



Geophysical Research Letters[®]



RESEARCH LETTER

10.1029/2021GL094662

On the Reliability of Global Seasonal Forecasts: Sensitivity to Ensemble Size, Hindcast Length and Region Definition

R. Manzanas^{1,2} , V. Torralba³ , Ll. Lledó⁴ , and P. A. Bretonnière⁴ 

¹Departamento de Matemática Aplicada y Ciencias de la Computación (MACC), Universidad de Cantabria, Santander, Spain, ²Grupo de Meteorología y Computación, Universidad de Cantabria, Unidad Asociada al CSIC, Santander, Spain, ³Fondazione Centro Euro-Mediterraneo Sui Cambiamenti Climatici (CMCC), Climate Simulations and Predictions Division, Bologna, Italy, ⁴Barcelona Supercomputing Center (BSC), Barcelona, Spain

Key Points:

- KP1 The new SEAS5 from the European Center for Medium Weather Forecasts increases the reliability of the previous System4 for global seasonal predictions of temperature and precipitation
- KP2 The reliability of probabilistic seasonal forecasts can vary substantially due to the ensemble size and the length of the available hindcast
- KP3 The newly defined IPCC-AR6 land reference regions are adequate for the verification of seasonal forecast reliability

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

R. Manzanas,
rodrigo.manzanas@unican.es

Citation:

Manzanas, R., Torralba, V., Lledó, L., & Bretonnière, P. A. (2022). On the reliability of global seasonal forecasts: Sensitivity to ensemble size, hindcast length and region definition. *Geophysical Research Letters*, 49, e2021GL094662. <https://doi.org/10.1029/2021GL094662>

Received 2 FEB 2022

Accepted 24 AUG 2022

Author Contributions:

Conceptualization: R. Manzanas

Data curation: R. Manzanas, P. A. Bretonnière

Formal analysis: R. Manzanas, V. Torralba, Ll. Lledó

Investigation: R. Manzanas, V. Torralba, Ll. Lledó

Methodology: R. Manzanas

Software: R. Manzanas

Supervision: R. Manzanas

Validation: R. Manzanas

© 2022 The Authors.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Abstract One of the key quality aspects in a probabilistic prediction is its reliability. However, this property is difficult to estimate in the case of seasonal forecasts due to the limited size of most of the hindcasts that are available nowadays. To shed light on this issue, this work presents a detailed analysis of how the ensemble size, the hindcast length and the number of points pooled together within a particular region affect the resulting reliability estimates. To do so, we build on 42 land reference regions recently defined for the IPCC-AR6 and assess the reliability of global seasonal forecasts of temperature and precipitation from the European Center for Medium Weather Forecasts SEAS5 prediction system, which is compared against its predecessor, System4. Our results indicate that whereas longer hindcasts and larger ensembles lead to increased reliability estimates, the number of points that are pooled together within a homogeneous climate region is much less relevant.

Plain Language Summary Seasonal climate forecasts provide information on the average conditions that can be expected for the next months (up to a year) and can help decision making in different socio-economic sectors such as agriculture, energy and health (among others). However, predictability at this time-scale is in general limited, so the actual usefulness of seasonal forecasts must be carefully evaluated before they are used in practical applications. In this aspect, reliability—which measures how well/bad the forecast probability for a particular event fits with its actual occurrence—is a key property. This work assesses the reliability of global seasonal forecasts of temperature and precipitation using the latest operational seasonal forecasting system from European Center for Medium Weather Forecasts. Our results show that reliability is generally better for temperature than for precipitation. Moreover, we demonstrate that reliability is sensitive to the number of retrospective forecasts (known as hindcast) and ensemble members (from which forecast probabilities are obtained) available. Finally, we also demonstrate that the new IPCC-AR6 land reference regions are adequate for seasonal verification purposes. These findings are important for a fair interpretation of the reliability of seasonal forecasts which are obtained for specific regions/seasons/systems building on different experimental frameworks.

1. Introduction

Seasonal climate forecasts have become essential for several socio-economic sectors such as agriculture, renewable energy, water management, insurance or public health (Befort et al., 2019; Ceglar et al., 2018; Lowe et al., 2016; Pechlivanidis et al., 2020; Torralba et al., 2017). These sectors have started to integrate seasonal forecasts in their decision-making processes because this information can lead to better and timely management of risks related to climate variability (Bruno-Soares et al., 2018; Buontempo et al., 2014). Several institutions around the world use coupled ocean-atmosphere general circulation models (GCMs) to produce seasonal forecasts on a regular basis. These forecasts are issued in a probabilistic way, which accounts for the uncertainties coming from the initial conditions and model formulation (Slingo & Palmer, 2011). The delivery of suitable information regarding the quality of these predictions is essential to help stakeholders in making better informed decisions (Alessandrini et al., 2013; Doblas-Reyes et al., 2013).

Seasonal forecast systems have been traditionally evaluated in terms of skill scores (see, e.g., Manzanas et al., 2014, 2017, 2020; Nikulin et al., 2018) and also their ability to reproduce the large-scale modes of variability (e.g., Stockdale et al., 2015). However, it is also key that these forecasts are statistically reliable (Palmer, 2002; Weisheimer & Palmer, 2014). Reliability measures the agreement between the forecast probabilities for a certain

Visualization: R. Manzanas

Writing – original draft: R. Manzanas,
V. Torralba, L.I. Lledó

Writing – review & editing: R.
Manzanas, V. Torralba, L.I. Lledó

event/category (e.g., the temperature being warmer than normal) and the actually observed frequency of occurrence of that event (Mason & Stephenson, 2008). For example, if we collect a large sample of forecasts indicating a 70% of probability, we would ideally expect to observe the event in the ~70% of the cases. Previous studies have shown that seasonal forecasts are reliable for particular regions and seasons of the year (see, e.g., Manzanas et al., 2018, 2019; Nikulin et al., 2018; Weisheimer & Palmer, 2014), but they can be also affected by over/under-confidence problems (see, e.g., Baker et al., 2018; Becker & Van Den Dool, 2016; Johnson et al., 2019). To overcome these issues, the use of large ensembles has been recommended (Eade et al., 2014; Manzanas et al., 2019). However, one of the major limitations in seasonal forecasting is the small sample size (i.e., number hindcast years and ensemble members) from which estimates of forecast quality (e.g., reliability) can be computed. Furthermore, while most skill metrics for seasonal forecasts are directly computed on a grid point basis, measuring reliability requires aggregation of many nearby grid points in order to obtain larger samples (Manzanas et al., 2018; Matsueda et al., 2016; Verfaillie et al., 2020; Weisheimer & Palmer, 2014). Although the scientific community has already recognized the importance of the sample size in seasonal forecast verification (Kumar, 2009; Lledó et al., 2020; Manzanas et al., 2019; Siebert et al., 2016), a systematic evaluation on how this factor may affect reliability estimates at a global scale is still lacking.

The aim of the present work is to fill this gap of knowledge by providing a detailed analysis of the influence on reliability of the ensemble size, the hindcast length and the number of points aggregated within a (homogeneous climate) region. To do this, we consider 2-m temperature and precipitation from SEAS5, the current operational seasonal forecasting system from European Center for Medium Weather Forecasts (ECMWF), and compare its reliability with that obtained from its predecessor, the System4.

This comparison allows us to quantify the improvement of SEAS5 over System4. Moreover, the systematic evaluation of the effect that the available sample size has on the reliability estimates allows for fairer interpretation of the results obtained when different seasonal forecasting systems are compared. In this regard, note that, although, the ensemble size and hindcast length of a prediction system are ultimately determined by the producing center, the number of points needed for the characterization of the reliability in a particular region is a choice to be made during verification.

