

ARTICLE OPEN



Multi-annual prediction of drought and heat stress to support decision making in the wheat sector

Balakrishnan Solaraju-Murali ¹✉, Nube Gonzalez-Reviriego¹, Louis-Philippe Caron ^{1,2}, Andrej Ceglar ³, Andrea Toreti ³, Matteo Zampieri ³, Pierre-Antoine Bretonnière¹, Margarita Samsó Cabré¹ and Francisco J. Doblas-Reyes ^{1,4}

Drought and heat stress affect global wheat production and food security. Since these climate hazards are expected to increase in frequency and intensity due to anthropogenic climate change, there is a growing need for effective planning and adaptive actions at all timescales relevant to the stakeholders and users in this sector. This work aims at assessing the forecast quality in predicting the evolution of drought and heat stress by using user-relevant agro-climatic indices such as Standardized Precipitation Evapotranspiration Index (SPEI) and Heat Magnitude Day Index (HMDI) on a multi-annual timescale, as this time horizon coincides with the long-term strategic planning of stakeholders in the wheat sector. We present the probabilistic skill and reliability of initialized decadal forecast to predict these indices for the months preceding the wheat harvest on a global spatial scale. The results reveal the usefulness of the study in a climate services context while showing that decadal climate forecasts are skillful and reliable over several wheat harvesting regions.

npj Climate and Atmospheric Science (2021)4:34; <https://doi.org/10.1038/s41612-021-00189-4>

INTRODUCTION

Near-term decadal climate predictions represent an important source of information with the potential to support decision-making and strengthen the resilience of a wide range of socio-economic sectors that are heavily influenced by climate variability, change, and extremes¹. For example, agriculture and climate-sensitive energy sectors, such as hydroelectricity, wind, and solar energies, can greatly benefit from this type of information^{2,3}. However, up to now, only a few studies^{4–7} have explored the usability of these readily available near-term decadal climate forecasts for specific sectoral needs.

The entire agricultural sector (food-producing systems, markets, etc.) is influenced by climate variability, unfavorable weather, and climate extremes (e.g., droughts and heatwaves^{8,9}). In this context, climate risk management¹⁰ for capacity adaptation, planning, risk-mitigation, and early response to climate extremes can benefit from decadal climate predictions. The advantage of using these predictions compared to climate projections relies on the temporal coherence of the main large scale climate patterns with observations¹¹ and the possibility to assess their quality by evaluating the performance of retrospective decadal predictions against observations over the recent historical period¹. The latter aspect facilitates a retrospective risk assessment of recent past droughts and heatwaves relevant for the agricultural sector, improving the communication of the risks associated with those extreme events for informed decision-making.

Climate information at decadal timescale holds the potential for being of great value for a broad range of users in the agricultural sector, especially in its producing component, who are affected by climate variability and who would benefit from additional sources of information. Wheat is the most harvested cereal crop in the world (harvested area in 2019; www.fao.org/faostat), contributing to the local food security of many countries. Unfavorable conditions and extreme weather such as heat or water stress affect both grain yield and quality, especially when they occur in

the most sensitive development stages of the plant^{12,13}. For instance, heat stress occurring during the flowering stage affects floret fertility causing a reduction of the grain number¹⁴, whereas heat stress that occurs during the grain filling stage reduces grain quality and size¹⁵. Similarly, drought conditions can affect either grain number, when the stress occurs before or during the flowering stage, or grain size, when the drought occurs after the flowering stage due to its impact on grain filling¹⁶.

In the framework of the EU H2020-MED-GOLD project, the durum wheat community has been engaged in the co-design of innovative sectoral climate services based on specific users' needs and opportunities. This process has revealed some key decision areas that could be improved by having reliable decadal climate information available well in advance, such as risk estimates for the global market, investments in agro-management infrastructure and tools, planning of supply chain contracts or targeted breeding programmes³. Strategic tools in policies such as the EU Common Agricultural Policy, designed to address EU climate ambition, may also benefit from initialized decadal climate predictions.

Recent advances in the understanding and forecasting of the climate have resulted in skillful decadal climate predictions^{1,11,17}. Nevertheless, for the provision of an end-to-end climate service focused on some of the above-mentioned specific needs of the wheat sector, efforts are still needed to convert available climate predictions into usable information by users. This study aims to address this necessary step for the provision of a climate service targeted to the wheat sector based on decadal predictions, including the estimation of specific indices accounting for both drought and heat stresses, their bias adjustment and their forecast quality assessment. Thus, the main objective of this paper is to assess the skill and reliability in predicting the Standardized Precipitation Evapotranspiration Index (SPEI¹⁸) and the Heat Magnitude Day Index (HMDI⁹), two indices that have been associated with wheat yield⁹, on a multi-annual timescale over

¹Barcelona Supercomputing Center (BSC), Barcelona, Spain. ²Ouranos, 550 Sherbrooke St W, Montreal, QC, Canada. ³European Commission, Joint Research Centre, Ispra, Italy. ⁴Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. ✉email: balakrishnan.solaraju@bsc.es

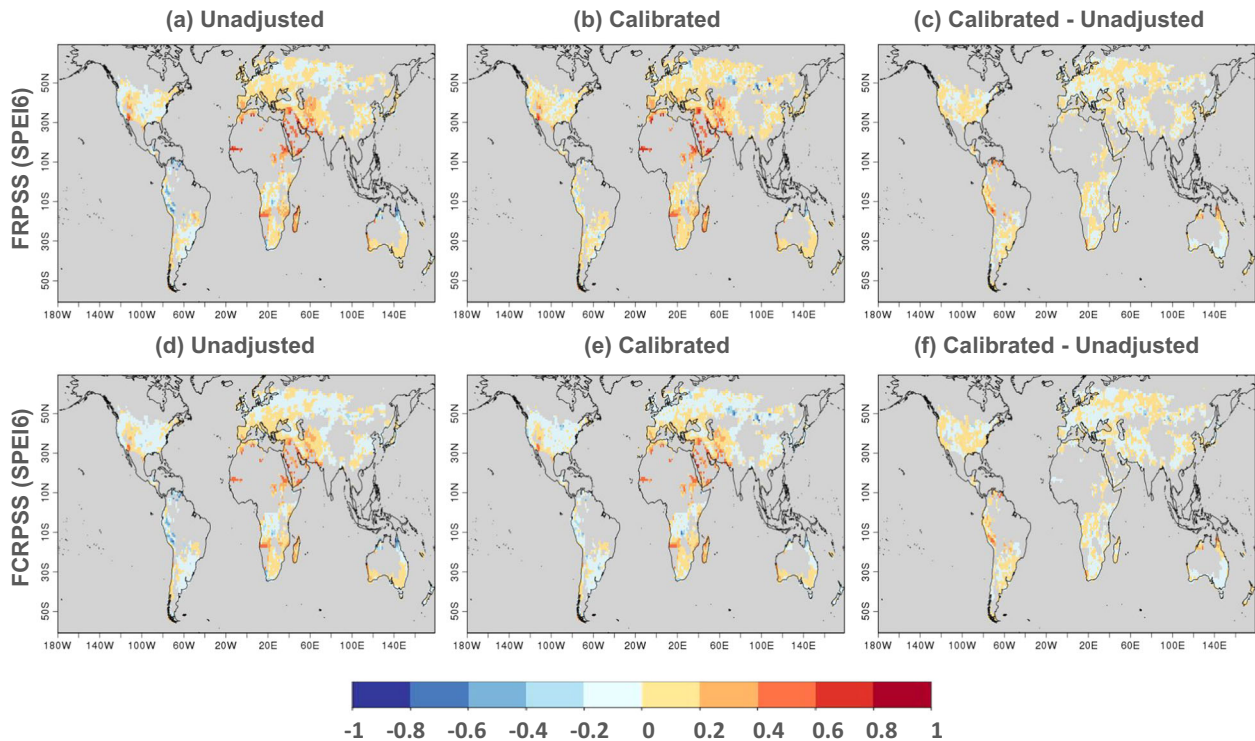


Fig. 1 Predictive skill of the drought index. Fair Ranked Probability Skill Score (FRPSS) for tercile events (top) and Fair Continuous Ranked Probability Skill Score (FCRPSS) for the full probability distribution (bottom) of unadjusted (**a, d**) and calibrated (**b, e**) SPEI6 forecast averaged over forecast years 1–5, with respect to the reference climatology during the wheat harvest months for the period 1961–2018. **c, f** presents the difference between calibrated and unadjusted forecast.

