Climate Dynamics manuscript No.
(will be inserted by the editor)

Multi-Model Skill Assessment of Seasonal Temperature and Precipitation Forecasts over Europe

Niti Mishra · Chloé Prodhomme · Virginie Guemas

Received: date / Accepted: date

Abstract There is now a wide range of forecasts and observations of seasonal climatic conditions that can be used across a range of application sectors, including hydrological risk forecasting, planning and management. As we rely more on seasonal climate forecasts, it becomes essential to also assess its quality to ensure its intended use. In this study, we provide the most comprehensive assessment of seasonal temperature and precipitation ensemble forecasts of the EUROSIP multi-model forecasting system over Europe. The forecasts from the four individual climate models within the EUROSIP are assessed using both deterministic and probabilistic approaches. One equally and two unequally Weighted Multi-Models (WMMs) are also constructed from the individual models, for both climate variables, and their respective forecasts are also assessed.

Consistent with existing literature, we find limited seasonal climate prediction skill over Europe. A simple equally WMM system performs better than both unequally WMM combination systems. However, the equally WMM system does not always outperform the single best model within the EUROSIP multi-model. Based on the results, it is recommended to assess seasonal temperature and precipitation forecast of individual climate models as well as their multi-model mean for a comprehensive overview of the forecast skill.

Keywords Seasonal Climate Forecast · Probability Ensemble Forecast · Weighted Multi-Model · Forecast Verification

Department of Earth Sciences, Barcelona Supercomputing Centre
Carrer de Jordi Girona, 29-31, 08034 Barcelona, Spain
E-mail: niti.mishra@bsc.es
E-mail: chloe.prodhomme@bsc.es
E-mail: virginie.guemas@bsc.es

1 Introduction

Recent hydrological extreme events demonstrate the vulnerability of European society to water-related natural hazards and there is a strong evidence that climate change will worsen these events (Lavell et al. 2012, chap1; National Academies of Sciences and Medicine 2016). The impacts of these hydrological extreme events can be reduced by early-warning design support systems (Wanders and Wood 2016). Hydrological simulations in these support systems rely on initial land surface conditions from upstream river flow, snow cover, soil moisture and/or skillful seasonal prediction of continental meteorological conditions, such as temperature and precipitation (Wanders and Wada 2015; Yuan et al. 2016). The predictability of precipitation and temperature is exploited particularly for long-term hydrological forecasts (Velázquez et al. 2009; Yuan et al. 2016) and thus, high-quality Seasonal Climate Forecasts (SCFs) are essential for the success of seasonal hydrological forecasting based on climate models.

SCFs are forecasts of climate conditions at timescales of a few weeks up to a few months, for statistics such as monthly/seasonal averages of temperature and/or precipitation or frequency of occurrences of extreme events. SCFs are possible due to the long-term predictability of the oceanic circulation (i.e. up to a few years) and by the fact that the variability in tropical Sea Surface Temperature (SST) has a significant global impact on the atmospheric circulation (Balmaseda and Anderson 2009; Doblas-Reyes et al. 2013). Considerable efforts have been made in the field to better represent the coupled ocean-atmospheric dynamics and to improve the operational Climate Forecast Systems (CFSs) such as the National Centers for Environmental Prediction (NCEP; Saha et al., 2014) and the Predictive Ocean Atmosphere Model for Australia (POAMA; Colman, 2005), which are single-model CFSs as well as the European Multimodel Seasonal to Interannual Prediction (EUROSIP; Vitart et al. 2007; EUROSIP 2016; Stockdale 2013) system, which comprises of four independent CFSs.

Generally, forecast skill of seasonal climatic conditions in areas influenced by ENSO is higher than in the extra-tropical regions (Alexander et al. 2002; Kumar et al. 2013; Palmer et al. 2004; Sordo et al. 2008). In Europe, stratospheric processes (Bell et al. 2009), snow cover (Senan et al. 2016), soil moisture (Prodhomme et al. 2016) and sea-ice (Guemas et al. 2016) are also proven to be effective sources of predictability. Recently, the North Atlantic Oscillation (NAO) has also been reported as an important source of predictability for European winter climate (Athanasiadis et al. 2017; Scaife et al. 2014). Yet, the overall seasonal forecast skill over Europe for surface variables is still quite low (Arribas et al. 2011; Kim et al. 2012; Scaife et al. 2014).

An ensemble forecast is a set of forecasts that generate a range of future climate possibilities. Ensemble forecasts are often preferred over deterministic ones because they can convey the uncertainties that arise due to the inability to accurately model atmospheric dynamics and the initial condition uncertainty (Hawkins and Sutton 2009, 2011; Lorenz 1963; Palmer et al. 2004; Tebaldi and Knutti 2007). To obtain ensemble forecast, a climate model is run multiple times, each time with slightly different initial conditions and with slightly perturbed numerical models. Each forecast in an ensemble, known as a member, are then used to calculate the probability distribution of the potential near-term future climate (Bröcker and Smith 2008; Fortin et al. 2006; Wilks 2006).

In addition to ensemble prediction, combining ensembles of multiple CFSs to make climate predictions has also garnered a lot of attention (Doblas-Reyes et al. 2003; Palmer et al. 2005; Rodrigues et al. 2014; Weigel et al. 2010; Yun et al. 2003). The EUROSIP multi-model, which became operational in 2005, is the result of DEMETER (Palmer et al. 2004) and other research projects that confirmed the scientific benefits of combining forecasts from several climate models. Such multi-model predictions address issues of structural uncertainties within models that arise due to incomplete physical parameterizations and numerical approximations (Palmer et al. 2004). In general, equally weighting each of the CFSs has been recognized to have consistently better performance than that of the individual models (DelSole 2007; Hagedorn et al. 2005; Kharin and Zwiers 2002; Peng et al. 2002). This is because random individual model errors tend to compensate one another and the robust predictable signal tend to stand out by averaging across a number of models. This is particularly important for medium-to-long range forecasting, where the timescale over which model errors accumulate are much longer and can significantly degrade long-term forecasts. On the other hand, improved performance of unequally Weighted Multi-Model (WMM) systems have also been reported (Krishnamurti et al. 2000; Robertson et al. 2004; Rodrigues et al. 2014; Wanders and Wood 2016). Intuitively, it makes sense to give more (less) weight to forecasts from a model that has consistently better (poor) historical performance. However, consensus on the optimal way of weighing the different models has yet to be reached (DelSole et al. 2013; Tebaldi and Knutti 2007).

Whether we use deterministic, probabilistic or some weighted combination of forecasts from multiple models, they are ultimately beneficial only if they have skill and can add value to the users. The objective of this study is to assess the seasonal forecasting skill of each of the forecast system in the EUROSIP multi-model and to compare it with that of equally and unequally WMMs in order to provide users in hydrology an overview on current potential and/or limits of seasonal temperature and precipitation predictability over Europe. A systematic investigation across different models of EUROSIP and for different seasons specifically over the European region is still lacking and this study aims to highlight the need for further studies by contributing to the limited extant literature. The assessment is done for winter and summer, temperature and precipitation forecasts over a period of 21 years (1992-2012) in terms of the Anomaly Correlation Coefficient (ACC) for deterministic forecasts and the Continuous Ranked Probability Skill Score (CRPSS) for probabilistic forecasts on each grid point.

The article is structured as follows: Section 2 describes the datasets used, the methods applied to assess the forecast skill and to construct the WMMs.

Section 3 presents the subsequent results followed by a discussion and a final conclusion in Section 4 and 5, respectively.

2 Data and Methods

2.1 Data

This study relies on a comprehensive set of seasonal temperature and precipitation forecasts from the EUROSIP multi-model over the European region specified as 20°W-70°E and 25°N-75°N, for the period 1992-2012. Four individual CFSs – Global Seasonal forecasting system version 5 (Glosea5) from Met Office, System4 of European Center for Medium Range Weather Forecasts (ECMWF), System2 of NCEP and System5 of Meteo France (MF), are integrated into one common EUROSIP multi-model framework (EUROSIP 2016). These are the common choice of operational multi-model in the European region (Soares and Dessai 2015; Stockdale 2013) and we select the longest available hindcast period in common for this study. The number of ensemble members and the horizontal resolutions of these four climate models are given in Table 1. More details on the dynamical cores and the physical parameterizations of individual models within EUROSIP can be found in their corresponding documentations (MacLachlan et al. 2015; Molteni et al. 2011; Saha et al. 2014; Voldoire et al. 2013).

[Table 1 about here.]

The reference dataset for temperature is obtained from the ERA-Interim (ERAINT) database, which includes a 4D variational analysis with a 12-hour analysis window (Dee et al. 2011). The spatial resolution of the dataset is $\cong$ 80 km (T255 spectral) on a reduced Gaussian grid with 60 vertical levels from the surface up to 0.1 hPa (Dee et al. 2011). The results (not presented) are insensitive to the comparison with the observation dataset from Global Historical Climatological Network (GHCN 2.2). For precipitation, the reference dataset is provided by the Global Precipitation Climatology Project (GPCP), which comprises a gridded analysis based on gauge measurements and satellite estimates of precipitation (Adler et al. 2003).

The original values of both forecasts and observations are interpolated using a bilinear interpolation to match the coarsest grid among each climate variable. The coarsest grid is chosen as a preferred grid for such interpolation method (Starks et al. 2003). All computations are done on grid point by grid point basis. The sea points are masked and only data over land is assessed.

The forecasts from the four models and the reference datasets are available with monthly averages of daily mean temperature and precipitation values for the period 1992-2012. The study is performed at seasonal (average of three months) timescale for winter and summer seasons. Winter consists of forecasts from December to February (DJF) while summer consists of forecasts from June to August (JJA). All forecasts are initialized around the first day of the month preceding the target season. These particular seasons and years are selected for study because a homogeneous history of hindcasts on a monthly timescale across all four participating forecast systems is available only for this time period. This is an important limitation of the study because a short time series of 21 years usually cannot accurately account for the sensitivity of climate system performance to the chaotic nature of climate, which differs greatly within the various regions of Europe. The data limitation also extends to the verification metrics used and the methodologies applied to combine forecasts as statistics derived from limited number of data is expected to suffer from uncertainty due to sampling error. Longer common period hindcasts are essential in studies to allow the results of analysis to be extended.

All calculations in this study are applied to the forecast anomalies computed with respect to model's own climatology. Therefore, the ability to predict departures from the seasonal cycle is measured rather than the absolute values of temperature and precipitation. Thus, the model bias does not appear in the verification metrics (or only indirectly since it might affect the variability).

2.2 Methods of Verification

2.2.1 Anomaly Correlation Coefficient (ACC)

In this study, for all deterministic forecasts i.e. the ensemble mean, ACC is used to assess the forecast skill. ACC is the most widely used skill metric for SCF quality (Doblas-Reyes et al. 2013; Fricker et al. 2013; Scaife et al. 2014), due to its invariant property (i.e. not affected by certain data transformation). ACC assesses the degree of linear correspondence between the target forecast anomalies and the anomalies of the observed climate variable. Additionally, for linearly re-calibrated forecasts, the squared ACC is equivalent to the mean squared skill score (Siegert et al. 2017). It is worth noting however, that correlation coefficient are extremely noisy in smaller sample sizes, meaning small changes in forecast values can impact correlation skill significantly. We test the significance of ACC at 5% significance level, controlling the False Discovery Rate (FDR) (Benjamini and Hochberg 1995).

