Clim Dyn (2014) 42:597–616
DOI 10.1007/s00382-013-1779-8

Multi-model calibration and combination of tropical seasonal sea surface temperature forecasts

Luis Ricardo Lage Rodrigues · Francisco Javier Doblas-Reyes · Caio Augusto dos Santos Coelho

Received: 3 September 2012 / Accepted: 19 April 2013 / Published online: 30 April 2013

Abstract Different combination methods based on multiple linear regression are explored to identify the conditions that lead to an improvement of seasonal forecast quality when individual operational dynamical systems and a statistical–empirical system are combined. A calibration of the post-processed output is included. The combination methods have been used to merge the ECMWF System 4, the NCEP CFSv2, the Météo-France System 3, and a simple statistical model based on SST lagged regression. The forecast quality was assessed from a deterministic and probabilistic point of view. SSTs averaged over three different tropical regions have been considered: the Niño3.4, the Subtropical Northern Atlantic and Western Tropical Indian SST indices. The forecast quality of these combinations is compared to the forecast quality of a simple multi-model (SMM) where all single models are equally weighted. The results show a large range of behaviours depending on the start date, target month and the index considered. Outperforming the SMM predictions is a difficult task for linear combination methods with the samples currently available in an operational context. The difficulty in the robust estimation of the weights due to the small samples available is one of the reasons that limit the potential benefit of the combination methods that assign unequal weights. However, these combination methods showed the capability to improve the forecast reliability and accuracy in a large proportion of cases. For example, the Forecast Assimilation method proved to be competitive against the SMM while the other combination methods outperformed the SMM when only a small number of forecast systems have skill. Therefore, the weighting does not outperform the SMM when the SMM is very skilful, but it reduces the risk of low skill situations that are found when several single forecast systems have a low skill.

Keywords Seasonal prediction · Calibration and combination · Probabilistic prediction · Forecast verification

L. R. L. Rodrigues (✉) · F. J. Doblas-Reyes
Institut Català de Ciències del Clima (IC3),
Doctor Trueta 203, 08005 Barcelona, Spain
e-mail: luis.rodrigues@ic3.cat

F. J. Doblas-Reyes
Institució Catalana de Recerca i Estudis Avançats (ICREA),
Passeig Lluís Companys 23, 08010 Barcelona, Spain

C. A. S. Coelho
Centro de Previsão de Tempo e Estudos Climáticos, Instituto Nacional de Pesquisas Espaciais (CPTEC/INPE), Rodovia Presidente Dutra Km 40, Cachoeira Paulista,
SP 12630-000, Brazil

1 Introduction

Due to the chaotic nature of the climate system and the inadequacy of current forecast systems, quantifying uncertainty plays an important role in climate forecasting (Palmer 2000). Dealing with uncertainty will help decision makers making better decisions on whether or not to take any action given a probability forecast for an event. The unavoidable uncertain character of weather and climate prediction forces climate forecasts to be formulated in a probabilistic way, as has been recognized for more than a century (Murphy and Winkler 1984). In addition, the probabilistic formulation requires an appropriate assessment of how reliable (i.e. whether the forecast uncertainty is accurate) the forecasts are (Slingo and Palmer 2011).

Two of the main sources of uncertainty in climate prediction are the lack of perfect knowledge of the initial

conditions of the climate system and the inability to perfectly model this system (Curry and Webster 2011; Knutti 2010; Slingo and Palmer 2011). The first source of uncertainty is usually addressed by generating a set of several independent forecasts with slightly different initial conditions using dynamical models, the so called ensemble technique (Gneiting and Raftery 2005; Palmer 2000). The ensemble technique does not take into account the model imperfections (e.g. model-specific biases, both in the mean state and in the internal variability), and for this reason, ensemble forecasts performed with an individual system are usually over-confident (i.e. under-dispersive) (Slingo and Palmer 2011). These limitations come from the simplification of the fluid dynamic equations required to solve them numerically, the limited spatial and temporal resolution of the models, which implies that some of the important climate variables are solved through parameterization, and the lack of perfect knowledge of all single aspects of the climate system physics (Curry and Webster 2011; Palmer 2000). Three techniques have been used to deal with the model inadequacy problem: the perturbed-parameter, the stochastic-physics, and the multi-model techniques (Doblas-Reyes et al. 2009 and references therein). The multi-model approach, which is the one that will be used in this paper, considers the combination of different forecast systems, independently designed from one another. The perturbed-parameter approach creates ensembles by perturbing uncertain parameters in the physical parameterizations of a single forecast system and the stochastic-physics method treats sub-grid scale physical processes in a probabilistic way adding extra terms to the model equations using simplified linear and nonlinear stochastic models.

When applying the multi-model approach, a question that immediately arises is to find the best way to combine the predictions made with the different forecast systems (Knutti 2010). It has been demonstrated that combining several dynamical forecast systems with equal weights (or simple multi-model) has, on average, improved deterministic and probabilistic forecast quality with respect to the single models (Doblas-Reyes et al. 2005; Hagedorn et al. 2005; Palmer et al. 2004; Tippett and Barnston 2008; Wang et al. 2009). Doblas-Reyes et al. (2005) explored several combination methods to merge several dynamical models setting different weights to each one based on their past performance and using different flavours of multiple linear regression. However, the small sample size typically available in climate prediction produces results with the combination methods that assign unequal weights that are not conclusive or robust, making the simple multi-model (SMM) a particularly successful benchmark (Doblas-Reyes et al. 2005). Other studies attempted to use more sophisticated combination methods and concluded that it is difficult to improve the SMM forecasts (DelSole et al. 2012; Kug et al. 2007, 2008; Tippett and Barnston 2008).

In a slightly different framework Coelho et al. (2004) used a Bayesian method to combine the European Centre for Medium-Range Weather Forecasts (ECMWF) dynamical model with a simple statistical model based on lagged regression to estimate calibrated probabilistic forecasts for the Niño3.4 index. Stephenson et al. (2005) generalized this method to deal with more than one model and more than one variable so that it could be used with several dynamical systems. They applied this Bayesian method to equatorial Pacific sea surface temperature (SST) grid point predictions produced by seven coupled forecast systems in the Development of a European Multi-model Ensemble System for Seasonal to Inter-Annual Prediction (DEMETER; Palmer et al. 2004) and showed improved forecast skill compared to individual forecast systems and the simple multi-model.

The predictability of the surface temperature and precipitation patterns, which plays an important role on human activities, is to a certain degree linked to our ability to predict the boundary conditions of the climate system such as the SST, especially in the tropics (Goddard et al. 2001; Shukla 1998). The El Niño Southern Oscillation (ENSO) is the most important source of predictability at seasonal timescale; therefore, the assessment of skill of ENSO SST predictions is a fundamental requirement for any seasonal forecasting system (Stockdale et al. 2011). It impacts the circulation and precipitation patterns in the Pacific Ocean itself and in several other regions around the globe (Ropelewski and Halpert 1987). Other tropical ocean basins such as the tropical SST over the Atlantic and Indian Oceans also have a major impact on the climate variability of the surrounding regions (Goddard et al. 2001). For instance, the SST anomalies over the tropical Atlantic region directly influence the position of the Intertropical Convergence Zone (ITCZ), which plays a role on the precipitation patterns over northern northeastern Brazil and western Africa, while the western Indian Ocean SST anomalies have impacts on the climate of eastern parts of the African continent. Another interesting feature is that the SST variability of the Atlantic and Indian basins is somehow linked to that of the tropical Pacific (Goddard et al. 2001). Therefore, an important tool used for operational seasonal predictions are the ocean climate indices that can be linked to major patterns of climate variability (Doblas-Reyes et al. 2013).

This study addresses several innovative aspects of climate forecasting. Firstly, it compares three different operational dynamical forecast systems: the ECMWF seasonal forecast system 4 (S4; Molteni et al. 2011), the National Centers for Environmental Prediction (NCEP) climate forecasting system version 2 (CFSv2; Saha et al. 2013) and the Météo-

France System 3 (MF3; Batté and Déqué 2011). These are some of the dynamical seasonal forecast systems available to the users of this type of climate information. A simple statistical model based on lagged regression (Coelho et al. 2004) is also used as an additional model in the combination procedure. Secondly, the study uses and compares several methods to combine the above single forecast systems in different ways: the multiple linear regression methods described in Doblas-Reyes et al. (2005) and the Bayesian method described in Stephenson et al. (2005). The simple multi-model, where the systems are put together with equal weighting, is used as a benchmark. The aim is to assess how the Bayesian method compares with the multiple linear regression methods and a simple multi-model. To the authors knowledge such a comprehensive comparison has not been performed so far. However, this study goes a bit farther than those two papers, and several others that were recently published (Hagedorn et al. 2005; Palmer et al. 2004; Tippett and Barnston 2008; Kug et al. 2007, 2008). In this paper the impact of those combination methods on a series of operational forecast systems, which is an aspect of the problem not dealt with in the past, was investigated. In particular, this implied considering the differences in how the systems are developed in a real-time basis, and how the combination affects predictions that are carried out regularly, with one start date per month. Finally, a comprehensive quality assessment of the climate forecasts both from a deterministic and probabilistic point of view is performed considering all possible start dates and lead time up to 7 months, which is the limit of the forecast time allowed by both S4 and MF3. Given the large amount of cases considered in this study, for simplicity the forecast quality assessment of both the combinations and the single forecast systems is carried out for SST averaged over three different tropical regions: the Niño3.4 SST index (170°W–120°W, 5°S–5°N), the Subtropical Northern Atlantic (SNA) SST index (55°W–15°W, 5°N–25°N), and the Western Tropical Indian ocean (WTI) SST index (50°E–70°E, 10°S–10°N).

In Sect. 2, the data sources are described, as well as the six combination methods used to combine the predictions from the four forecast systems and the scores employed to evaluate their quality. In Sect. 3 the forecast quality of the single forecast systems and their combination is described and the sources of improvement discussed. Finally, a summary and the main conclusions can be found in Sect. 4.

2 Data and methods

2.1 Seasonal predictions

Retrospective dynamical seasonal forecasts from three forecast systems based on coupled ocean–atmosphere models have been used in this study: S4 (Molteni et al. 2011), CFSv2 (Saha et al. 2013) and the MF3 (Batté and Déqué 2011). A simple statistical model based on lagged regression has been used as an empirical forecast system for both benchmarking and combination with the dynamical methods.

The atmospheric component of S4 is the cycle 36r4 of the ECMWF Integrated Forecast System (IFS) (Kim et al. 2012; Molteni et al. 2011). It has a horizontal resolution of about 80 km and 91 vertical levels, extending up to about 0.01 hPa. The ocean component of S4, the Nucleus for European Modelling of the Ocean (NEMO) version 3.0, has a horizontal resolution of about 1° with equatorial refinement and 42 vertical levels, 18 of which are in the upper 200 m. S4's hindcasts have 15 ensemble members, all starting in burst mode on the first day of every month at 0 UTC. Their integrations are 7-month long and cover the period 1981–2010.

CFSv2 uses the NCEP Global Forecast System (GFS), with horizontal resolution of about 100 km and 64 vertical levels, as its atmospheric component (Kim et al. 2012; Saha et al. 2013; Yuan et al. 2011). Its ocean component is the Geophysical Fluid Dynamics Laboratory Modular Ocean Model version 4 (MOM4) and it has maximum horizontal resolution of 0.25° within 10° of the equator and 0.5° poleward and 40 vertical levels. CFSv2 hindcasts have 24 ensemble members, except November that has 28 members. The hindcasts are initialized in different days and times, being the ones initialized after the day 7 used as the lead time zero ensemble members of the next month. For example, the ensemble members for the target month of February at lead time zero have start dates in January 11th, 16th, 21st, 26th, 31st, and the February 5th (at the synoptic times 00, 06, 12 and 18 UTC) of the same year. Their integrations are 10-month long and cover the period 1982–2010.

MF3 uses the Action de Recherche Petite Echelle Grande Echelle (ARPEGE) version 4 as its atmospheric component (Batté and Déqué 2011). It has a horizontal resolution of about 300 km and 91 vertical levels, reaching high into the stratosphere. Its ocean component ORCA, is the global version of the Océan PArallélisé (OPA) model version 8.2, has horizontal resolution of about 2° and 31 vertical levels. MF3's hindcasts have 11 ensemble members, all starting in burst mode on the first day of every month at 0 UTC. Their integrations are 7-month long and cover the period 1981–2010.