the entire wheat harvest areas. Following this, we present two tailored potential climate service products that can be effectively used in the wheat sector.

RESULTS

Forecast quality assessment of drought index

A fundamental step in the provision of a climate service based on predictions is the assessment of the skill and reliability associated with the climate information used for the service. This step allows both providers and users to better understand the potential value held by such climate information. Here, we present the probabilistic skill (Fig. 1) and reliability (Fig. 2) in predicting the multi-annual drought events using a soil-moisture proxy agro-climatic indicator (SPEI6) over the wheat harvesting regions. The forecast skill is investigated using the fair ranked probability skill score (FRPSS) and the fair continuous ranked probability skill score (FCRPSS), while the reliability is assessed using reliability diagrams (see “Methods” section). FRPSS and FCRPSS provide a measure of the skill of decadal forecasts in predicting the probabilities of the categorical events (tercile in our case) and the full probability distribution of the estimated index, respectively. Here, we compare the skill and reliability of an unadjusted and calibrated index obtained with the decadal forecast.

SPEI6 is the standardized value of climate water balance accumulated over the six months prior to the wheat harvesting month at each grid point (Supplementary Fig. 1). The climate water balance is defined as the difference between monthly total precipitation and potential evapotranspiration (PET); the latter is estimated here with the Thornthwaite approach¹⁹. The optimal approach for PET estimation still remains an open question and depends on several factors, such as the spatio-temporal scale involved, the aim of the study and the availability of meteorological data. A number of studies^{20,21} have pointed out that

different choices could lead to varied estimates of drought frequency and severity over several regions globally. Here, the thornthwaite approach has been chosen for its simplicity. However, recognizing its limitations, we have also tested two alternative approaches, namely the Hargreaves²² and the modified Hargreaves²³ methods. The comparison of these three approaches shows that our results are rather insensitive to the choice of the methodology used to estimate PET (Supplementary Fig. 3). The influence of other advanced data-intensive techniques of PET estimation such as FAO-56 Penman-Monteith will be explored in future studies.

The results obtained with the unadjusted SPEI6 predictions show a positive value of FRPSS over most of the wheat-growing regions, except for a few areas over Africa and both North and South America (Fig. 1a); whereas FCRPSS (Fig. 1d) presents low skill scores over most of the regions except for the Iberian Peninsula, South Africa, Western United States, Australia and the Middle East. For the unadjusted forecasts, FRPSS tends to give better results than FCRPSS as it involves implicit bias-adjustment in its definition since the errors in the climatological variance are partly corrected while setting the thresholds for the terciles individually for the observation and the forecasts. On the other hand, the widespread poor estimates of FCRPSS are the result of uncorrected errors present in the climatological variance since no thresholds are defined for computing this skill measure.

For the near-term probabilistic decadal predictions to be useful, systematic errors of the forecasting system have to be adjusted. To do this, we apply the variance inflation technique introduced in the “Methods” section. This approach ensures an increased reliability of the probability predictions by adjusting the inter-annual variance of the forecast system to that of the observational reference in cross-validation mode. Figure 1e shows that the FCRPSS increases after calibration, particularly over the areas where negative values were found in the unadjusted estimate of

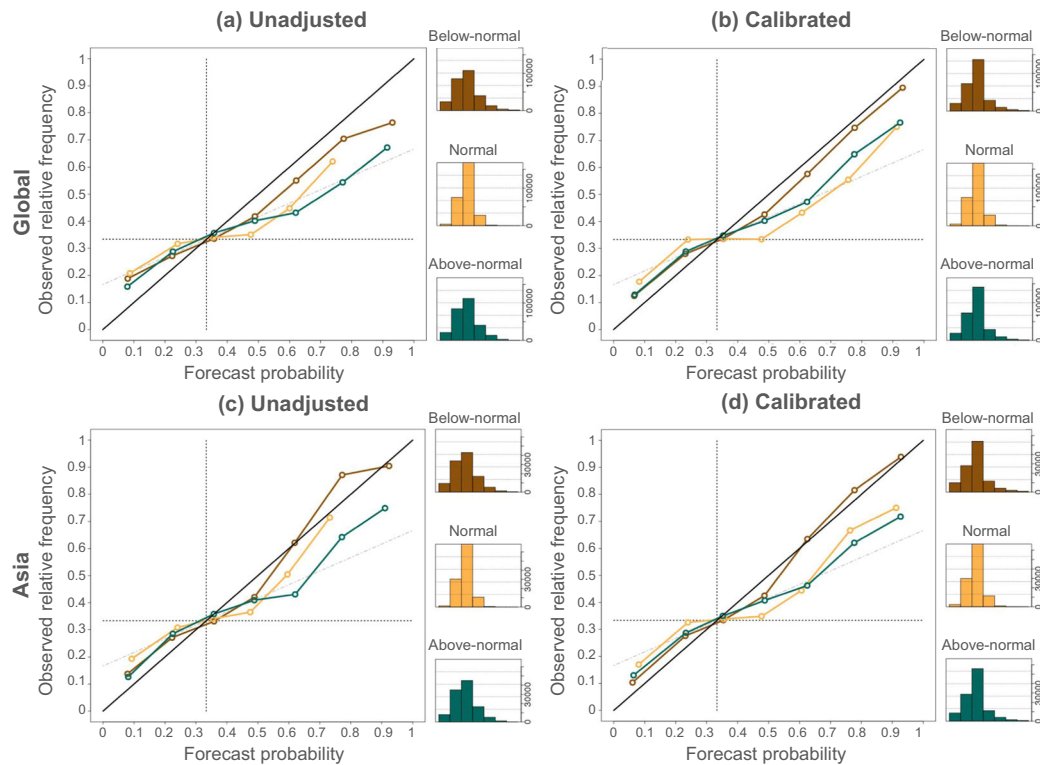


Fig. 2 Reliability of the drought index. Reliability diagrams (lines) for probabilistic categorical forecasts (tercile events) of unadjusted SPEI6 estimate (a, c) and for the calibrated index (b, d) over Global spatial domain (top) and Asia (bottom). The frequency of occurrence of each bin for each category is shown to the right of the panels.

the indices. The difference between the FRPSS and FCRPSS for the initialized decadal forecast before and after the calibration of the SPEI6 forecasts is presented in Fig. 1c and f, respectively. In general, there is an increase in FRPSS values over several regions, except for certain areas such as Eastern Europe and the Middle East (Fig. 1c). However, the absolute value of the FRPSS (Fig. 1b) over those regions with lower scores remains similar to the unadjusted forecast.

