Summer drought predictability in the Euro-Mediterranean region in seasonal forecasts

Summer Drought Predictability in the Euro-Mediterranean Region in Seasonal Forecasts

GIADA CERATO,[a] KATINKA BELLOMO,[b] AND JOST VON HARDENBERG[a,c]

[a] *Politecnico di Torino, Department of Environment, Land, and Infrastructure Engineering, Turin, Italy*
[b] *Department of Geosciences, University of Padova, Padova, Italy*
[c] *National Research Council, Institute of Atmospheric Sciences and Climate, Turin, Italy*

(Manuscript received 3 March 2025, in final form 3 October 2025, accepted 29 October 2025)

ABSTRACT: Seasonal drought forecasts are essential for risk management in climate-sensitive sectors, yet their performance across Europe remains uncertain. This study evaluates the ability of state-of-the-art seasonal forecast systems from the Copernicus Climate Change Service (C3S) to predict summer drought conditions in Europe using the June–August standardized precipitation evapotranspiration index (SPEI-3), which shows more spatially coherent and higher forecast skill across the region than the standardized precipitation index (SPI). Leveraging this superior performance, we adopt SPEI-3 as the reference drought indicator. We then implement a systematic multimetric evaluation framework to benchmark individual systems, their multimodel ensemble (MME), and to identify patterns of predictive skill across regions and lead times. Our findings reveal that when SPEI forecasts are initialized at the onset of the summer season, all models exhibit on average good quality in terms of correlation, accuracy, reliability, and discrimination skills, though with local variability. The performance is better for all models in southern Europe, indicating higher predictability of SPEI in that region compared to northern Europe, where summer drought variability is mainly driven by precipitation, which is inherently less predictable than temperature. Results show that, when a general analysis at the regional scale is needed, the MME offers the most robust solution, demonstrating more widespread significant skill compared to single models, up to a 1-month lead time. These findings highlight the value of SPEI-based ensemble forecasts for early summer drought detection and provide actionable insights into where and when seasonal predictions offer the greatest utility across Europe.

KEYWORDS: Drought; Forecast verification/skill; Seasonal forecasting; Predictability

1. Introduction

Meteorological drought is a complex, naturally occurring phenomenon, typically characterized by prolonged periods of lack of precipitation (Keyantash and Dracup 2002), often exacerbated by higher-than-normal temperatures (Vicente-Serrano et al. 2014). It frequently acts as the initial trigger for a chain of processes that can propagate into agricultural and hydrological droughts. This progression reflects a gradual depletion of water across different components of the hydrological cycle, beginning with reduced soil moisture that impairs plant growth (agricultural drought) and extending to declines in groundwater and streamflow levels, which characterize hydrological drought (Van Loon and Van Lanen 2012; Van Loon 2015). The Mediterranean region, recognized as a drought hotspot, has experienced an increase in the frequency and severity of droughts driven by climate change (IPCC 2021; Spinoni et al. 2017). In contrast, drought trends in northern Europe appear more ambiguous, with some studies suggesting a moderate reduction, while others report no clear patterns (Tuel and Eltahir 2020; Spinoni et al. 2017, 2019; Ionita and Nagavciuc 2021).

Summer droughts in Europe develop through distinct mechanisms shaped by regional climatic regimes. In the southern Mediterranean, precipitation deficits typically accumulate during winter and spring, setting the stage for hydrological stress that is later intensified by strong evaporative demand during summer (Teuling et al. 2013; Seneviratne et al. 2012). In contrast, in more temperate northern areas, summer precipitation plays a central role in the seasonal water balance, and shortfalls can rapidly lead to meteorological drought (IPCC 2021; Schumacher et al. 2024). Recent analyses confirm that meteorological droughts in southern Europe are most likely to occur during the summer months in future scenarios, whereas agricultural droughts tend to peak in autumn, reflecting the cumulative impact of prolonged summer dryness. These findings (Essa et al. 2023) underscore the need to assess summer drought forecasting skill, particularly in a region facing lengthening summers and increasing evapotranspiration.

Reliable seasonal forecasts can thus provide valuable information not only about the potential for ongoing drought to persist or worsen but also about the likelihood of drought emergence. This is especially relevant for water-sensitive sectors such as agriculture and urban supply, which are strongly affected by seasonal climate variability (Pozzi et al. 2013; Portele et al. 2021; Sánchez-García et al. 2022; Terzago et al. 2023; Zellou et al. 2023). Therefore, recent advancements in monthly to seasonal forecasts have sparked growing interest in developing methods for drought monitoring across Europe (Dutra et al. 2013; Turco et al. 2017; Tramblay et al. 2020). However, seasonal forecasts in the midlatitudes face unique

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-25-0039.s1.

Corresponding author: Giada Cerato, giada.cerato@polito.it

DOI: 10.1175/JHM-D-25-0039.1

challenges because strong atmospheric variability introduces unavoidable uncertainty, limiting seasonal climate predictability (Vitart 2004; Doblas-Reyes et al. 2013; Weisheimer and Palmer 2014).

The use of seasonal forecasts for predicting the occurrence of a drought has been previously analyzed across different regions of the world, usually focusing on drought indices. The standardized precipitation index (SPI; McKee et al. 1993) has widely been used due to its simplicity and reliance solely on precipitation data (Dutra et al. 2013, 2014; Ma et al. 2015; Lavaysse et al. 2015, 2020). Lavaysse et al. (2015) assessed the performance of ECMWF's extended-range and seasonal forecasting systems in predicting meteorological droughts using the SPI for a 1-month accumulation period, demonstrating that up to 40% of events can be correctly anticipated 1 month in advance. Dutra et al. (2013, 2014) showed that the skill of SPI forecasts generally declines at longer lead times, particularly for shorter accumulation periods, with meaningful predictability often limited to the first 1–3 months depending on SPI time scale and region. More recently, Lavaysse et al. (2020) demonstrated that global SPI-based early warning systems using ensemble thresholds can provide skillful forecasts of unusually dry conditions up to 6 months ahead. Numerous studies have explored the use of the SPI within statistical and hybrid drought prediction frameworks, including approaches that integrate dynamical seasonal forecasts with observational products to improve skill at extended lead times (e.g., Dutra et al. 2013; Yuan et al. 2013; Dutra et al. 2014; Lavaysse et al. 2015; Zellou et al. 2023). Torres-Vázquez et al. (2024) proposed a high-resolution probabilistic forecasting system on observational resampling of SPI, which allowed for skillful drought predictions several months in advance over Spain, highlighting the potential of empirical, persistence-based methods as cost-effective alternatives to dynamical systems in certain regions. Last, precipitation-based approaches have also proven effective in operational contexts. Arnone et al. (2020), for instance, developed a methodology that uses SEAS5 seasonal forecasts of cumulative precipitation to estimate drought probability and associated forecast reliability, with successful application in some Mediterranean regions. However, the overall predictability of SPI-based forecasts remains limited at seasonal time scales (Hao et al. 2018).

The standardized precipitation evapotranspiration index (SPEI; Vicente-Serrano et al. 2010) considers surface temperature through its effect on potential evapotranspiration. It therefore represents a more robust index than the SPI in regions where drought stress is increased by global warming. Importantly, the SPEI has proven improved capability to identify summer drought impacts on various hydrological variables as compared with the SPI (Vicente-Serrano et al. 2012; Marcos et al. 2015), and correlates well with vegetation stress and agricultural impacts across Europe, particularly in warmer and drier regions, underlying its utility for both drought monitoring and impact assessment (Stagge et al. 2015; Bachmair et al. 2018). While the SPI is limited by design as a purely meteorological index that neglects temperature and land–surface feedbacks, the SPEI (despite its advantages) has received comparatively less attention in forecasting applications (Marcos et al. 2015; Turco et al. 2017; Brands et al. 2025). Shyrokaya et al. (2025) evaluated the forecast skill of both indices across Europe using ECMWF's SEAS5 and found that SPEI outperforms SPI in summer over southern Europe, while SPI tends to perform better in northern Europe during winter and spring, possibly reflecting regional differences in the predictability of temperature and precipitation anomalies. Consistently, evidence from multimodel seasonal forecasting systems shows that temperature anomalies are generally more predictable than precipitation (Sánchez-García et al. 2018; Mishra et al. 2019; Calì Quaglia et al. 2022; Buontempo et al. 2022; Manzanas et al. 2022), possibly encouraging the use of indices that incorporate temperature in their formulation.

In this study, we investigate the seasonal predictability of European summer meteorological drought conditions using state-of-the-art seasonal forecast systems currently available through the Copernicus Climate Change Service (C3S). Our aim is twofold: 1) to identify the most predictable drought index for forecasting summer drought conditions across Europe and 2) to rigorously assess the probabilistic skill of current C3S systems based on that index. We focus on the SPI and the SPEI, both calculated over 3 months for the June–August (JJA) season, with forecast lead times extending up to one season (i.e., initialized on 1 March). We first investigate which drought indicator exhibits the highest potential forecast skill in our domain, demonstrating that the SPEI provides the most spatially coherent and skillful signal across the region. We then examine the strengths and limitations of current C3S seasonal forecast systems, including the multimodel ensemble (MME), in predicting summer drought conditions as represented by SPEI-3. The multifaceted nature of probabilistic forecasts implies that no single metric is sufficiently comprehensive to single out the best forecast system (Murphy 1993; Jolliffe and Stephenson 2011; Wilks 2011; Bradley et al. 2019). Thus, we employ a multimetric evaluation framework to compare the relative merits in predicting the summer SPEI-3 of each individual seasonal prediction system participating in the C3S and their MME. Our analysis includes five widely used verification metrics: the anomaly correlation coefficient, the Brier score, the area under the receiver operating characteristic curve score, the fair continuous ranked probability score, and rank histograms. Each technique evaluates different aspects of forecasts, providing a comprehensive skill assessment. By applying this comprehensive set of metrics, we aim to clarify where and how seasonal forecasts can offer reliable information while also advancing the scientific understanding of drought predictability in a region characterized by substantial forecast uncertainty.

2. Data and methods

a. *Seasonal forecast systems*

We use outputs from all the seasonal prediction systems available from the C3S Climate Data Store (2018) that provide precipitation, minimum and maximum temperature data at $1° \times 1°$ resolution and daily frequency. Our analysis uses the latest version of the models, incorporating both hindcasts

TABLE 1. Summary and details of the forecast systems. List of individual seasonal forecast systems and their multimodel ensemble: acronym, originating center, last version of the prediction system, number of ensemble members in the hindcast, available years to perform the evaluation procedure (if two periods are indicated, the second one refers to the operational forecast), reference paper.

Acronym	Originating center	Model's last version	Hindcast members	Available years	References
ECMWF	European Centre for Medium-Range Weather Forecasts	SEAS5	25	1993–2023	Johnson et al. (2019)
MF	Météo-France	System 8	25	1993–2018, 2022–23	Battè et al. (2021)
UKMO	U.K. Met Office	GloSea6-GC3.2	14	1993–2016, 2021–23	Williams et al. (2018)
CMCC	Centro Euro-Mediterraneo sui Cambiamenti Climatici	CMCC-SPS3.5	40	1993–2016, 2021–23	Gualdi et al. (2020)
DWD	Deutscher Wetterdienst	GCFS 2.1	30	1993–2019, 2021–23	Fröhlich et al. (2021)
ECCC	Environment and Climate Change Canada	GEM5.2-NEMO	10	1993–2020, 2022–23	Lin et al. (2021)
MME	—	—	60	1993–2016, 2022–23	—

and operational forecasts to maximize the verification period. Table 1 lists all individual forecast systems used in the study, the model version, the size of the hindcast ensemble, the verification time ranges, and a reference technical paper. Additionally, we construct an MME where each seasonal forecast system contributes equally. The MME is built by selecting 10 ensemble members from each forecast system, sampled randomly from the available members (see Table 1 for a count of members available for each model).