The paper has been structured as follows. Section 2 describes the seasonal forecasts and the observational references used, and presents the methodological framework considered. Section 3 discusses the sensitivity of the reliability to the ensemble size (Section 3.2), hindcast length (Section 3.3) and region definition (Section 3.4). Finally, Section 4 presents the main conclusions obtained from this work.

2. Data and Methodology

2.1. Observational Reference

Observations of 2-m temperature and precipitation from the EWEMBI data set (Lange, 2019) have been used as reference for evaluating the seasonal forecast reliability. EWEMBI has global coverage with a horizontal resolution of 0.5°—here we have applied a land-sea mask to focus exclusively on land grid points—and it is available for the 1979–2016 period, providing daily data for 41 meteorological variables based on a combination of different data sources. For the variables used in this work, 2-m temperature and precipitation, EWEMBI is based on the WATCH forcing methodology applied to ERA-Interim reanalysis data (WFDEI: Weedon et al. (2014)). For the particular case of precipitation, WFDEI has been bias adjusted with respect to GPCC v5 and v6 (Schneider et al., 2018). The EWEMBI data set was compiled to support the bias correction of climate input data for the impact assessments carried out in phase 2b of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b: Frieler et al., 2017), which has contributed to the 2018 IPCC special report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse emission pathways (<https://www.ipcc.ch/sr15/>).

In addition to EWEMBI, we have also considered another data set of observed 2-m temperature and precipitation in order to assess the effect that observational uncertainty may have on the estimation of seasonal reliability, the gridded Climate Research Unit Time Series (CRU TS, version 4.04: (Harris et al., 2020)), developed by the University of East Anglia. CRU TS has global coverage (only for land regions) with a resolution of 0.5° and provides monthly data for 1901–2019. This data set has been produced using angular-distance weighting interpolation from an extended network of weather stations distributed worldwide. Note thus its convenience as alternative observational reference to EWEMBI, which is mostly based on reanalysis data.

In all cases, for each variable, season and grid point, the 33rd and 66th climatological percentiles in 1981–2016 (the common period between observations and hindcasts) have been computed and used to separate the observations in three tercile categories (below normal or T1, normal or T2, and above normal or T3). The reliability assessment is based on tercile events, because most of the seasonal forecast products tailored to end-user applications are currently given as the percentage of ensemble members falling into the below normal, near normal, and above normal categories of the climatological distribution.

2.2. Seasonal Hindcasts

Seasonal forecasts of 2-m temperature and precipitation from the current operational seasonal forecasting system at the ECMWF, SEAS5 (Johnson et al., 2019), have been used. SEAS5 is based on the Integrated Forecast System (IFS Cycle 43r1) atmospheric model coupled to the Nucleus for European Modeling of the Ocean (NEMO 3.4.1) ocean model. SEAS5's forecasts are issued the first day of each month and span seven months into the future at 6-hourly time resolution and a spatial resolution of ~ 36 km. Hindcasts are available for the period 1981–2016 (i.e., 36 years) with 51 ensemble members for the start dates of February, May, August, and November. Note that SEAS5 has been selected for this study because it is the operational seasonal forecast system with the longest hindcast and the largest ensemble, which allows for robust experimentation with increasing sample sizes.

For comparison purposes, some analyses have been also carried out for the ECMWF System4 (Molteni et al., 2011), the previous version of SEAS5, whose operational lifecycle ended in November 2017. For this model we have combined hindcasts (1981–2010) and operational forecasts (2011–2016) to cover the full period available for SEAS5, 1981–2016. System4 has a spatial resolution of ~ 80 km and provides 51 members for the start dates of February, May, August, and November.

In this work we focus on one-month lead forecasts for December-January-February (DJF), March-April-May (MAM), June-July-August (JJA), and September-October-November (SON), corresponding to the start dates of November, February, May, and August, respectively. For the sake of direct comparison between SEAS5 and System4, both models have been interpolated to the EWEMBI observational grid through a nearest neighbor approach. In all cases, as for the observations, the 33rd and 66th percentiles for the 1981–2016 hindcast have been computed for each variable, season and grid. For a given year, the probability for each tercile category is given by counting the number of members (out of the total available) falling in that category.

2.3. Reliability Categories

The reliability of a prediction system measures how closely the forecast probabilities of a certain event (here, one of the three tercile categories: T1, T2, and T3) correspond to the actual observed frequency of that event. Typically, the forecast probabilities are plotted against the observed relative frequencies for a number of probability bins in the reliability diagram (see, e.g., Doblas-Reyes et al., 2008; Frías et al., 2018). As in previous works (see, e.g., Johnson et al., 2019), the number of probability bins used in all the analyses undertaken in this study equals the number of ensemble members considered plus one. Weisheimer and Palmer (2014) first proposed a methodology to translate the information contained in the reliability diagram into an easy-to-interpret scale with five categories, which was later modified by Manzananas et al. (2018) by including a sixth reliability category: *perfect* (green), *still very useful* (blue), *marginally useful +* (dark yellow), *marginally useful* (yellow), *not useful* (orange) and *dangerously useless* (red). This classification is based on the relative position of the reliability line (a weighted linear regression of the observed frequencies) with respect to the *perfect reliability* line, the *no-skill* line, and the *no-resolution* line, as well as on the 75% uncertainty range around it. The latter is derived in this work from 1000 bootstrapped reliability lines which are obtained by randomly resampling members, gridboxes, and years both in the observations and in the predictions. Note that this bootstrapping process makes the computation of reliability categories highly demanding in terms of computational resources and times. Note also that using different tercile thresholds for hindcasts and observations implicitly corrects systematic biases that may be present in the model predictions (see, e.g., Manzananas, 2020). Therefore, reliability is a non-sensitive to bias metric. The reader is referred to the aforementioned references for further details on reliability diagrams and reliability categories.

One problem that arises when computing the reliability diagrams and the corresponding reliability categories is the small sample sizes available for each probability bin. In our case, for a single grid point, season and category

one would have only 36 SEAS5 forecasts (one per year), which may result in bins with very few instances. To make the computation more stable to the random effects that may appear due to the bootstrapping step followed, it is usual to pool forecasts of nearby grid points in a single reliability diagram to derive the corresponding areal reliability category. For instance, Weisheimer and Palmer (2014) and Verfaillie et al. (2020) considered 21 regions over land for this purpose. In this study, we have used the new set of 42 land reference regions which have been defined in Iturbide et al. (2020) for the Sixth Assessment Report of the IPCC (two regions over the Antarctic continent have been omitted). The reader is referred to Figure S1 and Table S1 in Supporting Information S1 for details about these regions, including the number of grid points contained in each of them at the 0.5° resolution used here.

3. Results

3.1. Comparison of SEAS5 and System4 in Terms of Reliability

Although some recent works have preliminarily analyzed the reliability of SEAS5 for particular regions and/or seasons (see Gubler et al., 2020; Johnson et al., 2019), a global, systematic assessment is still lacking. Figures 1 and 2 shows the reliability categories obtained for precipitation and temperature in DJF, MAM, JJA, and SON over the 42 IPCC-AR6 reference regions considered when EWEMBI is used as observational reference. For each season, the maps in the first row correspond to the SEAS5. The second row provides a comparison between SEAS5 and System4, expressed as the number of categories changed in the former, with respect to the latter. Blue (red) colors indicate thus that SEAS5 improves (worsens) the reliability found for System4. For instance, passing from *not useful* in System4 to *marginally useful +* in SEAS5 (or from *perfect* to *still very useful*) would be represented by a 2 (−1). For both models, SEAS5 and System4, the full hindcast (51 members and 36 years) is considered in this experiment. For the sake of brevity, we only include here results for the below and above normal categories (T1 and T3, respectively). In agreement with previous works (Kharin & Zwiers, 2003; Van Den Dool & Toth, 1991; Yang et al., 2021), forecasts for the near normal category (T2) exhibit poorer quality in all cases.