To understand the reason behind the skill of SPEI6 over the wheat harvesting regions, we compared the forecast skill of the climate variables used to compute the index (six-month accumulated potential evapotranspiration and precipitation; hereafter PET6 and P6, respectively) and investigate the influence of these climate variables on SPEI6. From this assessment, it emerges that most of the predictive skill in forecasting SPEI6 is linked to the following two factors: (1) the high predictive skill exhibited by the decadal forecast system in predicting the individual climate variables (see Supplementary Fig. 15a and b) and (2) the forecast system's ability to correctly capture the influence of PET6 and P6 on the index (compare Supplementary Fig. 16b, e and 16a, d). These results are in line with the conclusion presented in Solaraju-Murali et al.⁷. A detailed explanation on the contribution of the climate variables to the overall skill of estimated indices can be found in the Supplementary material.

The reliability diagrams (Fig. 2a and b) present the reliability of the predicted probabilities of tercile categories over the wheat harvesting regions globally. Reliability diagrams are particularly useful to identify whether a forecasting system under- or over-forecasts any particular category. For a forecast to be reliable, the averaged forecast probabilities should match the observed relative frequencies; thus, the colored lines in the graph should fall close to the diagonal. For SPEI6, the brown (green) line, corresponding to below-normal (above-normal) category, demonstrates the reliability in predicting the dryness (wetness) event by

the forecast system. The unadjusted predictions (Fig. 2a) provide a reliable forecast for the below-normal and above-normal categories. The calibration improved the reliability of the forecast for both below and above-normal categories (Fig. 2b).

To further study the reliability on a regional level, the global domain is divided into six different regions: North America, South America, Europe, Africa, Asia and Australia (see Supplementary Fig. 6). For clarity purposes, we choose to focus on Asia, as China and India are the global leaders in wheat production (e.g., Fig. 1 in Toreti et al.²⁴). The results for the other regions can be found in Supplementary Figs. 7 and 8. The reliability estimates for all three categories run along the diagonal (Fig. 2c). In this case, the reliability curves of the calibrated forecast show features that are similar to those of the unadjusted forecasts (Fig. 2d). This stands in contrast to the other regions for which the calibration improves the reliability, particularly for below- and above-normal categories (Supplementary Fig. 8). The forecast is generally unreliable for the normal category except for Asia and Africa (Supplementary Fig. 7), meaning that there is no clear relationship between the forecast probabilities and the frequency of the observed normal SPEI6 values over most regions. The calibration does not seem to improve the reliability for this category (Supplementary Fig. 8). This result is not entirely surprising since it has been shown that predicting the normal category tends to be more difficult as such events are known to have weak driving signals²⁵.

Forecast quality assessment of heat stress index

While the impact of drought is substantial in several countries producing wheat, Zampieri et al.⁹ have also shown that the heat stress presents a comparable or stronger negative influence on the wheat yield globally. To explore the skill in predicting heat stress events at the multi-annual timescale, we evaluate the HMDI3 index during the stages of wheat growth that are most

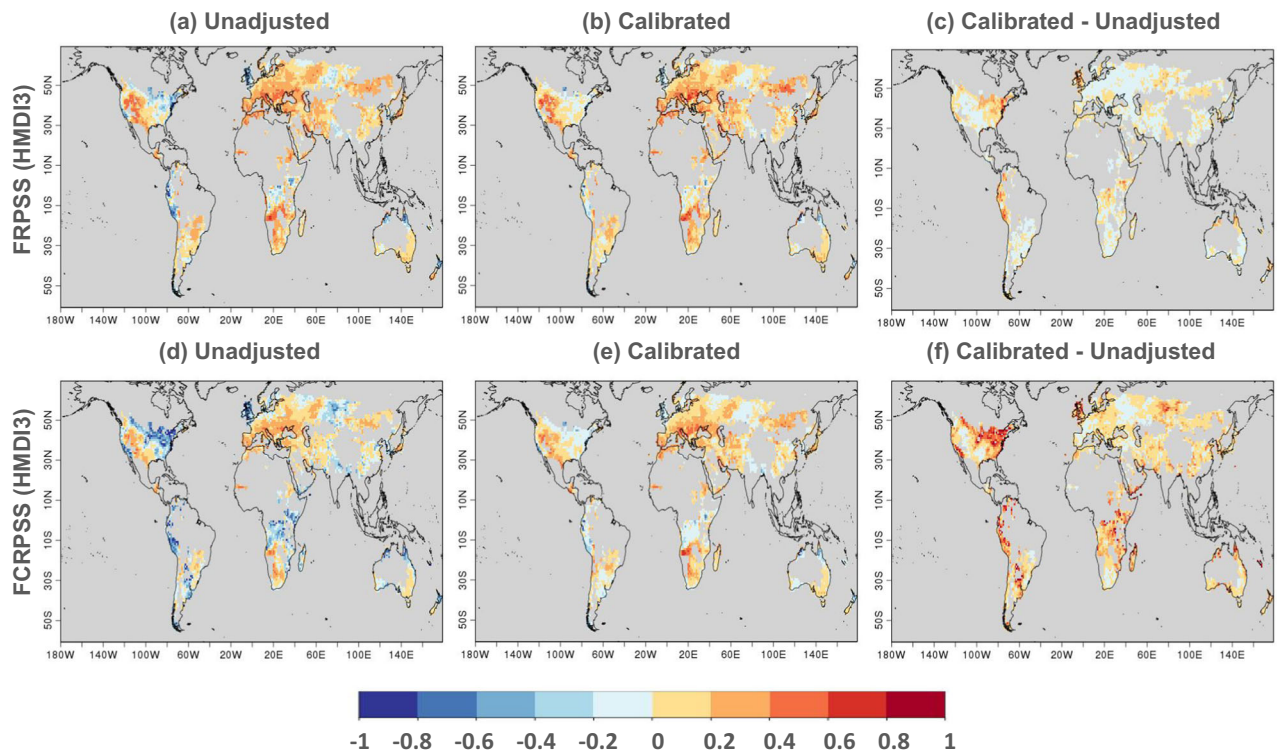


Fig. 3 Predictive skill of the heat stress index. Same as Fig. 1 but for the Heat Magnitude Day index (HMDI3).

sensitive to heat stress (i.e., 3 months before harvesting, roughly corresponding to the period including flowering and grain filling).

Maps of FRPSS and FCRPSS for HMDI3 are shown in Fig. 3. The best scores are achieved over Europe, the Western United States and South Africa with unadjusted forecasts (Fig. 3a and d). On the other hand, negative scores are found over the Eastern United States, South America and Angola. The calibrated forecast displays similar FRPSS patterns as the uncorrected forecast (compare Fig. 3a and b), but shows an increase over the Eastern United States, Ecuador, Peru, Columbia, the United Kingdom, part of Africa and Northern Australia, and with a slight decrease over the other regions (Fig. 3c). On the other hand, FCRPSS increased globally after applying the calibration (Fig. 3f). The improvement is considerably higher over the areas where the unadjusted forecast has high negative score values (compare Fig. 3d and f). The skill at predicting HMDI3 is associated with the decadal forecast systems ability to predict the maximum temperature during 3 months prior to wheat harvesting (Supplementary Fig. 15c).

