Temperature and precipitation seasonal forecasts over the Mediterranean region: added value compared to simple forecasting methods

Climate Dynamics
https://doi.org/10.1007/s00382-021-05895-6

Temperature and precipitation seasonal forecasts over the Mediterranean region: added value compared to simple forecasting methods

Filippo Calì Quaglia[1] · **Silvia Terzago**[2] · **Jost von Hardenberg**[3,2]

Received: 30 January 2021 / Accepted: 16 July 2021

Abstract
This study considers a set of state-of-the-art seasonal forecasting systems (ECMWF, MF, UKMO, CMCC, DWD and the corresponding multi-model ensemble) and quantifies their added value (if any) in predicting seasonal and monthly temperature and precipitation anomalies over the Mediterranean region compared to a simple forecasting method based on the ERA5 climatology (CTRL) or the persistence of the ERA5 anomaly (PERS). This analysis considers two starting dates, May 1st and November 1st and the forecasts at lead times up to 6 months for each year in the period 1993–2014. Both deterministic and probabilistic metrics are employed to derive comprehensive information on the forecast quality in terms of association, reliability/resolution, discrimination, accuracy and sharpness. We find that temperature anomalies are better reproduced than precipitation anomalies with varying spatial patterns across different forecast systems. The Multi-Model Ensemble (MME) shows the best agreement in terms of anomaly correlation with ERA5 precipitation, while PERS provides the best results in terms of anomaly correlation with ERA5 temperature. Individual forecast systems and MME outperform CTRL in terms of accuracy of tercile-based forecasts up to lead time 5 months and in terms of discrimination up to lead time 2 months. All seasonal forecast systems also outperform elementary forecasts based on persistence in terms of accuracy and sharpness.

Keywords Seasonal forecasts · Temperature · Precipitation · Skill scores · Forecast verification · Mediterranean

1 Introduction

Seasonal forecasts of atmospheric variables like near-surface air temperature and precipitation are attractive for a variety of applications in different economic and socially relevant sectors, including hydropower and wind energy production (Torralba et al. 2017; Clark et al. 2017), management of water resources (Svensson et al. 2015), fire risk, agriculture, transports (Palin et al. 2016) and shipping, health (Lowe et al. 2017), and in general, hazardous weather events which can cause serious economic damages (Morss et al. 2008). In all these cases, a reliable indication of mean climate conditions a few months ahead can be associated with a well defined economic value (Bruno Soares et al. 2018, and references therein). Several recent and ongoing research projects explore the potential of seasonal forecasts in providing added value to specific applications in different economic sectors (Graça 2019; Hewitt et al. 2013). To this end, the first step is to test whether seasonal forecasts of the main climatic variables have some skill per-se, i.e. whether they

This paper is a contribution to the MEDSCOPE special issue on the drivers of variability and sources of predictability for the European and Mediterranean regions at subseasonal to multi-annual time scales. MEDSCOPE is an ERA4CS project co-funded by JPI Climate. The special issue was coordinated by Silvio Gualdi and Lauriane Batté.

✉ Silvia Terzago
s.terzago@isac.cnr.it

[1] Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Via Torino 155, 30172 Venice-Mestre, Italy

[2] Institute of Atmospheric Sciences and Climate, National Research Council of Italy (CNR-ISAC), Corso Fiume, 4, 10133 Turin, Italy

[3] Department of Environment, Land and Infrastructure Engineering, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy

Published online: 20 September 2021

can predict the observed climate anomalies better than climatology, in a specific area of study.

Previous literature findings show that seasonal climate prediction has progressed considerably in the last decades. The tropics remain the region where seasonal forecasts are most successful (Doblas-Reyes et al. 2013); outside this region, predictability is generally lower, and forecast skill can drop considerably. For example, seasonal predictability over the Mediterranean region is influenced by the North Atlantic Oscillation (Athanasiadis et al. 2017; Dunstone et al. 2016), by other teleconnections such as El Niño (Frías et al. 2010), by processes taking place in the stratosphere and by specific initialisation of soil moisture (Prodhomme et al. 2016) and sea ice (Guemas et al. 2016). However there is little understanding of the incremental skill gained through teleconnections and which are the still missing sources of predictability and mechanisms responsible for low forecast skill (National Academies of Sciences Engineering and Medicine 2010). For example, Mishra et al. (2019) assessed temperature and precipitation ensemble seasonal forecast of the EUROSIP multi-model forecasting system (Stockdale 2012) over Europe and found, on average, limited prediction skills for precipitation. Sánchez-García et al. (2018) provided a detailed report of different probabilistic scores for temperature and precipitation forecasts over different European regions, comparing previous generation forecast systems and finding higher scores for temperature rather than for precipitation. A more recent paper by Johnson et al. (2019) provided a worldwide analysis of the skill of ECMWF System 5 with respect to System 4, mainly focusing on the spatial distribution of the continuous rank probability score and of the anomaly correlation at the global scale. They report a decrease in the skill of the SST forecast in the Northwest Atlantic, which may impact the prediction of the North Atlantic Oscillation. While the prediction of ENSO is quite good, issues in the lower stratosphere and at the tropopause are also reported, which could influence the ability to extend forecast skill to the extra-tropics. Dunstone et al. (2016) and Scaife et al. (2014) focused their work on the predictability of the North Atlantic Oscillation, which profoundly influences North American and European winter climate, finding the UKMO and the HadGEM3-GC2 models to have skill in NAO prediction up to the following season.

Other interesting results come from research projects aiming at bridging the gap between research and applications and exploring the potential of seasonal forecasts: DEMETER (Palmer et al. 2004) pointed at developing a well-validated European coupled multi-model ensemble forecast system for reliable seasonal to interannual predictions; ENSEMBLES (van der Linden and Mitchell 2009) focused on the assessment of climate model uncertainties; EUPORIAS (Hewitt et al. 2013) and MEDGOLD (Graça 2019) developed working prototypes of climate services addressing the need of specific users, with the latter being focused on sustainable agriculture and food systems. An additional recent project, ERA4CS MEDSCOPE, of which this work is a part (https://www.medscope-project.eu/), focuses on the evaluation of the seasonal climate predictability over the Mediterranean region and the exploitation of seasonal forecasts for the development of climate services for different economic sectors.

Given their probabilistic nature, seasonal forecasts describe a range of possible evolutions of climate and require appropriate ensemble verification tools to assess their quality. Many different metrics have been developed (Vannitsem et al. 2018; Jolliffe and Stephenson 2011; Wilks 2011), each of which addresses a different characteristic of the forecast, i.e. reliability, resolution, ability to discriminate events and non-events, ability to reproduce a meaningful ensemble, among others. Standard scoring metrics of probabilistic forecasts are affected by (i) improper estimates of probabilities from small-sized ensembles, (ii) an insufficient number of forecast cases, and (iii) imperfect reference values due to uncertainties in observation and reanalysis data (Doblas-Reyes et al. 2003). These issues can be partly alleviated using a suitably large area, the longest available hindcast period and different scoring rules for evaluating the features of a probabilistic forecast from different perspectives (as also suggested by, i.e.,WMO 2018; Wilks 2011; Murphy 1993).

In this paper we focus on the Mediterranean region, an area of transition representative of the mid-latitudes, where seasonal forecasts are challenging and where accurate investigation of the skills and limits of the current seasonal forecast systems is crucial to improve models. Moreover the Mediterranean is well known as a hotspot area for climate change, where enhanced warming is expected to impact food security, water availability and ecosystems (Cramer et al. 2020). In this context reliable seasonal predictions are essential to provide early warning on extreme seasons and to enable decision-makers to take actions that reduce the impacts. With this analysis we aim to: (i) provide an assessment of the skill of current state-of-the-art seasonal forecast systems at predicting temperature and precipitation anomalies over the Mediterranean region, focusing on the winter and summer seasons which are relevant for many applications in the energy sector, water management, agriculture and Alpine ski sector; (ii) evaluate the evolution in time of the forecast skill at the monthly resolution, to identify the maximum lead time at which the forecast can still be considered useful; (iii) provide both model developers and stakeholders with detailed information on the skill and limitations of the current seasonal forecast models over the Mediterranean region, to be used as a guidance for improving forecast systems and making the best use of their outputs in practical applications.

We consider the seasonal forecast systems available in the Copernicus Climate Data Store (C3S) archive, and

we perform a multi-model assessment of the skill of near-surface air temperature and precipitation forecasts over the Mediterranean domain, evaluating a selection of representative skill scores, each of which can test a different feature of the forecast ensemble. The skill of the forecast systems is quantified with respect to a simple forecasting method based on climatology, and the added value of the forecast systems is assessed at different lead times. In addition to the analysis of individual forecast systems, we evaluate the Multi-Model Ensemble (MME) stacking together all the members (equally-weighted) of all forecast systems. Compared to previous studies (Mishra et al. 2019), this analysis at a monthly scale allows to look with a finer temporal detail at the differences among the forecast systems, and sets the basis to narrow the effort in searching for different sources of predictability through the individuation of the time scale at which seasonal forecasts become drastically less skilful (Board et al. 2016). We also assess the impact of the ensemble size and of temporal averaging on the forecast system performances. Linking the performances of the seasonal forecast systems to their modelling schemes requires a deep knowledge of individual forecast systems and is out of the scope of the paper.

The paper is structured as follows: Sect. 2 briefly describes the seasonal forecast systems, the two simple forecasting methods based on climatology and persistence, and the reference data used in this study; Sect. 3 presents the different skill scores used and which questions they address; Sect. 4 shows the results of the evaluation of temperature and precipitation anomaly forecasts; Sections 5 and 6 discuss the results and draw the conclusions.

2 Data

2.1 Model datasets

The present analysis considers five seasonal forecast systems available in the C3S archive (retrieved on October 18th, 2018) which provide near-surface air temperature and precipitation data at monthly temporal resolution and at 1° by 1° spatial resolution: European Centre for Medium-range Weather Forecast System 5 (ECMWF), Météo France System 6 (MF), UK Met Office GloSea5-GC2 (UKMO), Centro Euro-Mediterraneo sui Cambiamenti Climatici SPS3 (CMCC) and Deutscher Wetterdienst GCFS 2.0 (DWD); please refer to Table 1 for the model details. For all forecast systems, we consider all available hindcasts initialised on May 1st and November 1st, and issued for the 6 months ahead. We indicate as lead time 0 the month in which the forecast is initialized (May or November). Lead time 1 is the 1st month after the initialization (June or December), and so on. When dealing with seasonal anomalies, we never use the term "lead time" as it would be confusing: DJF and JJA seasonal anomalies are calculated considering the monthly values from the forecasts initialized in November and May, respectively. We analyse the longest period common to all systems, i.e. 22 years from 1993 to 2014.

In addition, we also consider blended forecast systems and simpler approaches defined as follows:

- The multi-model ensemble (MME) including all the available ensemble members of the seasonal forecast systems cited above (148) transformed into anomalies with respect to each model's climatology
- The multi-model ensemble small (MMES), similar to MME but including only five ensemble members for each seasonal forecast system randomly chosen among all the available members (25 ensemble members in total)

Table 1 Seasonal forecast systems and simpler approaches considered in this study

Acronym	Prediction system	Institution	Ens. size	References
ECMWF	SEAS5	European Centre for Medium-Range Weather Forecasts	25	Johnson et al. (2019)
MF	System 6	Météo-France	25	Dorel et al. (2017)
UKMO	GloSea5-GC2	UK Met Office	28	Maclachlan et al. (2015)
CMCC	CMCC-SPS3	Centro Euro-Mediterraneo sui Cambiamenti Climatici	40	Sanna (2017)
DWD	GCFS 2.0	Deutscher Wetterdienst	30	Fröhlich et al. (2020)
MME	Multi-Model Ensemble	–	148	Section 2.1
MMES	Multi-Model Ensemble Small	–	25	Section 2.1
PERS	Persistence	–	30	Section 2.1
CTRL	Control	–	21	Section 2.1

- A persistence forecast (PERS) generated from the ERA5 anomaly at lead time 0: the forecast for each following month is the ERA5 anomaly at lead time 0, to which we applied a Gaussian anomaly kernel-dressing to obtain an ensemble forecast (Smith et al. 2015). The kernel dressing is performed for each starting date, lead time and grid point by estimating a Gaussian distribution using 2 parameters: (1) the mean, represented by the deterministic persistence forecast (the ERA5 anomaly at lead time 0), and (2) the standard deviation, represented by the root mean square of the residuals of the deterministic persistence forecast (difference between the ERA5 anomalies at that lead time and the ERA5 anomalies at lead time 0) calculated over the remaining 21 starting dates following an out-of-sample approach. From the resulting distribution 30 values are randomly selected to generate the ensemble. We verified that the use of a Gaussian distribution for generating the PERS ensemble forecasts is adequate for both temperature and precipitation by performing a Kolmogorov-Smirnov test on the residuals. The residuals follow a Gaussian distribution for both temperature and precipitation
- A climatological control forecast (CTRL) generated from the ERA5 anomalies by choosing, for each starting date and lead time, all the historical ERA5 values except for the one corresponding to that date, in order to form an ensemble of 21 members (1 less than the number of forecasts). This simple forecast, based on the observed climatology, is also employed as the reference forecast for the evaluation of the skill scores (see Sect. 3)

We consider and analyse all these datasets at monthly scale.

2.2 Reference dataset

To evaluate the seasonal forecast systems and simpler approaches described in Sect. 2.1, we employ the ERA5 reanalysis (Hersbach et al. 2020) as a reference dataset. ERA5 2-m air temperature and total precipitation data at 0.25° spatial resolution and monthly temporal resolution have been downloaded from the Copernicus Climate Data Store archive and upscaled to match the grid of the seasonal forecast systems at 1° resolution. The upscaling has been performed with a first-order conservative remapping using the Climate Data Operator command line tools (Schulzweida 2019). In order to compare forecast and reanalysis data, temperature and precipitation fields are considered in °C and mm/day, respectively.

2.3 Domain of study

We focus on the Mediterranean area as the domain of study (11°W–37°E; 31°N–52°N). Seasonal forecast fields include 22 gridpoints in latitude by 49 gridpoints in longitude, each representing an area of 1° by 1°.

3 Forecast verification methods

Probabilistic forecasts can be evaluated considering their quality, i.e the correspondence between the forecasts and the matching observations, and/or their value, i.e. the incremental economic value and/or other benefits realised by decision-makers through the use of the forecast (Murphy 1993). This study will focus on the "quality" aspect. An assessment of the "value" is fundamental but also specific for many sectorial applications and we leave it outside the scope of our analysis.

Table 2 summarises the scores used in this paper to assess the quality of the forecast. Each score is an attempt to measure one or more "attributes" of the forecast quality, following Murphy (1993).

The overall analysis is conducted considering for each forecast system and for simpler approaches monthly anomalies of air temperature and precipitation. Temperature and precipitation anomalies are calculated as the difference between a forecast and the corresponding model climatology. In order to remove the effects of temporal trends on the forecast skill, detrended anomalies are employed. For each forecast system, ensemble member, lead time and gridpoint, the anomalies are detrended by removing a linear function of time, obtained by least-squares regression of the model ensemble mean over time.