2.2.2 Continuous Ranked Probability Skill Score (CRPSS)

The second metric selected to assess forecast skill is the Continuous Ranked Probability Score (CRPS), which is a standard measure for assessing the accuracy and reliability aspects of probabilistic forecasts. CRPS evaluates the predictive skill of the full probability distribution of forecast obtained from the ensemble members (Hersbach 2000; Matheson and Winkler 1976). Such evaluation is desirable since climate forecasts are used as forcings in models such as the hydrological models (Boucher et al. 2009; Candille and Talagrand 2005; Gneiting et al. 2005; Murphy 1969).

Given that F is the cumulative density function of ensemble forecasts and y is the value that actually occurred, the CRPS is defined as:

$$CRPS(F, y) = \int_{-\infty}^{\infty} \left[F(t) - H(t-y)\right]^2 \, dt, \tag{1}$$

where $H(t-y)$ denotes the Heaviside function that takes the value of 0 when t < y and 1 otherwise (Hersbach 2000; Matheson and Winkler 1976). Thus, the CRPS measures the difference between the predicted and observed cumulative distributions. For deterministic forecasts, the average CRPS becomes the mean absolute error and therefore, has similar interpretation.

The skill score based on CRPS is CRPSS, computed as:

$$CRPSS = \frac{CRPS_f - CRPS_{clim}}{CRPS_{perf} - CRPS_{clim}}, \tag{2}$$

where $CRPS_f$, $CRPS_{clim}$ and $CRPS_{perf}$ stand for CRPS of the forecast in question, of the reference/benchmark forecast and that of the perfect forecast, respectively. In this study, climatology is used as the reference forecast, which refers to the average conditions over some recent reference period. Skill scores below 0 are unskillful compared to a naïve climatological forecast. Those equal to 0 are no better than that of climatology and anything above 0 (up to 1) signals an improvement upon climatology. The standard deviation of the skill score is approximated by propagation of uncertainty and the significance is measured at 95% confidence interval.

2.2.3 Fair Continuous Ranked Probability Skill Score (FCRPSS)

One drawback of the CRPS is that it inflates the score for models with higher number of ensemble members. To correct for this, Ferro et al., (2014; 2008) recommended the Fair Continuous Ranked Probability Score (FCRPS), which evaluates the underlying ensemble distribution and is independent of the empirical distribution of the ensemble members (Fricker et al. 2013). Results of FCRPSS (skill score based on FCRPS) are also provided.

2.3 Methods for Weighted Multi-Model (WMM) Combination

An important objective of this study is to combine forecasts from dynamical systems to estimate a single optimal forecast with an aim to understand benefits of such combination on the overall forecast quality. Separate models are established for each season and grid cell independently. Three methods of combinations are used in this study:

2.3.1 Multi-Model Mean (MMM)

The first combination approach consists of an equally WMM system, which is obtained by averaging ensembles of each CFS and then again, averaging these four ensemble means to obtain a multi-model ensemble mean anomalies. Hereinafter, this method is referred to as the Multi-Model Mean (MMM). This is one of the most commonly used method to combine forecasts of independent CFSs (DelSole et al. 2013; Kharin and Zwiers 2002; Krishnamurti et al. 2000). The basic idea behind this approach is the assumption that each individual CFS is equally likely to represent the truth whatever its performance (Wanders and Wood 2016).

2.3.2 Best OLS Combination Method (BOCM)

Various forms of regression have been tested on seasonal and weather forecasts to obtain optimal weights based on historical performance of the model (DelSole et al. 2013; Kharin and Zwiers 2002; Rodrigues et al. 2014; Weigel et al. 2008). The second method uses the Ordinary Least Squares (OLS) regression technique to obtain optimal weights. 15 possible OLS models are built out of the ensemble mean of each of the four available CFSs - one Multiple Linear Regression (MLR) model with all four CFSs, four MLR models with only three CFSs, six MLR models with only two CFSs and four linear regression models with only one CFS. For each of these 15 OLS combinations, ensemble mean of the participating CFS(s) are regressed onto the corresponding observations and their respective weights are the regression coefficients estimated from the data. Out of the 15 possible OLS models, the one that has highest correlation with the observation dataset is chosen as the Best OLS Combination Model (BOCM) for each grid point.

2.3.3 Correlation As Weight Method (CAWM)

The final weighted combination method uses as weights the ACC value between ensemble mean anomalies of each CFS and the anomalies of the observation. While correlation does not take into account the system performance in terms of variance, it is often the value relied upon for forecast verification. In addition, correlations are indicative of model performance and thus, it is reasonable to think of correlation values as potentially trustworthy weights. Note that this method may choose a CFS with only a minor correlation improvement among the competing CFSs. Here, the ACC value of each CFS is first multiplied to its respective forecast value. They are then added together and divided by the sum of their ACCs to standardise the forecast value. Hereinafter, this model is referred to as Correlation As Weights Model (CAWM).

2.3.4 Evaluation of Optimal Weights

Historical data is required to not only build statistical models but to also evaluate them. In this study there are only 21 years of records, which is considered not long enough to develop and validate regression-based model. However, a homogeneous history of hindcasts on a monthly timescale across all four participating forecast systems is available only for this time period. This is a major limitation of this study, albeit common in seasonal climate forecasting (Kumar 2009; Shi et al. 2015). To address the issue of small sample size, we apply leave-one-out cross-validation procedure in both WMMs (Efron 1983; Molinaro et al. 2005). This means for each forecast year, the model weights are estimated from the other 20 years of data and a seasonal forecast is made for that year. The process is repeated over each of the 21 years and the resulting hindcasts are then compared with the corresponding observation. An accuracy estimate obtained using leave-one-out cross-validation is known to be almost unbiased but has high variance (Chapelle et al. 2002; Efron 1983).

Another possible reason for unstable weights in linear regressions is the collinearity among the predictors, which can be dealt with by ridge regression (DelSole et al. 2013). However, multicollinearity among EUROSIPs CFSs was found not to be high enough to pursue further (See Fig. S1-2 in supplementary section).

Finally, the weights in both WMMs are constrained meaning both zero or negative weights are not allowed in the model. Model with unconstrained weight was tested by Wanders and Wood (2016) but omitted eventually due to poorer performance. This is because unconstrained models give rise to overconfident estimates of weights when the number of sampling years is small. Besides, it is reasonable to assume constrained weights because a CFS that consistently lacks skill for any given region can be removed from the combined model.

3 Forecast Skill Assessment of EUROSIP

3.1 Assessment of Individual Model Ensemble Mean Anomalies

3.1.1 Strength and Weakness of the Individual Models

This section focuses on the evaluation of the prediction skill of the individual CFSs of EUROSIP in terms of ACC. Figure 1 shows the ACC of seasonal temperature anomalies for both winter and summer seasons. There is a difference in skill exhibited by the four models between the two seasons. Glosea5 has some skill over Europe during winter, although it is not statistically significant. ECMWF has high statistically significant skill over the British Isles, Southern Sweden and parts of Central Europe during winter. NCEP exhibits some statistically significant skill over the North-Eastern Europe (close to Barents sea) and the skill of MF is significantly higher during winter over the Western Europe, the British Isles and the south of Scandinavia.

During summer, Glosea5 has notably higher, statistically significant skill over the Northern Scandinavia (close to the Norwegian sea) and the Southern Europe. ECMWF exhibits higher statistically significant skill notably over large parts of the Southern and the Eastern Europe. For NCEP, the summer seasonal temperature skill is higher over the East-Central Europe. MF exhibits limited skill (although not significant) mostly over the Southern regions during summer.

[Fig. 1 about here.]

[Fig. 2 about here.]

Figure 2 shows the results of the same evaluation as that of Figure 1 but for seasonal precipitation. It can be noted that the skill across Europe for seasonal precipitation is very low and sporadic. The skill is higher in winter for Glosea5 and ECMWF, mostly concentrated over the Eastern Europe. During winter, NCEP shows some significant skill over the Northern Scandinavia and MF exhibits significant skill over most of Scandinavia and over the British Isles. Higher skill in winter could be because precipitation is hard to both observe and to forecast, given its high variability during the dry months of summer. Conversely, winter precipitation are more dependent on large scale circulation, such as the NAO that has recently shown to have predictability and could be one of the sources of skill here (Scaife et al. 2014; Trigo et al. 2002).

For summer, Glosea5 and ECMWF exhibit significant skill over the Mediterranean region. NCEP has some significant skill over North-Eastern Europe. The summer seasonal precipitation skill of MF is notably low over Europe. The skill pattern for seasonal precipitation must be considered with caution however, because in regions with limited rain, small changes in observed precipitation can greatly impact correlation values. Thus, more evidence is needed to make conclusions about EUROSIP's seasonal precipitation forecast skill over Europe.

3.1.2 Utilizing Differences of Individual Model Skill

An important benefit of using multiple models is their potential capability to complement each other. It is unknown a priori which model performs best in which region. Thus, the different levels of skill of the different models can be exploited in an operational context. Figure 3 shows which model has the highest correlation at each grid point. While MF has an overall relatively low seasonal prediction skill in summer, it is in fact the only model among the EUROSIP multi-model that has high skill during winter over central Europe and Southern Scandinavia for seasonal temperature and over the British Isles and south of Scandinavia (close to North and Baltic seas) for seasonal precipitation. Thus, if a strategy to choose the best model for each region is adopted, MF would add value to the overall EUROSIP multi-model as the preferred model for these regions for the winter season.

[Fig. 3 about here.]

For seasonal temperature during summer, it can be seen that different models are preferred over different regions of Europe. Glosea5 is the preferred model for seasonal temperature over the Scandinavian region, ECMWF over the Mediterranean region and NCEP over the British Isles and East-Central Europe. The skill for seasonal precipitation forecast is mostly noisy and scattered among the four models. The superior correlation of MF for winter over Western Europe, the British Isles and parts of Scandinavia can be noticed.

3.2 Assessment of Probabilistic Ensemble Forecasts

3.2.1 Ensemble Performance of Individual Models

While the ensemble mean in general is the best available estimate of future conditions, the Probabilistic Ensemble Forecasts (PEFs) can provide further information about the distribution of the potential outcome of a prediction. This can be verified using a skill score based on the CRPS, which assesses the relative improvement of the PEFs over climatology to reliably and accurately predict differing observations (Gneiting 2011). In this section, the PEFs drawn from each individual EUROSIP model are assessed using the CRPSS. Figure 4 shows the prediction skill of individual EUROSIP models based on CRPSS for seasonal temperature. During winter, ECMWF exhibits low but significant skill over the British Isles. Other than that, the skill is very limited for all models during winter. For summer, the models exhibit CRPS skill over similar regions as where they exhibited ACC skill, although the skill based on CRPS is lower due to its stringent scoring rule. Glosea5 exhibits skill over Northern Scandinavia and Southern Europe. ECMWF exhibits skill over the Mediterranean regions and South-Eastern Europe. NCEP exhibits very limited significant skill over the East-Central Europe and MF does not exhibit statistically significant CRPS skill over Europe.