The reliability diagram for HMDI3 over both global and Asian wheat-growing regions is presented in Fig. 4. Prior to the calibration (Fig. 4a and c), the below- and above-normal categories for both the domains are found to be reliable, whereas the normal category is under-confident when the forecast provides high probability estimates. The calibration seems to improve the reliability for all the categories (Fig. 4b and d). The results of the calibrated HMDI3 forecasts are similar for the rest of the regions considered in this study (Supplementary Fig. 10).

Impact of initialization

We further evaluated the decadal predictions against an ensemble of uninitialized historical simulations having the same number of ensemble members in order to explore the added value of initialization. The historical simulations are primarily intended to estimate the forced response of the climate system. For assessing the improvement in the predictive skill of decadal forecasts, the skill metrics are estimated by using the historical simulations as the

baseline forecast instead of the climatology. As for SPEI6, the decadal forecast presents positive skill (thus an improvement with respect to non-initialized forecasts) over most of the wheat-producing region, particularly over Australia, Central Europe, South Africa and Eastern United States (Fig. 5a, c and Supplementary Fig. 11). As for HMDI3, the improvement is limited to fewer regions, namely Western and Central United States, South America, South Africa, and southern Australia (Fig. 5b, d and Supplementary Fig. 12). The reliability of both indices was found to have slightly improved over the uninitialized simulation, particularly for Europe and Australia (compare Supplementary Figs. 13 and 14 with Supplementary Figs. 8 and 10).

The origin of the skill was also investigated by looking at the climate variables used to build the indices. The results (Supplementary Fig. 15d–f) point to an increase in skill in the initialized climate simulations over the regions for which we found an improvement of the decadal prediction skill in Fig. 5. In addition, the influence of PET6 and P6 on the estimated SPEI6 is evaluated by comparing historical simulations (Supplementary Fig. 16c and 16f) with observations and decadal predictions. Results demonstrate that over certain regions (for example, Central Europe, Eastern United states, Peru, Australia and the United Kingdom), the historical simulations reproduce that relationship worse than decadal predictions (Supplementary Fig. 16b and 16e) when compared with the observations (Supplementary Fig. 16a and 16d).

Application of decadal prediction for the wheat sector

An illustration of a potential climate service product for the wheat sector based on categorical events (terciles, in our case) is presented in Fig. 6. It shows the calibrated forecast of multi-annual averaged SPEI6 (Fig. 6a) and HMDI3 (Fig. 6b) for the hindcast period 2014–2018 initialized in November 2013 along with the corresponding observational values (Fig. 6c, d). In addition, we display the time series of the multi-annual predictions produced for each period and the correspondence between predictions and reanalysis over three durum wheat harvesting areas in Italy: Jesi

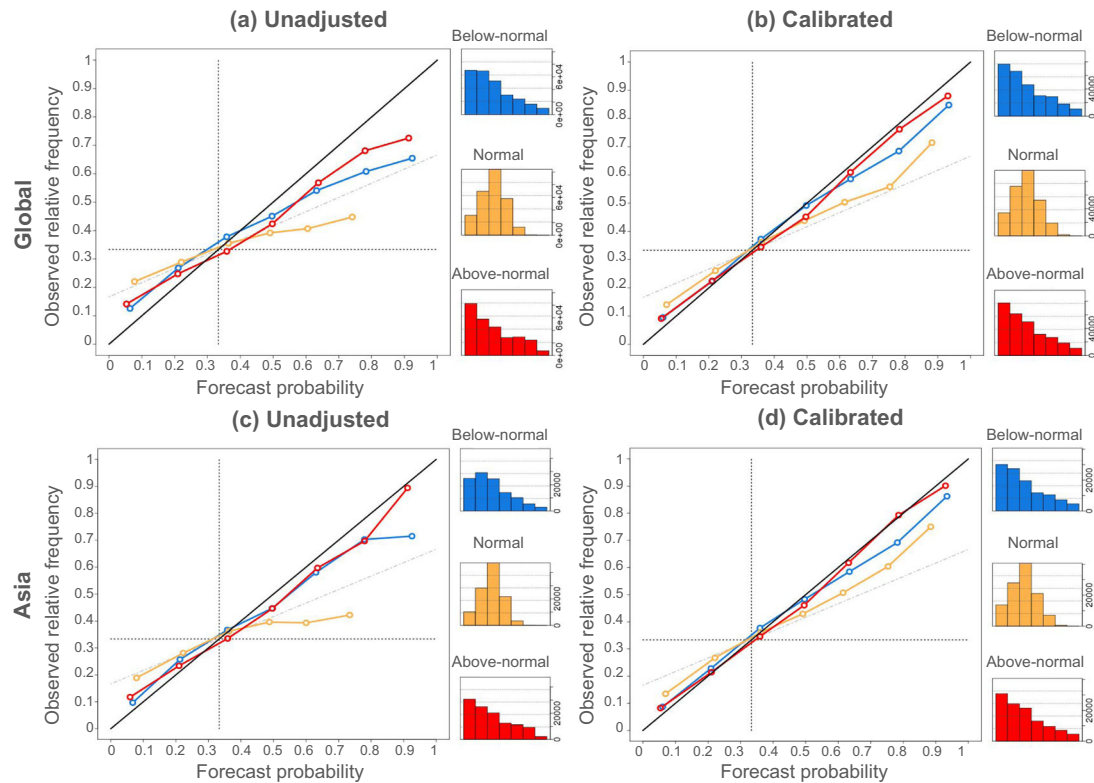


Fig. 4 Reliability of the heat stress index. Same as Fig. 2 but for the Heat Magnitude Day Index index (HMDI3).

(Fig. 6e, f), Ravenna (Supplementary Fig. 17a, b) and Foggia (Supplementary Fig. 17c, d). The three locations are selected by choosing the grid point in which each of the region falls. These areas were identified by the stakeholders in the EU H2020-MED-GOLD project to be among the leading regions used for the production of durum wheat in Italy and have also become, along with rest of the Mediterranean regions, prone to increasing temperatures and more frequent and intense extreme events such as droughts and heatwaves³. Figures 6a–d presents the predicted likelihood map (in %) of the most likely tercile (labeled as below-normal, normal and above-normal) of observed drought and heatwave occurrences using the indices corresponding to the wheat harvesting season (Supplementary Fig. 1) and for the regions where calibrated forecasts return positive FRPSS. Darker shades of brown (blue), yellow and green (red) correspond to a higher probability of occurrence of the event falling into the lower, middle, and upper tercile category of SPEI6 (HMDI3), respectively. The decadal predictions for both indices show an increase in drought (below-normal category) and heatwave (above-normal category) events over most of the wheat-growing regions prior to the harvest months for the period 2014–2018. Comparing Fig. 6a, c with Fig. 6b, d, we found a good agreement between forecasted and observed SPEI6 and HMDI3.

Multi-annual calibrated forecasts using decadal prediction (in gray) for SPEI6 and HMDI3 along with the observed values (in black) for the years 1961–2013 are presented in Fig. 6e and 6f over Jesi, Italy. It communicates probabilistic decadal forecasts in an intuitive way by issuing probabilities using the available ensemble member forecasts. Regarding SPEI6, we found the forecast to be able to predict the transition between the wet and dry conditions and to capture the slow variability of the observed drought conditions over Jesi (Fig. 6e). On the other hand, HMDI3 tends to show an increase in the heat stress events from the early 1990's (the positive trend in Fig. 6f, with more events at the above-normal category), as found in the observation.

