Weighting of NMME temperature and precipitation forecasts across Europe

Louise J. Slater[1,2], Gabriele Villarini[1], and A. Allen Bradley[1]

[1]IIHR-Hydroscience & Engineering, The University of Iowa, Iowa City, Iowa, USA

[2]Department of Geography, Loughborough University, Loughborough, UK

Corresponding author: Louise J. Slater (l.slater@lboro.ac.uk)

Abstract

Multi-model ensemble forecasts are obtained by weighting multiple General Circulation Model (GCMs) outputs to heighten forecast skill and reduce uncertainties. The North American Multi-Model Ensemble (NMME) project facilitates the development of such multi-model forecasting schemes by providing publicly-available hindcasts and forecasts online. Here, temperature and precipitation forecasts are enhanced by leveraging the strengths of eight NMME GCMs (CCSM3, CCSM4, CanCM3, CanCM4, CFSv2, GEOS5, GFDL2.1, and FLORb01) across all forecast months and lead times, for four broad climatic European regions: Temperate, Mediterranean, Humid-Continental and Subarctic-Polar. We compare five different approaches to multi-model weighting based on the equally weighted eight single-model ensembles (EW-8), Bayesian updating (BU) of the eight single-model ensembles (BU-8), BU of the 94 model members (BU-94), BU of the principal components of the eight single-model ensembles (BU-PCA-8) and BU of the principal components of the 94 model members (BU-PCA-94). We assess the forecasting skill of these five multi-models and evaluate their ability to predict some of the costliest historical droughts and floods in recent decades. Results indicate that the simplest approach based on EW-8 preserves model skill, but has considerable biases. The BU and BU-PCA approaches reduce the unconditional biases and negative skill in the forecasts considerably, but they can also sometimes diminish the positive skill in the original forecasts. The BU-PCA models tend to produce lower conditional biases than the BU models and have more homogeneous skill than the other multi-models, but with some loss of skill. The use of 94 NMME model members does not present significant benefits over the use of the 8 single model ensembles. These findings may provide valuable insights for the development of skillful, operational multi-model forecasting systems.

Keywords

Temperature; Precipitation; NMME; Forecasts; Multi-models; Europe

1. Introduction

In recent decades there has been growing interest in leveraging the skill of forecasts from multiple Global Circulation Models (GCMs) to improve climate predictions (e.g., Hagedorn et al., 2005; Weigel et al., 2008). Early multi-model projects such as the Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER) or the Ensemble-Based Predictions of Climate Changes and their impacts (ENSEMBLES) project provided GCM hindcasts (i.e., model forecasts that are produced by running the models in the past) to facilitate the development of multi-model weighting schemes based on the strengths and weaknesses of the different models. More recent international schemes like the North American Multi-Model Ensemble (NMME) and the operational European Seasonal-to-Interannual Prediction (EuroSIP) projects also provide near-real time forecasts to allow the development of multi-model forecasting applications (Kirtman et al., 2014).

The NMME is a collaborative forecasting system or ‘prediction experiment’ that began in 2011 (Kirtman et al., 2014), to which U.S. (NOAA/NCEP, NOAA/GFDL, IRI, NCAR, NASA) and Canadian (CMC) modeling centers (see **Table 1** for explanation of acronyms) contribute real time seasonal-to-interannual predictions. The NMME is based on the recognition that multi-model ensemble approaches generate better forecasts than any single model ensemble (e.g., Doblas-Reyes et al., 2005, Hagedorn et al., 2005, Kirtman and Min, 2009).

Before developing any multi-model ensembles, an important first step has been the evaluation of NMME model skill to understand the strengths and weaknesses of the different GCMs. Because of the large volumes of data that are produced within the NMME (**Table 1**), global-scale studies have focused on the evaluation of model skill at specific lead times (Becker et al., 2014; Mo and Lettenmaier, 2014), or for specific seasons (Wang, 2014), models (Jia et al., 2015; Saha et al., 2014), or climate quantities (Barnston and Lyon, 2016; Mo and Lyon, 2015). Regional evaluations of NMME forecast skill have focused principally on North America (Infanti and Kirtman, 2016), the United States (Misra and Li, 2014; Roundy et al., 2015; Slater et al., 2017), the southeastern United States (Infanti and Kirtman, 2014), but also China (Ma et al., 2015a, 2015b), Iran (Shirvani and Landman, 2016) and South Asia (Sikder et al., 2015). Thus, most of the effort of the NMME model skill evaluation has been over the USA, and far less attention has been paid to Europe, with some exceptions, such as Thober et al. (2015), who used NMME forecasts as input for the mesoscale hydrologic model (mHM).

Existing NMME multi-model approaches have mostly used equal weighting schemes, giving the same weight to each single-model ensemble (i.e., the mean of each model's members) or to all of the individual members, irrespective of their skill (Becker et al., 2014; Hagedorn et al., 2005; Slater et al., 2017; Tian et al., 2014). The predictive skill of these equally weighted multi-models tends to be greater than or equal to the skill of the best model within the ensemble (Becker et al., 2014; DelSole and Tippett, 2014; Hagedorn et al., 2005; Ma et al., 2015a; Slater et al., 2017; Thober et al., 2015; Wood et al., 2015). Generally, multi-model ensembles can outperform single-model ensembles when the individual models are overconfident, so the multi-model widens the ensemble spread and reduces the average ensemble-mean error (Weigel et al., 2008). However, the equal weights approach has limitations. First, it presumes that the models are independent, and so it accentuates the "region of model agreement" (Olson et al., 2016), assuming that the model biases will cancel out, and that the average forecast will be more skillful than that of any single-model ensemble (Knutti et al., 2010). If the models are not independent, the multi-model will over-strengthen the forecasts issued by similar models (Olson et al., 2016). This is particularly true in the case of the NMME, where many of the participating models are different versions of similar models, e.g., CCSM3 and CCSM4, CanCM3 and CanCM4, or GFDL2.1 and FLORb01 (**Table 1**), so the forecasts exhibit notable similarities (e.g., Slater et al., 2017). Another problem is that of reproducing the correct dispersion (Raftery et al., 2005): single-model ensembles are likely to be underdispersive (Arritt and Rummukainen, 2011), as are multi-model ensembles when the models are correlated among themselves. Multi-model averages are thus likely to impoverish the forecast signal (Knutti et al., 2010).

Overall, therefore, two of the main challenges in developing a solid multi-model approach are (1) to define an objective procedure that weights the contribution of each model based on historical performance, and (2) to eliminate the biases arising from models that perform similarly, because consolidation of information in multi-model approaches can only be better than the best individual model if the information is independent (Van den Dool, 2007).

To address the first of these aims, we use Bayesian updating (BU). Various approaches can be used to post-process ensemble forecasts based on their historical performance (e.g., Krishnamurti et al., 1999; Rajagopalan et al., 2002; Scheuerer and Büermann, 2014), but Bayesian schemes have gained increasing attention in recent years (e.g., Coelho et al. 2004; Hodyss et al., 2016) as they generally improve the sharpness of the forecasts and can be updated as new information becomes available. For example, Madadgar et al. (2016) developed a multivariate Bayesian model based on copula functions to predict drought as a function of atmosphere-ocean teleconnections and showed that the multi-model Bayesian forecasts performed considerably better than the initial NMME forecasts. In BU, each individual forecast adjusts the prior probability of the forecast variable, defined by the sample climatology of historical observations (Bradley et al., 2015). By expressing the observed values of the historic record in terms of their likelihood, given the forecasts made by each model, Bayesian approaches take full advantage of the historical record length. Thus, they circumvent one of the principal limitations of GCM forecasts, which is the shortness of the hindcast and forecast records.

To address the second challenge and reduce the multicollinearity and biases that may arise from including similar models within the ensemble, we propose a method based on principal components analysis (PCA). Instead of applying the BU approach to the single-model forecasts directly, we first compute the principal components among the available models, before conducting BU on the principal components. Thus, we aim to reduce any biases arising from model similarities and to simplify the Bayesian methodology by pooling together all of the single-model ensemble hindcasts (or the individual model member hindcasts).

This paper therefore describes an experiment to leverage the strengths of eight NMME models over the full range of forecast months and lead times by optimizing the available hindcast/forecast data following an approach based on BU of the climate forecasts. We aim to answer the following questions:

1) What is the skill of eight state-of-the-art NMME single-model ensembles in forecasting precipitation and temperature across Europe? Are they able to forecast extended periods of extreme temperature and extreme precipitation?

2) Can we develop a Bayesian approach for multi-model forecasting that leverages the strengths of the individual models, and reduces any biases and errors?

3) Does the Bayesian multi-model forecast improve when we use all of the 94 individual model members, instead of the eight single-model ensembles (based on mean values of the corresponding members)?

The remainder of the paper is organized as follows. **Section 2** describes the data and the European regions used in the study. **Section 3** describes the forecast verification metrics, the BU, the principal components approach, and the diagnosis of eight extreme precipitation and temperature events. **Section 4** describes and discusses the skill of the eight single-model ensembles, the EW-8 model, the BU models, the BU-PCA models, and compares the skill of all the multi-models in forecasting extreme events. Given the imperfect nature of the models and their strengths and weaknesses over different forecast months, lead times, and regions, Section 5 concludes by comparing the multi-models and discussing the best procedures for producing multi-model forecasts with optimized skill over longer lead times.

2 Data

2.1 NMME forecast temperature and precipitation data

The models and variables that are made available in the NMME are centralized in online repositories. We downloaded the data from IRI/Lamont Doherty Earth Observatory (LDEO) Climate Data Library (http://iridl.ldeo.columbia.edu/) in a netCDF format, on regular 1°×1° grids. We focus on eight single model ensembles, referred to as CCSM3, CCSM4, CanCM3, CanCM4, CFSv2, GEOS5, GFDL2.1 and FLORb01, and the 94 members of those models (see **Table 1** for model description and acronym definitions). The models have between 6 and 24 members each, and the forecasts are produced for varying lead times, ranging from 0.5 to 11.5 months (see caption of **Table 1** for a description of lead times).

Temperature and precipitation data were obtained for all model members and for all lead times, and tailored to the boundaries shown in **Figure 1**. The hindcast/forecast data for CFSv2, CanCM3 and CanCM4 were downloaded separately and combined. The netCDF files are five-dimensional, with longitude, latitude, lead, member, and forecast reference time.