The comparison between SEAS5 and System4 reveals that the former outperforms the latter in terms of reliability, as shown by the predominance of the blue color in most of the regions and seasons. In general, forecast reliability is higher for temperature than for precipitation. Indeed, Figure 2 suggests that seasonal forecasts of temperature might be safely used for practical applications in several regions and seasons for which the perfect reliability category is found. Differently, according to Figure 1, reliability is in general poor for precipitation, especially outside the tropics. Note that the results obtained here for System4 should not be directly compared with those presented in Weisheimer and Palmer (2014), who used a shorter 30-year period and a different set of 21 regions.

To investigate the effect that observational uncertainty may have on the results shown in Figures 1 and 2, the maps in the first rows of Figures S2 and S3 in Supporting Information S1 have been calculated considering CRU TS (instead of EWEMBI) as observational reference for precipitation and temperature, respectively. Moreover, the maps in the second row show the differences in reliability—expressed as the number of categories changed—with respect to those displayed in the first rows of Figures 1 and 2 (i.e., using EWEMBI as reference). Blue (red) colors indicate thus that the reliability of SEAS5 improves (worsens) when CRU TS is used for validation instead of EWEMBI.

These two figures reveal that the choice of observational reference may indeed lead to slightly different results for particular variable-season-region-tercile combinations (please note a detailed analysis of these differences is out of the scope of this paper). Nevertheless, most of the changes found (either improvements or worsening) are scattered worldwide and limited to one reliability category, which in some cases might be even explained by the uncertainty inherent to the calculation of reliability (recall bootstrapping needs to be applied). Importantly, none of the two datasets considered leads systematically to better results, neither for precipitation nor for temperature.

3.2. Sensitivity to the Ensemble Size

The SEAS5's hindcast have 51 ensemble members for the specific start dates considered in this work (February, May, August, and November). However, this ensemble size is halved for the rest of start dates. Furthermore, it is common that the operational configuration of the different seasonal forecast systems has more ensemble

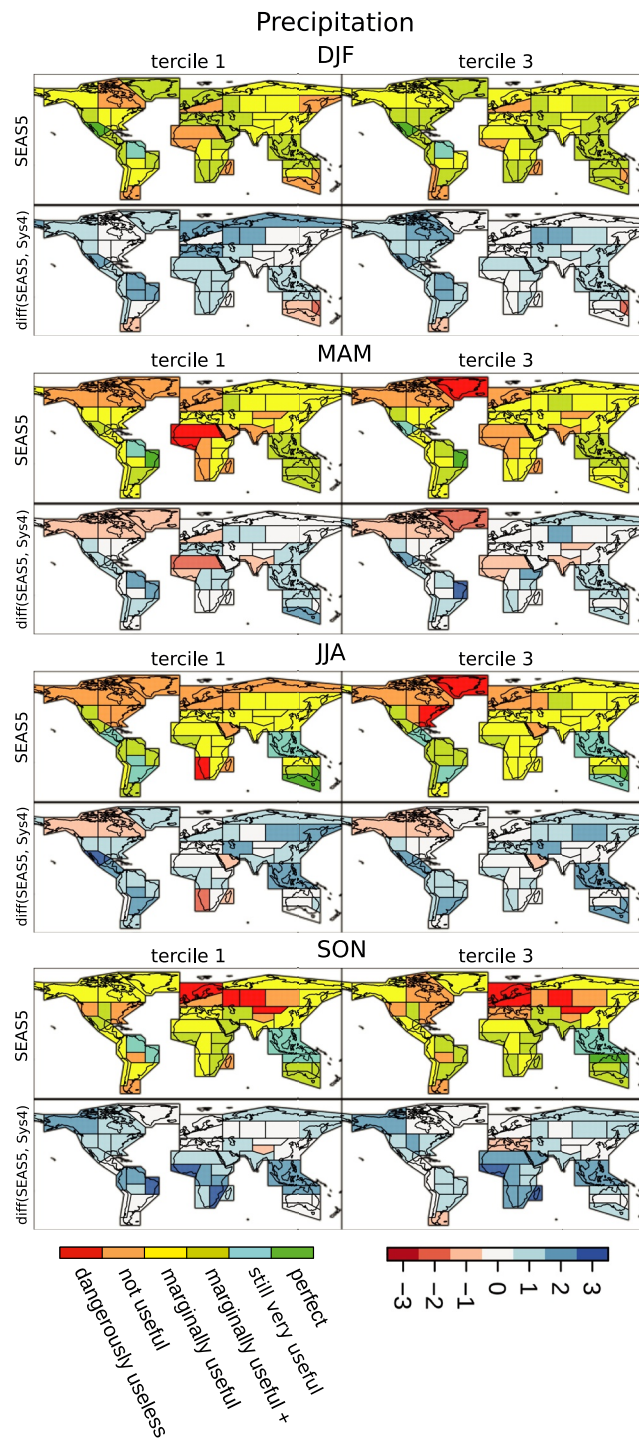


Figure 1. Reliability categories obtained for precipitation in December-January-February, March-April-May, June-July-August, and September-October-November (from top to bottom) over the 42 IPCC-AR6 reference regions considered when EWEMBI is used as observational reference. For each season, the maps in the first row correspond to the SEAS5. The second row provides a comparison between SEAS5 and System4, expressed as the number of categories changed in the former, with respect to the latter. Blue (red) colors indicate that SEAS5 improves (worsens) the reliability found for System4. For both models, SEAS5 and System4, the full hindcast (51 members and 36 years) is considered in this experiment, so direct comparison is fair. For brevity, results are only shown for the below and above normal categories (T1 and T3, respectively).

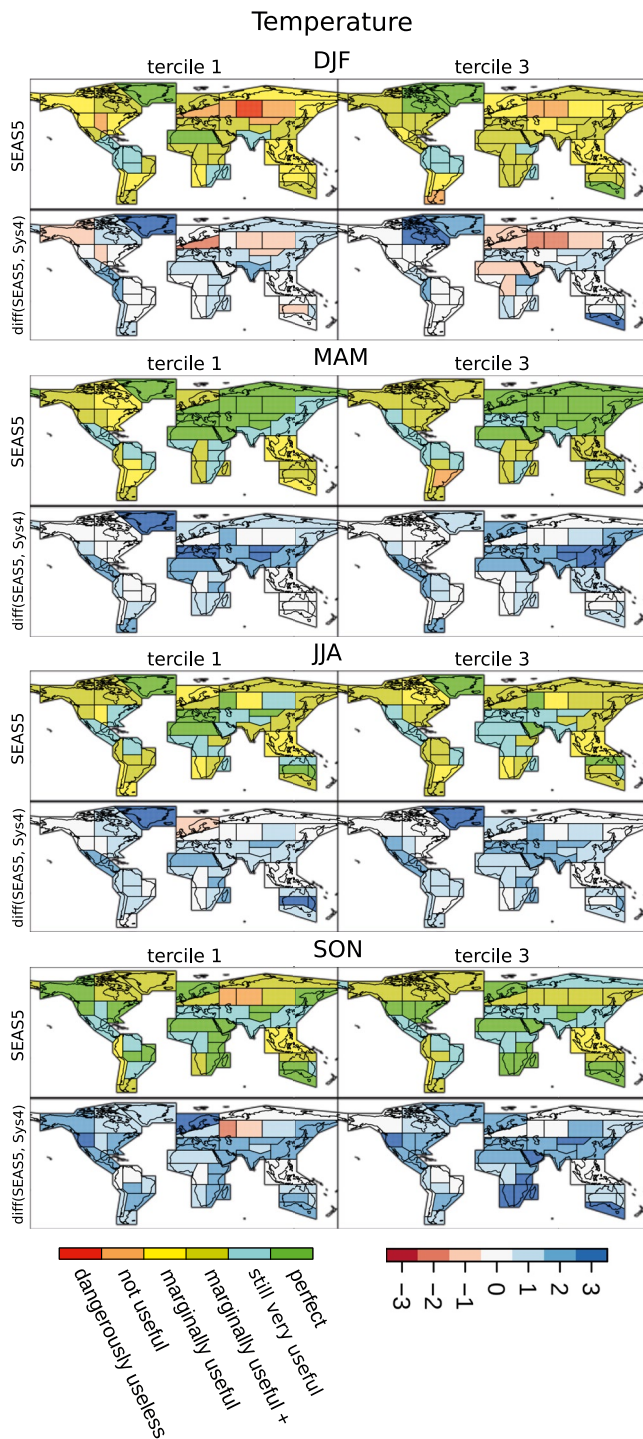


Figure 2. As Figure 1 but for temperature.

