Choices in the Verification of S2S Forecasts and Their Implications for Climate Services

ANDREA MANRIQUE-SUÑÉN, NUBE GONZALEZ-REVIRIEGO, VERÓNICA TORRALBA, AND NICOLA CORTESI

Barcelona Supercomputing Center, Barcelona, Spain

FRANCISCO J. DOBLAS-REYES

Barcelona Supercomputing Center, and Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain

(Manuscript received 26 February 2020, in final form 14 July 2020)

ABSTRACT

Subseasonal predictions bridge the gap between medium-range weather forecasts and seasonal climate predictions. This time scale is crucial for operations and planning in many sectors such as energy and agriculture. For users to trust these predictions and efficiently make use of them in decision-making, the quality of predicted near-surface parameters needs to be systematically assessed. However, the method to follow in a probabilistic evaluation of subseasonal predictions is not trivial. This study aims to offer an illustration of the impact that the verification setup might have on the calculation of the skill scores, thus providing some guidelines for subseasonal forecast evaluation. For this, several forecast verification setups to calculate the fair ranked probability skill score for tercile categories have been designed. These setups use different number of samples to compute the fair RPSS as well as different ways to define the climatology, characterized by different time periods to average (week or month). These setups have been tested by evaluating 2-m temperature in ECMWF-Ext-ENS 20-yr hindcasts for all of the initializations in 2016 against the ERA-Interim reanalysis. Then, the implications on skill score values of each of the setups are analyzed. Results show that to obtain a robust skill score several start dates need to be employed. It is also shown that a constant monthly climatology over each calendar month may introduce spurious skill score associated with the seasonal cycle. A weekly climatology bears similar results to a monthly running-window climatology; however, the latter provides a better reference climatology when bias adjustment is applied.

1. Introduction

Subseasonal predictions aim to fill the gap between short- to medium-range meteorological forecasts (up to ~10 days) and seasonal predictions (up to one year ahead). At the meteorological scale, a correct characterization of the initial state of the atmosphere is crucial for an accurate forecast. At seasonal time scales, however, the influence of the atmospheric initial conditions is progressively reduced and other slowly evolving components of the system, which are also initialized, gain importance, such as the ocean and the land surface. The subseasonal time scale falls between these two time ranges, when the effect of the atmospheric initial conditions has substantially decreased and the influence of the ocean state is still not dominant (Brunet et al. 2010; Mariotti et al. 2018). While El Niño–Southern Oscillation (ENSO) is the most important source of predictability at seasonal time scales (Doblas-Reyes et al. 2013), the Madden–Julian oscillation (MJO) has been recognized as the main source of predictability in the tropics and extratropics (Zhang 2013) at the subseasonal range. Other potential sources of predictability are the soil moisture (Koster et al. 2011), the state of the ocean, the snow cover, the sea ice state (Jeong et al. 2013; Thomas et al. 2016), the stratosphere–troposphere interactions (Domeisen et al. 2020) and tropical–extratropical teleconnections (Vitart and Robertson 2015).

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

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-20-0067.s1.

Corresponding author: Andrea Manrique-Suñén, andrea.manrique@bsc.es

DOI: 10.1175/MWR-D-20-0067.1

Although subseasonal predictions began in the 2000s and are currently issued by several operational centers across the world, there is not a common strategy for the generation of subseasonal forecast products at this time scale. This is partly because in some cases subseasonal predictions have derived from numerical weather prediction systems by performing longer integrations up to several weeks [e.g., European Centre for Medium-Range Weather Forecasts ensemble extended forecast, ECMWF-Ext-ENS (Vitart 2004; Vitart et al. 2008)], while in other cases subseasonal predictions are generated by the same system that produces seasonal climate predictions [e.g., National Centers for Environmental Prediction Climate Forecast System (NCEP CFS; Saha et al. 2014)]. Operational forecast centers do not only perform subseasonal simulations for the future, but also produce hindcast datasets. Hindcasts are forecasts issued for past dates using the same prediction system over a sufficiently long sample (around 20 years typically). Hindcasts are needed to estimate the quality of the predictions and also to identify and correct model biases that prevent the direct use of the predictions. The approaches followed by each center to produce the hindcasts are very diverse, depending on the strategies already developed for their medium-range or seasonal systems, but also as a function of the computational resources available to them. This heterogeneity in the systems configurations complicates the design of a common strategy for the generation of homogeneous subseasonal forecast products and a systematic comparison of their forecast quality. There are initiatives to facilitate and foster the use of subseasonal predictions. The Subseasonal to Seasonal Prediction (S2S) Project is an international initiative promoting subseasonal prediction and its exploitation for scientific purposes (Robertson et al. 2015). The project maintains a database that makes available subseasonal predictions and hindcasts from 11 operational centers (Vitart et al. 2017). With a focus on research to operations transfer, the Subseasonal Experiment (SubX) is a project in service of developing better operational subseasonal forecasts with a database including real-time forecasts from seven operational centers (Pegion et al. 2019).

Subseasonal predictions have attracted much interest in many societal sectors because operational decisions are often taken with an anticipation of 1–4 weeks. This is the case for energy, agriculture, public health, humanitarian aid preparedness, and so on. A review of the potential applications of S2S forecasts in industry and societal sectors can be found in White et al. (2017). For an appropriate uptake of subseasonal predictions in a decision-making context, it is necessary to understand the probabilistic nature of these predictions as well as their expected quality. Similar to seasonal predictions, subseasonal predictions express the likelihood of changes with respect to the normal climatic conditions. For example, a prediction that provides the probabilities of the weekly average temperature being in each of the climatological terciles categories: colder than normal, normal, or warmer than normal. The performance of such a probabilistic prediction needs to be evaluated with probabilistic measures (Jolliffe and Stephenson 2011). Specifically, the added value of the predictions is assessed by comparing them to the forecasts that would otherwise be employed by users. This is quantified with skill scores, which are a relative measure of the quality compared to a benchmark or reference forecast. The reference forecast for a target period can be persistence (i.e., persisting the conditions at the time of issuing the forecast) or a climatological forecast (i.e., the mean observed conditions of the target period averaged over a set of years in the past).

The skill of subseasonal predictions has been shown to vary throughout the year (Weigel et al. 2008). From the point of view of the user it is advantageous to know the seasonal variation of skill, to increase the usability of skillful predictions. A forecast quality assessment for periods shorter than a year at subseasonal time scales is not a trivial procedure. One of the main challenges resides in the small sample size available for a probabilistic evaluation, due to the small number of ensemble members and short hindcast periods. The forecast quality assessment can be performed on the forecasts or on the hindcasts. Forecasts date back to no longer than 20 years and they experience changes every time the system is upgraded. The advantage of a hindcast is that it is produced with a 'frozen' version of the system, providing around 20 years of consistent forecasts for evaluation, although hindcasts usually have fewer ensemble members than the forecasts. To bypass the problems that arise from the reduced sample size, different strategies have been employed in previous studies, for example it is common to aggregate several start dates (across a month, a season, or a full year), to compute a skill score. ECMWF-Ext-ENS was first evaluated by Vitart (2004) based on the 51 member forecasts from 2002–03 for surface variables: 2-m temperature, precipitation and mean sea level pressure. Despite the short verification period, the study revealed that, for days 12–18, the ECMWF-Ext-ENS performed generally better than both climatology and persistence. In this study, hindcasts spanning 12 years with 5 members were also employed to assess the potential seasonality of the skill using the ROC score. Weigel et al. (2008) designed a verification setup to jointly assess hindcast and forecasts and provided a probabilistic evaluation of 2-m temperature predictions

of ECMWF-Ext-ENS system as of 2006. The skill score employed, the discrete ranked probability skill score, was specially developed to be insensitive to the sampling errors due to small ensemble size (Weigel et al. 2007). In this analysis, skill was computed for each of the 52 target weeks of the year and then was aggregated annually, per month or per season. These results highlighted the seasonal and regional dependencies of the system's skill. Employing the same skill metric Vitart (2014) showed the positive evolution of the skill from 2002 to 2011 using a common period of hindcasts issued during each year. All of these studies used a reanalysis as observational reference; more recently, Monhart et al. (2018) carried out an evaluation of ECMWF-Ext-ENS against 2-m temperature and precipitation observations from in situ stations.