We analyze the boreal summer season averaged over June, July, and August. To assess the dependence of forecast skill on the initialization date, we conduct an evaluation process across different lead times, considering forecasts initialized on June, May, April, and 1 March. Here, we use the expression "0-month lead time" when forecasts are initialized on 1 June. "One-month lead time" refers to forecasts initialized 1 month before the beginning of the aggregation period, thus on 1 May, and so on. This nomenclature applies to all forecast systems except for the UKMO forecast system, for which hindcasts are initialized on the 1st, 9th, 17th, and 25th of each month (with seven members for each initialization day), and forecast data are initialized each day of the month (with two members per day). For UKMO hindcasts, the output data produced with initialization days 25 May and 1 June are grouped into an ensemble of 14 members and are referred to as 0-month lead time. For longer lead times, the ensembles are built following the same logic, but with a 1-month lag from each other. Instead, for UKMO operational forecasts, 0-month lead time consists of predictions initialized from 25 May to 1 June, including eight different initialization days and forming an ensemble of 16 members. The 1-month lead time prediction ensemble involves April (a 30-day month) and thus includes only seven initialization days (14 ensemble members) from 15 April to 14 May. The "3-month lead time" forecasts involve forecast output from the 23 February to 1 March, thus including 7 starting dates and 14 members (16 members in leap years). Since operational forecasts typically have larger ensembles than hindcasts, for years evaluated on the operational forecasts we randomly subselect a sample of members that matches the size of the hindcast ensembles. This procedure avoids the effects of a larger ensemble size on the scores and ensures consistency across different time periods.

b. *Reference data*

We use the E-OBS gridded observational dataset (Cornes et al. 2018) as a reference, examining years 1993 through 2023. We analyze precipitation, minimum and maximum temperature data at 0.25° spatial resolution and daily frequency. The temperature and precipitation fields are interpolated to a common 1° × 1° latitude-/longitude-resolution grid, to match the grid of the seasonal forecast systems, by means of a first-order conservative remapping. We also tested ERA5 reanalysis data (Hersbach et al. 2020) for comparison and found generally consistent results. Despite some gridpoint-specific sensitivity to the choice of reference dataset, the main patterns of predictability remained largely unchanged. We ultimately selected E-OBS due to its direct observational basis and its widespread use in European climate studies. We use this dataset both to check the correspondence between the forecast products and the observed event and to construct a heuristic climatology-based forecast system (see section 3b) that is used as a benchmark to compare forecast skills of the C3S models.

c. *Summer droughts characterization with SPEI and SPI*

The SPEI is a statistical drought indicator that compares the water balance (WB), calculated as precipitation minus potential evapotranspiration (PET), at a particular location with its long-term WB distribution. We estimate the PET using the method formulated by Hargreaves and Samani (1985), due to its reliance solely on temperature data. The method requires daily minimum and maximum temperature, along with an empirical estimate of solar radiation based on latitude and day of the year. Because temperature enters the PET calculation nonlinearly via the Hargreaves equation, we corrected the forecasted data of temperature using a leave-one-out cross validation (i.e., excluding the current year when computing the corrective parameters) prior to calculating the PET. The correction consists of subtracting, for each grid point, the long-term mean of the model monthly temperature from the daily data and then adding the long-term mean of monthly temperature calculated from the historical E-OBS data. The SPEI was calculated using a 3-month time scale (SPEI-3) in the month of August, accumulating the WB data in the JJA season. We corrected temperature but not precipitation

because the SPEI inherently adjusts for mean and variance biases in P-PET by standardizing the WB anomalies. To do so, we fit the input data to a statistical distribution, then transform these values into a normal distribution space with a mean of 0 and a standard deviation of 1, using an inverse normal (Gaussian) function. Data fitting was performed using a log-logistic distribution, as it is widely adopted in the literature and recommended by the methodological guidelines of Vicente-Serrano et al. (2010), to whom we point the reader for a more detailed explanation of the SPEI calculation. The three parameters of the distribution are computed using the maximum likelihood method. The SPI is calculated analogously, but using precipitation only and fitting the data to a gamma distribution, following the standard formulation of McKee et al. (1993). Before the forecast skill evaluation, to avoid artificial skill as an effect of temporal trends, data are detrended by removing their linear trend over the full historical time series, obtained by least squares regression of each model's ensemble at each grid point.

d. *Forecast skill evaluation*

Our evaluation process includes one deterministic score [the anomaly correlation coefficient (ACC), which considers the ensemble mean of the forecasts], three probabilistic scores [the Bier score (BS), the area under the receiver operating characteristic curve (AUC), and the fair continuous ranked probability score (FCRPS)], and the rank histograms (RH) to check the quality of the forecast ensemble. Please refer to Calì Quaglia et al. (2022) for a more comprehensive discussion of the properties of these scores in the context of seasonal predictions. The study domain spans the coordinates (11°W–33°E, 34°–70°N). We excluded North Africa from the analysis due to substantial data gaps in the observational dataset in that region and to focus the study on the European continent, where E-OBS coverage is more robust and consistent.

The ACC (Jolliffe and Stephenson 2011) measures the spatial correlation between forecasted and observed anomalies on a gridpoint basis, indicating their degree of agreement. It serves as a measure of potential utility since it ignores unconditional bias and reliability issues (i.e., the alignment between forecast probabilities and observed relative frequencies) in the forecast distribution. In this study, the ACC is applied to SPEI-3, SPI-3, 3-month accumulated precipitation, and the daily temperature range (Tmax–Tmin). The computation of anomalies is performed for the meteorological variables, but not for the two drought indices (SPEI and SPI), as they are inherently expressed as standardized anomalies. Therefore, for SPEI-3 and SPI-3, the ACC is computed as the Pearson correlation coefficient in time between the ensemble mean forecast and the corresponding observational reference values. Statistical significance is assessed using a one-sided Student's t test, with a false rejection rate set at $q = 0.05$. The ACC is used here to benchmark the forecast skill of the underlying meteorological variables employed in the construction of SPEI and SPI. This comparative evaluation is carried out exclusively for the ACC and is not extended to the probabilistic metrics presented in later sections.

The BS (Brier 1950) and the AUC (Jolliffe and Stephenson 2011) evaluate forecast skills for single-category events and are recommended by the WMO in the Standardized Verification System for Long-Range Forecasts. In our analysis, when using categorical scores, we refer to the issued probabilities conditioned on the event SPEI-3 < -0.8 to capture drought conditions (Svoboda et al. 2002). After trend removal, moderate drought events become rarer (approximately 21.1% in a nondetrended dataset), making them effectively more extreme. The BS and the AUC test the early warning skill for drought occurrence of the forecast systems.

The BS calculates the squared difference between the probabilities of having SPEI-3 < -0.8 and the binary observation (yes/no), averaged over time at each grid point. It is a measure of the overall accuracy of the forecast, in that it quantifies the average correspondence between issued probabilities and actual outcomes. It can be partitioned into three terms reflecting reliability, resolution, and uncertainty (Wilks 2011).

The AUC measures the forecast system's ability to distinguish events from nonevents (discrimination), using the area under the ROC curve as a summary statistic (Mason and Graham 1999). The ROC curve displays the hit rates against false alarm rates for a set of varying probability thresholds. It can be considered as a measure of potential usefulness because it is conditioned by the observations (i.e., given that a drought occurred, what was the corresponding forecast?), but it is not sensitive to bias.

Unlike the BS and the AUC, the continuous ranked probability score (CRPS; Hersbach 2000) evaluates the entire forecast distribution rather than verifying a specific event; thus, it is not restricted to the case of drought alone. It measures the distance between the cumulative distribution function of the forecasted SPEI-3 values and the observed values, represented as a step function that jumps to 1 at the point where the variable equals the observations. The CRPS is a measure of the overall accuracy of the forecast and indirectly measures sharpness (i.e., the tendency to forecast extreme probabilities near 0% or 100%). In this study, we use the FCRPS, a refined version of the CRPS that accounts for the dimension of the ensemble size (Ferro et al. 2008) to avoid artificial inflation of the forecast skill.

For all lead times, each metric is calculated annually for the available verification years (see Table 1) and then averaged over time. Since the scores are not normalized, they are successively transformed into a skill score (SS), except for the ACC, which is easily interpretable on its own. An SS allows us to compare the performance of a dynamical forecast to a benchmark (typically an ad hoc prediction system that requires minimal effort to be produced), translating the forecast quality into a gain or loss (or percentage improvement or worsening) over much simpler and computationally less expensive prediction products. This standardized way to evaluate the performance of the forecast systems also facilitates easier comparison among them. SS is calculated as

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

where S is the score of the dynamical forecast system, S_{ref} is the score of the reference forecast, and S_{perf} is the perfect

TABLE 2. Seasonal forecast verification process. List of verification metrics and their acronyms, type (deterministic or probabilistic), and focus of the verification (ensemble mean or forecast distribution), attributes of forecast quality assessed by the specific metric, SS transformation, and reference papers.

Verification metric	Type and focus of the verification	Attribute	SS	Reference papers
Anomaly correlation coefficient (ACC)	Deterministic, ensemble mean of SPEI-3 forecasts	Correlation	—	Jolliffe and Stephenson (2011)
Brier score (BS)	Probabilistic, categorical (conditioned on SPEI-3 < -0.8)	Accuracy, resolution, reliability, uncertainty	Eq. (1)	Brier (1950), Wilks (2011)
Area under the ROC curve score (AUCS)	Probabilistic, categorical (conditioned on SPEI-3 < -0.8)	Discrimination	Eq. (2)	Jolliffe and Stephenson (2011), Wilks (2011)
Fair continuous ranked probability score (FCRPS)	Probabilistic, multicategorical (not conditioned)	Accuracy, sharpness	Eq. (1)	Hersbach (2000), Ferro et al. (2008)
Rank histogram (RH)	Probabilistic, Multicategorical (not conditioned)	Ensemble quality, bias	—	Hamill and Colucci (1997), Wilks (2011)

score. Here, as a reference forecast, we use a heuristic prediction procedure based on the E-OBS climatology. For each year, the probabilistic distribution of SPEI-3 values at each grid point consists of the resampled SPEI-3 values reconstructed in the EOB-S dataset from 1993 to 2023 in that specific grid point, excluding the current year's value. As a result, our heuristic climatology-based prediction system (CLM) is made up of a probabilistic ensemble of 30 members (one less than the full set of years under consideration). Positive SS indicates grid points where the dynamical forecast outperforms the CLM, SS values near zero indicate no added value compared with the CLM, and negative SS indicates grid points where the dynamical forecast underperforms relative to the CLM.

The SS for AUC (AUCSS) is computed differently, as the percentage of improvement over a random classifier, and is derived using the following formula (Wilks 2011):

$$\text{AUCSS} = 2(\text{AUC} - 0.5). \tag{2}$$

In this study, an SS is deemed significant at a given grid point if the forecasts are significant at the 95% confidence level compared to a random forecast. To do so, we perform a bootstrap test with 1000 repetitions. In each iteration, a random forecast is generated by substituting the correct forecast for a given year with a randomly sampled forecast drawn from all available ensemble members and years, excluding the correct year and allowing repetitions. We then compute the SS for this random prediction ensemble. If the SS from the dynamical system falls outside the 95th percentile of the random distribution, we consider the forecast skill to be statistically significant. This procedure is applied independently at each grid point; therefore, the 95th percentile threshold is computed locally, based on the variability and forecast distribution specific to each location.

Last, we use rank histograms to assess whether the probability distribution of forecasts well represents the observed variability. The RH ranks the observation among the ordered set of ensemble forecasts, with a flat histogram indicating a well-calibrated ensemble. Deviations from the ideally flat histogram suggest systematic bias, though a flat RH does not guarantee a skillful prediction system (i.e., the RH for a climatology-based forecast is flat by construction). In our study, the RH helps evaluate whether forecasts tend to over- or underpredict SPEI-3. It also serves as a tool to identify a forecast ensemble that is overconfident (too small) or underconfident (too big). Table 2 provides a summary of all the verification metrics used in the study, the type (deterministic/probabilistic) and target of their verification, the measured attribute of forecast quality, the equation to compute the skill score (if any), and relevant references.

To simplify model performance interpretation, we categorized the models based on their skill at a lead time of 0 month. For each metric that provides a numerical score [ACC, Brier skill score (BSS), AUCSS, and fair continuous ranked probability skill score (FCRPSS)], we compute the area-weighted mean values over different regions. Because skill scores give an indication of the degree of improvement or worsening over a heuristic forecast, the performance categories are defined as follows: Very good refers to forecasts that, on average in the region, provide a clear added value with respect to the heuristic benchmark, good indicates moderate improvement, fair refers to forecasts providing comparable information, and poor describes forecasts that provide no added value. For the ACC, we have retained the same terminology, although in this context good does not imply added value relative to a benchmark, because the ACC is not a normalized comparative metric. Performance categories were defined using empirically determined, metric-specific thresholds detailed in Table S1 in the online supplemental material, accounting for differences in metric formulations: AUCSS is evaluated against a random classifier, BSS and FCRPSS are benchmarked against a climatological prediction system, and ACC is not an SS but a raw correlation measure.