members than the corresponding hindcasts that are used for forecast quality assessment. Therefore, it is important to assess how the ensemble size available may affect reliability.

Figure 3 shows the number of regions exhibiting each of the six reliability categories (blocks of bars in different colors) as a function of the ensemble size (for 10, 30, and 51 members), for precipitation and temperature (in columns) along the different seasons (in rows), when the SEAS5 is validated against EWEMBI. For the sake of robustness, we have considered five different subsets of 10 members and five other subsets of 30 members, all of them randomly chosen out of the 51 available members. Within each block of bars, the first two represent the mean result obtained across the five subsets of years considered, with the error bar accounting for the corresponding standard deviation. To provide a summary of the full picture, the results displayed are summed in all cases along the three terciles (T1, T2, and T3). In general, in agreement with previous works (see, e.g., Hagedorn et al., 2005; Manzanas et al., 2019), this figure evidences that reliability improves as the number of ensemble members considered increases. In particular, for the case of precipitation, the number of regions exhibiting the marginally useful + category increases in general when passing from 10 to 30 and 51 ensemble members. A very similar behavior is also found for temperature, for which the number of regions with still very useful and perfect reliability categories tends to increase with the ensemble size.

Similar results are found when EWEMBI is substituted by CRU TS for verification (see Figure S4 in Supporting Information S1), which suggests that our findings are robust to the choice of observational reference.

3.3. Sensitivity to the Hindcast Length

Most of seasonal forecast systems produce a hindcast which typically covers less than 25 years since its generation is computationally very expensive and consistent oceanic observations for initialization are usually not available before 1993. For this reason it is important to assess how the hindcast length available for verification may affect reliability. SEAS5 constitutes an ideal test-bed to do this, since this model provides the longest-to-date hindcast, covering a 36-year period.

Figure 4 shows the number of regions exhibiting each of the six reliability categories (blocks of bars in different colors) as a function of the hindcast length (for 15, 25, and 36 years), for precipitation and temperature (in columns) along the different seasons (in rows), when the SEAS5 is validated against EWEMBI. For the sake of robustness, we have considered five different subsets of 15 years and five other subsets of 25 years, all of them randomly chosen out of the 36 available years. Within each block of bars, the first two represent the mean result obtained across the five subsets of years considered, with the error bar accounting for the corresponding standard deviation. As in Figure 3, the results displayed are summed in all cases along the three terciles (T1, T2, and T3).

Although the dispersion of the results (i.e., the length of the error bars) is larger in this figure than in Figure 3, it is still clear that the number of regions with dangerously useless and not useful reliability categories is reduced for both temperature and precipitation when the number of years increases, whereas marginally useful and still very useful categories become more frequent. Interestingly, a reduction in the number of regions categorized with perfect reliability is encountered

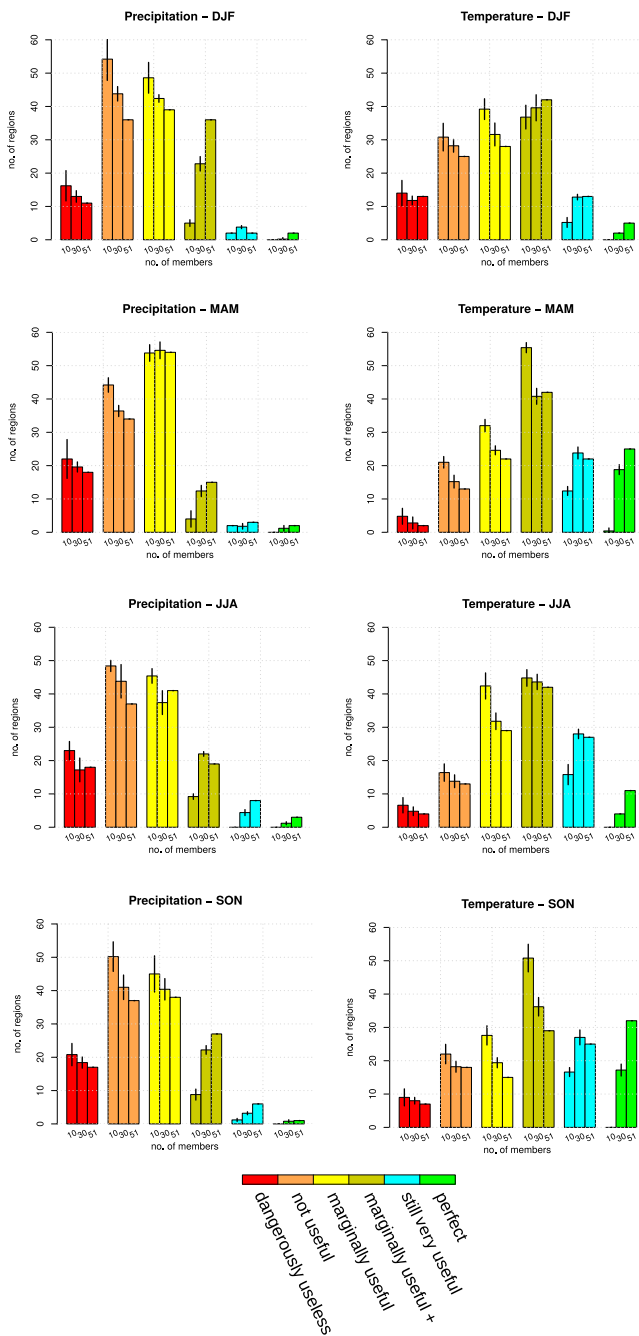


Figure 3. Number of regions exhibiting each of the six reliability categories (blocks of bars in different colors) as a function of the ensemble size (for 10, 30, and 51 members), for precipitation and temperature (in columns) along the different seasons (in rows), when the SEAS5 is validated against EWEMBI. In all cases, results are summed along the three terciles (T1, T2, and T3).

for the case of temperature in DJF when the number of years considered is increased. This suggests that short hindcasts may lead to overestimated reliability estimates merely due to sampling uncertainty (recall that bootstrapping is applied to quantify the uncertainty range around the regressed reliability line). In agreement with previous works (see, e.g., Manzanas et al., 2019), these findings suggest the importance of having long hindcasts for a robust estimation of forecast reliability.

These conclusions are not much altered when EWEMBI is replaced by CRU TS for verification (see Figure S5 in Supporting Information S1), which again suggests that our findings are robust to the choice of observational reference.