Table 2 List of metrics considered in this study with the target features addressed and the main references

Score	Attribute	References
ACC	Association	Jolliffe and Stephenson (2011), Wilks (2011)
RKH	Ensemble quality	Hamill and Colucci (1997), Anderson (1996)
BS	Resolution, reliability, accuracy	Wilks (2011), Mason (2004)
AUC	Discrimination	Jolliffe and Stephenson (2011)
CRPS	Accuracy, sharpness	Hersbach (2000)

3.1 Anomaly correlation coefficient (association)

The anomaly correlation coefficient (ACC) describes the strength of the linear relationship between forecast and observed anomalies (also referred to as association). It is widely exploited in seasonal forecast verification and is the only deterministic score considered in this paper. Here it is intended as the Pearson correlation computed in time between the ensemble mean forecast anomalies and the ERA5 reference anomalies at each point of the domain, for the winter and the summer seasons. ACC ranges between −1 and 1. ACC is not sensitive to bias, so it does not guarantee accuracy. The confidence interval is computed by a Fisher transformation and the significance level relies on a one-sided student-T distribution. Significance is assessed at 95% confidence level.

3.2 Rank histograms (ensemble quality)

Rank histograms (RH, Hamill and Colucci 1997; Anderson 1996) measure the ensemble quality, and, in detail, whether the probability distribution of observations is well represented by the ensemble. Rank histograms show the frequency of the rank of the observed value relative to values from the ensemble forecast, sorted in increasing order. If the forecast distribution reliably reproduces the distribution of possible outcomes, then the observed value should be a random draw from this same distribution, and it should occur in each of the bins an equal number of times (Hamill 2002). Therefore, the proportion of the total number of observations in each bin should follow a uniform distribution (Troccoli et al. 2008), and the perfect RH should be flat, with each bin assuming the same value. U-shaped and reversed U-shaped distribution suggest a too narrow ensemble spread (underdispersion) and too wide ensemble spread (overdispersion), respectively. An asymmetric shape means that the ensemble under- or overestimates the reference value. RHs are normalised with respect to the perfect value $1/(n+1)$, where n is the number of ranks that is equal the number of ensemble members) for the sake of comparison among different forecast systems. RHs give information on the ensemble quality, highlighting possible issues of over- or underdispersion and biases. RHs do not indicate skillful or sharp forecasts, in fact climatological forecasts show flat rank histograms (by definition) but they are not useful. For this reason, RHs have to be used in combination with other metrics for a comprehensive skill assessment. In our analysis, to summarise information over the different lead times, RHs are presented in the form of heatmaps as a function of rank and lead time.

3.3 Brier score (accuracy, reliability, resolution, uncertainty)

The Brier score (BS) is a strictly proper scoring rule for forecast verification, and it represents the mean square error of the probability forecast for a binary event, for example rain/no rain (Wilks 2011; Mason 2004, and references therein). It is a measure of the overall accuracy of the forecast, that is the average correspondence between individual forecasts and the observations (Wilks 2011). It can be partitioned into three components: (i) reliability, i.e. the extent to which forecast probabilities match the observed relative frequencies and the associated error is small, (ii) resolution, i.e. the degree to which a forecast can separate different outcomes (a forecast based on climatology has no resolution), (iii) the uncertainty, i.e. the degree of variability in the observations, which is independent of the forecast (Hersbach 2000).

A common way to display seasonal predictions is by means of tercile-based forecasts, showing the probabilities to have anomalies in the lower, middle or upper tercile of the distribution (WMO 2018). In our analysis we transform continuous forecasts into tercile-based forecasts and then we test their overall accuracy using the original definition of the Brier Score for multi-category forecasts (Brier 1950, eq. 2). This scalar is a measure of accuracy and it is calculated for each model, lead time and grid point.

3.4 Area under the ROC curve (discrimination)

The area under the receiver operating characteristic (ROC) curve, abbreviated as AUC (Jolliffe and Stephenson 2011), allows to evaluate binary forecasts, similarly to the Brier Score.

The AUC measures the discrimination, i.e. the ability of the forecast to discriminate between events and non-events. If forecast probabilities issued when an event occurs tend to be higher than those issued when such event does not occur, the probability forecasts have discrimination (Bradley and Schwartz 2011; Bradley et al. 2019). AUC is not sensitive to bias, so it does not provide information on the forecast reliability. A biased forecast may still have good discrimination and produce a good ROC curve, which means that it may be possible to improve the forecast through calibration. The ROC can thus be considered as a measure of potential usefulness.

Given an ensemble forecast for a binary event, for example temperature anomaly in the upper tercile, the ROC curve shows the hit-rate (HR) against the false-alarm rate (FAR) for different probability thresholds. The Area Under the ROC Curve (AUC) resulting from the display of the pairs of HR and FAR for different thresholds is calculated separately for each tercile and then averaged over the three terciles.

3.5 Continuous ranked probability score (accuracy, sharpness)

The continuous ranked probability score (CRPS) is a measure of the overall accuracy of the ensemble forecast (Bradley et al. 2019). It indirectly measures also the sharpness of the forecast, in that among several accurate forecasts it rewards those with smallest ensemble spread. The CRPS is defined as the difference between the cumulative distribution function (CDF) of a forecast and the respective observation, the latter being represented by a Heaviside step function. When the CDF of the forecasts well approximates the step function, it produces relatively small integrated squared differences, resulting in good CRPS score. Brier score and CRPS are complementary measures: the former provides information on the accuracy of tercile-based forecasts, the latter evaluates the overall accuracy and sharpness of the forecast distribution, considering the entire permissible range of values for the considered variable. A drawback of the CRPS is that an increasing ensemble size inflates it (Ferro 2014; Ferro et al. 2008). Therefore the Fair CRPS (FCRPS), a modified version of the CRPS addressing this issue, is implemented and considered in this study.

3.6 Skill scores

In this study, unless expressly noted otherwise, the scores are presented as skill scores (SS). The skill scores directly indicate the skill of the forecast with respect to the climatological forecast (CTRL, see Sect. 2.1 for details). Values of the skill scores generally range between negative (performances worse than climatological forecast) and positive values (improvements with respect to the climatological forecast). A value of 1 indicates perfect forecasts while null values indicate no improvements with respect to the climatological forecast. The choice of using the climatology as a benchmark is widely diffused, but persistence could be another possibility (not explored in this study).

BS and FCRPS are calculated for each starting date, lead time and grid point, then averaged over all starting dates and converted into skill scores as follows:

$$SS = \frac{S - S_{ref}}{S_{perf} - S_{ref}}, \tag{1}$$

where SS is the value of the skill score, S is the value of the score of the forecast against the observations, S_{ref} is the value of the score of the climatological forecast against the observations and S_{perf} is the value of the score in the theoretical case that forecasts perfectly match observations. The score of the climatological forecast S_{ref} is calculated using the CTRL approach (Sect. 2.1). We will call the resulting skill scores BSS and FCRPSS respectively in the following.

The AUC Skill Score (AUCSS), instead, is derived using the following formula (Wilks 2011, Equation 8.46):

$$AUCSS = 2 \cdot AUC - 1. \tag{2}$$

The spatial variability of the BSS, FCRPSS and AUCSS as a function of the lead time is presented in the form of maps and summarised by means of boxplots, where for each box, the lower hinge, the median, and the upper hinge correspond to the first, second, and third quartiles of the distribution, respectively, while the lower and upper whiskers extend from the bottom and the top of the box up to the furthest datum within 1.5 times the interquartile range. If there are any data beyond that distance, they are represented individually as points.

All the analyses are performed in R v3.6.3 (R Core Team 2019) using the following packages: s2dverification v2.9.0 (Manubens et al. 2018), easyVerification v0.4.4 (Bhend et al. 2017), SpecsVerification v0.5-3 (Siegert 2020) and verification v1.42 (NCAR - Research Applications Laboratory 2015).

4 Results

In this Section, we present the results obtained for the different skill scores defined in Sect. 3. The skill scores are calculated with respect to the reference forecast based on the climatology (CTRL) and using ERA5 as reference for the observed climate. Anomaly correlation coefficients are obtained from seasonal anomalies, while the other scores are calculated on a monthly basis (Sect. 3).

4.1 Anomaly correlation

4.1.1 Winter

For each forecast system considered in this study Fig. 1 shows the time correlations of the ensemble mean temperature anomaly forecasts with respect to ERA5 for the winter season. Figure 2 shows the same plots but for precipitation. They have been derived using forecasts initialised on November 1st and averaging over the December–January–February period, i.e. lead times 1–3. All models show significant temperature correlation patterns in the Western Mediterranean and off the Atlantic coast. Most models also show a significant correlation pattern over North Africa, except for the ECMWF model. UKMO shows widespread and significant temperature anomaly correlation over central Europe (Northern Italy, France and Germany), including the Alpine region. Large and significant correlations are often found over the sea. Since sea-surface temperature (SST)

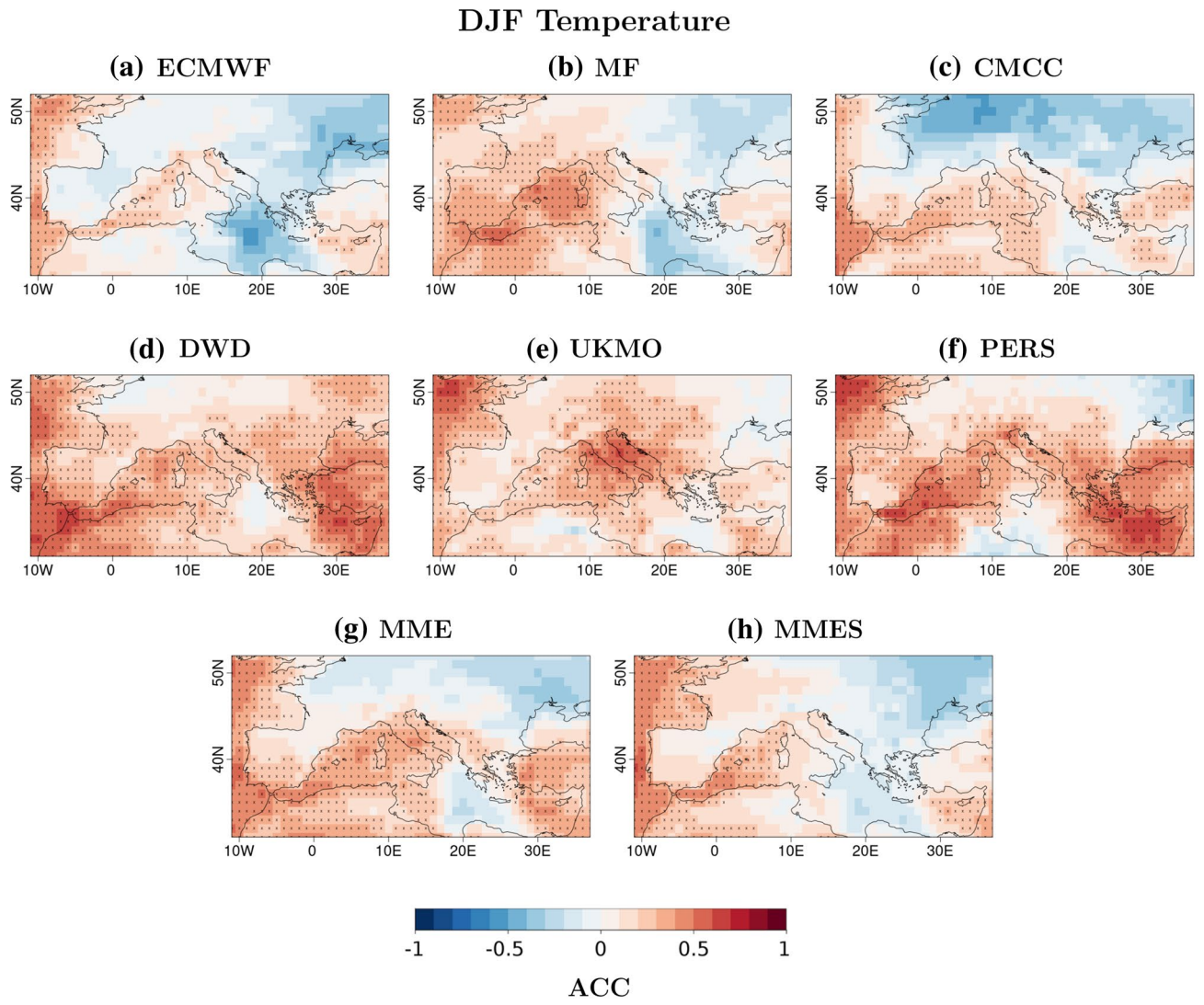

Fig. 1 Anomaly correlation coefficients of winter (DJF) near-surface air temperature forecasts with respect to ERA5, for all the forecast systems and simpler approaches listed in Table 1. Significant correlations (95% confidence level) are indicated by stippling. Forecasts are initialised on November 1st and refer to the hindcast period 1993–2014. The ACC map for CTRL is omitted since it provides trivial information

affects the overlying air temperature, the slower variability of sea-surface temperature with respect to land-surface temperature is likely to improve air temperature predictability over the sea. The slow variability of air temperature anomalies over the few months following the forecast initialisation date is supported by the high and significant correlation obtained when forecasts are based on persistence (PERS) (Fig. 1f). Indeed, the forecasts based on persistence outperform those provided by the seasonal forecast systems in terms of anomaly correlation.

Anomaly correlations of winter precipitation with respect to ERA5 are patchier compared to those found for temperature (Fig. 2). A common feature among all forecast systems and the two Multi-Models Ensembles (MME, MMES) is the relatively high and significant anomaly correlation over the Iberian Peninsula and the Eastern Mediterranean. Most of them (MF, CMCC, UKMO, MME and MMES) show high and significant correlations also over the Alpine mountain range. ECMWF shows poor correlation with observations over the Central Mediterranean. Winter precipitation forecasts based on persistence (PERS) show no significant correlation with ERA5 except for few gridpoints over the Iberian peninsula/Atlantic coast and North Africa/Central Mediterranean.