2.2 Reference temperature and precipitation data and region outline

As reference data, we used observed temperature and precipitation data (E-OBS) from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) (Haylock et al., 2008; Hewitt and Griggs, 2004), which are provided through the ECA&D project (http://www.ecad.eu). We downloaded E-OBS v13 (June 2016 release) at a 0.25 × 0.25 degree resolution, and aggregated the data to 1°×1° grids to match the resolution and spatial extent of the NMME data. We then defined four European regions based on Köppen climate categories and tailored the region outlines to include only the grid cells where both NMME and E-OBS data were available (**Figure 1**).

3 Methods

3.1 Forecast verification

Forecast skill can be quantified using a variety of approaches. Here, we use the mean square error (MSE) skill score SS_{MSE} (e.g., Hashino et al., 2007) to assess the accuracy of the forecast relative to observed temperature and precipitation, because it allows us to evaluate the conditional and unconditional biases in the models separately. The MSE skill score can be written as

$$SS_{MSE} = 1 - \frac{MSE}{\sigma_x^2}, \qquad (1)$$

where σ_x represents the standard deviation of the observations. If the forecasts are probabilistic, rather than deterministic, then the SS_{MSE} is equivalent to a Brier skill score (Brier, 1950). A skill score of 1 indicates a perfect forecast; a skill score of zero indicates that the forecast accuracy is the same as using the long-term climatological averages; and a skill of less than zero indicates that the skill is below that of the climatology. The value of SS_{MSE} can be decomposed into three components (Murphy and Winkler, 1992)

$$SS_{MSE} = \rho_{fx}^2 - \left[\rho_{fx} - \frac{\sigma_f}{\sigma_x}\right]^2 - \left[\frac{\mu_f - \mu_x}{\sigma_x}\right]^2, \qquad (2)$$

where ρ_{fx} is the Pearson correlation coefficient between observations and forecasts and quantifies the degree of linear dependence between the two; μ_f and μ_x are the forecast and observation means, respectively, and σ_f is the standard deviation of the forecasts. Based on this decomposition, the coefficient of determination (denoted by R^2) reflects the forecast accuracy in the absence of biases, and is referred to as the *potential skill* (PS), or 'inflated' skill that might be achieved in the absence of biases. The second term in the right side of equation (2) quantifies the conditional biases and it is referred to as the *slope reliability* (SREL). The last term quantifies the unconditional biases and it is referred to as the *standardized mean error* (SME).

Forecast verification using the MSE skill score and its decomposition in equation (2) produces a more realistic diagnostic of the forecast skill compared to taking the correlation coefficient at face value. The decomposition of the skill in different sources of bias provides information on model strengths and weaknesses, which may be useful for model developers and/or forecast users. In general, the unconditional biases (large SME) can easily be removed with bias-correction methods (Hashino et al., 2007) while the conditional biases (large SREL) tend to require more sophisticated calibration. Any forecasts with low PS will have limited predictability, even if the biases are eliminated.

3.2. Bayesian updating (BU)

Post-processing of ensemble forecasts is a common approach for removing forecast biases and reducing model error (National Academy of Sciences, 2006). BU of climate model forecasts is an implementation of Bayes' theorem, in which the climatological probability distribution of a forecast variable, *Y* (e.g., precipitation or temperature), can be updated using newly-available information (e.g., the precipitation or temperature NMME forecasts).

Bayesian approaches were successfully introduced as part of the DEMETER project to enhance sea surface temperature and precipitation forecasts (Coelho 2004; Luo et al., 2007). In hydrologic forecasting, Bayesian merging has been used to develop a multimodel seasonal hydrologic ensemble prediction system (Luo and Wood, 2008), to obtain probabilistic streamflow forecasts (Wang et al., 2013), or to weight the forecasts using a climate index such as the El Niño-Southern Oscillation or Pacific Decadal Oscillation (Bradley et al., 2015). However, BU has not yet been implemented in a systematic fashion over large regions to see if it is possible to enhance NMME precipitation or temperature forecasts.

Here, we implement BU to leverage the forecasting skill of the eight NMME single model-ensembles or of the 94 individual model members based on their performance for every month of the year and for every lead time. Before any forecast is made, our best estimate of the probability of different outcomes is defined by the climatology (i.e., the probability distribution of historical outcomes), represented here by the prior climatological density function $f(y)$. After a climate model forecast θ is issued, the updated (or posterior) density function is given by Bayes' theorem to be

$$f(y \mid \theta) = \frac{f_\theta(\theta \mid y) f(y)}{f_\theta(\theta)}, \tag{3}$$

where $f_\theta(\theta)$ is the unconditional density of θ, and $f_\theta(\theta \mid y)$ is the likelihood function. The posterior density $f(y \mid \theta)$ describes the conditional distribution of the variable given the climate model forecast θ, and therefore represents a probability distribution forecast of the outcome. Analytical solutions to equation (3) are available when the prior density and the likelihood function are normally distributed (i.e., Gaussian). Here we apply BU to a data sample (rather than to density functions). Let $\{y_i, i=1, \ldots, N\}$ represent the historical observations of Y, i.e., a sample drawn from the prior density $f(y)$. We represent a sample drawn from the posterior density $f(y \mid \theta)$ (Smith and Gelfand, 1992) using the likelihood function $f_\theta(\theta \mid y)$. By definition, the likelihood function $f_\theta(\theta \mid y)$ is the distribution of a given model forecast θ conditioned on a particular outcome y for the same month.

For example, to apply BU to the eight NMME models (or 94 members), we treat each model (or member) sequentially. Beginning with one model, one month, one lead time, and one region (e.g., NASA January forecasts at Lead 0.5 in the Atlantic region), we first hypothesize a linear relationship between the forecasts (θ) and observations (y) across all years (e.g., Luo et al. 2007) as

$$\theta = \alpha + \beta y + \varepsilon, \tag{4}$$

where α and β are the intercept and slope parameters (bias and scaling error in the model), respectively, and ε is the Gaussian residual model error. Using every observation for the given month (e.g., January E-OBS observations from 1950 to 2015), excluding the actual forecast observation, we estimate the parameters α and β by linear regression. For any given outcome y, the expected value of a corresponding forecast $\bar{\theta}(y)$ using a simple linear regression model is

$$\bar{\theta}(y) = \alpha + \beta y \cdot \tag{5}$$

We assume that the residual model errors ε are normally distributed with mean zero and constant variance σ^2 and can then write the likelihood function $f_\theta\ (\theta \mid y)$ as a Gaussian density function

$$f_\theta(\theta \mid y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(\theta - \bar{\theta}(y))^2}{\sigma^2}} \ . \tag{6}$$

The likelihood function is then computed for each historical monthly observation y_i in the historical sample (excluding the forecast month) to obtain a weight w_i for each observation as

$$w_i = \frac{f_\theta(\theta \mid y_i)}{\sum_{j=1}^{N} f_\theta(\theta \mid y_j)} \ . \tag{7}$$

The weight w_i represents the likelihood of observing outcome y_i given the climate forecast θ(Smith and Gelfand, 1992), and the sum of the weights w_i is equal to 1. The collection of weights for a given month (e.g., from 1950 to 2015, minus the forecast year) is therefore analogous to a discrete probability distribution forecast for the given model (or model member). In other words, the weights show the likelihood of each discrete historical outcome given the climate model forecasts. If all the weights are equal (i.e., 1/*N*), they produce the same distribution as the prior distribution before BU, so the output is equivalent to a climatology forecast (i.e., the average historical conditions for the same months) and the model forecast is automatically ignored. For models with a weak relationship between forecasts and observations, the weights will be close to 1/*N*, indicating that each outcome is nearly equally likely. For models with a strong, significant relationship between forecasts and observations, each historical outcome y_i receives a different weight, and the unequal weighting grows as the PS increases. Any weights greater than 1/*N* indicate that the outcome is more likely than the climatology given the forecast; any weights smaller than 1/*N* indicate that the outcome is less likely. We repeat this procedure for each forecast individually.

To combine the eight single-model ensembles (or 94 model members) into a multi-model forecast, we apply the BU sequentially to each model, and then combine their weights to produce a multi-model weight. Assuming that the single-model forecasts are independent (Luo et al. 2007), the multi-model weight w_i^* is the product of the eight model weights for each observation y_i in the historical sample, normalized to produce a set of multi-model weights that sum to 1 (Bradley et al. 2015)

$$w_i^* = \frac{\prod_{k=1}^{8} w_i^k}{\sum_{j=1}^{N}\left(\prod_{k=1}^{8} w_j^k\right)}, \tag{8}$$

where w_i^k is the *i*-th weight for the *k*-th model. For a given forecast (e.g. January 1982) we have 66 multi-model weights (e.g., one for each historical observation for January from 1950 to 2015, minus the forecast year). The final multi-model forecast $\bar{y}$ is the expected value of the Bayesian updated probability distribution, defined by the weighted average:

$$\bar{y} = \sum_{i=1}^{N} w_i^* y_i \,. \tag{9}$$

The multi-model forecast weight is thus a normalized product of all the weights for the individual models. It is important to note that a model with relative weights that are all 1/*N* (a climatology forecast) has no effect at all on the multi-model weights; in other words, if a model has no PS, it is as if the model is automatically ignored. The method, as an application of Bayes' theorem, produces bias-corrected ensemble climate forecasts by optimally merging climate forecasts from multiple models based on their performance for specific months and lead times.

Four of our multi-models are based on BU: (1) BU of the eight single-model ensemble forecasts (BU-8); (2) BU of the 94 individual model members (BU-94); (3) BU of the principal components of the eight single-model ensemble forecasts (BU-PCA-8), and (4) BU of the principal components of the 94 model members (BU-PCA-94). Our rationale for differentiating between the eight single-model ensembles and the 94 individual model members is to assess whether the individual members actually do produce an enhanced model forecast in comparison with the single-model ensembles. This question is important, as the single-model ensemble forecasts are much faster to prepare and compute for a given region in comparison with the model members. Thus, if their skill is comparable to that of the members, model forecasts may be obtained much faster.

3.3. BU of principal components

In multi-models BU-8 and BU-94, we make the assumption that the errors from the eight single-model ensembles are independent, so the BU is applied sequentially for each model, and the multi-model forecast weight is a normalized product of all the weights for the single-model ensembles for every given month and lead time (as described above). As a result, the forecasts have a tendency to highlight any consensus among the models, regardless of whether or not the single-model forecasts are correct (e.g., Olson et al., 2016).