3.4. Sensitivity to the Region Definition

The 42 land reference regions considered in this work have been defined in Iturbide et al. (2020) for the assessment of climate change according to their climatic homogeneity in terms of the Köppen–Geiger classification. In this section we investigate if these regions might also be adequate for seasonal forecast verification purposes—recall that reliability is usually assessed by pooling together points over large regions (Buizza & Leutbecher, 2015; Verfaillie et al., 2020; Weisheimer & Palmer, 2014).—To do this, Figure 5 shows the reliability category obtained across the 42 reference regions (in columns) as a function of the number of grid points considered (in rows) within the region for precipitation, when the SEAS5 is validated against EWEMBI. Independent subsets of 25%, 50%, 75% grid points were randomly selected (only once) and kept fixed for the entire experiment. The different seasons are displayed for top to bottom. Within each season, results are shown for T1 and T3 (similar conclusions are obtained for T2). Figure 6 is the equivalent to Figure 5 but for temperature. In both cases, the full hindcast available (51 members and 36 years) from SEAS5 was employed.

Although marginal changes in reliability (either increases or decreases) can be found for specific seasons and terciles in particular regions—note the detailed analysis of these changes is out of the scope of this work,—these two figures evidence that, overall, reliability is not strongly affected by the number of grid points aggregated within a given region, regardless of the variable and the season considered.

This overall conclusion remains valid when EWEMBI is substituted by CRU TS for verification (see Figures S6 and S7 in Supporting Information S1), corroborating the choice of observational reference has little effect on the key outcomes from the present study.

4. Conclusions

Previous works have already explored the influence that the small sample sizes currently available in seasonal forecasting may have on skill, but the impact on reliability has not been explored so far at a global scale. This work aims to fill this gap by providing a systematic evaluation on how the seasonal forecast reliability is affected by different potential sources of uncertainty: ensemble size, hindcast length and number of aggregated points (within a homogeneous climate region). To do this, we analyze one-month lead seasonal forecasts of temperature and precipitation for the four main boreal seasons (DJF,

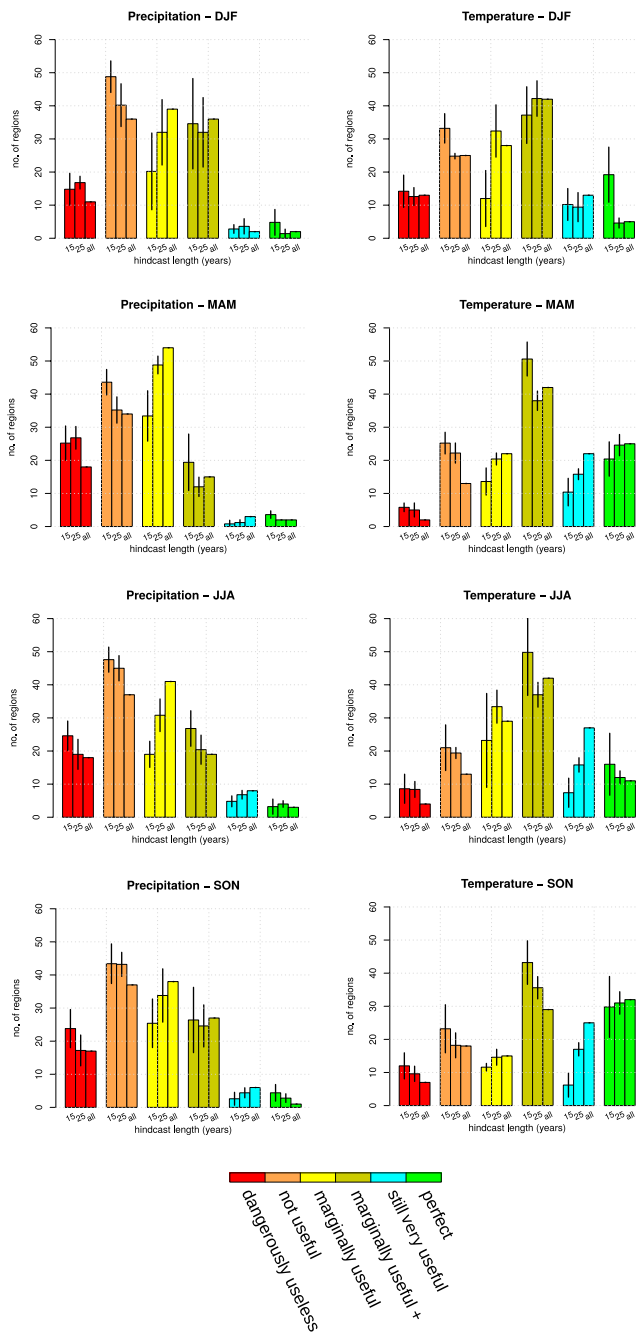


Figure 4. Number of regions exhibiting each of the six reliability categories (blocks of bars in different colors) as a function of the hindcast length (for 15, 25, and 36 years), for precipitation and temperature (in columns) along the different seasons (in rows), when the SEAS5 is validated against EWEMBI. In all cases, results are summed along all the three terciles (T1, T2, and T3).

MAM, JJA, and SON) over the new set of 42 land reference regions defined for the AR6 of the IPCC, which cover the entire globe. To facilitate the interpretation of the results obtained, we employ the six reliability categories first introduced in Weisheimer and Palmer (2014) and later modified by Manzanas et al. (2018), which provide an easy to interpret and communicate ranking.

First, we have compared the reliability provided by SEAS5, the current operational seasonal forecast system from ECMWF, with that from its previous version, System4. Our results show that SEAS5 outperforms System4 for most of the regions and seasons. For temperature, perfect reliability is found for particular regions and seasons for which seasonal forecasts may be directly used in practical applications and/or operational climate services. Differently, reliability is in general poor for precipitation, especially outside the tropics. These results indicate that the improvements undertaken in SEAS5 (with respect to System4) have efficiently helped to increase reliability. However, there is still room for further enhancement, which in some occasions may be achieved through proper calibration of the raw model outputs (see, e.g., Hemri et al., 2020; Manzanas et al., 2019).

Second, building on the SEAS5's full hindcast (51 members and 36 years), we have quantified how reliability estimates change with different ensemble sizes (10, 30, and 51 members) and hindcast lengths (15, 25, and 36 years). On the one hand, we have found that reliability improves in general as the number of ensemble members increases, which indicates that larger ensembles allow for a better representation of the forecast uncertainty, leading to more reliable probabilities. On the other hand, increasing the number of years available for verification leads also to some improvement of reliability, although this effect is not so clear. Indeed, our results suggest that forecast reliability can be overestimated in short hindcasts as a mere consequence of sampling uncertainty.

Third, we have also demonstrated that new IPCC-AR6 land reference regions defined in Iturbide et al. (2020) are useful not only for climate change assessments, but also for the verification of seasonal forecast systems in terms of reliability. Moreover, the fact that a reduced subset of randomly selected points is enough to accurately infer the reliability of these entire regions allows for drastically reducing computational costs, which are typically high as a consequence of the bootstrapping procedure followed. This consideration turns especially important as the spatial resolution of the newer forecasting systems increases.

Finally, the comparison between the figures shown in the manuscript for EWEMBI and the equivalent ones for CRU TS (Figures S1–S7 in Supporting Information S1) allows us to conclude that the choice of observational reference does not greatly affect any of the aforementioned conclusions, which can be useful to (a) model developers who need to fairly assess improvements in successive model versions, (b) climate scientists who need to fairly evaluate and understand the reliability of seasonal forecasts for specific regions/seasons/systems building on different experimental frameworks, and (c) climate services' developers who need to optimize the computational resources used for specific applications in which reliability plays a key role.