Another important issue in a forecast quality assessment is the construction of the climate distribution, which has important implications in the definition of a forecast product and its probabilistic forecast evaluation (Jolliffe and Stephenson 2011). The climate distribution is calculated both for the observational reference and for the forecasts over a defined number of years (typically the hindcast period). In the case of forecasts, the climate distribution varies with the lead time. The mean of the climate distribution is the climatology and it is used to calculate the anomalies which are obtained as deviations from this mean value. Inconsistencies in the computed climatology lead to differences in the anomalies, a list of relevant aspects related to the construction of the climatology is enumerated in the appendix of Anderson et al. (1999). The difference between the observed climatology and the forecast climatology indicates the model bias, which varies with lead time. The climate distribution is also used to derive the quantiles that delimit the categories (i.e., terciles). In addition, the construction of the climatology may affect the climatological forecast used as benchmark for the skill scores. Different approaches exist to construct the reference climatology, the eligibility of one or other may depend on the hindcast's start dates schedule and the intended use. These approaches go from simple average across years that may include weighting or smoothing (Pegion et al. 2019) to fitting methodologies (Epstein 1988; Narapusetty et al. 2009). For instance, NCEP employs harmonics to produce a daily climatology (Schemm et al. 1998; Johansson et al. 2007; Saha et al. 2014). The harmonic climatology introduces an inherent smoothing and interpolation which is useful when the start dates of the hindcast do not match those of the forecasts (Tippett et al. 2018). When the hindcasts are initialized on the same dates, the climatology can be computed by simple averaging over the hindcast years. However, the length of the aggregation period to average is not set; it could be monthly as is done in the context of seasonal forecasting, or alternatively it could be weekly. The second option is coherent with the forecast target period of subseasonal forecasts; however, the amount of available data is constrained by the frequency of start dates in the hindcast and the number of years available. To enhance the robustness of the climatology, it is possible to increase the period used to compute it by taking a temporal window centered on the target week. For example, a 3-week window was employed by Weigel et al. (2008) and Vigaud et al. (2017) to compute the tercile categories for 2-m temperature and for precipitation, respectively. Another aspect in the forecast evaluation in which the definition of reference climatology plays a role is the bias adjustment. Monhart et al. (2018) tested two different bias adjustment techniques on ECMWF-Ext-ENS forecasts and used a weekly climatology as reference. These examples show that although climatology is an important concept for probabilistic forecast evaluation, there is not a common procedure for its construction in subseasonal time scales.

Subseasonal forecast assessment studies imply some choices in the evaluation procedure, both to increase the sample size to compute the skill scores and in the way the climatology is computed. These choices may influence the resulting skill score as well as the bias adjustment procedure. This work is the first to illustrate the real impact that choices in the verification setup have on the results in terms of probabilistic skill scores. The aim of this study is therefore to highlight these elements and bring awareness of their implications in the context of subseasonal probabilistic verification. The results are relevant for the development of climate services at subseasonal time scales.

With this purpose, several possible setups for a probabilistic forecast evaluation of weekly 2-m temperature predictions (hindcast) from ECMWF-Ext-ENS are designed and intercompared. The implications that the different choices have on the skill scores are analyzed and explained. Particularly, the fair ranked probability skill score for tercile categories is considered, but also the fair continuous ranked probability skill score. This study analyses the sensitivity of forecast skill scores to choices in (i) the sample size (number of forecast–observation pairs) used for verification, (ii) the application of a bias adjustment to the forecasts, and (iii) the construction of the climatology (which affects the computation of anomalies, the benchmark forecast for skill scores and the bias adjustment).

The text is structured as follows, the data employed and the verification setups are described in section 2, the results and analyses of the implications of the verification setups on the skill scores are presented in section 3,

and the conclusions and discussion are presented in section 4.

2. Method

a. Datasets

Global subseasonal predictions of 2-m temperature from ECMWF-Ext-ENS system (Vitart 2004; Vitart et al. 2008) were used. This prediction system is the same as employed for medium-range forecasts, each forecast run consists of 51 ensemble members. To produce the extended forecast, the simulations initialized on Mondays and Thursdays at 0000 UTC are allowed to run for an extended time range of 46 days with coupled ocean–atmosphere. For each extended run, an "on the fly" hindcast of 11 members is computed with the same model version, initializing on the same date for the 20 previous years. For example, for the forecast of 4 January 2016, 20 hindcast runs are computed initializing on 4 January 1996, 4 January 1997, . . . , 4 January 2015. For this study, 20 years of hindcast associated with the predictions issued every Monday and Thursday in 2016 were considered, therefore spanning the period 1996–2015. The data were obtained from the S2S database (Vitart et al. 2017), with spatial resolution of 1.5° × 1.5° (http://s2sprediction.net/). The forecasts are provided as daily averages which are then averaged weekly for verification. Part of the analysis will focus on the predictions initialized in April, although also an annual analysis is performed to examine specific aspects related to seasonality. It should be noted that throughout the text, when the term forecast or prediction is used it refers to hindcast runs.

The reference for forecast evaluation and bias adjustment is the ERA-Interim reanalysis (Dee et al. 2011). This reanalysis is available for the 1979–2019 period with a horizontal resolution of T255 (~80 km). For this study the data for the period 1996–2015 have been interpolated to the S2S grid (1.5° × 1.5°) prior to computing the point by point validation and bias adjustment (when applied).

b. Forecast quality

To assess the quality of the ECMWF-Ext-ENS predictions of the 2-m temperature weekly averages against ERA-Interim reanalysis probabilistic skill scores were employed. A skill score is a relative performance measure that quantifies the added value of the predictions with respect to a set of standard control or reference forecasts (Wilks 2011), defined as

$$\mathrm{SS} = \frac{\mathrm{score} - \mathrm{score}_{\mathrm{ref}}}{\mathrm{score}_{\mathrm{perf}} - \mathrm{score}_{\mathrm{ref}}}, \qquad (1)$$

where score measures the performance of the forecast being evaluated, $\mathrm{score}_{\mathrm{ref}}$ measures the performance of the reference forecast, and $\mathrm{score}_{\mathrm{perf}}$ is the score that would be achieved by a perfect forecast. A skill score ranges from $-\infty$ to 1, where a negative value implies that the forecast is worse than the reference forecast and a positive value indicates an improvement with respect to the reference forecast. In the present study, the reference forecast is the climatological forecast, which is based on forecasting the mean value of the variable in the ERA-Interim reanalysis over the hindcast period. In forecast evaluation of seasonal predictions, the climatological forecast is computed as monthly or seasonal averages, as this time period matches with the forecast aggregation period. In the case of subseasonal predictions, for which the typical forecast aggregation period is one week, several approaches have been tested as described in the next subsection.

The probabilistic skill scores that have been employed are the ranked probability skill score (RPSS) and the continuous ranked probability skill score (CRPSS). The RPSS is the skill score associated with the RPS (ranked probability score), which is the sum of the square differences of the cumulative probabilities of forecast and observations for multiple-category ordinal events (Epstein 1969; Wilks 2011). In this case, three categories are defined by the 33rd and 66th percentiles of the climate distribution. Therefore, the predicted probabilities refer to the probability of temperatures being above-normal, normal, or below-normal conditions. A perfect forecast would predict the observed tercile with 100% probability and would score a RPS of zero. The RPS for a collection of forecast–observation pairs is the average of the individual scores. The CRPSS is the skill score associated with the CRPS, which measures the difference between predicted and observed cumulative distributions. Like in the case of RPS, a CRPS equal to zero denotes a perfect score and the CRPS for a collection of forecast–observation pairs is the average of the individual scores (Hersbach 2000). In this work the fair versions of the two skill scores, fair CRPSS and fair RPSS, introduced by Ferro (2014), have been employed. These scores are designed to reward ensembles with members that behave as though they and the verifying observation are sampled from the same distribution, thus compensates for small ensemble sizes and allows to compare systems with different ensemble size or hindcast and forecast of the same system. The sampling uncertainty of the skill score values has been assessed according to Bradley et al. (2008) for a confidence level of 95%. If the forecasts–observation pairs are identically distributed, but significantly correlated, the sample size is effectively reduced, and this method would underestimate the true sampling uncertainty (Bradley et al. 2008). Although the discussion focuses on fair RPSS (and in some case