The statistical model used, which is probabilistic in nature because it predicts a distribution of solutions, follows the statistical model used in Coelho et al. (2004), except that here the model was trained in two different ways. On the one hand, a 1-year out cross-validation method was applied using the period 1951–2010. This is

referred to as the cross-validation mode. On the other hand, in the forecast-mode statistical model the period 1951–1981 was used as the training period and forecasts were performed for the target years 1982–2010, extending the training period by 1 year at a time as in an operational context (Mason and Baddour 2008; Mason and Mimmack 2002). As for the forecast quality assessments of the dynamical forecast systems, verification statistics were computed for the target period 1982–2010. A brief description of the differences in skill between the statistical model predictions in forecast and cross-validation modes is provided in "Appendix 1". The statistical model developed in forecast mode was used in Sect. 3.

2.2 Observations

The Hadley Centre's Global Sea-Ice Coverage and Sea Surface Temperature v1.1 dataset (HadISST; Rayner et al. 2003) was used to estimate the coefficients in the statistical model analysis and for the forecast quality assessment of all forecast systems. It contains a set of monthly fields of global SST and sea ice concentration on a 1° latitude and longitude grid from 1871 onwards.

2.3 Multi-model combination methods

Several methods (Table 1) were used to combine the four forecast systems (i.e. S4, CFSv2, MF3 and the statistical model). A description of the methods can be found in "Appendix 2" and a summary in Table 1.

- The first combination method, which is referred to as simple multi-model (SMM), consisted in pooling S4, CFSv2, MF3 and the statistical model together with equal weights attributed to each model. Therefore, the predicted mean of the SMM is the average of the predicted mean of all single models while its probabilistic prediction is the average of the probabilistic predictions of the four forecast systems.
- The second combination was built up by performing a least-square multiple linear regression (MLR) of the observations on the anomaly values of the four forecast systems. The predicted standard deviation, which assumes a Gaussian error distribution, was computed using Eq. 11.
- The forecast assimilation (FA) is a Bayesian method for calibrating and combining predictions from several sources with a prior (historical) empirical information (Stephenson et al. 2005). It has been used to combine the four forecast systems. In one case, the statistical model predictions obtained in forecast mode were combined with the three dynamical systems having a climatological forecast as the prior information. This method is referred to as the FA-climatology (FAC). Using the FA with the four forecast systems having the climatology as the prior information would give the same forecasts as the MLR combination described below (Stephenson et al. 2005).
- A fourth combination was performed by combining the three dynamical systems using the FA and using the statistical model predictions as the prior information. This combination is referred to as the FA-statistical (FAS) method.
- Because of the unavoidable co-linearity of the predictors due to the positive linear correlation between the predicted mean of the four forecast systems, the MLR combination could introduce a large uncertainty in the estimated linear regression coefficients (Doblas-Reyes et al. 2005). To avoid this, a principal component analysis has been performed on the four forecast systems to estimate a new set of predictors. The leading principal components (PC) were used as predictors in a simple linear regression with the observations. When using only the leading PC the fifth combination, the PC1 combination is obtained.
- A multiple linear regression of the observations on the four leading PCs is referred to as PCA combination. The predicted standard deviation of the PC1 and PCA are estimated as in the MLR combination.

Ideally the training period used to estimate the combinations should be independent to avoid artificial skill in the forecast quality assessment; however, this is difficult in seasonal forecasting given the short time series available. In cross-validation, this independence can only be achieved by using a fairly large window (Mason and Baddour 2008). Therefore, the anomalies of the forecast systems, computed by subtracting the predicted mean of each system from their climatological value, as well as the combinations described above were obtained in three-years out cross-validation mode. It is worth noting that for the Niño3.4 index in the CFSv2 two climatological values were computed, one prior and one after the year 1999. This was done to deal with the change in the SST forecast bias in equatorial Pacific due to changes in the ocean reanalysis from which the ocean initial conditions for hindcasts are taken (Kumar et al. 2012).

2.4 Forecast quality assessment

The forecast quality of the SST predictions of all combinations and individual forecast systems described above was assessed from a deterministic and a probabilistic point of view. The correlation coefficient was used to assess the degree of linear association between the predicted mean and the observed SST indices. Several probabilistic

Table 1 Summary of the combination methods described in the Sect. 2.3

Combination method	Name	Predicted value	Predicted standard deviation
Simple multi-model	SMM	$\widehat{y}_j^{SMM} = \frac{1}{M}\sum_{i=1}^{M} x_{j,i}$. $M = 4$, which represents the four forecast systems at the jth target year	
Multiple Linear Regression	MLR	$\widehat{y}_j^{MLR} = \sum_{i=1}^{M} x_{j,i}a_i^j + a_0^j$, where the indices a_1^j and a_0^j were estimated by regressing $\mathbf{y}^j$ on $\mathbf{X}^j$.$\mathbf{y}^j$ is the $(N-3)\times 1$ vector of predictands and $\mathbf{X}^j$ the $(N-3)\times 4$ matrix of forecast system predictors for all N samples available, expect for the $j-1$th, jth and $j+1$th target years. $M = 4$, which represents the four forecast systems	$\widehat{s}_j^{MLR} = s_0^j\sqrt{1+\frac{1}{N-3}\mathbf{x}_j\left(\mathbf{S}_{xx}^j\right)^{-1}\mathbf{x}_j^T}$, where s_0^j the standard deviation of the regression residuals and $N-3$ is the cross-validated sample size
Principal Component 1 regression	PC1	$\widehat{y}_j^{PC1} = p_{j,1}a_1^j + a_0^j$, where the indices a_1^j and a_0^j were estimated by regressing $\mathbf{y}^j$ on $\mathbf{P}^j$.$\mathbf{y}^j$ is the $(N-3)\times 1$ vector of predictands and $\mathbf{P}^j$ the $(N-3)\times 1$ vector of the leading principal component predictor for all N samples available expect for the $j-1$th, jth and $j+1$th target years. $M = 1$, which represents the leading PCs	$\widehat{s}_j^{PC1} = s_0^j\sqrt{1+\frac{1}{N-3}p_j\left(S_{pp}^j\right)^{-1}p_j^T}$, where s_0^j the standard deviation of the regression residuals and $N-3$ is the cross-validated sample size
Principal Component Analysis regression	PCA	$\widehat{y}_j^{PCA} = \sum_{i=1}^{M} p_{j,i}a_i^j + a_0^j$, where the indices a_i^j and a_0^j were estimated by regressing $\mathbf{y}^j$ on $\mathbf{P}^j$. $\mathbf{y}^j$ is the $(N-1)\times 1$ vector of predictands and $\mathbf{P}^j$ the $(N-1)\times 4$ matrix of four leading principal component predictors for all N samples available expect for the $j-1$th, jth and $j+1$th target years. $M = 4$, which represents the four PCs	$\widehat{s}_j^{PCA} = s_0^j\sqrt{1+\frac{1}{N-3}\mathbf{p}_j\left(S_{pp}^j\right)^{-1}\mathbf{p}_j^T}$, where s_0^j the standard deviation of the regression residuals and $N-3$ is the cross-validated sample size
Forecast Assimilation-climatology	FAC	$\widehat{y}_j^{FAC} = y_b^j + L^j\left[\mathbf{x}_j - \mathbf{G}^j\left(y_b^j - y_0^j\right)\right]$, where $y_b^j = \overline{y}^j$. $M = 4$, which represents the three dynamical systems and the statistical model	$\widehat{s}_j^{FAC} = \left(\left(\mathbf{G}^j\right)^T\left(\mathbf{S}^{2,j}\right)^{-1}\mathbf{G}^j + \left(\mathbf{C}^j\right)^{-1}\right)^{-1}$, where $\mathbf{C}^j = \mathbf{S}_{yy}^j$
Forecast Assimilation-statistical	FAS	$\widehat{y}_j^{FAS} = y_b^j + \mathbf{L}^j\left[x_j - \mathbf{G}^j\left(y_b^j - y_0^j\right)\right]$, where y_b^j is the predicted mean by the statistical model for the jth target year. See Coelho et al. (2004) for details. $M = 3$, which represents the three dynamical systems	$\widehat{s}_j^{FAS} = \left(\left(\mathbf{G}^j\right)^T\left(\mathbf{S}^{2,j}\right)^{-1}\mathbf{G}^j + \left(\mathbf{C}^j\right)^{-1}\right)^{-1}$, where $\mathbf{C}^j$ is the predicted standard deviation by the statistical model for the jth target year. See Coelho et al. (2004) for details

Details can be found in "Appendix 2"

verification scores have been computed for dichotomous events of SST anomalies exceeding the median and the upper quartile of the climatological distribution.

The main measure of probabilistic forecast quality is the Brier score (BS), which can be defined as:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \quad (1)$$

Where p_i is the probability forecast and o_i is the observation, which is set to be one if the event happened and 0 if it did not happen, for the ith year. The BS could be generalized in the form of a skill score where the forecast of a given system is compared to a reference prediction system, which is usually a much simpler forecast such as the climatological frequency of the event. This generalization is called the Brier skill score (BSS), and could be written as $BSS = 1 - \frac{BS}{BS_c}$, where BS is the Brier score of a given system and BS_c is the Brier score of the reference forecast. Positive BSS means the BS of the system is better than the BS of the reference forecast.

The median and the upper quartile of the climatological distribution were estimated using ensemble members for the predictions of the dynamical forecast systems and all available years. Separate threshold estimates were obtained for the predictions and the observations to take into account that the predictions have systematic errors in the variability. Two different types of hindcasts were handled in this study: ensemble predictions (S4, CFSv2 and MF3) and sets of predictions defined by a forecast mean and standard deviation (all other forecast systems). For those forecast systems that did not have ensemble hindcasts, the normal forecast distribution of each year was sampled with size 10,000 to obtain samples from which to compute the median and quartiles of the corresponding climatological distributions. The 10,000 sample size was chosen because

it was found to provide robust estimates of the climatological probability density function (PDF). The robustness was estimated by calculating the BSS 1,000 times for the statistical model and for a given target month and lead time pair. These 1,000 estimations were performed with sample size 11, 51, 100, 1,000 and 10,000. The sample size 10,000 was chosen because it presented the smallest spread in the histogram of the 1,000 estimated values of the BSS. Finally, the probability forecasts were estimated using the estimated thresholds (median and upper quartile).

Other forecast quality attributes have also been analyzed, among them the reliability and resolution components of the BS (Mason and Stephenson 2008). The reliability component of the BS verifies the degree of correspondence between the frequency of events predicted by the system and the frequency of events that actually happened and measures the degree of trustworthiness of the predicted probabilities. The resolution, on the other hand, measures the ability of the forecasts to distinguish events that have forecast probabilities different from the climatological frequency. A third component of the BS is the uncertainty, which is associated with the uncertainty of the observations for a given event and does not depend on the predictions.

These three components of the BS are estimated by stratifying the forecast probabilities into a set of bins, the number of which is usually smaller than the number of possible forecast probabilities. However, depending on the number of bins used to stratify the forecast probabilities, the sum of the three components does not equal the BS computed using Eq. (1). Two additional components that account for the within-bin variance of the forecasts and the within-bin covariance between forecasts and observations are also needed to make the components of the BS less sensitive to the number of bins (Stephenson et al. 2008). These two extra components were added to the resolution component of the BS to make a generalized resolution term. The skill scores of the reliability and generalized resolution were computed as follows (Doblas-Reyes et al. 2005):

$$BSS_{REL} = 1 - \frac{BS_{REL}}{BS_{UNC}} \quad (2)$$

$$BSS_{GRES} = \frac{BS_{GRES}}{BS_{UNC}} \quad (3)$$

where BS_{REL} is the reliability component of the BS, BS_{GRES} is the generalized component of the BS, and BS_{UNC} is the uncertainty component of the BS.