We noted earlier that CRPS is known to inflate skill for models with larger ensemble size and ECMWF with the largest ensemble size indeed exhibits the highest skill based on CRPS over most of Europe for seasonal temperature. To verify whether this high skill of ECMWF is due to its larger ensemble size, we calculated the proposed FCRPSS (Ferro 2014; Fricker et al. 2013) for each individual models and show similar results in Figure 5. Additionally, we calculated CRPSS for all individual models using first, only 9 members, and then 15 members (See Fig. S3-4 in supplementary section). Based on these results, we note two things - (1) even with the reduced ensemble size, ECMWF still exhibits higher significant skill and (2) when accounting for the ensemble size, the skill of all CFSs remains over the same regions but is lowered. Thus, the higher skill of ECMWF cannot be attributed solely to its larger ensemble members. This is also true for Glosea5 and NCEP with ensemble size greater than that of MF as well as for MF, when comparing between its 9 and 15 ensemble members. Hence, these models are accurately and reliably capturing atmospheric dynamics to the extent that they perform better than climatology over the parts of Europe where they exhibit skill.

[Fig. 4 about here.]

[Fig. 5 about here.]

Figure 6 shows the results of the same evaluation as that of Figure 4 but for seasonal precipitation. It is seen from these maps that none of the individual CFSs accurately predict seasonal precipitation over Europe for both seasons. Additionally, no noticeable change in skill is found in terms of FCRPSS (see Figure 7) as well as CRPSS with just 9 and 15 ensembles members (see Fig. S5-6 in supplementary section). Thus, PEFs of EUROSIP for seasonal precipitation show very limited skill over Europe.

[Fig. 6 about here.]

[Fig. 7 about here.]

3.2.2 Multi-Model Ensemble Performance

The PEFs of the EUROSIP multi-model for seasonal temperature and precipitation are obtained by taking ensemble anomalies from all four CFSs (118 for winter and 114 for summer). Then, the CRPSS is calculated for the resulting multi-model PEFs to assess whether such multi-model provides higher prediction skill than the single best CFS. Based on the results shown in Figure 8, a superior predicting skill is not gained by this combination method with respect to the single best model. For both seasons and for both climate variables, the CRPSS of the multi-model PEFs is mostly lower than that of the single best model. For winter seasonal temperature over the British Isles, the CRPSS is much higher for ECMWF than it is for the multi-model PEFs. Similarly, significant skill exhibited by PEFs of Glosea5 over Northern Scandinavia for seasonal temperature during summer is not seen in the multi-model PEFs. Some significant skill is gained by multi-model PEFs in parts of Southern Europe and Central Europe during winter and summer, respectively. However, overall the decrease in skill is more evident when compared to that of the single best model. This is seen more clearly in Figure 9, where the maps show only the maximum positive CRPSS among the individual CFSs and the multi-model for seasonal temperature (not shown for seasonal precipitation due to very low skill).

[Fig. 8 about here.]

[Fig. 9 about here.]

3.3 Assessment of Optimal Method of Combining Forecasts

In this section, weighting each of the individual CFS of EUROSIP with different techniques is tested to determine to which extent the past performance of these CFSs can be utilized to make better predictions of seasonal climate. The maps here evaluate the prediction skill of the three WMM systems (as described in Section 2c) for seasonal temperature (Figure 10) and seasonal precipitation (Figure 11) over Europe. As seen in both figures, a simple MMM is the best combination for SCFs over most of Europe and its predicting skill is significantly higher during summer than winter. Some exceptions can be seen however. For example, BOCM model performs better over the British Isles (most notably over Ireland) during winter. During summer, the skill of BOCM is also higher over parts of Spain, Northern Scandinavia, north of Eastern Europe and over the south of Black sea. The CAWM model shows limited competing skill, although higher skill over the British Isles and along the coastlines of North sea over Western Europe.

When compared to that of the single best models on any given grid point, it is difficult to interpret the skill of these WMMs as it is sensitive to the location and the combination method. Over the entire region of Europe as shown in the map, notably fewer negative skill is exhibited by MMM during winter and by CAWM during summer. However, seasonal temperature skill exhibited during winter over East-Central Russia, the British Isles, Northern coastlines of Russia and Western Europe by Glosea5, ECMWF, NCEP and MF, respectively, is not surpassed by any of the WMMs. During summer, CAWM model exhibits higher skill over the Northern coastlines of Western Europe than that of NCEP, although the skill is not significant. Over south of Europe below Black sea, the skill exhibited by BOCM is higher only by a slight margin.

[Fig. 10 about here.]

[Fig. 11 about here.]

Based on these results, the additional post-processing of forecast data to obtain optimally weighted forecasts is hard to justify. Although, it is also important to note the limitation of these WMMs due to the small sample size. The superiority of a simple MMM technique compared to the unequally WMM in predicting seasonal climate could be attributed to the small sample size available, which for linear combinations methods, is usually not enough to attain robust weights. The coefficients in the linear regression models adapt to the unpredictable variability within the available historical training dataset and thus, performs poorly in the independent data when sample size is limited. Similarly, in the CAWM technique, estimated correlation coefficients can exhibit large uncertainty (Bellprat et al. 2017) and thus, stand as highly volatile weights. Besides, when atmospheric internal variability is large, as in the case of Europe, more information may be lost by inappropriate weighting and therefore, equal weighting may be safer to use. Finally, optimally WMM generally perform better than the single best models only when there is too little information provided by other CFSs within the multi-model (Rodrigues et al., 2014a; 2014b). However, in the case of EUROSIP, each CFS exhibit varied prediction skill based on climate variable, season and region.

4 Discussion

Most of the research on the assessment of forecast skill of SCFs by climate models is done globally and lower skill over the extratropics than in the tropics is reported. Very few studies focus exclusively on the European region. The H2020 IMPREX European project has offered an opportunity to study the performance of SCFs over Europe focusing on their usability in adaptation to water-related climate hazards. Robust assessment is needed to objectively evaluate whether SCFs are fit for purpose. To this end, this study aims to fill an important gap in the literature by quantifying the skill of SCFs over Europe using the EUROSIP multi-model and its component operational forecast systems – Glosea5, ECMWF, NCEP and MF.

The evaluation is done for seasonal temperature and precipitation forecasts over a period of 21 years (1992-2012) using ACC for deterministic forecasts and CRPSS for probabilistic forecast. The assessment is applied to forecast anomalies (with respect to the models own climatology) against that of observations provided by ERAINT and GPCP for temperature and precipitation, respectively. We also constructed one equally and two unequally WMM systems to evaluate the prospects of model weighting in the context of improving SCFs. Based on the results of this study, limited predictive skill of seasonal temperature and very low skill of seasonal precipitation forecast is found over Europe.

For seasonal temperature, the forecast skill of competing models differ based on both regions and seasons. The predictability is higher during summer than winter. The higher skill based on ACC during summer for seasonal temperature forecasts can be associated with the warming trend (results not shown) as the data has not been detrended (Doblas-Reyes et al. 2013). The higher skill during summer can also be because interannual variability tends to be weaker for temperature during summer (Doblas-Reyes et al. 2000). The skill of MF over the Western Europe during winter can be attributed to the extreme soil moisture conditions in South-Central Europe (van den Hurk et al. 2012). The low predictive skill for European winter was also reported by Wehrli et al. (2017), who assessed the seasonal temperature forecasts of ECMWF back to 1981. This low skill can be associated with blocking events or winter meteorological perturbations as well as the misrepresentation of teleconnections leading to erroneous NAO-related signals (Doblas-Reyes et al. 2003; Kim et al. 2012; Wehrli et al. 2017). Scaife et al.(2014), however, showed NAO forecasting skill in Glosea5 and thus, attributed the low skill over temperature to overdispersion, suggesting that increasing the ensemble size would increase forecast skill. However, we show that the skill of a multi-model PEF, which combines ensembles from all four models, did not surpass the skill of the single best model over most of the grid points.

In the case of winter precipitation, MF has some significant correlation skill over the British Isles, parts of Scandinavia and Italy. Some skill over parts of Eastern Europe is found in ECMWF and Glosea5 during winter, and in NCEP and MF during summer. Other than that, seasonal precipitation predictability over Europe is remarkably low. The results of this study are consistent with that of Brankovič and Palmer (2000), Doblas-Reyes et al. (2000), Graham et al. (2000), Lavers et al. (2009) and Wehrli et al. (2017), all of whom also assessed the forecast skill of seasonal precipitation and showed lower skill over the extratropics than in the tropics, although not negligible. This predictability for seasonal mean precipitation over Europe is attributed to ENSO and local North Atlantic SST forcings. (Doblas-Reyes et al. 2013; Frías et al. 2010; Lloyd-Hughes and Saunders 2002).

Additionally, our results highlighted the benefits of having a multi-model system given that the skill of each individual CFS varied across locations, seasons and climate variables. The difference in skill exhibited by different models also provides context to further examine and understand the climate phenomena that are differently represented by these models. However, the differing number of ensemble sizes among the CFSs complexifies the comparison between the models. In terms of probabilistic skill assessment using CRPS, it was highlighted by Ferro et al. (2014; 2008) that CRPS inflates the score for models with higher number of ensemble members. We tested the effect of ensemble size on superior predictability by models with higher number of members. First, we calculated the recommended FCRPSS that corrects for systematic bias in skill scores induced by the finite ensemble size (Ferro 2014; Ferro et al. 2008; Fricker et al. 2013). Then, we also compared the skill based on CRPS taking equal numbers (9 and 15) of ensemble members for all CFS. The results show that ECMWF still has higher prediction skill over most of Europe compared to other CFSs. Additionally, we show that increasing the ensemble size contributes to higher and more significant prediction skill, not only for ECMWF, but for all CFSs.

Varied levels of skill among the models has also garnered an interest to combine their forecasts based on historical performance of individual CFSs. In this study however, the multi-model PEFs of EUROSIP did not present a considerable improvement in CRPSS when compared to that of the single best model. This could be because the optimal set of models in the ensemble may vary in time for continuous forecasting scenarios as argued by Krikunov and Kovalchuk (2015). They suggested a method to dynamically select ensemble members in multi-model and demonstrated slight improvement of forecast skill. Further improvement opportunities also lie in the use of different machine learning techniques and artificial neural networks for such dynamic selection procedures.

In case of deterministic multi-models, the two unequally weighted methods, BOCM and CAWM, did not always outperform the simple equally weighted MMM. When compared to the single best model, the skill of WMMs is sensitive to the location and combination method and thus difficult to interpret. The limited sample size in this study can be attributed to this sensitivity. A sample size of 21 years is usually not enough to attain robust weights with linear regression or correlation methods. In principle, optimal weighting can optimize the predictive skill (DelSole et al. 2013). However, the ideal way to combine these forecasts is still far from being trivial. These results are also consistent with Rodrigues et al.(2014), who also found that equally or unequally WMM approaches does not always perform better, especially when a models with lower skill are present in a multi-model system.

In contrast to this, Wanders and Wood (2016) showed considerable improvements through unequally WMM. This can be attributed to their assessment of SCFs skill for longer time-period of 30 years (and over large-scale area averages), which yields temporally and spatially stable relationships and allows more robust estimation of weights. Thus, larger homogenous dataset of competing individuals CFSs is critical for the performance of WMMs. The Copernicus Climate Change Services (C3S) by the Copernicus programme comprise all year round seasonal climate forecasts from several state-of-the-art seasonal prediction systems for longer a period. This offers an opportunity to extended the analysis of this study in the near future.

