Climate Dynamics
https://doi.org/10.1007/s00382-018-4140-4

Calibration and combination of monthly near-surface temperature and precipitation predictions over Europe

Luis R. L. Rodrigues[1] · Francisco J. Doblas-Reyes[2,3] · Caio A. S. Coelho[4]

Received: 30 December 2016 / Accepted: 15 February 2018

Abstract
A Bayesian method known as the Forecast Assimilation (FA) was used to calibrate and combine monthly near-surface temperature and precipitation outputs from seasonal dynamical forecast systems. The simple multimodel (SMM), a method that combines predictions with equal weights, was used as a benchmark. This research focuses on Europe and adjacent regions for predictions initialized in May and November, covering the boreal summer and winter months. The forecast quality of the FA and SMM as well as the single seasonal dynamical forecast systems was assessed using deterministic and probabilistic measures. A non-parametric bootstrap method was used to account for the sampling uncertainty of the forecast quality measures. We show that the FA performs as well as or better than the SMM in regions where the dynamical forecast systems were able to represent the main modes of climate covariability. An illustration with the near-surface temperature over North Atlantic, the Mediterranean Sea and Middle-East in summer months associated with the well predicted first mode of climate covariability is offered. However, the main modes of climate covariability are not well represented in most situations discussed in this study as the seasonal dynamical forecast systems have limited skill when predicting the European climate. In these situations, the SMM performs better more often.

Keywords Climate prediction · Multimodel ensemble · Forecast quality assessment · Forecast assimilation

1 Introduction

Seasonal climate prediction attempts to predict monthly or 3-month statistical properties of climate with lead time ranging from 1 month to 1 year (Doblas-Reyes et al. 2013). On these timescales, surface climate variables are difficult to predict over extratropical land areas because the response of the atmosphere to the slowly varying components of the climate system is highly uncertain and subject to many sources of error (Mason et al. 1999; Goddard et al. 2001; Doblas-Reyes et al. 2013; Scaife et al. 2014).

Statistical and dynamical forecast systems have limited prediction skill over land, particularly in extratropical latitudes (Goddard et al. 2001; Doblas-Reyes et al. 2013). Barnett and Preisendorfer (1987) assessed the forecast quality of several statistical forecast systems to predict monthly and seasonal near-surface temperature over the United States. They showed that the statistical forecast systems that had persistence and sea surface temperature (SST) as predictors offered the highest prediction skill in summer among the statistical schemes examined in their study. The highest summer prediction skill was linked to a decadal northern

This paper is a contribution to the special collection on the North American Multi-Model Ensemble (NMME) seasonal prediction experiment. The special collection focuses on documenting the use of the NMME system database for research ranging from predictability studies, to multi-model prediction evaluation and diagnostics, to emerging applications of climate predictability for subseasonal to seasonal predictions. This special issue is coordinated by Annarita Mariotti (NOAA), Heather Archambault (NOAA), Jin Huang (NOAA), Ben Kirtman (University of Miami) and Gabriele Villarini (University of Iowa).

✉ Luis R. L. Rodrigues
luis.rodrigues@inpe.br

[1] Centro de Ciências do Sistema Terrestre, Instituto Nacional de Pesquisas Espaciais, Rod. Presidente Dutra Km 40, Cachoeira Paulista 12630-000, Brazil

[2] Barcelona Supercomputing Center-Centro Nacional de Supercomputación, C. Jordi Girona 29, Barcelona 08034, Spain

[3] ICREA, Pg. Lluís Companys 23, Barcelona 08010, Spain

[4] Centro de Previsão de Tempo e Estudos Climáticos, Instituto Nacional de Pesquisas Espaciais, Rod. Presidente Dutra Km 40, Cachoeira Paulista 12630-000, Brazil

Published online: 28 February 2018

hemisphere (NH) near-surface temperature variability. Prediction skill in statistical forecast systems linked to decadal near-surface temperature variability was also described in studies focused on northern Europe (Johansson et al. 1998) and globally (Eden et al. 2015). Eden et al. (2015) showed that seasonal prediction skill was considerably reduced when trends for near-surface temperature were removed prior to the forecast quality assessment.

Seasonal prediction skill by dynamical forecast systems is assessed routinely for operational systems. As an illustration of such studies, Graham et al. (2005) showed that the UK Met Office coupled (one-tier) dynamical forecast system has higher seasonal prediction skill than its uncoupled (two-tier) version, especially in tropical regions. This is because dynamical forecast systems have skill when predicting the SST anomalies in the tropical Pacific Ocean (Rodrigues et al. 2014a) as well as the corresponding atmospheric response both locally and remotely (Kim et al. 2012). On the other hand, dynamical forecast systems still have limitations to predict the interannual atmospheric variability in the extratropics (Kim et al. 2012) because many physical processes that could lead to prediction skill in these regions, such as an appropriate representation of sea ice, the land surface and the stratosphere is still under development (Arribas et al. 2011; Stockdale et al. 2015). Despite all these issues, climate modelers have made significant progress in predicting the North Atlantic Oscillation (NAO) and its links to surface climate variables at seasonal timescale (Scaife et al. 2014).

Because of imperfections of forecast systems in simulating the observed climate, seasonal climate prediction must account for the uncertainties associated with forecast system formulation. The multimodel ensemble (MME), a method that combines predictions derived from several forecast systems, has proven to produce better seasonal predictions than other methods by providing better estimates of the forecast uncertainty (Doblas-Reyes et al. 2009). This is particularly relevant to regions with low skill such as Europe. When using the multimodel approach, the question of how to formulate the most skillful combination arises. In this respect, several studies have demonstrated that the combination of predictions derived from several forecast systems with equal weights, referred to hereafter as simple multimodel ensemble (SMM), produces on average better predictions than the best single forecast system (Doblas-Reyes et al. 2005; Hagedorn et al. 2005). However, even the SMMs hardly produce skillful predictions over Europe (Doblas-Reyes et al. 2000, 2009, 2013; Wang et al. 2009).

To improve the SMM, Robertson et al. (2004) applied a Bayesian methodology to combine seasonal precipitation and near-surface temperature predictions derived from several dynamical forecast systems with a climatological distribution. Their combination method estimated weights for each dynamical forecast system independently at each season, variable and grid point based on the ranked probability skill score (RPSS). The idea was to assign more weight to the predictions that had higher RPSS and assign more weight to the climatological distribution in situations when predictions were found to be unskillful. They concluded that their Bayesian scheme was more skillful than the SMM and the single dynamical forecast systems. However, the main benefit of their method when predicting seasonal precipitation in extratropical regions was to bring much of the large area of negative RPSS values to near-zero skill.

Similarly, Coelho et al. (2006) applied a Bayesian methodology to combine summer precipitation predictions over South America derived from several dynamical forecast systems with a prior distribution estimated from a simple statistical forecast system. Their Bayesian method, referred to hereafter as Forecast Assimilation (FA; Coelho et al. 2004, 2006; Stephenson et al. 2005), requires the application of a dimension reduction technique to deal with the high dimensionality of a spatial MME with strong dependency between values of neighboring grid points (Stephenson et al. 2005). This is an aspect of the combination problem not dealt with in Robertson et al. (2004) where the weights were estimated independently at each grid point. Coelho et al. (2006) concluded that the FA combination improved the prediction skill over the SMM combination and the single forecast systems in terms of the Brier score (BS) and its components.

The aim of the present study is to apply the FA method to calibrate and combine monthly near-surface temperature and precipitation predictions over Europe, an extratropical region where seasonal prediction skill is low. The relative merits of this method are discussed. Six seasonal dynamical forecast systems produced the monthly-mean predictions: two from the European Seasonal to Interannual Prediction Project (EUROSIP; Vitart et al. 2007) and four from the North American Multimodel Ensemble (NMME; Kirtman et al. 2014). These are two independent international MME efforts that are currently producing real-time predictions. Exploring the feasibility of combining monthly predictions from two independent operational MME efforts is an important novelty of this study.

The seasonal dynamical forecast systems and the observational reference are described in Sect. 2. Section 2 also presents the combination and the forecast quality methods. The forecast quality assessment is discussed in Sect. 3 whereas the benefits and limitations of the FA method are described in Sect. 4. The conclusions are presented in Sect. 5.

2 Data and methods

2.1 Observations

The observational reference for precipitation is the version 2.2 of the Global Precipitation Climatology Project (GPCP)

Table 1 Forecast systems from the EUROSIP and NMME systems

Forecast system	Atmospheric component	Ocean component	Number of ensembles
S4	Integrated forecast system with 80 km resolution	Nucleus for European Modelling of the Ocean version 3.0	51
MF3	Action de Recherche Petite Echelle Grande Echelle version 4 with 300 km resolution	Océan PArallélisé model version 8.2	11
CFSv2	Global forecast system with 100 km resolution	Modular Ocean Model version 4	24 (May) 28 (Nov)
GFDL	GFDL atmospheric model with 200 km resolution	Modular Ocean Model version 4	10
CMC2	Canadian atmospheric model version 4 with 200 km resolution	Canadian ocean model version 4	10
NASA	Goddard earth observing system version 5 with 200 km resolution	Modular Ocean Model version 4	11

monthly satellite-gauge combination with 2.5° horizontal resolution (Huffman and Bolvin 2013). The European Centre for Medium Range Weather Forecasts (ECMWF) global atmospheric reanalysis ERA-Interim with 0.7° horizontal resolution (Dee et al. 2011) was used as the observational reference for near-surface temperature. Both datasets cover land and ocean for the period from January 1979 onwards. They were used in both the forecast quality assessment and the estimation of the FA combination parameters (i.e. calibration).

2.2 Seasonal dynamical forecast systems

Table 1 summarizes the characteristics of the six seasonal dynamical forecast systems used in this study. They are the ECMWF climate forecast system 4 (S4; Molteni et al. 2011; Kim et al. 2012) and Météo-France seasonal forecast system version 3 (MF3; Alessandri et al. 2011) from the EUROSIP and the National Center for Environmental Prediction (NCEP) climate forecast system version 2 (CFSv2; Yuan et al. 2011; Kim et al. 2012; Saha et al. 2014), Geophysical Fluid Dynamics Laboratory Climate Model version 2.1 (GFDL; Zhang et al. 2007), Canadian Meteorological Center seasonal forecast system version 2 (CMC2; Merryfield et al. 2013) and the National Aeronautics and Space Administration Global Modeling and Assimilation Office (NASA; Vernieres et al. 2012) from the NMME. The seasonal dynamical forecast system will be referred to hereafter simply as forecast system.