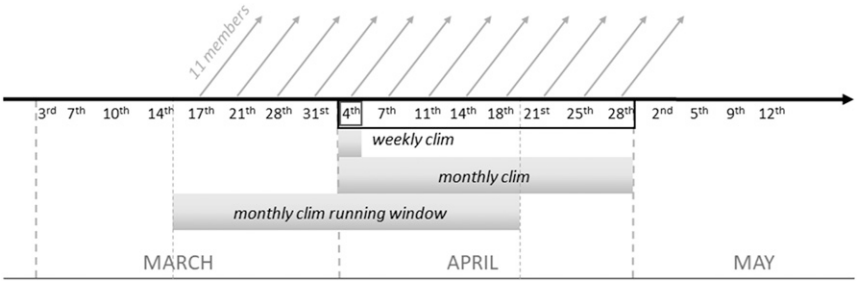

FIG. 1. Schematic showing the setup of ECMWF-Ext-ENS hindcast for 2016 start dates. The numbers along the timeline indicate the start dates (Mondays and Thursdays). The two options to aggregate start dates to compute skill scores are shown by rectangles outlined with a black solid line: single start date (4 April) or monthly start dates (all Mondays and Thursdays in April). The three different options to compute climatology are indicated by the filled gray boxes: weekly climatology, monthly climatology, and monthly climatology with running window.

CRPSS), a deterministic score has been computed for completeness, the correlation of the anomaly of the ensemble mean with respect to the observations (Ens Corr). For the computation of skill scores, the s2dverification (Manubens et al. 2018) and SpecsVerification R software packages have been employed (https://CRAN.R-project.org/package=s2dverification; https://CRAN.R-project.org/package=SpecsVerification).

c. *Forecast verification setups*

At subseasonal time scales, the forecast target period is typically 7 days and thus the forecasts are provided as weekly averages. The way to aggregate the forecast weeks can vary, some studies begin with forecast week 1 coverings days 1–7 (Vigaud et al. 2017) while in others forecast week 1 spans days 5–11 (e.g., Vitart 2004, 2014; Weigel et al. 2008). In this study the second definition is followed, and four different forecast weeks are considered: week 1 (days 5–11), week 2 (days 12–18), week 3 (days 19–25), and week 4 (days 26–32). With this convention, the first week is embedded in what is considered medium-range (up to 10 days) but the second week is completely beyond day 10 and initiates the extended range or subseasonal range (Vitart 2004). The analysis focuses on week 2, and results for the other forecast weeks in selected cases are included in the online supplemental material.

In this analysis the forecast skill is evaluated on the hindcast anomalies of 2-m temperature weekly averages. The performance measures (fair RPSS, fair CRPSS, and Ens Corr) have been computed for each grid point and for each of the four forecast weeks separately. The probabilistic verification of a subseasonal forecast presents many challenges as explained in the introduction, implying some choices related to the sample size for the computation of the skill scores and the construction of the climatology (Fig. 1) as well as the implementation of a bias adjustment procedure, if applied. In this work, eight different setups (summarized in Table 1) have been tested and their impact on the skill scores values has been investigated. A complete description of different aspects of the verification setups is detailed as follows.

1) NUMBER OF FORECAST–OBSERVATION PAIRS FOR SKILL SCORES

This aspect of the verification relates to the size of the sample size (or the number of forecast–observation pairs) needed to compute meaningful skill scores. The aim is to assess whether it is enough to use the hindcasts that start on the same day of the year or whether it is

TABLE 1. Setups for forecast verification (the numbers correspond to the panels in Fig. 2). Number of forecast–observation pairs is the number of aggregated start dates per year for the skill score calculation, climatology is the number of aggregated start dates used to define the climatology, and bias adjustment is whether a simple bias correction is applied to the forecasts.

Forecast verification setup	No. of forecast–observation pairs	Climatology	Bias adjustment
1	Single start date (1 start date)	Weekly (1 start date)	Raw
2	Single start date (1 start date)	Weekly (1 start date)	Simple bias correction
3	Monthly (8/9 start dates)	Weekly (1 start date)	Raw
4	Monthly (8/9 start dates)	Weekly (1 start date)	Simple bias correction
5	Monthly (8/9 start dates)	Monthly (8/9 start dates)	Raw
6	Monthly (8/9 start dates)	Monthly (8/9 start dates)	Simple bias correction
7	Monthly (8/9 start dates)	Monthly running window (9 start dates)	Raw
8	Monthly (8/9 start dates)	Monthly running window (9 start dates)	Simple bias correction

beneficial to increase the sample by pooling hindcasts whose start dates are nearby. Two different approaches have been tested to evaluate how the number of samples affects the values of skill scores.

The first approach involves a single start date. In this case the evaluation is performed for each forecast initialized on a specific start date over the 20 years of hindcast. This provides a total of 20 forecast–observation pairs to compute one skill score per start date. For example, in the case of 4 April the hindcasts initialized on 4 April from 1996 to 2015 are used (see Fig. 1, indicated by the black square around 4 April). This approach corresponds to verification setups 1 and 2 in Table 1.

The second approach uses monthly start dates. To increase the sample size, forecasts from all of the semiweekly start dates initialized in the same month for all the hindcast period are evaluated jointly, giving one skill score for each month. Since ECMWF-Ext-ENS forecasts are issued every Monday and Thursday, there are 8 or 9 start dates depending on the month. Consequently, for a 20-yr hindcast, this approach brings together 160 or 180 forecast–observation pairs to compute one skill score per month. For example, corresponding to the Mondays and Thursdays of the month of April 2016, the start dates 4, 7, 11, 14, 18, 21, 25, and 28 taken over the 1996–2015 period make a total of 160 forecast–observation pairs (see Fig. 1, indicated by the black rectangle). This aggregation of start dates is applied in verification setups 3–8 in Table 1.

2) AGGREGATION PERIOD FOR THE COMPUTATION OF THE CLIMATOLOGY

The definition of the climatology has important implications in forecast verification. The weekly anomalies used for verification are computed as deviations from the climatology over the 20 years of hindcast (1996–2015). Likewise, the weekly anomalies of the observations are calculated as deviations from the observed climatology over the same period. The definition of the climatology affects as well the climatological forecast benchmark used as reference for the skill scores and the adjustment of the biases, which is performed with respect to the observed climatology. To analyze the impact that the construction of the climatology has on the skills scores, three different approaches have been tested. They are described below and represented with the gray boxes in the schematic in Fig. 1. For each case, the same method has been employed to compute the climatology in the reanalysis and in the hindcast. In the case of the hindcast, the climatology is computed for each forecast time independently (weeks 1–4).

Weekly climatology, the most simple approach, uses only one start date every year to compute the weekly climatology (verification setups 1–4 in Table 1). The climatology is computed by averaging the values in the target week for the 20 years of hindcast. This generates a sample size of 20 for the observations. In the case of the hindcasts, with 11 ensemble members, the sample size is $20 \times 11 = 220$. This approach is indicated as *weekly clim* in Fig. 1.

The monthly climatology approach aims to increase the sample size to produce the climatology by employing forecast runs from all the start dates that fall within a calendar month (verification setups 5 and 6 in Table 1). Since ECMWF-Ext-ENS prediction system is launched on Mondays and Thursdays, there are eight or nine start dates per month. The sample size used for the estimation of the climatology therefore increases to at least $8 \times 20 = 160$ for the observed climatology and $8 \times 20 \times 11 = 1760$ for the hindcast climatology. This approach is indicated as *monthly clim* in Fig. 1 and in this example takes the runs initialized on 4, 7, 11, 14, 18, 21, 25, and 28 April.