All verification scores in the study are obtained from forecast anomalies to filter out the systematic biases existent in climate models that can lead to underestimation of true predictability. However, anomaly correction only partially removes these biases. The systematic errors in the model variability can be possibly accounted for by implementing other bias correction and calibration methods (Bazile et al. 2017; Crochemore et al. 2016; Torralba et al. 2017). We implemented a Simple Univariate Bias Correction (SUBC) method, which resulted in further decrease in the prediction skill (See Fig. S7-8 in supplementary section, not shown for precipitation due to the inappropriate Gaussian assumption). This is because SUBC directly adjusts the distribution of forecasts from CFSs against the observations to match their statistical properties (mean and standard deviation), which requires parameter estimation and further adds uncertainty in the forecast datasets (Bazile et al. 2017; Manzanas et al. 2017). Besides, the choice of bias correction method also depends on the climate variable and the intended use of its seasonal forecast. In addition, calibration of pattern errors and probability forecast adjustments generally also require longer training datasets.

5 Conclusion

This study evaluated forecasts from the EUROSIP multi-model database and highlighted the limited skill of forecasting seasonal mean temperature and precipitation over Europe for winter and summer. Based on these results, for a comprehensive assessment of the prediction skill of seasonal climate variables, it is recommended to analyze the prediction skill of both individual and combined multi-model systems to identify the one with the highest skill over any given area and/or season. The overall lack of forecast skill is because most of the mechanisms driving changes in seasonal temperature and precipitation over Europe are a topic of active research still. In addition, even if the mechanism are well known such as the NAO, they are challenging to represent by coupled global climate models. Thus, significant improvements in the operational climate forecast systems are required through an improvement of the initial conditions and the realism of the climate models. Improved predictability can benefit decision making based on these forecasts and better prepare the European society against the anticipated water-related climate hazards.

Acknowledgement

The authors thank Prof. Francisco J. Doblas Reyes, Alicia Sanchez Lorente, Veronica Torralba and Louis-Philippe Caron for discussions and suggestions during the forming of this paper. Thanks to Nicolau Manubens, whose incredible technical support allowed steady implementation of the experiments in this study. Thanks are also due to the anonymous reviewers and their critical reviews. Their comments and suggestions improved the content of this paper. The research leading to these results has received funding from the EU H2020 Framework Programme under grant agreement #641811 (IMPREX).

References

Adler, R. F., G. J. Huffman, A. Chang, R. Ferraro, P.-P. Xie, J. Janowiak, B. Rudolf, U. Schneider, S. Curtis, D. Bolvin, A. Gruber, J. Susskind, P. Arkin, and E. Nelkin (2003). The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979-Present). *Journal of Hydrometeorology 4*(6), 1147–1167. doi:10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2.

Alexander, M. A., I. Bladé, M. Newman, J. R. Lanzante, N.-C. Lau, and J. D. Scott (2002). The atmospheric bridge: The influence of ENSO teleconnections on air-sea interaction over the global oceans. *Journal of Climate 15*(16), 2205–2231. doi:10.1175/1520-0442(2002)015<2205:TABTIO>2.0.CO;2.

Arribas, A., M. Glover, A. Maidens, K. Peterson, M. Gordon, C. MacLachlan, R. Graham, D. Fereday, J. Camp, A. A. Scaife, P. Xavier, P. McLean, A. Colman, and S. Cusack (2011). The GloSea4 ensemble prediction system for seasonal forecasting. *Monthly Weather Review 139*(6), 1891–1910. doi10.1175/2010MWR3615.1.

Athanasiadis, P. J., A. Bellucci, A. A. Scaife, L. Hermanson, S. Materia, A. Sanna, A. Borrelli, C. MacLachlan, and S. Gualdi (2017). A multisystem view of wintertime NAO seasonal predictions. *Journal of Climate 30*(4), 1461–1475.

Balmaseda, M. and D. Anderson (2009). Impact of initialization strategies and observations on seasonal forecast skill. *Geophysical Research Letters 36*(1). doi:10.1029/2008GL035561.

Bazile, R., M.-A. Boucher, L. Perreault, and R. Leconte (2017). Verification of ECMWF System4 for seasonal hydrological forecasting in a northern climate. *Hydrology and Earth System Sciences Discussions 2017*, 1–22. doi:10.5194/hess-2017-387.

Bell, C. J., L. J. Gray, A. J. Charlton-Perez, M. M. Joshi, and A. A. Scaife (2009). Stratospheric communication of El Niño teleconnections to European winter. *Journal of Climate 22*(15), 4083–4096. doi:10.1175/2009JCLI2717.1.

Bellprat, O., F. Massonnet, S. Siegert, C. Prodhomme, D. Macias-Gómez, V. Guemas, and F. Doblas-Reyes (2017). Uncertainty propagation in observational references to climate model scales. *Remote Sensing of Environment*. https://doi.org/10.1016/j.rse.2017.06.034.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), 289–300.

Boucher, M.-A., L. Perreault, and F. Anctil (2009). Tools for the assessment of hydrological ensemble forecasts obtained by neural networks. *Journal of Hydroinformatics 11*(3-4), 297–307. doi:10.2166/hydro.2009.037.

Branković, C. and T. N. Palmer (2000). Seasonal skill and predictability of ECMWF PROVOST ensembles. *Quarterly Journal of the Royal Meteorological Society 126*(567), 2035–2067. doi:10.1002/qj.49712656704.

Bröcker, J. and L. A. Smith (2008). From ensemble forecasts to predictive distribution functions. *Tellus A 60*(4), 663–678. doi:10.1111/j.1600-0870.2008.00333.x.

Candille, G. and O. Talagrand (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society 131*(609), 2131–2150. doi:10.1256/qj.04.71.

Chapelle, O., V. Vapnik, and Y. Bengio (2002, Jul). Model selection for small sample regression. *Machine Learning 48*(1), 9–23.

Colman, R. A. (2005). BMRC Atmospheric Model (BAM) version 3.0: comparison with mean climatology. Research Report No. 108, Bur. Met. Australia.

Crochemore, L., M.-H. Ramos, and F. Pappenberger (2016). Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrology and Earth System Sciences 20*(9), 3601–3618. doi:10.5194/hess-20-3601-2016.

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kallberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system.

Quarterly Journal of the Royal Meteorological Society 137(656), 553–597. doi:10.1002/qj.828.

DelSole, T. (2007). A bayesian framework for multimodel regression. *Journal of Climate 20*(12), 2810–2826. doi:10.1175/JCLI4179.1.

DelSole, T., X. Yang, and M. K. Tippett (2013). Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quarterly Journal of the Royal Meteorological Society 139*(670), 176–183. doi:10.1002/qj.1961.

Doblas-Reyes, F. J., M. Déqué, and J.-P. Piedelievre (2000). Multi-model spread and probabilistic seasonal forecasts in PROVOST. *Quarterly Journal of the Royal Meteorological Society 126*(567), 2069–2087. doi:10.1002/qj.49712656705.

Doblas-Reyes, F. J., J. García-Serrano, F. Lienert, A. P. Biescas, and L. R. L. Rodrigues (2013). Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdisciplinary Reviews: Climate Change 4*(4), 245–268. doi:10.1002/wcc.217.

Doblas-Reyes, F. J., V. Pavan, and D. B. Stephenson (2003). The skill of multi-model seasonal forecasts of the wintertime North Atlantic Oscillation. *Climate Dynamics 21*(5), 501–514. doi:10.1007/s00382-003-0350-4.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association 78*(382), 316–331.

EUROSIP (2016). EUROSIP operational history. Accessed: 2017-04-08.

Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society 140*(683), 1917–1923. doi:10.1002/qj.2270.

Ferro, C. A. T., D. S. Richardson, and A. P. Weigel (2008). On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications 15*(1), 19–24. doi:10.1002/met.45.

Fortin, V., A.-C. Favre, and M. Saïd (2006). Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. *Quarterly Journal of the Royal Meteorological Society 132*(617), 1349–1369. doi:10.1256/qj.05.167.

Frías, M. D., S. Herrera, A. S. Cofiño, and J. M. Gutiérrez (2010). Assessing the skill of precipitation and temperature seasonal forecasts in Spain: windows of opportunity related to ENSO events. *Journal of Climate 23*(2), 209–220. doi:10.1175/2009JCLI2824.1.

Fricker, T. E., C. A. T. Ferro, and D. B. Stephenson (2013). Three recommendations for evaluating climate predictions. *Meteorological Applications 20*(2), 246–255. doi:10.1002/met.1409.

Gneiting, T. (2011). Quantiles as optimal point forecasts. *International Journal of Forecasting 27*(2), 197 – 207. https://doi.org/10.1016/j.ijforecast.2009.12.015".

Gneiting, T., A. E. Raftery, A. H. W. III, and T. Goldman (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review 133*(5), 1098–1118.

doi:10.1175/MWR2904.1.

Graham, R. J., A. D. L. Evans, K. R. Mylne, M. S. J. Harrison, and K. B. Robertson (2000). An assessment of seasonal predictability using atmospheric general circulation models. *Quarterly Journal of the Royal Meteorological Society 126*(567), 2211–2240. doi:10.1002/qj.49712656712.

Guemas, V., E. Blanchard-Wrigglesworth, M. Chevallier, J. J. Day, M. Déqué, F. J. Doblas-Reyes, N. S. Fućkar, A. Germe, E. Hawkins, S. Keeley, T. Koenigk, D. S. y Mélia, and S. Tietsche (2016). A review on Arctic sea-ice predictability and prediction on seasonal to decadal time-scales. *Quarterly Journal of the Royal Meteorological Society 142*(695), 546–561. doi:10.1002/qj.2401.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting i. basic concept. *Tellus A 57*(3), 219–233. doi:10.1111/j.1600-0870.2005.00103.x.

Hawkins, E. and R. Sutton (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society 90*(8), 1095–1107. doi:10.1175/2009BAMS2607.1.

Hawkins, E. and R. Sutton (2011, Jul). The potential to narrow uncertainty in projections of regional precipitation change. *Climate Dynamics 37*(1), 407–418. doi:10.1007/s00382-010-0810-6.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting 15*(5), 559–570. doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Kharin, V. V. and F. W. Zwiers (2002). Climate predictions with multimodel ensembles. *Journal of Climate 15*(7), 793–799.

Kim, H.-M., P. J. Webster, and J. A. Curry (2012, Dec). Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter. *Climate Dynamics 39*(12), 2957–2973. doi:10.1007/s00382-012-1364-6.

Krikunov, A. V. and S. V. Kovalchuk (2015). Dynamic selection of ensemble members in multi-model hydrometeorological ensemble forecasting. *Procedia Computer Science 66*(Supplement C), 220 – 227. 4th International Young Scientist Conference on Computational Science.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran (2000). Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate 13*(23), 4196–4216. doi:10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2.

Kumar, A. (2009). Finite samples and uncertainty estimates for skill measures for seasonal prediction. *Monthly Weather Review 137*(8), 2622–2631.

Kumar, A., M. Chen, and W. Wang (2013). Understanding prediction skill of seasonal mean precipitation over the tropics. *Journal of Climate 26*(15), 5674–5681. doi:10.1175/JCLI-D-12-00731.1.

