci.,* **8,** 281–291.

Katz, R., and M. Ehrendorfer, 2006: Bayesian approach to decision making using ensemble weather forecasts. *Wea. Forecasting,* **21,** 220–231.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science,* **285,** 1548–1550, doi:10.1126/science.285.5433.1548.

Kumar, A., B. Jha, Q. Zhang, and L. Bounoua, 2007: A new methodology for estimating the unpredictable component of seasonal atmospheric variability. *J. Climate,* **20,** 3888–3901.

Lorenz, E., 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus,* **17,** 321–333.

Marrocu, M., and P. A. Chessa, 2008: A multimodel/multianalysis limited-area ensemble: Calibration issues. *Meteor. Appl.,* **15,** 171–179.

Marsigli, C., F. Boccanera, A. Montani, and T. Paccagnella, 2005: The COSMO–LEPS mesoscale ensemble system: Validation of the methodology and verification. *Nonlinear Processes Geophys.,* **12,** 527–536.

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

Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.,* **63,** 71–116.

——, and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872.

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

Sloughter, J. M., A. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 3209–3220.

Trenberth, K., 1997: The definition of El Niño. *Bull. Amer. Meteor. Soc.,* **78,** 2771–2777.

Uppala, S. M., and Coauthors, 2005: The ERA-40 re-analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multimodel combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Meteor. Soc.,* **134,** 241–260, doi:10.1002/qj.210.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 627 pp.

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