The horizontal brown (blue) and green (red) lines represent the lower and upper tercile limits, which are the 33 and 66 percentiles of the climatological distribution, of SPEI6 (HMDI3). After defining the limits, an estimation of the forecast probabilities of each tercile category, presented in the last row for the forecast years 2014–2018, is carried out by counting the number of ensemble members falling into each category and dividing the count of each category with the total number of ensemble members generated. For instance, for the forecast years 2014–2018 in Fig. 6e, we can see that 28, 12, and 0 out of 40 members fell in the below-normal, normal and above-normal category respectively and hence, a 70, 30, and 0 percent probability event is assigned to the corresponding category for the considered forecast year. The time series and the skill scores for Ravenna and Foggia are presented in Supplementary Fig. 17. In both cases, the decadal forecast performs better than a climatological forecast for both SPEI6 and HMDI3, as evidenced by the positive skill scores.

DISCUSSION

Decadal climate predictions have been made available for users as a potential source of near-term climate information with the aim of supporting strategic decisions in key socio-economic sectors such as agriculture^{11,26}. The skill assessment performed here reveals that calibrated forecasts are more reliable in predicting SPEI6 and HMDI3 indices than unadjusted forecasts, which confirms that the calibration step is a necessary condition for using decadal climate information in action-oriented climate services. This is in line with previous studies that stressed the need for calibration to provide trustworthy and robust predictions^{27,28}.

Calibrated decadal predictions are generally more skillful than forecasts built on the observed past climatology in predicting the probabilities of the tercile categories and the full probability distribution of the multi-year averages of drought and heat stress indices over several wheat-growing regions. In addition, the initialization of the decadal forecasting system contributes to

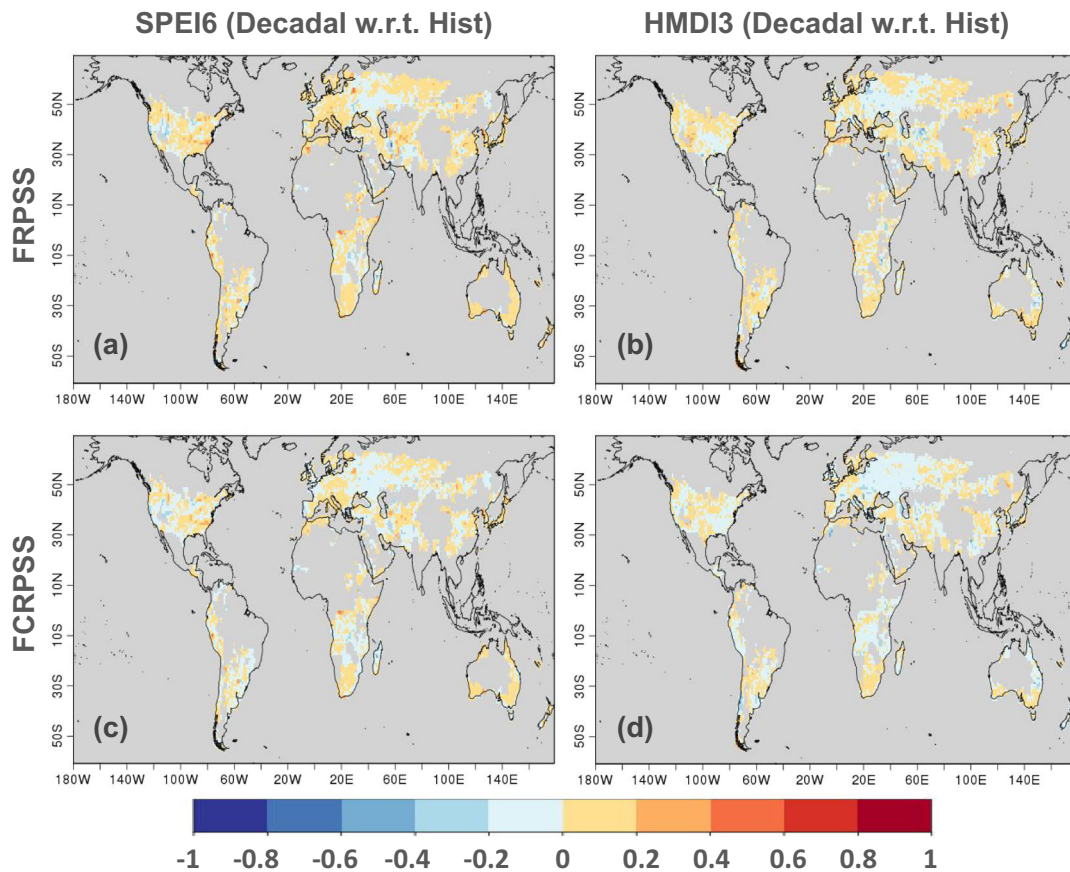


Fig. 5 Impact of initialization. Fair Ranked Probability Skill Score (FRPSS) for tercile events (top) and Fair Continuous Ranked Probability Skill Score (FCRPSS) for the full probability distribution (bottom) of calibrated SPEI6 (a, c) and HMDI3 (b, d) decadal forecast (Decadal) averaged over forecast years 1–5, with respect to the calibrated historical simulation (Hist) during the wheat harvest months for the period 1961–2018.

enhanced skill and reliability of the predicted drought and heat stress over the wheat producing regions. These encouraging results demonstrate the usefulness of decadal predictions and the possibility of using such predictions in climate services supporting risk reduction and adaptation strategies in the wheat sector. This implies that there exist opportunities to support wheat stakeholders in their decision-making processes, policies development, implementation and evaluation on a multi-annual timescale. The proposed methodology can be easily adapted to other crops, as for example maize and rice, or to other sectors where water management plays a fundamental role.

While the results presented here are promising, it is important to note that this assessment is based on a single decadal forecast system. Several studies have pointed out the need for large ensembles to achieve more reliable and skillful forecasts^{1,29}. To address this, a multi-model study with all the available state-of-the-art decadal forecasting systems will be explored in future analyses. In addition, we plan to investigate the impact of the slow, internally generated natural climate oscillations such as the Atlantic Multi-decadal variability (AMV) and the Pacific Decadal Oscillation (PDO) on the predicted indices used in this assessment, to determine whether the predictability of drought and heat stress can be linked to large-scale oceanic teleconnection patterns acting on multi-annual timescales.

Further efforts should also be devoted to the co-design of climate services, involving key users and relevant actors of the wheat sector, based on the obtained results. The co-design process should be oriented toward several important aspects such as: (i) the refinement of the products for fulfilling the particular user needs; (ii) effective visualization of the products to better

convey the information; (iii) dissemination material that discloses the main points regarding decadal predictions and climate service products in an easy and clear way. Efforts in properly tackling all these aspects will help the uptake of decadal information by stakeholders.