Here we attempt to reduce the conditional biases arising from similarities among the single model ensemble forecasts by developing a second approach based on principal components analysis (PCA), which is referred to as BU-PCA-8 and BU-PCA-94, respectively. Instead of computing a linear regression between the model forecasts and observations as described above, we first pool together the eight (or 94) model forecasts, and conduct a PCA using the 'prcomp' function from the base `stats` package in the open-source software R (R Core Team and contributors worldwide, 2016). If one model forecast is missing for a given lead time and month, then that entire model is removed from the calculation of the components. Additionally, the PCA must be conducted on complete data, so any month that is missing a forecast (from one or more models) is excluded from the analysis. The variables are centered and scaled prior to applying the PCA, and we retain all of the components. The linear relationship is then computed between the principal components and the observed data, and the BU procedure is applied in the same manner as before, but using the principal components instead of the single-model ensemble forecasts.

By implementing the principal components approach before the BU, we no longer have to assume independence of the single-model ensembles that are used in the weighting scheme. The BU gives more weight to the model components with high PS, and less to those with low or no PS, for every month and lead time. This BU-PCA approach is similar to other probability adjustment procedures (Stedinger and Kim, 2010) and can be thought of as a way of preconditioning the forecasts to reduce any over confidence arising from model similarity. The methodology can then be applied to other climate variables beyond precipitation and temperature, and the multi-model forecasts can be used as inputs for practical ensemble forecasting.

Note that for all five multi-models, the maximum number of forecasts (i.e. eight single-model ensemble forecasts, or 94 individual model member forecasts) is not always used because of the presence of gaps in the original forecast data. When computing the multi-model forecasts, we use as many forecasts as are available for the given month or lead time.

3.4. Extreme event diagnosis

To evaluate the skill of the NMME in predicting extremes, we focus on four extreme precipitation events (August 2002, August 2005, May-June 2010, May-June 2013) and four extreme temperature events (June-August 2003, June-July 2007, June-July 2010, March 2012), using the two- or three- month average when the event lasted more than one month. We selected events that lasted between one and three months to assess how well they were forecast by the single-model ensembles over multiple lead times, and how well they would have been forecast using our five multi-model weighting schemes. The events were chosen using the International Disaster Database from the Centre for Research on the Epidemiology of Disasters (Emergency Events Database, http://www.emdat.be), which records data on world mass disasters that have occurred since the beginning of the twentieth century. Using extreme observations to compare forecasts may not always be an appropriate strategy, as 'predicting calamity becomes a worthwhile strategy', and incorrect conclusions may be drawn (Lerch et al. 2017). Here, however, we use extreme events solely to draw qualitative conclusions regarding consistency of forecasts across lead times.

We start by defining the extent of the extreme event using the reference E-OBS data. For every one degree pixel, we compute the standardized anomaly for the selected season for every year between 1983 and 2015. The years 1983 to 2015 are retained because not all NMME models have forecasts before 1983. We plot the seasonal anomaly across the whole of Europe, and select all of the grid cells where the anomaly was greater than or equal to 1. We did this for every event with the exception of the Summer 2003 event which covered most of Europe, and where we set a threshold of 1.5. This threshold allowed us to reduce the event's spatial extent and to test the forecasting skill of NMME models over a range of extremes (the June-August 2003 temperature extreme was about 3.5, compared to about 1.7 for the 2012 March event). Based on the limits of the outlined event (**Figure 2**), we then compute the domain-averaged time-series of temperature or precipitation for the given months, from 1983 to 2015 (e.g., for the June-August 2007 temperature event, we have a time series of the June-August temperature anomaly for 1983, for 1984, and every year until 2015). The 95% confidence intervals are computed for the observed E-OBS anomaly (*x*) following the approach described in Stedinger et al. (1993, section 18.4.2.) as

$$x \pm 1.96 * \sqrt{\frac{1}{n}\left(1 + 0.5 * x^2\right)}, \qquad (10)$$

where *n* is the number of years in the E-OBS anomaly time series (here 33 years from 1983 to 2015) and the values represent the upper and lower limits, respectively, of the confidence interval.

Separately, we obtain the time series of NMME anomalies over the same region, using the same spatial boundaries (**Figure 2**). Domain-averaged anomaly time series are computed in the same manner as for the E-OBS data, but for every lead time. The seasonal forecast is computed as the sum of the forecasts initialized ahead of the entire season, for each of the eight single-model ensembles and for the 94 individual model members. Following the approach described in Slater et al. (2017), the seasonal forecast for a given event, such as the June-July 2010 extreme precipitation event, initialized in June and lasting for two months, would be computed as the sum of the 0.5- and the 1.5- month lead forecasts initialized in June. The forecast initialized one month earlier would be computed as the sum of the 1.5- and the 2.5-month forecasts initialized in May. Those forecasts are then computed as anomalies for comparison with the E-OBS anomalies time series. The BU approach is applied separately to the eight single-model ensemble seasonal forecasts or the 94 individual model member seasonal forecasts. The aim is to investigate how well the individual NMME models are able to forecast these climate extremes, and whether we can obtain improved, bias-corrected weighted model forecasts of these extremes over longer lead times.

4. Results

Using the skill score decomposition described in **Section 3.1** to evaluate the predictive skill of

NMME forecasts, we first measure the skill of the eight single-model ensembles (**Section 4.1**) before comparing that of the multi-model ensembles in subsequent sections. EW-8 is used as a benchmark (**Section 4.2**) against the two Bayesian models (BU-8 and BU-94 in **Section 4.3**) and the two Bayesian models with principal components (BU-PCA-8 and BU-PCA-94 in **Section 4.4**). Last, we finish by comparing the ability of these different multi-models to forecast a selection of eight extreme events that occurred in different European regions during the first two decades of the 21st century.

4.1. The eight single-model ensembles: low skill and high biases

We evaluate the predictive ability of the eight single-model ensembles (computed as the mean of each model's members, i.e. the simplest and fastest forecasting approach) through a decomposition of the skill score into PS, and the two main sources of bias, unconditional and conditional biases.

Across all four European regions and all lead times, the PS of the precipitation forecasts for individual months is relatively low, mostly ranging between 0 and 0.1 (**Supplementary Figure 1**). It tends to be higher at the shortest lead time (~0.2-0.4) for the models with good skill (e.g., CCSM4, CFSv2), and low, with random variations, across all other lead times. The forecasts are not markedly better in any given one of the four regions.

The precipitation skill score, or *actual* skill of the models, is mostly negative as a result of large unconditional biases, which are systematic errors in the model (i.e., a tendency to over- or under-predict), and tend to be seasonal (e.g., stronger biases in the winter months for CCSM3 and CCSM4 or stronger in the summer months for GEOS5). Their effect can be seen in the mirror-image between the skill score (blue) and the unconditional biases (red). Thus, the unconditional bias is clearly the primary source of bias across these eight models, as was also found in Bradley et al. (2015) and Slater et al. (2017). The conditional biases are also irregularly distributed across the different months of the year and lead times, and vary substantially from model to model.

The skill of temperature forecasts is also relatively poor across all four regions for individual months. Compared to precipitation, there is a more pronounced decrease in skill with increasing lead time, and relatively high forecast skill (>0.5) is obtained by many models at the shortest lead time (e.g., CFSv2; **Supplementary Figure 2**). The best PS tends to be found in the Mediterranean region during the summer months (e.g., CCSM4, FLORb01, CFSv2). The skill score is largely driven by the unconditional biases, which vary inconsistently: for some models like CFSv2, some of the biases grow with increasing lead time, whereas for others, they grow seasonally (e.g., for GEOS5 biases grow in the cold months for the Humid-Continental and Subarctic-Polar regions; or in the summer months for the Mediterranean region). The conditional biases, in contrast, tend to be randomly distributed.

Overall, the eight single-model ensemble forecasts for precipitation and temperature have relatively little skill beyond the shortest lead times (at the monthly scale), primarily due to the presence of unconditional biases, which tend to vary by season and lead time. Variations in the conditional biases also affect the skill score to a much lesser extent. Our aim is therefore to develop a systematic methodology that will allow us to eliminate these biases by leveraging the strengths of the different models over specific regions, months, and lead times.

4.2. EW-8: a substantial improvement over the raw forecasts

Our first multi-model takes the arithmetic mean of the eight single-model ensembles (which are computed as the arithmetic mean of the members; so each single-model ensemble may have between 6 and 24 members - see **Table 1**). This model can be thought of as eight equally weighted GCMs, and thus will be referred to as EW-8. The PS (R^2) is computed by correlating this arithmetic mean against the observed values. Previous work has shown that equally weighted NMME forecasts tend to be as good as or better than those of the best single-model ensemble (Becker et al., 2014; Slater et al., 2017). Therefore, here we use EW-8 as a 'least effort' benchmark against which to compare subsequent multi-models in **sections 4.3-4.5**. For comparison, we also compute the R^2 of the raw 94 model members ('94 mem'; see **Table 2**). For 94 mem, the R^2 is derived from the correlation between all 94 members and the observation. In contrast, for EW-8 we first compute the arithmetic mean of the 8 single-model ensembles, before computing the R^2 (so there is far less spread in the data).

Results indicate that the EW-8 forecast PS is much better than the raw 94 member PS (the raw 94 members have greater spread and larger conditional biases than the EW-8 averages). We chose to show the 0.5- and 5.5- month lead times in **Table 2** and **Figures 3-4** for the sake of parsimony and to compare the 'best' skill with the skill obtained after several months (once it is no longer affected by the initial conditions). Across all four regions at the 0.5-month lead time, the mean precipitation PS increases from R^2=0.15 for the 94 model members to R^2=0.38 for EW-8 (color circles top row of **Figure 3; Table 2**). A similar improvement can be found for precipitation at the 5.5-month lead time (94 members R^2=0.09; EW-8 R^2=0.27) (**Figure 3** and **Table 2**).