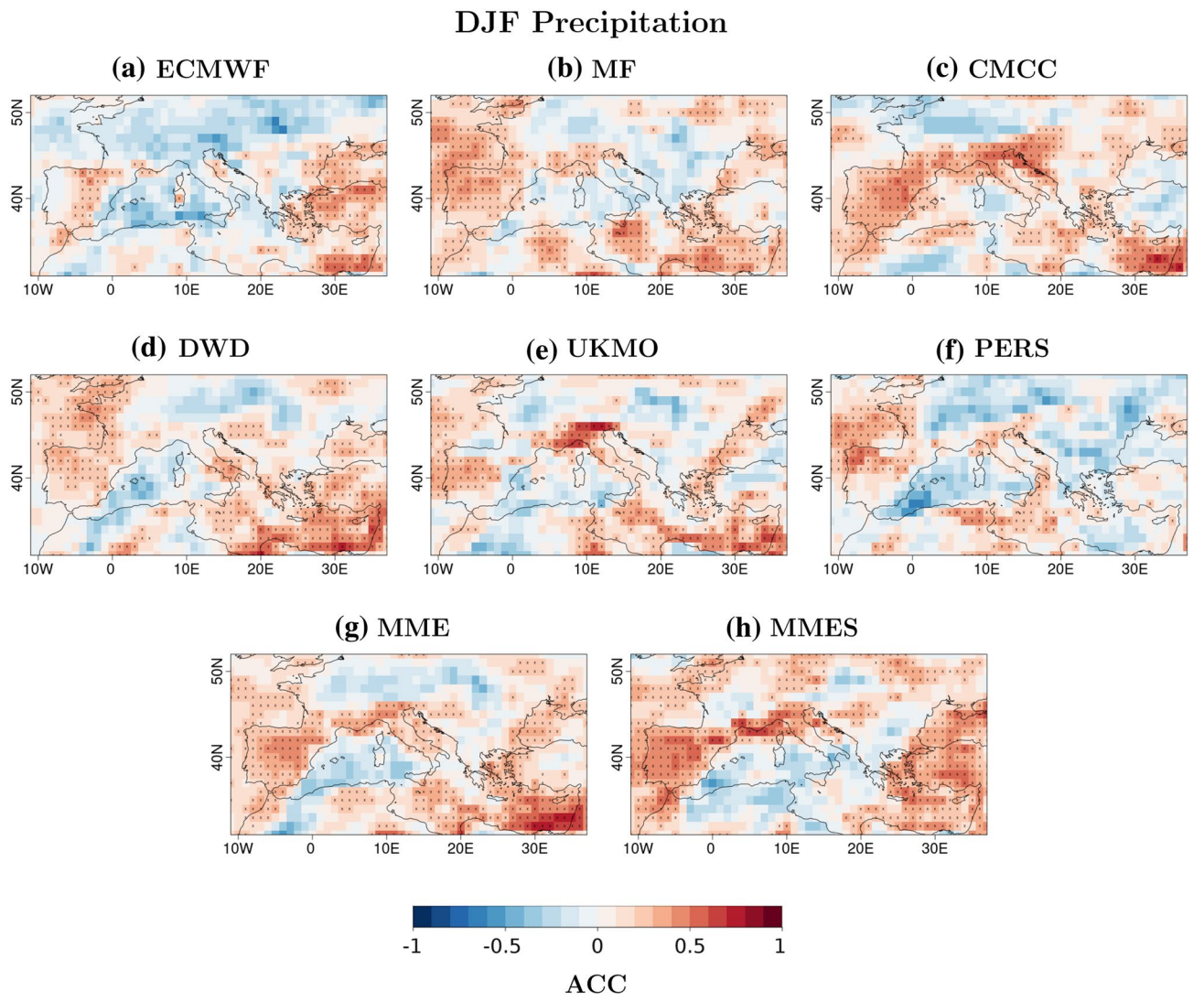

Fig. 2 The same as Fig. 1 but for DJF precipitation

4.1.2 Summer

Similarly to Fig. 1, Fig. 3 shows the correlation between air temperature anomaly forecasts and ERA5 for each model, but for the summer season, i.e. using forecasts initialised on May 1st and averaging monthly values over the June–July–August period, respectively lead time 1–3. Figure 4 shows the same plots but for precipitation.

All models show high and significant summer temperature anomaly correlations over the Eastern Mediterranean and Eastern Europe, and most of them also over the Iberian peninsula and North Africa. The best performing model in terms of summer temperature ACC is CMCC, which shows high and significant correlations over most of the domain. The persistence forecast (PERS) produces significant and widespread correlations over most of the Mediterranean domain, outperforming many forecast systems.

Correlations of summer precipitation anomalies (Fig. 4) are more patchy, nonetheless significant over some parts the Mediterranean region. All models show significant positive correlation over the Iberian peninsula, however this area receives scarce precipitation in summer and correlations based on very low precipitation values should be considered with caution. Some models show also significant positive correlation over Eastern Europe and the Black Sea coast. Significant correlations over these areas are also present in the MME and MMES, which outperform many individual models. Forecasts based on persistence (PERS) provide significant correlations at fewer gridpoints compared to the MME.

In conclusion, temperature anomaly forecasts based on persistence (PERS) have a high and significant correlation with ERA5 over most of the Mediterranean area in both winter and summer, outperforming individual seasonal forecast

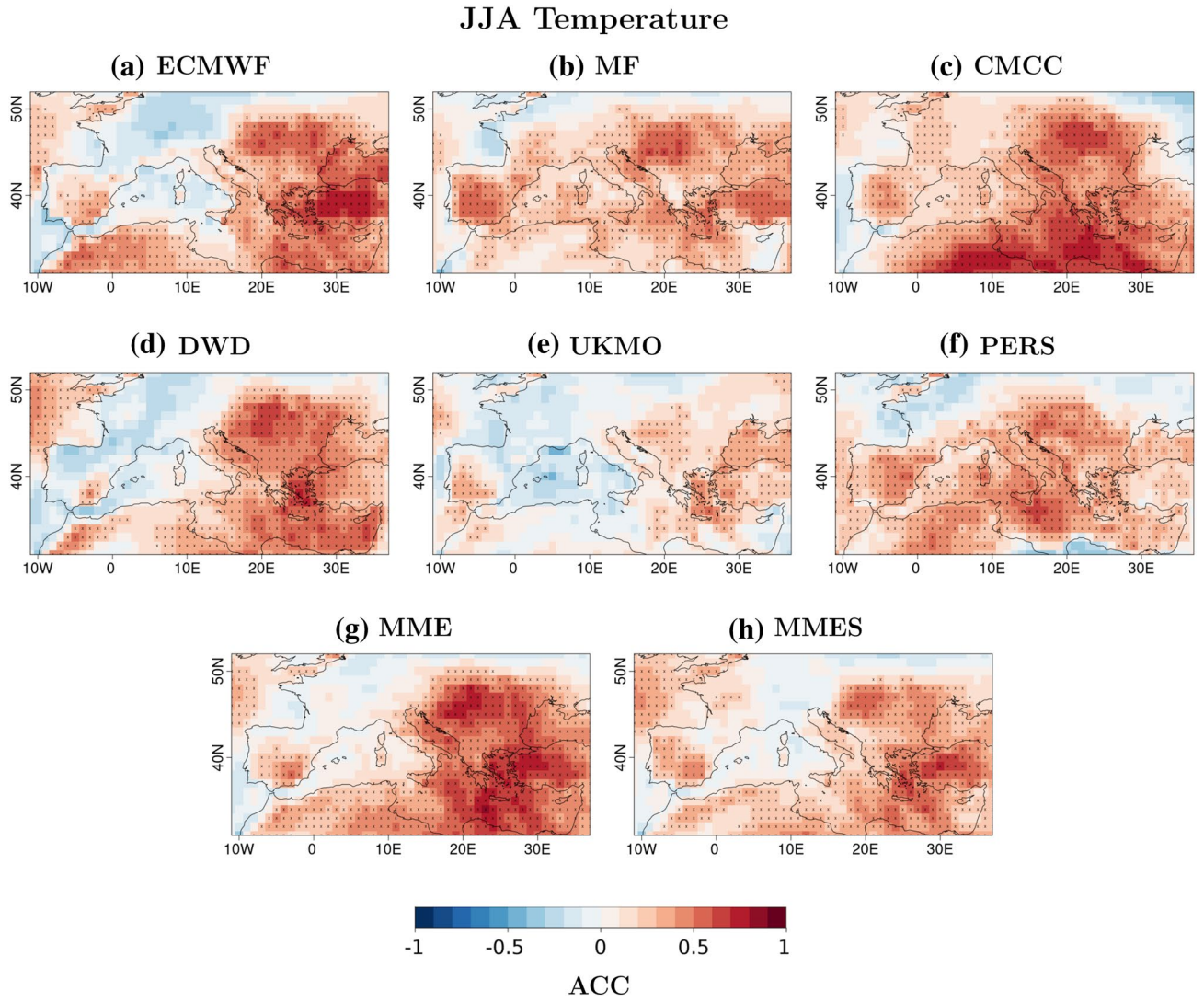

Fig. 3 Anomaly correlation coefficients of summer (JJA) near-surface air temperature forecasts with respect to ERA5, for all the forecast systems and simpler approaches listed in Table 1. Significant correlations (95% confidence level) are indicated by stippling. Forecasts are initialised on May 1st and refer to the hindcast period 1993–2014. The ACC map for CTRL is omitted since it provides trivial information

systems and the multi-model means. Compared to temperature, precipitation anomaly forecasts based on persistence (PERS) show lower correlations over most of the domain. By contrast, precipitation forecasts obtained with the multi-model mean (MME) show areas with significant correlation with observations, outperforming many individual forecast systems.

4.2 Rank histograms

Rank histograms for winter and summer temperature anomaly forecasts are reported in Fig. 5 panels a and b, respectively. For winter temperature anomalies, most models for most lead times show flat rank histograms, similarly to the case of a perfect forecast. The ensemble quality seems to be more a characteristic of a given model rather than a time-dependent feature (see, for example, ECMWF, MF and UKMO). Few are the exceptions: DWD and, to a smaller extent, CMCC show a U-shaped pattern at lead time 0, suggesting a too small ensemble spread at the beginning of the forecast. For summer temperature anomalies, CMCC and, to a smaller extent, UKMO, MME and MMES at lead times 0 and 1, show a reverse U-shaped pattern, indicating a too large ensemble spread and too large interannual variability.

RHs for winter precipitation anomalies are generally flat (Fig. 6a), except for CMCC and ECMWF, which present a peak at rank zero indicating that observed anomalies are more often lower than the forecast anomalies.

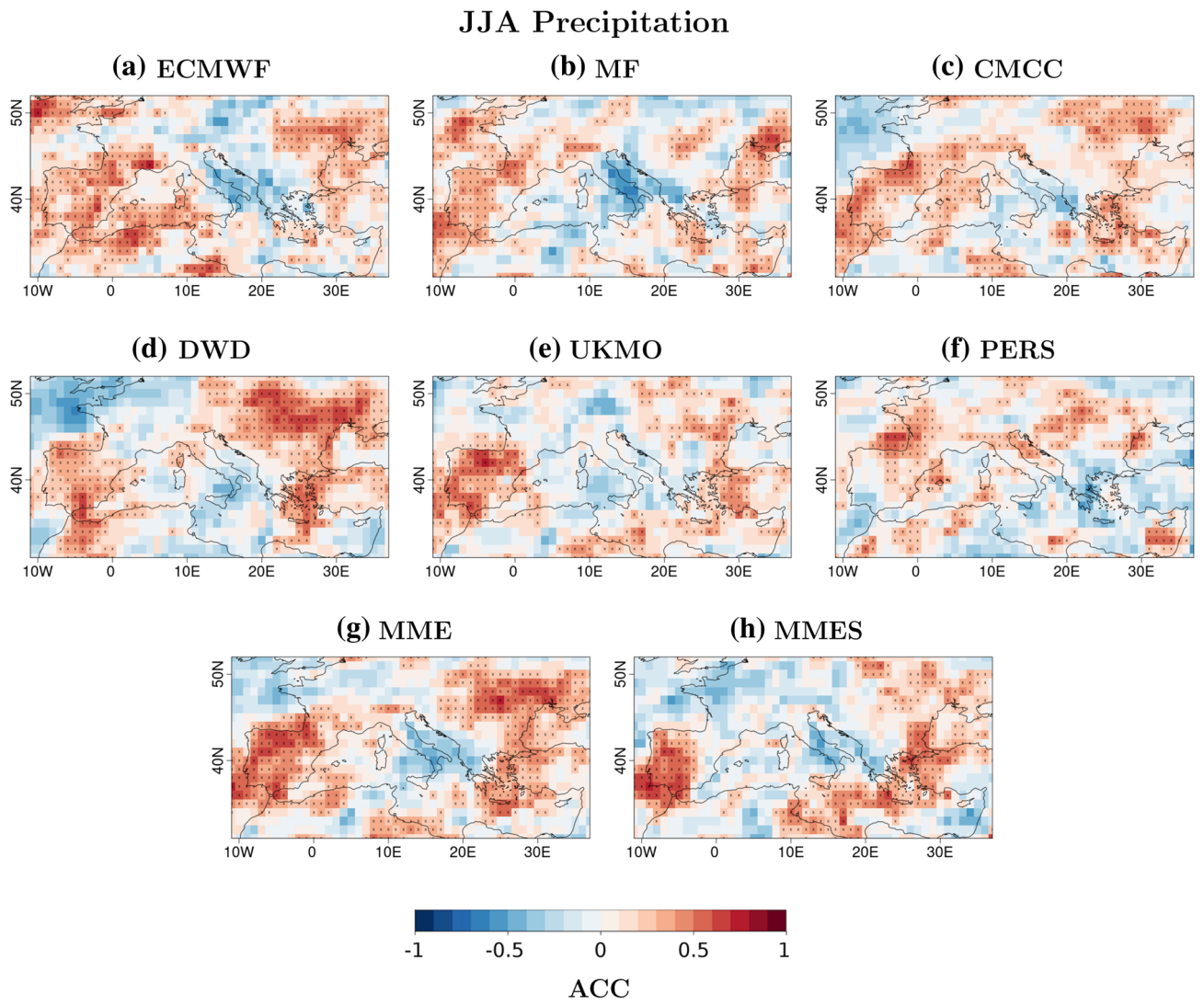

Fig. 4 The same as Fig. 3 but for JJA precipitation. Forecasts are initialised on May 1st

For summer precipitation anomalies (Fig. 6b), DWD, CMCC and to a lesser extent UKMO, show a U-shaped RH, indicating that the ensemble is underdispersive. These models generally underestimate the interannual variability of precipitation anomalies and the intensity of summer precipitation extremes (both dry and wet). On the other hand, ECMWF, MF, MME and MMES share an asymmetric pattern indicating that the observed anomalies are often lower than forecast anomalies: these models tend to overestimate the observed precipitation anomalies, and thus the amount of summer precipitation.

Overall, the MME and MMES ensembles show the best agreement with the observations, despite an overestimation of summer precipitation anomalies and a slight tendency towards overdispersion at lead times 0 and 1 for summer temperature.

For all seasons and variables, DWD has a strongly U-shaped histogram at lead time 0, suggesting a systematic underdispersion of the model ensemble at the beginning of the forecasting period.

Forecasts based on the climatology (CTRL, not shown) are by construction well calibrated for all variables and seasons. In fact, the CTRL ensemble forecast is made using ERA5 climatological values, and the outcome is a random value from the ensemble forecast. Forecasts based on persistence (PERS) are also well calibrated: this seems to suggest that an ensemble forecast built on the ERA5 anomaly at lead time 0 following the method described in Sect. 2.1 is well balanced compared to the observed anomalies at longer lead times.