3. Results

a. Deterministic skill for drought indices and meteorological variablescenter

To assess their deterministic performance in summer SPEI-3 forecasting, Fig. 1 depicts the ACC for each seasonal forecast system at 0-month lead time. The ACC ranges from -1 to 1, where 1 represents a perfect correlation, 0 indicates no correlation, and -1 represents an antiphase relationship with

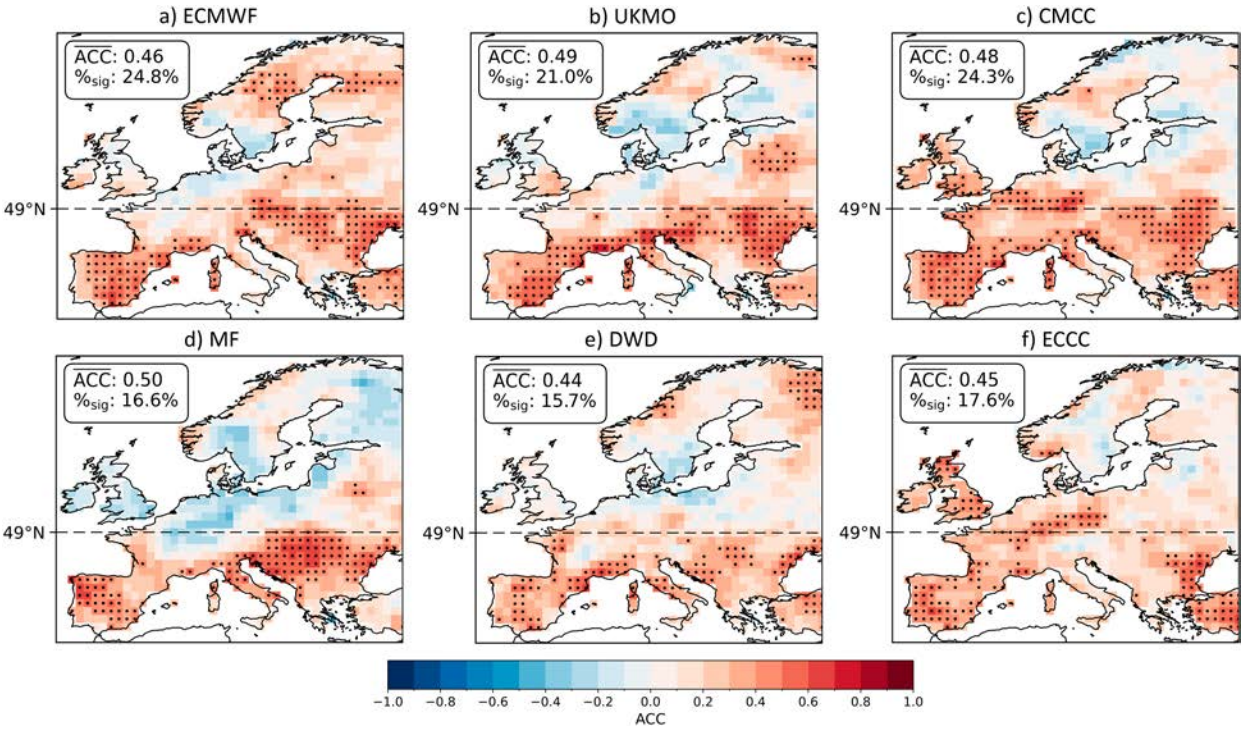

FIG. 1. ACC computed in time between seasonal forecasts (ensemble mean) and observations of summer SPEI-3. Panels show the spatial maps for all individual prediction systems for forecasts issued at 0-month lead time (on 1 Jun): (a) ECMWF, (b) UKMO, (c) CMCC, (d) MF, (e) DWD, and (f) ECCC. Stippling indicates statistical significance at the 95% level (p value < 0.05). Text boxes display the area-weighted mean ACC for points of statistical significance and the percentage of grid points with statistical significance relative to the total domain. The horizontal dashed line at 49°N marks the latitude boundary separating the MED and NEU regions, as defined in the text.

observations. Text boxes in the panels indicate the area-weighted mean ACC in points of statistical significance and the percentage of grid points with significant skill. Typically, an ACC of 0.6 represents a reasonable lower limit for delimiting field forecasts that are effective in capturing synoptical patterns (Wilks 2011).

ECMWF, UKMO, and CMCC exhibit the widest spatial extent of statistically significant skill (24.8%, 21.0%, and 24.3% of grid points, respectively), with large portions of southern and eastern Europe showing moderate to high positive correlation values. In regions of statistical significance, mean ACC scores range from 0.46 to 0.49, suggesting robust agreement with observed interannual variability across broad regions. On the other hand, Météo-France (MF), DWD, and Environment and Climate Change Canada (ECCC) yield moderate mean ACC values (up to 0.50 in MF), but their skill is more spatially confined, with significant areas concentrated in specific subregions (ranging from 15.7% to 17.7% of the domain). Notably, all models exhibit more widespread significant correlation patterns across the Mediterranean region (MED, defined here as southward of 49°N) at 0-month lead time. The strongest positive values are found in the Iberian Peninsula, Sardinia, and eastern Europe, specifically in between the Balkans and the Black Sea coast. Notwithstanding, correlation patterns vary across regions. Two main exceptions are Greece and southern Italy, where the ACC is either low or negative for most of the models, none of which exhibits statistical significance. UKMO, CMCC, and MF show good correlation over the Alpine region, despite its complex topography. The pattern of correlation is generally patchier in northern Europe (NEU, northward of 49°N), with extensive areas exhibiting negative ACC. Among the models, ECMWF performs best in the northernmost latitudes, showing large areas of positive ACC, although statistical significance is limited. Finally, ECCC and CMCC demonstrate similar significant ACC patterns, being the only models where forecasts of SPEI-3 are significant in the British Isles. All systems consistently show the lowest ACC in regions bordering the Baltic Sea.

Figure 2 represents the ACC between MME ensemble mean forecasts and observations for four variables: SPEI-3 (Fig. 2a), SPI-3 (Fig. 2b), precipitation (Fig. 2c), and temperature range (Tmax–Tmin) (Fig. 2d), across lead times from 0 to 3 months. At 0-month lead time, in regions of statistical significance (30.1% of grid points), the area-weighted mean ACC reaches 0.49, indicating moderate ability to capture inter annual SPEI-3 fluctuations. Because the MME reflects the average skill of its contributing models, the highest correlations are concentrated over the Iberian Peninsula and parts of eastern Europe, with ACC values between 0.6 and 0.7. In contrast, skill in NEU is generally low, suggesting limited forecast utility in this region. Note that the ECMWF model (Fig. 1a) alone displays stronger skill in the region. SPEI-3 exhibits a more extensive and coherent area of skill with respect to SPI-3, that reveals a fragmented pattern, with statistically

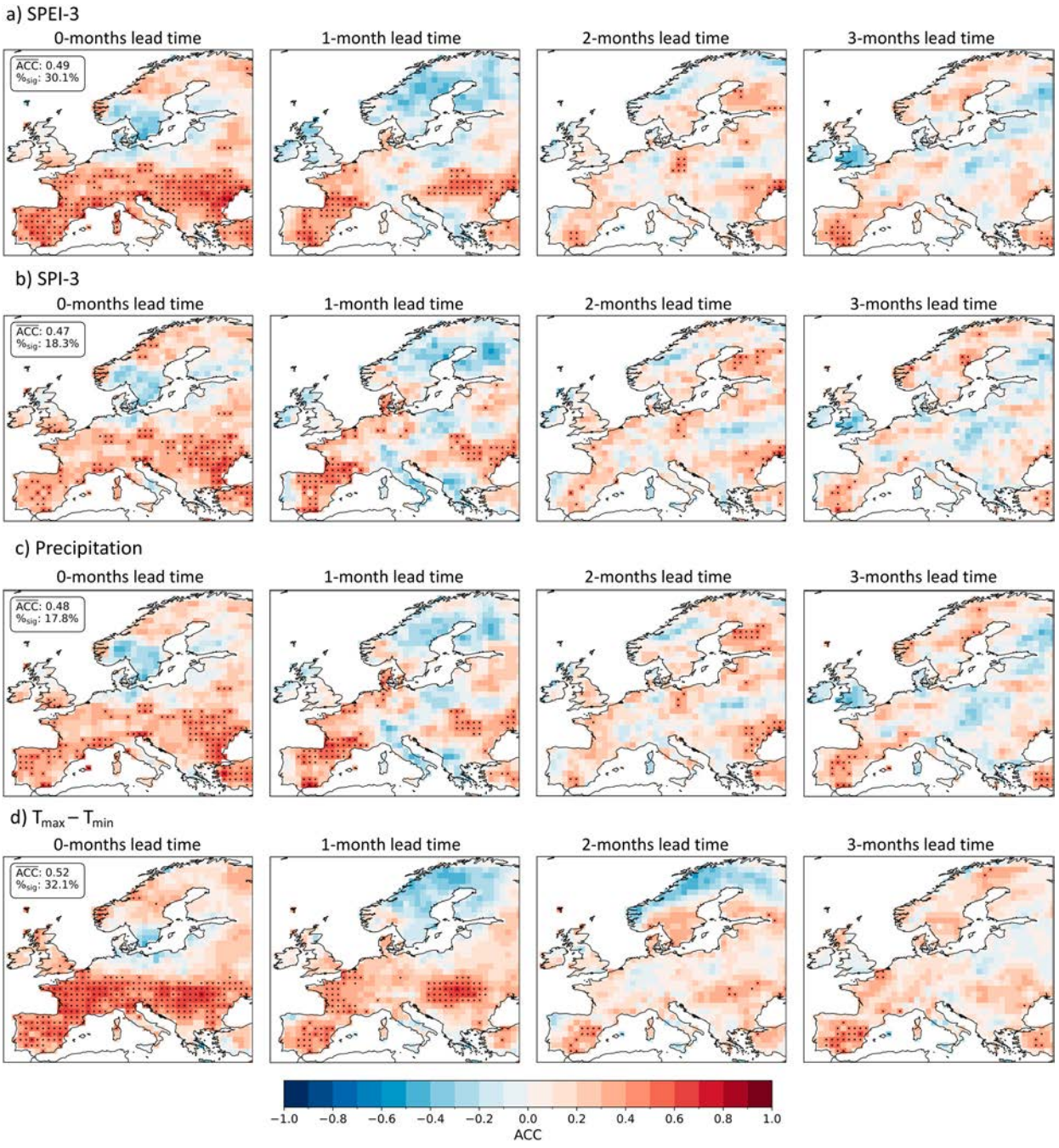

FIG. 2. ACC computed in time between MME forecasts (ensemble mean) and observations for drought indices and related variables. Each row shows results for a specific variable: (a) SPEI-3, (b) SPI-3, (c) precipitation, and (d) daily temperature range (Tmax–Tmin). Within each row, the four panels represent forecasts initialized at 0-, 1-, 2-, and 3-month lead times, respectively. Stippling indicates statistical significance at the 95% level (p value < 0.05). The text box in each leftmost panel reports the area-weighted mean ACC over statistically significant land grid points, along with the percentage of significant grid points relative to the total domain.

significant ACC mostly confined to parts of eastern Europe. Furthermore, we note a strong spatial agreement with 3-month precipitation ACC. This indicates that SPI-3 adds little predictive value beyond precipitation alone, while the inclusion of temperature in SPEI-3 provides added skill, particularly over central and southwestern Europe. In contrast, Tmax–Tmin emerges as the individual variable with the highest forecast skill among those evaluated (ACC = 0.52, 32.1% significant area). Additional analysis (not shown) demonstrates that this combined variable performs worse than both Tmax and Tmin when

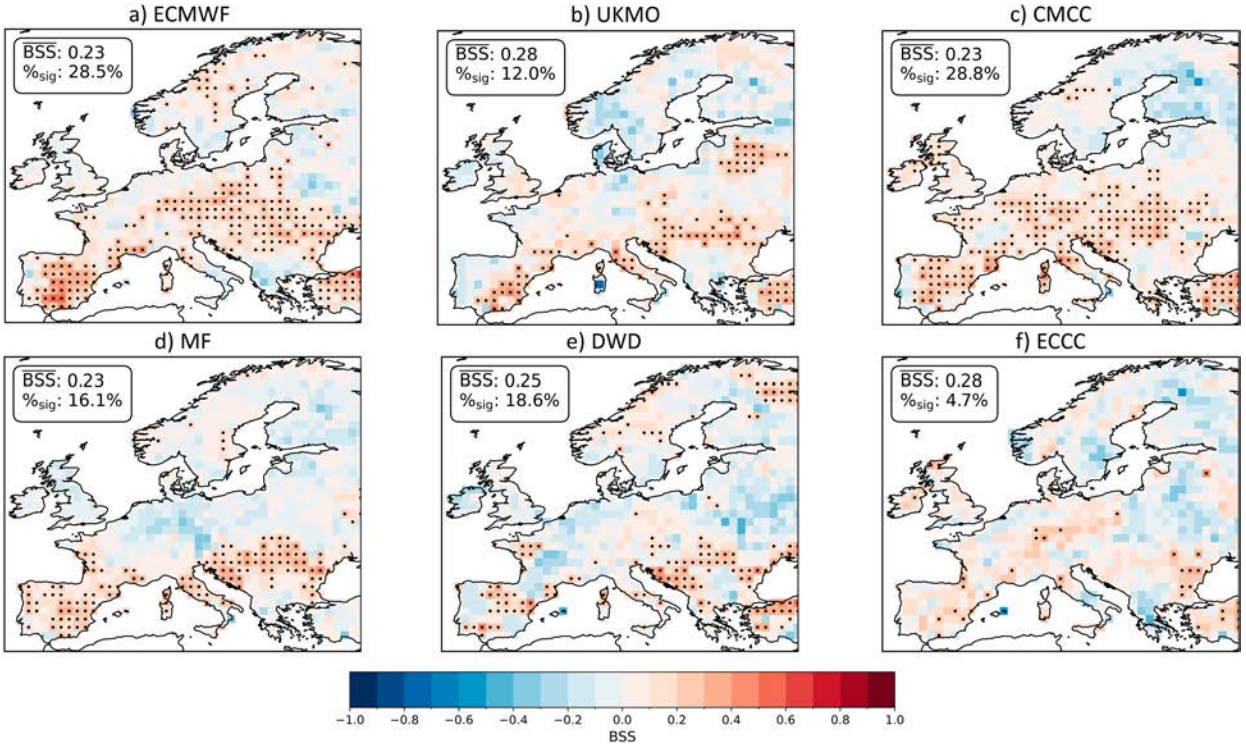

FIG. 3. BSS for probabilities of SPEI-3 < -0.8 with respect to CLM for all individual prediction systems issued at 0-month lead time (on 1 Jun): (a) ECMWF, (b) UKMO, (c) CMCC, (d) MF, (e) DWD, and (f) ECCC. Stippling indicates significant BSS values at the 95% confidence level over a random forecast according to a bootstrap test. Text boxes display the area-weighted mean BSS value for points of statistical significance and the percentage of significant grid points.