The statistical significance of the results is quantified with the p value, which was estimated using a nonparametric bootstrap method. The bootstrap procedure was used to resample the forecast-observation pairs randomly with replacement, keeping the forecast and observation pairs together (Mason 2008). This procedure was applied both to the statistic itself and the statistic difference between two forecast systems. Thus, a distribution of the statistic centered on the sample value of the statistic or the statistic difference was created, from where the p value was estimated. The sample size of the bootstrap was chosen to be 1,000. It was applied to the correlation coefficient, the BSS, and the reliability (BSSrel) and resolution (BSSgres) components of the BSS. The null hypothesis is that the statistic or the statistic difference is zero, while the alternative hypothesis is that the statistic is larger than zero for the skill score (i.e. one-tailed test) and different from zero for the skill score difference (i.e. two-tailed test).

3 Results

3.1 Niño3.4 index

Figure 1 shows monthly forecast anomalies of the Niño3.4 index for the four single forecast systems and the FAS for the period between 1982 and 2010. This illustration is for the target month of January and lead time 2 months. This means that the statistical model used the previous month of October of the previous year as the predictor, S4 and MF3 forecasts were started on the first of November while CFSv2 has its ensemble members started between the second week of October and the first week of November. The 95 % prediction interval for each forecast system, given by the predicted mean anomaly plus or minus 1.96 times the predicted standard deviation, and the mean climatology forecast are also displayed. For the dynamical forecast systems the predicted standard deviation is the standard deviation of all available ensembles. All forecast systems have a high linear correspondence with the observed anomalies. However, the FAS has higher correlation than all single forecast systems. Besides, most of the observations in the forecast systems, except for the MF3, fall inside the 95 % prediction interval meaning that these forecast systems are reliable. This is shown quantitatively by the large values of the reliability component (BSSrel) of the BSS of each forecast system that is displayed in the top left corner of each panel of Fig. 1. Note that even though many of the MF3 forecasts fall outside the prediction interval, it has a good reliability skill score. The FAS has the highest BSS of all forecast systems.

As operational systems have to provide a prediction starting at least once a month all year round, this analysis was extended to all months of the year and for the seven lead times available for S4 and MF3. The results for both

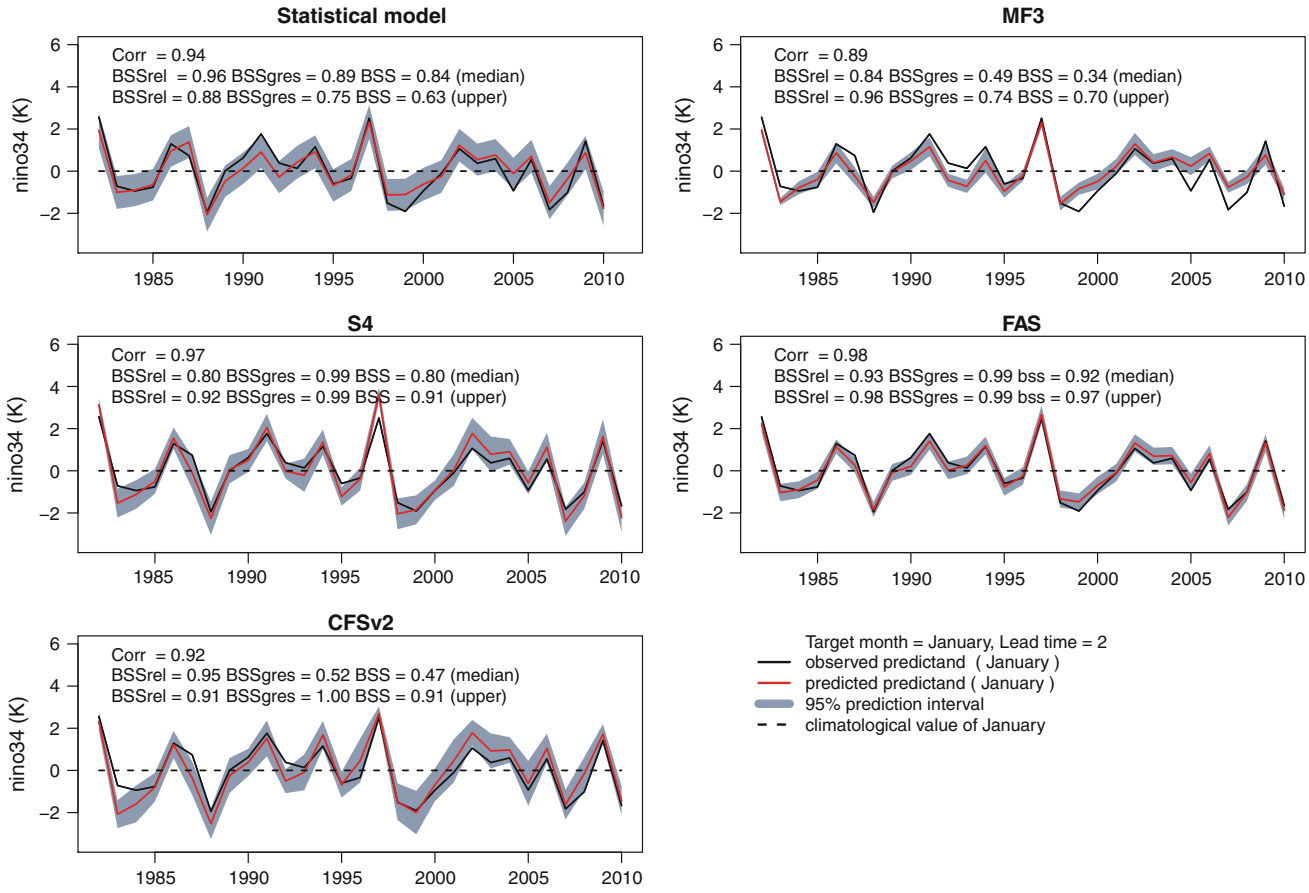

Fig. 1 Monthly forecast anomalies of Niño3.4 index for the statistical model, S4, CFSv2, MF3 and FAS. Forecasts are for the target month of January with lead time two. Observed values (*black solid line*), predicted values (*red solid line*), 95 % predicted interval (*grey area*) and the climatology value of January (*black dashed line*). Several scores are displayed in each panel: the correlation coefficient, and the Brier skill score and its reliability and resolution components for dichotomous events of SST anomalies exceeding the median and the upper quartile

deterministic and probabilistic scores are summarized in Figs. 2 and 3. Figure 2 shows the correlation coefficient of the Niño3.4 SST index mean prediction as a function of both target month and lead time for all forecast systems and combinations for the period between 1982 and 2010. The statistical model has the highest values of correlation for predictions produced during the boreal winter, when ENSO persistence is the strongest, followed by a period of decreasing skill for predictions produced during the boreal spring. This decrease in skill during the boreal spring is known as the spring barrier (e.g., Balmaseda et al. 1995; Goddard et al. 2001; Mason and Mimmack 2002; Stockdale et al. 2011). The lowest values of correlation were observed during the boreal summer for longer leads coinciding with the period of the year when ENSO typically changes from one phase to another.

A similar pattern is found for the S4 predictions, except that the correlation is higher, and the decrease of skill for predictions started during the boreal spring is much less important than in the statistical predictions. The superior performance of S3 over the older ECMWF forecast systems and persistence when predicting the Niño3.4 index has been shown previously (Stockdale et al. 2011). S4 and S3 have similar skill in terms of anomaly correlation when predicting the Niño3.4 index (Molteni et al. 2011). The CFSv2 predictions also show a less marked decrease in correlation across the spring barrier than the statistical model, but on average its skill is slightly lower than in S4.Kim et al. (2012) also found that S4 has higher correlation than CFSv2 when predicting the Niño3.4 index in the boreal winter with lead time 1 month; however, they did not apply the CFSv2 bias correction suggested by Kumar et al. (2012). Here we show that S4 outperforms CFSv2 even after applying the bias correction suggested by Kumar et al. (2012). On the other hand, MF3 predictions are less skilful than the other dynamical forecast systems and also have a decrease in correlation during the boreal spring although its correlation does not turn into negative correlation as in the statistical model. Similar results were found for the NCEP CFS version 1 (CFSv1) and a persistence

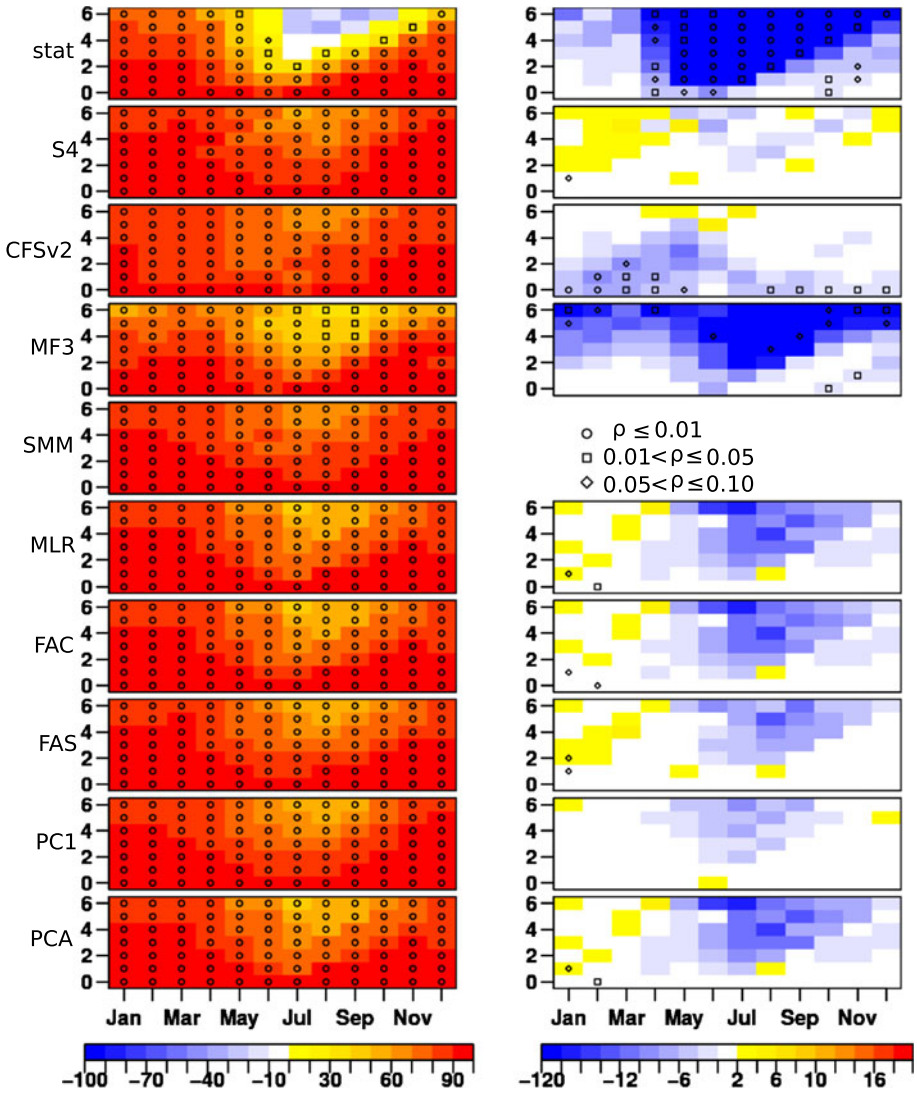

Fig. 2 (*Left column*) Correlation between the ensemble-mean predicted and observed Niño3.4 index as a function of target month (*horizontal axis*) and lead time (*vertical axis*) for the different forecast systems. (*Right column*) Correlation difference between each forecast system and the SMM. The predictions have been formulated over the period 1982–2010. The forecast systems used are, from top to bottom the statistical model, S4, CFSv2, MF3, SMM, MLR, FAC, FAS, PC1 and PCA. HadISST data are used to estimate the coefficients in the statistical model and for the forecast quality assessment. Circles are for p-values smaller than or equal 0.01, squares for p values between 0.05 and 0.01, and diamonds for p values between 0.10 and 0.05

model in previous studies (Saha et al. 2006; Sooraj et al. 2012). It is worth noting that ENSO skill may have a decadal dependence, i.e. skill may depend on the period of verification (Balmaseda et al. 1995). That is not addressed here. Moreover, persistence forecasts, though outperformed by more sophisticated statistical models of ENSO, are a tough standard to beat mainly when predicting short lead times (Goddard et al. 2001; Mason and Mimmack 2002). In any case, none of the three dynamical systems as well as none of the combinations, show any negative correlation as the statistical model does in boreal summer.