The forecast systems, except for CFSv2 and NASA, initialize all ensemble members in burst mode on the first day of every month at 0 UTC. CFSv2´s ensemble members are initialized in different days and times, being the ones initialized after the seventh day of the month used as the zero-month lead time ensemble members of the next month (Saha et al. 2014). For example, the ensemble for the target month of February at zero-month lead time have their members initialized 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. NASA initialized five ensemble members, one every 5 days, and six members in burst mode on whichever of the five-day starts is closest to the first of the month.[1]

A bilinear interpolation was performed on the predictions and observational references into a common 2.5° global grid prior to the combination and forecast quality assessment because they have different horizontal resolution. The combination and forecast quality assessment were performed in winter (NDJF) and summer (MJJA) months for predictions initialized in May and November, allowing the evaluation of prediction skill varies from zero to 3-month lead time. The hindcast period used in the analysis covers the 1982–2010 common period when all forecast systems had available data at the time this study started.

2.3 Combination methods

Two combination methods were used in this study. The first one is the simple multimodel ensemble (SMM; Doblas-Reyes et al. 2005; Hagedorn et al. 2005), a method that combines predictions derived from several forecast systems with equal weights. In this method, the predicted mean is the simple arithmetic average of the ensemble-mean of the six forecast systems computed independently at each grid point and can be written as:

$$\hat{\boldsymbol{y}}_i^{SMM} = \frac{1}{m}\sum_{j=1}^{m}\hat{\boldsymbol{m}}_{i,j},$$

where $\hat{\boldsymbol{m}}_{i,j}$ is a vector with q elements representing an ensemble-mean of the jth forecast system at the ith hindcast year at each grid point for a given variable, target month and lead time. m is the number of forecast systems (i.e. 6 forecast systems). q is the number of longitude points (i.e. 36 points) times the number latitude points (i.e. 22 points) covering

[1] http://gmao.gsfc.nasa.gov/research/climate/NMME/. Accessed 15 December 2016.

Europe and adjacent regions at the 2.5° resolution. The hat over the vector $\hat{\boldsymbol{m}}_{i,j}$ reminds that the ensemble-mean is a prediction (hindcast).

Similarly, the cumulative density function (CDF), used to quantify the quality of probabilistic predictions of the SMM, can be written as:

$$\hat{\boldsymbol{Y}}_i^{SMM} = \frac{1}{m}\sum_{j=1}^{m}\hat{\boldsymbol{F}}\left(\hat{\boldsymbol{E}}_{i,j}\right),$$

and

$$\hat{\boldsymbol{E}}_{i,j} = \begin{bmatrix} \hat{e}_{1,1} & \hat{e}_{1,2} & \cdots & \hat{e}_{1,q} \\ \hat{e}_{2,1} & \hat{e}_{2,2} & \cdots & \hat{e}_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{e}_{n,1} & \hat{e}_{n,2} & \cdots & \hat{e}_{n,q} \end{bmatrix},$$

where $\hat{\boldsymbol{F}}$ is the Gaussian kernel density estimate as a function of the ensemble members $\left(\hat{e}_{n,q}\right)$ for each forecast system independently and n is the number of ensemble members, which varies according to the forecast system as described in Table 1 (e.g. S4 has 51 ensemble members). Note that the vector $\hat{\boldsymbol{m}}_{i,j}$ is the row average of the matrix $\hat{\mathbf{E}}_{\mathrm{i,j}}$. For instance, $\hat{m}_{1,1,1} = \frac{1}{51}\sum_{k=1}^{51}\hat{e}_{\mathrm{k,1}}$ represents the first element of the ensemble-mean of S4 (j = 1) at the first hindcast year (i = 1), such as $\hat{\boldsymbol{m}}_{1,1} = c\left(\hat{m}_{1,1,1}, \hat{m}_{1,1,2}, \ldots, \hat{m}_{1,1,\mathrm{q}}\right)$. The ensemble-mean and kernel density estimates were computed prior to the combination to avoid assigning more weight to the forecast systems that have a larger number of ensemble members.

The second combination method attempts to assign more weight to the predictions derived from the better forecast systems over the hindcast period. The Forecast Assimilation (FA; Coelho et al. 2004, 2006; Stephenson et al. 2005), a Bayesian method that calibrates and combines predictions from several sources with a prior historical information, was used to assign the weights. Because of the high dimensionality of the multimodel ensemble of spatial fields compared to the number of hindcast years and strong dependency between values of neighboring grid points, a dimensionality reduction is applied on the original gridded predictions prior to the combination (Stephenson et al. 2005). The maximum covariance analysis (MCA) was the statistical technique used in this study for dimensionality reduction. First, the observations were organized in a matrix to place the grid points in the rows and the training years in the columns, such as:the first element of

$$\boldsymbol{Y}^{\mathbf{i}} = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,q} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{t,1} & y_{t,2} & \cdots & y_{t,q} \end{bmatrix},$$

where t is the number of training years. Note that the letters $\hat{y}/\hat{\boldsymbol{Y}}$ (with hats) represent the predictions or estimates of $y/\boldsymbol{Y}$ (observational references). The FA calibration and combination procedure was performed using 3-year-out cross-validation method to reduce the effect of artificial skill (Mason and Baddour 2008). Therefore, the number of training years equals the number of hindcast years (i.e. 1982–2010: 29 years) minus three, which are the target year, the previous year and the following year. The only exceptions were the first and last target years, when two hindcast years were removed: the target year and the following one for the former and the target year and the prior year for the later. Thus, $t = N - 2$ for the first and last target years and $t = N - 3$ otherwise, where N is the number of hindcast years. The cross-validation is indicated by the superscript in the matrix $\boldsymbol{Y}$. For instance, the matrix $\boldsymbol{Y}^1$ has the elements $y_{1,1}$, $y_{2,1}$ and $y_{t,1}$ representing the first grid point in the years 1984, 1985 and 2010 because the years 1982 and 1983 were removed in the cross-validation procedure whereas the matrix $\boldsymbol{Y}^2$ has the elements $y_{1,q}$, $y_{2,q}$ and $y_{t,q}$ representing the qth grid point in the years 1985, 1986 and 2010 because the years 1982, 1983 and 1984 were removed in the cross-validation procedure. Similarly, all predictions were organized in a matrix where the grid points of all forecast systems are placed in the rows and the training years in the columns, such as:

$$\boldsymbol{A} = \begin{bmatrix} \hat{\boldsymbol{m}}_{1,1} & \hat{\boldsymbol{m}}_{1,2} & \cdots & \hat{\boldsymbol{m}}_{1,m} \\ \hat{\boldsymbol{m}}_{2,1} & \hat{\boldsymbol{m}}_{2,2} & \cdots & \hat{\boldsymbol{m}}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\boldsymbol{m}}_{t,1} & \hat{\boldsymbol{m}}_{t,2} & \cdots & \hat{\boldsymbol{m}}_{t,m} \end{bmatrix},$$

where the vector $\hat{\mathbf{m}}_{1,1}$ have all grid points of the first forecast system in the first training year. The cross-covariance matrix can be then decomposed into the product of three matrices, such as:

$$\boldsymbol{Y}^{i^T}\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T,$$

where the columns of $\mathbf{U}_{q,q}$ are the orthonormal eigenvectors of $\left(\mathbf{Y}^{\mathbf{i}^{\mathrm{T}}}\mathbf{A}\right)\left(\mathbf{Y}^{\mathbf{i}^{\mathrm{T}}}\mathbf{A}\right)^{\mathrm{T}}$, the columns of $\mathbf{V}_{(q^*m),(q^*m)}$ are the orthonormal eigenvectors of $\left(\mathbf{Y}^{\mathbf{i}^{\mathrm{T}}}\mathbf{A}\right)^{\mathrm{T}}\left(\mathbf{Y}^{\mathbf{i}^{\mathrm{T}}}\mathbf{A}\right)$ and $\mathbf{D}_{q,(q^*m)}$ is a diagonal matrix containing the square roots of the eigenvalues of $\mathbf{U}$ or $\mathbf{V}$. Note that the product (q^*m) is the number of columns in the matrix $\boldsymbol{A}$ which is equal the number of grid points times the number of forecast systems and the superscript i was removed from the other matrices for simplicity. The expansion coefficients associated with the eigenvectors can be written as:

$$\boldsymbol{Z} = \boldsymbol{Y}^{\mathbf{i}}\boldsymbol{U},$$

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{V},$$

where $\boldsymbol{X}$ and $\mathbf{Z}$ are the right (predictions) and left (observations) expansion coefficients and η is the number of retained modes. The number of retained MCA modes that gives the best FA prediction varies according to the climate variable, start date, lead time and area. However, the FA combination estimated by retaining the three first modes gave the best predictions more often. After the multimodel ensemble of spatial-fields were reduced into η time series that accounted for most of the dataset covariability, the likelihood parameters for the prediction calibration, the slope $\mathcal{G}$, the intercept z_0 and the prediction error covariance $\mathcal{S}$, were estimated:

$$\mathcal{G} = \mathcal{S}_{XZ}{\mathcal{S}_{ZZ}}^{-1}$$
$$z_0 = -\left(\bar{x} - \bar{z}\mathcal{G}^T\right)\mathcal{G}\left(\mathcal{G}^T\mathcal{G}\right)^{-1}$$
$$\mathcal{S} = \mathcal{S}_{XX} - \mathcal{S}_{XZ}{\mathcal{S}_{ZZ}}^{-1}{\mathcal{S}_{XZ}}^T,$$