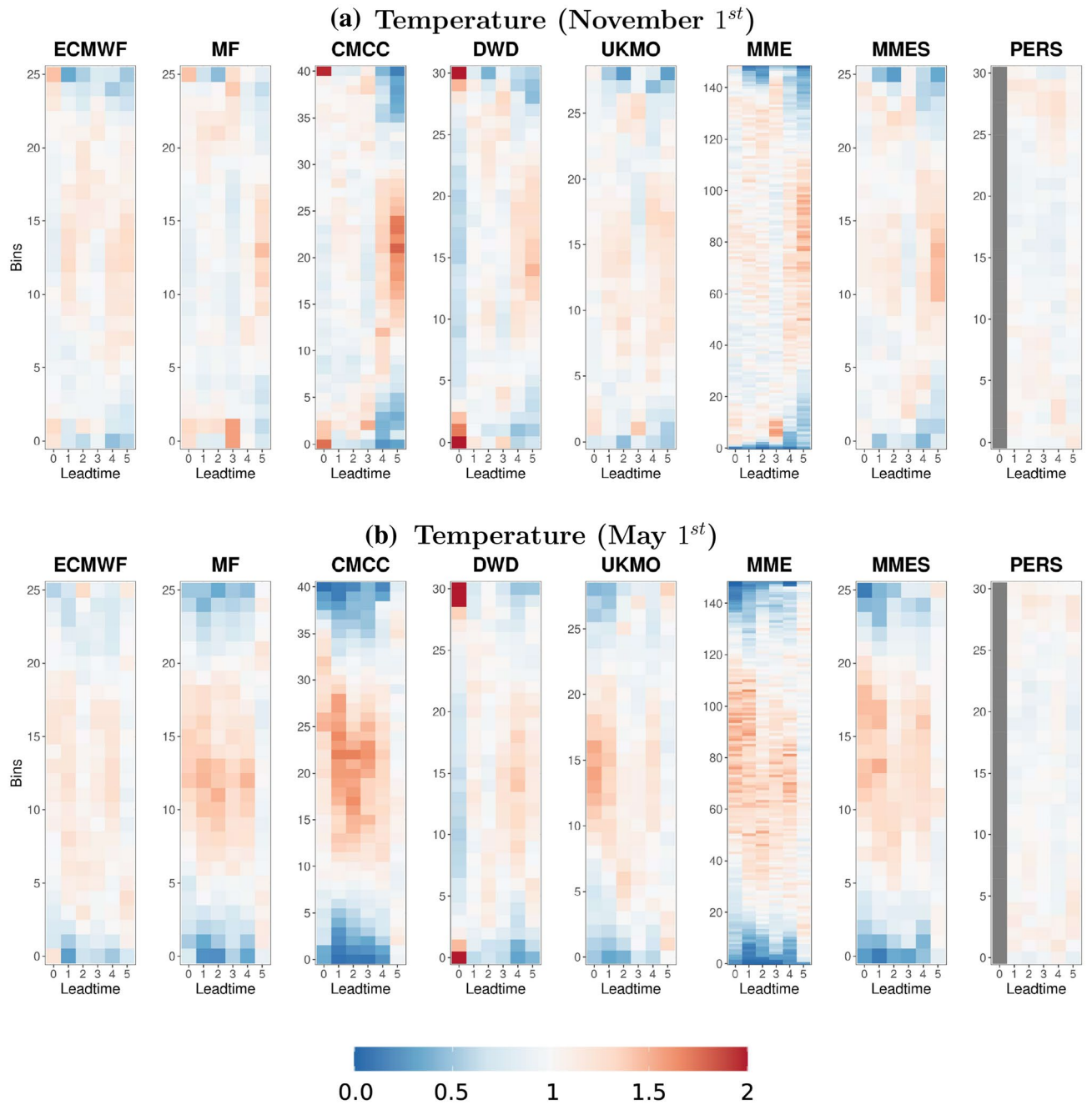

Fig. 5 Rank histograms as a function of lead time for temperature anomaly forecasts for each model and for November 1st (**a**) and May 1st (**b**) starting dates. The color indicates the normalized frequency of the rank of observations with respect to the ensemble. The forecast based on persistence (PERS) is not available at lead time 0, and it is reported in grey. CTRL is not shown

4.3 Brier skill score

Figure 7 summarises the Brier skill score statistics for temperature and precipitation anomaly forecasts for each model, season and lead time. We recall that positive BSS values indicate an improvement of the forecast system with respect to the climatological forecast (CTRL). The boxplots show the statistics of the distribution of the BSS values over the Mediterranean domain, so they are representative of the spatial variability of the skill score. For almost any forecast system, variable, season and lead time, positive BSS values are found in 75% up to 100% of the gridpoints. This percentage is lower, however above 50%, only for DWD at lead time 0, for which we have already presented an issue of underdispersion of the model ensemble for any variable and season (Sect. 4.2). Overall, all forecast systems show a

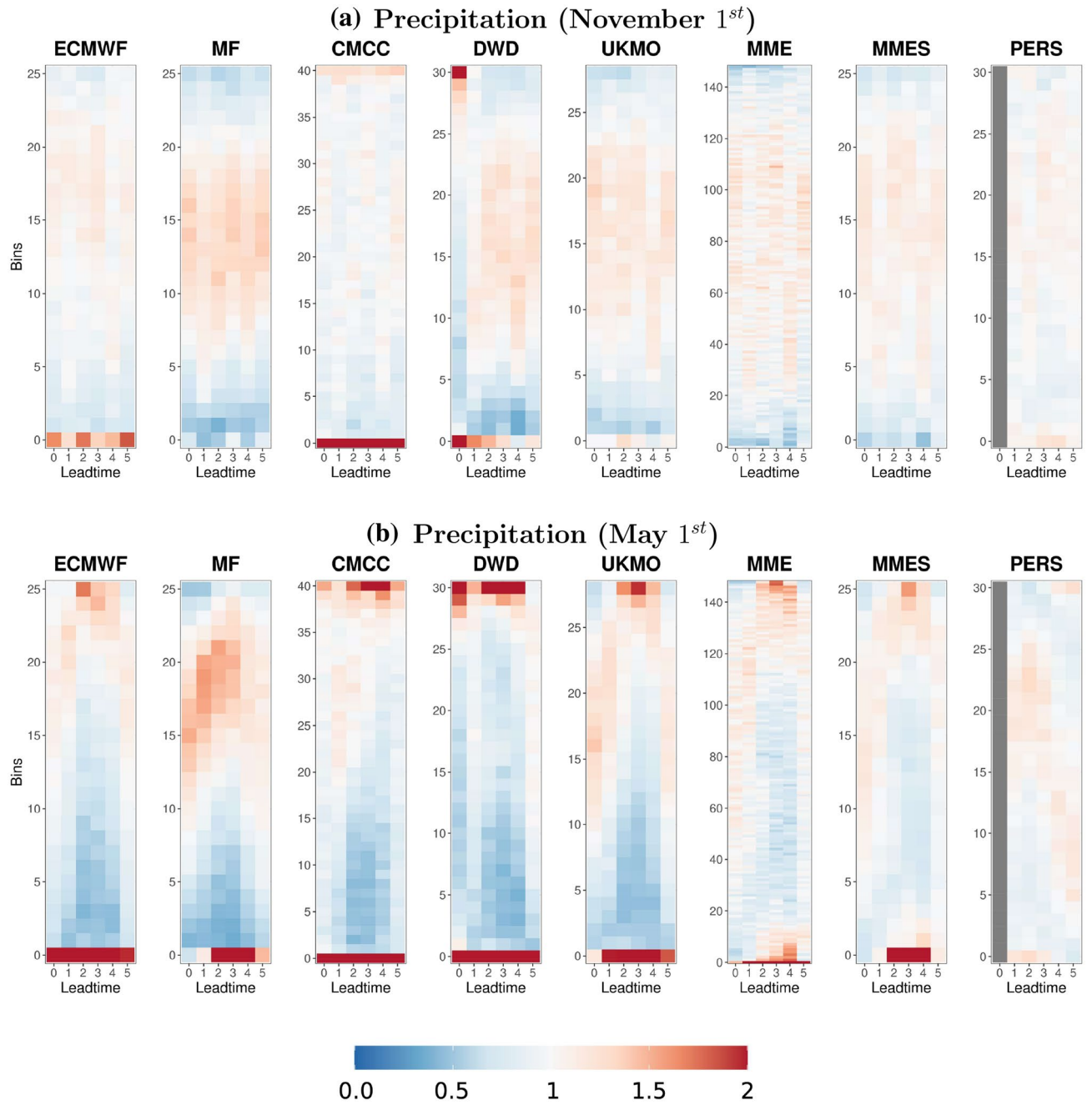

Fig. 6 The same as Fig. 5 but for precipitation

clear added value compared to the climatological forecast in terms of forecast error.

Generally, seasonal forecast systems have their highest BSS at lead time 0, then they show lower but stable values for longer lead times. The multi-model ensembles (MME and MMES) show comparable or higher median BSS with respect to individual models for each variable and lead time. In addition, MME has a smaller spread compared to any other forecast system, and positive BSS values are found for almost 100% of the gridpoints of the domain. MME shows a clear improvement with respect to the climatological forecasts in almost any point of the domain and for both temperature and precipitation. Temperature forecasts based on persistence (PERS) have progressively lower BSS, thus larger errors, at longer lead times. The median BSS for PERS temperature forecasts is positive at lead time 1 and close to zero or negative at longer lead times. The median BSS for PERS precipitation forecasts is negative for any

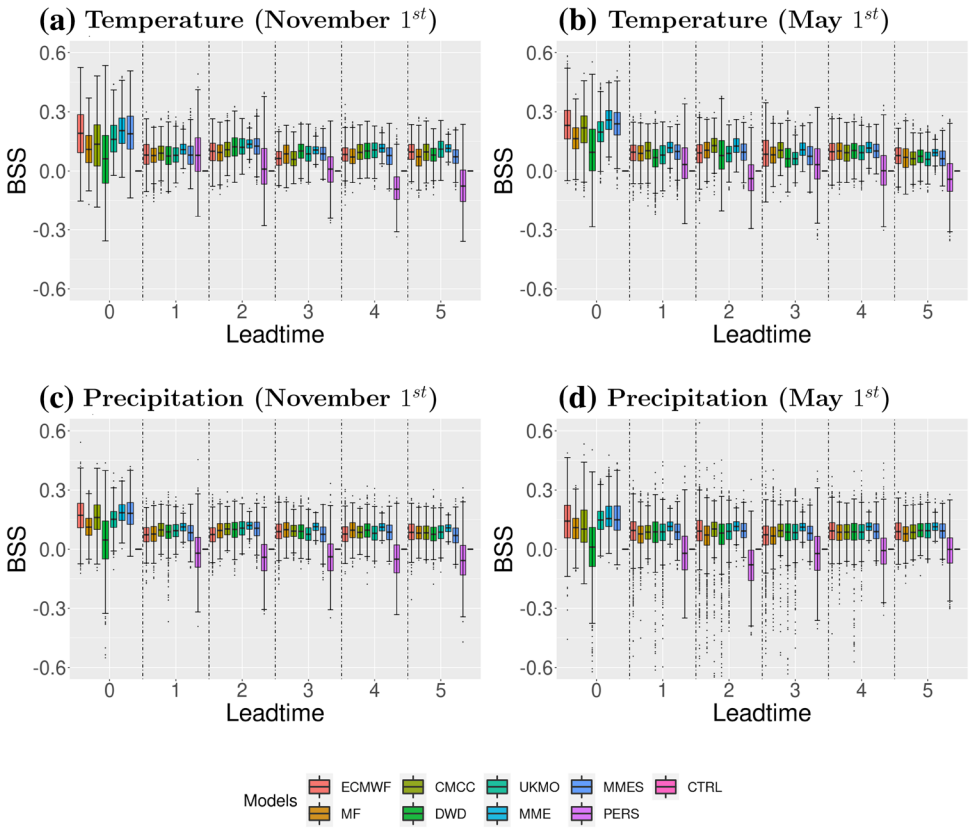

Fig. 7 Brier skill score of winter (**a**) and summer (**b**) temperature, winter (**c**) and summer (**d**) precipitation anomaly forecasts, for all models and lead times. The boxplots summarise the statistics of the distribution of the BSS over the Mediterranean domain

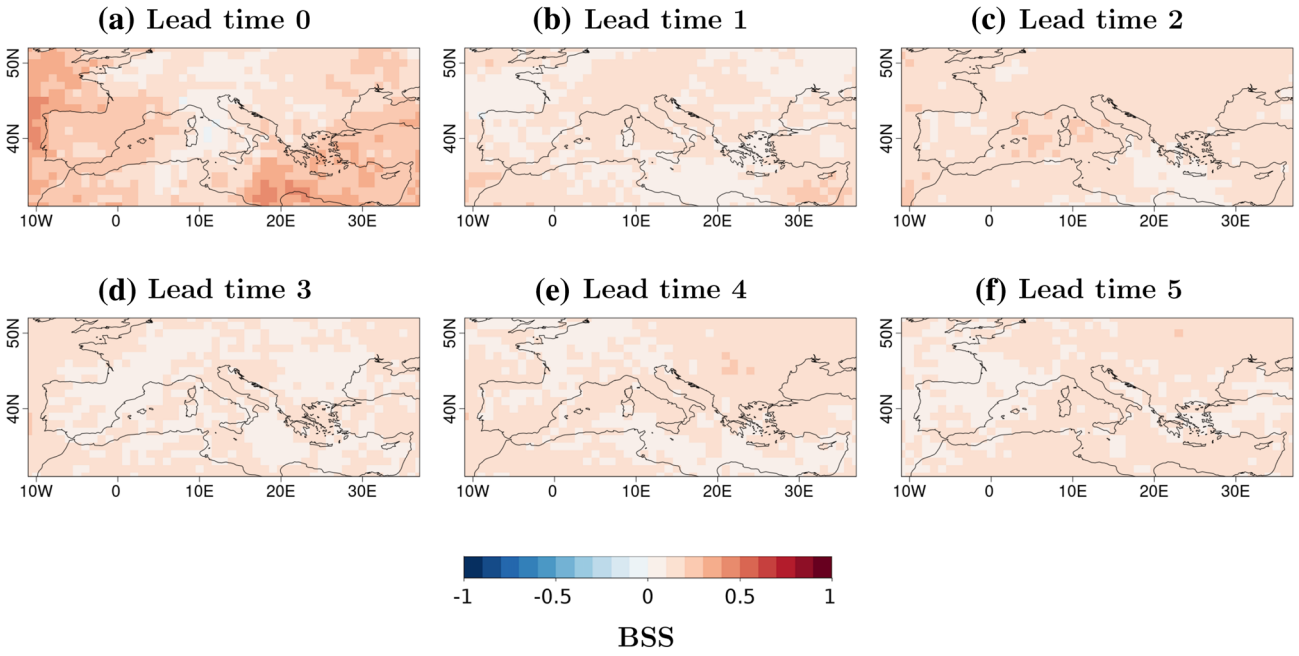

Fig. 8 Spatial pattern of the BSS for the Multi-Model Ensemble (MME) temperature anomalies forecasts for starting date November 1st and lead times 0–5

starting date and lead time, indicating that for most of the domain, the forecasts based on persistence are less reliable and with lower resolution than the climatological forecasts.