Notwithstanding future research still remaining, this assessment should be considered an essential first step in the development of climate services for agriculture, revealing that decadal climate forecasts are skillful and reliable over several wheat harvesting regions. Decadal climate predictions can already become a key element of agro-climate services in these regions to support informed decisions for a variety of end-users such as crop breeders, farmers, national and regional policymakers. Farmers can benefit from climate information on decadal timescale when it comes to decisions on investment in irrigation infrastructure, drainage systems, crop rotation and diversification strategies. Decadal climate predictions can help crop breeders to identify priorities in their breeding programs, such as drought tolerance and tolerance to high temperatures while looking for optimal responses (in terms of quantity and quality) to elevated atmospheric CO₂ concentration³⁰. Reliable decadal climate predictions have been also envisaged to anticipate the spread of new pests, pathogens and weeds and ultimately, in order to identify future suitable cultivation areas (see Fig. 3-3 in Manstretta et al.³¹). Given also the relevance of the decadal timescale for policy making and legislation, the results presented are of capital importance to provide a term of reference for the accuracy that can be achieved with state-of-the-art climate models.

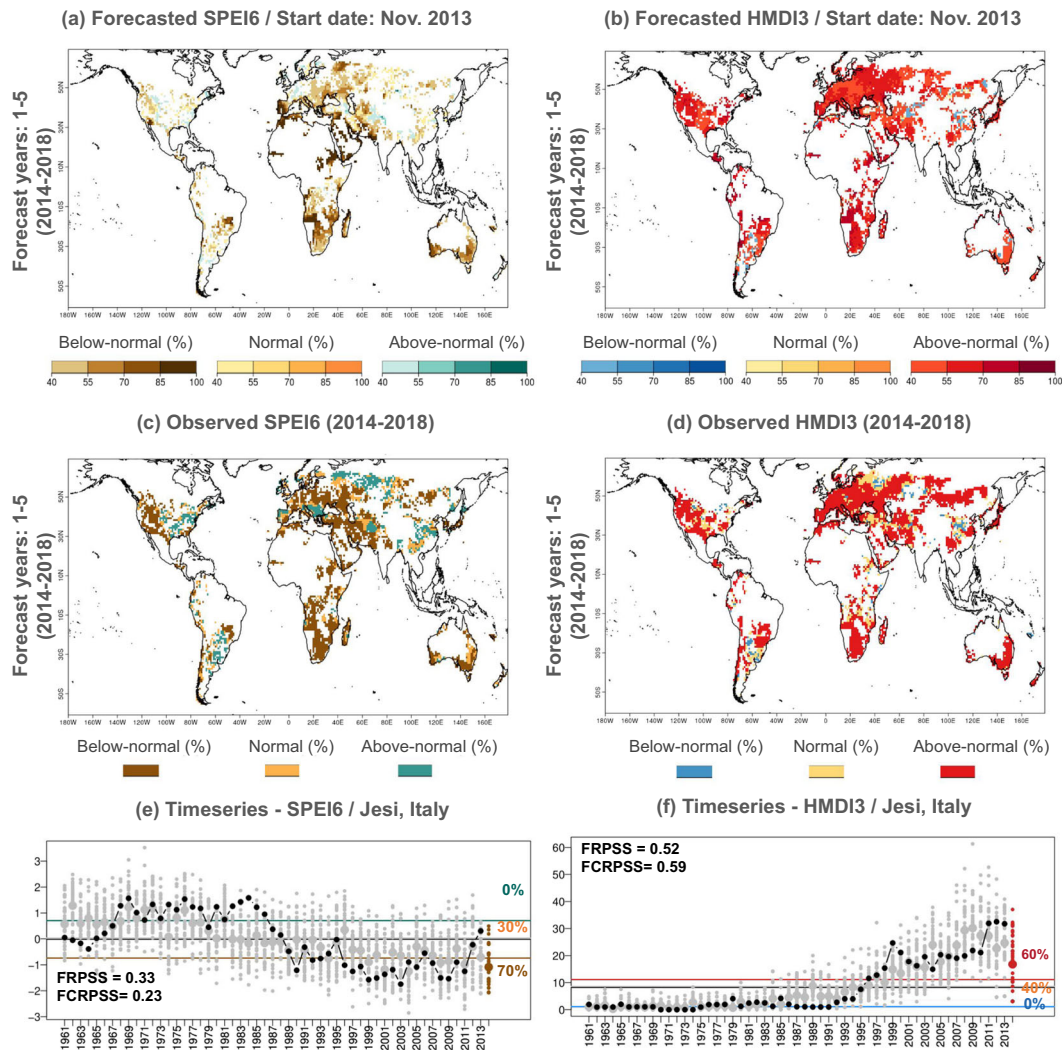


Fig. 6 Climate service product for the wheat sector. Multi-year probabilistic calibrated forecast (a, b) and observed (c, d) most likely tercile category of SPEI6 (left) and HMDI3 (right) for the year 2014–2018 over the wheat harvesting regions that presents positive FRPSS values. Time series of multi-year averaged SPEI6 (e) and HMDI3 (f) over Jesi, Italy (43.82° N, 13.75° E). The small (large) gray dots correspond to the ensemble members of the hindcasts (ensemble members mean) for each start date. The gray horizontal line shows the mean of the hindcast for the whole period (1961–2013). The brown (blue) and green (red) horizontal lines show its lower and upper observed terciles of SPEI6 (HMDI3), respectively. The black dots correspond to the observed SPEI6 values. The forecast is issued for the year 2014–2018, where the percentages are the fraction of ensemble members falling into each tercile category.

METHODS

Data

This study uses decadal hindcasts from the Community Earth System Model Decadal Prediction Large Ensemble (CESM-DPLE). This forecast system is developed by the National Center for Atmospheric Research (NCAR) in an effort to support research into near-term climate information. A summary of the experimental design can be found in Yeager et al.³² For this study, near-term initialized decadal experiments are considered. These are a set of 10-year long initialized hindcasts that were simulated by explicitly prescribing the contemporaneous state of the climate system at the start of the simulation (November 1 of each year from 1960 to 2014) and are forced under the Coupled Model Intercomparison Project phase 5 historical forcings prior to 2015 and RCP8.5 forcings thereafter. The experiment includes a total of 40 ensemble members at the 1-degree nominal resolution, and here we restrict our analysis to the first five forecast years over the global spatial domain where (and when) the winter wheat is harvested. In order to assess the impact of initialization, we have systematically compared the decadal hindcasts with the non-initialized historical simulations (CESM Large Ensemble³³). Such simulations are run by forcing the observed atmospheric composition changes (reflecting both anthropogenic and natural sources) as prescribed in decadal hindcasts, but

without including the details of the contemporaneous state of the climate system.

In this study, we assess the quality of the decadal hindcasts based on two-meter temperature and precipitation. For this, the reference datasets used are Japanese 55-year Reanalysis³⁴ (JRA-55) for two-meter temperature and GPCP Version 2018³⁵ for precipitation. The choice of JRA-55 was motivated by the long continuous coverage of this product (1958–present) and by its high temporal coherency³⁶, in particular regards to daily maximum temperatures. To have a broad estimation of the uncertainties associated with the observational data source, Supplementary Figs. 4–5 show the verification metrics FRPSS and FCRPSS of the estimated drought and heat stress indices calculated with ERA5 reanalysis^{37,38} for two-meter temperature and GPCP Version 2018 for precipitation dataset. The results obtained present a high consistency with the ones obtained in the main manuscript.

Drought and heat stress indices

In this assessment, we use two user-specific indices: SPEI, to assess the drought conditions, and HMDI, to study the occurrence of heat stress during the wheat growing season at the corresponding producing areas.