where $\bar{x}$ and $\bar{z}$ are the row averages of the matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively, $\mathcal{S}_{ZZ}$ is the $(\eta \times \eta)$ covariance matrix of $\boldsymbol{Z}$, $\mathcal{S}_{XX}$ is the $(\eta \times \eta)$ covariance matrix of $\boldsymbol{X}$, and $\mathcal{S}_{XZ}$ is the $\eta \times \eta$ cross-covariance matrix between $\boldsymbol{X}$ and $\boldsymbol{Z}$. The FA predicted mean and covariance were then estimated in the MCA space:

$$\hat{\boldsymbol{Z}}_i = \hat{z} + \mathcal{L}\left[\hat{\boldsymbol{x}}_i - \mathcal{G}\left(\bar{z} - z_0\right)\right]$$
$$\hat{\mathcal{S}}_i = \left(\mathcal{G}^T\mathcal{S}^{-1}\mathcal{G} + \mathcal{C}^{-1}\right)^{-1},$$

where $\hat{\boldsymbol{x}}_i = \hat{\boldsymbol{m}}_i\boldsymbol{V}$ is the predictor vector at the ith target year in the MCA space, $\hat{\boldsymbol{m}}_i = \left(\hat{m}_1\ \hat{m}_2\ \cdots\ \hat{m}_{(q*m)}\right)$ is the vector containing the ensemble-mean at each grid point for all forecast systems at the ith target year and $\mathcal{L}^{\mathrm{i}} = \mathcal{C}\mathcal{G}^{\boldsymbol{T}}\left(\mathcal{G}\mathcal{C}\mathcal{G}^{\boldsymbol{T}} + \mathcal{S}\right)^{-1}$ is the gain/weight matrix. In this study, the climatology ($\bar{z}$) was used as the prior information. The final step is to get the FA predictions in the geographical coordinate system:

$$\hat{\boldsymbol{Y}}_i^{FA} = \boldsymbol{U}\hat{\boldsymbol{Z}}_i$$
$$\hat{\boldsymbol{S}}_i^{FA} = \boldsymbol{U}\hat{\mathcal{S}}_i\boldsymbol{U}^{\boldsymbol{T}},$$

where $\hat{\mathbf{Y}}_{\mathrm{i}}^{\mathrm{FA}}$ and $\hat{\mathbf{S}}_{\mathrm{i}}^{\mathrm{FA}}$ are the FA predicted mean and covariance matrices.

The above equations show that the FA method will perform poorly if $\boldsymbol{U}$ and $\boldsymbol{V}$ are not well estimated. In other words, the multimodel ensemble must predict well the observed leading modes of covariability. Therefore, heterogeneous correlation matrices were computed by correlating the right (left) expansion coefficients with the observations (predictions) to illustrate the advantages and disadvantages of the FA. They are called heterogenous because we correlate $\boldsymbol{X}$ (predictions in the MCA space) with $\boldsymbol{Y}$ (observations) and $\boldsymbol{Z}$ (observations in the MCA space) with $\boldsymbol{A}$ (predictions). Representing the observed heterogeneous correlation matrices is extremely challenging given the small sample size in current operational forecast systems and their lack of skill in simulating the interannual extratropical climate covariability.

2.4 Forecast quality assessment

The forecast quality assessment was performed using two different measures: one for deterministic and another one for probabilistic predictions. The Pearson correlation coefficient between the predicted mean and the corresponding observation, which is a measure of association, was the deterministic measure used in this study. The temporal correlation coefficient over the hindcast period was computed independently for each grid point, target month and lead time as follows:

$$r = \frac{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}}\right)^2\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}},$$

where $\hat{y}_i$ and y_i are the predictions and observations and $\bar{\hat{y}}$ and $\bar{y}$ are their respective mean over N the hindcast years. Note that $\hat{y}_i$ and y_i are anomaly values computed in 3-year-out cross-validation mode and that their time averages are computed for all hindcast years (i.e., they are close to, but not exactly zero).

The correlation coefficient measures the quality of deterministic forecasts (i.e., how well the mean of the probability density function (PDF) is predicted), but provides no information about the quality of the forecast uncertainty (i.e., how well the spread of the PDF is predicted). The continuous ranked probability score (CRPS; Gneiting et al. 2005) was used to quantify the quality of probabilistic predictions. It compares the predicted CDF with a Heaviside function, which assigns probability density 0 to values smaller than the observation and 1 otherwise, and is defined on a continuous scale so that the reduction of probabilistic predictions into discrete probabilities of binary or categorical events is not needed. It can be generically defined as:

$$CRPS = \int_{-\infty}^{\infty}\left[\hat{F}(\chi) - F_0(\chi)\right]^2 d\chi,$$

where

$$F_0(\chi) = \begin{cases} 0, & \chi < y_i \\ 1, & \chi \geqslant y_i \end{cases},$$

is the Heaviside step function that jumps from 0 to 1 at the observation and $\hat{F}(\chi)$ is the predicted CDF. The CRPS has the advantage of being defined on a continuous scale and does not require reduction to discrete probabilities of binary events (Jolliffe and Stephenson 2012).

The CRPS was estimated differently for the two types of probability forecasts handled in this study. When ensemble

predictions are considered, the CRPS was estimated assigning equal weight to each ensemble member, such as:

$$\hat{p}_{i,j,k} \equiv \frac{k}{n}, \quad \text{for} \quad \hat{e}_{i,j,k} < \hat{e}_{i,j,k+1} < \cdots \hat{e}_{i,j,\mathrm{n}},$$

where $\hat{p}_{i,j,k}$ is a piecewise constant for the kth ensemble member and n is the number of ensemble members of the jth forecast system at the ith hindcast year. Note that $\hat{p}_{i,j,0} = -\infty$ and $\hat{p}_{i,j,m+1} = \infty$ are introduced for convenience. For instance, S4 has $\hat{p}_{i,j,1} = \frac{1}{51} \cdots = \hat{p}_{i,j,51} = \frac{51}{51} = 1$ given it is a cumulative distribution.

Therefore, the CRPS can be rewritten to deal with ensemble forecasts (Hersbach 2000):

$$\mathrm{CRPS} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{k}=1}^{n} \mathrm{c}_{\mathrm{i,j,k}},$$

where

$$\mathrm{c}_{\mathrm{i,j,k}} = \alpha_{\mathrm{k}} \hat{\mathrm{p}}_{\mathrm{i,j,k}}^{2} + \beta_{\mathrm{t}} \left(1 - \hat{\mathrm{p}}_{\mathrm{i,j,k}}\right)^{2}.$$

If the observation $\mathrm{y_i}$ falls between the lowest and highest ensemble member, then α_{k} and β_{k} can be estimated as follows:

$0 < k < n$	α_k	β_k
$y_i > \hat{e}_{i,j,k+1}$	$\hat{e}_{i,j,k+1} - \hat{e}_{i,j,k}$	0
$\hat{e}_{i,j,k+1} > y_i > \hat{e}_{i,j,k}$	$y_i - \hat{e}_{i,j,k}$	$\hat{e}_{i,j,k+1} - y_i$
$y_i < \hat{e}_{i,j,k}$	0	$\hat{e}_{i,j,k+1} - \hat{e}_{i,j,k}$

Otherwise as:

Outlier	α_k	β_k
$y_i < \hat{e}_{i,j,1}$	0	$\hat{e}_{i,j,k} - y_i$
$y_i > \hat{e}_{i,j,n}$	$y_i - \hat{e}_{i,j,k}$	0

The FA predictions, where sets of predictions defined by a forecast mean and standard deviation are considered, the CRPS was estimated as follows (Gneiting et al. 2005):

$$\mathrm{CRPS} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \hat{\sigma}_{\mathrm{i}} \left\{ \frac{\mathrm{y_i} - \hat{y}_{\mathrm{i}}}{\hat{\sigma}_{\mathrm{i}}} \left[2\mathrm{F}\left(\frac{\mathrm{y_i} - \hat{y}_{\mathrm{i}}}{\hat{\sigma}_{\mathrm{i}}} \right) - 1 \right] + 2\mathrm{f}\left(\frac{\mathrm{y_i} - \hat{y}_{\mathrm{i}}}{\hat{\sigma}_{\mathrm{i}}} \right) - \frac{1}{\sqrt{\pi}} \right\},$$

where F and f denote the CDF and PDF of the normal distribution with zero 0 and variance 1 at the normalized prediction error $\frac{y_i - \hat{y}_i}{\hat{\sigma}_i}$, respectively. $\hat{y}_i$ and $\hat{\sigma}_i$ are the elements of the matrices $\hat{\boldsymbol{Y}}_i^{FA}$ and $\hat{\boldsymbol{S}}_i^{FA}$, respectively. Note that the CRPS was computed independently at each grid point.

The CRPS can be computed in terms of skill score, such as:

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}}{\mathrm{CRPS_{CLIM}}},$$

where $CRPS_{CLIM}$ is the CRPS of the climatological distribution computed considering all but the target year as a large ensemble member.

The statistical significance at the 5% level was estimated using non-parametric bootstrap (Mason 2008; Jolliffe and Stephenson 2012) to quantify the sampling uncertainty of the verification measures. This method was chosen because it can be used in both correlation coefficient and CRPSS in a uniform fashion. In this procedure, the prediction-observation pairs are randomly resampled 1000 times with replacement, keeping the prediction and observation pairs together (Mason 2008). For the deterministic assessment, these pairs are composed by the predicted mean and the corresponding observation. For the probabilistic assessment, the pairs are made of the forecast CDF and the corresponding observation. The null hypothesis used to estimate the p-values was that the verification measure was zero, while the alternative hypothesis was that the verification measure was larger than zero (i.e. one-tailed test).

3 Forecast quality assessment of the forecast systems

In agreement with previous studies (Wang et al. 2009; Arribas et al. 2011; Kim et al. 2012; Doblas-Reyes et al. 2009, 2013), the forecast systems are often more skillful when predicting near-surface temperature (Fig. 1) than precipitation (Fig. 2). Figure 1 illustrates the correlation coefficient between the predicted and observed June near-surface temperature for predictions initialized in the previous May (1-month lead time) over the 1982–2010 hindcast period. Large areas of statistically significant positive correlation coefficient illustrate that the forecast systems are able to reproduce the main observed monthly near-surface temperature patterns over the North Atlantic, Mediterranean Sea, Europe and Middle East in summer months. However, all of them lack skill when predicting the June near-surface temperature over the Scandinavian region.