When comparing the precipitation forecast PS of the single-model ensembles across regions, for a given lead time, we find that the skill tends to be good in the Mediterranean region, but much poorer in the three other regions (**Table 2**), where there is greater seasonal variability. At the 0.5-month lead time, the magnitude of the improvement of the forecast skill between the 94 members and EW-8 (in absolute terms) is best in the Subarctic-Polar region, where the skill was one of the poorest to begin with. At the 5.5-month lead time, however, the precipitation forecasts have even larger initial spread and so EW-8 does not perform quite as well (see the Humid-Continental region).

The temperature forecasts tend to be more skillful than the precipitation forecasts and are relatively consistent across the four regions, although the skill decreases and becomes more variable in the cold months (**Figure 4**). The enhancement between the 94 members and EW-8 forecasts is smaller than for precipitation (e.g., R^2=0.91 for 94 members, to R^2=0.96 for EW-8 at the 0.5-month lead time on average), because there is less room for improvement (**Figure 4 and Table 2**). One reason for these high R^2 values is the ability of the models to reproduce the seasonality of temperature (e.g., July is warmer than January), so the skill is artificially inflated when observing all months together (in comparison with the skill that would be achieved on a month-by-month basis, and which can be seen in **Figures 3-5**). Hence, in future work, it may be worth studying the forecasts of anomalies (from their monthly mean) to eliminate the effect of seasonality.

For both temperature and precipitation, the breakdown of EW-8 in terms of PS and biases indicates that it performs as well as or better than the best single-model ensemble (**Figure 5 vs. Supplementary Figures 1-2**). The PS of the best single-model ensembles (e.g., CFSv2 precipitation) is mostly preserved. The skill score improves slightly (particularly in the Temperate region) but remains largely negative, indicating that there is still substantial room for improvement, namely by tackling the presence of unconditional biases in the model forecasts. Overall, therefore, EW-8 reduces the conditional biases, preserves the unconditional biases, and slightly improves the skill score (**Figure 5 vs. Supplementary Figures 1-2**).

4.3. BU: improved skill and removal of unconditional bias, at the expense of the conditional biases

Models BU-8 and BU-94 seek to address the issue of the unconditional biases in the models (i.e., the primary source of bias) by using the (unbiased) climatological distribution as a prior, and updating it (so the lack of bias is preserved). For precipitation, BU-8 clearly eliminates much of the single-model bias (see the first and second rows of each panel; **Figure 3**). The forecasts are sharply re-centered around the one-to-one line, particularly in the two regions with the strongest biases, Humid-Continental and Subarctic-Polar. When the bias is small, such as in the Mediterranean region, the bias removal is less noticeable, and BU-8 actually performs less well than EW-8 (**Table 2**). The PS is generally a little better in BU-8 than BU-94 (especially for longer leads); however the unconditional bias removal (SME) is better in BU-94 (**Figure 5**).

For temperature, the effect of BU-8 and BU-94 is similar, as the forecasts for each of the 12 months clearly re-center around the one-to-one line (**Figure 4**). The adjustment is most visible for the months that had the largest variability and error to begin with, such as the cold months (dark blue circles). However, the PS is not improved when all months are considered together (**Table 2**).

The skill score of BU-8 and BU-94 is notably ‘smoothed out’ in comparison with EW-8 (**Figure 5**) due to the unconditional bias removal. The BU conditional biases, however, are slightly worse than those of the eight single-model ensembles (**Figure 5** vs. **Supplementary Figures 1-2**). Because the BU models are based on the questionable assumption of independence across models, it is likely that the forecasts may be overconfident in comparison with the EW-8 forecasts. Thus, in BU-8 and BU-94, most of the bias is conditional, as is clearly visible in the mirror-image between the skill score and the SREL in **Figure 5**.

We hypothesize that the increase in conditional biases in BU-8 and BU-94 is due to the lack of independence among model forecasts. Models that behave similarly, such as CCSM3 and

CCSM4, or CanCM3 and CanCM4, will tend to produce overconfidence for specific months and lead times when the models concur, because all of the models are treated equally in the reweighting scheme. Therefore, we develop a multi-model based on PCA that will transform the potentially correlated forecasts from the eight single-model ensembles (or 94 individual model members) into a new set of linearly uncorrelated components, before conducting the BU.

4.4. BU-PCA: effectively removes negative skill but at the expense of positive skill

Instead of applying the weights on a model-by-model basis, we compute the principal components among the eight single-model ensembles (BU-PCA-8) and among the 94 model members (BU-PCA-94). For every lead time and every month, the model forecasts are pooled together across the entire forecast period (1982-2015), and the BU procedure is applied to the principal components, as described in **Section 3.3**. The scatterplots of the resulting forecasts (fourth and fifth rows in each panel in **Figures 3-4**) show that both BU-PCA models tend to re-center the forecasts around the one-to-one line, in the same manner as the two BU models (second and third rows), but they also “flatten” the forecast variance considerably (horizontally). The PCA procedure thus appears to reduce the conditional biases (compared to BU-8 and BU-94) by removing any overconfidence arising from similarities among single model ensemble forecasts (i.e., instead of applying BU to every model/member, it is applied to the principal components). Compared with EW-8, BU-PCA-8 and BU-PCA-94 still have slightly greater conditional biases (**Figure 5**) but the unconditional biases are notably reduced. Following the reduction of biases, the skill score of the BU-PCA models mirrors the PS much more closely than in EW-8, so there is less ‘room for improvement’ left in the difference between the PS and the skill score (**Figure 5**).

We compare the BU-PCA-8 and BU-PCA-94 forecasts to determine whether it is “worth” using all of the individual model members when producing a weighted model forecast. Our reasoning is that the use of individual members is likely to heighten model skill through the addition of new forecast signals (DelSole et al., 2014) while the use of single-model ensemble forecasts is more likely to impoverish the signal (Knutti et al., 2010). Interestingly, we find that at the shortest lead times (0.5-month lead), the PS of BU-PCA-8 is consistently better than that of BU-PCA-94. At the 5.5-month lead time, however, the reverse holds. These results suggest that when there is greater uncertainty in the model forecast (i.e., at longer lead times), it may be better to use all the model members than the single-model ensembles, within the BU-PCA approach (**Table 2**).

Thus, the five multi-models each have different biases: those in EW-8 are primarily unconditional; those of the two BU models are primarily conditional; and those of the two BU-PCA models the biases are relatively small and random, while the strong negative values in the skill score are virtually eliminated.

4.5. Skill of the five multi-models in forecasting extreme precipitation and temperature events

As a test of the ability of the five multi-models to predict extreme climate, we evaluate the magnitude of precipitation and temperature forecast anomalies for four extreme temperature and four extreme precipitation events (**Figure 6**). Previously, we found that the eight single-model ensembles were unable to forecast extreme precipitation and climate more than several months ahead of an extreme event's occurrence in different regions of the continental USA (Slater et al., 2017). Here, the 94 individual model members (grey lines) also tend to fluctuate between extremely high and low anomalies, with temperature and precipitation performing similarly. The 94 members rarely attain the observed anomaly, particularly when the anomaly is greater than 3. Even when they do, the forecasts appear to be random and rarely persist several months ahead of the event (e.g., 2002 August precipitation).

So how well do the five multi-models perform in comparison with the 94 individual model members? EW-8 (black line) is mediocre: it tends to forecast the sign of the anomaly correctly, but largely under-predicts the magnitude (**Figure 6**). BU-8 (magenta) and BU-94 (green) do better in estimating the magnitude of the anomaly (particularly for temperature), but are more likely to get the sign wrong. Thus, BU-8 is arguably less consistent than EW-8, likely because the single-model ensembles are treated independently, so any similarities among the models are over-strengthened (Olson et al., 2016), even when they are incorrect. BU-PCA-8 and BU-PCA-94 are both very inconsistent (especially BU-PCA-94), with abrupt variations from one lead time to the next, possibly because the BU-PCA approach brings the resulting forecasts closer to the climatological mean.

Overall, the skill of our multimodels is similar to that of other multi-model weighting techniques such as equal weights (Becker et al., 2014; Hagedorn et al., 2005; Slater et al., 2017), multiple linear regression (Doblas-Reyes et al., 2005), other Bayesian-based approaches (Rajagopalan et al., 2002; Robertson et al., 2004; Weigel et al., 2008), optimal weights (Wanders and Wood, 2016; Weigel et al., 2008) or genetic algorithms (Ahn and Lee, 2016). However, it is difficult to compare these multi-models in detail as most have been applied over different spatial and temporal resolutions, and often verified using different evaluation metrics. Overall, these results suggest that the 'conservative' approach would be to stick with the EW-8 model, which is both the fastest and simplest model forecast to produce.

5. Conclusions

We have evaluated the skill of eight NMME models and different weighting schemes in forecasting temperature and precipitation across Europe over the 1982-2015 period. The main findings of this paper can be summarized as follows:

- Individually, the eight single-model ensembles have little forecasting skill beyond the shortest lead times, primarily because of the large unconditional biases in the models, which vary seasonally. The conditional biases have less influence on the forecast skill because they tend to be irregularly distributed across the different months of the year and lead times.

- EW-8 is a simple, but effective method for improving forecast skill by taking the arithmetic mean of the single-model ensembles. EW-8 reduces the conditional biases, preserves the unconditional biases, and slightly improves the skill score and PS of the eight single-model ensembles. Overall, however, the skill score remains negative, so there is still vast room for improvement.

- BU-8 and BU-94 both homogenize model skill scores slightly across all lead times and forecast months by removing the unconditional biases. However, they do this at the expense of the conditional biases, which are accentuated in comparison with EW-8 (likely due to overfitting and/or model similarity). The improvements are most notable in the regions and months that exhibit the strongest biases to begin with.

- BU-PCA-8 and BU-PCA-94 transform the potentially correlated forecasts from the eight single-model ensembles and from the 94 individual model members into a new set of linearly uncorrelated components, before conducting the BU. In comparison with the two BU models, their unconditional biases are similar and the conditional biases are reduced. It appears overall that the principal components approach fixes the lack of independence across models, but brings the resulting forecasts closer to the climatological mean. In comparison with EW-8, the skill score is much more homogeneous (negative skill is dramatically reduced) but there is also some loss of skill.

Our results suggest that there is not much to be gained by using the full information provided by the 94 individual model members, in comparison with the single model ensembles (which take the mean of each model's members). In fact, the equally weighted (EW-8) model is considerably faster to compute than any other multi-model, and in the case of extreme precipitation and temperature events, its forecasts are more conservative, but less prone to major errors. Other studies have found that considerable skill improvement can be obtained using optimal weights (Wanders and Wood 2016) and in our case it remains to be determined in future work how the BU-PCA approach may be improved.