Lavell, A., M. Oppenheimer, C. Diop, J. Hess, R. Lempert, J. Li, R. Muir-Wood, and S. Myeong (2012). Climate change: new dimensions in disaster risk, exposure, vulnerability, and resilience. In C. Field, V. Barros, T. Stocker, D. Qin, D. Dokken, K. Ebi, M. Mastrandrea, K. Mach, G.-K.

Plattner, S. Allen, M. Tignor, and P. Midgley (Eds.), *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*, pp. 25–64. Cambridge, UK and New York, NY, USA: Cambridge University Press. doi:10.1596/978-0-8213-8845-7.

Lavers, D., L. Luo, and E. F. Wood (2009). A multiple model assessment of seasonal climate forecast skill for applications. *Geophysical Research Letters 36*(23). doi:10.1029/2009GL041365.

Lloyd-Hughes, B. and M. A. Saunders (2002). Seasonal prediction of European spring precipitation from El Niñosouthern oscillation and local sea-surface temperatures. *International Journal of Climatology 22*(1), 1–14. doi:10.1002/joc.723.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences 20*(2), 130–141. doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

MacLachlan, C., A. Arribas, K. A. Peterson, A. Maidens, D. Fereday, A. A. Scaife, M. Gordon, M. Vellinga, A. Williams, R. E. Comer, J. Camp, P. Xavier, and G. Madec (2015). Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society 141*(689), 1072–1084. doi:10.1002/qj.2396.

Manzanas, R., A. Lucero, A. Weisheimer, and J. M. Gutiérrez (2017, Apr). Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts? *Climate Dynamics*. doi:10.1007/s00382-017-3668-z.

Matheson, J. E. and R. L. Winkler (1976). Scoring rules for continuous probability distributions. *Management Science 22*(10), 1087–1096. doi:10.1287/mnsc.22.10.1087.

Molinaro, A. M., R. Simon, and R. M. Pfeiffer (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics 21 15*, 3301–7.

Molteni, F., T. Stockdale, M. Balmaseda, G. Balsamo, R. Buizza, L. Ferranti, L. Magnusson, K. Mogensen, T. Palmer, and F. Vitart (2011). The new ECMWF seasonal forecast system (System 4). *ECMWF Technical Memorandum* (656), 49.

Murphy, A. H. (1969). On the "Ranked Probability Score". *Journal of Applied Meteorology 8*(6), 988–989. doi:10.1175/1520-0450(1969)008<0988:OTPS>2.0.CO;2.

National Academies of Sciences, E. and Medicine (2016). *Attribution of Extreme Weather Events in the Context of Climate Change*. Washington, DC: The National Academies Press. doi:10.17226/21852.

Palmer, T., F. Doblas-Reyes, R. Hagedorn, and A. Weisheimer (2005). Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Philos. Trans. Roy. Soc. B*. doi:10.1098/rstb.2005.1750.

Palmer, T. N., F. J. Doblas-Reyes, R. Hagedorn, A. Alessandri, S. Gualdi, U. Andersen, H. Feddersen, P. Cantelaube, J.-M. Terres, M. Davey, R. Graham, P. Délécluse, A. Lazar, M. Déqué, J.-F. Guérémy, E. Díez, B. Orfila, M. Hoshen, A. P. Morse, N. Keenlyside, M. Latif, E. Maisonnave, P. Rogel, V. Marletto, and M. C. Thomson (2004). Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bulletin of the American Meteorological Society 85*(6), 853–872.

Peng, P., A. Kumar, H. van den Dool, and A. G. Barnston (2002). An analysis of multimodel ensemble predictions for seasonal climate anomalies. *Journal of Geophysical Research: Atmospheres 107*(D23), ACL 18–1–ACL 18–12. doi:10.1029/2002JD002712.

Prodhomme, C., F. Doblas-Reyes, O. Bellprat, and E. Dutra (2016, Aug). Impact of land-surface initialization on sub-seasonal to seasonal forecasts over Europe. *Climate Dynamics 47*(3), 919–935. doi:10.1007/s00382-015-2879-4.

Robertson, A. W., U. Lall, S. E. Zebiak, and L. Goddard (2004). Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. *Monthly Weather Review 132*(12), 2732–2744. doi:10.1175/MWR2818.1.

Rodrigues, L. R. L., F. J. Doblas-Reyes, and C. A. dos Santos Coelho (2014, Feb). Multi-model calibration and combination of tropical seasonal sea surface temperature forecasts. *Climate Dynamics 42*(3), 597–616. doi:10.1007/s00382-013-1779-8.

Rodrigues, L. R. L., J. García-Serrano, and F. Doblas-Reyes (2014). Seasonal forecast quality of the West African monsoon rainfall regimes by multiple forecast systems. *Journal of Geophysical Research: Atmospheres 119*(13), 7908–7930. doi:10.1002/2013JD021316.

Saha, S., S. Moorthi, X. Wu, J. Wang, S. Nadiga, P. Tripp, D. Behringer, Y.-T. Hou, H. ya Chuang, M. Iredell, M. Ek, J. Meng, R. Yang, M. P. Mendez, H. van den Dool, Q. Zhang, W. Wang, M. Chen, and E. Becker (2014). The ncep climate forecast system version 2. *Journal of Climate 27*(6), 2185–2208. doi:10.1175/JCLI-D-12-00823.1.

Scaife, A. A., A. Arribas, E. Blockley, A. Brookshaw, R. T. Clark, N. Dunstone, R. Eade, D. Fereday, C. K. Folland, M. Gordon, L. Hermanson, J. R. Knight, D. J. Lea, C. MacLachlan, A. Maidens, M. Martin, A. K. Peterson, D. Smith, M. Vellinga, E. Wallace, J. Waters, and A. Williams (2014). Skillful long-range prediction of European and North American winters. *Geophysical Research Letters 41*(7), 2514–2519. doi:10.1002/2014GL059637.

Senan, R., Y. J. Orsolini, A. Weisheimer, F. Vitart, G. Balsamo, T. N. Stockdale, E. Dutra, F. J. Doblas-Reyes, and D. Basang (2016, Nov). Impact of springtime Himalayan–Tibetan Plateau snowpack on the onset of the Indian summer monsoon in coupled seasonal forecasts. *Climate Dynamics 47*(9), 2709–2725. doi:10.1007/s00382-016-2993-y.

Shi, W., N. Schaller, D. MacLeod, T. N. Palmer, and A. Weisheimer (2015). Impact of hindcast length on estimates of seasonal climate predictability. *Geophysical Research Letters. 42*(5), 15541559.

Siegert, S., O. Bellprat, M. Ménégoz, D. B. Stephenson, and F. J. Doblas-Reyes (2017). Detecting improvements in forecast correlation skill: Statistical testing and power analysis. *Monthly Weather Review 145*(2), 437–450.

Soares, M. B. and S. Dessai (2015). Exploring the use of seasonal climate forecasts in Europe through expert elicitation. *Climate Risk Management 10*, 8 – 16.

Sordo, C., M. D. Frías, A. S. C. n. S. Herrera, and J. M. Gutiérrez (2008). Interval-based statistical validation of operational seasonal forecasts in Spain conditioned to El NiñoSouthern Oscillation events. *Journal of Geophysical Research: Atmospheres 113*(D17). doi:10.1029/2007JD009536.

Starks, P. J., J. D. Ross, and G. C. Heathman (2003). Modelling the spatial and temporal distribution of soil moisture at watershed scales using remote sensing and GIS. In V. H. Singhroy, D. T. Hansen, R. R. Pierce, and A. I. Johnson (Eds.), *Spatial methods for solution of environmental and hydrologic problems - science, policy and standardization*, pp. 58–75. West Conshohocken, PA: ASTM International.

Stockdale, T. (2013). The EUROSIP system - a multi-model approach. In *Seminar on Seasonal prediction: science and applications, 3-7 September 2012*, Shinfield Park, Reading, pp. 257–268. ECMWF: ECMWF.

Tebaldi, C. and R. Knutti (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 365*(1857), 2053–2075. doi:10.1098/rsta.2007.2076.

Torralba, V., F. J. Doblas-Reyes, D. MacLeod, I. Christel, and M. Davis (2017). Seasonal climate prediction: A new source of information for the management of wind energy resources. *Journal of Applied Meteorology and Climatology 56*(5), 1231–1247. doi:10.1175/JAMC-D-16-0204.1.

Trigo, R. M., T. J. Osborn, and J. M. Corte-Real (2002). The North Atlantic Oscillation influence on Europe: climate impacts and associated physical mechanisms. *Climate Research 20*, 9–17. doi:10.3354/cr020009.

van den Hurk, B., F. Doblas-Reyes, G. Balsamo, R. D. Koster, S. I. Seneviratne, and H. C. Jr (2012, Jan). Soil moisture effects on seasonal temperature and precipitation forecast scores in europe. *Climate Dynamics 38*(1), 349–362.

Velázquez, J. A., T. Petit, A. Lavoie, M.-A. Boucher, R. Turcotte, V. Fortin, and F. Anctil (2009). An evaluation of the Canadian global meteorological ensemble prediction system for short-term hydrological forecasting. *Hydrology and Earth System Sciences 13*(11), 2221–2231. doi:10.5194/hess-13-2221-2009.

Vitart, F., M. R. Huddleston, M. Déqué, D. Peake, T. N. Palmer, T. N. Stockdale, M. K. Davey, S. Ineson, and A. Weisheimer (2007). Dynamically-based seasonal forecasts of Atlantic tropical storm activity issued in June by EUROSIP. *Geophysical Research Letters 34*(16). doi10.1029/2007GL030740.

Voldoire, A., E. Sachez-Gomez, D. S. y Melia, B. Decharme, C. Cassou, S. Sénési, S. Valcke, I. Beau, A. Alias, M. Chevallier, M. Déqué, J. Deshayes, H. Douville, E. Fernandez, G. Madec, E. Maisonnave, M.-P. Moine, S. Planton, D. Saint-Martin, S. Szopa, S. Tyteca, R. Alkama, S. Belamari, A. Braun, L. Coquart, and F. Chauvin (2013, May). The CNRM-CM5.1 global climate model: description and basic evaluation. *Climate Dynamics 40*(9), 2091–2121. doi:10.1007/s00382-011-1259-y.

Wanders, N. and Y. Wada (2015). Decadal predictability of river discharge with climate oscillations over the 20th and early 21st century. *Geophysical Research Letters 42*(24), 10,689–10,695. doi:10.1002/2015GL066929.

Wanders, N. and E. F. Wood (2016). Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations. *Environmental Research Letters 11*(9), 094007. doi:10.1088/1748-9326/11/9/094007.

Wehrli, K., J. Bhend, and M. A. Liniger (2017). Systematic quality assessment of an operational seasonal forecasting system. Technical Report MeteoSwiss, 263, 52 pp.

Weigel, A. P., R. Knutti, M. A. Liniger, and C. Appenzeller (2010). Risks of model weighting in multimodel climate projections. *Journal of Climate 23*(15), 4175–4191.

Weigel, A. P., M. A. Liniger, and C. Appenzeller (2008). Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society 134*(630), 241–260. doi:10.1002/qj.210.

Wilks, D. S. (2006). Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorological Applications 13*(3), 243–256. doi:10.1017/S1350482706002192.