The correlation coefficient is sensitive to linear trends. Therefore, it was also computed after the linear trend was removed from the observations and predictions (not shown). The linear trend was computed for the forecast system ensemble mean, not the ensemble members

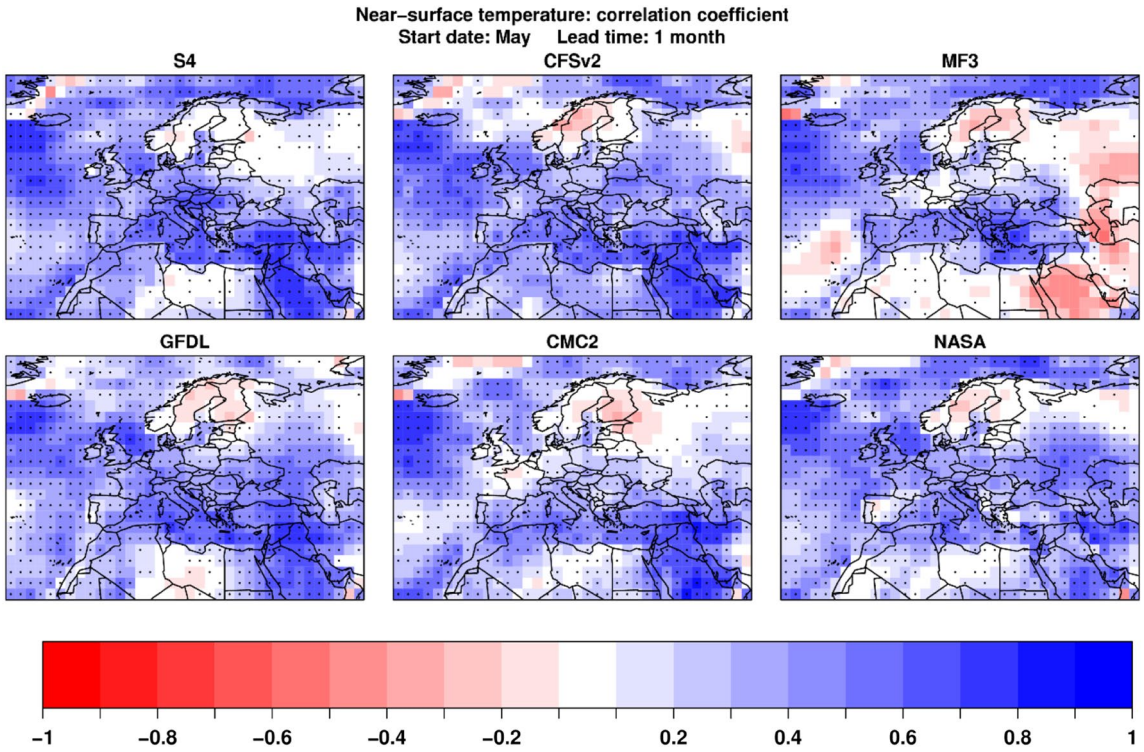

Fig. 1 Correlation coefficient for the June near-surface temperature ensemble-mean predictions initialized in the previous May (1-month lead time) for the hindcast period 1982–2010. The dots are placed where the correlation coefficient is statistically significantly larger than zero at the 5% level using a non-parametric bootstrap procedure. See text for details

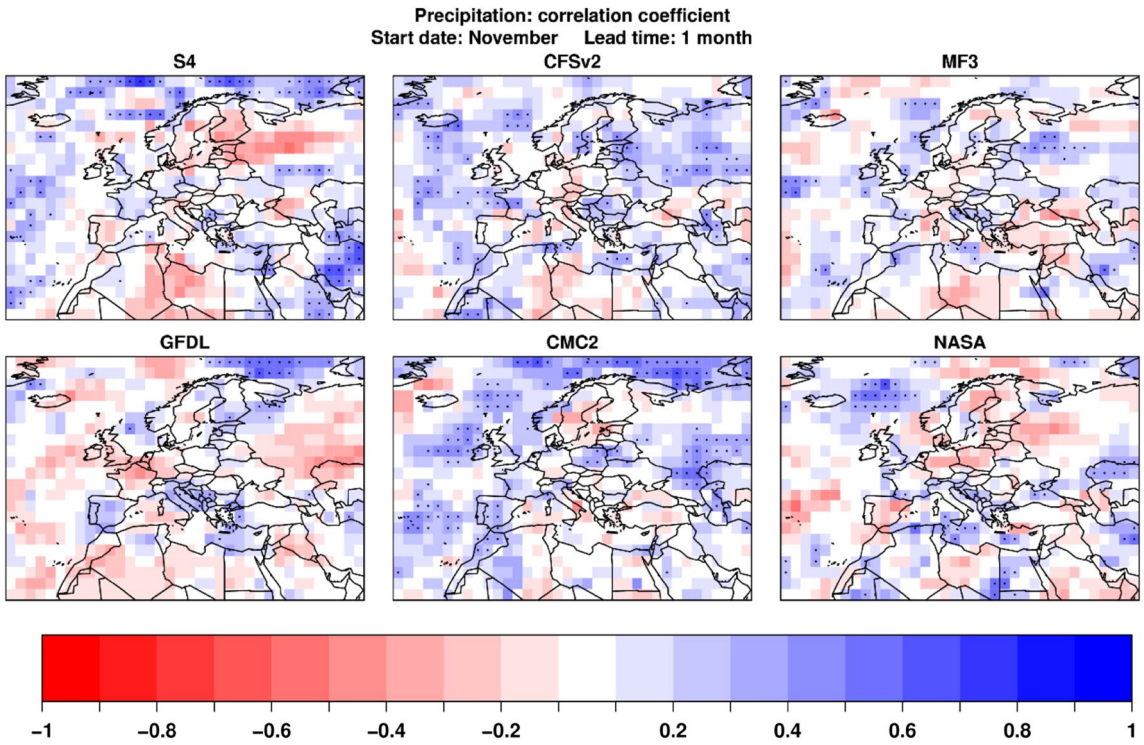

Fig. 2 As Fig. 1, but for the December precipitation ensemble-mean predictions initialized in the previous November (1-month lead time)

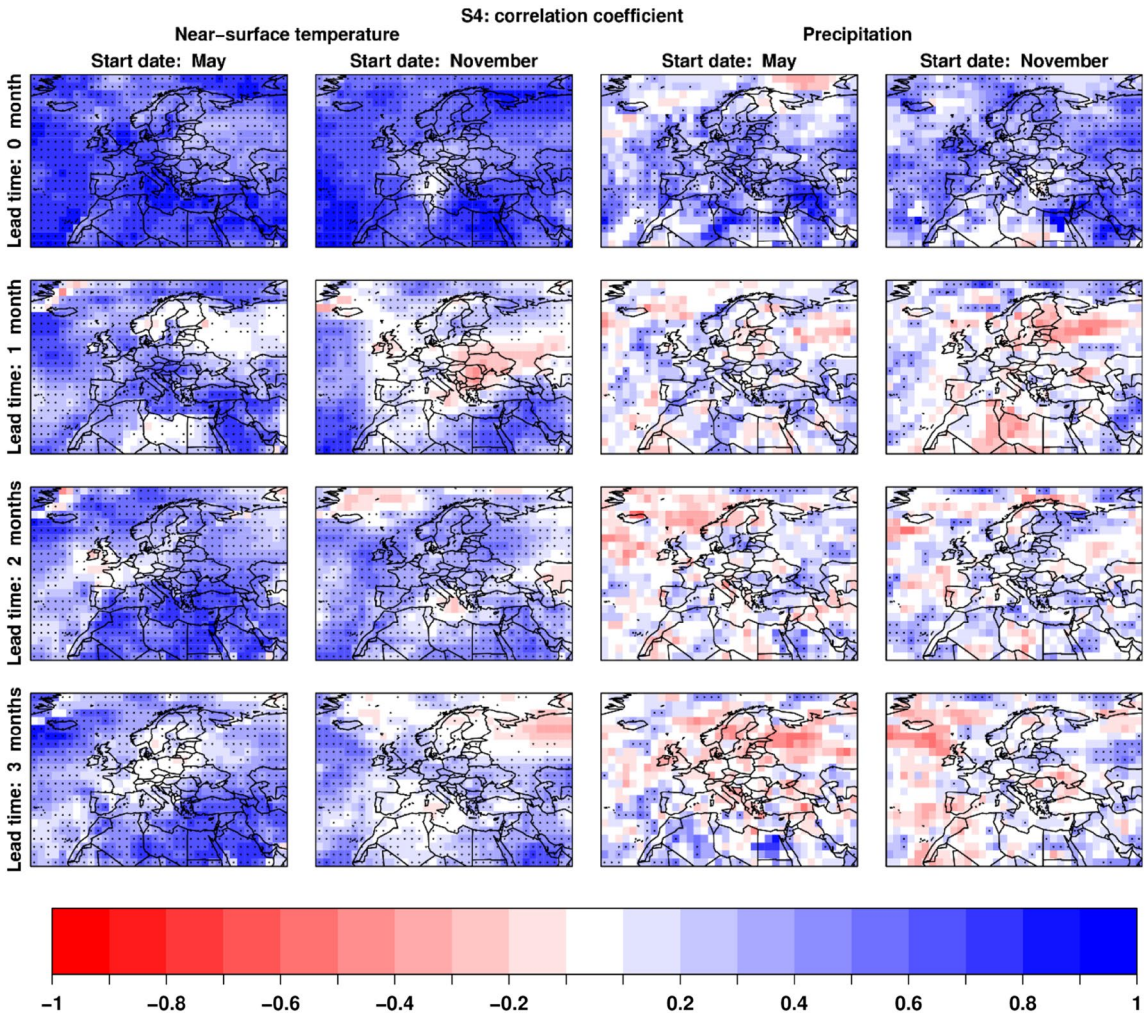

Fig. 3 Correlation coefficient of near-surface temperature (two left columns) and precipitation (two right columns) for S4 ensemble-mean predictions. Predictions are initialized in May (first and third columns) and November (second and fourth columns) and valid for zero to 3-month lead time (from top to bottom) for the hindcast period 1982–2010. The dots are placed where the correlation coefficient is statistically significantly larger than zero at the 5% level using a non-parametric bootstrap procedure. See text for details

individually because this is the most robust way to estimate the predicted trend. All forecast systems experienced a generalized reduction in the correlation compared to the ones displayed in Fig. 1, especially over continental Europe and the Mediterranean region, although statistically significantly positive values remained in North Atlantic Ocean and Middle-East. This shows that although the forecast systems have skill in predicting the June near-surface temperature, much of the large positive correlation coefficient values described in Fig. 1 were due to both predictions and observations having a linear trend.