Acknowledgments and Data

The authors thank the NMME program partners and acknowledge the help of NCEP, IRI and NCAR personnel in creating, updating and maintaining the NMME archive, with the support of NOAA, NSF, NASA and DOE. Three reviewers are thanked for their insightful comments that helped improve the quality of the paper. This study was supported by NOAA's Climate Program Office's Modeling, Analysis, Predictions, and Projections Program, Grant #NA15OAR4310073. Gabriele Villarini and Louise Slater also acknowledge financial support from the Broad Agency Announcement (BAA) Program and the Engineer Research and Development Center (ERDC)–Cold Regions Research and Engineering Laboratory (CRREL) under Contract No. W913E5-16-C-0002.

The data supporting the conclusions can be obtained from the corresponding author, l.slater@lboro.ac.uk

References

Ahn, J.-B., Lee, J., 2016. A new multimodel ensemble method using nonlinear genetic algorithm: An application to boreal winter surface air temperature and precipitation prediction. J. Geophys. Res. Atmos. 121, 9263–9277. doi:10.1002/2016JD025151

Arritt, R.W., Rummukainen, M., 2011. Challenges in Regional-Scale Climate Modeling. Bull. Am. Meteorol. Soc. 92, 365–368. doi:10.1175/2010BAMS2971.1

Barnston, A.G., Lyon, B., 2016. Does the NMME Capture a Recent Decadal Shift toward Increasing Drought Occurrence in the Southwestern United States? J. Clim. 29, 561–581. doi:10.1175/JCLI-D-15-0311.1

Becker, E., den Dool, H. Van, Zhang, Q., 2014. Predictability and Forecast Skill in NMME. J. Clim. 27, 5891–5906. doi:10.1175/JCLI-D-13-00597.1

Bradley, A.A., Habib, M., Schwartz, S.S., 2015. Climate index weighting of ensemble streamflow forecasts using a simple Bayesian approach. Water Resour. Res. 51, 7382–7400. doi:10.1002/2014WR016811

Brier, G.W., 1950. Verification of forecasts expersses in terms of probaility. Mon. Weather Rev. 78, 1–3. doi:10.1126/science.27.693.594

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

DelSole, T., Nattala, J., Tippett, M.K., 2014. Skill improvement from increased ensemble size and model diversity. Geophys. Res. Lett. 41, 7331–7342. doi:10.1002/2014GL060133

DelSole, T., Tippett, M.K., 2014. Comparing Forecast Skill. Mon. Weather Rev. 142, 4658–4678. doi:10.1175/MWR-D-14-00045.1

Doblas-Reyes, F.J., Hagedorn, R., Palmer, T.N., 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting - II. Calibration and combination. Tellus 57, 234–252. doi:10.1111/j.1600-0870.2005.00104.x

Faber, B.A., Stedinger, J.R., 2001. Reservoir optimization using sampling SDP with ensemble streamflow prediction (ESP) forecasts. J. Hydrol. 249, 113–133. doi:10.1016/S0022-1694(01)00419-X

Hagedorn, R., Doblas-Reyes, F.J., Palmer, T.N., 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. Tellus, Ser. A Dyn. Meteorol. Oceanogr. 57, 219–233. doi:10.1111/j.1600-0870.2005.00103.x

Hashino, T., Bradley, A.A., Schwartz, S.S., 2007. Evaluation of bias-correction methods for ensemble streamflow volume forecasts. Hydrol. Earth Syst. Sci. 11, 939–950. doi:10.5194/hess-11-939-2007

Haylock, M.R., Hofstra, N., Klein Tank, A.M.G., Klok, E.J., Jones, P.D., New, M., 2008. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. J. Geophys. Res. 113, D20119. doi:10.1029/2008JD010201

Hewitt, C.D., Griggs, D.J., 2004. Ensembles-based predictions of climate changes and their impacts. Eos, Trans. Am. Geophys. Union 85, 566. doi:10.1029/2004EO520005

Hodyss, D., Satterfield, E., McLay J., Hamill T.M., Scheuerer, M. 2016. Inaccuracies with multimodel postprocessing methods involving weighted, regression-corrected forecasts. Mon. Wea. Rev., 144, 1649-1668.

Infanti, J.M., Kirtman, B.P., 2016. North American rainfall and temperature prediction response to the diversity of ENSO. Clim. Dyn. 46, 3007–3023. doi:10.1007/s00382-015-2749-0

Infanti, J.M., Kirtman, B.P., 2014. Southeastern U.S. Rainfall Prediction in the North American Multi-Model Ensemble. J. Hydrometeorol. 15, 529–550. doi:10.1175/JHM-D-13-072.1

Jia, L., Yang, X., Vecchi, G.A, Gudgel, R.G., Delworth, T.L., Rosati, A., Stern, W.F., Wittenberg, A.T., Krishnamurthy, L., Zhang, S., Msadek, R., Underwood, S., Kapnick, S., Zeng, F., Anderson, W.G., Balaji, V., Dixon, K., 2015. Improved Seasonal Prediction of Temperature and Precipitation over Land in a High-resolution GFDL Climate Model. J. Clim. 5, 2044–2062. doi:10.1175/JCLI-D-14-00112.1

Kirtman, B.P., Min, D., 2009. Multimodel ensemble ENSO prediction with CCSM and CFS. Monthly Weather Review, 137(9), pp.2908-2930.

Kirtman, B.P., Min, D., Infanti, J.M., Kinter, J.L., Paolino, D.A., Zhang, Q., Van Den Dool, H., Saha, S., Mendez, M.P., Becker, E., Peng, P., Tripp, P., Huang, J., Dewitt, D.G., Tippett, M.K., Barnston, A.G., Li, S., Rosati, A., Schubert, S.D., Rienecker, M., Suarez, M., Li, Z.E., Marshak, J., Lim, Y.K., Tribbia, J., Pegion, K., Merryfield, W.J., Denis, B., Wood, E.F., 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. doi:10.1175/BAMS-D-12-00050.1

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., Meehl, G.A., 2010. Challenges in combining projections from multiple climate models. J. Clim. 23, 2739–2758. doi:10.1175/2009JCLI3361.1

Krishnamurti, T.N., Kishtawal, C.M., LaRow, T.E., 1999. Improved weather and seasonal climate forecasts from multimodel superensemble. Science. 285(5433), 1548–1550. doi:10.1126/science.285.5433.1548

Lerch, S., Thorarinsdottir, T.L., Ravazzolo F. Gneiting, T., 2016. Forecaster's dilemma: Extreme events and forecast evaluation. Statistical Science, 32(1), pp.106-127.

Luo, L., Wood, E.F., 2008. Use of Bayesian Merging Techniques in a Multimodel Seasonal Hydrologic Ensemble Prediction System for the Eastern United States. J. Hydrometeorol. 9, 866–884. doi:10.1175/2008JHM980.1

Luo, L., Wood, E.F., Pan, M., 2007. Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. J. Geophys. Res. Atmos. 112, 1–13. doi:10.1029/2006JD007655

Ma, F., Ye, A., Deng, X., Zhou, Z., Liu, X., Duan, Q., Xu, J., Miao, C., Di, Z., Gong, W., 2015a. Evaluating the skill of NMME seasonal precipitation ensemble predictions for 17 hydroclimatic regions in continental China. Int. J. Climatol. 144, n/a-n/a. doi:10.1002/joc.4333

Ma, F., Yuan, X., Ye, A., 2015b. Seasonal drought predictability and forecast skill over China. J. Geophys. Res. Atmos. 120, 8264–8275. doi:10.1002/2015JD023185

Madadgar, S., AghaKouchak, A., Shukla, S., Wood, A.W., Cheng, L., Hsu, K.-L., Svoboda, M., 2016. A hybrid statistical-dynamical framework for meteorological drought prediction: Application to the southwestern United States. Water Resour. Res. 51, 9127–9140. doi:10.1002/2015WR018547

Misra, V., Li, H., 2014. The seasonal climate predictability of the Atlantic Warm Pool and its teleconnections. Geophys. Res. Lett. 661–666. doi:10.1002/2013GL058740.Received

Mo, K.C., Lettenmaier, D.P., 2014. Hydrologic prediction over Conterminous U.S. using the National Multi Model ensemble. J. Hydrometeorol. 140429111703004. doi:10.1175/JHM-D-13-0197.1

Mo, K.C., Lyon, B., 2015. Global Meteorological Drought Prediction using the North American Multi-Model Ensemble. J. Hydrometeorol. 150310071054006. doi:10.1175/JHM-D-14-0192.1

Murphy, A.H., Winkler, R.L., 1992. Diagnostic verification of probability forecasts. Int. J. Forecast. 7, 435–455. doi:10.1016/0169-2070(92)90028-8

National Academy of Sciences, 2006. Completing the Forecast: Characterizing and Communicating

Uncertainty for Better Decisions Using Weather and Climate Forecasts, National Research Council Committee on Estimating and Communicating Uncertainty in Weather and Climate Forecasts.

Olson, R., Fan, Y., Evans, J.P., 2016. A simple method for Bayesian model averaging of regional climate model projections: Application to southeast Australian temperatures. Geophys. Res. Lett. 43, 7661–7669. doi:10.1002/2016GL069704

Peel, M.C., Finlayson, B.L., McMahon, T.A., 2007. Updated world map of the Köppen-Geiger climate classification. Hydrol. Earth Syst. Sci. 11, 1633–1644. doi:10.5194/hess-11-1633-2007

R Core Team and contributors worldwide, 2016. The R Stats Package.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Mon. Weather Rev. 133, 1155–1174. doi:10.1175/MWR2906.1

Rajagopalan, B., Lall, U., Zebiak, S.E., 2002. Categorical Climate Forecasts through Regularization and Optimal Combination of Multiple GCM Ensembles. Mon. Weather Rev. 130, 1792–1811. doi:10.1175/1520-0493(2002)130<1792:CCFTRA>2.0.CO;2

Robertson, A.W., Lall, U., Zebiak, S.E., Goddard, L., 2004. Improved Combination of Multiple Atmospheric GCM Ensembles for Seasonal Prediction. Mon. Weather Rev. 132, 2732–2744. doi:10.1175/MWR2818.1

Roundy, J.K., Yuan, X., Schaake, J., Wood, E.F., 2015. A Framework for Diagnosing Seasonal Prediction through Canonical Event Analysis. Mon. Weather Rev. 143, 2404–2418. doi:10.1175/MWR-D-14-00190.1

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.T., Chuang, H.Y., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M.P., 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. doi:10.1175/JCLI-D-12-00823.1

Scheuerer, M., Büermann, L., 2014. Spatially adaptive post-processing of ensemble forecasts for temperature. J. R. Stat. Soc. C, 63(3), 405-422.