evaluated separately. Given that the combination of Tmax and Tmin into a single variable tends to reduce skill, we opted not to use the Penman–Monteith equation for PET estimation (Beguería et al. 2013). While more physically based, Penman–Monteith requires additional meteorological inputs (e.g., radiation, humidity, wind speed) that would introduce further layers of uncertainty. Overall, our result reinforces the added value of SPEI-3 for forecasting applications, strengthening related meteorological drought predictability skill particularly over central and southwestern Europe. In northern regions, although SPI may occasionally outperform SPEI, the correlation remains low overall, confirming the poor predictability of summer drought indices in NEU. Areas of negative correlation in SPEI and SPI at 0-month lead time often co-occur with regions where precipitation shows little to no skill or even negative correlation. Therefore, poor precipitation predictability is a major contributor to the reduced skill of both indices. However, Tmax–Tmin also exhibits some negative skill, particularly around the Baltic Sea. This further reinforces the negative correlation observed in SPEI at northern latitudes, where both temperature and precipitation drivers may be poorly captured by the forecasts.

As expected, ACC values decline with increasing lead time (see Fig. S1 for individual models). At 1-month lead time, although the overall skill decreases across all variables, Tmax–Tmin retains a more spatially coherent correlation pattern compared to precipitation, which appears increasingly scattered and noisy. This relative stability in the temperature signal contributes to preserving some degree of spatial consistency in SPEI-3 forecasts at this lead time. Certain areas such as the Iberian Peninsula and the eastern Mediterranean continue to show significant SPEI-3 and SPI-3 correlation skill, the latter being weaker and more localized. At longer lead times, all variables begin to show broader regions where the forecasts are in antiphase with observations, indicating poor predictive skill. Nevertheless, Tmax–Tmin tends to exhibit fewer grid points with negative correlation compared to precipitation, confirming its relatively greater robustness. Considering this, SPEI-3 is adopted as the reference index for the remainder of the study.

b. *Probabilistic skill for single-category events (SPEI-3 < -0.8)*

The BSS measures forecast accuracy relative to CLM based on the mean squared error of the issued probabilities for the event SPEI-3 < -0.8. Positive values indicate grid points where the forecast outperforms CLM, and vice versa. Figure 3 shows the spatial maps of BSS for forecasts issued at 0-month lead time for each model. Text boxes in the panels indicate the area-weighted mean BSS in points of statistical significance and the percentage of grid points with significant skill.

ECMWF (Fig. 3a) and CMCC (Fig. 3c) generally perform the best, with significant skill across 28.5% and 28.8% ($\overline{\text{BSS}} = 0.23$) of the land grid points, respectively. Maps in

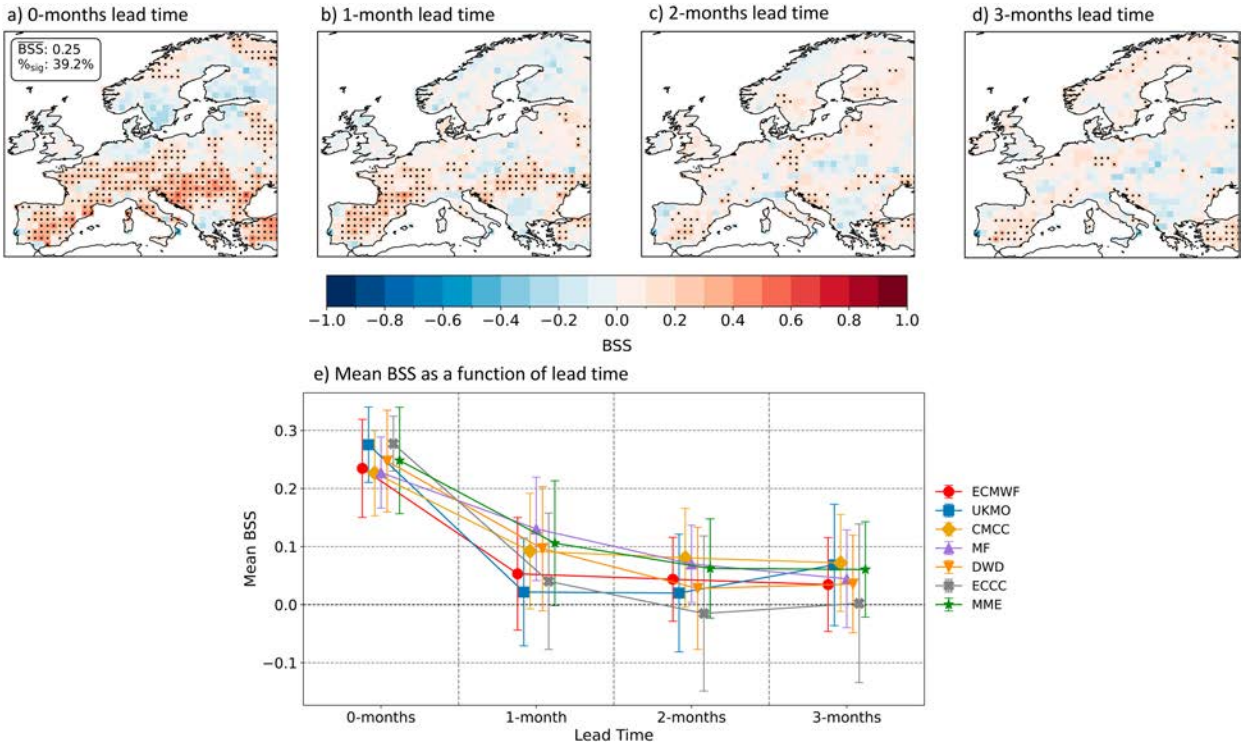

FIG. 4. BSS for probabilities of SPEI-3 < -0.8 with respect to CLM at different lead times. Panels show the spatial maps for MME forecasts issued at (a) 0-, (b) 1-, (c) 2-, and (d) 3-month lead time. Stippling indicates significant BSS values at 95% confidence level over a random forecast according to a bootstrap test. (e) The area-weighted average BSS value for each model in points which show statistical significance at 0-month lead time [indicated by stippling in Fig. 3 for individual models and in (a) of the present figure for the MME] across different lead times. Error bars indicate the standard deviation of the BSS values in regions of statistical significance. To enhance clarity and prevent overlap, offsets have been applied to the x values of each model's data points.

Fig. 3 suggest that these models perform similarly in drought forecasting, with BSS values generally ranging from 0.1 to 0.4 over southern, central, and southeastern Europe. UKMO (Fig. 3b), MF (Fig. 3d), and DWD (Fig. 3e) exhibit comparatively lower skill in central Europe but shows strong skill in some regions localized in southeastern Europe and the Iberian Peninsula. These models achieve $\overline{\text{BSS}} \approx 0.23$–$0.28$, covering 12%–18.6% of grid points with significant skill, confirming their usefulness in specific subregions. ECCC shows statistical significance in less than 5% of the domain, demonstrating poor probabilities accuracy. However, it is worth noting that its low skill in categorical forecasting can be low due to the small ensemble size (10 members), because smaller ensembles do not allow for a good representation of the forecast uncertainty, leading to less reliable probabilities (Manzanas et al. 2022). Across all models, reduced skill is consistently observed over the Scandinavian Peninsula, where ACC and BSS values are generally low or negative.

Figure 4a through Fig. 4d shows spatial maps of the BSS for the MME for forecasts issued from 0- to 3-month lead time. Figure 4e displays the area-weighted average BSS for all systems in grid points that are significant at lead time 0 (shown in Fig. 3), calculated for lead times ranging from 0 to 3 months.

The map of the MME skill at 0-month lead time (Fig. 4a) provides a consolidated view of the forecast performance across the different models. In points of statistical significance (39.2% of the domain, highest proportion across models), the mean BSS is 0.25, demonstrating a clear added value compared to CLM in terms of drought forecasting accuracy and a notable geographic extent of significant skill. As the lead time increases, probabilities' accuracy declines in each single model (Fig. S2 and Fig. 4e) and the MME (Figs. 4b–d), though BSS values and spatial patterns of significant skill remain stable from 1- to 3-month lead time. At lead time 1 month, the MME issues accurate probabilities in the Iberian Peninsula (where it demonstrates significant skill at all lead times) and some areas in central and eastern Europe. Among individual models, CMCC and MF achieve the best skill at lead time 1-month in western Europe, while DWD performs better further east. At longer lead times, CMCC consistently retains good skill in western Europe, enhancing the overall performance of the MME. Since a 3-month averaging period means higher skill in June influences scores through August, we conducted an additional analysis (not shown) assessing the BSS for monthly components (SPEI-1 for June, July, and August individually) for the CMCC system. Results demonstrate that the forecasts issued on 1 June exhibit the highest skill in June itself, with slightly positive values persisting across July and August in central Europe, the Iberian Peninsula, and parts of eastern Europe, and neutral and slightly negative

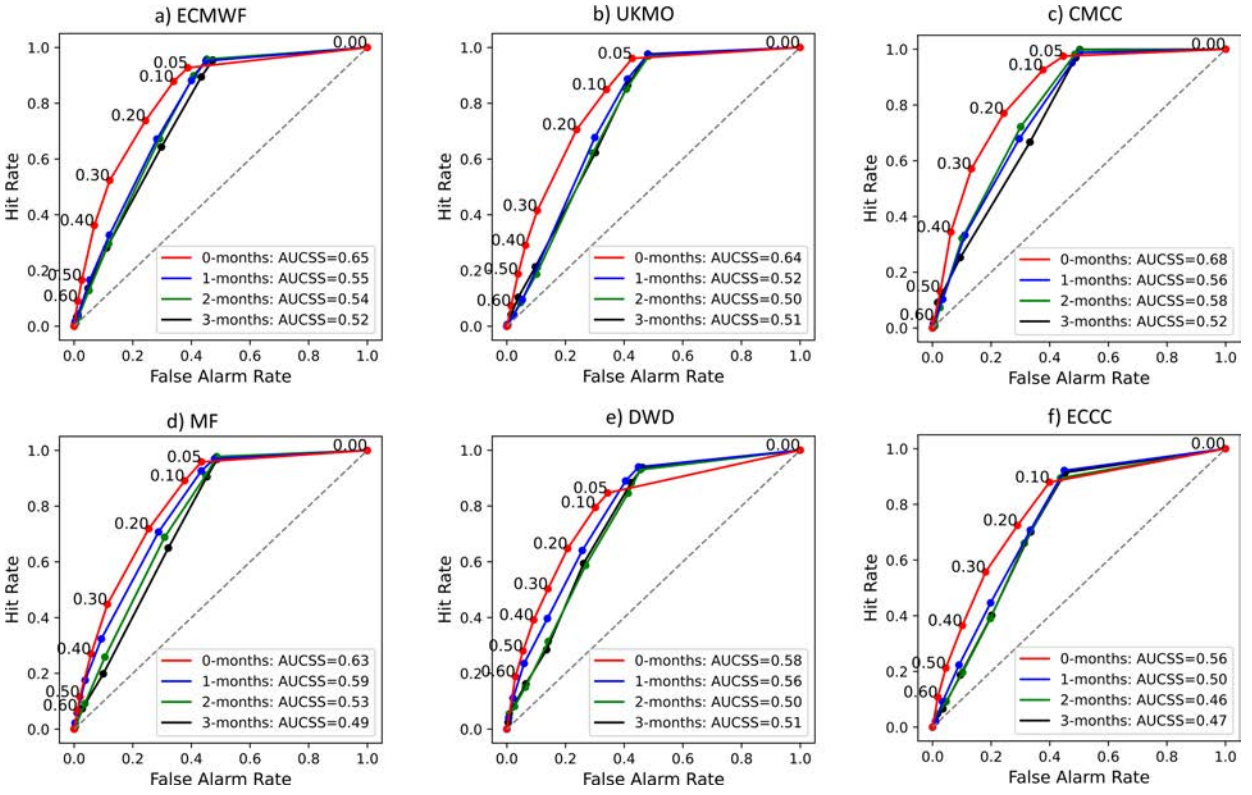

FIG. 5. ROC curves for seasonal forecasts of SPEI-3 < -0.8 in the MED (south of 49°N) for (a) ECMWF, (b) UKMO, (c) CMCC, (d) MF, (e) DWD, and (f) ECCC for forecasts issued at 3-month lead time (black), 2-month lead time (green), 1-month lead time (blue), and 0-month lead time (red).The associated AUCSS values for the different lead times are provided in the text box. The gray dashed line represents the performance of a random classifier (AUC = 0.5). Numbers on the red line correspond to the probability thresholds at which the hit rate and false alarm rate have been evaluated to build the curves.

values elsewhere. This aspect could be particularly important for certain sectoral applications because it suggests that forecasts may be useful for extended periods in the future (more than 1 month from the issuance date) for users who do not require highly detailed temporal resolution.