The correlation coefficient for the December precipitation predictions initialized in the previous November display a different picture (Fig. 2): only a small fraction of the forecast systems has statistically significantly positive correlation coefficients. They are mostly located over the North Atlantic Ocean. The lack of prediction skill in extra-tropical regions has been documented in previous studies (Doblas-Reyes et al. 2000; Wang et al. 2009; Arribas et al. 2011; Kim et al. 2012).

Figure 3 illustrates how the correlation coefficient evolves with the lead time for the S4 near-surface temperature and precipitation predictions initialized in May and November from lead time 0–3 months. S4 was chosen for the illustration because it was often the most skillful forecast system (not shown). For both climate variables, the correlation coefficient decreases rapidly from lead time 0–1 month. However, the number of grid points with statistically significantly positive correlation coefficient for precipitation almost vanishes in continental areas at lead times 1, 2 and

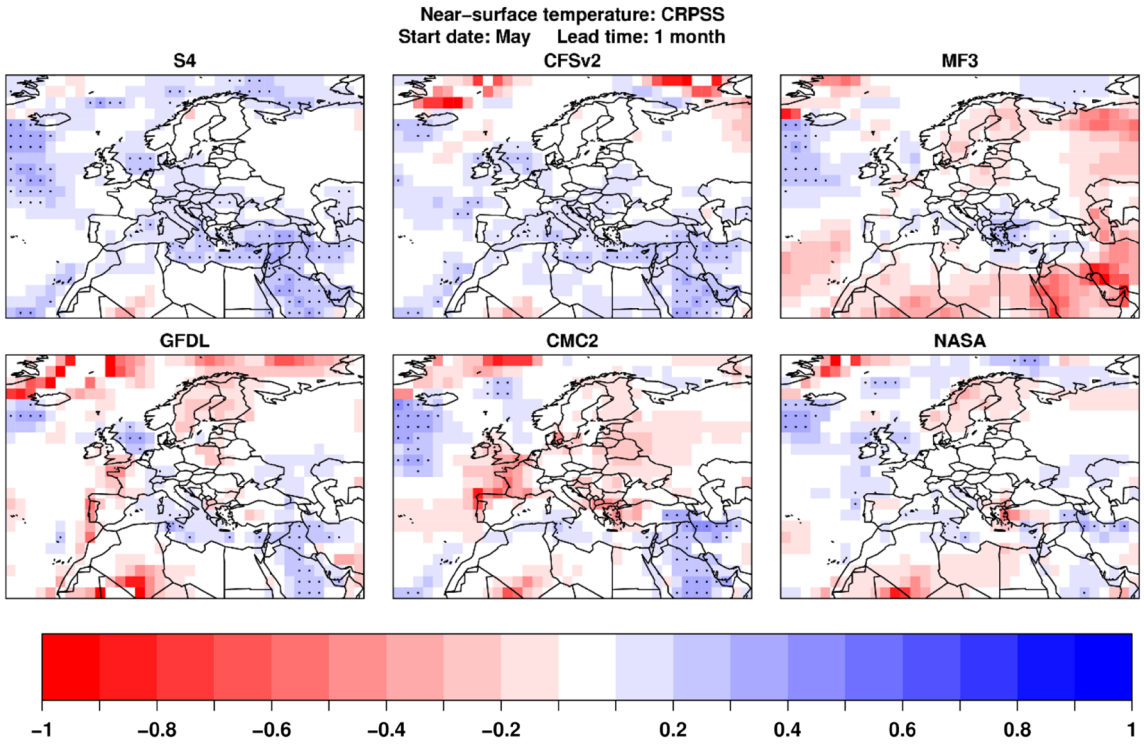

Fig. 4 As Fig. 1, but for the CRPSS

3 months. Some of the other five forecast systems do not display statistically significantly positive precipitation correlation coefficient over continental Europe even at 0-month lead time (not shown). On the other hand, there are statistically significantly positive near-surface temperature correlations over many grid points at the four lead times analyzed, especially over the ocean (first and second columns of Fig. 3). In addition, near-surface temperature prediction skill does not always decrease linearly with lead time, as the skill in boreal winter for 1-month lead time is lower compared to 2 and 3-months lead time (second columns of Fig. 3). This might be because prediction skill at each month is associated with different atmospheric phenomena (Barnett and Preisendorfer 1987) or because the role of the trend changes with the calendar month.

The CRPSS estimated for the June near-surface temperature predictions initialized in the previous May (1-month lead time) shows that the forecast systems are outperformed by the climatological forecast in several instances (Fig. 4). Except for a few grid points, none of the forecast systems displays statistically significantly positive CRPSS grid points over continental Europe and only S4 and CFSv2 display more positive than negative CRPSS areas over the studied region. Even for near-surface temperature predictions initialized in and valid for May, only half of the six forecast systems (S4, CFSv2 and NASA) have more positive than negative CRPSS grid points over continental Europe, similar to predictions initialized in May and valid for July and August as well as for predictions initialized in November and valid for the winter months (not shown). Only S4 and CFSv2 have positive precipitation CRPSS at zero-month lead time, both in summer and winter. However, these values go to zero or take negative values at lead times 1, 2 and 3 months.

It should be borne in mind that the CRPSS is a more stringent skill measure than the correlation coefficient because it requires not only that the predicted signal has the correct sign, but also that the uncertainty is correctly predicted (the correlation coefficient is not sensitive to errors in the predicted interval), which is usually not the case in current forecast systems. In this sense, the CRPS penalizes predictions that are either over-dispersive (i.e. prediction interval is wide) or under-dispersive (i.e. prediction interval is narrow) and unreliable (i.e. observations fall often outside the prediction interval). This is especially challenging when the reference is a perfectly reliable climatological prediction (Arribas et al. 2011), since a good CRPS occur when predictions are both under-dispersive and reliable. Therefore, red spots are observed in several grid points in most panels of Fig. 4, being S4 the only exception. Two reasons could explain these red spots: while the CFSv2 ensemble system presents over-dispersion in some situations, as for the grid points north of Iceland, the MF3, GFDL, CMC2 and NASA ensemble systems are under-dispersive and unreliable. Penalties

are especially strong for ensemble systems that have a small ensemble size because they are directly proportional to the probability densities, which are in turn inversely proportional to the number of ensemble members (e.g. an ensemble system with 10 members will be able to resolve probability density intervals of 1/10 whereas one with 51 members will distinguish probability density intervals of 1/51).

4 FA predictions

4.1 Modes of covariability

As described previously, the first step to combine predictions produced by several forecast systems using the FA method is to apply a dimension reduction technique on the gridded data (Stephenson et al. 2005; Coelho et al. 2006). This is necessary because the number of grid points in a multimodel ensemble is much higher than the number of hindcast years and the dependency between values at neighboring grid points (Stephenson et al. 2005). The MCA was used to reduce the data dimensionality on the two climate variables (precipitation and near-surface temperature), two start dates (May and November) and four lead times (from 0 to 3 months). The large amount of cases makes a detailed description of every single one unfeasible. Therefore, only one case, which corresponds to the June near-surface temperature covariability modes for predictions initialized in the previous May (one-month lead time), will be described. This case will illustrate that the FA method is able to produce skillful predictions when the forecast systems can predict properly the leading covariability modes, but poorly otherwise.

The MCA modes will be shown below as the heterogeneous correlation maps computed by correlating the expansion coefficients of the left field (i.e. the observation) with the original data of the right field (i.e. the predictions) and vice versa. Thus, all fields will have values between -1.0 and 1.0 making a comparison between predictions and observations easier. For instance, one can easily identify that current forecast systems overestimate the positive near-surface temperature trend in June over the Mediterranean Sea and Northern Africa (first column of Fig. 5). In addition, the correlation coefficient between the expansion coefficients (time series) associated with the MCA modes and several climate indices were computed to quantify their association with the European climate variability. The teleconnection patterns were: NAO, Artic Oscillation (AO), East Atlantic (EA), East Atlantic/Western Russia (EAWR), Scandinavian and Polar/Eurasian (PE). These climate indices summarize in a single number the main patterns of climate variability that affect the European climate.

The first observed June near-surface temperature covariability mode displays positive values over most of the area, except for a few grid points in the Eurasian region (top left panel of Fig. 5). The expansion coefficient associated with this mode (red line of the left panel of Fig. 6) has null or near null correlation coefficient with all physical teleconnection patterns listed above. Similarly, S4, CFSv2, GFDL, CMC2 and NASA (first column of Fig. 5) simulate widespread positive values in the region, but all these systems overestimate the magnitude of the values compared to observations. This mode accounts for 51% of the squared covariance between the observations and the predictions at 1-month lead time.

The second observed near-surface temperature MCA mode, which accounts for 19% of the squared covariance, displays a strong dipole with positive correlation coefficient values over central Europe and negative values centered over Russia (top central panel of Fig. 5). The expansion coefficient associated with this covariability mode (red line of the central panel of Fig. 6) has a low correlation coefficient with the NAO (0.31), AO (0.34) and EA (0.40), and near null correlation coefficient with all other analyzed teleconnection indices. These three climate indices were correlated with the observed grid point near-surface temperature anomalies for the period 1982–2010. The correlation coefficient between EA and the near-surface temperature anomalies (not shown) exhibited a pattern like the second MCA mode displayed in the top central panel of Fig. 5. When analyzing the predictions, we found that all forecast systems underestimate the magnitude of this MCA mode. Besides, most of them fail to emulate the regional patterns from the observational reference, especially the strong negative values centered over Russia (second column of Fig. 5). S4 is an exception to this by simulating a dipole similar to the one in the observations, although the positive values over central Europe are shifted southward whereas the negative values over Russia are shifted northward.