Shirvani, A., Landman, W.A., 2016. Seasonal precipitation forecast skill over Iran. Int. J. Climatol. 36, 1887–1900. doi:10.1002/joc.4467

Sikder, M.S., Chen, X., Hossain, F., Roberts, J.B., Robertson, F., Shum, C.K., Turk, F.J., 2015. Are General Circulation Models Ready for Operational Streamflow Forecasting for Water Management in Ganges and Brahmaputra River basins? J. Hydrometeorol. 150911144418005. doi:10.1175/JHM-D-14-0099.1

Slater, L.J., Villarini, G., Bradley, A.A., 2017. Evaluation of the skill of North-American Multi-Model Ensemble (NMME) Global Climate Models in predicting average and extreme precipitation and temperature over the continental USA. Clim. Dyn. doi:10.1007/s00382-016-3286-1

Smith, A.F.M., Gelfand, A.E., 1992. Bayesian Statistics without Tears: A Sampling–Resampling Perspective. Am. Stat. 46, 84–88. doi:10.1080/00031305.1992.10475856

Stedinger, J.R., Kim, Y.O., 2010. Probabilities for ensemble forecasts reflecting climate information. J. Hydrol. 391, 9–23. doi:10.1016/j.jhydrol.2010.06.038

Stedinger JR, Vogel RM, Foufoula-Georgiou E, 1993. Frequency analysis of extreme events, in: Maidment, D.R. (Ed.), Handbook of Hydrology. McGrawHill Book Company, New York.

Thober, S., Kumar, R., Sheffield, J., Mai, J., Schäfer, D., Samaniego, L., 2015. Seasonal Soil Moisture Drought Prediction over Europe using the North American Multi-Model Ensemble (NMME). J. Hydrometeorol. 150904104740009. doi:10.1175/JHM-D-15-0053.1

Tian, D., Martinez, C.J., Graham, W.D., Hwang, S., 2014. Statistical Downscaling Multimodel Forecasts for Seasonal Precipitation and Surface Temperature over the Southeastern United States. J. Clim. 27, 8384–8411. doi:10.1175/JCLI-D-13-00481.1

Van den Dool, H., 2007. Empirical methods in short-term climate prediction, Oxford University Press. doi:10.1029/2005GL023422

Wanders, N., Wood, E.F., 2016. Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations. Environ. Res. Lett. 11, 94007. doi:10.1088/1748-9326/11/9/094007

Wang, H., 2014. Evaluation of monthly precipitation forecasting skill of the National Multi-model Ensemble in the summer season. Hydrol. Process. 28, 4472–4486. doi:10.1002/hyp.9957

Wang, H., Reich, B., Lim, Y.., 2013. A Bayesian approach to probabilistic streamflow forecasts. J. Hydroinformatics 15, 381–391. doi:10.2166/hydro.2012.080

Weigel, A.P., Liniger, M.A., Appenzeller, C., 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q. J. R. Meteorol. Soc. 134, 241–260. doi:10.1002/qj.210

Wood, E.F., Schubert, S.D., Wood, A.W., Peters-Lidard, C.D., Mo, K.C., Mariotti, A., Pulwarty, R.S., 2015. Prospects for Advancing Drought Understanding, Monitoring, and Prediction. J. Hydrometeorol. 16, 1636–1657. doi:10.1175/JHM-D-14-0164.1

Figures and tables

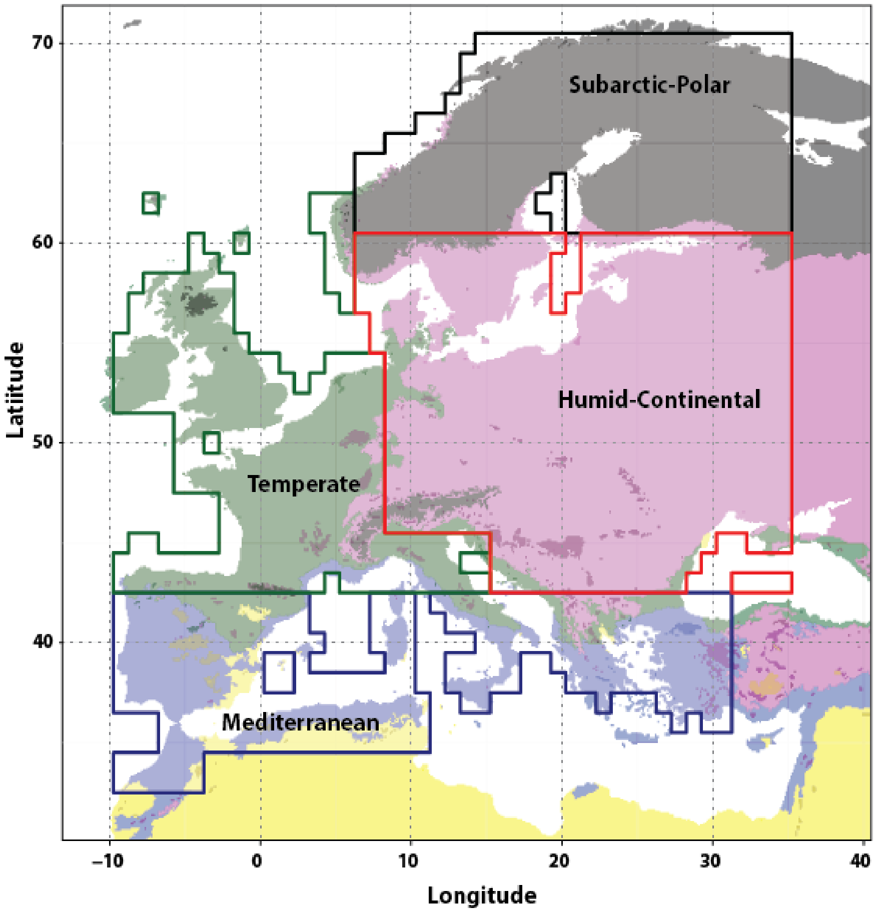

Figure 1. Map of the four biophysical European regions used in the study. Region outlines are based on similar Köppen climate regions and then tailored to the grid cells of the NMME/E-OBS data (E-OBS data are regridded to the same resolution as NMME data, see Section 2). The Temperate region is based on Köppen categories $Cw_{a\text{-}c}$ and $Cf_{a\text{-}c}$; the Subarctic-Polar region is based on $Df_{c,d}$, Dw_c, $Ds_{c,d}$, ET, and EF; the Mediterranean region is based on $Cs_{a,b}$; and the Humid-Continental region is based on $Df_{a,b}$, $Dw_{a,b}$, and $Ds_{a,b}$ (see Peel et al. (2007)).

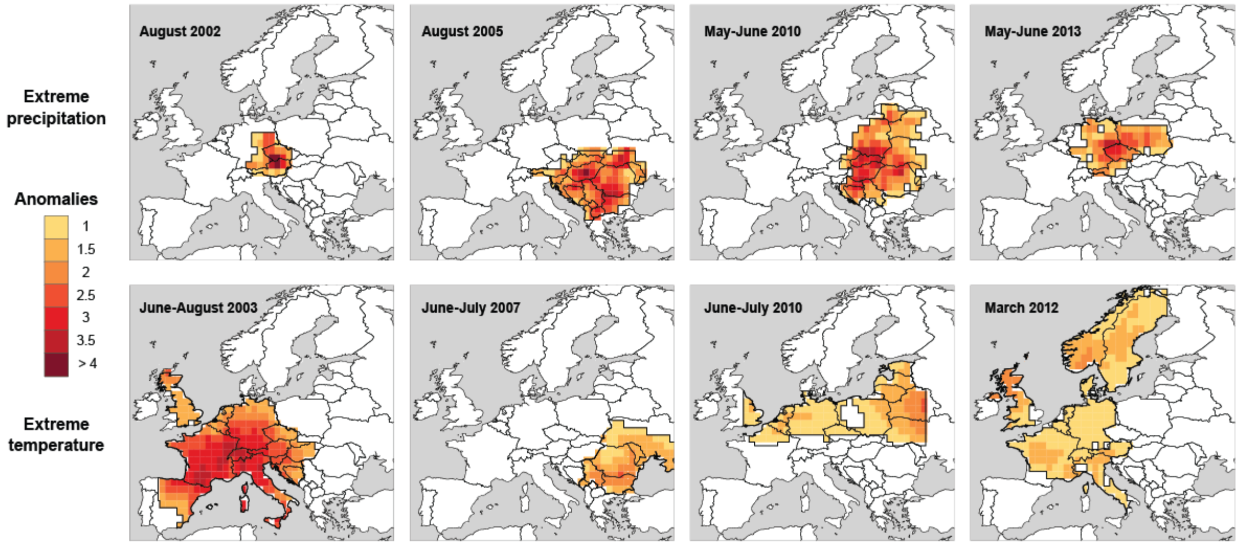

Figure 2. Location of four extreme precipitation and four extreme temperature events across continental Europe. The spatial extent of each event is indicated with a thick black outline, and the magnitude of the climatological anomaly is displayed as yellow/red shades (with darker reds indicating greater anomalies). The anomaly is computed on a pixel-by-pixel level at the monthly or seasonal scale across Europe. Extreme precipitation events are shown across the top row: August 2002, August 2005, May-June 2010, May-June 2013. Extreme temperature events are displayed across the bottom row: June-August 2003, June-July 2007, June-July 2010 and March 2012.