The third approach is monthly climatology with running window. As an alternative approach, a running window centered in the target week has been tested (verification setups 7 and 8 in Table 1). This is done by employing four start dates before and four after the start date for which the climatology is computed (including both Mondays and Thursdays). This approach collects nine start dates, producing a sample size of $9 \times 20 = 180$ for the observed climatology and $9 \times 20 \times 11 = 1980$ for the hindcast climatology. This methodology is indicated as *monthly clim running window* in Fig. 1. This approach employs a similar amount of data as the monthly climatology approach, with the difference that in this case the monthly period is centered on the target week.

3) BIAS ADJUSTMENT

Climate predictions have different statistical properties to the observed climate (i.e., mean, standard deviation), for this reason a bias adjustment or calibration against a reference dataset is needed (Doblas-Reyes et al. 2005). This postprocessing is particularly important when essential climate variables are used as inputs in impact models or to compute climate indices (Torralba et al. 2017). Bias adjustment procedures aim to correct the mean model biases and under/over dispersion of the ensemble members. Many statistical methods have been developed for calibration and bias adjustment (variance inflation, quantile mapping, etc.) although typically they have been designed and applied on seasonal forecasts (Manzanas et al. 2019). For subseasonal forecasts these methodologies have had very little testing (Vigaud et al. 2017; Monhart et al. 2018).

In this study we focus on a simple method for bias adjustment and implement it in three different ways. These correspond to adjusting the predictions to three

observed climate distributions defined by the three alternatives to construct the climatology explained above. The effect of these implementations in forecast skill scores is analyzed. The method chosen is a simple bias correction, as applied on seasonal predictions of wind speed (Torralba et al. 2017) and of temperature and precipitation (Manzanas et al. 2019). This method assumes a Gaussian distribution for both the forecasts and the observations, and corrects the mean and the standard deviation of the forecasts to those of the observations. The ECMWF-Ext-ENS 2-m temperature hindcasts were adjusted to ERA-Interim reanalysis data. To account for the increase of the bias with lead time the method is applied to each forecast week (from 1 to 4) independently. Additionally, a leave-one-out cross-validation framework is used to ensure that it replicates an operational context in which information of the observations of a particular week is not available when adjusting the prediction of that specific week. The bias adjustment procedure is described by the following equation:

$$y_{ij} = (x_{ij} - \overline{x})\frac{\sigma_{\text{ref}}}{\sigma_e} + \overline{o}, \tag{2}$$

where y_{ij} is the adjusted weekly prediction of member j for the year i, $(x_{ij} - \overline{x})$ is the weekly anomaly obtained by subtracting the ensemble mean $\overline{x}$ over the hindcast period to the predicted weekly value x_{ij}, and σ_{ref} and σ_e are the interannual standard deviations of the reference dataset and of the forecasts including the ensemble members, respectively. The climatological reference mean is denoted as $\overline{o}$. The aggregation period used to compute the means and standard deviations is the same as that of the chosen climatology (i.e., weekly, monthly, or monthly running window) (as indicated in column 3 of Table 1).

3. Results

Probabilistic verification of the ECMWF-Ext-ENS subseasonal forecasts of the 2-m temperature is performed for each of the different verification setups described in Table 1. Results for the fair RPSS for the month of April and forecast week 2 (days 12–18) are presented in Fig. 2. Overall, some common areas of enhanced skill scores can be identified in all panels of this figure although panels 1 and 2 are noisier than the rest (see the following subsections for further details). In general, higher skill score values are found over sea than over land. The most skillful areas over the ocean are the eastern equatorial Pacific (El Niño region), eastern North Pacific (Gulf of Alaska), tropical Atlantic (north and south) and the North Sea. Over land, the areas that show highest RPSS are the Amazon basin, central Africa, Australia, and to a lesser extent, the eastern part of North America and northeast Asia. These regional patterns of RPSS for temperature are generally consistent with previous studies that evaluated older versions of ECMWF-Ext-ENS against persistence or ERA-40 reanalysis (Vitart 2004; Weigel et al. 2008). Results of the fair CRPSS and the Pearson correlation of the ensemble mean anomaly for the same month and forecast time are included in Figs. S1 and S2 in the online supplemental material, respectively. In general, fair CRPSS values (Fig. S1) show similar patterns to fair RPSS (Fig. 2) but with slightly lower skill values. The correlation (Fig. S2) is positive and mostly above 0.5 globally, except for verification setups 1 and 2, for which some negative correlations are found in some specific areas. In the following subsections, the effect of the different aspects of the verification setups on these skill scores is analyzed.

a. *Effect of number of forecast–observation pairs for skill scores*

While fair RPSS in panels 1 and 2 in Fig. 2 was computed using one single start date per year (specifically 4 April) and thus employing 20 forecast–observation pairs from the hindcast for this specific date, the remaining panels (Fig. 2, panels 3–8) included all the start dates within April to compute the skill score (160 forecast–observation pairs). Thus, the skill scores in panels 1 and 2 are not directly comparable with the others, but the resulting fair RPSS maps give an indication of the robustness of the skill scores.

When only one single start date per year is used, the fair RPSS map presents a very noisy spatial distribution, as seen in Fig. 2 panels 1 and 2 for the 4 April start date (the other single start dates in April result in equally noisy fair RPSS maps). This implies that with a 20-yr hindcast, one single start date per year is not enough to produce robust probabilistic skill scores. On the contrary, the rest of the panels in Fig. 2, which employ eight start dates, result in values of fair RPSS more spatially consistent. Similar results are found for fair CRPSS (supplemental Fig. S1). This shows that concatenating several contiguous start dates is a good approach to compensate for the short hindcast size and produce robust skill scores.

b. *Effect of climatology aggregation on simple bias correction*

Although the simple bias correction removes the lead-dependent biases and corrects the variance with respect to the reference, skill scores may be slightly degraded due the cross validation (Barnston and van den Dool 1993). In Fig. 2, the panels on the left-hand side show the RPSS