The third observed mode (top right panel of Fig. 5) shows a pattern with positive values over most of Europe and the Norwegian Sea and negative values over some parts of the North Atlantic Ocean, northern Africa, and Middle East. This mode accounts for 10% of the squared covariance. None of the climate indices described above has correlation coefficient higher than 0.3 with the expansion coefficient associated with this MCA mode. All forecast systems capture the negative values over the North Atlantic Ocean, but most of them fail to predict the pattern over the continents. The similarities between the predicted and observed modes are partly constrained by the need of the MCA modes to be orthogonal to each other, so that it is increasingly difficult for the modes beyond the first one to bear similarities.

The expansion coefficients associated with the first three observed covariability modes in summer months (MJJA) are similar (Fig. 6 is an illustration for June). However, the

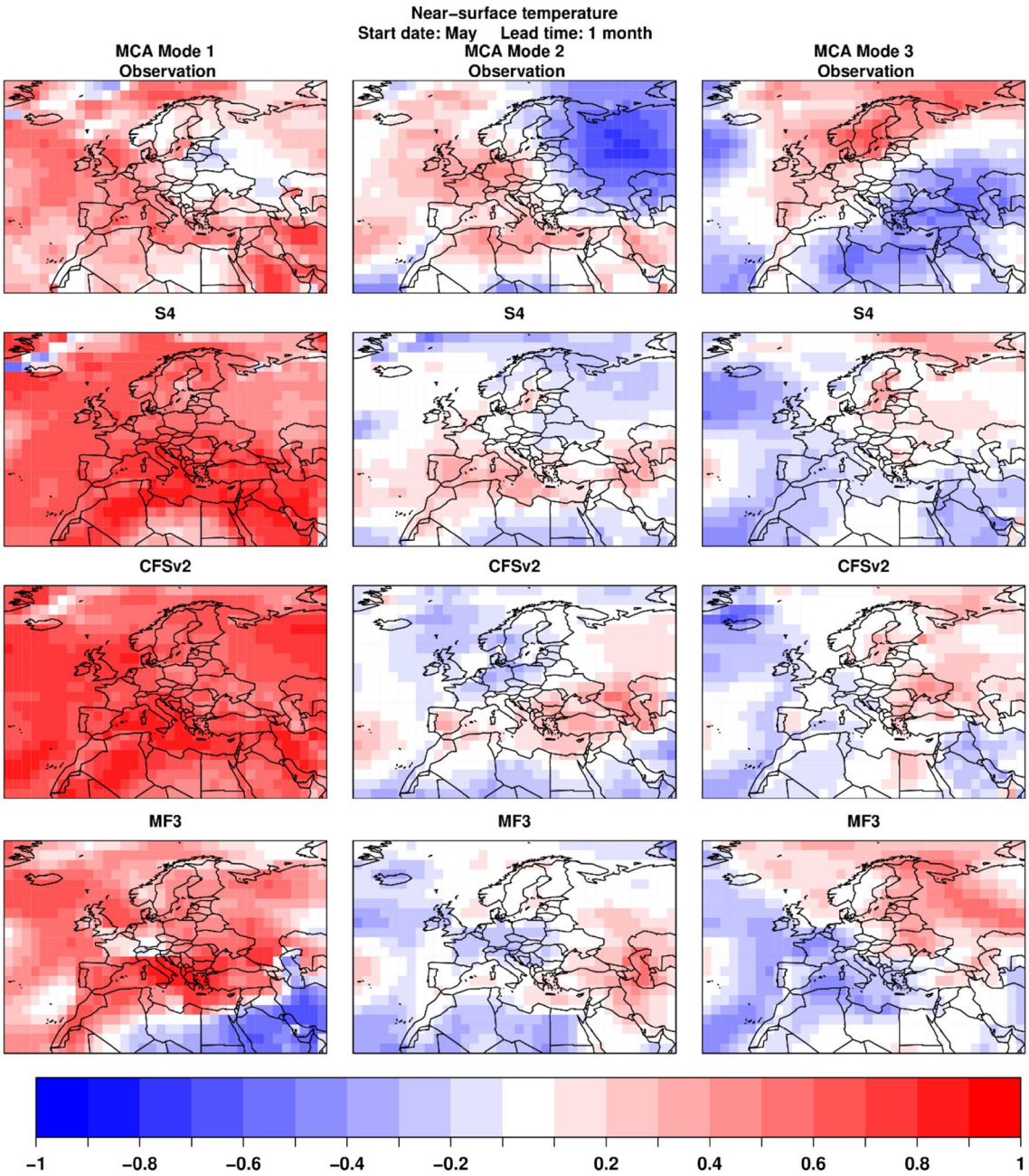

Fig. 5 Heterogeneous correlation maps for the observed and predicted June near-surface temperature. Predictions are initialized in the previous May (1-month lead time) for the hindcast period 1982–2010. The expansion coefficients of the left field (i.e. the observation) are correlated with the original data of the right field (i.e. the forecast systems) and vice versa. Results are shown for the first three leading MCA modes

covariance explained by each mode varies from May (predictions at zero-month lead time) to August (predictions at 3-month lead time): while the first three modes account respectively for 33%, 24% and 17% of the squared covariance in May (the three modes together account for 74%), these numbers change to 66%, 10% and 7%, respectively, in August (the three modes accounts for 83%). That is, not only there is an increase in the total squared covariance accounted for the first three modes with lead time (from 74 to 83%), but also the share of the first mode alone doubles (from 33 to 66%) while the share of the second and third modes summed up are significantly reduced (from 41 to 17%). This makes sense as most of the forecast systems consistently increase the positive values of the first covariability mode with lead time (not shown), which represents the trend that

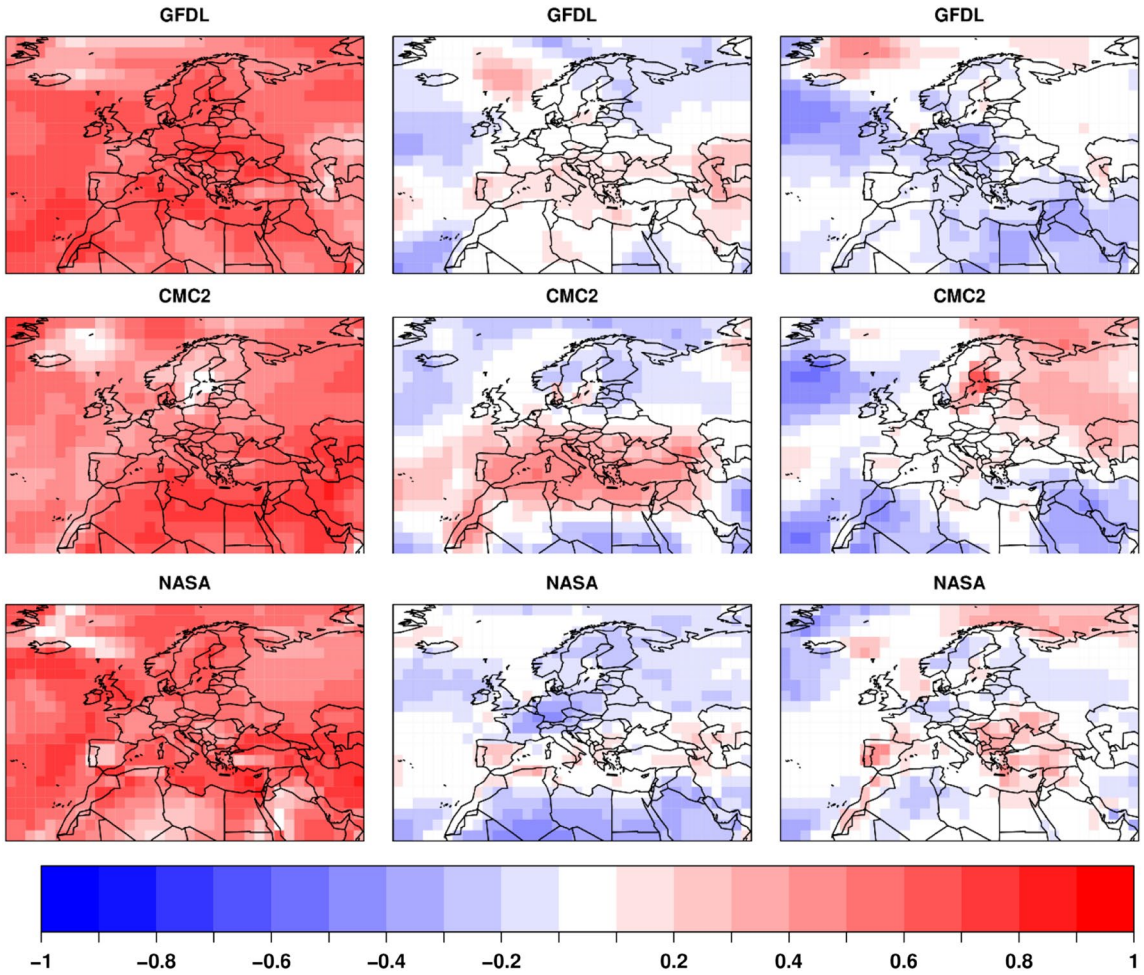

Fig. 5 (continued)

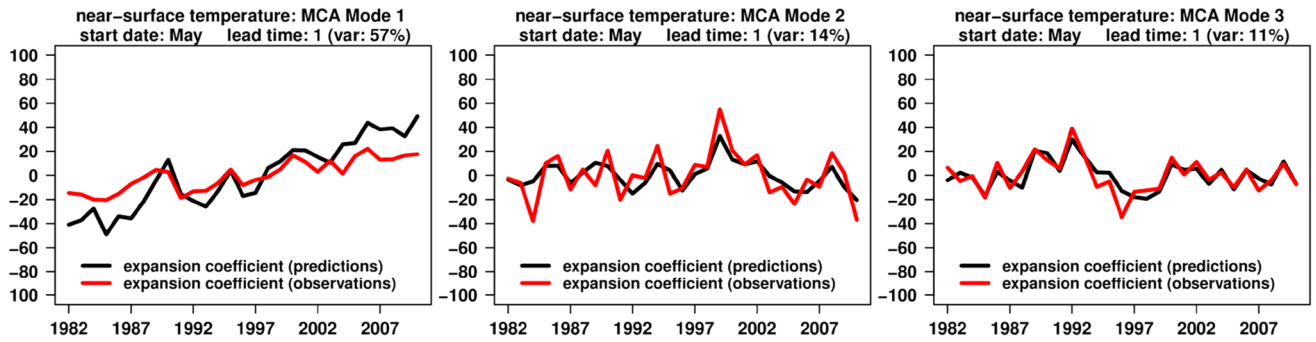

Fig. 6 First three expansion coefficients of the left (observations, red line) and right (predictions, black line) fields for the June near-surface temperature. Predictions are initialized in the previous May (1-month lead time)

takes an increasingly larger role in the relationship between the observations and the predictions.