Yuan, X., F. Ma, L. Wang, Z. Zheng, Z. Ma1, A. Ye, and S. Peng (2016). An experimental seasonal hydrological forecasting system over the yellow river basin – part 1: Understanding the role of initial hydrological conditions. *Hydrology and Earth System Sciences 20*(6), 2437–2451. doi:10.5194/hess-20-2437-2016.

Yun, W. T., L. Stefanova, and T. N. Krishnamurti (2003). Improvement of the multimodel superensemble technique for seasonal forecasts. *Journal of Climate 16*(22), 3834–3840. doi:10.1175/1520-0442(2003)016<3834:IOTMST>2.0.CO;2.

List of Figures

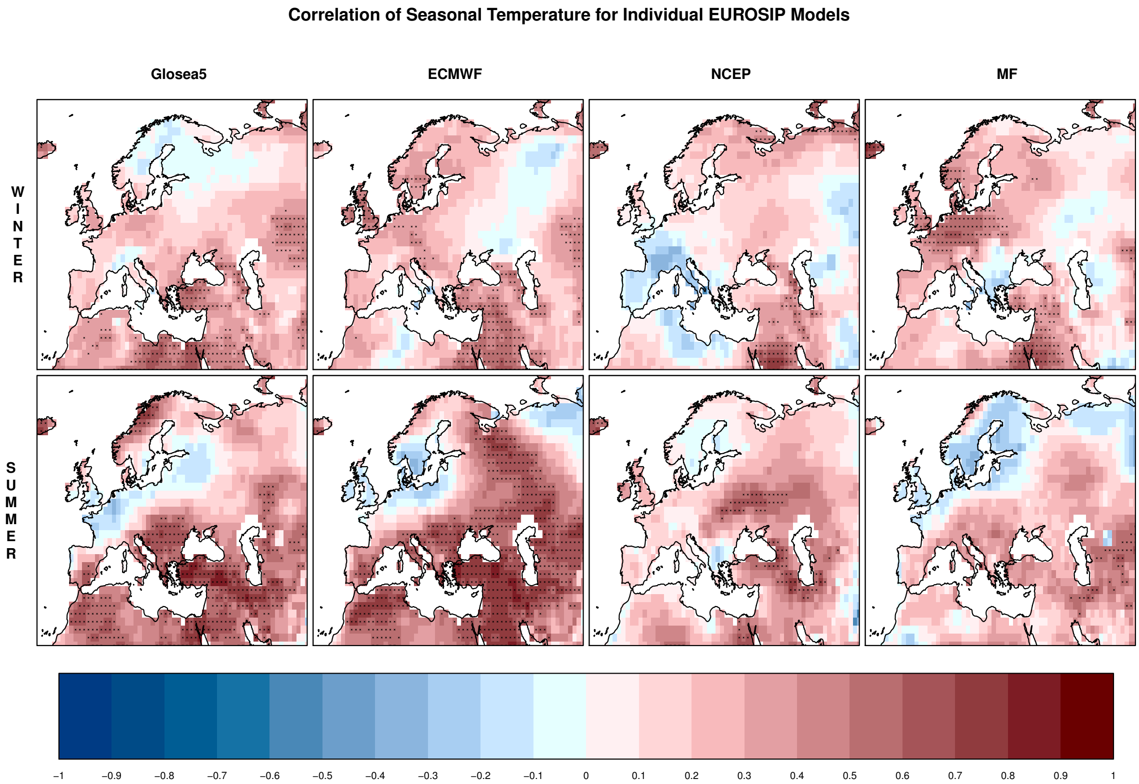

Fig. 1: Anomaly Correlation Coefficient (ACC) between the predicted ensemble mean of each individual climate model of EUROSIP and the observed seasonal winter (DJF; top row) and summer (JJA; bottom row) temperature obtained from ERAINT over the European region (20° W-70° E and 25° N-75° N) for the period 1992-2012. The individual climate models are Glosea5, ECMWF, NCEP and MF (from left to right; see details in section 2). Forecasts are initialized in November for DJF and in May for JJA. Areas covered in red are indicative of positive correlation, while areas covered in blue indicate negative correlation. Dots in each grid point indicate significant positive correlation at 5% significance level using one-sided Student t-test and controlling for false discovery rate.

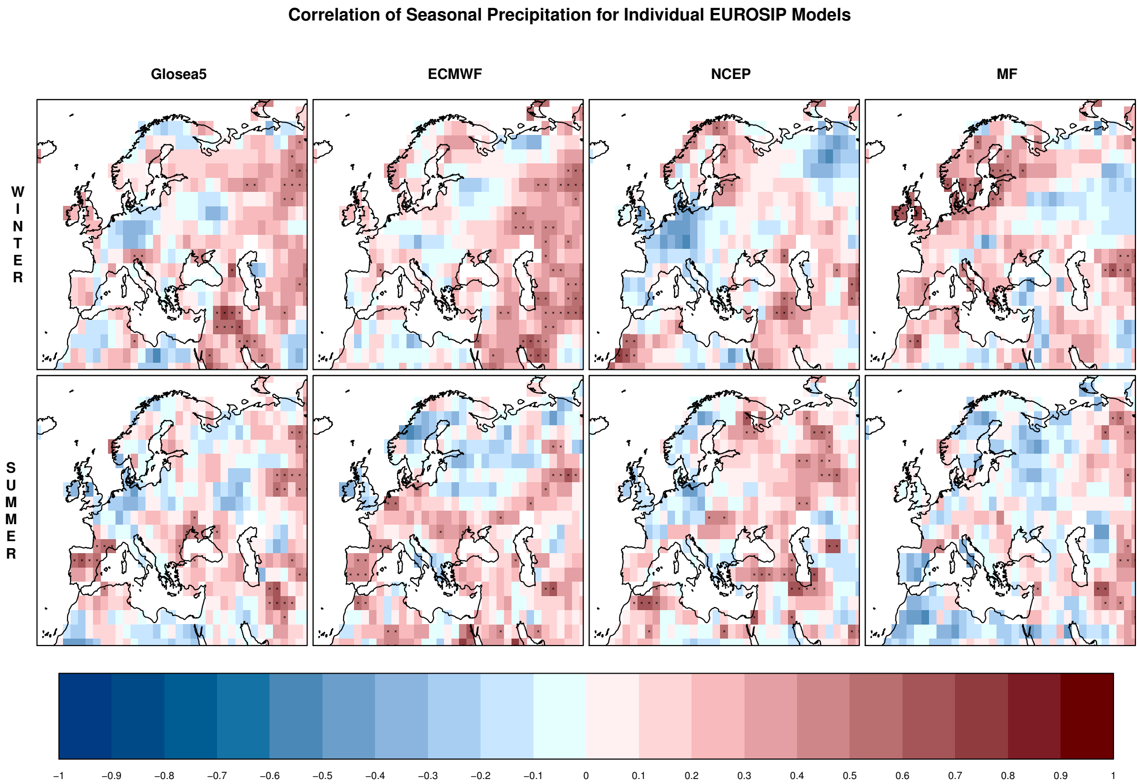

Fig. 2: Same as Fig.1 but for precipitation and reference data obtained from GPCP

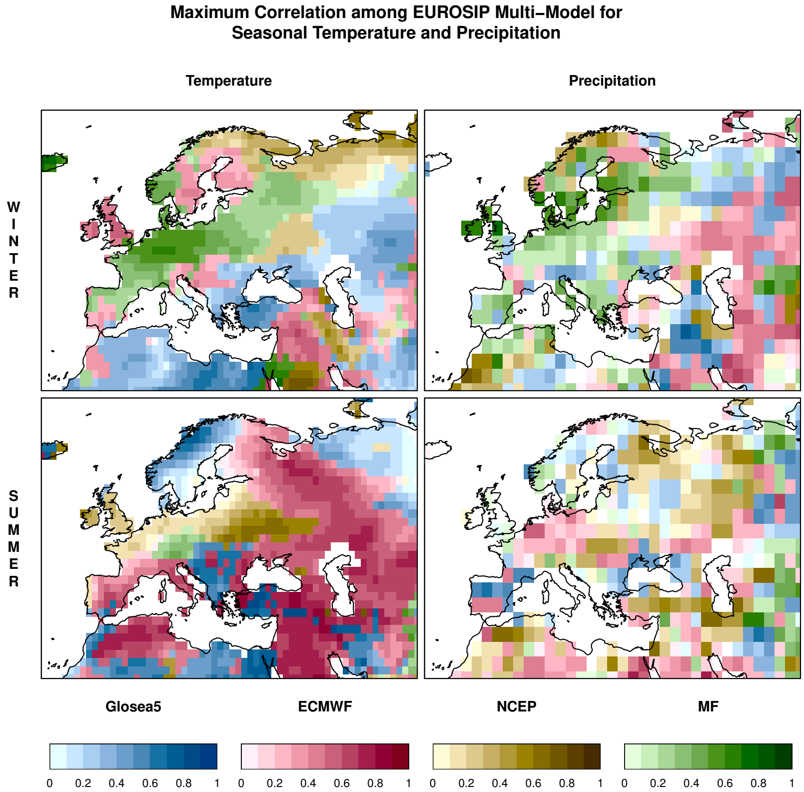

Fig. 3: Maximum positive Anomaly Correlation Coefficient (ACC) among the four individual models from EUROSIP. ACC for each model is calculated between their respective predicted ensemble mean anomalies and the anomalies of the observed temperature obtained from ERAINT (left) and of precipitation obtained from GPCP (right) for winter (DJF; top row) and summer (JJA; bottom row) seasons over the period 1992-2012. Blue, red, yellow and green colors indicate that the maximum correlation is obtained for GloSea5, ECMWF, NCEP and MF respectively. Negative or 0 correlations appear in white.