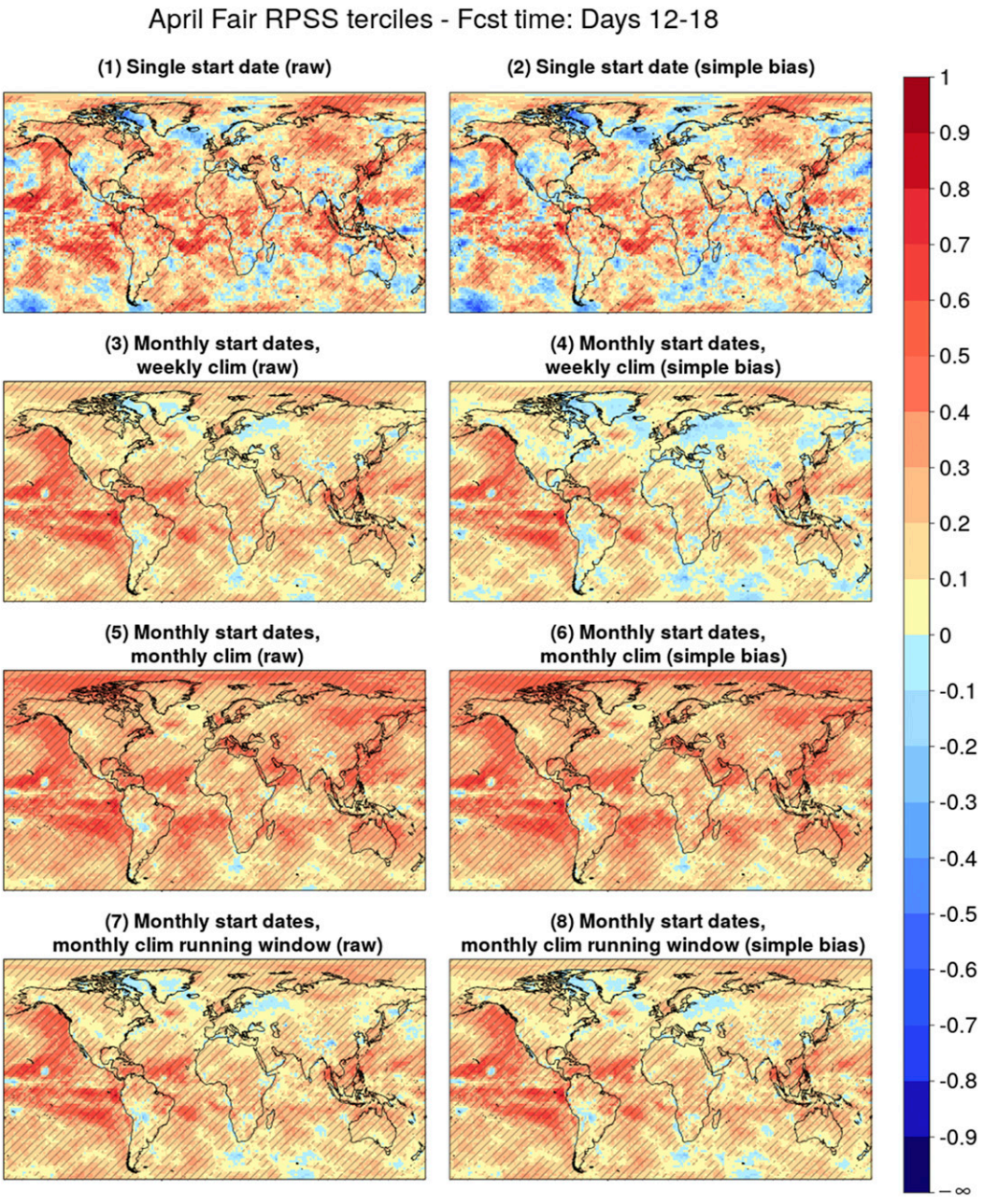

FIG. 2. Fair RPSS for terciles of 2-m temperature hindcasts from ECMWF-Ext-ENS in the 1996–2015 period for forecast days 12–18 (week 2) in April. In panels 1 and 2 the fair RPSS is computed using a single start date (4 April), whereas in panels 3–8 the skill scores are computed using all start dates in a month (eight start dates: 4, 7, 11, 14, 18, 21, 25 and 28 Apr). The climatology is constructed weekly in panels 1–4, monthly in panels 5 and 6, and with a monthly running window in panels 7 and 8. (left) Results for raw forecasts, and (right) results for forecasts that have been adjusted with a simple bias correction. The verification setups used for each panel are summarized in Table 1. The ERA-Interim reanalysis has been employed as reference. Hatching indicates skill scores that are significantly higher than zero (95% confidence level).

of raw forecasts while the corresponding panels on the right-hand side show the RPSS after the simple bias correction. The effect of the simple bias correction on the fair RPSS can be identified by visual inspection. The drop in skill score is particularly evident when a weekly climatology is employed in the simple bias correction (Fig. 2, panels 1–4). Focusing on panel 3 some areas of negative RPSS which can be identified for the raw

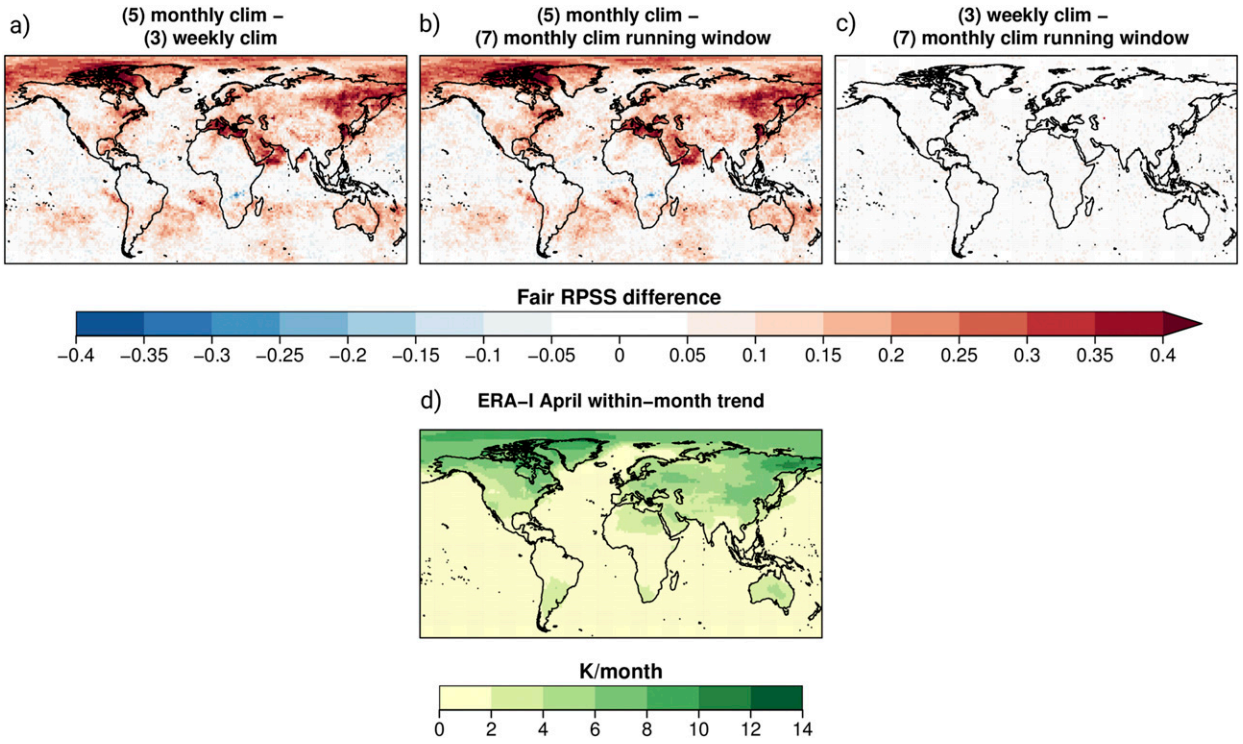

FIG. 3. Differences in fair RPSS values for 2-m temperature for April, forecast days 12–18: (a) employing a monthly climatology vs a weekly climatology (panel 5 minus panel 3 of Fig. 2), (b) employing a monthly climatology vs a running-window monthly climatology (panel 5 minus panel 7 of Fig. 2), (c) employing a weekly climatology vs a running-window monthly climatology (panel 3 minus panel 7 of Fig. 2), and (d) within-month trend in 2-m weekly temperatures in April (years 1996–2015).

forecast (eastern Europe, Greenland, northeast Asia, and South America), increase their spatial extent after the simple bias correction, as seen in panel 4) The degradation of fair RPSS when applying the simple bias correction based on a weekly climatology is of up to 0.1 depending on the area (see maps of differences in fair RPSS between raw and bias adjusted predictions in Figs. S3a,b in the online supplemental material). On the other hand, the use of a climatology based on a larger sample for the simple bias correction, either with a monthly climatology (Fig. 2, panels 5 and 6) or on a monthly running-window climatology (Fig. 2, panels 7 and 8) results in little changes in the fair RPSS. Although there is still some small degradation, changes in skill score are below 0.05 (supplemental Figs. S3c,d).

These results show that when applying a simple bias correction to weekly averages, it is beneficial to employ a climatology that has been generated from an aggregation period that is longer than just one week. This is most likely because a larger aggregation period, reduces the sampling uncertainty and produces more robust estimates of the mean climatology and standard deviation of the climatology used in the simple bias correction procedure [Eq. (1)]. The same conclusions can be derived from the maps of fair CRPSS (supplemental Fig. S1).