The expansion coefficients associated with the covariability modes described above were used to estimate the parameters needed to combine several models using the FA method. Figure 6 illustrates the expansion coefficients associated with the first three June near-surface temperature covariability modes. The expansion coefficient associated with the first mode displays a positive trend over the studied region, which is spatially consistent as most of

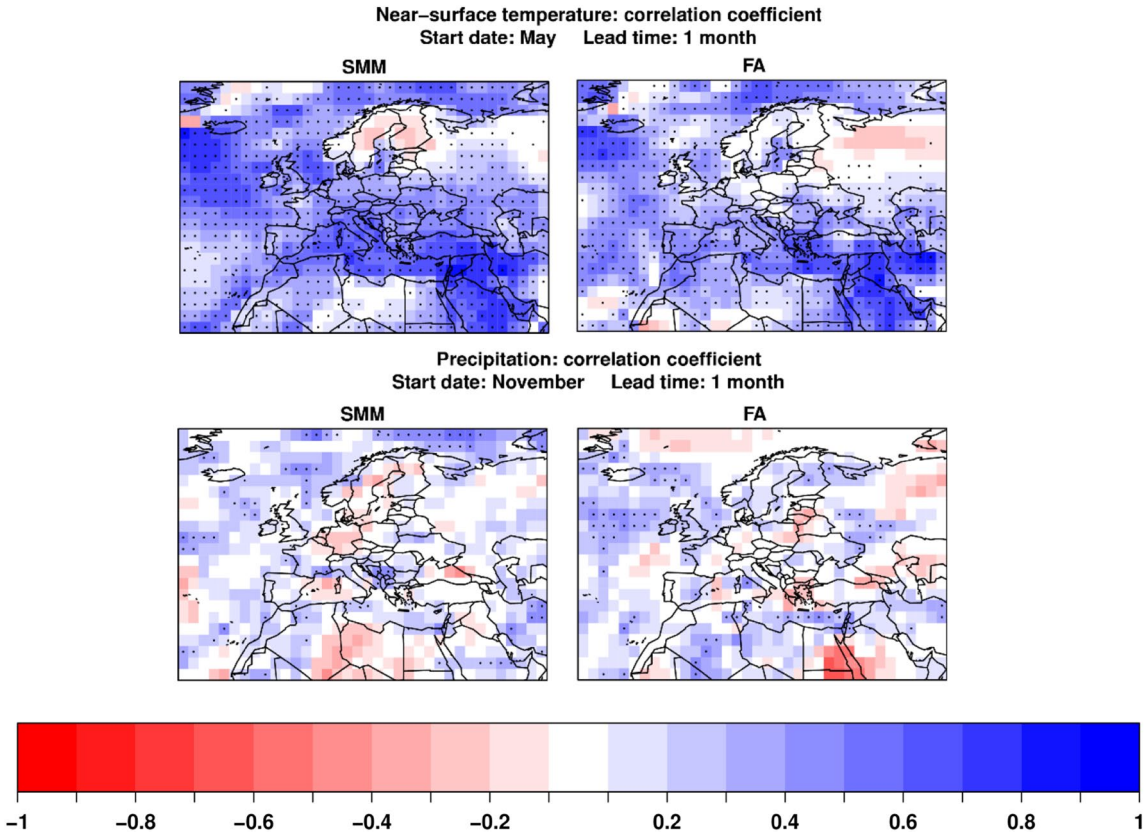

Fig. 7 As Fig. 1 (first row) and Fig. 2 (second row), but for the multimodel ensembles

the grid points have the same sign (first column of Fig. 5). This is consistent with the well-known observed warming trend over Europe in summer months (EEA 2015). On the other hand, both the second and the third expansion coefficients are dominated by interannual variability.

4.2 Forecast quality of combined and calibrated predictions

This section discusses the forecast quality of the SMM and FA combination methods. The correlation coefficient displays an overall similar pattern for the June near-surface temperature predictions initialized in the previous May (1-month lead time) and the December precipitation predictions initialized in the previous November (1-month lead time) with both methods. On the other hand, local differences can be found in some regions (Fig. 7). For instance, FA outperforms SMM when predicting the June near-surface temperature over the Scandinavian Peninsula, but not over the Eurasian region (upper panels of Fig. 7). The weak performance of the FA over the Eurasian region was linked to the failure of most forecast systems in predicting the weak negative (strong positive) values in the first (second) observed near-surface temperature MCA mode (first and second columns of Fig. 5). More specifically, the large (small) negative area of the top right panel of Fig. 7 over central Russia (western edge of Russia/Baltic countries) coincides with the strong positive (weak negative) values of the top central (left) panel of Fig. 5. As expected, the FA method works in regions where predictions used in the combination correctly emulate the sign and magnitude of the observed covariability modes.

Much of the positive correlation coefficient values described in Figs. 1 and 7 might be explained because the predicted and observed linear trend had the same sign and not necessarily due to the ability of the forecast systems to predict the interannual near-surface temperature covariability. Therefore, SMM and the FA methods were also estimated using linearly detrended data to measure the correlation coefficient sensitivity to the trend (not shown). We found that although the spatial correlation coefficient patterns were generally similar to the original data in both methods, there was a substantial reduction in the magnitude of the correlation values and statistically significantly positive correlation coefficient values disappeared in several grid points in western Europe. Additionally, the CRPSS for the June near-surface temperature probabilistic predictions initialized in the previous May (1-month lead time) illustrates

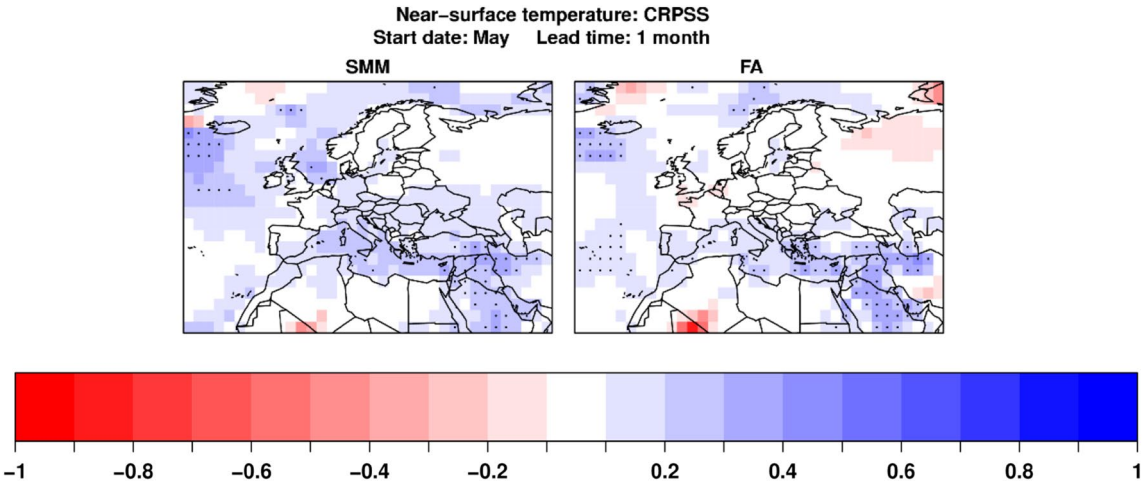

Fig. 8 As Fig. 4, but for the multimodel ensembles

the similarities in skill patterns but with differences in the magnitude of the values (upper panels of Figs. 7, 8). In the latter case, the FA combination presents statistically significantly positive CRPSS over the North Atlantic Ocean, Mediterranean Sea and Middle East, an overall similar spatial pattern compared to the SMM. The performance of the FA method is closely linked to the required dimension reduction explained in the previous section.

The number of grid points with negative CRPSS for the June near-surface temperature with 1-month lead time is substantially reduced in both combination methods (Fig. 8) compared to all single forecast systems (Fig. 4), except S4. However, only modest improvement is found when comparing only to S4 in this illustration. For instance, SMM has higher CRPSS (although not statistically significant positive) than S4 over eastern Ukraine, the Black Sea and Southwestern Russia. This shows that a multimodel approach is not necessarily the best prediction strategy in all cases (Rodrigues et al. 2014b). On the other hand, the best single forecast system is not always the same (e.g. CFSv2 is the best one in Fig. 2). Therefore, evaluate different multimodel strategies, especially considering operational forecast systems and objective methods that eliminates or downweights predictions from unskillful forecast system, might be of great value to users of climate information.

5 Summary and conclusions

Monthly near-surface temperature and precipitation predictions produced by several forecast systems were calibrated and combined using the Forecast Assimilation (FA) method. The FA method was then compared to the simple multimodel ensemble (SMM). The FA method calibrates and combines predictions from multiple sources by assigning unequal weights based on their forecast quality over the historical period. In addition, differently from other combination methods that assigns unequal weights, the FA method attempts to handle the high dimensionality of a spatial multimodel ensemble field (i.e. number of grid points times the number of forecast systems) compared to the number of independent hindcast years and the strong dependency between values of neighboring grid points (Stephenson et al. 2005). This study focused on Europe, a region where prediction skill is low and where the need for improvement of forecast quality has been expressed in the context of the developing climate services. Predictions initialized in May and November with lead times up to 4 months aimed to cover the boreal summer (MJJA) and winter (NDJF) months. Another important aspect of this study was to use six forecast systems that belong to two international MME efforts (EUROSIP and NMME), most of them publicly available.