We compute the ROC curves by pooling together all grid points within the MED region (Fig. 5), to increase event samples for each probability threshold, as recommended by the WMO guidelines. Probability thresholds for building the curves vary by models. For instance, ECCC can only assign numerical probabilities that are multiples of 10, as it includes only 10 members. Corresponding spatial maps of AUCSS for each model and all lead times can be found in the supplementary information (SI) for scrutiny (Fig. S4). Any forecasting system with a negative AUCSS performs no better than a random classifier (gray dashed lines in Fig. 5). Here, the AUC measures the ability of the forecast systems to discriminate the event SPEI-3 < -0.8.

All ROC curves for the MED region exceed the random classification line, indicating some discrimination skill for each model at all lead times. CMCC (Fig. 5c) demonstrates the best discrimination skills at 0-month lead time, with an AUCSS of 0.68. ECMWF (Fig. 5a), UKMO (Fig. 5b), and MF (Fig. 5d) slightly underperform CMCC, with AUCSS of 0.65, 0.64, and 0.63 at lead time 0, respectively. All models exhibit a strong drop in skill at lead time 1. Overall, DWD (Fig. 5e) and ECCC (Fig. 5f) demonstrate reduced performance at all lead times. Although the AUCSS generally decreases for longer forecast periods, it does not show a steady decline with lead time. For example, UKMO, DWD, and ECCC show higher AUCSS at 3-month than at 2-month lead. Figure S3 shows the ROC curves for the NEU region. All curves are above the random classification line. ECMWF performs best, yielding an AUCSS of 0.57 at 0-month lead time. However, overall discrimination skill is low in the region.

Figure 6a through Fig. 6d shows the spatial maps of AUCSS for the MME for forecasts issued from 0- to 3-month lead time, while Figs. 6e and 6f display ROC curves for the MED and NEU regions. The spatial pattern of positive AUCSS broadly aligns with BSS across models and lead times (see also Figs. S2 and S4), but statistical significance over a random forecast is more rarely found (18.5% of the total domain with a mean AUCSS of 0.69). As mentioned, the AUCSS alone is not sensitive to bias, but regions where both the AUCSS and the BSS are positive ensure that the forecast systems have good quality in terms of discrimination, accuracy, reliability, and resolution. As shown in Fig. 6a, at 0-month lead time, the MME displays significant skill across

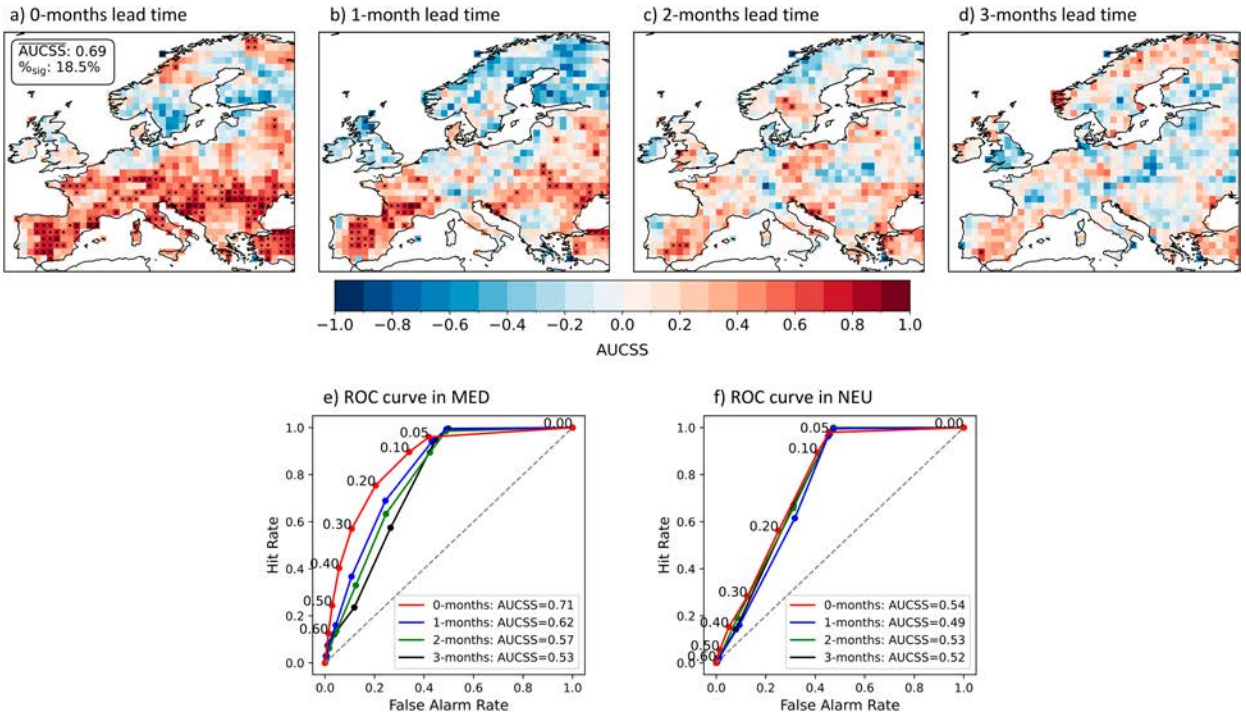

FIG. 6. AUCSS for MME for seasonal forecasts of SPEI-3 < -0.8. Panels show the spatial maps for MME issued at (a) 0-, (b) 1-, (c) 2-, and (d) 3-month lead time. Stippling indicates significant AUCSS values at the 95% confidence level over a random forecast according to a bootstrap test. (e),(f) The corresponding ROC curves aggregated over the MED and NEU, respectively. Each line represents a different lead time: 0 month (red), 1 month (blue), 2 months (green), and 3 months (black).

the southern MED region, excluding Greece and southern Italy. Northward, as expected, discrimination decreases sharply. With increasing lead time, while the BSS mostly turns neutral or slightly negative across much of the domain, the AUCSS reveals widespread negative skill.

Figure 6e demonstrates good discrimination skill for the MME at lead time 0 in the MED region, with an AUCSS of 0.71. Therefore, to maximize the hit rate while keeping the false alarm rate within a reasonable range, it is advisable to expect an SPEI-3 < -0.8 when the system issues a probability (at 0-month lead time) exceeding 10%–20% for the occurrence of such event. An analogous reasoning can be performed for the NEU region. However, the low AUCSS, combined with the low forecast reliability and resolution (Fig. 4f and Fig. S4), indicates that seasonal forecasts in that area should be interpreted with caution.

c. *Probabilistic forecast skills for SPEI-3*

So far, our analysis has focused on probabilistic forecasts for the categorical occurrence of a drought event. In contrast, the FCRPS evaluates the entire probability distribution of forecasted SPEI-3, evaluating the model's accuracy across dry and wet conditions. Figure 7 shows the FCRPSS of each forecast system at 0-month lead time. Text boxes in the panels indicated the area-weighted mean FCRPSS in points of statistical significance and the percentage of significant grid points.

Among all models, CMCC (Fig. 7c) and UKMO (Fig. 7b) exhibit the widest spatial coverage of statistically significant skill, with 36.1% and 33.2% of grid points exceeding the 95% significance level, respectively. These two models show spatially coherent skill especially over southern and southeastern Europe, including Turkey and the Balkans. ECMWF (Fig. 7a) and MF (Fig. 7d) show more localized but still significant clusters of skill, particularly over the western Mediterranean and parts of central Europe. The spatial extent of significant skill for ECCC (Fig. 7c) is limited (14.4%), suggesting strong performance only in some areas, such as parts of Spain and southern Italy. It stands out as the best performing model over the British Isles in terms of overall accuracy. However, categorical scores (Fig. 3f) indicate not significant skill in specifically capturing drought events in the region. In contrast, DWD (Fig. 7e) shows the lowest average FCRPSS and the smallest proportion of significant grid points (9.0%), indicating limited added value in probabilistic drought prediction compared to CLM. All models consistently lack overall accuracy in regions around the Baltic Sea.

Figure 8a through Fig. 8d depicts the spatial maps of FCRPSS from MME forecasts from 0- to 3-month lead time. Figure 8e displays the area-weighted average FCRPSS for all systems in significant grid points at lead time 0 (as shown in Fig. 7), calculated for lead times ranging from 0 to 3 months. For MME forecasts issued at lead time 0 in regions of statistical significance (36.6% of the domain), FCRPSS ranges between 0.06 and 0.31, with a mean value of 0.14. Despite being the system with the highest number of grid points with statistically significant skill across the domain, the MME does not

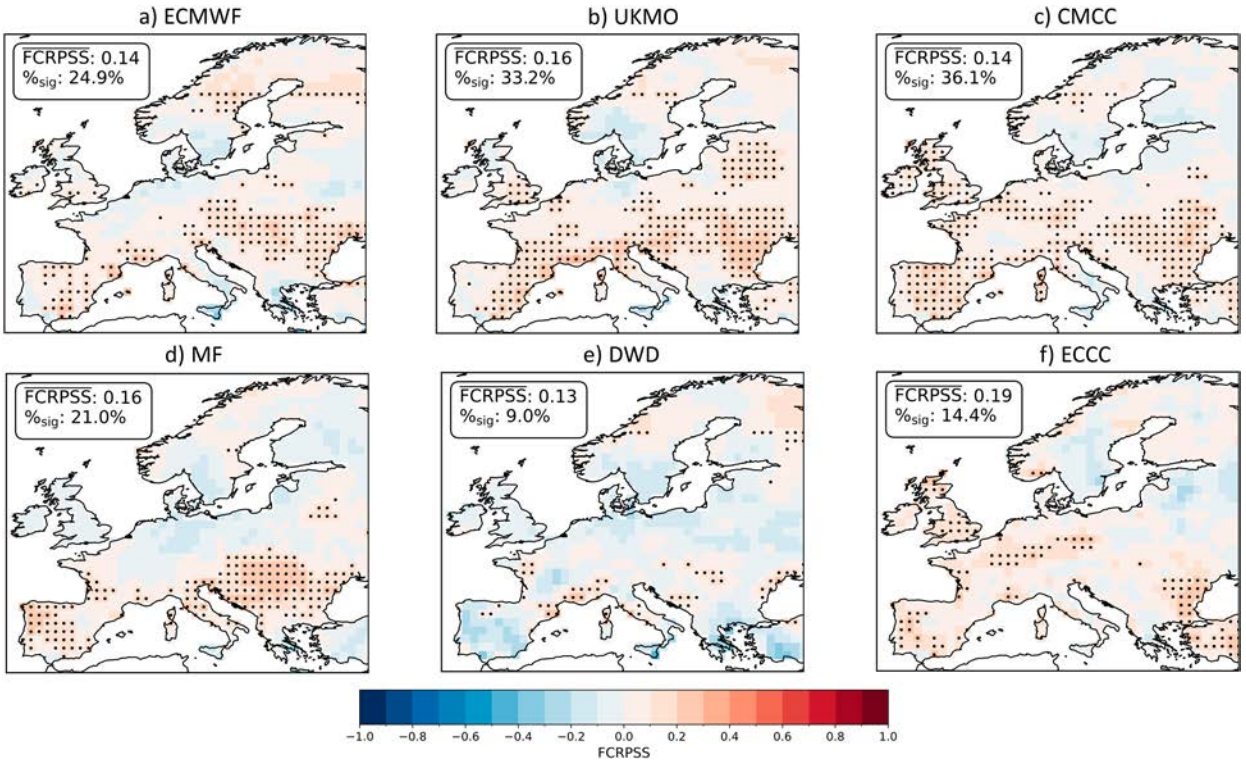

FIG. 7. FCRPSS for seasonal forecasts of summer SPEI-3 with respect to CLM. Panels show the FCRPSS spatial maps for all individual prediction systems issued at 0-month lead time (on 1 Jun): (a) ECMWF, (b) UKMO, (c) CMCC, (d) MF, (e) DWD, and (f) ECCC. Stippling indicates significant FCRPSS values at the 95% confidence level over a random forecast according to a bootstrap test. Text boxes display the area-weighted mean FCRPSS value at points of statistical significance and the percentage of significant grid points.

show the highest FCRPSS in specific regions, which are achieved by individual models, as expected from averaging effects in multimodel systems. At 1-month lead time, significant FCRPSS values persist over scattered areas, notably across the Iberian Peninsula and parts of eastern Europe. As the lead time increases, skill levels decline substantially: FCRPSS values become mostly neutral, and statistically significant areas become rare, although mean scores remain slightly positive, indicating residual improvement over CLM. Notably, CMCC outperforms all models and the MME in the Iberian Peninsula at all lead times We note that the mean FCRPSS of all systems except DWD never falls below 0 (Fig. 8e) that demonstrates average improvement over CLM at all lead times. DWD configures as the least accurate model in SPEI-3 forecasting at all lead times. Despite the drop in mean skill, the MME and several systems retain positive FCRPSS values up to 3-month lead, albeit large uncertainty ranges indicate strong variability in forecast performance across lead times.

d. Rank histograms

Figure 9 depicts the RH for all forecast systems and lead times, aggregating MED region data points. The frequency of the rank of observations is normalized to facilitate comparison across forecast systems. The normalization is based on the ideal frequency of $1/(n + 1)$, where n is the number of ensemble members of each system. The value corresponds to the frequency with which the observation would fall into each single bin in a perfect ensemble, where issued probabilities match observed frequencies. RHs assess the quality of the ensemble forecast. A uniform distribution indicates a well-calibrated ensemble, although it does not ensure good predictive skill. Deviations from uniformity, such as U-shaped distributions or peaks at the ends, suggest issues like underdispersion or systematic biases. RHs for the NEU region are shown in Fig. S6.