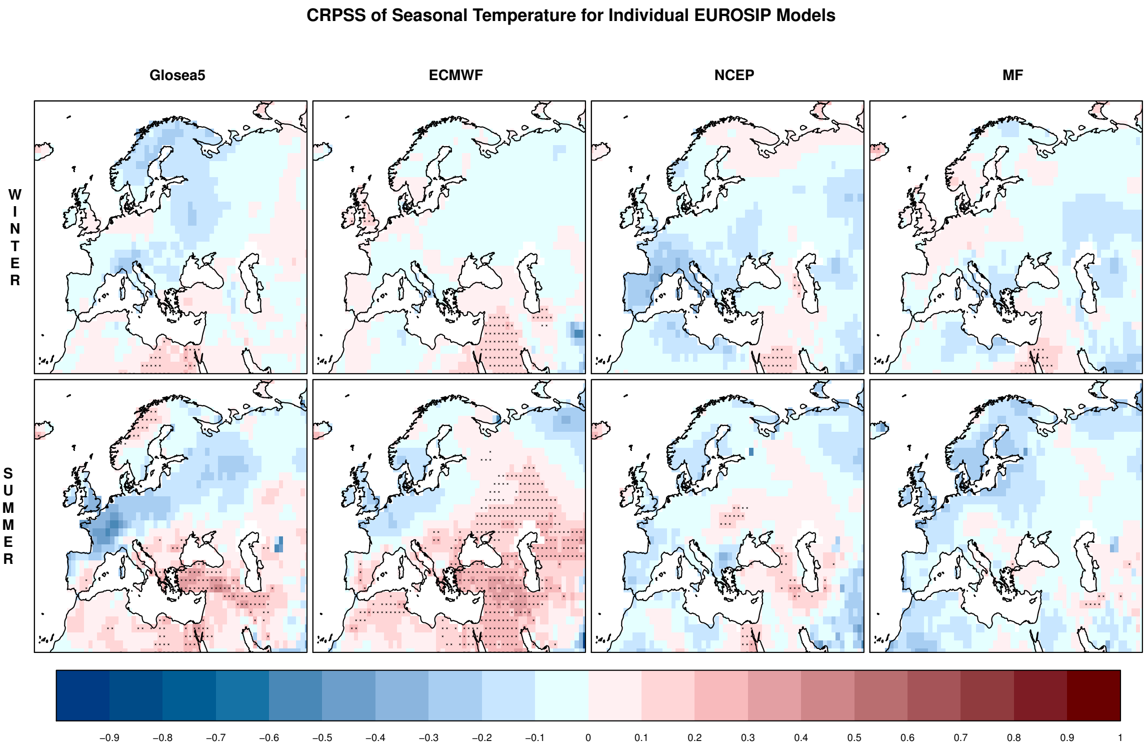

Fig. 4: CRPSS of the probability ensemble forecasts of each individual climate model of EUROSIP with climatology used as reference forecast obtained from ERAINT for winter (DJF; top row) and summer (JJA; bottom row) seasonal temperature over the European region (20° W-70° E and 25° N-75° N) for the period 1992-2012. The individual climate models are Glosea5, ECMWF, NCEP MF (from left to right). Forecasts are initialized in November for DJF and in May for JJA. Areas covered in red are indicative of positive CRPSS, suggesting skill better than climatology. Areas covered in blue indicate worse skill than climatology. Dots in each grid point indicate significant positive CRPSS using the standard deviation of the skill score, approximated by propagation of uncertainty at 95% confidence interval.

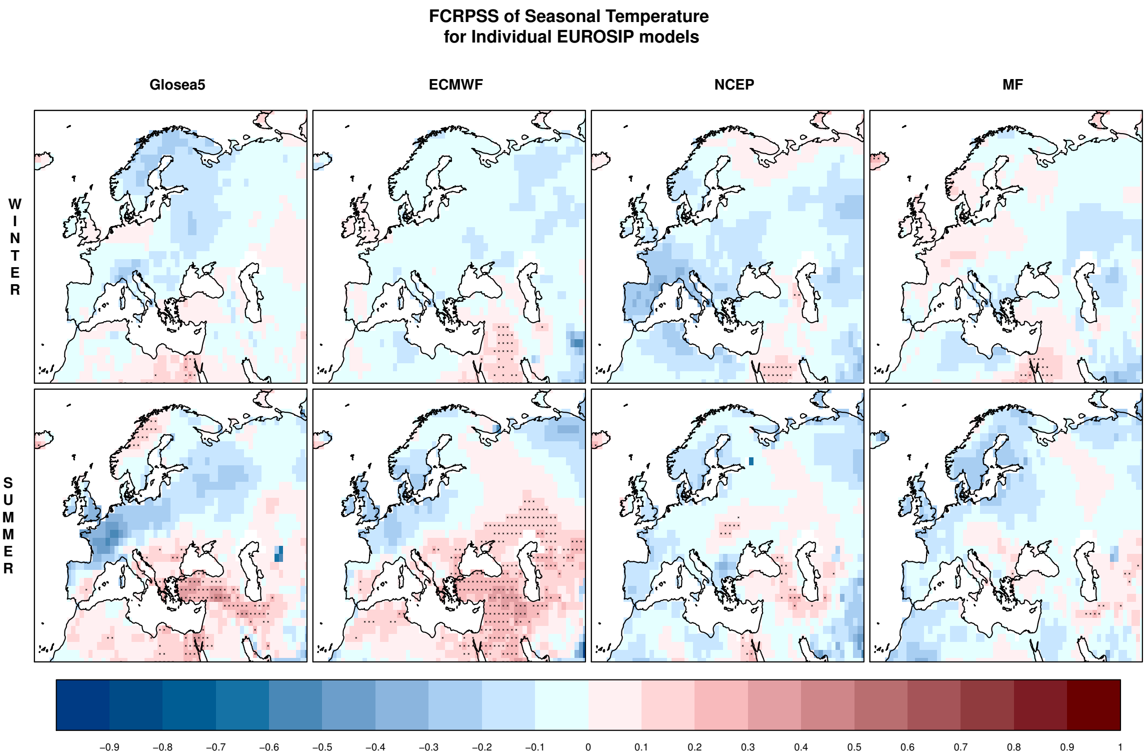

Fig. 5: FCRPSS of the probability ensemble forecasts of each individual climate model of EUROSIP with climatology obtained from ERAINT used as reference forecast for winter (DJF; top row) and summer (JJA; bottom row) seasonal temperature over the European region (20° W-70° E and 25° N-75° N) for the period 1992-2012. The individual climate models are Glosea5, ECMWF, NCEP and MF (from left to right). Forecasts are initialized in November for DJF and in May for JJA. Areas covered in red are indicative of positive CRPSS, suggesting skill better than climatology. Areas covered in blue indicate worse skill than climatology. Dots in each grid point indicate significant positive CRPSS using the standard deviation of the skill score, approximated by propagation of uncertainty at 95% confidence interval.

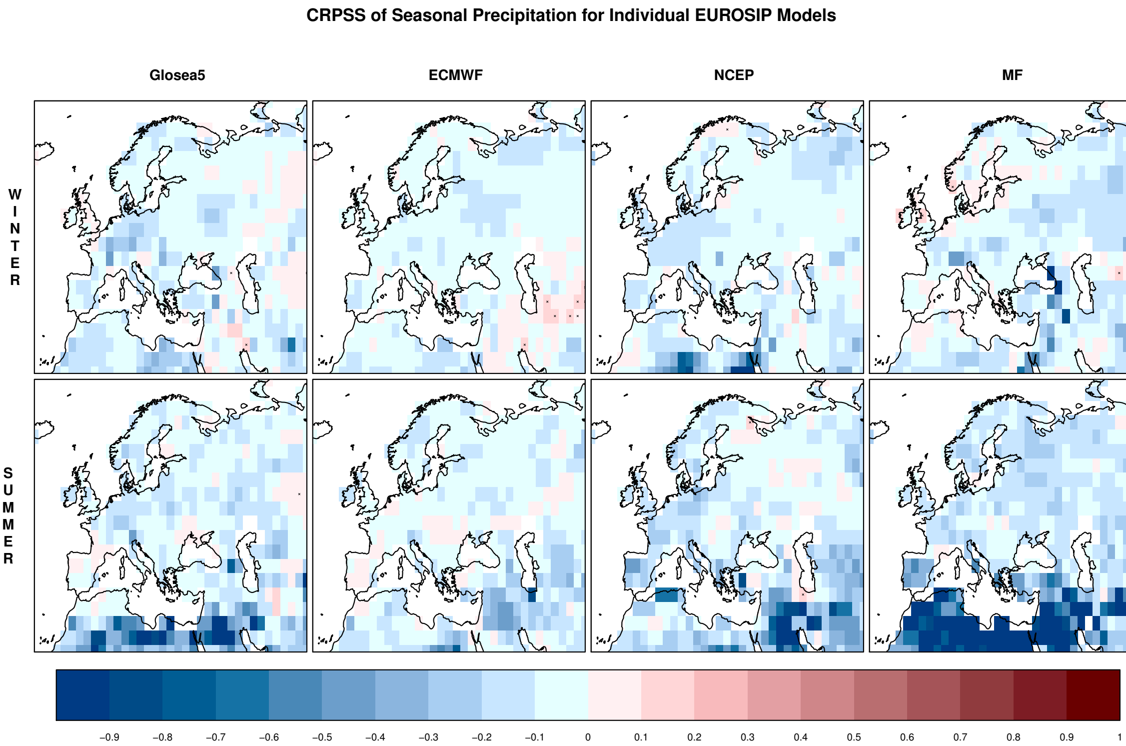

Fig. 6: Same as Fig.4 but for precipitation and reference data obtained from GPCP.

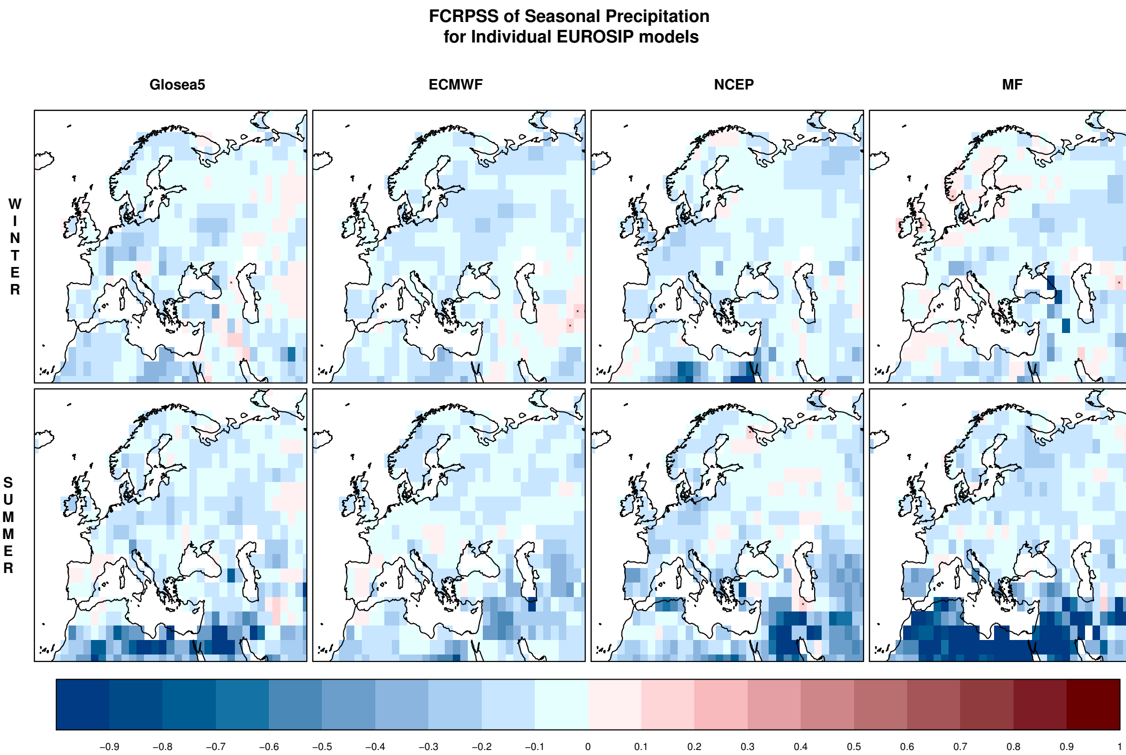

Fig. 7: Same as Fig. 5 but for precipitation and reference forecast obtained from GPCP.