Firstly, a forecast quality assessment was performed on the single forecast systems using a deterministic and a probabilistic forecast quality measure. We show that the forecast systems have higher skill when predicting near-surface temperature than precipitation. In addition, the near-surface temperature predictions have higher correlation in summer (MJJA) than in the winter months (NDJF), which may be partly explained because predictions and observations have the same long-term linear trend sign. The correlation decreased over many grid points when the trend is removed from the predictions and observations, although the positive correlation values prevailed. The CRPSS shows lower values than the correlation because it evaluates not only the predicted mean but also the predicted uncertainty, which can be expressed by a prediction interval. This is specially challenging when the reference climatological prediction is perfectly reliable (Arribas et al. 2011). Therefore, even if the forecast systems are able to predict the correct sign of the

near-surface temperature trend they might still show negative CRPSS if they do not estimate correctly the forecast uncertainty. Yet statistically significantly positive CRPSS was found over several grid points for near-surface temperature predictions, but not for precipitation predictions. This shows that despite improvements in the forecast systems formulation, they still have difficulties to produce skillful precipitation predictions in extratropical regions.

In the second part of the study, we show that the FA method generally presents less skillful predictions than the SMM over Europe. The reduced FA skill is associated with the way it accounts for the high dimensionality of the multimodel gridded data and the strong dependency between values of neighboring grid points (Stephenson et al. 2005; Coelho et al. 2006), a feature not accounted for in other multimodel studies (Robertson et al. 2004; Doblas-Reyes et al. 2005). This is done by applying the MCA on the cross-covariance of the observations and predictions to reduce the multimodel ensemble into a few modes that explain most of the covariability. An illustration of how the predicted covariability modes affect the FA skill can be seen by comparing Fig. 5 with Figs. 7 and 8. In this illustration, large area with negative correlation coefficient (right hand columns of Fig. 7) and CRPSS (right hand column of Fig. 8) over Russia is found in a region where most forecast systems do not predict the correct sign of the anomalies in the first and second observed covariability mode (first row of Fig. 5). On the other hand, in agreement with previous studies (Coelho et al. 2006), the FA have skill and can even beat the SMM in regions where the covariability modes are well predicted by the forecast systems. In the above illustration, statistically significant positive correlation coefficient and CRPSS are found in parts of Scandinavian peninsula, North Atlantic, the Mediterranean Sea and Middle-East (right hand columns of Figs. 7, 8), regions where most forecast systems produce a spatial pattern similar to the observations (Fig. 5).

Seasonal forecast systems have predictive skill over South America mainly because they are able to represent the impact of the slowly varying components of the climate system on the atmospheric surface climate variables, especially linked to the ENSO teleconnections. In this situation, the FA method works as well as or better than the SMM (Coelho et al. 2006). However, as illustrated previously, the forecast systems were not able to reproduce the main observed near-surface temperature and precipitation covariability modes over Europe and adjacent regions in a joint MCA. Therefore, the FA method might not be the best combination strategy for the European climate.

Acknowledgements The authors thank NOAA, NCEP, IRI and NCAR personnel in creating, updating and maintaining the NMME archive. The NMME project and data dissemination is supported by NOAA, NSF, NASA and DOE. Météo-France and ECMWF are appreciated for making available their seasonal prediction hindcasts. This study was supported by the Seventh Framework Programme SPECS project (contract 308378) and the H2020 EUCP project (contract 776613). CASC was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) processes 304586/2016-1. LRLR and CASC acknowledge the support of FAPESP, process 2015/50687-8 (CLIMAX project). The authors acknowledge two anonymous reviewers for their useful comments and suggestions.

References

Alessandri A, Borrelli A, Navarra A, Arribas A, Déqué M, Rogel P, Weisheimer A (2011) Evaluation of probabilistic quality and value of the ENSEMBLES multimodel seasonal forecasts: comparison with DEMETER. Mon Weather Rev 139:581–607

Arribas A, Glover M, Maidens A, Peterson K, Gordon M, MacLachlan C, Graham R, Fereday D, Camp J, Scaife AA, Xavier P, McLean P, Colman A, Cusack S (2011) The GloSea4 ensemble prediction system for seasonal forecasting. Mon Weather Rev 139:1891–1910

Barnett TP, Preisendorfer R (1987) Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. Mon Weather Rev 115:1825–1850

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

Coelho CAS, Stephenson DB, Balmaseda M, Doblas-Reyes FJ, van Oldenborgh GJ (2006) Toward an integrated seasonal forecasting system for South America. J Clim 19:3704–3721

Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Holm EV, Isaksen L, Kallberg P, Kohler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette J-J, Park B-K, Peubey C, Rosnay P, Tavolato C, Thepaut J-N, Vitart F (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q J R Meteorol Soc 137:553–597

Doblas-Reyes FJ, Déqué M, Piedelievre JP (2000) Multi-model spread and probabilistic seasonal forecasts in PROVOST. Q J R Meteorol Soc 126:2069–2087

Doblas-Reyes FJ, Hagedorn R, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting. Part 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. WIREs Clim Change 4:245–268

Eden JM, van Oldenborgh GJ, Hawkins E, Suckling EB (2015) A global empirical system for probabilistic seasonal climate prediction. Geosci Model Dev Discuss 8:3941–3970

EEA (2015) European Environment Agency: Global and European temperatures (CSI 012/CLIM 001) Assessment. http://www.eea.europa.eu/data-and-maps/indicators/global-and-european-temperature-1/assessment. Accessed 7 Aug 2015

Gneiting T, Raftery AE, Westveld AH, Goldman T (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon Weather Rev 133:1098–1118

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

Graham RJ, Gordon M, McLean PJ, Ineson S, Huddleston MR, Davey MK, Brookshaw A, Barnes RTH (2005) A performance comparison of coupled and uncoupled versions of the Met Office seasonal prediction general circulation model. Tellus A 57:320–339

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

Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast 15:559–570

Huffman GJ, Bolvin DT (2013) GPCP Version 2.2 combined precipitation data set documentation, internet publication, pp 1–46. http://www1.ncdc.noaa.gov/pub/data/gpcp/gpcp-v2.2/doc/V2.2_doc.pdf. Accessed 16 Nov 2012

Johansson A, Barnston A, Saha S, van den Dool H (1998) On the level and origin of seasonal forecast skill in northern Europe. J Atmos Sci 55:103–127

Jolliffe IT, Stephenson DB (2012) Forecast verification: a practitioner's guide in atmospheric science, Second edn. Wiley, Chichester

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

Kirtman BP, Min D, Infanti JM, Kinter JL III, Paolino DA, Zhang Q, van den Dool H, Saha S, Mendez MP, Becker E, Peng P, Tripp P, Huang J, DeWitt DG, Tippett MK, Barnston AG, Li S, Rosati A, Schubert SD, Rienecker M, Suarez M, Li ZE, Marshak J, Lim Y-K, Tribbia J, Pegion K, Merryfield WJ, Denis B, Wood EF (2014) The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bull Am Meteorol Soc 95:585–601

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 Academic Publishers, Dordrecht, pp 167–206

Mason SJ, Goddard L, Graham NE, Yulaeva E, Sun L, Arkin PA (1999) The IRI seasonal climate prediction system and the 1997/98 El Niño Event. Bull Am Meteorol Soc 80:1853–1873

Merryfield WJ, Lee W-S, Boer GJ, Kharin VV, Scinocca JF, Flato GM, Ajayamohan RS, Fyfe JC, Tang Y, Polavarapu S (2013) The canadian seasonal to interannual prediction System. part i: models and initialization. Mon Weather Rev 141:2910–2945

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 Technical Memorandum 656. http://www.ecmwf.int/publications/library/do/references/list/14. Accessed 20 Dec 2012

Robertson AW, Lall U, Zebiak SE, Goddard L (2004) Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. Mon Weather Rev 132:2732–2744

Rodrigues LRL, Doblas-Reyes FJ, Coelho CAS (2014a) Multi-model calibration and combination of tropical seasonal sea surface temperature forecasts. Clim Dyn 42:597–616

Rodrigues LRL, García-Serrano J, Doblas-Reyes FJ (2014b) Seasonal forecast quality of the West African monsoon rainfall regimes by multiple forecast systems. J Geophys Res 119:7908–7930

Saha S, Moorthi S, Wu X, Wang J, Nadiga S, Tripp P, Behringer D, Hou Y-T, Chuang H, Iredell M, Ek M, Meng J, Yang R, Mendez MP, van den Dool H, Zhang Q, Wang W, Chen M, Becker E (2014) The NCEP climate forecast system version 2. J Clim 27:2185–2208

Scaife AA, Arribas A, Blockley E, Brookshaw A, Clark RT, Dunstone N, Eade R, Fereday D, Folland CK, Gordon M, Hermanson L, Knight JR, Lea DJ, MacLachlan C, Maidens A, Martin M, Peterson AK, Smith D, Vellinga M, Wallace E, Waters J, Williams A (2014) Skillful long-range prediction of European and North American winters. Geophys Res Lett 41:2514–2519

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

Stockdale TN, Molteni F, Ferranti L (2015) Atmospheric initial conditions and the predictability of the Arctic Oscillation. Geophys Res Lett 42:1173–1179

Vernieres G, Rienecker MM, Kovach R, Keppenne CL (2012) The GEOS-iODAS: Description and evaluation. NASA Technical Report Series on Global Modeling and Data Assimilation TM2012-104606 30

Vitart F, Huddleston MR, Déqué M, Peake D, Palmer TN, Stockdale TN, Davey MK, Ineson S, Weisheimer A (2007) Dynamically-based seasonal forecasts of Atlantic tropical storm activity issued in June by EUROSIP. Geophys Res Lett 34:L16815

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:L1340

Zhang S, Harrison MJ, Rosati A, Wittenberg A (2007) System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. Mon Weather Rev 135:3541–3564