In the MED region, ECMWF shows a flatter histogram at shorter lead times, with a slight overforecasting bias (evidenced by the observation more frequently falling into the first rank), indicating a tendency to predict wetter-than-observed conditions. At a 3-month lead time, ECMWF's ensemble becomes underdispersed, with a spread too narrow to capture the whole observed variability of SPEI. UKMO, CMCC, MF, and ECCC display overdispersed ensembles, particularly at shorter lead times. This suggests that the spread of ensemble members is too wide compared to the actual year-to-year variability. Among these, UKMO and ECCC stand out for issuing relatively less biased ensembles, despite having fewer ensemble members than the other systems. Both produce rank histograms close to a uniform distribution (especially for forecasts issued at 3-month lead time), suggesting good calibration to observed data in the MED region, a finding also supported by the FCRPSS performance. DWD exhibits the most pronounced bias, characterized by a

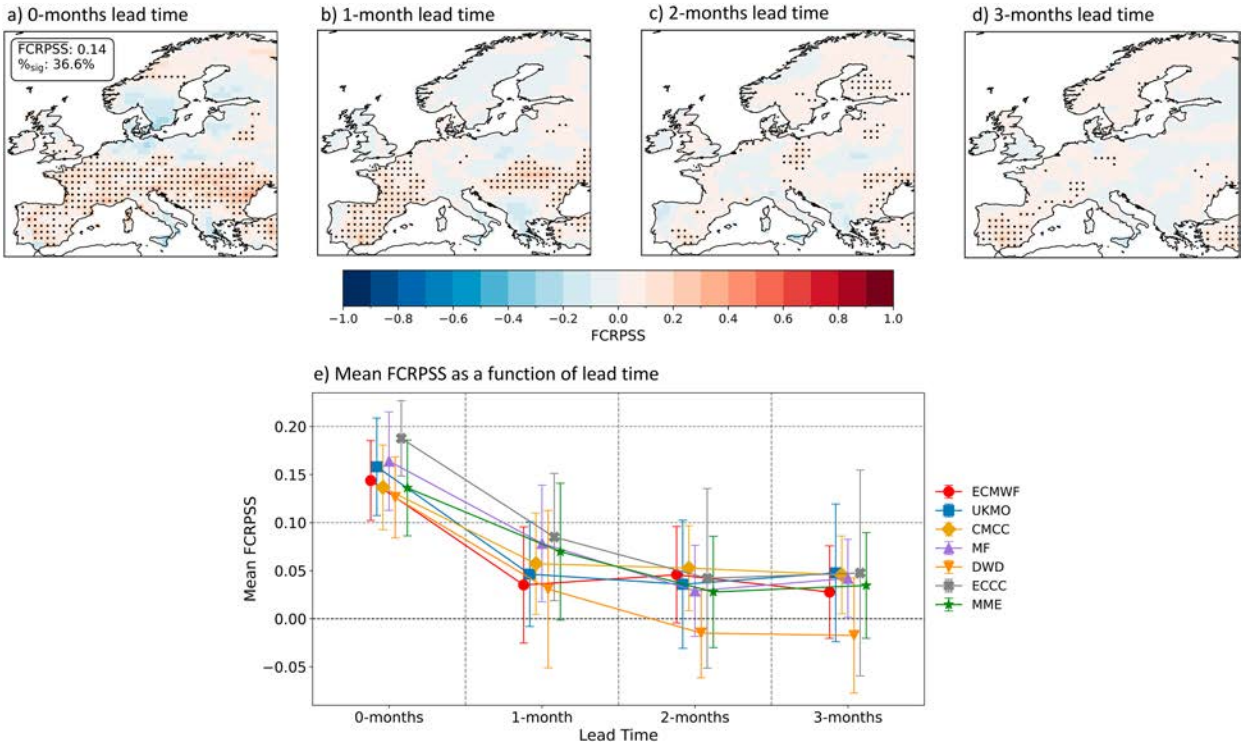

FIG. 8. FCRPSS of summer SPEI-3 forecasts with respect to CLM as a function of lead time. Panels on the first row show the spatial maps for MME forecasts issued at (a) 0-, (b) 1-, (c) 2-, and (d) 3-month lead time. Stippling indicates significant FCRPSS values at the 95% confidence level over a random forecast according to a bootstrap test. (e) The area-weighted mean FCRPSS value for each model in points which show statistical significance at 0-month lead time [indicated by stippling in Fig. 7 for individual models and in (d) of the present figure for the MME] across different lead times. Error bars indicate the standard deviation of the FCRPSS values in regions of statistical significance. To enhance clarity and prevent overlap, offsets have been applied to the x values of each model's data points.

strong underdispersion across all lead times, which indicates potential issues in extreme event predictions (both dry and wet). The MME shows a mixed behavior that does not fit neatly into a single category. At lead time 0, the observation most often falls in the first rank, yet the distribution is bimodal and skewed toward the higher ranks. At 3-month lead time, instead, the MME reveals an underdispersion bias, with stronger tendencies toward underforecasting. In NEU (Fig. S6), ECMWF, UKMO, and ECCC yield notably uniform histograms. CMCC and MF show an overforecasting bias at lead time 3 months that might

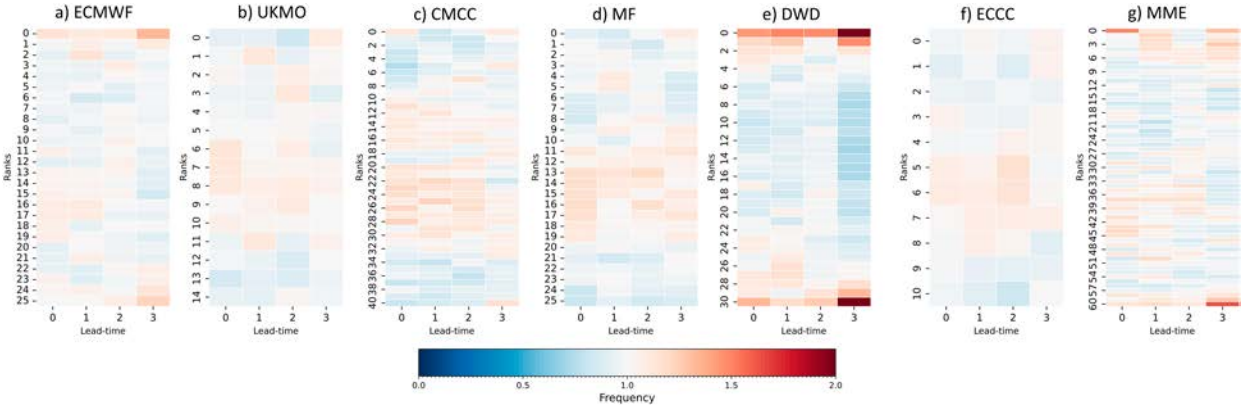

FIG. 9. Rank histograms for (a) ECMWF, (b) UKMO, (c) CMCC, (d) MF, (e) DWD, (f) ECCC, and (g) MME in the MED region (below 49°N) as a function of lead time for SPEI-3 forecasts in JJA. The color indicates the normalized frequency of the rank of observations with respect to the ensemble. Normalization is computed relative to the ideal frequency of $1/(n + 1)$, where n is the number of ensemble members of each system. An ideally flat rank histogram would have a normalized frequency of 1 across all bins.

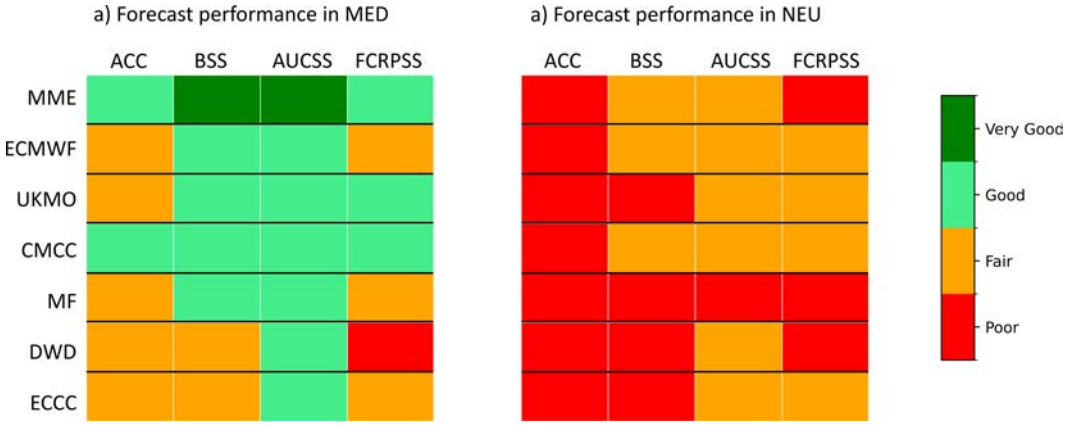

FIG. 10. Performance categories for all forecast systems across (a) the MED region and (b) the NEU region at 0-month lead time. Empirical thresholds defining the categories can be found in Table S1.

reflect a broader issue in capturing drought conditions, though their histograms are nearly flat at shorter leads. DWD remains the most biased system, showing underdispersion at 3-month lead time, albeit to a lesser extent than in the MED region. The MME produces almost flat histograms at short lead times, though an underdispersion bias persists at 3-month lead time.

e. *Summary of systems performance*

Figure 10 provides a comparative synthesis of the forecast skill of individual models and the MME across multiple scores (ACC, BSS, AUCSS, FCRPSS) in the two spatial domains (MED and NEU), as defined by a latitudinal threshold at 49°N. Each column captures a different facet of probabilistic or deterministic forecast skill.

Across MED, the MME emerges with generally superior performance. Forecast skill tends to degrade in NEU compared to MED, though the degradation is not uniform across models or metrics. Notably, the classification results remain stable when adjusting the boundary separating MED and NEU by a few degrees northward or southward, indicating low sensitivity to the precise choice of the latitudinal threshold. Models such as CMCC and UKMO consistently achieve high rankings across different metrics, while others (e.g., DWD and ECCC) exhibit more metric-dependent skill. The MME, however, remains the highest-ranked option under most scores, making it a strong candidate for operational and decision-support applications in climate forecasting over Europe. Figure S8 in the SI depicts the mean score (averaged over ACC, BSS, AUCSS, and FCRPSS) for the MME at lead time 0 month, for a spatial visualization of the areas where the MME has, on average, major skill.

4. Discussion

This study analyzed the full ensemble of state-of-the-art seasonal forecast systems provided by the C3S to evaluate their skill in predicting drought conditions across Europe in the boreal summer, a season during which high temperature anomalies drive strong evaporative demand in southern regions, while precipitation deficits in northern regions (where summer is typically wetter) can lead to rapid onset of meteorological droughts. We first carried out a preliminary assessment of the two most widely used drought indicators in seasonal prediction, SPI and SPEI. Consistent with previous research (e.g., Shyrokaya et al. 2025), we found that SPEI exhibits higher predictability than SPI in summer in southern Europe (i.e., southward of 49°N), as measured by the ACC. This difference is particularly evident at short lead times, when models more effectively capture temperature-related variables (e.g., daily temperature range) than precipitation. Importantly, we found that Tmax–Tmin retains strong and spatially coherent skill even at short lead times, contributing to the stability of SPEI forecasts where precipitation skill is more fragmented (see Calì Quaglia et al. 2022; Manzanas et al. 2022 for an analysis of Tmean and precipitation alone).