The SMM, which is used as the reference standard in the comparison with all the other forecast systems, has higher correlation than the statistical model, CFSv2 and MF3 more often than not (right panel of Fig. 2). On the other hand, the SMM has higher correlation than S4 only at longer leads in the boreal summer and fall. As discussed in previous studies (Hagedorn et al. 2005) the SMM could be outperformed by the best single forecast system on some of the aspects of the prediction (here the Niño3.4 index in the boreal winter). On the other hand, as it will be shown in the following sections, the SMM has an overall better performance than the four single forecast systems when all aspects of the prediction (i.e. the three analyzed regions, all target month and lead time pairs) are taken into account.

All combinations show a similar skill pattern. They outperform the SMM only in a few target month and lead time pairs especially during the boreal winter. On the other hand, they have lower correlation than the SMM in the other months of the year, especially for leads longer than 2 months. The forecast quality of the SMM is usually difficult to improve using multiple linear regression because of the small number of single forecast systems and short time series used to estimate the regression coefficients (Doblas-Reyes et al. 2005). This could also help explaining the similarities between the MLR and PCA predictions. S4 has the best overall correlation for the Niño3.4 predictions, that is, it has higher correlation than

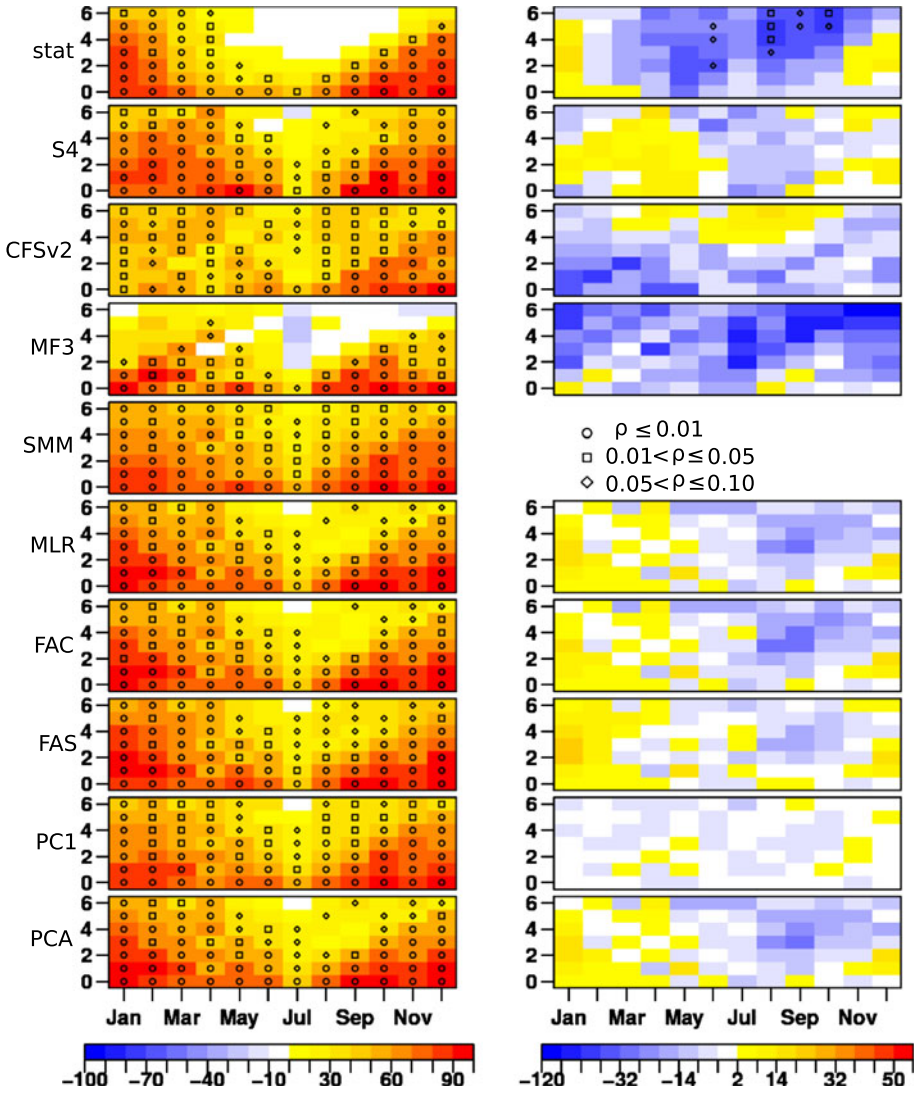

Fig. 3 As Fig. 2, but for the BSS of the Niño3.4 SST index anomalies exceeding the median

all the other single forecast systems and combinations more often than not.

Because of the inherent uncertainty involved in climate forecasting (Mason and Mimmack 2002) the quality of the probabilistic forecasts were also assessed and will be described below. The BSS with respect to climatology for the SST anomalies exceeding the median for the Niño3.4 index is shown in Fig. 3. The patterns of skill are similar to those of the correlation coefficient, except for the smaller magnitude of the values. Wang et al. (2009) also found that probabilistic forecast skill scores display similar patterns as those of deterministic scores. Positive BSS for most of the target month and lead time pairs in all single forecast systems and their combinations show that predictions are more skillful than the climatology. As for the mean predictions, the probabilistic predictions for this index also show the lowest skill at the target period of the boreal summer, especially for longer leads (i.e. for predictions with start dates in boreal spring). This agrees with previous studies (Tippett and Barnston 2008).

The differences between the BSS of the single forecast systems and the combinations with the SMM have all a pattern similar to that of the correlation shown in Fig. 2. One difference is that the SMM beats all single forecast systems, including S4, more often than not. On the other hand, the other combinations are more competitive when assessing their probabilistic skill in comparison with their deterministic counterpart although the SMM performs better than all of them more often than not. The only exception to this is given by the FAS predictions. The only season where the unequal combinations beat the SMM more often than not is the boreal winter. This is achieved by improved resolution skill score, a highly desirable feature that shows that unequal weighting can also improve the accuracy of the predictions and not just the reliability. This result provides evidence that unequal combination can

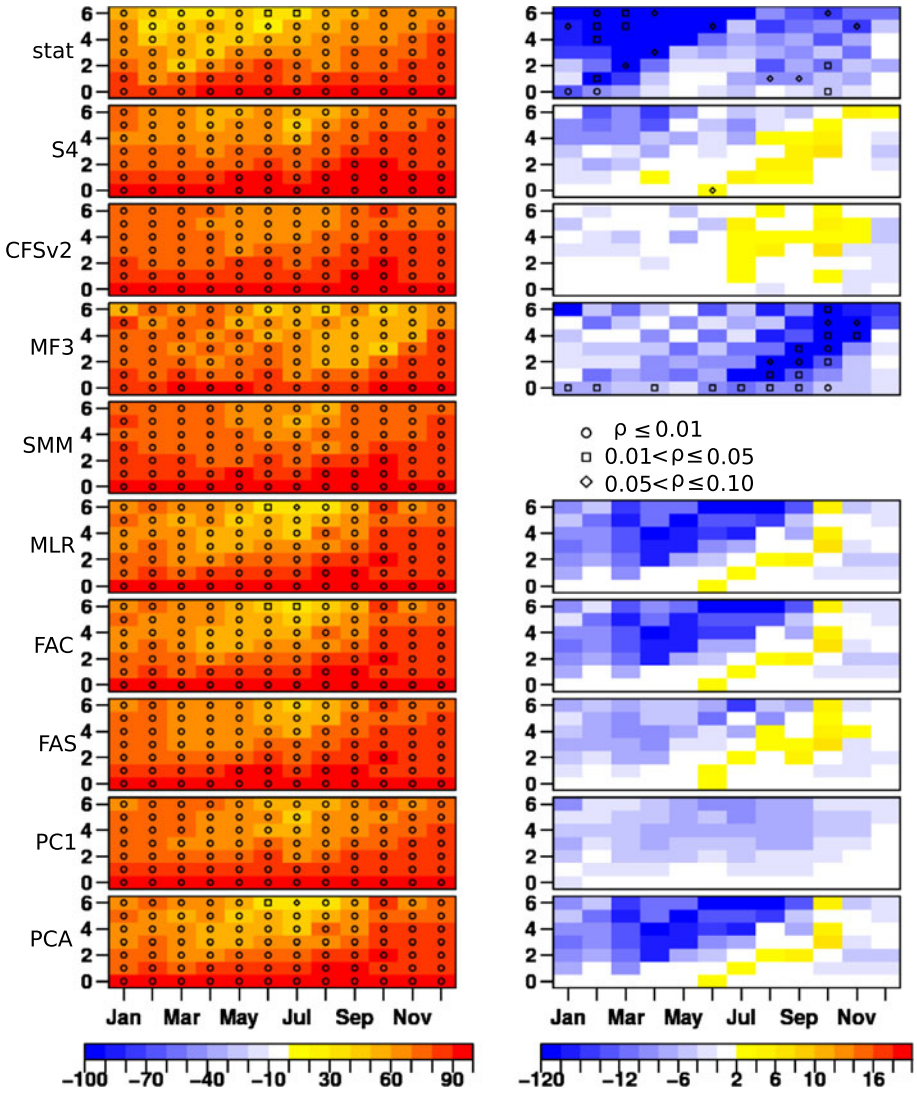

Fig. 4 Same as Fig. 2, but for the correlation coefficient of the SNA SST index anomalies

indeed improve predictions over that of the SMM even with the limited sample size typical of seasonal forecasting.

The BSS for the event defined as the anomalies of the Niño3.4 SST index exceeding the upper quartile were also analyzed (not shown). The BSS has similar patterns to those of the SST anomalies exceeding the median shown in Fig. 3. One difference is that the SMM is more difficult to beat when predicting the SST anomalies exceeding the upper quartile.

3.2 Subtropical North Atlantic index

The correlation coefficient of the predictions of the SNA SST index for all forecast systems and combinations show positive values in all target month and lead time pairs (Fig. 4). This is also observed in the Niño3.4 index predictions, except for the statistical model where negative correlations are observed in the boreal summer and beginning of fall at leads four, five and six. This confirms that there is considerable SST memory in these two ocean regions for the lead times considered here. For all forecast systems and combinations the correlation coefficient is higher in the Niño3.4 index than in the SNA index more often than not. All these findings agree with previous studies, although they used slightly different areas to represent the tropical northern Atlantic SST region (Sooraj et al. 2012; Stockdale et al. 2011).

The skill of the SNA index varies seasonally. All forecast systems reach a maximum peak in correlation in December. After the peak the correlation starts decreasing, reaching relatively lower values of correlation during boreal spring. The SMM has higher correlation than all forecast systems and combinations during all seasons more often than not, except for the S4 and CFSv2 in the boreal summer and fall (right panel of Fig. 4). This shows how difficult it is to improve the SMM ensemble-mean predictions in all cases using more sophisticated combination methods that assign unequal weights to each forecast system (DelSole et al. 2012; Doblas-Reyes et al. 2005).