SPEI is the standardized accumulated climate water balance defined as the difference between monthly precipitation and potential evapotranspiration (PET) over a given period of time. The computation of the SPEI can be divided into two steps for both the forecast and observations. First, we derive the monthly climatic water balance accumulated over the six months prior to the harvesting month for each individual forecast years, as the heat and associated evaporation during this period were identified to have frequently caused drought events with pronounced negative impacts over the global wheat-producing regions⁹. The accumulated values for the forecast years 1–5 are then averaged for each start date and ensemble member of the hindcast dataset. In the second step, we standardize the multi-annual averaged six-month accumulated climate balance values (referred to as SPEI6) by fitting a suitable probability distribution. For this, we use the three-parameter shifted log-logistic probability distribution function to fit the values. The parameters used to build the distribution were computed by taking into account the values of all the ensemble members and start dates. Positive (negative) values of SPEI6 correspond to periods of wet (dry) conditions. The PET calculation is carried out with the Thornthwaite equation as described in Vicente-Serrano et al.¹⁸ because of its simplicity and the relatively minimal amount of data required.

We derive the HMDI as suggested by Zampieri et al.⁹, which is commonly used to study the impact of heatwave and heat-related conditions on wheat production at a global scale. Heatwave is defined here as a period, not interrupted by more than 3 days, with maximum daily temperatures exceeding their 90th percentiles for the period 1961–2018. HMDI is defined as the sum of the daily magnitudes (M_d) of the consecutive days when the heatwaves have occurred during the critical phenological stage of wheat growth that is sensitive to heat stress (three months prior to harvesting).

$$M_d(T_d) = \begin{cases} \frac{T_d - T_{25p}}{T_{75p} - T_{25p}}, & \text{if } T_d > T_{90p} \text{ for three consecutive days} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Here, T_d is the maximum daily temperature on day 'd' of the heatwave. T_{25p} , T_{75p} , and T_{90p} are the 25th, 75th and 90th percentile of daily maximum temperature, respectively, computed for all days of the year centered on a 2-week window for the reference period 1961–2018. As the production of wheat is found to be adversely influenced by the prolonged high temperatures occurring in the three months prior to harvest⁹, we computed the HMDI over this target period (referred to as HMDI3) for individual years.

Since we aim to assess the forecast systems skill at the multi-annual timescale, the forecasted HMDI3 and its corresponding observed values are averaged over 5 years. For example, for the forecast initialized on November 1960, the multi-annual average of HMDI3 (forecast years 1–5) for a specific region, where the harvest is performed in July is the average of the 1961–1965 July HMDI3 values, and for the forecast initialized on November 1961, the average is obtained with the 1962–1966 July HMDI3 values, and so on. Values greater than zero represent a heatwave, where low (high) values present low-intensity (high-intensity) heatwave events.

Subsequently, the forecasts are calibrated in cross-validation mode using the variance inflation technique presented in Doblas-Reyes et al.³⁹. The inflation approach adjusts the forecasted indices to have a similar interannual variance to that of the reference dataset at every grid point and by preserving the ensemble-mean correlation between the forecast and reference dataset. In this case, the Gaussian assumption is a limitation of the approach because the estimated indices can be non-Gaussian. For instance, HMDI has a skewed distribution, primarily due to the finite left-end bound and a right skewness with a heavy/exponential tail. For this reason, a square root transformation of the HMDI was applied prior to calibration. A detailed description of the steps involved in the computation of the indices is present in the supplementary material.

Both the SPEI6 and HMDI3 indices are evaluated over the wheat-producing areas for both the forecasts and observation. The crop calendar for this assessment has been retrieved from MIRCA2000⁴⁰. It provides the harvest months of wheat globally. For each grid box, it is important to note that the evaluation matches the period of the harvest (see Supplementary Fig. 1) and therefore, the evaluation will correspond to different times of the year for different regions.

Skill scores and reliability diagram

We perform the skill assessment using the ranked probability score (RPS) and the Continuous Ranked probability score (CRPS). RPS is the squared distance between the cumulative probabilities of the categorical forecast (tercile events in our case) and its corresponding observational reference,

whereas CRPS measures the integrated squared difference between the forecast and observed cumulative distribution function. Often, it is of interest to evaluate forecasts with respect to a baseline. This baseline is usually a simpler and/or cheaper alternative (e.g., climatology). Such an assessment provides users with information on the added value of the decadal prediction system against this alternative approach. For this reason, the probabilistic skill measures are reformulated as skill scores (RPSS, CRPSS) by comparing the score obtained from the decadal forecasts to the corresponding score obtained from a reference forecast, in our case a climatological forecast.

However, probabilistic measures such as the RPSS and CRPSS require a large ensemble size to produce robust results. For example, Müller et al.⁴¹ and Weigel et al.⁴² have shown that the skill as measured by probabilistic metrics such as the RPSS and CRPSS is strongly influenced by the ensemble size, with lower estimates being associated with smaller ensemble size. To address this, we use the fair version of these skill scores⁴³. Fair scores (FRPSS and FCRPSS) provide an estimate of what the skill would be if the forecast system had an infinite number of ensemble members. The skill score is 1 for a perfect forecast and 0 for the forecasts that do not perform any better than the reference forecast. Negative values indicate that the forecast system performs worse than the reference.

Aside from the probabilistic skill measures, another crucial aspect, from a user perspective, is the reliability of the forecast system. The reliability diagram is the tool most commonly used to assess the reliability of decadal climate predictions. The reliability diagram aims at evaluating whether the forecast probabilities match the observed relative frequencies of occurrences. For example, events forecasted to occur 70% of the time should occur, on average, 70% of the time such forecast is issued. Evaluating the reliability is a critical step, as it allows users to assess whether they can trust the probabilities of the prediction system.

DATA AVAILABILITY

The data that support the findings of this study are openly available and they can be found in the respective references.

CODE AVAILABILITY

The code used for this assessment is available from the corresponding author on reasonable request.

Received: 15 January 2021; Accepted: 13 May 2021;

Published online: 01 June 2021

REFERENCES

- Smith, D. M. et al. Robust skill of decadal climate predictions. *npj Clim. Atmos. Sci.* **2**, 13 (2019).
- Bruno Soares, M., Alexander, M. & Dessai, S. Sectoral use of climate information in Europe: a synoptic overview. *Clim. Serv.* **9**, 5–20 (2018).
- Ceglar, A. et al. Clisagri: an R package for agro-climate services. *Clim. Serv.* **20**, 100197 (2020).
- Hanlon, H. M., Hegerl, G. C., Tett, S. F. & Smith, D. M. Near-term prediction of impact-relevant extreme temperature indices. *Clim. Change*. **132**, 61–76 (2015).
- Caron, L. P. et al. How skillful are the multiannual forecasts of Atlantic hurricane activity? *Bull. Am. Meteorol. Soc.* **99**, 403–414 (2018).
- Paxian, A. et al. User-oriented global predictions of the GPCC drought index for the next decade. *Meteorol. Z.* **28**, 3–21 (2019).
- Solaraju-Murali, B., Caron, L. P., Gonzalez-Reviriego, N. & Doblas-Reyes, F. J. Multi-year prediction of European summer drought conditions for the agricultural sector. *Environ. Res. Lett.* **14**, 124014 (2019).
- Chatzopoulos, T., Pérez Domínguez, I., Zampieri, M. & Toreti, A. Climate extremes and agricultural commodity markets: a global economic analysis of regionally simulated events. *Weather Clim. Extrem.* **27**, 100193 (2020).
- Zampieri, M., Ceglar, A., Dentener, F. & Toreti, A. Wheat yield loss attributable to heat waves, drought and water excess at the global, national and subnational scales. *Environ. Res. Lett.* **12**, 064008 (2017).
- Selvaraju, R. Climate risk assessment and management in agriculture. *Proc. Joint FAO/OECD Workshop* **23**, 71–90 (2012).
- Doblas-Reyes, F. J. et al. Initialized near-term regional climate change prediction. *Nat. Commun.* **4**, 1715 (2013).
- Powell, J. P. & Reinhard, S. Measuring the effects of extreme weather events on yields. *Weather Clim. Extrem.* **12**, 69–79 (2015).