Successively, we used a comprehensive multiscore approach to evaluate several attributes of the SPEI forecasts quality in relation to the lead time (Murphy 1993). Probabilistic scores were converted into skill scores, benchmarking dynamical models against heuristic predictions. We used a dataset that combines hindcasts and operational forecasts from the same model version (details are provided in Table 1). While verification usually rely on hindcasts, the inclusion of operational forecasts allowed us to extend the time period to perform the verification process, enhancing the statistical robustness of the results.

In absolute terms, no forecast system consistently outperforms the others. However, in specific regions one may assign more credibility to certain systems. Notably, for all models the spatial distribution of statistically significant skill (which is lead-time dependent) is consistent across the different metrics. We found that the MME often outperforms individual models (with gridpoint-specific differences) across the full range of skill metrics, because multimodel ensembles tend to provide more reliable forecasts by averaging out biases inherent in single models (Krishnamurti et al. 2000). While it does not always show the highest gridpoint skill, it has the highest percentage of significant skill grid points, demonstrating good

accuracy, discrimination, and reliability in SPEI-3 forecasting, making it the most robust solution for continental-scale applications. Overall, most skillful forecasts are predominantly found over the Iberian Peninsula, the Balkans, and parts of the eastern Mediterranean. In contrast, in northern Europe, characterized by greater synoptic variability, forecast skill remains more limited and spatially fragmented for all models and the MME (see tables in Fig. 10). This latitudinal gradient in forecast skill is evident in both SPI-3 and SPEI-3 forecasts. However, it becomes more pronounced with SPEI-3, suggesting that the inclusion of temperature information reinforces the north–south contrast in drought predictability.

Conditioning SPEI-3 forecasts to a threshold of −0.8 (i.e., drought occurrence), we evaluated the quality of categorical probabilistic forecasts, uncovering small changes in the spatial patterns of significant skill with respect to ACC that considers the performance of deterministic forecasts. Since the BSS indicates that the most reliable forecasts of droughts are in grid points within the MED region, we focus on this region to construct the ROC curves in Fig. 5, as the AUCSS is not sensitive to biases by itself. All models demonstrate good discrimination abilities across all lead times. Importantly, these ROC curves can assist in identifying optimal probability thresholds for triggering adaptation responses by balancing hit rates and false alarm rates. Certainly, the selection of an appropriate threshold should consider sector-specific risk tolerance. We refer the reader to Wilks (2011) to expand on this concept in a seasonal forecasting context.

The FCRPSS, the only multicategorical probabilistic score used, assesses overall forecast accuracy beyond drought cases. It reveals that, on average, all models outperform the heuristic model in terms of accuracy and sharpness in vast regions at lead time 0 months. To confirm this, most RH shows quite flat bins, indicating good quality of the ensembles. The main exception is DWD, which underperforms the heuristic forecast in several regions and owns a strongly underdispersed ensemble at all lead times, underscoring its difficulties in capturing SPEI-3 interannual variability, as previously observed also by Calì Quaglia et al. (2022) for forecasts of temperature and precipitation anomalies. On the contrary, ECCC's good performance in FCRPSS and RH suggests the quality of its forecast ensemble, with potential gains from increasing ensemble size. The consistency in CMCC, MF, DWD, and ECCC RH shapes across lead time suggests that errors likely stem from the model's underlying climatology rather than issues specific to initialized runs. Therefore, tuning efforts should focus primarily on improving the representation of the model's mean state rather than targeting lead-time-specific corrections. The skill scores decline sharply after 0-month lead time but remain stable from lead times 1 to 3, despite some month-to-month fluctuations. Few regions show consistently good skills across lead times, such as the Iberian Peninsula and regions bordering the Black Sea, depending on the model. Previous studies have identified temperature anomalies in the Iberian Peninsula as predictable (Calì Quaglia et al. 2022; Frías et al. 2010), which may contribute to the predictability of SPEI-3 there. These seasonal forecast skills are likely influenced by large-scale teleconnections like ENSO (Doblas-Reyes et al. 2013; Ma et al. 2015), possibly making it a key factor in improving SPEI-3 forecasts. Frías et al. (2010) specifically showed that La Niña tends to drive predictability for hot events in summer along the Mediterranean coast. Improvements over the heuristic forecast are modest at longer lead times, making the use of seasonal forecasts questionable in low-predictability areas.

Our study demonstrates that seasonal drought forecasts using SPEI-3 provide valuable early warnings for the drought conditions in Europe, especially at shorter lead times. Our results broadly align with findings from Turco et al. (2017), who assessed summer SPEI-6 predictability in Europe using ECMWF SEAS4, with predictions augmented by merging forecasts with ERA5 data. They found that integrating observed data enhances forecast accuracy over extended lead times. Unlike Turco et al. (2017), our analysis relies solely on the forecast model outputs without such merging, employing a broader set of models and a different range of verification metrics. Although integration is particularly essential for improving forecast accuracy over such extended time scales, our analysis demonstrates the feasibility of good forecast accuracy when data are aggregated over a shorter time period.

Our findings can support existing hydrological forecasting systems that use seasonal climate information to drive hydrological models. For instance, Wanders et al. (2019) developed a global-scale hydrological forecasting system based on seasonal climate forecasts. While our analysis focuses on the predictability of meteorological drought conditions through SPEI, integrating such predictability insights with hydrological models could help assess the likelihood of drought propagation from meteorological to hydrological and agricultural impacts. Additionally, extending forecasts beyond meteorological drought to include soil moisture and hydrological indicators (Shyrokaya et al. 2025) could further bridge the gap between climate forecasts and sector-specific drought risks. This represents a promising direction for future research, particularly in regions where our results suggest significant seasonal predictability.

To further improve regional drought predictions, it is also crucial to understand how large-scale climate teleconnections influence local drought conditions. Persistent atmospheric circulation patterns, such as blocking events, have been shown to play a key role in the development of summer droughts and heat extremes (Hoskins and Woollings 2015; Pfahl and Wernli 2012; Schaller et al. 2018; Brunner et al. 2018; Sousa et al. 2018). Atmospheric blocking, in turn, is influenced by low-frequency variability modes, such as the summer NAO (Woollings et al. 2008; Athanasiadis et al. 2014), that hold predictive potential. The summer NAO also influences persistent temperature anomalies (Folland et al. 2009; Ossó et al. 2017) and could therefore enhance their predictability.

5. Conclusions

In conclusion, our work shows that dynamical seasonal forecasts from C3S offer moderate-to-good skill in anticipating summer SPEI-3 conditions, particularly in southern Europe. As drought indices are inherently sensitive to regional climate variability (Peres et al. 2023), future work should also consider the development of

tailored indicators that reflect the evolving precipitation and temperature regimes of the Euro-Mediterranean region. Integrating ensemble-based, multi-index approaches (such as combining SPI and SPEI) could better capture drought risk and support decision-making. By understanding the spatial and temporal sensitivity of each index to climate drivers and linking drought forecasts to broader climate patterns, seasonal drought predictions for the Euro-Mediterranean could become both more accurate and more skillful for risk management.

Acknowledgments. We acknowledge the Copernicus Climate Change Service (C3S) for providing the seasonal forecast data utilized in this study. The C3S is implemented by the European Centre for Medium-Range Weather Forecasts (ECMWF) on behalf of the European Union. We also thank the contributing modeling institutions (listed in Table 1 of this paper) for producing and making available their forecast datasets. This study was carried out within the RETURN Extended Partnership and received funding from the European Union Next-GenerationEU [National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.3 D.D. 1243 2/8/2022, PE0000005]. K. B. was supported by the "The Geosciences for Sustainable Development" project (Budget Ministero dell'Università e della Ricerca–Dipartimenti di Eccellenza 2023–27 C93C23002690001). The authors declare that they have no relevant financial or nonfinancial interests to disclose. J. V. H., G. C., and K. B. conceived the project. G. C. retrieved and analyzed the data. All authors contributed to the discussion and interpretation of the results. G. C. wrote the first draft of this manuscript, and all authors contributed to edit and finalize the manuscript. All authors read and approved the final manuscript. All code for data analysis associated with the current work can be requested by writing to the corresponding author.

Data availability statement. All data used in this study are publicly available and can be accessed through the Copernicus Climate Data Store (CDS) at https://cds.climate.copernicus.eu/datasets/seasonal-original-single-levels.

REFERENCES

Arnone, E., M. Cucchi, S. D. Gesso, M. Petitta, and S. Calmanti, 2020: Droughts prediction: A methodology based on climate seasonal forecasts. *Water Resour. Manage.*, **34**, 4313–4328, https://doi.org/10.1007/s11269-020-02623-3.

Athanasiadis, P. J., and Coauthors, 2014: The representation of atmospheric blocking and the associated low-frequency variability in two seasonal prediction systems. *J. Climate*, **27**, 9082–9100, https://doi.org/10.1175/JCLI-D-14-00291.1.

Bachmair, S., M. Tanguy, J. Hannaford, and K. Stahl, 2018: How well do meteorological indicators represent agricultural and forest drought across Europe? *Environ. Res. Lett.*, **13**, 034042, https://doi.org/10.1088/1748-9326/aaafda.

Battè, L., L. Dorec, C. Ardilouze, and J. F. Guérémy, 2021: Documentation of the METEO-FRANCE seasonal forecasting system 8. Copernicus Climate Change Service Tech. Rep., 118 pp.

Beguería, S., S. M. Vicente-Serrano, F. Reig, and B. Latorre, 2013: Standardized precipitation evapotranspiration index (SPEI) revisited: Parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. *Int. J. Climatol.*, **34**, 3001–3023, https://doi.org/10.1002/joc.3887.

Bradley, A. A., J. Demargne, and K. J. Franz, 2019: Attributes of forecast quality. *Handbook of Hydrometeorological Ensemble Forecasting*, Q. Duan et al., Eds., Springer, 849–892, https://doi.org/10.1007/978-3-642-39925-1_2.

Brands, S., and Coauthors, 2025: Seasonal drought predictions in the mediterranean using the SPEI index: Paving the way for their operational applicability in climate services. *Climate Services*, **38**, 100555, https://doi.org/10.1016/j.cliser.2025.100555.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2.

Brunner, L., N. Schaller, J. Anstey, J. Sillmann, and A. K. Steiner, 2018: Dependence of present and future European temperature extremes on the location of atmospheric blocking. *Geophys. Res. Lett.*, **45**, 6311–6320, https://doi.org/10.1029/2018GL077837.

Buontempo, C., and Coauthors, 2022: The Copernicus climate change service: Climate science in action. *Bull. Amer. Meteor. Soc.*, **103**, E2669–E2687, https://doi.org/10.1175/BAMS-D-21-0315.1.

Calì Quaglia, F., S. Terzago, and J. von Hardenberg, 2022: Temperature and precipitation seasonal forecasts over the Mediterranean region: Added value compared to simple forecasting methods. *Climate Dyn.*, **58**, 2167–2191, https://doi.org/10.1007/s00382-021-05895-6.

Cornes, R. C., G. Van Der Schrier, E. J. M. Van Den Besselaar, and P. D. Jones, 2018: An ensemble version of the E-OBS temperature and precipitation data sets. *J. Geophys. Res. Atmos.*, **123**, 9391–9409, https://doi.org/10.1029/2017JD028200.

C3S Climate Data Store, 2018: Seasonal forecast daily and subdaily data on single levels. CDS, accessed 16 June 2024, https://doi.org/10.24381/cds.181d637e.

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 Interdiscip. Rev.: Climate Change*, **4**, 245–268, https://doi.org/10.1002/wcc.217.

Dutra, E., F. Di Giuseppe, F. Wetterhall, and F. Pappenberger, 2013: Seasonal forecasts of droughts in African basins using the Standardized Precipitation Index. *Hydrol. Earth Syst. Sci.*, **17**, 2359–2373, https://doi.org/10.5194/hess-17-2359-2013.

——, and Coauthors, 2014: Global meteorological drought—Part 2: Seasonal forecasts. *Hydrol. Earth Syst. Sci.*, **18**, 2669–2678, https://doi.org/10.5194/hess-18-2669-2014.

Essa, Y. H., M. Hirschi, W. Thiery, A. M. El-Kenawy, and C. Yang, 2023: Drought characteristics in Mediterranean under future climate change. *npj Climate Atmos. Sci.*, **6**, 133, https://doi.org/10.1038/s41612-023-00458-4.

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. *Meteor. Appl.*, **15**, 19–24, https://doi.org/10.1002/met.45.

Folland, C. K., J. Knight, H. W. Linderholm, D. Fereday, S. Ineson, and J. W. Hurrel, 2009: The summer North Atlantic Oscillation: Past, present, and future. *J. Climate*, **22**, 1082–1103, https://doi.org/10.1175/2008JCLI2459.1.