The spatial patterns of the BSS for the MME temperature anomalies forecasts initialised on November 1st for the 6 months ahead are shown in Fig. 8. The BSS is positive over all the domain at all lead times, confirming an added value of the MME forecast with respect to the CTRL forecast. At lead time 0 the MME approach shows positive BSS values, especially over the Southern and Eastern Mediterranean Sea and the Atlantic Ocean. At longer lead times, BSS is still positive with a more homogeneous pattern, indicating that the forecast error is slightly variable over the domain.

4.4 Area under the ROC curve skill score

The AUC skill score (AUCSS) quantifies the forecast discrimination, i.e. the ability of the forecast to discriminate between events and non-events, in our case to predict whether of not anomalies will fall in a given tercile. AUCSS statistics for each model, variable, season and lead time, averaged over the three terciles, are reported in Fig. 9 and the spatial pattern of MME at different lead times is depicted in Fig. 10.

Figure 9 shows that positive AUCSS values are obtained at lead time 0 for 75–100% of the domain for each variable, each season and for almost all models (except for MF, for which this percentage is slightly lower). At longer lead times the median AUCSS decreases: it remains positive or close to zero up to lead time 2, and it is generally higher for temperature than for precipitation. From lead time 3 models show both positive and negative median AUCSS values depending on the season, variable and lead time. Compared to BSS, AUCSS shows larger variability among different models and lead times.

Temperature forecasts based on persistence (PERS) provide comparable or slightly better scores than individual models up to lead time 3. In winter overall best results are obtained with the persistence model over Western Europe/Western Mediterranean and, to a lesser extent, over the Eastern Mediterranean (positive values up to lead time 2, not shown). Precipitation forecasts based on persistence have median AUCSS comparable to 0 at all lead times, for both winter and summer, revealing no added value with respect to the climatological forecast.

Compared to the BSS, the AUCSS of the seasonal forecast systems shows limited improvement with respect to the simple climatological forecasts. However, the skill is neither uniform in space nor over the three terciles, with higher values for the lower/upper terciles than for the middle tercile. As an example, we report the spatial pattern of AUCSS for temperature anomaly forecasts based on the MME, start date

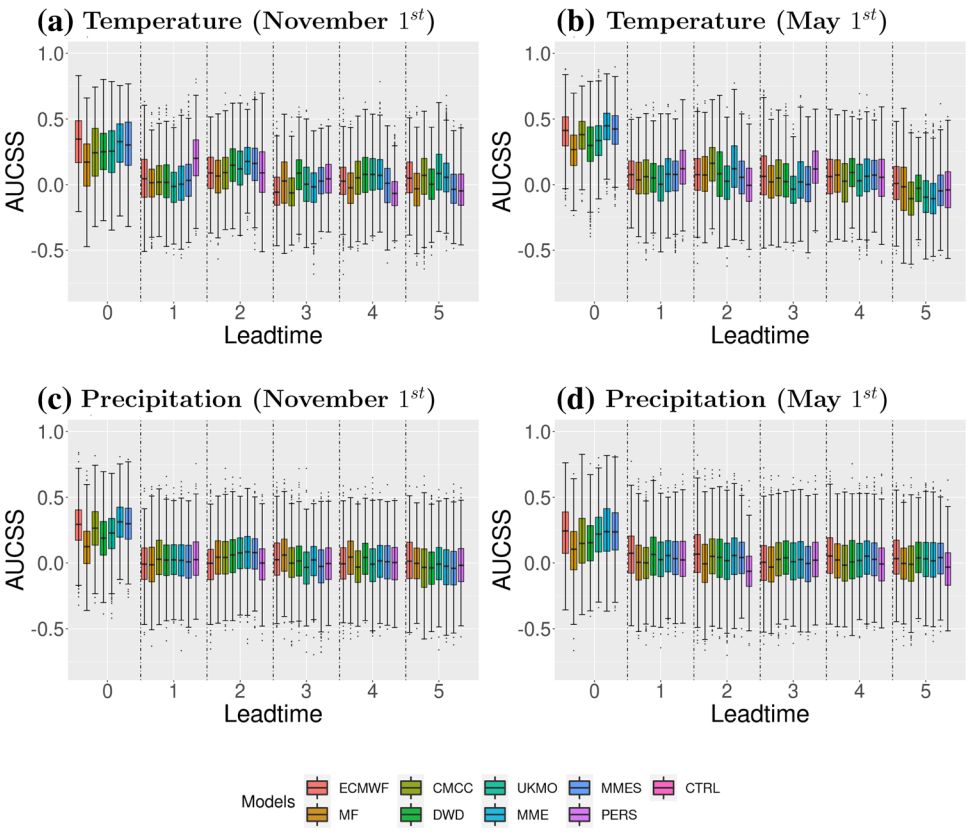

Fig. 9 Area under the ROC curve skill score (AUCSS) of winter (**a**) and summer (**b**) temperature, winter (**c**) and summer (**d**) precipitation anomaly forecasts, for all models and lead times. The boxplots summarise the statistics of the distribution of the AUCSS over the Mediterranean domain

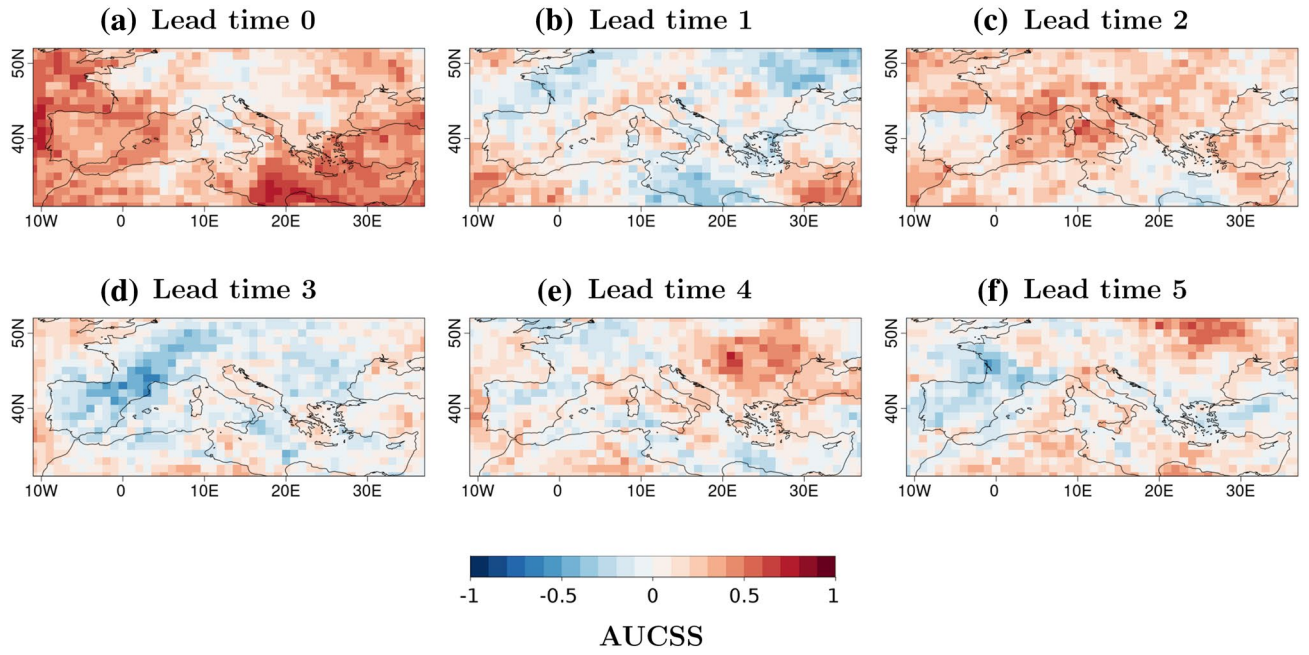

Fig. 10 Spatial pattern of the AUCSS for the multi-model ensemble (MME) temperature anomalies forecasts for the starting date November 1st and lead times 0–5

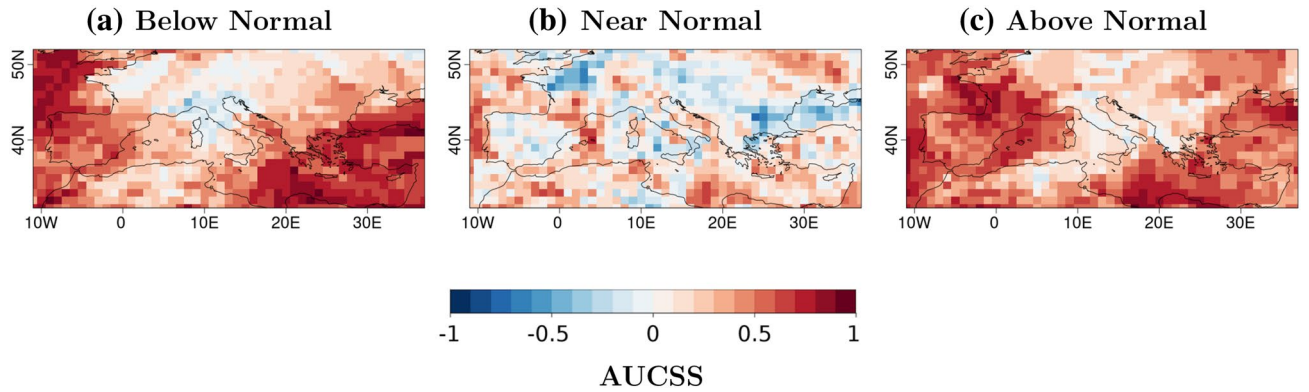

Fig. 11 AUCSS maps for MME temperature anomaly forecasts at lead time 0, starting date November 1st and for the three terciles: below normal (**a**), near normal (**b**), above normal (**c**)

November 1st, lead time 0, for the three terciles (Fig. 11). While for the middle tercile the mean AUCSS over the domain is 0.10, for the lower and upper terciles it reaches values of 0.41 and 0.44 respectively. When the AUCSS is averaged over the lower and upper terciles only, the median AUCSS over the domain slightly increases, especially for temperature, but also in this case the added value of the forecasts compared to the climatological forecast is visible up to lead time 2.

This analysis shows that seasonal forecasts provide a clear improvement with respect to the climatological forecast in terms of discrimination at lead time 0 and limited improvements up to lead time 2.

4.5 Fair continuous ranked probability skill score

The analysis of the FCRPSS is used to evaluate the overall accuracy and sharpness of the forecast. The forecast systems considered in this study show similar FCRPSS evolution for both temperature and precipitation and slight differences depending on the season (Figs. 12, 13).

At lead time 0 the median FCRPSS is generally positive, except for DWD for which we have already discussed low skill in terms of RH and BSS at the beginning of the forecast period. So, apart from this model, forecast systems generally have higher accuracy compared to the climatological forecast at lead time zero. At longer lead times the median

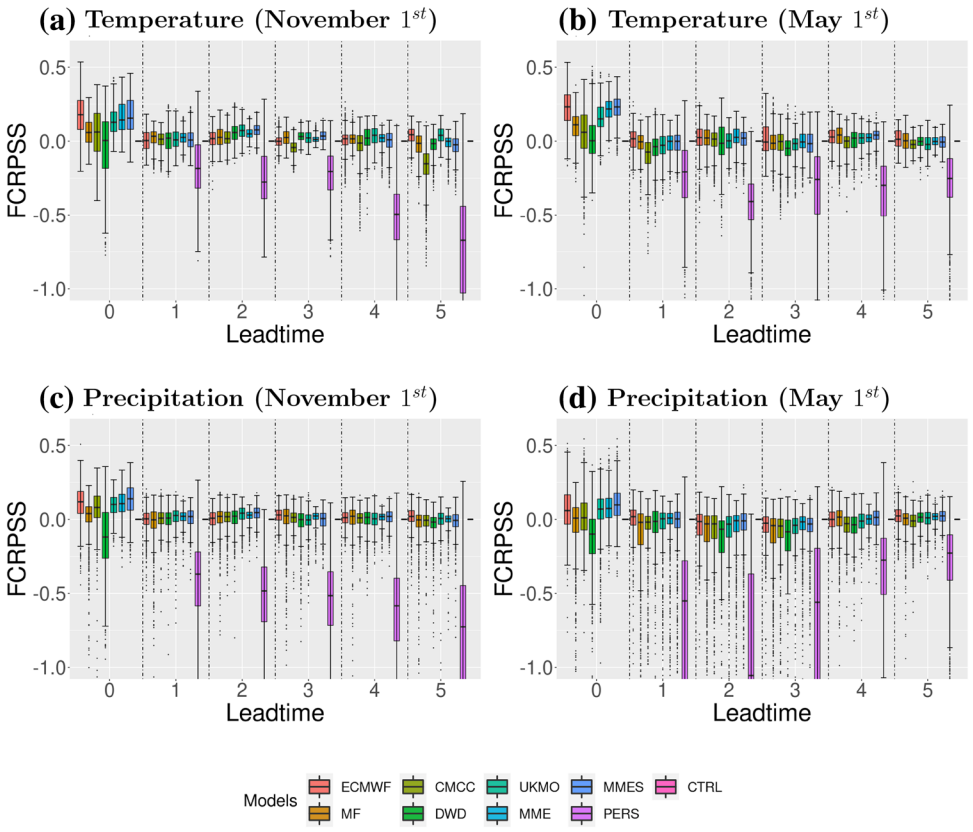

Fig. 12 Fair continuous ranked probability skill score (FCRPSS) of winter (**a**) and summer (**b**) temperature, winter (**c**) and summer (**d**) precipitation anomaly forecasts, for all models and lead times. The boxplots summarise the statistics of the distribution of the FCRPSS over the Mediterranean domain