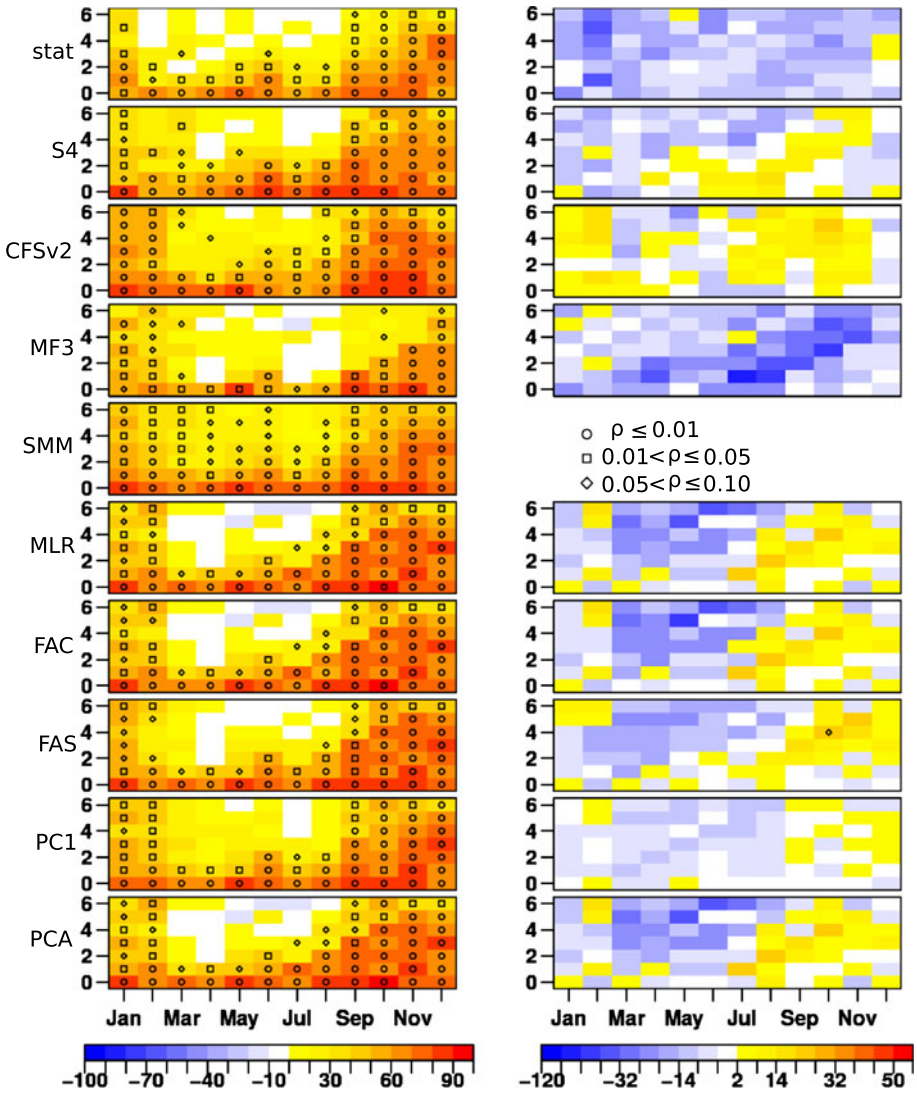

Fig. 5 As Fig. 2, but for the BSS of the SNA SST index anomalies exceeding the median

The BSS of the SNA SST index anomalies exceeding the median have similar patterns as the correlation counterpart, except for the lower magnitude (Fig. 5). The SMM has higher BSS than the statistical model, S4 and MF3 more often than not. CFSv2 beats the SMM in terms of BSS more frequently than not, but S4 beats the SMM only during the boreal summer and fall in some leads. It is important to note that S4 shows noticeable improvements in skill when compared to S3 and persistence over the tropical Atlantic region (Molteni et al. 2011). The SMM outperforms all combinations methods more often than not; however, it is observed that the unequal combinations do a good job during the target months between August and November. The decomposition of the BSS shows that the resolution skill score term explain most of the pattern of the BSS in all systems, that is, whenever the SMM has higher (smaller) BSS than a single forecast system or combination it also performs better (worse) in terms of BSSres. All forecast systems and combinations have similar BSSrel, except for the SMM that performs better in a few target month and lead time pairs.

The BSS of the SNA SST index anomalies exceeding the upper quartile shows similar patterns as the ones in Fig. 5, except that they are smaller in magnitude (not shown). As for the Niño3.4 index, predictions of the SNA SST index anomalies exceeding the upper quartile are less skillful than when predicting the event of exceeding the median. In addition, in agreement with the results discussed above, the SMM has higher BSS than all forecast system more often than not. For the Niño3.4 and SNA indices it is more difficult to improve SMM forecasts in terms of BSS for more extreme events, such as the ones above the upper quartile, than for events above the median.

3.3 Western Tropical Indian index

All forecast systems show positive correlation for the WTI index predictions in almost all target month and lead time

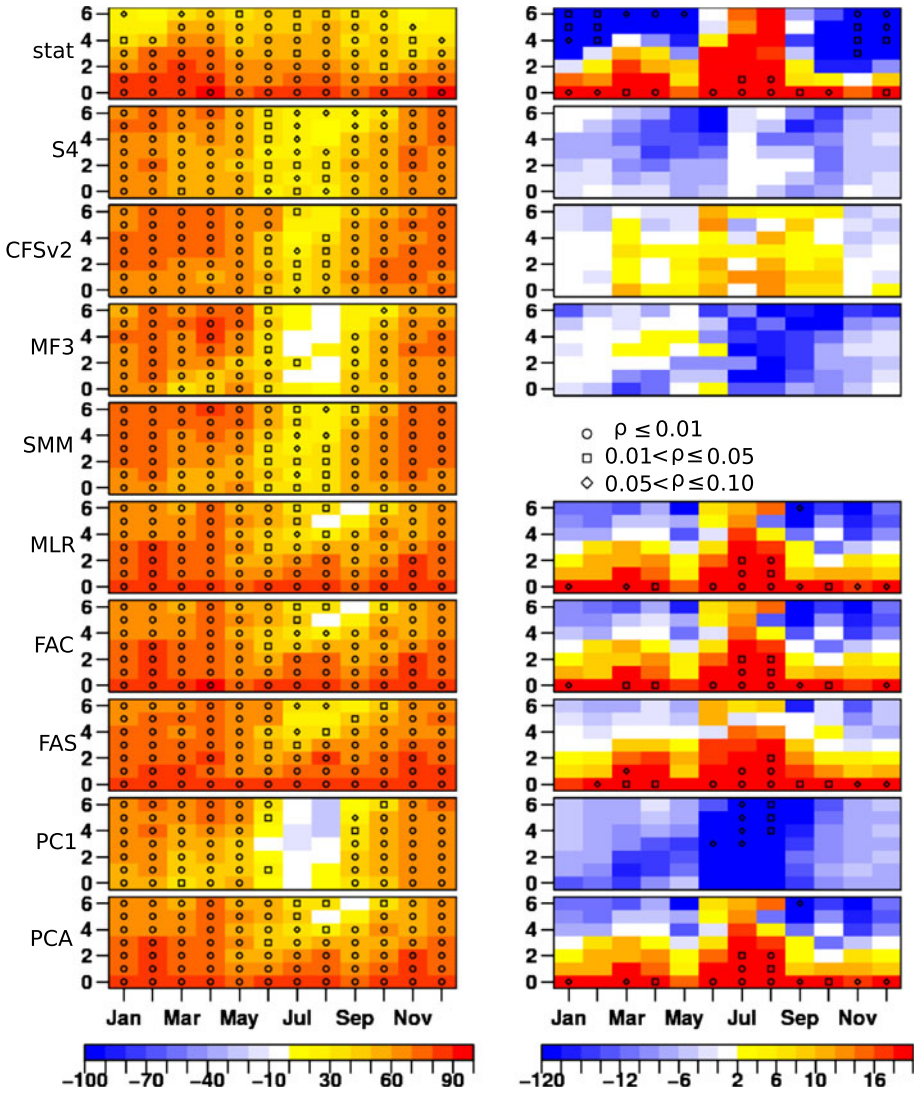

Fig. 6 As Fig. 2, but for the correlation coefficient of the WTI SST index anomalies

pairs (Fig. 6). This shows both that there is considerable SST memory in the region and that the three dynamical forecast systems analyzed here are able to reproduce well the inter-annual SST variability in the Western Indian Ocean. The predictability of the WTI index also varies seasonally, but unlike the two indices described above the skill of the statistical and the three dynamical forecast systems vary differently. The three dynamical forecast systems have the highest correlation during the target months between November and May, and a significant drop in skill in the target months of the boreal summer, especially for longer lead times. This rapid decrease in correlation was observed previously in S3 (Stockdale et al. 2011), CFSv1 (Sooraj et al. 2012), the Climate Prediction and its Application to Society (CliPAS) and the DEMETER multi-model (Wang et al. 2009). On the other hand, the statistical model has two peaks in correlation, one in the boreal spring and another one in the boreal summer. Wang et al. (2009) showed that while the SST predictions in the WTI and East Tropical Indian (ETI; 90°–110°E, 10°S-Equator) have some useful skill both in dynamical and statistical forecast systems and their combinations, the skill for the Indian Ocean Dipole SST index (SST at ETI minus SST at WTI; Saji et al. 1999), which has influence in the surround continental regions, is reduced.

The statistical model has higher correlation than the SMM at the first two lead times in all target months and during the boreal summer also at longer leads when the three dynamical forecast systems have relatively lower skill. CFSv2 is the only dynamical forecast system that outperforms the SMM more often than not. On the other hand, S4 and MF3 have systematically lower correlation than the SMM. All combinations, except for the PC1, outperform the SMM more often than not and this coincides with the target month and lead time pairs where the statistical model performs well. This shows that in some situations the combination methods that assign unequal weights can in fact lead to improvement in skill over that of

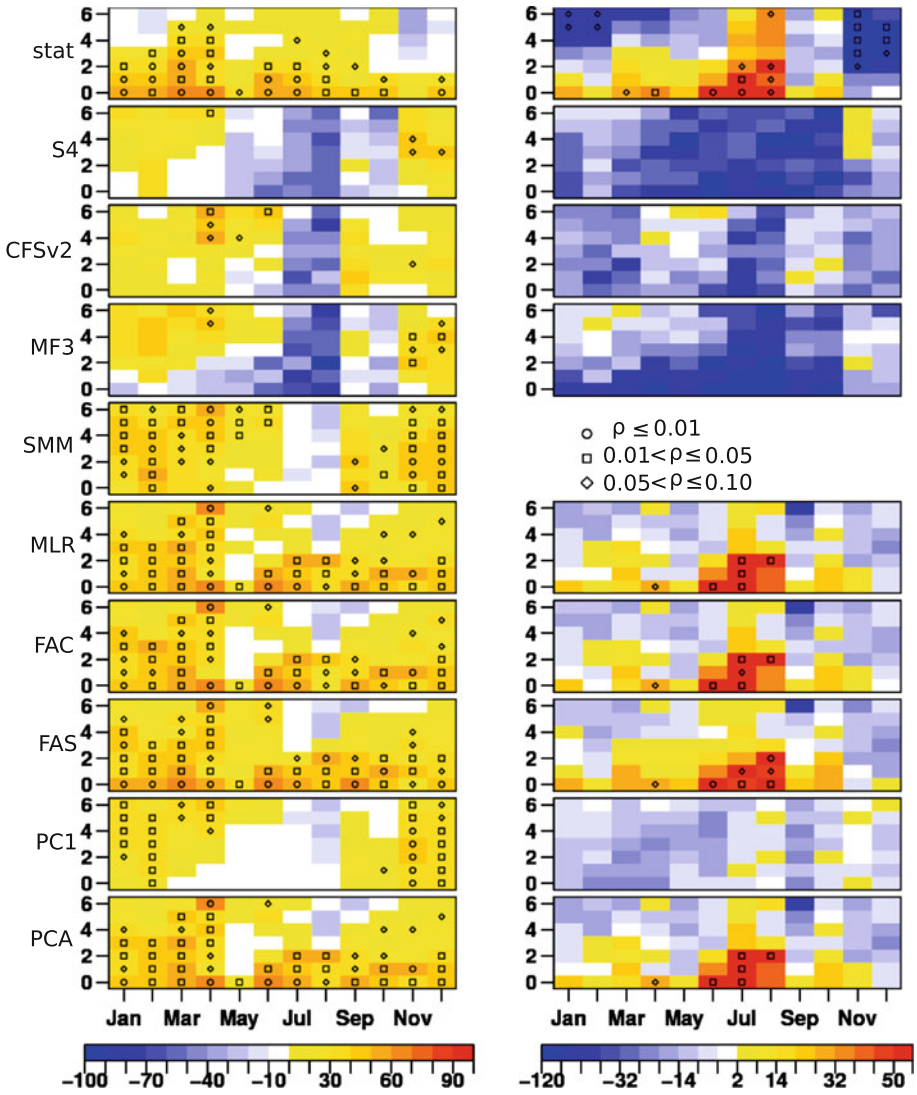

Fig. 7 As Fig. 2, but for the BSS of the WTI SST index anomalies exceeding the median

the SMM. In contrast to the Niño3.4 and SNA analyses, the inclusion of the statistical model information in the WTI index adds skill to the combinations.