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. *J. Climate*, **23**, 209–220, https://doi.org/10.1175/2009JCLI2824.1.

Fröhlich, K., and Coauthors, 2021: The German Climate Forecast System: GCFS. *J. Adv. Model. Earth Syst.*, **13**, e2020MS002101, https://doi.org/10.1029/2020MS002101.

Gualdi, S., and Coauthors, 2020: The new CMCC Operational Seasonal Prediction System. Fondazione CMCC Tech. Rep., 34 pp., https://www.cmcc.it/wp-content/uploads/2020/09/TN0288-csp-09-2020-1.pdf.

Hamill, T. M., and S. J. Colucci, 1997: Verification of eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327, https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

Hao, Z., V. P. Singh, and Y. Xia, 2018: Seasonal drought prediction: Advances, challenges, and future prospects. *Rev. Geophys.*, **56**, 108–141, https://doi.org/10.1002/2016RG000549.

Hargreaves, G. H., and Z. A. Samani, 1985: Reference crop evapotranspiration from temperature. *Appl. Eng. Agric.*, **1** (2), 96–99, https://doi.org/10.13031/2013.26773.

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

——, and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hoskins, B., and T. Woollings, 2015: Persistent extratropical regimes and climate extremes. *Curr. Climate Change Rep.*, **1**, 115–124, https://doi.org/10.1007/s40641-015-0020-8.

Ionita, M., and V. Nagavciuc, 2021: Changes in drought features at the European level over the last 120 years. *Nat. Hazards Earth Syst. Sci.*, **21**, 1685–1701, https://doi.org/10.5194/nhess-21-1685-2021.

IPCC, 2021: *Climate Change 2021: The Physical Science Basis.* V. Masson-Delmotte et al., Eds., Cambridge University Press, 2392 pp., https://doi.org/10.1017/9781009157896.

Johnson, S. J., and Coauthors, 2019: SEAS5: The new ECMWF seasonal forecast system. *Geosci. Model Dev.*, **12**, 1087–1117, https://doi.org/10.5194/gmd-12-1087-2019.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2011: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Wiley, 254 pp.

Keyantash, J., and J. A. Dracup, 2002: The quantification of drought: An evaluation of drought indices. *Bull. Amer. Meteor. Soc.*, **83**, 1167–1180, https://doi.org/10.1175/1520-0477-83.8.1167.

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. *J. Climate*, **13**, 4196–4216, https://doi.org/10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2.

Lavaysse, C., J. Vogt, and F. Pappenberger, 2015: Early warning of drought in Europe using the monthly ensemble system from ECMWF. *Hydrol. Earth Syst. Sci.*, **19**, 3273–3286, https://doi.org/10.5194/hess-19-3273-2015.

——, T. Stockdale, N. McCormick, and J. Vogt, 2020: Evaluation of a new precipitation-based index for global seasonal forecasting of unusually wet and dry periods. *Wea. Forecasting*, **35**, 1189–1202, https://doi.org/10.1175/WAF-D-19-0196.1.

Lin, H., and Coauthors, 2021: The Canadian Seasonal to Interannual Prediction System version 2 (CanSIPSv2). *Wea. Forecasting*, **35**, 1317–1343, https://doi.org/10.1175/WAF-D-19-0259.1.

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

Manzanas, R., V. Torralba, L. Lledó, and P. A. Bretonnière, 2022: On the reliability of global seasonal forecasts: Sensitivity to ensemble size, hindcast length and region definition. *Geophys. Res. Lett.*, **49**, e2021GL094662, https://doi.org/10.1029/2021GL094662.

Marcos, R., M. Turco, J. Bedia, M. Llasat, and A. Provenzale, 2015: Seasonal predictability of summer fires in a Mediterranean environment. *Int. J. Wildland Fire*, **24**, 1076–1084, https://doi.org/10.1071/WF15079.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713–725, https://doi.org/10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2.

McKee, T. B., N. J. Doesken, and J. Kleist, 1993: The relationship of drought frequency and duration to time scales. *Proc. Eighth Conf. on Applied Climatology*, Anaheim, CA, Amer. Meteor. Soc., 179–184, https://climate.colostate.edu/pdfs/relationshipofdroughtfrequency.pdf.

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

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

Ossó, A., R. Sutton, L. Shaffrey, and B. Dong, 2017: Observational evidence of European summer weather patterns predictable from spring. *Proc. Natl. Acad. Sci. USA*, **115**, 59–63, https://doi.org/10.1073/pnas.1713146114.

Peres, D. J., B. Bonaccorso, N. Palazzolo, A. Cancelliere, G. Mendicino, and A. Senatore, 2023: A dynamic approach for assessing climate change impacts on drought: An analysis in southern Italy. *Hydrol. Sci. J.*, **68**, 1213–1228, https://doi.org/10.1080/02626667.2023.2217332.

Pfahl, S., and H. Wernli, 2012: Quantifying the relevance of atmospheric blocking for co-located temperature extremes in the Northern Hemisphere on (sub-)daily time scales. *Geophys. Res. Lett.*, **39**, L12807, https://doi.org/10.1029/2012GL052261.

Portele, T. C., C. Lorenz, B. Dibrani, P. Laux, J. Bliefernicht, and H. Kunstmann, 2021: Seasonal forecasts offer economic benefit for hydrological decision making in semi-arid regions. *Sci. Rep.*, **11**, 10581, https://doi.org/10.1038/s41598-021-89564-y.

Pozzi, W., and Coauthors, 2013: Toward global drought early warning capability: Expanding international cooperation for the development of a framework for monitoring and forecasting. *Bull. Amer. Meteor. Soc.*, **94**, 776–785, https://doi.org/10.1175/BAMS-D-11-00176.1.

Sánchez-García, E., J. Voces Aboy, and E. Rodríguez Camino, 2018: Verification of six operational seasonal forecast systems over Europe and North Africa. AEMET Tech. Rep., 105 pp., https://seasonal.meteo.fr/sites/data/Documentation/doc_generale/AEMET_MEDCOF_model_verification.pdf.

——, and Coauthors, 2022: Co-design of sectoral climate services based on seasonal prediction information in the Mediterranean. *Climate Serv.*, **28**, 100337, https://doi.org/10.1016/j.cliser.2022.100337.

Schaller, N., J. Sillmann, J. Anstey, E. M. Fischer, C. M. Grams, and S. Russo, 2018: Influence of blocking on northern European and western Russian heatwaves in large climate model ensembles.

Environ. Res. Lett., **13**, 054015, https://doi.org/10.1088/1748-9326/aaba55.

Schumacher, D. L., and Coauthors, 2024: Detecting the human fingerprint in the summer 2022 western–central European soil drought. *Earth Syst. Dynam.*, **15**, 131–154, https://doi.org/10.5194/esd-15-131-2024.

Seneviratne, S. I., and Coauthors, 2012: Changes in climate extremes and their impacts on the natural physical environment. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*, C. B. Field et al., Eds., Cambridge University Press, 109–230, https://doi.org/10.1017/CBO9781139177245.006.

Shyrokaya, A., F. Pappenberger, G. Messori, I. Pechlivanidis, H. Cloke, and G. Di Baldassarre, 2025: How good is my drought index? Evaluating predictability and ability to estimate impacts across Europe. *Environ. Res. Lett.*, **20**, 034051, https://doi.org/10.1088/1748-9326/adb869.

Sousa, P. M., R. M. Trigo, D. Barriopedro, P. M. M. Soares, and J. A. Santos, 2018: European temperature responses to blocking and ridge regional patterns. *Climate Dyn.*, **50**, 457–477, https://doi.org/10.1007/s00382-017-3620-2.

Spinoni, J., G. Naumann, and J. V. Vogt, 2017: Pan-European seasonal trends and recent changes of drought frequency and severity. *Global Planet. Change*, **148**, 113–130, https://doi.org/10.1016/j.gloplacha.2016.11.013.

——, and Coauthors, 2019: A new global database of meteorological drought events from 1951 to 2016. *J. Hydrol.*, **22**, 100593, https://doi.org/10.1016/j.ejrh.2019.100593.

Stagge, J. H., L. M. Tallaksen, L. Gudmundsson, A. F. Van Loon, and K. Stahl, 2015: Candidate distributions for climatological drought indices (SPI AND SPEI). *Intl. J. Climatol.*, **35**, 4027–4040, https://doi.org/10.1002/joc.4267.

Svoboda, M., and Coauthors, 2002: The Drought Monitor. *Bull. Amer. Meteor. Soc.*, **83**, 1181–1190, https://doi.org/10.1175/1520-0477-83.8.1181.

Terzago, S., G. Bongiovanni, and J. von Hardenberg, 2023: Seasonal forecasting of snow resources at Alpine sites. *Hydrol. Earth Syst. Sci.*, **27**, 519–542, https://doi.org/10.5194/hess-27-519-2023.

Teuling, A. J., and Coauthors, 2013: Evapotranspiration amplifies European summer drought. *Geophys. Res. Lett.*, **40**, 2071–2075, https://doi.org/10.1002/grl.50495.

Torres-Vázquez, M. Á., and Coauthors, 2024: Probabilistic predictions for meteorological droughts based on multi-initial conditions. *J. Hydrol.*, **640**, 131662, https://doi.org/10.1016/j.jhydrol.2024.131662.

Tramblay, Y., and Coauthors, 2020: Challenges for drought assessment in the Mediterranean region under future climate scenarios. *Earth. Sci. Rev.*, **210**, 103348, https://doi.org/10.1016/j.earscirev.2020.103348.

Tuel, A., and E. A. B. Eltahir, 2020: Why is the Mediterranean a climate change hot spot? *J. Climate*, **33**, 5829–5843, https://doi.org/10.1175/JCLI-D-19-0910.1.

Turco, M., A. Ceglar, C. Prodhomme, A. Soret, A. Toreti, and J.-D.-R. Francisco, 2017: Summer drought predictability over Europe: Empirical versus dynamical forecasts. *Environ. Res. Lett.*, **12**, 084006, https://doi.org/10.1088/1748-9326/aa7859.

Van Loon, A. F., 2015: Hydrological drought explained. *Wiley Interdiscip. Rev.: Water*, **2**, 359–392, https://doi.org/10.1002/wat2.1085.

——, and H. A. J. Van Lanen, 2012: A process-based typology of hydrological drought. *Hydrol. Earth Syst. Sci.*, **16**, 1915–1946, https://doi.org/10.5194/hess-16-1915-2012.

Vicente-Serrano, S. M., S. Beguería, and J.-I. López-Moreno, 2010: A multiscalar drought index sensitive to global warming: The standardized precipitation evapotranspiration index. *J. Climate*, **23**, 1696–1718, https://doi.org/10.1175/2009JCLI2909.1.

——, and Coauthors, 2012: Performance of drought indices for ecological, agricultural, and hydrological applications. *Earth Interact.*, **16** (1), 1–27, https://doi.org/10.1175/2012EI000434.1.

——, and Coauthors, 2014: Evidence of increasing drought severity caused by temperature rise in southern Europe. *Environ. Res. Lett.*, **9**, 044001, https://doi.org/10.1088/1748-9326/9/4/044001.

Vitart, F., 2004: Monthly forecasting at ECMWF. *Mon. Wea. Rev.*, **132**, 2761–2779, https://doi.org/10.1175/MWR2826.1.

Wanders, N., S. Thober, R. Kumar, M. Pan, J. Sheffield, L. Samaniego, and E. F. Wood, 2019: Development and evaluation of a pan-European multimodel seasonal hydrological forecasting system. *J. Hydrometeor.*, **20**, 99–115, https://doi.org/10.1175/JHM-D-18-0040.1.

Weisheimer, A., and T. N. Palmer, 2014: On the reliability of seasonal climate forecasts. *J. Roy. Soc. Interface*, **11**, 20131162, https://doi.org/10.1098/rsif.2013.1162.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 704 pp.

Williams, K. D., and Coauthors, 2018: The Met Office Global Coupled Model 3.0 and 3.1 (GC3.0 and GC3.1) configurations. *J. Adv. Model. Earth Syst.*, **10**, 357–380, https://doi.org/10.1002/2017MS001115.

Woollings, T., B. Hoskins, M. Blackburn, and P. Berrisford, 2008: A new Rossby wave–breaking interpretation of the North Atlantic Oscillation. *J. Atmos. Sci.*, **65**, 609–626, https://doi.org/10.1175/2007JAS2347.1.

Yuan, X., E. F. Wood, N. W. Chaney, J. Sheffield, J. Kam, M. Liang, and K. Guan, 2013: Probabilistic seasonal forecasting of African drought by dynamical models. *J. Hydrometeor.*, **14**, 1706–1720, https://doi.org/10.1175/JHM-D-13-054.1.

Zellou, B., N. El Moçayd, and E. H. Bergou, 2023: Towards improved drought prediction in the Mediterranean region—Modeling approaches and future directions. *Nat. Hazards Earth Syst. Sci.*, **23**, 3543–3583, https://doi.org/10.5194/nhess-23-3543-2023.