In addition, similar results are found in terms of reliability (See Fig. S4 in the online supplemental material), a monthly running-window climatology for the simple bias correction (verification setup 8) results in similar or improved reliability than the raw forecasts (verification setup 7), while a weekly climatology (verification setup 4) degrades reliability with respect to raw forecasts (verification setup 3). This effect is also related to sampling uncertainty (Tippett et al. 2014).

c. *Effect of climatology aggregation on skill scores*

To analyze the effect of the climatology aggregation on fair RPSS, we focus on verification setups 3, 5 and 7 (Table 1), all of which employ the same start dates for verification. Among them, the highest fair RPSS is found for verification setup 5, which uses a monthly climatology (Fig. 2, panel 5), while the weekly climatology and the monthly running-window climatology (Fig. 2, panels 3 and 7) result in lower values of fair RPSS. This is evident in the maps of differences of fair RPSS (Fig. 3); fair RPSS differences between using verification setup 5 and verifications setups 3 and 7 are

positive and very similar (Figs. 3a,b), while the differences between fair RPSS with verification setups 3 and 7 are close to zero (Fig. 3c). This comparison implies that verification setups 5 and 7, which use the same size of aggregation period to compute the climatology but displaced with respect to the target week (see Fig. 1, *monthly clim* and *monthly clim running window*), result in important differences in fair RPSS (Fig. 3b). On the other hand, verification setups 3 and 7, which use different sizes of aggregation period but both centered on the target week (see Fig. 1, *weekly* and *monthly clim running window*), result in similar fair RPSS (Fig. 3c).

The spatial distribution of the differences in fair RPSS values between the setup employing the monthly climatology and the other two (Figs. 3a,b) reveals that the largest differences appear in mid- and high latitudes, particularly in the Northern Hemisphere. There are marked differences of up to 0.4 in fair RPSS in the north of Canada, the Mediterranean Sea, the Arabian Peninsula and northeast Asia. The fact that these differences occur at midlatitudes in both hemispheres, although more clearly in the Northern Hemisphere, gives an indication that the representation of the seasonal cycle in the climatology might be playing a role. In fact, the maps of differences in Figs. 3a and 3b bear a strong resemblance with the within month trend of weekly 2-m temperatures throughout April (Fig. 3d). This value has been computed as the difference between the weekly climatology of the first and the last weeks of April (and is presented in absolute values). This explains why the largest differences in Figs. 3a and 3b appeared in the Northern Hemisphere, in particular, the differences found in the north of Canada and northeast Asia. Also the differences in some inland regions in the midlatitudes in the Southern Hemisphere, like Australia and southern South America are explained by the seasonal cycle of temperature throughout April. Nevertheless, the differences found over the Mediterranean Sea and the Arabian Peninsula (Figs. 3a,b) cannot be explained by the imprint of the seasonal cycle of temperatures.

To further investigate the seasonal dependence of the fair RPSS, the spatial average over a limited area in North America (35°–65°N, 240°–280°E) for each month of the year is presented in Fig. 4a (black lines). The *x* axis indicates the month when the forecasts were initialized and the results correspond to forecast days 12–18 (second week). The fair RPSS obtained when considering verification setup 5, with monthly climatology (black solid line on Fig. 4a), is superior to the fair RPSS obtained when using verification setup 7, with the monthly running-window climatology (black dotted line) for all months. Results for verification setup 3, with the weekly climatology, are almost identical to results for verification setup 7 for all months (not shown for clarity; see Fig. S5 in the online supplemental material). The largest differences in the fair RPSS obtained with verification setups 5 and 7 occur in May, September and October, periods with significant changes in temperature due to the seasonal cycle. Furthermore, the values of fair RPSS for September and October obtained with verification setup 5 exceed the fair RPSS obtained for January, February and December, months when other studies have reported the highest skill scores (Weigel et al. 2008; Wang and Robertson 2019). Similar results are found for the other forecast times (Fig. S5). To further investigate where these high fair RPSS values come from, the decomposition of the skill score in the absolute scores [i.e., fair RPS, see Eq. (1)] is analyzed (Fig. 4a, red lines).

The values of fair RPS of ECMWF-Ext-ENS (solid and dotted red lines, Fig. 4a) show an annual evolution equal to that of the fair RPSS (black lines), with verification setup 5 (red solid line) again showing better values than ones for verification setup 7 (red dotted line), particularly in May, September and October (note that fair RPS has the *y* axis on the right-hand side and is negatively oriented (i.e., a perfect forecast would score zero). For the climatological forecasts, their corresponding fair RPS (solid and dotted "R" red lines, Fig. 4a) stay constant throughout the year (around 0.44) independent of the verification setup. The reason for this constant value is that by definition, for categorical forecasts, the climatological forecast predicts the same probability for each category, in this case, one-third for each tercile category. Consequently, the differences in fair RPSS values found between verification setup 5 and 7 cannot be attributed to the construction of the climatological forecast used as benchmark. The other aspect of the assessment in which the choice of the aggregation to compute the climatology may play a role is in the definition of the anomalies. When employing verification setup 5, the weekly anomalies are computed with respect to a monthly climatology, and consequently there might be an imprint of the seasonal cycle in the weekly anomalies, particularly in transition months. Since the assessment is performed on weekly anomalies, the increased fair RPSS found for some transition months with verification setup 5 might be indicating a good match of the seasonal cycle in both forecasts and observations.

In terms of fair CRPSS, verification setup 5 (monthly climatology) also results in higher skill scores than verification setup 7 (monthly running-window climatology) for all months (solid and dotted black lines, Fig. 4b). The annual evolutions of fair CRPSS with the two verifications setups are analogous to those of the fair RPSS (solid and dotted black lines in Fig. 4a), with the largest differences again found for May, September and October. However, in this case, the fair CRPS values computed with

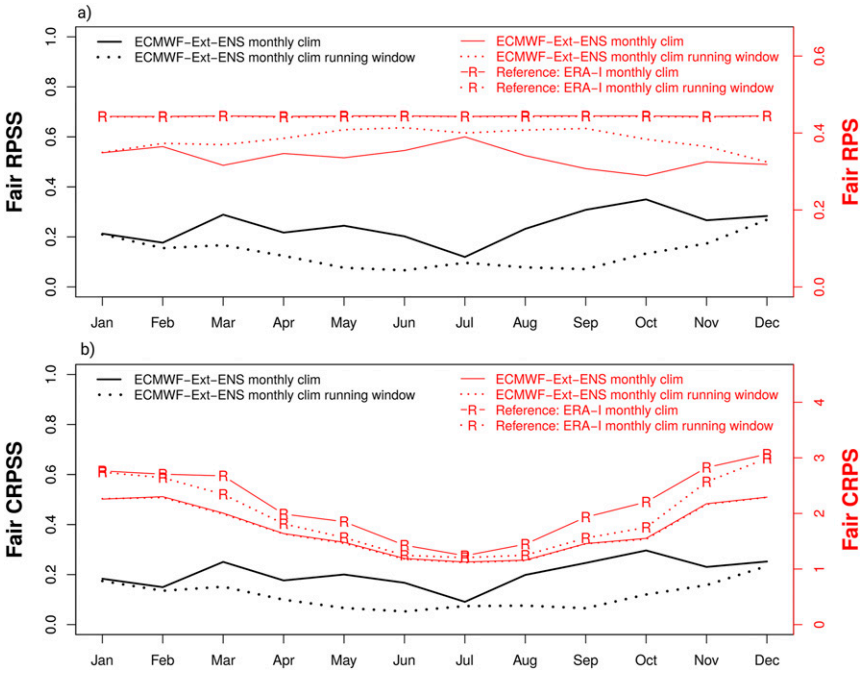

FIG. 4. (a) Annual evolution of fair RPSS for ECMWF-Ext-ENS 2-m temperatures for a region in North America (35°–65°N, 240°–280°E) for days 12–18, with a monthly climatology (verification setup 5) represented by the solid black line, and with a monthly running-window climatology (verification setup 7) represented by the dotted black line. In red, decomposition of the fair RPSS: annual evolution of fair RPS for ECMWF-Ext-ENS, with a monthly climatology (verification setup 5) represented by the solid red line, and with a monthly running-window climatology (verification setup 7) represented by the dotted red line, and annual evolution of fair RPS for the climatological forecast based on ERA-Interim, with a monthly climatology (verification setup 5) represented by the solid red line with "R," and with a monthly running-window climatology (verification setup 7) represented by the dotted red line with "R." (b) As in (a), but for fair CRPSS (black lines) and its decomposition in fair CRPS (red lines).

verification setups 5 and 7 agree for all months (solid and dotted red lines in Fig. 4b), indicating the same score in absolute terms. In this case, the differences in the fair CRPSS arise from differences in the climatological forecasts used as benchmark. The fair CRPS of the climatological forecasts based on a monthly climatology (solid "R" red line) are worse than the ones based on a monthly running-window climatology (dotted "R" red line). Consequently, the apparently better fair CRPSS found when employing verification setup 5, particularly during the transition months, is the result of a comparison against a worse benchmark. The low fair CRPS associated with the climatological forecast in verification setup 5 is a consequence of using a monthly climatology as a forecast for the individual weeks of the month, and therefore not characterizing intramonth differences associated with the seasonal cycle.