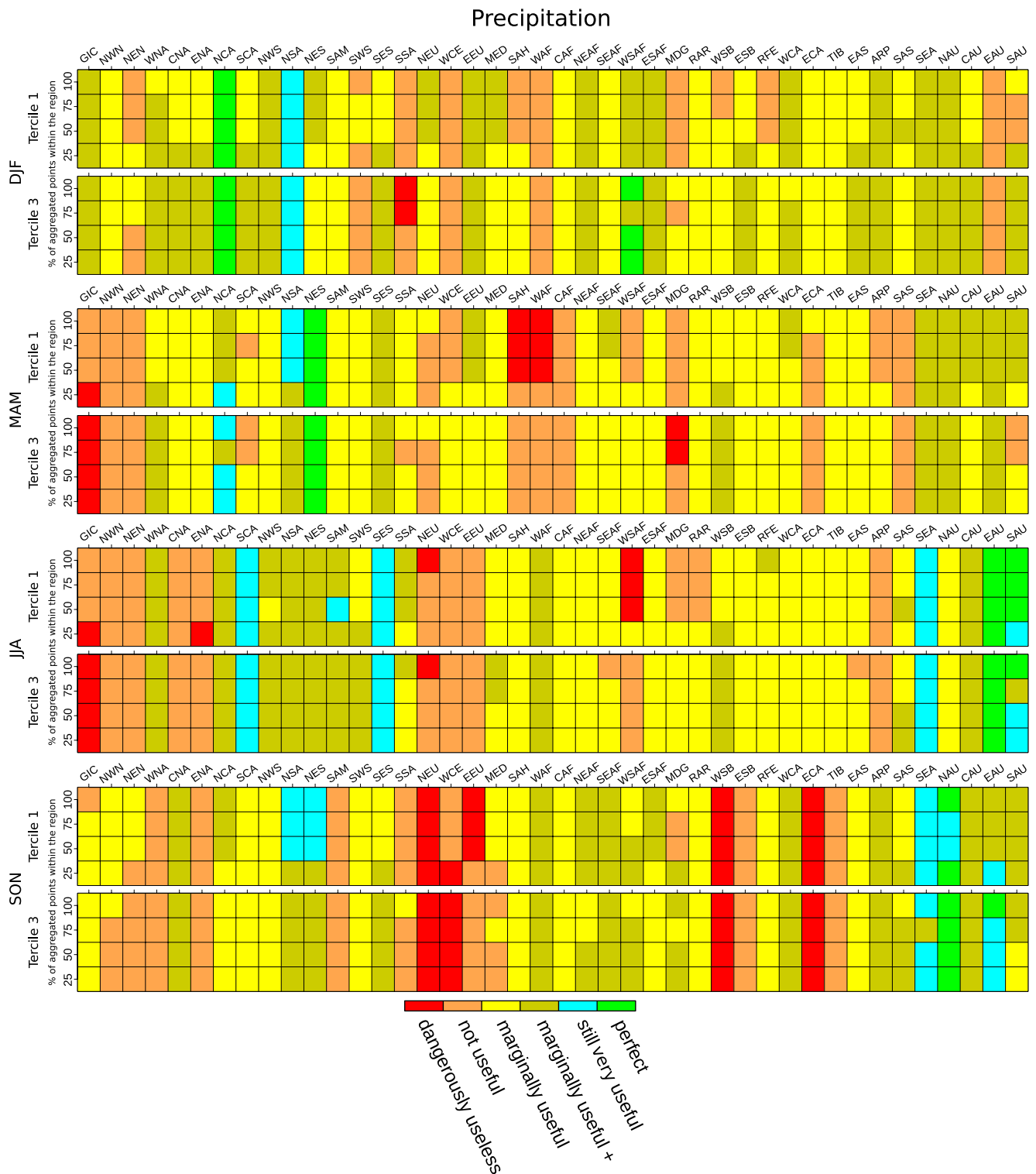


Figure 5. Reliability category obtained across the 42 regions analyzed (in columns; see Figure S1 and Table S1 in Supporting Information S1 for details) as a function of the number of grid points aggregated (25%, 50%, 75%, and 100%, in rows) within the region for precipitation, when the SEAS5 is validated against EWEMBI. The different seasons are displayed for top to bottom. Within each season, results are given for T1 and T3. The full hindcast available (51 members and 36 years) was used for this analysis.

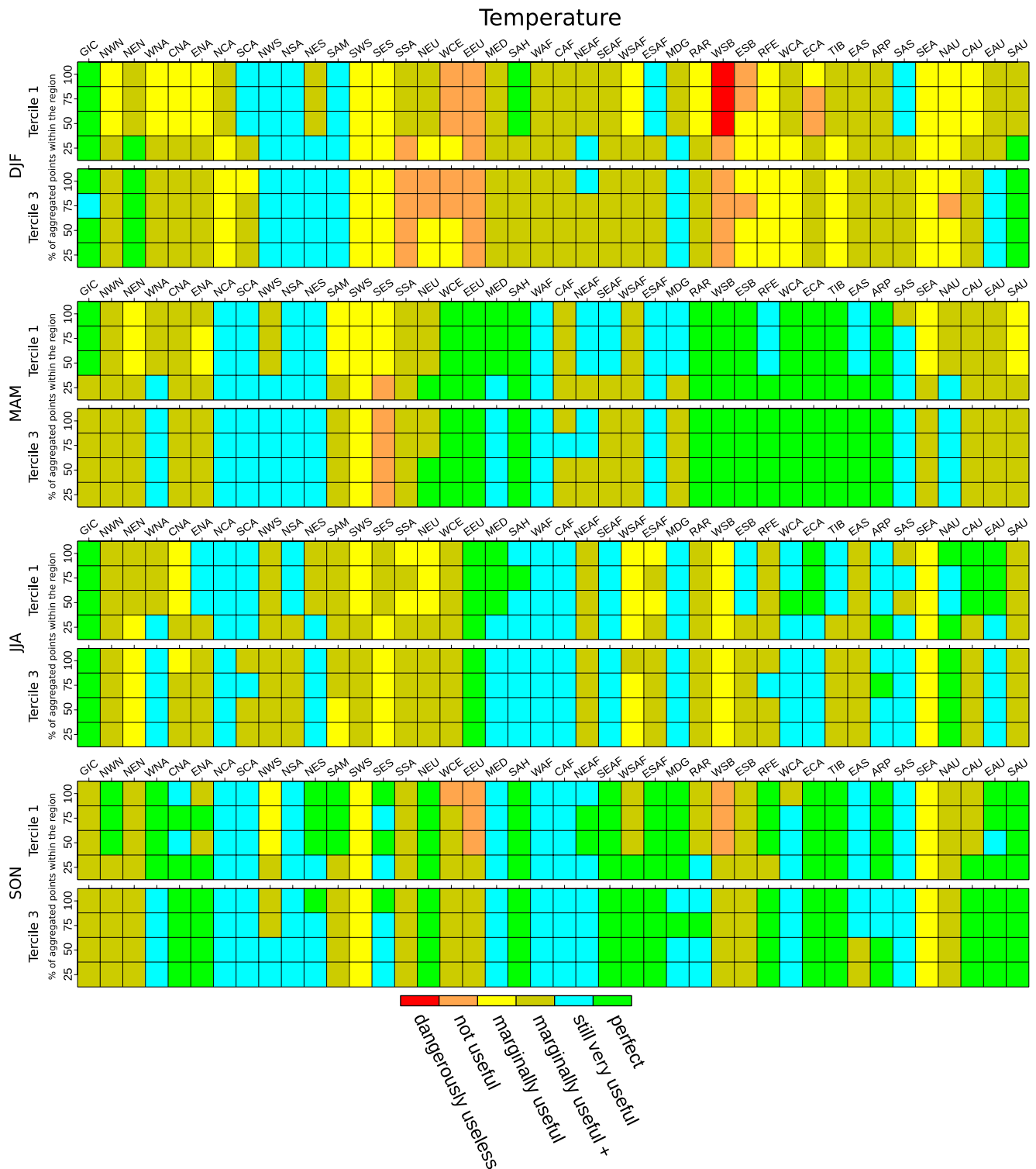


Figure 6. As Figure 5 but for temperature.