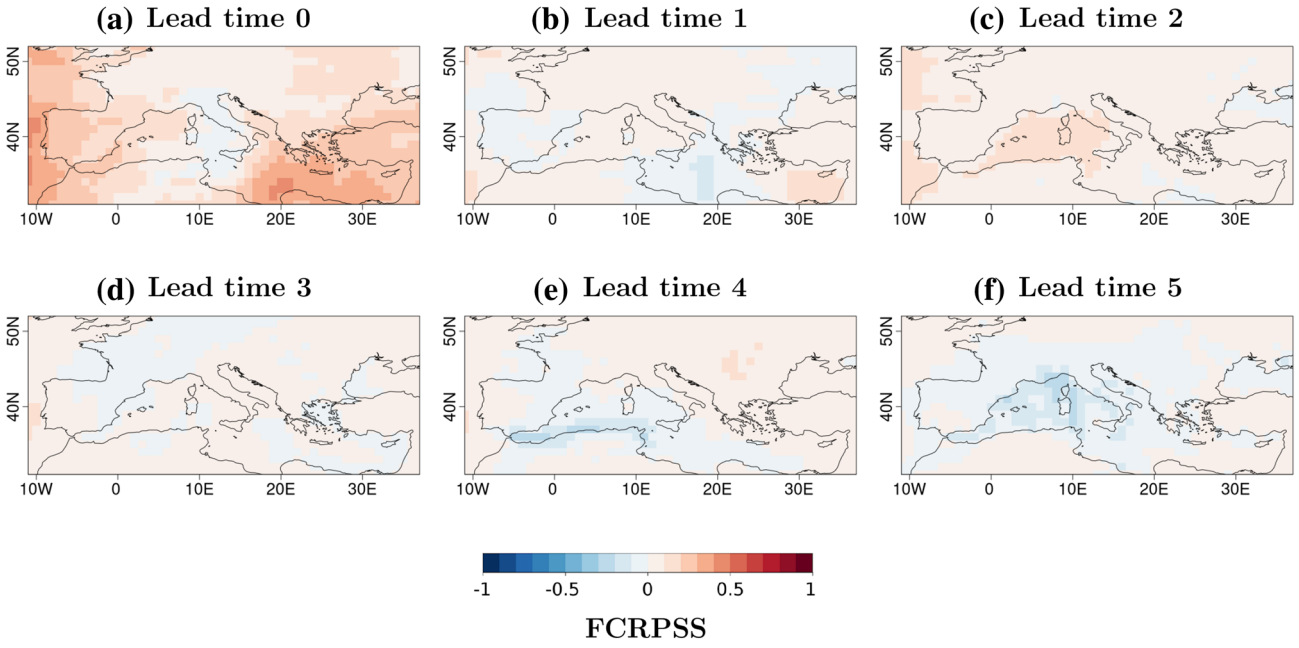

Fig. 13 Spatial pattern of the FCRPSS for the multi-model ensemble (MME) temperature anomalies forecasts for starting date November 1st and lead times 0–5

FCRPSS is generally positive up to lead time 2 in winter and close to zero with both positive and negative values in summer. In particular, the analysis of FCRPSS confirms some limitations of the forecast systems in predicting the correct distribution of summer precipitation anomalies, as already highlighted in Sect. 4.2 after the analysis of rank histograms.

The multi-model ensembles (MME and MMES) median FCRPSS show slightly better performances than the forecast systems in winter, with positive values up to lead time 4, and comparable performances in summer. The persistence forecast (PERS) show negative scores, decreasing in time, for both the median and the 75% percentile FCRPSS, indicating remarkably lower skill than individual seasonal forecast systems and MME over most (> 75%) of the domain. This analysis reveals that forecast systems outperform simple forecasts based on persistence at all lead times in terms of accuracy and sharpness.

5 Discussion

The present study provides an overall assessment of the skill of five state-of-the-art seasonal forecast systems at forecasting monthly temperature and precipitation anomalies over the Mediterranean region at different lead times. The main question which we addressed in this study is whether the most advanced seasonal forecast systems, or multi-model ensembles based on them, outperform elementary forecast approaches: (i) a climatological forecast (CTRL), which has been set as the benchmark; (ii) a persistence forecast (PERS) based on the persistence of ERA5 anomalies at lead times after the first month.

5.1 Overview on skill scores

Our results shows that seasonal forecast systems generally have higher skills in predicting temperature rather than precipitation anomalies, in agreement with previous studies (Sánchez-García et al. 2018). We find different correlation patterns depending on the season and variable. In winter (DJF) we find significant temperature anomaly correlations between the seasonal forecast systems and ERA5 over the Atlantic, Western Mediterranean, Iberian peninsula and Alps, which are areas typically influenced by the NAO (Hurrell 1995; Rodriguez-Puebla et al. 1998; Qian et al. 2000; Goodess and Jones 2002; Trigo et al. 2004; Terzago et al. 2013; López-Moreno and Vicente-Serrano 2008; Lopez-Bustins et al. 2008); in summer (JJA) significant temperature anomaly correlations are mainly over the Eastern Mediterranean, an area of intense land-atmosphere coupling where a reliable initialization of soil moisture can help improving summer air temperature forecasts (Ardilouze et al. 2017), and over the Iberian peninsula.

Precipitation anomalies are predicted with more limited skills than temperature anomalies: precipitation anomalies are significantly correlated in winter (DJF) at few gridpoints over the Iberian Peninsula, the Alps, Eastern Mediterranean and the Black Sea coasts; in summer (JJA) over the Iberian Peninsula and the Black Sea coasts. The highest correlation with the ERA5 reference is found for the multi-model mean (MME) that generally outperforms individual models. Precipitation forecasts based on PERS show no significant correlation over most of the domain, indicating that PERS is not a reliable method for precipitation forecasts.

Significant correlations of winter precipitation anomalies are found over the Alps and over the Iberian Peninsula, both areas affected by the NAO (Hurrell 1995; Rodriguez-Puebla et al. 1998; Qian et al. 2000; Goodess and Jones 2002; Trigo et al. 2004; Terzago et al. 2013; López-Moreno and Vicente-Serrano 2008; Lopez-Bustins et al. 2008) which is skillfully predicted by most seasonal forecast systems (Lledó et al. 2020). Over the Iberian peninsula temperature and precipitation anomalies are as well generally significantly correlated with ERA5, although summer precipitation is generally low and correlation scores based on very low precipitation values should be considered with caution. Overall, these results indicate that the climate of the Iberian Peninsula is more predictable than others in the Mediterranean area, probably owing to the influence of teleconnections like NAO and ENSO, that has been found to increase the predictability of dry events in spring/winter and hot events in summer (Frías et al. 2010).

This analysis highlighted specific features and limitations of individual forecast systems. For example, ECMWF for the winter season shows no significant correlation over most of the Mediterranean area, probably owing to limitations in reproducing the sea-surface temperatures in the Northwest Atlantic and the North Atlantic Oscillation (Johnson et al. 2019).

The ensemble spread of the seasonal forecast systems is generally appropriate, as shown by rank histograms. There are few exceptions: (i) the DWD forecast system, showing underdispersion at the beginning of the forecasting period (lead time 0) for all seasons and variables; (ii) summer precipitation ensemble forecasts are found to be underdispersive or to overestimate the observed anomalies. Since this signal is generalised for all forecast systems excluding MF, the issue could also be in the reference ERA5 data that may not be well suited to evaluate summer precipitation of forecast systems running at a coarser resolution (Rivoire et al. 2021).

Tercile-based forecasts from seasonal forecast systems show overall lower forecast errors and higher accuracy compared to the climatological forecast, as shown by the positive median BSS for almost any model, variable, season and lead time. When considering the MME we observe positive BSS in almost 100% of the gridpoints of the domain, so MME

performs even better than individual models, which show an improvement with respect to the reference forecast in 75–100% of the gridpoints.

Moving from the evaluation of the accuracy of tercile-based forecasts (through the BSS) to the evaluation of the accuracy of the forecast distribution (through the FCRPSS), in this second case we find that the improvement with respect to the climatological forecast is evident for lead time 0, limited up to lead time 2 and null for summer precipitation for selected models.

Combining the results for BSS and FCRPSS, the added value of the seasonal forecast systems compared to simpler methods is clear when predicting the probability of each tercile, while it is more limited when predicting the probabilities for the entire permissible range of values for the considered variables at lead times longer than 0.

When looking at the skill scores for the Mediterranean region (boxplots), they are generally stable from lead time 1 to 4. However some fluctuations from month to month can also be found: for example, winter temperature AUCCS at lead time 2 is higher than at lead time 1 for most models. This might be related to the monthly resolution of this analysis that amplifies the noise with respect to the signal. Please refer to Sect. 5.3 for a discussion of the impact of time averaging on the skill scores.

Another interesting feature is that tercile-based forecasts show better scores (higher accuracy, higher discrimination) for the upper and lower terciles than for the central tercile, see for example Fig. 11. This has been found also in other studies (Athanasiadis et al. 2017) and it has been related to the better performances of the models at forecasting large- rather than small-amplitude anomalies.

5.2 The persistence approach

Temperature anomaly forecasts based on PERS show high and significant correlation with observations, often outperforming individual models. On the contrary, precipitation anomaly forecasts based on PERS show no significant correlation over most of the domain. Tercile-based forecasts based on persistence have generally larger errors and lower accuracy than the climatological forecasts over most of the domain at long lead times for temperature, and at all lead times for precipitation. Despite the fact that PERS forecasts have been proven to have well balanced ensembles (flat rank histograms), the distribution of all PERS forecast values (for all starting dates, all lead times, all ensemble members and all gridpoints) has a different shape with respect to that of observations, with lower amplitude and higher tails (not shown). Overall, despite its simplicity the persistence forecast is of limited usefulness for climate services since it might be inaccurate and it does not provide a reliable measure of uncertainty.

5.3 Sensitivity of skill scores to temporal averaging

The analysis of anomaly correlation coefficients (ACCs) between seasonal forecast systems and ERA5 presented in this study has been performed on seasonal means, i.e. monthly temperatures and precipitation are averaged over

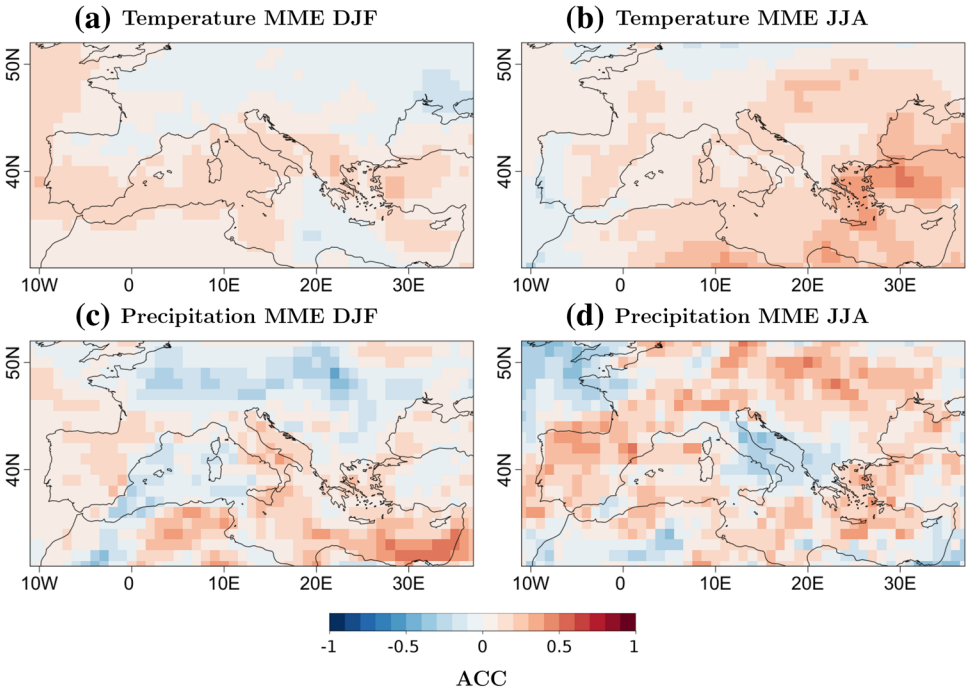

Fig. 14 Differences between the ACC calculated on seasonal anomalies and the ACC computed as the average of the three corresponding monthly ACCs for winter (**a**) and summer (**b**) temperature, winter (**c**) and summer (**d**) precipitation, for the multi-model ensemble

DJF and JJA and then transformed into seasonal anomalies. Since we are also interested in the forecast skills at the monthly scale (as we did for the other skill scores which we considered), we also evaluated monthly ACCs, then averaged them over the DJF and JJA seasons and finally compared them to the seasonal anomalies of Figs. 1, 2, 3 and 4. Figure 14 shows the differences between seasonal ACCs and seasonally-averaged monthly ACCs for MME temperature and precipitation forecasts in both seasons. The differences are generally small and patterns depend on the specific season and variable considered. In order to clarify the difference between seasonal ACCs and the corresponding seasonally-averaged monthly ACCs we summarize the statistics of the two scores over the Mediterranean domain by means of boxplots. Moreover we extend the analysis to the other skill scores considered in the study: BSS, AUCSS, FCRPSS. In Fig. 15 for each skill score and each forecast system we compare (i) the scores obtained from seasonally-averaging monthly scores, to (ii) seasonal scores, for temperature. The width of the boxplot shows the spatial variability of the score over the Mediterranean domain. In general, seasonal scores have slightly higher median (in the case of ACC) or comparable median and larger spread compared to seasonally-averaged monthly scores. These considerations generally apply to any forecast system and season (with few exceptions) and are valid also for precipitation (not shown). These results only partially agree with Buizza and Leutbecher (2015), who showed that the forecast skill horizon (i.e the lead time when the ensemble forecast ceases to be more skillful than the climatological distribution) of medium range/monthly ensemble forecasts is considerably longer for time- and spatially-averaged fields than for grid-point, instantaneous fields. Their analysis concluded that time- and space-averaging are a basic way to filter out the less predictable components of the climate system and to focus on the more predictable components. In our analysis a relative improvement related to time averaging is found only for ACC, while for the probabilistic scores the difference is small, at least in average, over the Mediterranean region. So, we do not find a clear advantage of using seasonal anomalies rather than seasonally-averaged monthly anomalies. Our results seems to leave some room for the exploitation of seasonal predictions at the monthly time scale without systematically losing forecast skills.

5.4 Sensitivity of skill scores to trend removal

All our analyses are performed on time-detrended temperature and precipitation anomalies. In a similar paper by Mishra et al. (2019) seasonal anomalies were not detrended before calculating ACC. Limited to UKMO temperature ACC we can compare the results of the two studies. We observe a different behaviour depending on the season: in summer we obtain similar temperature correlation patterns but with a notable reduction of the areas with significant correlation; in winter we obtain remarkably different patterns with correlations of the opposite sign for example over the Alps. It is important to note that in the previous study the reference data were ERA-Interim instead of ERA5: part of the difference we find might be explained by the different reference data used, although the two reanalyses are expected to have similar behaviours. In any case, discrepancies among the results of these studies suggest the importance of removing the trend from the original data in order to avoid overestimation of the anomaly correlation, as also suggested by Sánchez-García et al. (2018). In particular that work emphasised the importance of detrending anomalies specifically when considering long timeseries of hindcasts; in the detailed results for different regions, they found the trend removal to be either responsible for a decrease in the overall skill or completely ineffective. We analysed the effects of detrending on the ACC of temperature and precipitation forecasts. We evaluated the difference between ACC calculated on (i) the original seasonal forecast data (without detrending) and (ii) the residuals (detrended data). Results (not shown) indicate that at lead time 1, the difference is generally very small, meaning that ACCs of original and detrended data are comparable. At longer lead times the difference becomes positive over most of the domain, and higher ACC is obtained when data are not detrended. The trend in the original data artificially increases the correlation between two variables, masking the real correlation. The trend contributes to improving the overall skill of the model, especially at longer lead times, when the skill of the model is lower, and especially for temperature. Precipitation, instead, is less affected by the trend removal being its trend actually smaller than for temperature.