To sum up, the breakdown of fair RPSS and fair CRPSS into their components has revealed that the use of a monthly climatology based on a calendar month (verification setup 5) can lead to a misleading inflation of the skill scores for two reasons. In the case of the probabilistic assessment of tercile categories (fair RPSS), the problem originates in the fact that a monthly climatology is used as a reference to compute anomalies for a one week period, while not being representative of a weekly subset, an issue raised by Hamill and Juras (2006). In this case, the seasonal cycle of temperature leads to a common seasonal signal both in the forecast and observed weekly anomalies which is rewarded by the skill scores. This results in unrealistic high skill score values both in the fair RPSS and in the fair RPS, associated with a good forecast of the seasonal cycle seen in the weekly anomalies. On the other hand, in the probabilistic assessment of the full probability distribution forecast, the problem arises from the definition of the climatological forecast used as a benchmark. When using a monthly climatology as forecast (verification setup 5) the climatological forecast is not good for transition months, since again, the monthly climatology is not representative

of the weekly climatology. Thus, the climatological reference is easier to beat by the ECMWF-Ext-ENS forecast, leading to an apparent increase of CRPSS in the transition months. For these reasons, a calendar month climatology as described in verification setup 5 should be avoided when analyzing weekly averages.

Although the annual analysis has focused on verification setups 5 and 7, skill scores obtained with a weekly climatology (verification setup 3) are analogous to those of shown for verification setup 7 in this subsection. However, the disadvantage of verification setup 3 is that the weekly climatology is not robust enough to estimate the parameters needed for the simple bias correction, as demonstrated in the previous subsection.

4. Conclusions and discussion

This work reviews the methodology for probabilistic forecast assessment of subseasonal predictions of 2-m temperatures weekly anomalies. The analysis is performed on hindcasts produced by ECMWF-Ext-ENS during 2006 and uses two common probabilistic skill scores (fair RPSS and fair CRPSS). The predictions are assessed with eight different verification setups that combine various relevant aspects for probabilistic evaluation. These factors are the sample size (number of forecast–observation pairs) used to compute the probabilistic skill scores and the construction of the climatology in terms of the averaging aggregation period and its position, as well as the implications of the construction of the reference climatology when a simple bias correction is applied to the forecasts.

One of the main messages that emerges from this work is to inform the subseasonal forecast community about the relevance of the selection of both the sample size and position to calculate the climatology, independently of whether it is done in an operational context or not. In addition, it emphasizes the importance of providing a detailed and transparent description of the methodology when issuing a forecast product and its corresponding verification. Following the results of this study, it can be concluded that in the computation of a weekly climatology for a variable with a marked seasonal cycle in mid- and high latitudes, such as 2-m temperature in this study, the averaging period should be centered in the target week. Additionally, it is demonstrated that for a correct bias adjustment of weekly forecasts with respect to a reference, the period used to define the reference climatology should be longer than one week, to ensure a robust estimation of the adjustment parameters. In particular, the specific conclusions are detailed as follows.

- Computing a skill score for a single start date from weekly averages forecasts over a 20-yr hindcast does not produce a robust skill score result. It becomes therefore necessary to concatenate several start dates to increase the sample size for a forecast quality assessment (over a month, season, etc.). Since other subseasonal systems have similar hindcast lengths, this result applies to the verification of other systems.
- The reference climatology to bias adjust weekly averages should span a longer period than one week. When applying a simple bias correction to weekly forecasts, the use of a reference climatology based on a month period has shown to degrade less the skill scores than the use of a weekly climatology. This is because a larger sample size reduces the sampling uncertainty and provides more robust statistical parameters employed in the bias adjustment (i.e., climatology and standard deviation).
- The period to compute the climatology should be centered in the target week. The use of a monthly climatology computed on calendar months has shown to lead to misleading skill score in the transition months for both fair RPSS for tercile categories and fair CRPSS. This is an illustration of the issue of the representativeness of the climatology raised by Hamill and Juras (2006). In this case, the monthly climatology is not representative of a one week subset. Interestingly, although the effect on each of the skill scores is the same (an artificial increase of skill scores in the transition months), the reason behind it varies for each skill score measure. In the case of fair RPSS, the increased skill score is indicating a good match in the seasonal cycle of the weekly anomalies in the forecast and reanalysis. On the other hand, for fair CRPSS, the increase in skill score is due to the monthly climatology not being a good estimate for a climatological weekly forecast and therefore easier to beat.

These conclusions have direct implications for a climate service based on subseasonal predictions. First, a climate forecast product itself is defined with reference to the system's climate distribution. For example, in the case of a forecast product that provides the probabilities of each of the three tercile categories, the tercile boundaries will vary depending on the aggregation period used to define the system's climate distribution and consequently will lead to different predicted tercile probabilities. Second, the forecast quality assessment of a prediction product is evaluated in terms of skill scores (in particular the fair RPSS for terciles is widely used) relative to a benchmark such as a climatological reference from a reanalysis, as presented here. As it has been shown, an inappropriate setup can lead to a misleading

judgement of the skill score. This highlights the importance of producing a specific climatology for subseasonal predictions, since the monthly climatology typically used in seasonal predictions is not appropriate for weekly target periods. Finally, this work has shown the importance of using a reference climatology that is larger than one week for a robust bias adjustment, which is a fundamental step in the provision of climate forecasts to stakeholders.

This study is intended to evidence the impact that the construction of the climatology has on the skill scores, in terms of window length (one week versus one month) and the relative position of the period (calendar month versus monthly running window). Therefore, the aggregation periods for the computation of climatology verification setups presented in this study were designed to allow a clean comparison of these elements, independently of the current ECMWF-Ext-ENS real-time schedule. Additionally, for consistency with each verification setup the anomalies and the bias adjustment were computed or applied using the same definition of the climatology. To make a specific recommendation for a real-time implementation of a climate service based on ECMWF-Ext-ENS is out of the scope of this study, since in such an operational setting, the timing when the on-the-fly hindcasts for each real-time forecast are made available imposes a constraint on the size of the window that can be used to construct the climatology for the target week. For instance, the monthly running-window climatology (9 start dates, 4 before and 4 after the forecast initialization) that has shown to give the best results for anomalies and bias adjustment cannot be implemented operationally with the current ECMWF-Ext-ENS setting which provides the hindcast three weeks prior to the real-time forecast. Following these premises, future work will be devoted to defining strategies that are compatible with the operational settings, considering the use of shorter running windows to define the climate distribution.

Acknowledgments. The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under Grants 7767874 (S2S4E) and 641811 (IMPREX). We acknowledge the ECMWF for producing the ERA-Interim reanalysis and ECMWF-Ext-ENS predictions and the WWRP/WCRP S2S Project for making ECMWF-Ext-ENS predictions available through its database (http://s2sprediction.net/). We acknowledge the use of the SpecsVerification (https://CRAN.R-project.org/package=SpecsVerification), s2dverification https://CRAN.R-project.org/package=s2dverification), and startR (https://CRAN.R-project.org/package=startR) R software packages. The authors thank Pierre-Antoine Bretonnière, Margarida Samsó, and Julia Giner for data downloading and processing and Nicolau Manubens, An Chi-Ho, Núria Pérez-Zanón, Javier Vegas, and Lluís Palma for technical support.

Data availability statement. The data employed for this work have been produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). The ERA-Interim reanalysis (Dee et al. 2011) data are available online (https://apps.ecmwf.int/datasets/). The ECMWF-Ext-ENS forecasts are available from the S2S Project database through its two archiving centers: ECMWF (https://apps.ecmwf.int/datasets/data/s2s/levtype=sfc/type=cf/) and CMA (http://s2s.cma.cn/index).