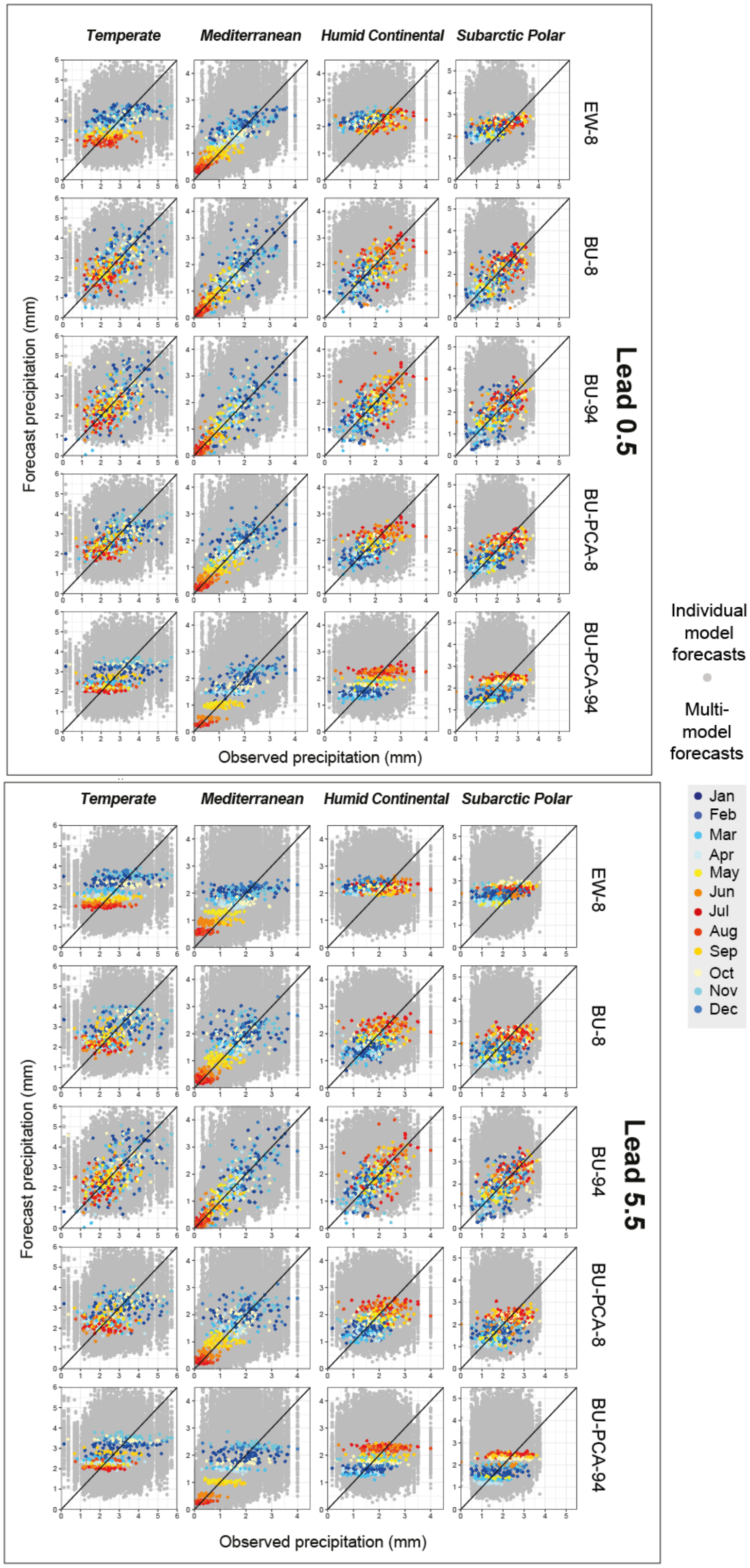

Figure 3. Comparison of the NMME precipitation forecasts before and after multi-model weighting for the 0.5 lead time

(top panel) and the 5.5 lead time (bottom panel). For each of the four regions (columns), five types of weighting procedures are compared (rows): equal weights of the eight single-model ensembles (EW-8), BU of the eight single-model ensembles (BU-8), BU of the 94 model members (BU-94), BU of the principal components of the eight single-model ensembles (BU-PCA-8), and BU of the principal components of the 94 model members (BU-PCA-94). Grey background circles indicate the pooled forecasts from the 94 individual model members (i.e., no distinction is made among the different model members in the figure). Color circles represent the different months of the year, ranging from winter (blue) to summer (red). The one-to-one line is shown in the foreground to highlight the biases in the different approaches.

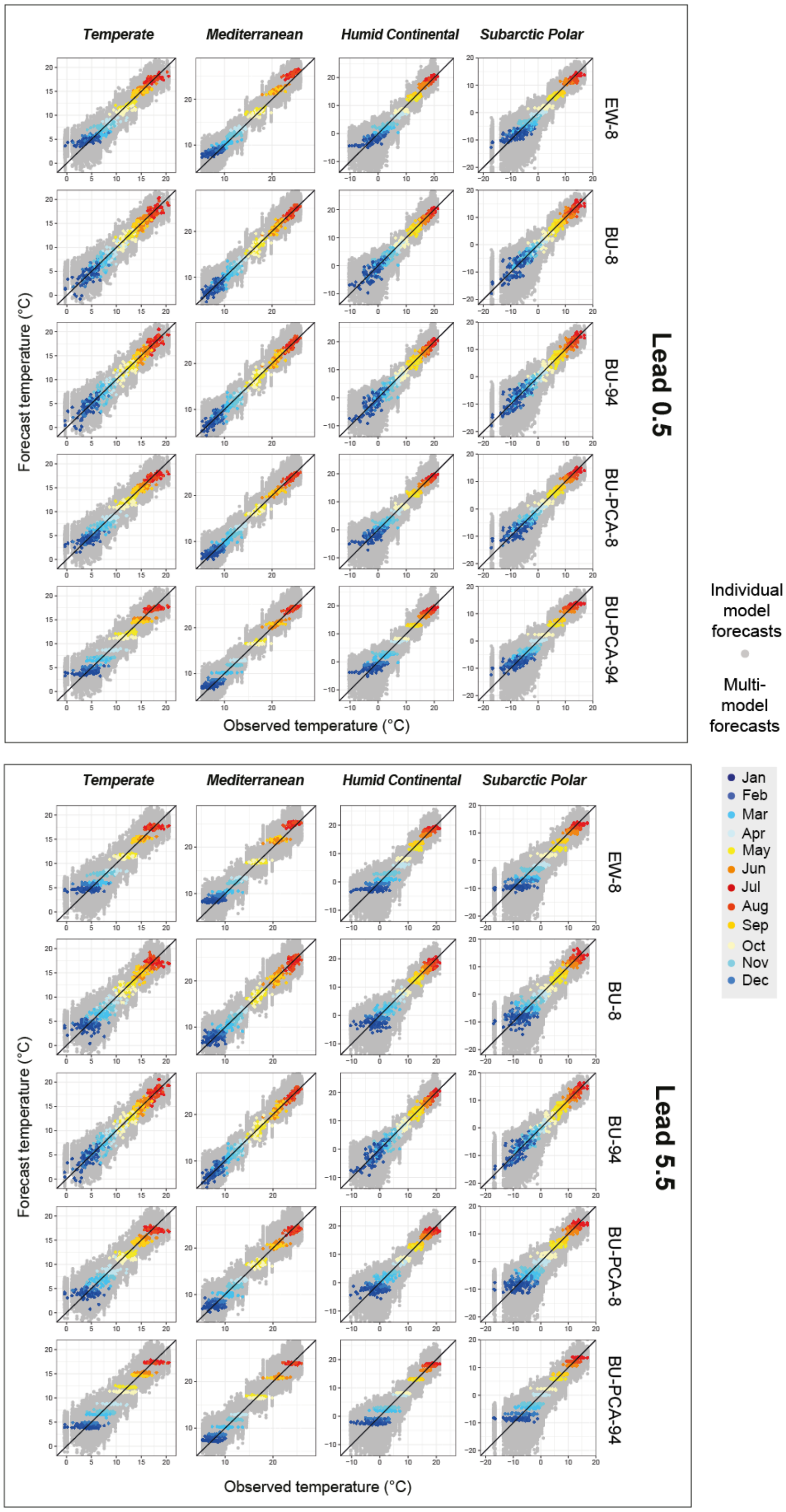

Figure 4. Same as Figure 3 but for temperature.