5.5 MME and sensitivity to the ensemble size

The choice of the detailed scheme for constructing the Multi Model Ensemble (MME) and the Multi Model Ensemble Small (MMES) has been affected by the scientific community's debate around this topic. In their review paper Tebaldi and Knutti (2007) state that seasonal forecasts have better skill, higher reliability and consistency when several independent models are combined. The most significant benefit is seen in the consistently better performance of the multi-model when considering all aspects of the predictions (Hagedorn et al. 2005). Different authors have compared equally and unequally weighting schemes (Hemri et al. 2020; Vigaud et al. 2019; Mishra et al. 2019; Weisheimer et al. 2009; Pavan and Doblas-Reyes 2000), never reaching strong conclusions. Due to this ongoing discussion spanning

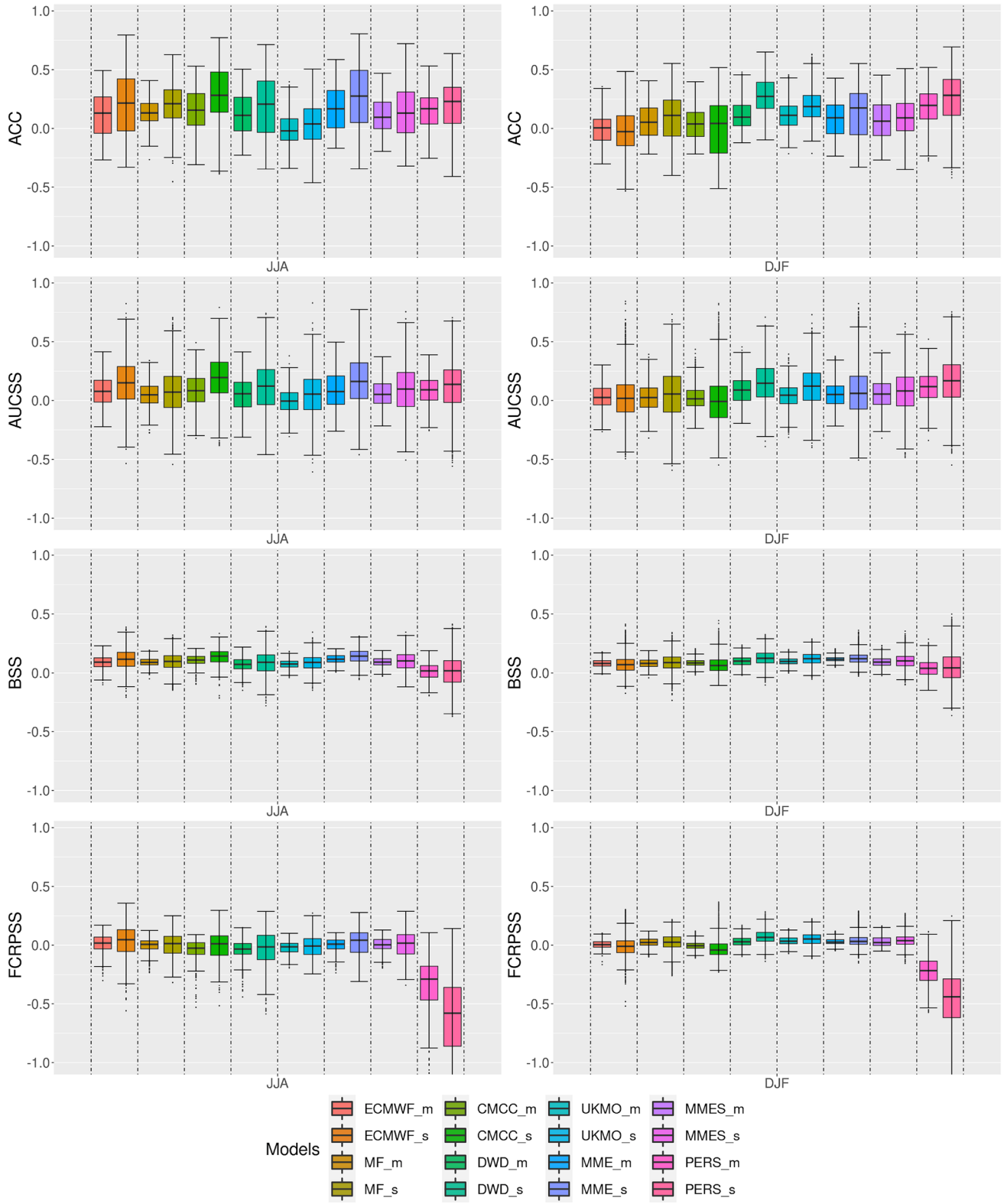

Fig. 15 Seasonally-averaged monthly skill scores (_m) compared to seasonal scores (_s) for each forecast system, for summer (left column) and winter (right column) temperature

the last 20 years and more, in this study the MME and the MMES have been built with an equally weighting scheme. In order to consider the different ensemble sizes of the models, the MMES has been built using a subset of 5 ensemble members randomly chosen for each model, leading to an equally weighting scheme (ensemble size wise). The advantage of this approach is twofold: first, it allows to obtain an ensemble with comparable size as the other forecast systems; second, it enables the estimation of how much the ensemble size impacts on the skill scores retrievals through the comparison of MME to MMES. The latter tries to deal with the limitations imposed by computational power, counting 25 ensemble members instead of 148 and allowing the estimation of the skill with a smaller ensemble.

The MME well summarizes correlation patterns common to several individual models, often providing higher correlations than individual models, especially in the case of precipitation. According to our results, MMES, which is assembled with 25 ensemble members randomly chosen in the number of 5 from each model, shows similar temperature and precipitation anomaly correlation coefficient patterns as MME (including all 148 members from the 5 different models). Winter precipitation ACC pattern is slightly more widespread in MMES than in MME, with more significant values over the Eastern and Northern part of the domain. On the contrary, MMES shows slightly lower ACC values for summer temperature and a slightly less widespread pattern for winter temperature, compared to MME.

Forecast errors of the seasonal forecast systems considered in this study are generally lower than that of the climatological forecast. When considering the MME we observe better performances with respect to the climatological forecast in almost 100% of the gridpoints of the domain, so MME shows a clear added value in almost any point of the domain and for both temperature and precipitation. MMES shows comparable median score and larger spread compared to MME. MME and MMES rank histograms are similar to each other for both variables. Comparable values are also found for BSS and FCRPSS, which accounts for differences in the ensemble size. All these features suggest that forecasts based on MME or on a subset of each model's ensemble members (MMES) provides overall similar performances, and the use of MMES would reduce the computational time needed for analysis and the data storage requirements, keeping many of the advantages of the MME.

5.6 Seamless predictions: sub-seasonal to seasonal forecasts

There is a general tendency toward a seamless prediction approach that would join different timescales, from meteorological to subseasonal to seasonal. An example of such an approach is discussed in Dirmeyer and Ford (2020) in which a weighting scheme for the transition from forecast at different timescales is proposed. This "weather-climate prediction gap" (Mariotti et al. 2018) lies at the lower edge of the seasonal forecast timescale and the upper edge of the meteorological one. The approach chosen in this article allows us to make available to the scientific community a monthly analysis that can be more easily compared to subseasonal forecasts (from 2 weeks to 2 months) helping in filling this time gap. Concurrently, the monthly analysis allows looking with a finer detail at the differences in response among the forecast systems, setting the basis for narrowing the effort in searching for different sources of predictability by identifying the time scale at which seasonal forecasts become drastically less skilful (Board et al. 2016).

6 Conclusions

This study provides an overview of the skill of 5 state-of-the-art seasonal forecast systems (ECMWF, MF, UKMO, CMCC, DWD) and the corresponding Multi-Model Ensembles (MME and MMES) at forecasting monthly temperature and precipitation anomalies over the Mediterranean region, focusing on the winter and summer seasons. All our analyses are performed on detrended temperature and precipitation anomalies. The work has been carried out in the frame of the ERA4CS MEDSCOPE project (https://www.medscope-project.eu/) and it is motivated by the increasing interest in the use of seasonal forecasts for developing climate services, and the consequent need to assess added value and limitations of these products over the specific domain of interest. We assess the improvement of each forecast system compared to a very simple forecast based on the climatology, which is set as a benchmark, and we used a multi-score approach to evaluate different features of the probabilistic forecast in relation to the lead time.

Temperature anomalies are found to be more predictable than precipitation anomalies. Anomaly correlation patterns vary across different forecast systems however some frequently occurring features can be highlighted. Temperature anomaly correlations are significant over large areas mainly over the Western Mediterranean in winter and over the Eastern Mediterranean in summer. Precipitation correlations are lower and patchier, although significant over some regions: in winter at few grid points over the Iberian Peninsula, the Alps and Eastern Mediterranean; in summer over the Iberian Peninsula and the Black Sea coasts.

Individual forecast systems are found to outperform the reference forecast based on climatology in the following features:

- higher accuracy (lower forecast errors) for tercile-based forecasts for any variable, season and lead time, in 75% up to 100% of the gridpoints of the domain;
- higher discrimination up to lead time 2 months in most (> 50%) of the domain and on average over the three terciles. However, the discrimination is higher for the lower/upper terciles than for the middle tercile.

Since forecast skill varies in space and time across different models, for climate services applications we recommend the use of an ensemble of models, together with the MME forecast. In fact, MME summarizes the common features of individual forecast systems, often outperforming single models in terms of anomaly correlations with respect to the ERA5 reference. MME generally provides the best anomaly correlation with observed precipitation anomalies.

A simple forecast methods based on persistence has been found to outperform seasonal forecast systems in terms of anomaly correlation with temperature observations. However, the persistence method shows no skill in forecasting precipitation anomalies. Moreover the persistence ensemble forecast has lower accuracy and lower sharpness than the reference forecast and individual models.

Overall, despite their limitations, seasonal forecast systems show an added value with respect to simple forecast methods based on the climatology or persistence, although the added value is not uniform over the Mediterranean area.

The present evaluation of the climate predictability on seasonal time scales over the Mediterranean can set the basis for the development of applications and tools for climate services dedicated to various end-users in the water management, agriculture, and energy production, enhancing the awareness of strengths and limitations of individual seasonal forecast system outputs.

Further steps beyond this analysis should go towards an assessment of the sources of predictability responsible for the skill score patterns: in particular it would be interesting to explore the predictability in the Mediterranean area conditioned on NAO+/NAO- and El Niño/ La Niña conditions.

A limitation of this and similar studies is the length of the hindcasts period: the availability of a larger set of forecasts would allow for more robust performance evaluations. This issue should be taken into account when planning hindcast simulations.

Concerning the idea of the persistence forecast, many methods for kernel dressing at a different level of complexity are available, and a more thorough analysis should be carried out since the use of different approaches could affect the skill of the resulting model. Regarding applications, an economic value assessment could be carried out to better focus the attention on different climate service sectors. However such evaluations are left for a separate in-depth analysis.

Acknowledgements We would like to thank two anonymous reviewers for their suggestions.

Author Contributions Conceptualization: all authors; Methodology: all authors; Formal analysis and investigation: FCQ; Writing—original draft preparation: ST with contribution of FCQ; Writing—review and editing: ST and JvH; Funding acquisition: JvH; Supervision: JvH and ST.

Funding This work was performed in the framework of the MEDSCOPE (MEDiterranean Services Chain based On climate PrEdictions) ERA4CS project (Grant Agreement No. 690462) funded by the European Union.

Availability of data and material. Model datasets are available upon request in the framework the MEDSCOPE project (www.medscope-project.eu/).

Code availability Code written for the analysis is available upon request to the corresponding author.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Anderson JL (1996) A method for producing and evaluating probabilistic forecasts from ensemble model integrations. J Clim 9(7):1518–1530. https://doi.org/10.2307/26201352

Ardilouze C, Batté L, Bunzel F, Decremer D, Déqué M, Doblas-Reyes FJ, Douville H, Fereday D, Guemas V, MacLachlan C, Müller W, Prodhomme C (2017) Multi-model assessment of the impact of soil moisture initialization on mid-latitude summer predictability. Clim Dyn 49(11–12):3959–3974. https://doi.org/10.1007/s00382-017-3555-7

Athanasiadis PJ, Bellucci A, Scaife AA, Hermanson L, Materia S, Sanna A, Borrelli A, MacLachlan C, Gualdi S (2017) A multisystem view of wintertime NAO seasonal predictions. J Clim 30(4):1461–1475. https://doi.org/10.1175/JCLI-D-16-0153.1

Bhend J, Ripoldi J, Mignani C, Mahlstein I, Hiller R, Spirig C, Liniger M, Weigel A, Jimenez JB, De Felice M, Siegert S, Sedlmeier K (2017) easyVerification

Board OS, of Sciences Engineering NA, Medicine et al (2016) Next generation earth system prediction. National Academies Press, Washington. https://doi.org/10.17226/21873

Bradley AA, Demargne J, Franz KJ (2019) Attributes of forecast quality. In: Duan Q, Pappenberger F, Wood A, Cloke HL, Schaake JC (eds) Handbook of hydrometeorological ensemble forecasting. Springer, Berlin, pp 849–892. https://doi.org/10.1007/978-3-642-39925-1_2

Bradley AA, Schwartz SS (2011) Summary verification measures and their interpretation for ensemble forecasts. Mon Weather Rev 139(9):3075–3089. https://doi.org/10.1175/2010MWR3305.1

Brier GW (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78(1):1–3. https://doi.org/10.1175/1520-0493(1950)078<3c0001:VOFEIT>3e2.0.CO;2

Bruno Soares M, Daly M, Dessai S (2018) Assessing the value of seasonal climate forecasts for decision-making. Wiley Interdiscip Rev Clim Change. https://doi.org/10.1002/wcc.523

Buizza R, Leutbecher M (2015) The forecast skill horizon. Q J R Meteorol Soc 141(693):3366–3382. https://doi.org/10.1002/qj.2619

Clark RT, Bett PE, Thornton HE, Scaife AA (2017) Skilful seasonal predictions for the European energy industry. Environ Res Lett. https://doi.org/10.1088/1748-9326/aa57ab

Cramer W, Guiot J, Marini K (eds) (2020) MedECC: climate and environmental change in the Mediterranean basin—current situation and risks for the future. First Assessment Report, Union for the Mediterranean, Plan Bleu, UNEP/MAP, France