13. Dolferus, R., Ji, X. & Richards, R. A. Abiotic stress and control of grain number in cereals. *Plant Sci.* **181**, 331–341 (2011).
14. Moriondo, M., Giannakopoulos, C. & Bindri, M. Climate change impact assessment: the role of climate extremes in crop yield simulation. *Clim. Change.* **104**, 679–701 (2011).
15. Harkness, C. et al. Adverse weather conditions for UK wheat production under climate change. *Agric. For. Meteorol.* **282–283**, 107862 (2020).
16. Dong, B. et al. Effects of drought stress on pollen sterility, grain yield, abscisic acid and protective enzymes in two winter wheat cultivars. *Front. Plant Sci.* **8**, 1008 (2017).
17. Merryfield, W. J. et al. Current and emerging developments in subseasonal to decadal prediction. *Bull. Am. Meteorol. Soc.* **101**, E869–E896 (2020).
18. Vicente-Serrano, S. M., Beguería, S. & López-Moreno, J. I. A multiscale drought index sensitive to global warming: The standardized precipitation evapotranspiration index. *J. Clim.* **23**, 1696–1718 (2010).
19. Thornthwaite, C. W. An approach toward a rational classification of climate. *Geogr. Rev.* **38**, 55–94 (1948).
20. Beguería, S., Vicente-Serrano, S. M., Reig, F. & Latorre, B. Standardized precipitation evapotranspiration index (spei) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. *Int. J. Climatol.* **34**, 3001–3023 (2014).
21. Sheffield, J., Wood, E. F. & Roderick, M. L. Little change in global drought over the past 60 years. *Nature* **491**, 435–438 (2012).
22. Hargreaves, G. H. Defining and using reference evapotranspiration. *J. Irrig. Drain. Eng.* **120**, 1132–1139 (1994).
23. Droogers, P. & Allen, R. G. Estimating reference evapotranspiration under inaccurate data conditions. *Irrig. Drain. Syst.* **16**, 33–45 (2002).
24. Toreti, A., Cronie, O. & Zampieri, M. Concurrent climate extremes in the key wheat producing regions of the world. *Sci. Rep.* **9**, 5493 (2019).
25. Van Den Dool, H. M. & Toth, Z. Why do forecasts for "near normal" often fail? *Weather Forecast.* **6**, 76–85 (1991).
26. Kushnir, Y. et al. Towards operational predictions of the near-term climate. *Nat. Clim. Change.* **9**, 94–101 (2019).
27. Bellprat, O., Guemas, V., Doblas-Reyes, F. & Donat, M. G. Towards reliable extreme weather and climate event attribution. *Nat. Commun.* **10**, 1732 (2019).
28. Verfaillie, D. et al. How reliable are decadal climate predictions of near-surface air temperature? *J. Clim.* **34**, 697–713 (2021).
29. Scaife, A. A. & Smith, D. A signal-to-noise paradox in climate science. *npj Clim. Atmos. Sci.* **1**, 1–8 (2018).
30. Toreti, A. et al. Narrowing uncertainties in the effects of elevated CO₂ on crops. *Nat. Food* **1**, 775–782 (2020).
31. Manstretta, V. et al. Report on the identified specific needs and opportunities. *H2020 MED-GOLD Deliv. 4.1* 1–18 (2019). URL www.med-gold.eu/wp-content/uploads/docs/776467_MED-GOLD_DEL4.1_Report-on-the-identified-specific-needs-and-opportunities.pdf
32. Yeager, S. G. et al. Predicting near-term changes in the earth system: A large ensemble of initialized decadal prediction simulations using the community earth system model. *Bull. Am. Meteorol. Soc.* **99**, 1867–1886 (2018).
33. Kay, J. E. et al. The community earth system model (cesm) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Am. Meteorol. Soc.* **96**, 1333–1349 (2015).
34. Kobayashi, S. et al. The JRA-55 reanalysis: general specifications and basic characteristics. *J. Meteorol. Soc. Japan.* **93**, 5–48 (2015).
35. Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A. & Ziese, M. Gpcp full data monthly product version 2018 at 0.5°: Monthly land-surface precipitation from rain-gauges built on gts-based and historical data. *Global Precipitation Climatology Centre* (2018).
36. Ceglar, A., Toreti, A., Balsamo, G. & Kobayashi, S. Precipitation over Monsoon Asia: a comparison of reanalyses and observations. *J. Clim.* **30**, 465–476 (2017).
37. Bell, B. et al. ERA5 hourly data on single levels from 1950 to 1978 (preliminary version). *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)* (2020).
38. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
39. Doblas-Reyes, F. J., Hagedorn, R. & Palmer, T. The rationale behind the success of multi-model ensembles in seasonal forecasting - II. Calibration and combination. *Tellus A* **57**, 234–252 (2005).
40. Portmann, F. T., Siebert, S. & Döll, P. MIRCA2000-Global monthly irrigated and rainfed crop areas around the year 2000: a new high-resolution data set for agricultural and hydrological modeling. *Global Biogeochem. Cy.* **24**, GB1011 (2010).
41. Müller, W. A., Appenzeller, C., Doblas-Reyes, F. J. & Liniger, M. A. A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *J. Clim.* **18**, 1513–1523 (2005).
42. Weigel, A. P., Liniger, M. A. & Appenzeller, C. The discrete Brier and ranked probability skill scores. *Mon. Weather Rev.* **135**, 118–124 (2007).
43. Ferro, C. A. Fair scores for ensemble forecasts. *Q. J. R. Meteorol. Soc.* **140**, 1917–1923 (2014).

ACKNOWLEDGEMENTS

We would like to acknowledge financial support from the European Union's Horizon 2020 Research and Innovation programme (MED-GOLD; Grant No. 776467, EUCP; Grant No. 776613 and FOCUS-Africa; Grant No. 869575) and from the Ministerio de Economía y Competitividad (MINECO) as part of the project CLINSA (Grant No. CGL2017-85791-R). This study has also received support from C3S_34c (contract number: ECMWF/COPERNICUS/2019/C3S_34c_DWD) of the Copernicus Climate Change Service (C3S) operated by ECMWF. B.S.M. acknowledges additional financial support from the Marie Skłodowska-Curie fellowship (Grant No. 713673) and from a fellowship of 'la Caixa' Foundation (ID 100010434). The fellowship code is LCF/BQ/IN17/11620038.

AUTHOR CONTRIBUTIONS

F.J.D.R. initiated this work. B.S.M. carried out the analysis with input from all authors. A.C. and M.T. provided crop calendar dataset and contributed to numerical implementation. B.S.M. and N.G.R. drafted the manuscript with extensive contributions from L.P.C., A.C., A.T., and M.Z. P.A. and M.S.C. provided support with the data collection and formatting.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41612-021-00189-4>.

Correspondence and requests for materials should be addressed to B.S.-M.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021