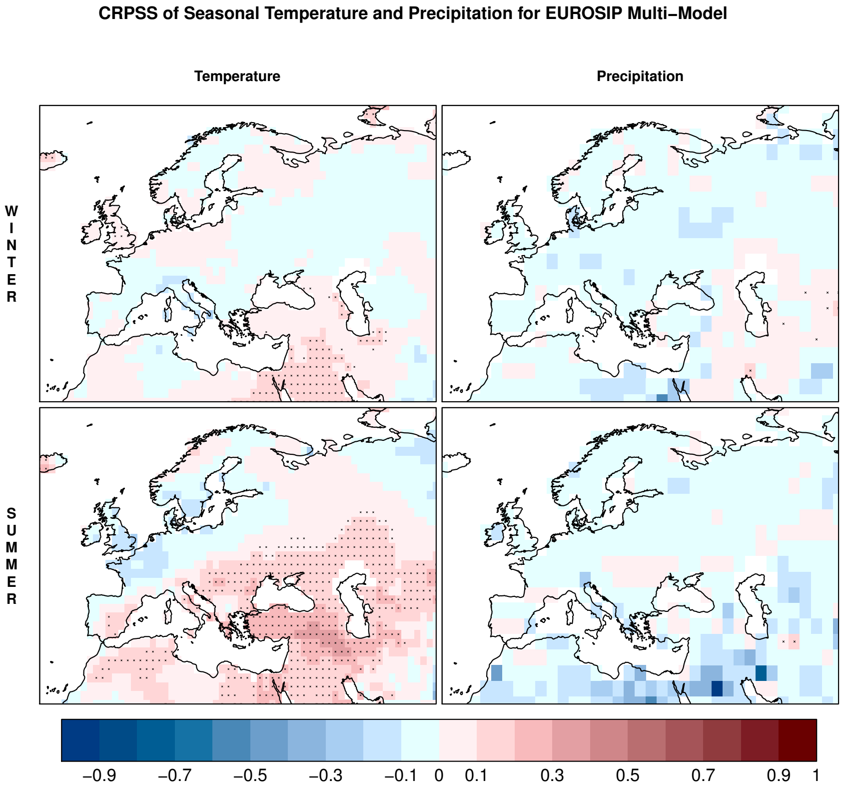

Fig. 8: CRPSS of the probability ensemble forecasts from all four individual climate models of EUROSIP (Glosea5, ECMWF, NCEP and MF) treated as one single model with climatology used as reference forecast obtained from ERAINT for temperature (left) and GPCP for precipitation (right) for winter (DJF; top row) and summer (JJA; bottom row) seasons over the European region (20° W-70° E and 25° N-75° N) for the period 1992-2012. Forecasts are initialized in November for DJF and in May for JJA. Areas covered in red are indicative of positive correlation suggesting skill better than climatology. Areas covered in blue indicate worse skill than climatology. Dots in each grid point indicate significant positive CRPSS using the standard deviation of the skill score, approximated by propagation of uncertainty at 95% confidence interval.

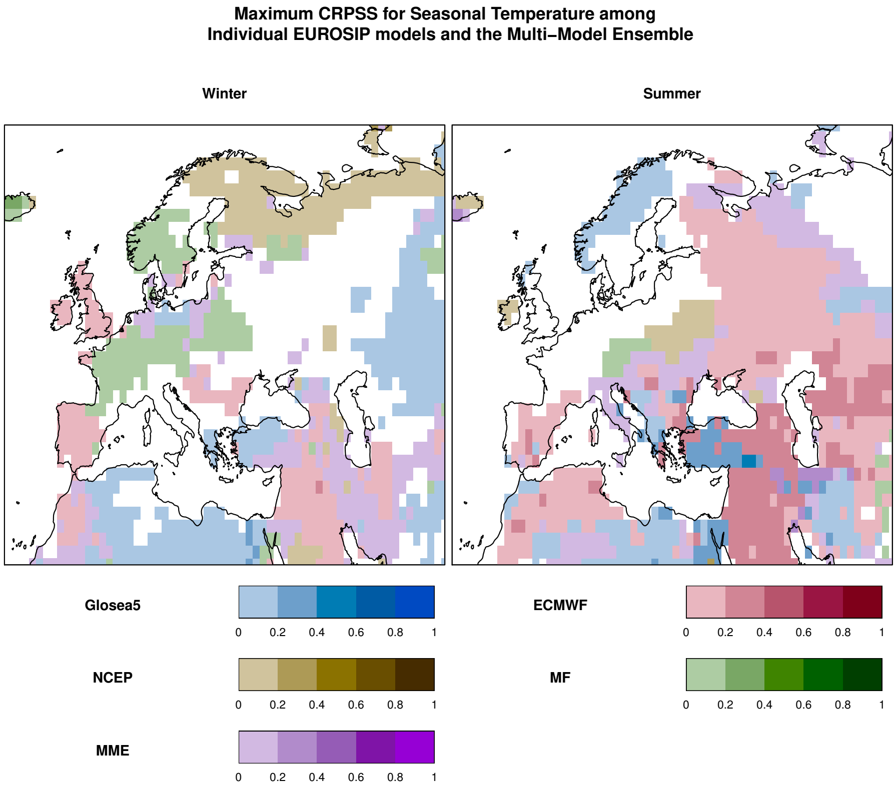

Fig. 9: Maximum positive CRPSS among the four individual models from EUROSIP and the multi-model with climatology used as reference forecast obtained from ERAINT for winter (DJF; left) and summer (JJA; summer) over the European region (20° W-70° E and 25° N-75° N) for the period 1992-2012. Forecasts are initialized in November for DJF and in May for JJA. Blue, red, yellow, green and purple colors indicate that the maximum CRPSS is obtained for GloSea5, ECMWF, NCEP, MF and the multi-model, respectively. Negative or 0 correlations appear in white.

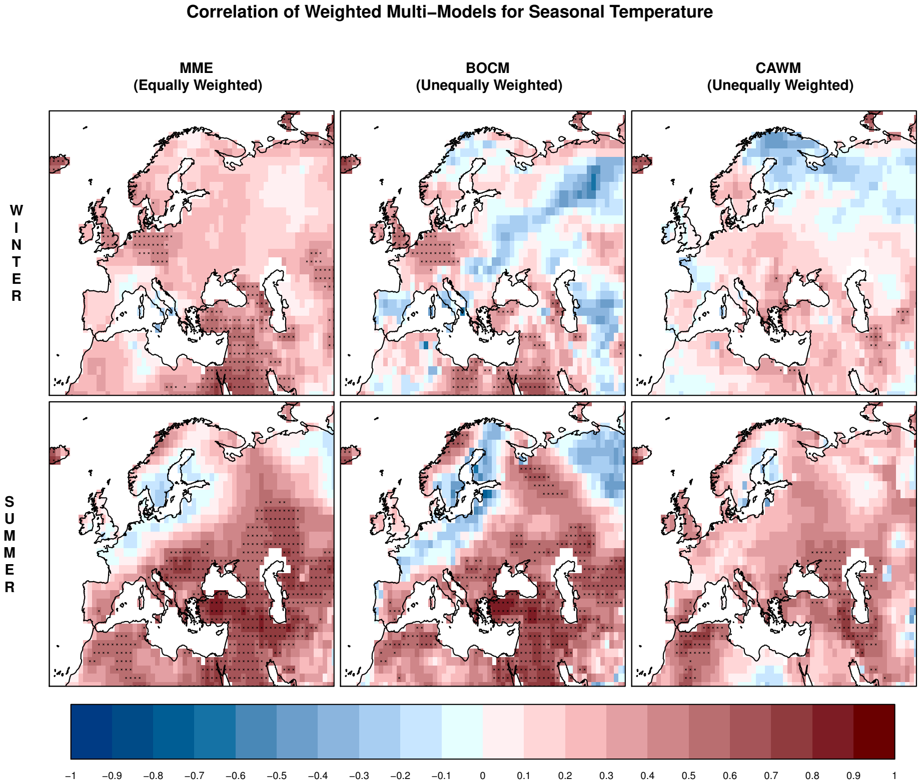

Fig. 10: Anomaly Correlation Coefficient (ACC) between the predictions obtained from the three Weighted Multi-Model (WMM) systems of EUROSIP - Multi-Model Mean (MMM), Best OLS Combination Model (BOCM) and Correlation As Weights Model (CAWM; from left to right) and the observed seasonal winter (DJF; top row) and summer (JJA; bottom row) temperature obtained from ERAINT, respectively, over the European region (20° W-70° E and 25° N-75° N) for the period 1992-2012. Areas covered in red are indicative of positive correlation, while areas covered in blue indicate negative correlation. Dots in each grid point indicate significant positive correlation at 5% significance level using one-sided Students t-test. Details on the construction of each WMM system are given in Section 2c.

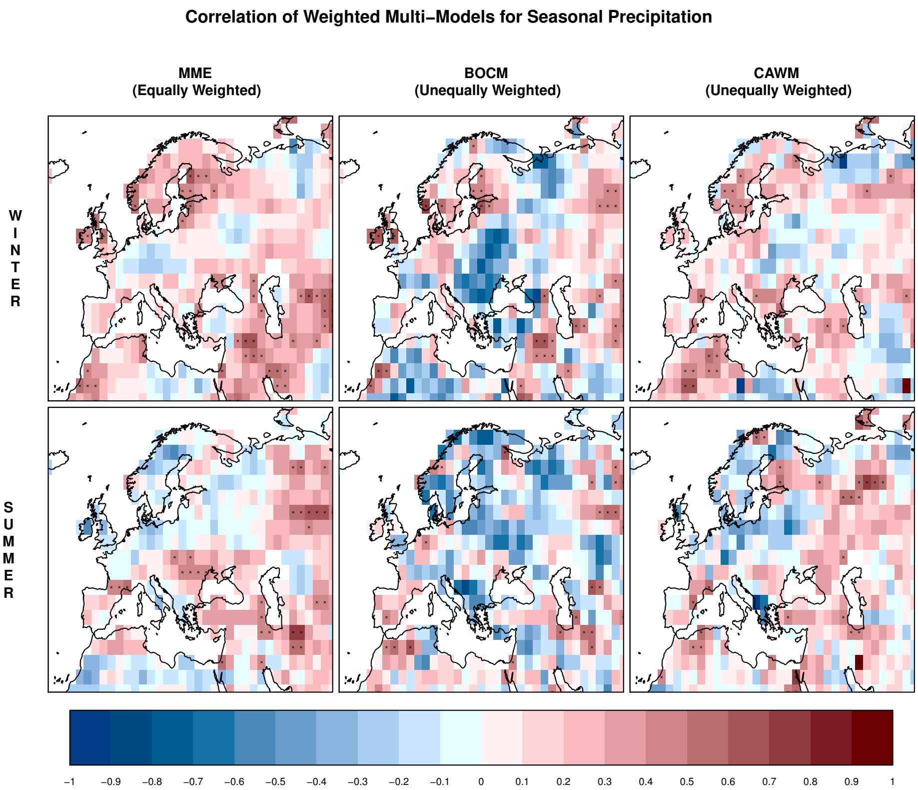

Fig. 11: Same as Fig.10 but for precipitation and reference data obtained from GPCP.

List of Tables

Table 1: Individual climate models of EUROSIP multi-model seasonal forecasting system

Climate Model	No. of Ensemble Members	Resolution (in Gaussian grid)
Glosea5	24	512x256
ECMWF	51	432x325
Meteo France	15	256x128 for temperature
		360x181 for precipitation
NCEP	28 for winter	384x190
	24 for summer	

Reference Dataset	Resolution (in Gaussian grid)
ERA-Interim for temperature	512x256
GPCP for precipitation	144x72