Data Availability Statement

SEAS5 data was retrieved from the MARS archive (<https://confluence.ecmwf.int/display/COPSRV/MARS+archive>) following the ECMWF data policy. CRU TS v4.04 was downloaded from <https://catalogue.ceda.ac.uk/uuid/89e1e34ec3554dc98594a5732622bce9>. Differently, System4 and EWEMBI were obtained from the User

Data Gateway (UDG), a THREDDS-based service from the Santander Climate Data Service that provides access to a wide catalogue of popular climate datasets: <http://meteo.unican.es/udg-tap/home>. See Cofiño et al. (2018) for further information. Finally, reliability categories can be computed (and plotted) with the function `reliabilityCategories` included in the `visualizeR` package (Frías et al., 2018), which forms part of `climate4R` (Iturbide et al., 2019), a bundle of R packages developed by the Santander Meteorology Group for transparent climate data access, post-processing and visualization.

Acknowledgments

This research has been partially supported by the AfriCultuReS (“Enhancing Food Security in African Agricultural Systems with the Support of Remote Sensing”) and FOCUS-Africa projects, which received funding from the European Union’s Horizon 2020 Research and Innovation Framework Programme under grant agreements No. 77465 and 869575, respectively.

References

- Alessandrini, S., Sperati, S., & Pinson, P. (2013). A comparison between the ECMWF and COSMO Ensemble Prediction Systems applied to short-term wind power forecasting on real data. *Applied Energy*, *107*, 271–280. <https://doi.org/10.1016/j.apenergy.2013.02.041>
- Baker, L., Shaffrey, L., Sutton, R., Weisheimer, A., & Scaife, A. (2018). An intercomparison of skill and overconfidence/underconfidence of the wintertime North Atlantic Oscillation in multimodel seasonal forecasts. *Geophysical Research Letters*, *45*(15), 7808–7817. <https://doi.org/10.1029/2018GL078838>
- Becker, E., & Van Den Dool, H. (2016). Probabilistic seasonal forecasts in the North American Multimodel Ensemble: A baseline skill assessment. *Journal of Climate*, *29*(8), 3015–3026. <https://doi.org/10.1175/JCLI-D-14-00862.1>
- Befort, D. J., Wild, S., Knight, J. R., Lockwood, J. F., Thornton, H. E., Hermanson, L., et al. (2019). Seasonal forecast skill for extratropical cyclones and windstorms. *Quarterly Journal of the Royal Meteorological Society*, *145*(718), 92–104. <https://doi.org/10.1002/qj.3406>
- Bruno-Soares, M., Alexander, M., & Dessai, S. (2018). Sectoral use of climate information in Europe: A synoptic overview. *Climate Services*, *9*, 5–20. <https://doi.org/10.1016/j.cliser.2017.06.001>
- Buizza, R., & Leutbecher, M. (2015). The forecast skill horizon. *Quarterly Journal of the Royal Meteorological Society*, *141*(693), 3366–3382. <https://doi.org/10.1002/qj.2619>
- Buontempo, C., Hewitt, C. D., Doblas-Reyes, F. J., & Dessai, S. (2014). Climate service development, delivery and use in Europe at monthly to inter-annual timescales. *Climate Risk Management*, *6*, 1–5. <https://doi.org/10.1016/j.crm.2014.10.002>
- Ceglar, A., Toreti, A., Prodhomme, C., Zampieri, M., Turco, M., & Doblas-Reyes, F. J. (2018). Land-surface initialisation improves seasonal climate prediction skill for maize yield forecast. *Scientific Reports*, *8*(1), 1–9. <https://doi.org/10.1038/s41598-018-19586-6>
- Cofiño, A., Bedia, J., Iturbide, M., Vega, M., Herrera, S., Fernández, J., et al. (2018). The ECOMS User Data Gateway: Towards seasonal forecast data provision and research reproducibility in the era of climate services. *Climate Services*, *9*, 33–43. <https://doi.org/10.1016/j.cliser.2017.07.001>
- Doblas-Reyes, F. J., Coelho, C. A. S., & Stephenson, D. B. (2008). How much does simplification of probability forecasts reduce forecast quality? *Meteorological Applications*, *15*(1), 155–162. <https://doi.org/10.1002/met.50>
- Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues, L. R. L. (2013). Seasonal climate predictability and forecasting: Status and prospects. *Wiley Interdisciplinary Reviews: Climate Change*, *4*(4), 245–268. <https://doi.org/10.1002/wcc.217>
- Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., & Robinson, N. (2014). Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, *41*(15), 5620–5628. <https://doi.org/10.1002/2014GL061146>
- Frías, M., Iturbide, M., Manzananas, R., Bedia, J., Fernández, J., Herrera, S., et al. (2018). An R package to visualize and communicate uncertainty in seasonal climate prediction. *Environmental Modelling & Software*, *99*, 101–110. <https://doi.org/10.1016/j.envsoft.2017.09.008>
- Frieler, K., Lange, S., Piontek, F., Reyer, C. P. O., Schewe, J., Warszawski, L., et al. (2017). Assessing the impacts of 1.5°C global warming – Simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b). *Geoscientific Model Development*, *10*(12), 4321–4345. <https://doi.org/10.5194/gmd-10-4321-2017>
- Gubler, S., Sedlmeier, K., Bhend, J., Avalos, G., Coelho, C. A. S., Escajadillo, Y., et al. (2020). Assessment of ECMWF SEAS5 seasonal forecast performance over South America. *Weather and Forecasting*, *35*(2), 561–584. <https://doi.org/10.1175/WAF-D-19-0106.1>
- Hagedorn, R., Doblas-Reyes, F. J., & Palmer, T. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus A: Dynamic Meteorology and Oceanography*, *57*(3), 219–233. <https://doi.org/10.3402/tellusa.v57i3.14657>
- Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, *7*(1), 109. <https://doi.org/10.1038/s41597-020-0453-3>
- Hemri, S., Bhend, J., Liniger, M. A., Manzananas, R., Siebert, S., Stephenson, D. B., et al. (2020). How to create an operational multi-model of seasonal forecasts? *Climate Dynamics*, *55*(5), 1141–1157. <https://doi.org/10.1007/s00382-020-05314-2>
- Iturbide, M., Bedia, J., Herrera, S., Baño-Medina, J., Fernández, J., Frías, M., et al. (2019). The R-based climate4R open framework for reproducible climate data access and post-processing. *Environmental Modelling & Software*, *111*, 42–54. <https://doi.org/10.1016/j.envsoft.2018.09.009>
- Iturbide, M., Gutiérrez, J. M., Alves, L. M., Bedia, J., Cerezo-Mota, R., Giménez, E., et al. (2020). An update of IPCC climate reference regions for subcontinental analysis of climate model data: Definition and aggregated datasets. *Earth System Science Data*, *12*(4), 2959–2970. <https://doi.org/10.5194/essd-12-2959-2020>
- Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., et al. (2019). SEAS5: The new ECMWF seasonal forecast system. *Geoscientific Model Development*, *12*(3), 1087–1117. <https://doi.org/10.5194/gmd-12-1087-2019>
- Kharin, V. V., & Zwiers, F. W. (2003). On the ROC score of probability forecasts. *Journal of Climate*, *16*(24), 4145–4150. [https://doi.org/10.1175/1520-0442\(2003\)016<4145:otsrop>2.0.co;2](https://doi.org/10.1175/1520-0442(2003)016<4145:otsrop>2.0.co;2)
- Kumar, A. (2009). Finite samples and uncertainty estimates for skill measures for seasonal prediction. *Monthly Weather Review*, *137*(8), 2622–2631. <https://doi.org/10.1175/2009MWR2814.1>
- Lange, S. (2019). Earth2Observe, WFDEI and ERA-interim data Merged and Bias-corrected for ISIMIP (EWEMBI). *GFZ Data Services*. <https://doi.org/10.5880/PIK.2019.004>
- Lledó, L., Cionni, I., Torralba, V., Bretonnière, P.-A., & Samsó, M. (2020). Seasonal prediction of Euro-Atlantic teleconnections from multiple systems. *Environmental Research Letters*, *15*(7), 074009. <https://doi.org/10.1088/1748-9326/ab87d2/meta>
- Lowe, R., García-Díez, M., Ballester, J., Creswick, J., Robine, J.-M., Herrmann, F. R., & Rodó, X. (2016). Evaluation of an early-warning system for heat wave-related mortality in Europe: Implications for sub-seasonal to seasonal forecasting and climate services. *International Journal of Environmental Research and Public Health*, *13*(2), 206. <https://doi.org/10.3390/ijerph13020206>
- Manzananas, R. (2020). Assessment of model drifts in seasonal forecasting: Sensitivity to ensemble size and implications for bias correction. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001751. <https://doi.org/10.1029/2019MS001751>