Dirmeyer PA, Ford TW (2020) A technique for seamless forecast construction and validation from weather to monthly time scales. Mon Weather Rev 148(9):3589–3603. https://doi.org/10.1175/MWR-D-19-0076.1

Doblas-Reyes FJ, Pavan V, Stephenson DB (2003) The skill of multi-model seasonal forecasts of the wintertime North Atlantic Oscillation. Clim Dyn 21(5–6):501–514. https://doi.org/10.1007/s00382-003-0350-4

Doblas-Reyes FJ, García-Serrano J, Lienert F, Biescas AP, Rodrigues LR (2013) Seasonal climate predictability and forecasting: status and prospects. Wiley Interdiscip Rev Clim Change 4(4):245–268. https://doi.org/10.1002/wcc.217

Dorel L, Ardilouze C, Déqué M, Batté L, Guérémy JF (2017) Documentation of the METEO-FRANCE Pre-Operational seasonal forecasting system. Tech. rep., Copernicus Climate Change Service

Dunstone N, Smith D, Scaife A, Hermanson L, Eade R, Robinson N, Andrews M, Knight J (2016) Skilful predictions of the winter North Atlantic Oscillation one year ahead. Nat Geosci 9(11):809–814. https://doi.org/10.1038/ngeo2824

Ferro CAT (2014) Fair scores for ensemble forecasts. Q J R Meteorol Soc 140(683):1917–1923. https://doi.org/10.1002/qj.2270

Ferro CAT, Richardson DS, Weigel AP (2008) On the effect of ensemble size on the discrete and continuous ranked probability scores. Meteorol Appl 15(1):19–24. https://doi.org/10.1002/met.45

Frías MD, Herrera S, Cofiño AS, Gutiérrez JM (2010) Assessing the skill of precipitation and temperature seasonal forecasts in Spain: windows of opportunity related to ENSO events. J Clim 23(2):209–220. https://doi.org/10.1175/2009JCLI2824.1

Fröhlich K, Dobrynin M, Isensee K, Gessner C, Paxian A, Pohlmann H, Haak H, Brune S, Früh B, Baehr J (2020) The German climate forecast system: GCFS. J Adv Model Earth Syst. https://doi.org/10.1002/essoar.10502582.2

Goodess CM, Jones PD (2002) Links between circulation and changes in the characteristics of Iberian rainfall. Int J Climatol 22(13):1593–1615. https://doi.org/10.1002/joc.810

Graça A (2019) The MED-GOLD project: advanced user-centric climate services for higher resilience and profitability in the grape and wine sector. BIO Web Conf 12:01005. https://doi.org/10.1051/bioconf/20191201005

Guemas V, Blanchard-Wrigglesworth E, Chevallier M, Day JJ, Déqué M, Doblas-Reyes FJ, Fučkar NS, Germe A, Hawkins E, Keeley S, Koenigk T, Salas y Mélia D, Tietsche S (2016) A review on Arctic sea-ice predictability and prediction on seasonal to decadal time-scales. Q J R Meteorol Soc 142(695):546–561. https://doi.org/10.1002/qj.2401

Hagedorn R, Doblas-Reyes FJ, Palmer T (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. Tellus A 57(3):219–233. https://doi.org/10.1111/j.1600-0870.2005.00103.x

Hamill TM (2002) Interpretation of rank histograms for verifying ensemble forecasts. Mon Weather Rev 129(3):550–560. https://doi.org/10.1175/1520-0493(2001)129<3c0550:iorhfv>3e2.0.co;2

Hamill TM, Colucci SJ (1997) Verification of Eta-RSM short-range ensemble forecasts. Mon Weather Rev 125(6):1312–1327. https://doi.org/10.1175/1520-0493(1997)125<3c1312:VOERSR>3e2.0.CO;2

Hemri S, Bhend J, Liniger MA, Manzanas R, Siegert S, Stephenson DB, Gutiérrez JM, Brookshaw A, Doblas-Reyes FJ (2020) How to create an operational multi-model of seasonal forecasts? Clim Dyn 55(5–6):1141–1157. https://doi.org/10.1007/s00382-020-05314-2

Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast 15(5):559–570. https://doi.org/10.1175/1520-0434(2000)015<3c0559:DOTCRP>3e2.0.CO;2

Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, Simmons A, Soci C, Abdalla S, Abellan X, Balsamo G, Bechtold P, Biavati G, Bidlot J, Bonavita M, De Chiara G, Dahlgren P, Dee D, Diamantakis M, Dragani R, Flemming J, Forbes R, Fuentes M, Geer A, Haimberger L, Healy S, Hogan RJ, Hólm E, Janisková M, Keeley S, Laloyaux P, Lopez P, Lupu C, Radnoti G, de Rosnay P, Rozum I, Vamborg F, Villaume S, Thépaut JN (2020) The ERA5 global reanalysis. Q J R Meteorol Soc 146(730):1999–2049. https://doi.org/10.1002/qj.3803

Hewitt C, Buontempo C, Newton P (2013) Using climate predictions to better serve society's needs. Eos Trans Am Geophys Union 94(11):105–107. https://doi.org/10.1002/2013EO110002

Hurrell JW (1995) Decadal trends in the North Atlantic oscillation: regional temperatures and precipitation. Science 269(5224):676–679. https://doi.org/10.1126/science.269.5224.676

Johnson SJ, Stockdale TN, Ferranti L, Balmaseda MA, Molteni F, Magnusson L, Tietsche S, Decremer D, Weisheimer A, Balsamo G, Keeley SP, Mogensen K, Zuo H, Monge-Sanz BM (2019) SEAS5: the new ECMWF seasonal forecast system. Geosci Model Dev 12(3):1087–1117. https://doi.org/10.5194/gmd-12-1087-2019

Jolliffe IT, Stephenson DB (eds) (2011) Forecast verification: a practitioner's guide in atmospheric science. Wiley, New York

Lledó L, Cionni I, Torralba V, Bretonniere PA, Samsó M (2020) Seasonal prediction of Euro-Atlantic teleconnections from multiple systems. Environ Res Lett. https://doi.org/10.1088/1748-9326/ab87d2

Lopez-Bustins JA, Martin-Vide J, Sanchez-Lorenzo A (2008) Iberia winter rainfall trends based upon changes in teleconnection and circulation patterns. Glob Planet Change 63(2–3):171–176. https://doi.org/10.1016/j.gloplacha.2007.09.002

López-Moreno JI, Vicente-Serrano SM (2008) Positive and negative phases of the wintertime North Atlantic oscillation and drought occurrence over Europe: a multitemporal-scale approach. J Clim 21(6):1220–1243. https://doi.org/10.1175/2007JCLI1739.1

Lowe R, Stewart-Ibarra AM, Petrova D, García-Díez M, Borbor-Cordova MJ, Mejía R, Regato M, Rodó X (2017) Climate services for health: predicting the evolution of the 2016 dengue season in Machala, Ecuador. Lancet Planet Health 1(4):e142–e151. https://doi.org/10.1016/S2542-5196(17)30064-5

Maclachlan C, Arribas A, Peterson KA, Maidens A, Fereday D, Scaife AA, Gordon M, Vellinga M, Williams A, Comer RE, Camp J, Xavier P, Madec G (2015) Global Seasonal forecast system

version 5 (GloSea5): a high-resolution seasonal forecast system. Q J R Meteorol Soc 141(689):1072–1084. https://doi.org/10.1002/qj.2396

Manubens N, Caron LPP, Hunter A, Bellprat O, Exarchou E, Fučkar NS, Garcia-Serrano J, Massonnet F, Ménégoz M, Sicardi V, Batté L, Prodhomme C, Torralba V, Cortesi N, Mula-Valls O, Serradell K, Guemas V, Doblas-Reyes FJ (2018) An R package for climate forecast verification. Environ Model Softw 103:29–42. https://doi.org/10.1016/j.envsoft.2018.01.018

Mariotti A, Ruti PM, Rixen M (2018) Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. npj Clim Atmos Sci 1(1):4. https://doi.org/10.1038/s41612-018-0014-z

Mason SJ (2004) On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. Mon Weather Rev 132(7):1891–1895. https://doi.org/10.1175/1520-0493(2004)132<1891:OUCAAR>2.0.CO;2

Mishra N, Prodhomme C, Guemas V (2019) Multi-model skill assessment of seasonal temperature and precipitation forecasts over Europe. Clim Dyn 52(7–8):4207–4225. https://doi.org/10.1007/s00382-018-4404-z

Morss RE, Lazo JK, Brown BG, Brooks HE, Ganderton PT, Mills BN (2008) Societal and economic research and applications for weather forecasts: priorities for the North American THORPEX Program. Bull Am Meteorol Soc 89(3):335–346. https://doi.org/10.1175/BAMS-89-3-335

Murphy AH (1993) What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather Forecast 8(2):281–293. https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2

National Academies of Sciences Engineering and Medicine (2010) Assessment of intraseasonal to interannual climate prediction and predictability. National Academies Press, Washington, D.C. https://doi.org/10.17226/12878

NCAR–Research Applications Laboratory (2015) Verification. https://cran.r-project.org/web/packages/verification/

Palin EJ, Scaife AA, Wallace E, Pope ECD, Arribas A, Brookshaw A (2016) Skillful seasonal forecasts of winter disruption to the U.K. transport system. J Appl Meteorol Climatol 55(2):325–344. https://doi.org/10.1175/JAMC-D-15-0102.1

Palmer TN, Alessandri A, Andersen U, Cantelaube P, Davey M, Délécluse P, Déqué M, Díez E, Doblas-Reyes FJ, Feddersen H, Graham R, Gualdi S, Guérémy JF, Hagedorn R, Hoshen M, Keenlyside N, Latif M, Lazar A, Maisonnave E, Marletto V, Morse AP, Orfila B, Rogel P, Terres JM, Thomson MC (2004) Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). Bull Am Meteorol Soc 85(6):853–872. https://doi.org/10.1175/BAMS-85-6-853

Pavan V, Doblas-Reyes FJ (2000) Multi-model seasonal hindcasts over the Euro-Atlantic: skill scores and dynamic features. Clim Dyn 16(8):611–625. https://doi.org/10.1007/s003820000063

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

Qian B, Corte-Real J, Xu H (2000) Is the North Atlantic Oscillation the most important atmospheric pattern for precipitation in Europe? J Geophys Res Atmos 105(D9):11901–11910. https://doi.org/10.1029/2000JD900102

R Core Team (2019) R: a language and environment for statistical computing

Rivoire P, Martius O, Naveau P (2021) A comparison of moderate and extreme ERA-5 daily precipitation with two observational data sets. Earth Space Sci. https://doi.org/10.1002/essoar.10505726.1

Rodriguez-Puebla C, Encinas AH, Nieto S, Garmendia J (1998) Spatial and temporal patterns of annual precipitation variability over the Iberian Peninsula. Int J Climatol 18(3):299–316. https://doi.org/10.1002/(SICI)1097-0088(19980315)18:3<299::AID-JOC247>3.0.CO;2-L

Sánchez-García E, Voces Aboy J, Rodríguez Camino E (2018) Verification of six operational seasonal forecast systems over Europe and North Africa. Tech. rep., AEMET

Sanna A (2017) RP0285-CMCC-SPS3: the CMCC Seasonal Prediction System 3. Tech. rep., Fondazione CMCC

Scaife AA, Arribas A, Blockley E, Brookshaw A, Clark RT, Dunstone N, Eade R, Fereday D, Folland CK, Gordon M, Hermanson L, Knight JR, Lea DJ, MacLachlan C, Maidens A, Martin M, Peterson AK, Smith D, Vellinga M, Wallace E, Waters J, Williams A (2014) Skillful long-range prediction of European and North American winters. Geophys Res Lett 41(7):2514–2519. https://doi.org/10.1002/2014GL059637

Schulzweida U (2019) Climate data operator (CDO). https://doi.org/10.5281/zenodo.3991594

Siegert S (2020) SpecsVerification: forecast verification routines for ensemble forecasts of weather and climate

Smith LA, Du H, Suckling EB, Niehörster F (2015) Probabilistic skill in ensemble seasonal forecasts. Q J R Meteorol Soc 141(689):1085–1100. https://doi.org/10.1002/qj.2403

Stockdale T (2012) The EUROSIP system. Tech. rep., ECMWF. http://www.ecmwf.int/sites/default/files/elibrary/2013/12429-eurosip-system-multi-model-approach.pdf

Svensson C, Brookshaw A, Scaife AA, Bell VA, Mackay JD, Jackson CR, Hannaford J, Davies HN, Arribas A, Stanley S (2015) Long-range forecasts of UK winter hydrology. Environ Res Lett. https://doi.org/10.1088/1748-9326/10/6/064006

Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos Trans R Soc A Math Phys Eng Sci. https://doi.org/10.1098/rsta.2007.2076

Terzago S, Fratianni S, Cremonini R (2013) Winter precipitation in Western Italian Alps (1926–2010). Meteorol Atmos Phys 119(3–4):125–136. https://doi.org/10.1007/s00703-012-0231-7

Torralba V, Doblas-Reyes FJ, MacLeod D, Christel I, Davis M (2017) Seasonal climate prediction: a new source of information for the management of wind energy resources. J Appl Meteorol Climatol 56(5):1231–1247. https://doi.org/10.1175/JAMC-D-16-0204.1

Trigo RM, Pozo-Vázquez D, Osborn TJ, Castro-Díez Y, Gámiz-Fortis S, Esteban-Parra MJ (2004) North Atlantic oscillation influence on precipitation, river flow and water resources in the Iberian Peninsula. Int J Climatol 24(8):925–944. https://doi.org/10.1002/joc.1048

Troccoli A, Harrison M, Anderson DLT, Mason SJ (2008) Seasonal climate: forecasting and managing risk. In: Nato science series: IV: earth and environmental sciences, vol 82. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-6992-5

van der Linden P, Mitchell J (2009) ENSEMBLES: climate change and its impacts: summary of research and results from the ENSEMBLES project. Tech. rep., Met Office Hadley Centre, Exeter

Vannitsem S, Wilks DS, Messner JW (2018) Statistical postprocessing of ensemble forecasts. Elsevier. https://doi.org/10.1016/C2016-0-03244-8

Vigaud N, Tippett MK, Yuan J, Robertson AW, Acharya N (2019) Probabilistic skill of subseasonal surface temperature forecasts over North America. Weather Forecast 34(6):1789–1806. https://doi.org/10.1175/WAF-D-19-0117.1

Weisheimer A, Doblas-Reyes FJ, Palmer TN, Alessandri A, Arribas A, Déqué M, Keenlyside N, MacVean M, Navarra A, Rogel P (2009) ENSEMBLES: a new multi-model ensemble for seasonal-to-annual predictions—skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. Geophys Res Lett. https://doi.org/10.1029/2009GL040896

Wilks DS (2011) Statistical methods in the atmospheric sciences. Academic Press Inc, London

WMO (2018) Guidance on verification of operational seasonal climate forecasts. Tech. Rep. WMO-1220, World Meteorological Organization (WMO)