The BSS of the WTI SST index anomalies exceeding the median are shown in Fig. 7. Except for the statistical model during the boreal summer, when the dynamical forecast systems perform worse than climatology, the SMM outperforms all forecast systems more often than not. The statistical model has higher BSS than the SMM during the target months of July and August at all lead times and also during the first two leads in the first six target months of the year. For the other target month and lead time pairs the SMM has higher BSS than the statistical model more often than not. The SMM has systematically higher BSS than all dynamical forecast systems in all seasons of the year and lead times. Among all forecast systems and combinations the FAS is the only one that has higher BSS than the SMM more often than not. On the other hand, the PC1 is the only combination that has systematically lower BSS than the SMM while the other combinations have higher BSS than the SMM at the target month and lead time pairs when the statistical model performs well. The decomposition of the BSS shows that the combination methods that assign unequal weights have higher reliability skill score than the SMM more often than not, but they only perform better than the SMM in terms of resolution skill score during the boreal summer (not shown).

The BSS of the WTI SST index anomalies exceeding the upper quartile shows similar results as those of Fig. 7. However, all dynamical forecast systems perform worse than the climatology more often than when predicting the same index exceeding the median. The boreal summer is the most difficult season to improve the climatological probability forecasts and the statistical model is the only single forecast system that has skill. All combinations also have skilful probabilistic predictions when predicting the WTI index anomalies exceeding the upper quartile. Besides, the WTI is the only index when it is easier to

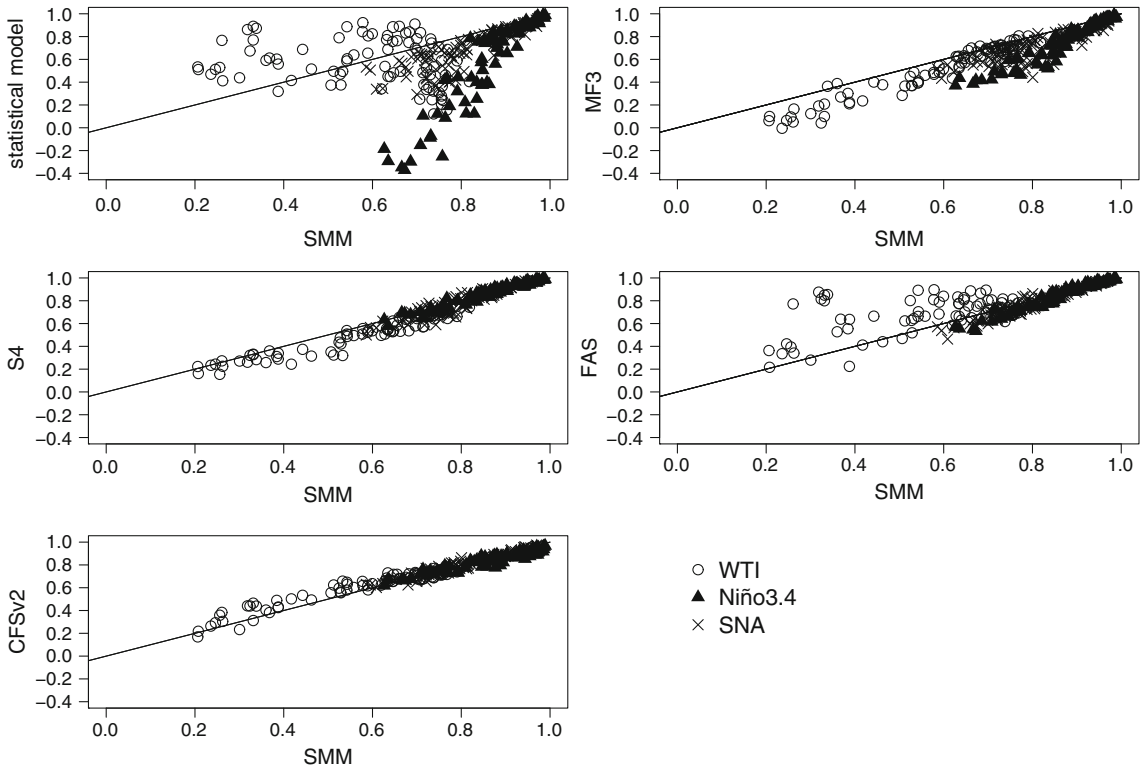

Fig. 8 Scatterplots of the correlation coefficient for the statistical model, S4, CFSv2, MF3 and FAS versus the SMM. Results are for twelve target months, seven lead times and three indices. Each *symbol* represents the correlation for one index: WTI (*circle*), Niño3.4 (*triangle*) and SNA (*cross*)

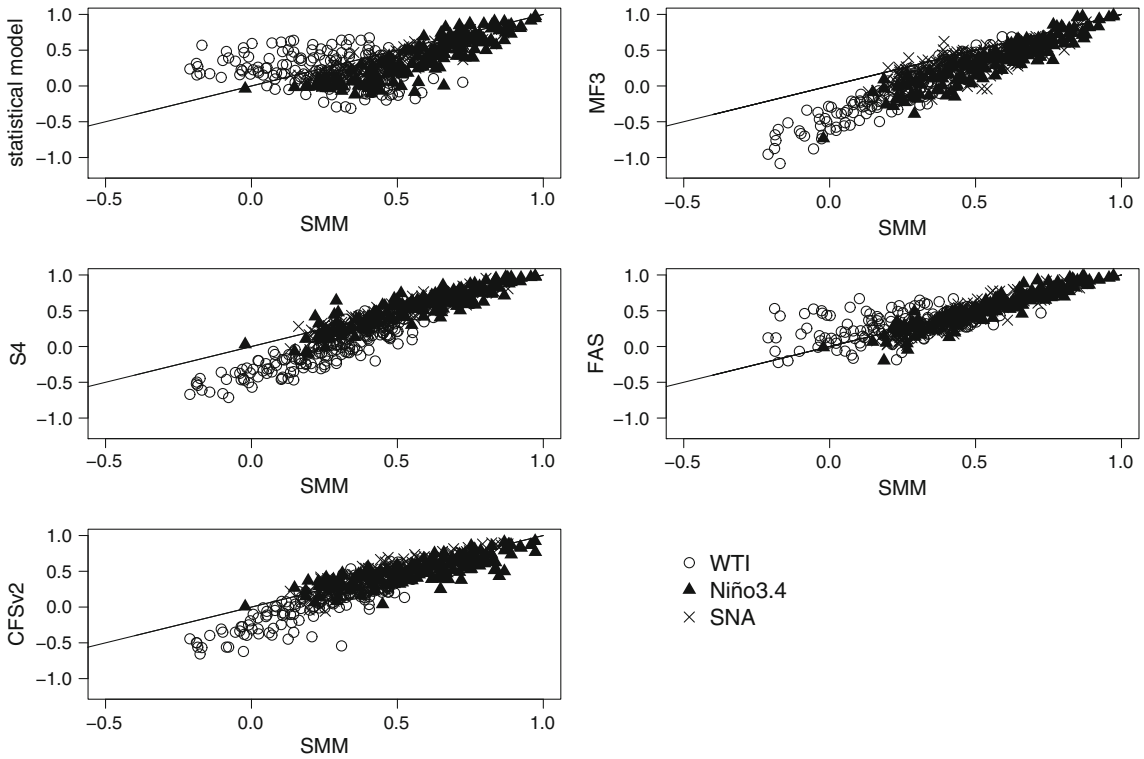

Fig. 9 As Fig. 8, but for the BSS. Results also include two events: anomalies above the median and above the upper quartile

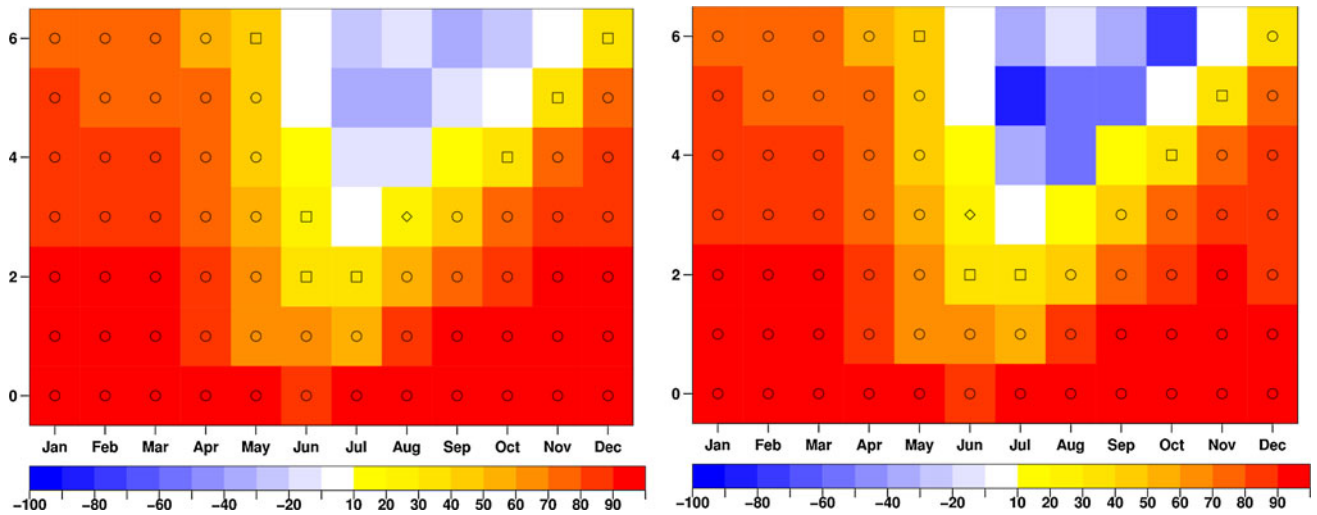

Fig. 10 Correlation between the predicted and observed Niño3.4 index as a function of target month (*horizontal axis*) and lead time (*vertical axis*) for the statistical model trained in forecast mode (*left column*) and in cross-validation mode (*right column*). Predictions have been formulated over the period 1982–2010. HadISST data are used to estimate the coefficients in the statistical model and for the forecast quality assessment. The symbols are for the p values (see text for details). Circles are for p values smaller than or equal 0.01, squares for p values between 0.05 and 0.01, and diamonds for p values between 0.10 and 0.05

outperform the SMM when predicting events above the upper quartile.

4 Discussion

The previous results give a detailed account of the different performance of the single forecast systems and the impact of the combination methods. It was seen that there is no single forecast system that provides the best results for all cases. In fact, while one system is better for Niño3.4 (S4), a different one is the best overall for the WTI (statistical). Surprisingly, simple empirical models can still provide useful predictive information, even when compared to the recently developed state-of-the-art dynamical forecast systems. As it is impossible to choose a single system to provide climate information to the users, an approach to integrate the different sources into a single prediction is necessary. These combination methods have different properties for specific target month and lead time pairs, not only because they combine a different set of single systems, but also because they calibrate the probabilistic predictions differently, as it has been found in the analysis of the reliability.

A more integrated view of the advantages of the set of combination methods considered is required. The scatterplots of the correlation (Fig. 8) and the BSS (Fig. 9) summarize the results described above. Predictions for all indices, target months, lead times and events of the probabilistic forecasts have been included to obtain a general picture of the performance. Each symbol in the scatterplots represents one of the three analyzed regions: WTI (circle), Niño3.4 (triangle) and SNA (cross).

The SMM has higher correlation than all single forecast systems and combinations, except for the FAS, more often than not (not shown). This is seen in Fig. 8 by the number of symbols that fall below the diagonal more frequently than above it in all forecast systems and combinations, except in the FAS. However, this superiority of the SMM over the single forecast systems is not found in all single aspect of the forecast as noted previously in this study and, for instance, in Hagedorn et al. (2005). As mentioned above, if only the Niño3.4 index is considered then S4 ensemble-mean predictions are more skilful than the SMM (Fig. 2). Moreover, if only the target months of July and August for the predictions of the WTI index are considered then the statistical model and the CFSv2 are more skilful than the SMM.

It is interesting to note that the SMM fails to beat all forecast systems when it has correlation smaller than 0.6 (Fig. 8). In these cases, all forecast systems and combinations, except S4, MF3 and PC1, have higher correlation than the SMM. The combination methods almost never show a negative correlation. This result is significant because it illustrates an important property of appropriate weighting methods: they reduce the risk of providing poor predictions in cases where the single forecast systems have low or negative correlation.