- Manzanas, R., Frias, M. D., Cofiño, A. S., & Gutiérrez, J. M. (2014). Validation of 40 year multimodel seasonal precipitation forecasts: The role of ENSO on the global skill. *Journal of Geophysical Research: Atmospheres*, *119*(4), 1708–1719. <https://doi.org/10.1002/2013JD020680>
- Manzanas, R., Gutiérrez, J. M., Bhend, J., Hemri, S., Doblas-Reyes, F. J., Penabaz, E., & Brookshaw, A. (2020). Statistical adjustment, calibration and downscaling of seasonal forecasts: A case-study for Southeast Asia. *Climate Dynamics*, *54*(5), 2869–2882. <https://doi.org/10.1007/s00382-020-05145-1>
- Manzanas, R., Gutiérrez, J. M., Bhend, J., Hemri, S., Doblas-Reyes, F. J., Torralba, V., et al. (2019). Bias adjustment and ensemble recalibration methods for seasonal forecasting: A comprehensive intercomparison using the C3S dataset. *Climate Dynamics*, *53*(3–4), 1287–1305. <https://doi.org/10.1007/s00382-019-04640-4>
- Manzanas, R., Gutiérrez, J. M., Fernández, J., van Meijgaard, E., Calmanti, S., Magariño, M. E., et al. (2017). Dynamical and statistical downscaling of seasonal temperature forecasts in Europe: Added value for user applications. *Climate Services*. <https://doi.org/10.1016/j.cliser.2017.06.004>
- Manzanas, R., Lucero, A., Weisheimer, A., & Gutiérrez, J. M. (2018). Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts? *Climate Dynamics*, *50*(3), 1161–1176. <https://doi.org/10.1007/s00382-017-3668-z>
- Mason, S., & Stephenson, D. (2008). How do we know whether seasonal climate forecasts are any good? In A. Troccoli, M. Harrison, D. L. T. Anderson, & S. J. Mason (Eds.), *Seasonal climate: Forecasting and managing risk* (Vol. 82, pp. 259–289). Springer Netherlands.
- Matsueda, M., Weisheimer, A., & Palmer, T. (2016). Calibrating climate change time-slice projections with estimates of seasonal forecast reliability. *Journal of Climate*, *29*(10), 3831–3840. <https://doi.org/10.1175/JCLI-D-15-0087.1>
- Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., et al. (2011). *The new ECMWF seasonal forecast system (System 4)*. European Centre for Medium-Range Weather Forecasts. Retrieved from <https://www.ecmwf.int/en/elibrary/11209-new-ecmwf-seasonal-forecast-system-system-4>
- Nikulin, G., Asharaf, S., Magariño, M. E., Calmanti, S., Cardoso, R. M., Bhend, J., et al. (2018). Dynamical and statistical downscaling of a global seasonal hindcast in eastern Africa. *Climate Services*, *9*, 72–85. <https://doi.org/10.1016/j.cliser.2017.11.003>
- Palmer, T. N. (2002). The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society: A Journal of the Atmospheric Sciences, Applied Meteorology and Physical Oceanography*, *128*(581), 747–774. <https://doi.org/10.1256/0035900021643593>
- Pechlivanidis, I., Crochemore, L., Rosberg, J., & Bosshard, T. (2020). What are the key drivers controlling the quality of seasonal streamflow forecasts? *Water Resources Research*, *56*(6), e2019WR026987. <https://doi.org/10.1029/2019WR026987>
- Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., & Ziese, M. (2018). GPCP monitoring product: Near real-time monthly land-surface precipitation from rain-gauges based on SYNOP and CLIMAT data. https://doi.org/10.5676/DWD_GPCP/MP_M_V6_100
- Siebert, S., Stephenson, D. B., Sansom, P. G., Scaife, A. A., Eade, R., & Arribas, A. (2016). A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability? *Journal of Climate*, *29*(3), 995–1012. <https://doi.org/10.1175/jcli-d-15-0196.1>
- Slingo, J., & Palmer, T. (2011). Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *369*(1956), 4751–4767. <https://doi.org/10.1098/rsta.2011.0161>
- Stockdale, T. N., Molteni, F., & Ferranti, L. (2015). Atmospheric initial conditions and the predictability of the Arctic oscillation. *Geophysical Research Letters*, *42*(4), 1173–1179. <https://doi.org/10.1002/2014GL062681>
- Torralba, V., Doblas-Reyes, F. J., MacLeod, D., Christel, I., & Davis, M. (2017). Seasonal climate prediction: A new source of information for the management of wind energy resources. *Journal of Applied Meteorology and Climatology*, *56*(5), 1231–1247. <https://doi.org/10.1175/JAMC-D-16-0204.1>
- Van Den Dool, H. M., & Toth, Z. (1991). Why do forecasts for “near normal” often fail? *Weather and Forecasting*, *6*(1), 76–85. [https://doi.org/10.1175/1520-0434\(1991\)006<0076:wdfino>2.0.co;2](https://doi.org/10.1175/1520-0434(1991)006<0076:wdfino>2.0.co;2)
- Verfaillie, D., Doblas-Reyes, F. J., Donat, M. G., Pérez-Zanón, N., Solaraju-Murali, B., Torralba, V., & Wild, S. (2020). How reliable are decadal climate predictions of near-surface air temperature? *Journal of Climate*, *34*(2), 1–57. <https://doi.org/10.1175/JCLI-D-20-0138.1>
- Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., & Viterbo, P. (2014). The WFDEI meteorological forcing data set: WATCH forcing data methodology applied to ERA-interim reanalysis data. *Water Resources Research*, *50*(9), 7505–7514. <https://doi.org/10.1002/2014WR015638>
- Weisheimer, A., & Palmer, T. N. (2014). On the reliability of seasonal climate forecasts. *Journal of the Royal Society Interface*, *11*(96), 20131162. <https://doi.org/10.1098/rsif.2013.1162>
- Yang, D., Tang, Y., Yang, X., Ye, D., Liu, T., Feng, T., et al. (2021). A theoretical relationship between probabilistic relative operating characteristic skill and deterministic correlation skill in dynamical seasonal climate prediction. *Climate Dynamics*, *56*(11–12), 3909–3932. <https://doi.org/10.1007/s00382-021-05678-z>