REFERENCES

Anderson, J., H. van den Dool, A. G. Barnston, W. Chen, W. Stern, and J. Ploshay, 1999: Present-day capabilities of numerical and statistical models for atmospheric extratropical seasonal simulation and prediction. *Bull. Amer. Meteor. Soc.*, **80**, 1349–1362, https://doi.org/10.1175/1520-0477(1999)080<1349:PDCONA>2.0.CO;2.

Barnston, A. G., and H. M. van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate*, **6**, 963–977, https://doi.org/10.1175/1520-0442(1993)006<0963:ADICVS>2.0.CO;2.

Bradley, A. A., S. S. Schwartz, and T. Hashino, 2008: Sampling uncertainty and confidence intervals for the Brier score and Brier skill score. *Wea. Forecasting*, **23**, 992–1006, https://doi.org/10.1175/2007WAF2007049.1.

Brunet, G., and Coauthors, 2010: Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. *Bull. Amer. Meteor. Soc.*, **91**, 1397–1406, https://doi.org/10.1175/2010BAMS3013.1.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus*, **57A**, 234–252, https://doi.org/10.3402/tellusa.v57i3.14658.

——, J. García-Serrano, F. Lienert, A. P. Biescas, and L. R. L. Rodrigues, 2013: Seasonal climate predictability and forecasting: Status and prospects. *Wiley Interdiscip. Rev.: Climate Change*, **4**, 245–268, https://doi.org/10.1002/wcc.217.

Domeisen, D. I. V., and Coauthors, 2020: The role of the stratosphere in subseasonal to seasonal prediction: 2. Predictability arising from stratosphere–troposphere coupling. *J. Geophys. Res. Atmos.*, **125**, e2019JD030923, https://doi.org/10.1029/2019JD030923.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985–987, https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2.

——, 1988: A spectral climatology. *J. Climate*, **1**, 88–107, https://doi.org/10.1175/1520-0442(1988)001<0088:ASC>2.0.CO;2.

Ferro, C. A. T., 2014: Fair scores for ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **140**, 1917–1923, https://doi.org/10.1002/qj.2270.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, https://doi.org/10.1256/qj.06.25.

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

Jeong, J.-H., H. W. Linderholm, S.-H. Woo, C. Folland, B.-M. Kim, S.-J. Kim, and D. Chen, 2013: Impacts of snow initialization on subseasonal forecasts of surface air temperature for the cold season. *J. Climate*, **26**, 1956–1972, https://doi.org/10.1175/JCLI-D-12-00159.1.

Johansson, Å., C. Thiaw, and S. Suranjana, 2007: CFS retrospective forecast daily climatology in the EMC/NCEP CFS public server. NOAA Doc., 27 pp., http://cfs.ncep.noaa.gov/cfs.daily.climatology.doc.

Jolliffe, I. T., and D. B. Stephenson, 2011: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed. John Wiley and Sons, 292 pp.

Koster, R. D., and Coauthors, 2011: The second phase of the Global Land–Atmosphere Coupling Experiment: Soil moisture contributions to subseasonal forecast skill. *J. Hydrometeor.*, **12**, 805–822, https://doi.org/10.1175/2011JHM1365.1.

Manubens, N., and Coauthors, 2018: An R package for climate forecast verification. *Environ. Modell. Software*, **103**, 29–42, https://doi.org/10.1016/j.envsoft.2018.01.018.

Manzanas, R., J. M. Gutiérrez, J. Bhend, S. Hemri, F. J. Doblas-Reyes, V. Torralba, E. Penabad, and A. Brookshaw, 2019: Bias adjustment and ensemble recalibration methods for seasonal forecasting: A comprehensive intercomparison using the C3S dataset. *Climate Dyn.*, **53**, 1287–1305, https://doi.org/10.1007/s00382-019-04640-4.

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

Monhart, S., C. Spirig, J. Bhend, K. Bogner, C. Schär, and M. A. Liniger, 2018: Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. *J. Geophys. Res. Atmos.*, **123**, 7999–8016, https://doi.org/10.1029/2017JD027923.

Narapusetty, B., T. Delsole, and M. K. Tippett, 2009: Optimal estimation of the climatological mean. *J. Climate*, **22**, 4845–4859, https://doi.org/10.1175/2009JCLI2944.1.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. *Bull. Amer. Meteor. Soc.*, **100**, 2043–2060, https://doi.org/10.1175/BAMS-D-18-0270.1.

Robertson, A. W., A. Kumar, M. Peña, and F. Vitart, 2015: Improving and promoting subseasonal to seasonal prediction. *Bull. Amer. Meteor. Soc.*, **96** (3), ES49–ES53, https://doi.org/10.1175/BAMS-D-14-00139.1.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, https://doi.org/10.1175/JCLI-D-12-00823.1.

Schemm, J.-K. E., H. M. van den Dool, J. Huang, and S. Saha, 1998: Construction of daily climatology based on the 17-year NCEP/NCAR reanalysis. *Proc. First WCRP Int. Conf. on Reanalyses*, Silver Spring, MD, World Meteorological Organization, 290–293.

Thomas, J. A., A. A. Berg, and W. J. Merryfield, 2016: Influence of snow and soil moisture initialization on sub-seasonal predictability and forecast skill in boreal spring. *Climate Dyn.*, **47**, 49–65, https://doi.org/10.1007/s00382-015-2821-9.

Tippett, M. K., T. DelSole, and A. G. Barnston, 2014: Reliability of regression-corrected climate forecasts. *J. Climate*, **27**, 3393–3404, https://doi.org/10.1175/JCLI-D-13-00565.1.

——, L. Trenary, T. DelSole, K. Pegion, and M. L. L'Heureux, 2018: Sources of bias in the monthly CFSv2 forecast climatology. *J. Appl. Meteor. Climatol.*, **57**, 1111–1122, https://doi.org/10.1175/JAMC-D-17-0299.1.

Torralba, V., F. J. Doblas-Reyes, D. MacLeod, I. Christel, and M. Davis, 2017: Seasonal climate prediction: A new source of information for the management of wind energy resources. *J. Appl. Meteor. Climatol.*, **56**, 1231–1247, https://doi.org/10.1175/JAMC-D-16-0204.1.

Vigaud, N., A. W. Robertson, and M. K. Tippett, 2017: Multimodel ensembling of subseasonal precipitation forecasts over North America. *Mon. Wea. Rev.*, **145**, 3913–3928, https://doi.org/10.1175/MWR-D-17-0092.1.

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

——, 2014: Evolution of ECMWF sub-seasonal forecast skill scores. *Quart. J. Roy. Meteor. Soc.*, **140**, 1889–1899, https://doi.org/10.1002/qj.2256.

——, and A. W. Robertson, 2015: Sub-seasonal to seasonal prediction: Linking weather and climate. *Seamless Prediction of the Earth System: From Minutes to Months*, G. Brunet, S. Jones, and P. M. Ruti, Eds., WMO-1156, World Meteorological Organization, 385–401.

——, and Coauthors, 2008: The new VarEPS-monthly forecasting system: A first step towards seamless prediction. *Quart. J. Roy. Meteor. Soc.*, **134**, 1789–1799, https://doi.org/10.1002/qj.322.

——, and Coauthors, 2017: The Subseasonal to Seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

Wang, L., and A. W. Robertson, 2019: Week 3–4 predictability over the United States assessed from two operational ensemble prediction systems. *Climate Dyn.*, **52**, 5861–5875, https://doi.org/10.1007/s00382-018-4484-9.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **135**, 118–124, https://doi.org/10.1175/MWR3280.1.

——, D. Baggenstos, M. A. Liniger, F. Vitart, and C. Appenzeller, 2008: Probabilistic verification of monthly temperature forecasts. *Mon. Wea. Rev.*, **136**, 5162–5182, https://doi.org/10.1175/2008MWR2551.1.

White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, https://doi.org/10.1002/met.1654.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Zhang, C., 2013: Madden–Julian Oscillation: Bridging weather and climate. *Bull. Amer. Meteor. Soc.*, **94**, 1849–1870, https://doi.org/10.1175/BAMS-D-12-00026.1.