The scatterplots of the BSS show that the SMM has higher skill than the four single forecast systems more frequently than not (Fig. 9). In agreement with previous studies (Hagedorn et al. 2005) the SMM probabilistic

predictions do have an overall improved reliability and resolution when compared to the single forecasting systems (not shown). The statistical model hardly gets values of BSS below −0.4, while the dynamical forecast systems do worse for low values (Fig. 9). This could be explained because the statistical models are calibrated by construction (Mason and Baddour 2008) while the three dynamical systems are not and tend to be overconfident (Slingo and Palmer 2011). None of the weighting methods tends to show higher BSS than the SMM, but when the SMM has a low BSS the weighted predictions tend to be better.

5 Summary and conclusions

The chaotic nature of the climate system implies that forecast uncertainty must be quantified (Palmer 2000). The uncertainties in climate forecasts are due to both the initial conditions and model inadequacy (Slingo and Palmer 2011). The first source of uncertainty is addressed by generating an ensemble of forecasts, while model inadequacy has been addressed in this paper using the multi-model method (Doblas-Reyes et al. 2009). It has been shown that the quantification of these two sources of uncertainty leads to more reliable probabilistic forecasts (Coelho et al. 2004; Doblas-Reyes et al. 2005, 2009; Hagedorn et al. 2005; Stephenson et al. 2005; Wang et al. 2009).

Traditionally, multi-model predictions are built by merging the different single systems, avoiding the question of how best to combine them depending on their past performance. Hence, the question of whether there exists an optimal way to combine the different forecast systems remains unanswered. In this study the Bayesian method described in Stephenson et al. (2005) is compared with the multiple linear regression methods described in Doblas-Reyes et al. (2005) and a simple combination method where all forecast systems are combined with equal weight attributed to each of them. However, this study goes a bit farther than those two papers, and several others that were recently published (Hagedorn et al. 2005; Palmer et al. 2004; Tippett and Barnston 2008; Kug et al. 2007, 2008). In this paper the impact of those combination methods on a series of operational forecast systems, which is an aspect of the problem not dealt with in the past, was investigated. In particular, this implied considering the differences in how the systems are developed in a real-time basis, and how the combination affects predictions that are carried out regularly, with one start date per month. Three operational dynamical seasonal forecast systems were used: the ECMWF System 4, the NCEP CFSv2 and the Météo-France System 3. The statistical model described in Coelho et al. (2004) is used as an additional model for both benchmarking and to increase the number of systems in the combination procedure.

The predictability of the climate system is to a large extent linked to our ability to predict its boundary conditions, such as the SST. Therefore, the forecast quality assessment of the SSTs for deterministic and probabilistic predictions of all forecast systems and combinations was analyzed. Given the large amount of cases considered in this study, for simplicity the forecast quality assessment is carried out for SST averaged over three different tropical regions: the tropical Pacific Ocean, the tropical Atlantic Ocean and the tropical Indian Ocean.

The SMM, which is used as the reference forecast, has often higher correlation than all single forecast systems. However, for a specific aspect of the forecast the SMM can be outperformed by the best single forecast system as noted in earlier studies (Hagedorn et al. 2005; Wang et al. 2009). For instance, S4 has higher correlation than the SMM more often than not when only the Niño3.4 predictions are considered (Fig. 2). On the other hand, the statistical model has higher correlation than the SMM more often than not when only the WTI predictions in the boreal summer are considered (Fig. 6). This shows that empirical systems can still provide useful information. The SMM has also higher mean prediction correlation than all combination methods that assign unequal weights, except for the FAS, more often than not, making it a benchmark difficult to beat. Even the Forecast Assimilation method, which includes and generalizes previous calibration methods and had been proved to be competitive against the SMM when predicting the equatorial Pacific SST for the four target months available in the DEMETER project (Stephenson et al. 2005), performed only as good as the SMM in terms of correlation when all cases are considered (all SST indices, target month and lead time pairs). This could be explained by the low number of forecast systems (four) and short samples (29 years) used to estimate the regression coefficients (Doblas-Reyes et al. 2005).

The SMM outperforms the four single forecast systems analyzed here more often than not also in terms of BSS. It has been found that the higher BSS of the SMM predictions when compared to the single forecast systems is the result of improved reliability and resolution, which is in agreement with previous studies despite the slightly different definition of reliability and resolution used. The probabilistic predictions of the SMM are often better than those of the combination methods that assign unequal weights in terms of BSS. However, some of the results shown here give light to further research on how to improve the SMM predictions using combination methods that assign unequal weights. For example, FAS deterministic and probabilistic predictions are often competitive against the SMM, the combination methods that assign unequal weights improve

the SMM predictions when only a fraction of all single forecast systems have skill as in the case of the SNA index predictions in the boreal fall (Fig. 5) or in the case WTI index predictions in the boreal summer (Fig. 7). Therefore, the weighting does not outperform the SMM when the SMM is very skilful, but it reduces the risk of low skill situations that are found when several single forecast systems have a low skill.

Many questions concerning the calibration and combination procedure still remain to be addressed in future studies. For instance, how the inclusion of other dynamical forecast systems will affect the above results, the way the combination methods will behave if applied to less predictable regions such as the extratropics or to other variables such as air temperature or precipitation, or how the methods described above will behave in a combination of spatial fields instead of time series. In all cases, these questions and their solutions will have to take into account the requirements of the users of the climate information.

Acknowledgments LRLR would like to thank Virginie Guémas, Javier Garcia-Serrano, Salvador Pueyo (IC3, Spain) for useful discussions, and the National Center for Atmospheric Research (NCAR) Advanced Study Program (ASP) to support his attendance to the 2010 ASP summer colloquium "Forecast Verification in the Atmospheric Sciences and Beyond". CASC was supported by Conselho Nacional de Desenvolvimento Científico e Tencológico (CNPq) process 306664/2010-0. This study was supported by the Spanish MINECO-funded RUCSS project (CGL2010-20657), the European Union's FP7-funded QWeCI (ENV-2009-1-243964) and SPECS (ENV-3038378), and the Catalan Government. The authors would also like to thank the two anonymous reviewers for their useful comments.

Appendix 1

Figure 10 shows the correlation with the observations of the hindcasts produced by the statistical model for the Niño3.4 index in forecast and cross-validation modes for the period between 1982 and 2010. Similar patterns are found, including the decrease in correlation (blue squares) for 4–6 month lead predictions produced early in the year, particularly during the northern hemisphere spring. This feature is known as the spring barrier (Balmaseda et al. 1995). The correlation is also very similar. The main differences are found in the boreal summer (June–July–August) for leads longer than 4 months. In these cases, the statistical model in forecast mode performs better (i.e. has higher correlation) than the statistical model trained in cross-validation mode, which justified the use of the former approach in the results shown in this paper. It has been shown in previous studies that the predictive skill estimation methods using simple or multiple regression in cross-validation mode could introduce negative bias in negative correlations (Barnston and Van den Dool 1993).

Appendix 2

This appendix presents some of the equations used for the combination methods described in Sect. 2.3. The vector of predictands can be written as:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

where y_j is the predictand at the jth target year for a given target month and N is the total number of years. The vector of predictands corresponds to the time series of the predicted SST index. For example, y_j could be an observed Niño3.4 SST index in June 2000. The matrix of predictors can be written as:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,M} \end{bmatrix}$$

where $x_{j,i}$ is the predictor for the ith forecast system and jth target year for a given target month and lead time and M is the number of forecast systems. For example, $x_{j,i}$ could be an ensemble-mean prediction of the Niño3.4 SST index by S4 for the target month of June and target year 2000 with lead time one. Note that both predictands and predictors refer to anomalies computed using the 3 year-out cross-validation method, that is, for each case the mean was computed using all years, but the $j-1$th, jth and $j+1$th target years.

(a) *Simple multi-model*

The SMM was computed by applying equal weights to all models as follows:

$$\hat{y}_j^{SMM} = \frac{1}{M}\sum_{i=1}^{M} x_{j,i} \tag{4}$$

where $\hat{y}_j^{SMM}$ is the ensemble-mean SMM prediction at the jth target year.

(b) *Multiple linear regression (MLR)*

A multiple linear regression of the observations on the anomaly values of the four forecast systems was performed to estimate the linear combination of the different forecast systems. The multiple linear regression can be expressed in matrix form as follows:

$$\mathbf{y} = \mathbf{M}\mathbf{a} + \varepsilon \tag{5}$$

where

$$\mathbf{M} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \cdots & x_{N,M} \end{bmatrix}$$

and

$$\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_M \end{bmatrix}$$

where a_0 and a_i are the least-squares estimates of the intercept and the slope parameters of the ith forecast system and ε is the vector of residuals. Note that $\mathbf{M}$ is an extension of the matrix of predictors.

The least-squares estimate of $\mathbf{a}$ was obtained using the 3 year-out cross-validation method by minimizing the sum of the squared error $SSE = \left(\mathbf{y}^j - \mathbf{M}^j\mathbf{a}^j\right)^T\left(\mathbf{y}^j - \mathbf{M}^j\mathbf{a}^j\right)$, and has the following standard solution:

$$\mathbf{a}^j = \left(\left(\mathbf{M}^j\right)^T\mathbf{M}^j\right)^{-1}\left(\mathbf{M}^j\right)^T\mathbf{y}^j \tag{6}$$

where the superscript j in indicates that all years, but the $j-1$th, jth and $t+1$th target years were included in the regression.

Thus, the predicted value $\hat{y}_j$ for the MLR combination at the jth target year is then estimated as:

$$\hat{y}_j^{MLR} = \sum_{i=1}^{M} x_{j,i}a_i^j + a_0^j \tag{7}$$

The covariance matrix of the model predictions, necessary to estimate the predicted standard deviation, was computed as follows:

$$\mathbf{S}_{xx}^j = \frac{1}{N-3}\left(\mathbf{X}^j\right)^T\mathbf{X}^j = \begin{bmatrix} s_{1,1}^{2,j} & s_{1,2}^{2,j} & \cdots & s_{1,M}^{2,j} \\ s_{2,1}^{2,j} & s_{2,2}^{2,j} & \cdots & s_{2,M}^{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ s_{M,1}^{2,j} & s_{M,2}^{2,j} & \cdots & s_{M,M}^{2,j} \end{bmatrix} \tag{8}$$

where the superscript j indicates that all years, but the $j-1$th, jth and $j+1$th target years were included in the computation of the variances-covariances, and $s_{i,i}^{2,j}$ and $s_{i,m}^{2,j}$ are the variances and covariances given by:

$$s_{i,i}^{2,j} = \frac{1}{N-3}\sum_{k=1}^{N-1}\left(x_{k,k}\right)^2 \quad \forall k \neq j-1, j, j+1 \tag{9}$$

$$s_{i,m}^{2,j} = \frac{1}{N-3}\sum_{k=1}^{N-1}\left(x_{k,i}\cdot x_{k,m}\right) \quad \forall k \neq j-1, j, j+1,$$

$$\text{where } m \text{ is a forecast system, such as } m \neq i \tag{10}$$

The predicted forecast uncertainty for each target year was computed as in Doblas-Reyes et al. (2005):

$$\hat{s}_j^{MLR} = s_0^j\sqrt{1 + \frac{1}{N-3}x_j\left(S_{xx}^j\right)^{-1}x_j^T} \tag{11}$$

where s_0^j is the standard deviation of the regression residuals and $\mathbf{x}_j$ is the vector of predictors at the jth target year, such as $\mathbf{x}_j = [\,x_{j,1} \quad x_{j,2} \quad \cdots \quad x_{j,M}\,]$.

(c) *Principal component multiple linear regression*

A principal component analysis was performed on $\mathbf{X}^j$ aiming at finding a new set of predictors that were uncorrelated from each other. This new set of predictors was used to estimate the PC1-regression (PC1) and the PCA-regression (PCA) combinations. The aim of using the principal component analysis is to avoid introducing a large uncertainty in the estimated linear regression coefficients due to colinearity (Doblas-Reyes et al. 2005).