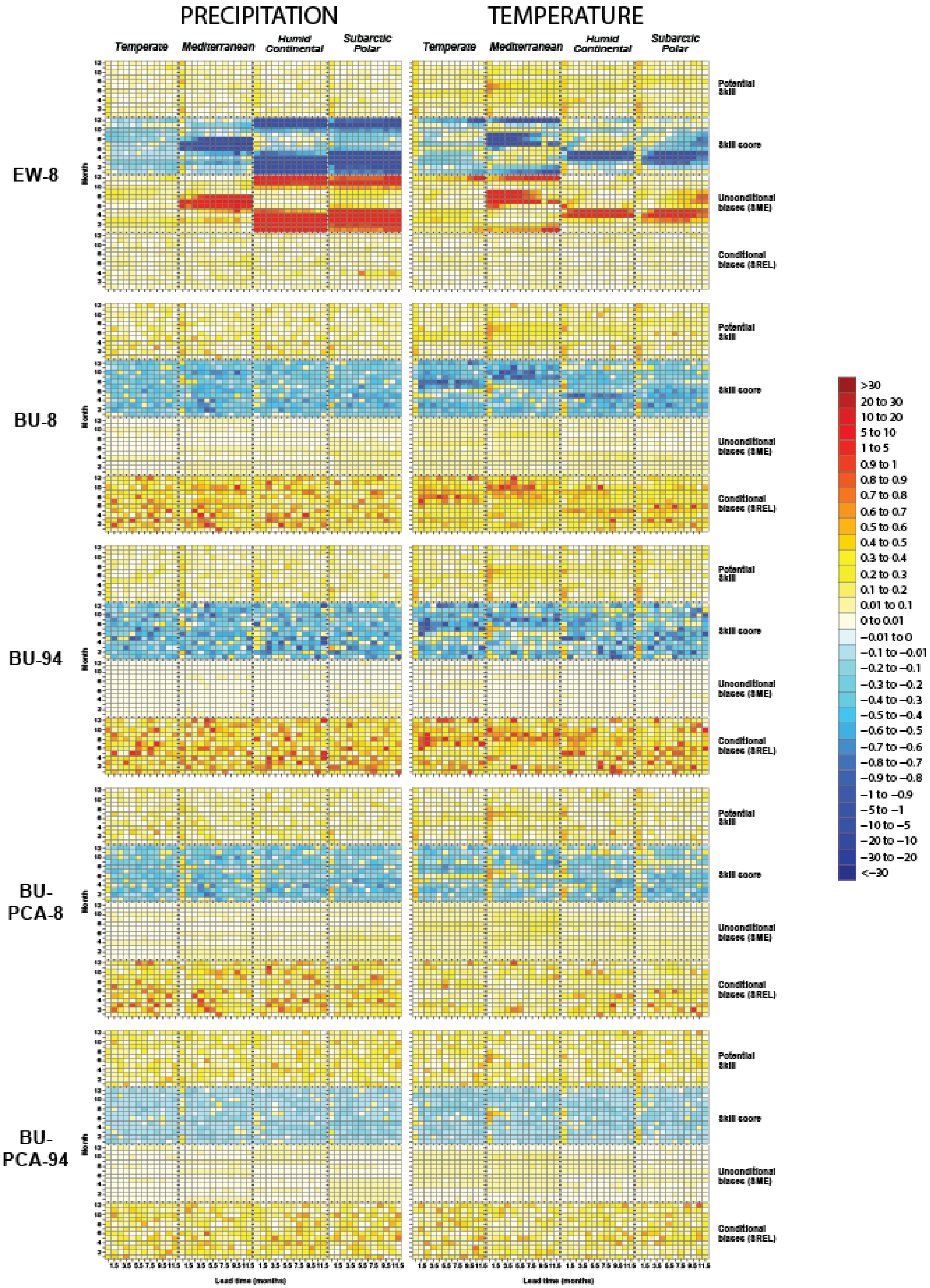

Figure 5. Summary color maps comparing the skill of the five

multi-models for precipitation (left column) and temperature (right column) forecasts. The skill is shown for five multi-models computed using a) Equal weights of the eight single-model ensembles (EW-8, row 1), b) BU of the eight single-model ensembles (BU-8, row 2), c) BU of the 94 model members (BU-94, row 3), d) BU of the principal components of the eight single-model ensembles (BU-PCA-8, row 3), and e) BU of the principal components of the 94 model members (BU-PCA-94, row 4). The potential skill, skill score, unconditional biases (SME) and conditional biases (SREL) (rows) are shown for all four European regions (columns), lead times (x-axes) and months of the year (y-axes). Colors range from negative (blue shades) to neutral (white shades) to positive (red shades).

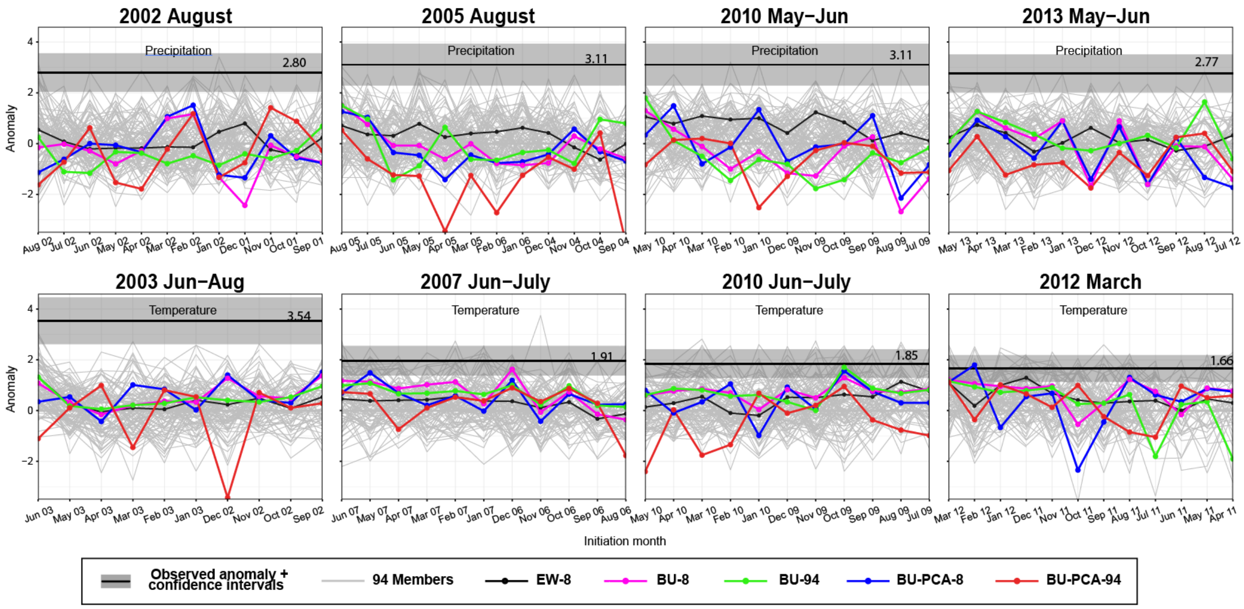

Figure 6. Skill of the 94 NMME model members (grey lines) and of the five multi-models (color lines) in predicting eight individual extreme precipitation/ temperature events, against the observed climatology. The extreme precipitation and temperature events are the same as those represented in **Figure 2**. The horizontal black line indicates the observed E-OBS climatological anomaly, together with the 95% confidence intervals (grey areas; see **Section 3.4**). The anomalies forecast by the 94 individual model members are indicated as thin grey lines in the background. The anomalies of the five multi-models are shown in black (EW-8), magenta (BU-8), green (BU-94), blue (BU-PCA-8) and red (BU-PCA-94). Note that not necessarily all 94 members are always present (some models have gaps, so the multi-models are computed using the available data).

Table 1. Characteristics of the eight NMME models. The available period does not reflect the presence of gaps in the forecasts. The number of ensemble members indicates the largest number of members per GCM and is not reflective of missing data for one or more members. The 0.5-lead time is the shortest available lead time and refers to the forecast for a month issued at the beginning of the month itself (e.g., the 0.5 lead time forecast for January 2000 is issued at the beginning of January 2000). NMME Phase I and Phase II refer to the timescales of the NMME project. The Phase I project was funded in 2011 by NOAA; Phase II was funded in 2012-2013 as an inter-agency project by NOAA, the National Science Foundation, the Department of Energy and NASA. New variables and models were released as part of Phase II, and were made available in 2014.

Model name	Modeling Center	Available Period	Ensemble Size	Lead Times (months)	NMME Phase I	NMME Phase II
CCSM3 (version 3)	NCAR / COLA / RSMAS	1982 - Present	6	0.5 – 11.5	✓	
CCSM4 (version 4 – subset of CESM)	NCAR / COLA / RSMAS	1982 - Present	10	0.5 – 11.5		✓
CanCM3 (3rd Generation)	CMC	1981 - Present	10	0.5 – 11.5	✓	✓
CanCM4 (4th Generation)	CMC	1981 - Present	10	0.5 – 11.5	✓	✓
CFSv2 (version 2)	NOAA / NCEP	1982 – Present	28 (24 used; 4 incomplete)	0.5 – 9.5	✓	✓
GEOS5 (version 5)	NASA / GMAO	1981 - Present	12	0.5 – 8.5	✓	✓
GFDL2.1 (version 2.1)	NOAA / GFDL	1982 - Present	10	0.5 – 11.5	✓	
FLORb01 (version 2.5)	NOAA / GFDL	1982 - Present	12	0.5 – 11.5		✓

Model and modeling center acronyms

CanCM	Canadian Coupled Global Climate Model
CESM	NCAR's Community Earth System Model (successor of CCSM)
CCSM	Community Climate System Model
CFS	Climate Forecast System
COLA	Center for Ocean–Land–Atmosphere Studies
CMC	Environment Canada's Meteorological Service of Canada - Canadian Meteorological Centre
GEOS	Goddard Earth Observing System Model
GFDL	NOAA's Geophysical Fluid Dynamics Laboratory
GMAO	NASA's Global Modeling and Assimilation Office
IRI	International Research Institute for Climate and Society, part of Columbia University's Earth Institute
NCAR	National Center for Atmospheric Research
NCEP	NOAA's National Centers for Environmental Prediction
NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NOAA	National Oceanic and Atmospheric Administration
RSMAS	Rosenstiel School for Marine and Atmospheric Science, University of Miami

Table 2. Coefficients of determination (R^2) for the 94 individual model members ('94 mem') and the five multi-models, when pooling forecasts for all months against E-OBS observed data (1982-2015). These R^2 values correspond to the grey and color scatter plots shown in **Figures 3 and 4**. See **Section 4.2** for a discussion of the difference between 94 mem and EW-8.

		0.5-month lead time						5.5-month lead time					
		RAW: 94 mem.	**EW-8**	**BU-8**	**BU-94**	**BU-PCA-8**	**BU-PCA-94**	**RAW: 94 mem.**	**EW-8**	**BU-8**	**BU-94**	**BU-PCA-8**	**BU-PCA-94**
Precipitation	Temperate	0.11	0.29	0.30	0.29	0.31	0.25	0.07	0.23	0.16	0.12	0.17	0.24
	Mediter-ranean	0.40	0.71	0.68	0.68	0.70	0.69	0.24	0.59	0.54	0.48	0.57	0.60
	Humid-Continental	0.03	0.13	0.36	0.37	0.36	0.30	0.01	0.03	0.22	0.16	0.23	0.27
	Subarctic-Polar	0.07	0.39	0.38	0.39	0.38	0.33	0.04	0.24	0.23	0.21	0.23	0.27
	Means	0.15	0.38	0.43	0.43	0.44	0.39	0.09	0.27	0.29	0.24	0.30	0.34
Temperature	Temperate	0.90	0.95	0.95	0.95	0.96	0.95	0.87	0.93	0.92	0.92	0.93	0.94
	Mediter-ranean	0.95	0.98	0.98	0.98	0.98	0.98	0.92	0.98	0.97	0.97	0.97	0.98
	Humid-Continental	0.91	0.96	0.96	0.96	0.96	0.96	0.87	0.94	0.94	0.94	0.94	0.94
	Subarctic-Polar	0.87	0.95	0.95	0.95	0.96	0.95	0.83	0.93	0.91	0.92	0.92	0.93
	Means	0.91	0.96	0.96	0.96	0.97	0.96	0.87	0.94	0.94	0.94	0.94	0.95