The eigenvalue decomposition of the covariance matrix $\mathbf{S}_{xx}^j$ can be written as

$$\mathbf{S}_{xx}^j\,\mathbf{E}^j = \boldsymbol{\lambda}^j\,\mathbf{E}^j \tag{12}$$

where $\mathbf{E}^j$ and $\boldsymbol{\lambda}^j$ are the eigenvectors and eigenvalues of $\mathbf{S}_{xx}^j$, respectively.

The principal components (PC), $\mathbf{P}^j$ are given by

$$\mathbf{P}^j = \mathbf{X}^j \cdot E^j \tag{13}$$

The PCs for the jth target year $\mathbf{p}_j$ was computed by multiplying the matrix of eigenvectors $\mathbf{E}^j$ by the vector of predictor $\mathbf{x}_j$ at the jth target year.

The output of the principal components analysis of the four forecast systems was four PCs. In the first case, the PC that explained the largest variance was used to compose $\mathbf{M}$ and the steps B.3 to B.8 were performed to estimate the predicted value $\hat{y}_j^{PC1}$ and the predicted standard deviation $\hat{s}_j^{PC1}$ of the PC1 combination. In the second case, both PCs were used to compose $\mathbf{M}$ and the predicted value $\hat{y}_j^{PCA}$ and the predicted standard deviation $\hat{s}_j^{PCA}$ of the PCA combination were estimated.

(d) *Forecast assimilation (FA)*

The FA is a Bayesian approach that combines the dynamical model predictions with prior historical information to produce calibrated probabilistic forecasts (Stephenson et al. 2005). It can be expressed as:

$$y_j|x_j = N(y_j, s_j) \tag{14}$$

The predicted index $\widehat{y}_j$ and the predicted standard deviation $\widehat{s}_j$ can be written as follows:

$$\widehat{y}_j = y_b^j + \mathbf{L}^j\left[\mathbf{x}_j - \mathbf{G}^j\left(y_b^j - y_0^j\right)\right] \tag{15}$$

$$\widehat{s}_j = \left(\left(\mathbf{G}^j\right)^T\left(\mathbf{S}^{2,j}\right)^{-1}\mathbf{G}^j + \left(\mathbf{C}^j\right)^{-1}\right)^{-1} \tag{16}$$

where the $\mathbf{L}^j = \mathbf{C}^j\left(\mathbf{G}^j\right)^T\left(\mathbf{G}^j\mathbf{C}^j\left(\mathbf{G}^j\right)^T + \mathbf{S}^{2,j}\right)^{-1}$ is the gain/weight matrix. The slope $\mathbf{G}^j$, the intercept y_0^j and the prediction error covariance $\mathbf{S}^j$ matrices were estimated using the least-squares estimation of the regression of the four forecast systems on the observations. They are given by:

$$\mathbf{G}^j = \mathbf{S}_{xy}^j\left(\mathbf{S}_{yy}^j\right)^{-1} \tag{17}$$

$$y_0^j = -\left(\overline{\mathbf{x}}^j - \overline{\mathbf{y}}^j\left(\mathbf{G}^j\right)^T\right)\mathbf{G}^j\left(\left(\mathbf{G}^j\right)^T\mathbf{G}^j\right)^{-1} \tag{18}$$

$$\mathbf{S}^j = \mathbf{S}_{xx}^j\left(\mathbf{S}_{yy}^j\right)^{-1}\left(\mathbf{S}_{xy}^j\right)^T \tag{19}$$

where $\mathbf{S}_{yy}^j$ is the covariance matrix of the observations, and $\mathbf{S}_{xy}^j$ is the cross-covariance matrix. $\overline{\mathbf{x}}^j$ and $\overline{y}^j$ are defined as:

$$\overline{\mathbf{x}}^j = \frac{1}{N-3}\sum_{k=1}^{N-1}\left(\mathbf{x}_k\right)^T \quad \forall k \neq j-1, j, j+1 \tag{20}$$

$$\overline{y}^j = \frac{1}{N-3}\sum_{k=1}^{N-1}\mathbf{y}^j \quad \forall k \neq j-1, j, j+1 \tag{21}$$

The reader will note that the computations are done in cross-validation mode, as above. The mean (y_b^j) and covariance ($\mathbf{C}^j$) matrices of the normally-distributed prior were computed in two different ways:

- In the first case, the expected values and the predicted standard deviation from the statistical model predictions were used as y_b^j and $\mathbf{C}^j$, respectively. This combination was called the FA-statistical (FAS).
- In the second case, the matrix of predictors $\mathbf{x}^j$ is made of the three forecast systems: the statistical model and the three dynamical forecast systems, and the prior distribution was estimated using the climatological information, such as:

$$y_b^j = \overline{y}^j$$

$$\mathbf{C}^j = \mathbf{S}_{yy}^j$$

This last combination is referred to as the FA-climatology (FAC).

References

Balmaseda MA, Davey MK, Anderson DLT (1995) Decadal and seasonal dependence of ENSO prediction skill. J Clim 8:2705–2715

Barnston AG, Van den Dool HM (1993) A degeneracy in cross-validated skill in regression-based forecasts. J Clim 6:963–977

Batté L, Déqué M (2011) Seasonal predictions of precipitation over Africa using coupled ocean-atmosphere general circulation models: skill of the ENSEMBLES project multimodel ensemble forecasts. Tellus A 63:283–299

Coelho CAS, Pezzulli S, Balmaseda M, Doblas-Reyes FJ, Stephenson DB (2004) Forecast calibration and combination: a simple Bayesian approach for ENSO. J Clim 17:1504–1516

Curry JA, Webster PJ (2011) Climate science and the uncertainty monster. Bull Am Meteorol Soc 92:1667–1682

DelSole T, Yang X, Tippett MK (2012) Is unequal weighting significantly better than equal weighting for multi-model forecasting? Q J R Meteorol Soc 139:176–183

Doblas-Reyes FJ, Hagedorn R, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. Tellus A 57:234–252

Doblas-Reyes FJ, Weisheimer A, Déqué M, Keenlyside N, McVean M, Murphy JM, Rogel P, Smith D, Palmer TN (2009) Addressing model uncertainty in seasonal and annual dynamical seasonal forecasts. Q J R Meteorol Soc 135:1538–1559

Doblas-Reyes FJ, Garcia-Serrano J, Lienert F, Pinto-Biescas A, Rodrigues LRL (2013) Seasonal climate predictability and forecasting: status and prospects. WIRE Clim Change (in press)

Gneiting T, Raftery AE (2005) Weather forecasting with ensemble methods. Science 310:248–249

Goddard L, Mason SJ, Zebiak SE, Ropelewski CF, Basher R, Cane MA (2001) Current approaches to seasonal to interannual climate predictions. Int J Climatol 21:1111–1152

Hagedorn R, Doblas-Reyes FJ, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. Tellus A 57:219–233

Kim HM, Webster PJ, Curry JA (2012) Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter. Clim Dyn 39:2957–2973

Knutti R (2010) The end of model democracy? Editor Clim Change 102:395–404

Kug J-S, Lee J-Y, Kang I-S (2007) Global sea surface temperature prediction using a multi-model ensemble. Mon Weather Rev 135:3239–3247

Kug J-S, Lee J-Y, Kang I-S, Wang B, Park C-K (2008) Optimal multi-model ensemble method in seasonal climate prediction. Asian Pac J Atmos Sci 44:259–267

Kumar A, Chen M, Zhang L, Wang W, Xue Y, Wen C, Marx L, Huang B (2012) An analysis of the nonstationarity in the bias of sea surface temperature forecasts for the NCEP climate forecast system (CFS) version 2. Mon Weather Rev 140:3003–3016

Mason SJ (2008) Understanding forecast verification statistics. Meteorol Appl 15:31–40

Mason SJ, Baddour O (2008) Statistical modeling. In: Troccoli A, Harrison MSJ, Anderson DLT, Mason SJ (eds) Seasonal climate: forecasting and managing risk. Springer, Dordrecht, pp 167–206

Mason SJ, Mimmack GM (2002) Comparison of some statistical methods of probabilistic forecasting of ENSO. J Clim 15:8–29

Mason SJ, Stephenson DB (2008) How can we know whether the forecasts are any good? In: Troccoli A, Harrison MSJ, Anderson DLT, Mason SJ (eds) Seasonal climate: forecasting and managing risk. Springer, Dordrecht, pp 259–289

Molteni F, Stockdale T, Balmaseda M, Balsamo G, Buizza R, Ferranti L, Magnusson L, Mogensen K, Palmer T, Vitart F (2011) The new ECMWF seasonal forecast system (System 4). ECMWF Tech Memo 656:51. http://www.ecmwf.int/publications/library/do/references/list/14. Accessed 20 Dec 2012

Murphy AH, Winkler RL (1984) Probability forecasting in meterology. J Am Stat Assoc 79:489–500

Palmer TN (2000) Predicting uncertainty in forecasts of weather and climate. Rep Prog Phys 63:71–116

Palmer TN, Alessandri A, Andersen U, Cantelaube P, Davey M, Délécluse P, Déqué M, Díez E, Doblas-Reyes FJ, Feddersen H, Graham R, Gualdi S, Guérémy J-F, Hagedorn R, Hoshen M, Keenlyside N, Latif M, Lazar A, Maisonnave E, Marletto V, Morse AP, Orfila B, Rogel P, Terres J-M, Thomson MC (2004) Development of a European multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). Bull Am Meteorol Soc 85:853–872

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res 108:4407–4444

Ropelewski CF, Halpert M (1987) Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. Mon Weather Rev 115:1606–1626

Saha S, Nadiga S, Thiaw C, Wang J, Wang W, Zhang Q, Van den Dool HM, Pan HL, Moorthi S, Behringer D, Stokes D, Peña M, Lord S, White G, Ebisuzaki W, Peng P, Xie P (2006) The NCEP Climate Forecast System. J Clim 19:3483–3517

Saha S, Moorthi S, Wu X, Wang J, Nadiga S, Tripp P, Pan HL, Behringer D, Hou YT, Chuang H, Iredell M, Ek M, Meng J, Yang R (2013) The NCEP Climate Forecast System version 2. J Clim. http://cfs.ncep.noaa.gov/

Saji HN, Goswami BN, Vinayachandran PN, Yamagata T (1999) A dipole mode in the tropical Indian Ocean. Nature 401:360–363

Shukla J (1998) Predictability in the midst of chaos: a scientific basis for climate forecasting. Science 282:728–731

Slingo J, Palmer TN (2011) Uncertainty in weather and climate prediction. Philos Trans R Soc A 369:4751–4767

Sooraj KP, Annamalai H, Kumar A, Wang H (2012) A comprehensive assessment of CFS seasonal forecasts over the tropics. Weather Forecast 27:3–27

Stephenson DB, Coelho CAS, Doblas-Reyes FJ, Balmaseda M (2005) Forecast Assimilation: a unified framework for the combination of multi-model weather and climate predictions. Tellus A 57:253–264

Stephenson DB, Coelho CAS, Jolliffe IT (2008) Two extra components in the Brier score decomposition. Weather Forecast 23:752–757

Stockdale TN, Anderson DLT, Balmaseda MA, Doblas-Reyes FJ, Ferranti L, Mogensen K, Palmer TN, Molteni F, Vitart F (2011) ECMWF seasonal forecast System 3 and its prediction of sea surface temperature. Clim Dyn 37:455–471

Tippett MK, Barnston AG (2008) Skill of multimodel ENSO probability forecasts. Mon Weather Rev 136:3933–3946

Wang B, Li J-Y, Kang I-S, Shukla J, Park C-K, Kumar A, Schemm J, Cocke S, Kug J-S, Luo J–J, Fu X, Yun W-T, Alves O, Jin E, Kinter J, Kirtman B, Krishnamurti T, Lau N, Lau W, Liu P, Pegion P, Rosati T, Schubert S, Stern W, Suarez M, Yamagate T (2009) Advance and prospectus of seasonal prediction: assessment of the APCC/CliPAS 14 model ensemble retrospective seasonal prediction (1980–2004). Clim Dyn 33:93–117

Yuan X, Wood EF, Luo L, Pan M (2011) A first look at Climate Forecast System version 2 (CFSv2) for hydrological seasonal prediction. Geophys Res Lett 38:L13402. doi:10.1029/2011GL047792