AGU100 ADVANCING EARTH AND SPACE SCIENCE

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE

10.1029/2017JD027923

Key Points:

- We provide a station based verification of downscaled and bias-corrected subseasonal temperature and precipitation predictions for entire Europe
- Promising skill is found for weekly mean temperature up to 19-24 days lead time, but limited to lead days 5-11 for precipitation
- Seasonal and spatial variations in the skill of the forecasts are discussed in the context of subseasonal to seasonal predictions

Correspondence to:
S. Monhart,
samuel.monhart@wsl.ch

Citation:
Monhart, S., Spirig, C., Bhend, J., Bogner, K., Schär, C., & Liniger, M. A. (2018). Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. *Journal of Geophysical Research: Atmospheres*, *123*, 7999–8016. https://doi.org/10.1029/2017JD027923

Received 19 OCT 2017
Accepted 14 JUL 2018
Accepted article online 24 JUL 2018
Published online 15 AUG 2018

Skill of Subseasonal Forecasts in Europe: Effect of Bias Correction and Downscaling Using Surface Observations

S. Monhart[1,2,3], C. Spirig[2], J. Bhend[2], K. Bogner[1], C. Schär[3], and M. A. Liniger[2]

[1]Mountain Hydrology and Mass Movements, Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Birmensdorf, Switzerland, [2]Federal Office of Meteorology and Climatology MeteoSwiss, Zurich, Switzerland, [3]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

Abstract Subseasonal predictions bridge the gap between medium-range weather forecasts and seasonal climate predictions. This time horizon is of crucial importance for many planning purposes, including energy production and agriculture. The verification of such predictions is normally done for areal averages of upper-air parameters. Only few studies exist that verify the forecasts for surface parameters with observational stations, although this is crucial for real-world applications, which often require such predictions at specific surface locations. With this study we provide an extensive station-based verification of subseasonal forecasts against 1,637 ground based observational time series across Europe. Twenty years of temperature and precipitation reforecasts of the European Centre for Medium-Range Weather Forecasts Integrated Forecasting System are used to analyze the period of April 1995 to March 2014. A lead time and seasonally dependent bias correction is performed to correct the daily temperature and precipitation forecasts at all stations individually. Two bias correction techniques are compared, a mean debiasing method and a quantile mapping approach. Commonly used skill scores characterizing different aspects of forecast quality are computed for weekly aggregated forecasts with lead times of 5–32 days. Overall, promising skill is found for temperature in all seasons except spring. Temperature forecasts tend to show higher skill in Northern Europe and in particular around the Baltic Sea, and in winter. Bias correction is shown to be essential in enhancing the forecast skill in all four weeks for most of the stations and for both variables with QM generally performing better.

1. Introduction

In recent years large progress has been made in numerical weather prediction (Bauer et al., 2015). Major advances have been seen in data assimilation methods and model developments, resulting in the possibility to provide skillful predictions at lead times beyond a week toward subseasonal. In particular, the advent of ensemble predictions made long-range predictions feasible as they can at least partly capture the chaotic nature of the atmospheric system. Currently, several meteorological services around the globe run operational meteorological prediction systems for subseasonal to seasonal forecast horizons (e.g., Buizza et al., 2005; Li & Robertson, 2015). Seasonal predictions are mainly driven by atmosphere-ocean coupling processes and boundary conditions, whereas the quality of the forecasts at shorter lead times is mainly driven by the initial conditions. Thus, subseasonal forecasts covering lead times up to several weeks are located at the transition from weather forecasting to climate predictions. The recently launched World Weather Research Programme/World Climate Research Programme initiative on subseasonal to seasonal prediction aims to further improve these predictions by coordinating the efforts on a global scale (Robertson et al., 2015).

For the subseasonal forecast horizon, different expressions are used in the literature. In the early stage of subseasonal forecasting the European Centre for Medium-Range Weather Forecasts (ECMWF) referred to monthly forecasts for the subseasonal time scales up to 32 days. Nowadays, these forecasts are referred to as extended range to better distinguish it from medium-range forecasts. In recent publications and research projects the term subseasonal forecasts has been established for these forecast horizons, also in line with a further extension of the lead times. For example, the latest model cycle of the ECMWF runs up to 46 days lead time. Throughout this paper, the expression subseasonal forecasts is used for forecasts covering lead times up to 6 weeks.

Several verification studies have analyzed the performance of the existing subseasonal prediction systems operated by the different weather services. Buizza and Leutbecher (2015) showed for the ECMWF subseasonal prediction system that skill horizons of up to 2 to 3 weeks lead time can be achieved for the upper air variables geopotential height, temperature, and wind components. They define the skill horizon as the lead time when the ensemble ceases to be more skillful than a climatological distribution, retrieved from the ERA-Interim reanalysis, using the continuous ranked probability score (CRPS) as a metric. Another study by Saha et al. (2014) verified the subseasonal prediction of the National Centers for Environmental Prediction Climate Forecast System Version 2 in terms of predicting a Madden-Julian Oscillation (MJO) index. Prediction skill for the MJO in terms of anomaly correlation stays above 0.5 for 2 to 3 weeks for the new Climate Forecast System Version 2 system, whereas it was only 1 week in the old system. In a comparison of three different systems, Li and Robertson (2015) found skill in precipitation forecasts in terms of anomaly correlation over the Maritime Continent and the equatorial Pacific and Atlantic Oceans. In their comparison the ECMWF subseasonal forecast system performed best, the system that is being analyzed in the current manuscript.

An early verification study for the subseasonal ECMWF predictions by Weigel et al. (2008) concluded that the forecasts are generally not worse than climatology and do outperform persistence. The verification is performed against reanalysis data. In addition, the authors reported skill beyond a lead time of 18 days for some ocean regions and tropical South America and Africa. Since then, the performance of the ECMWF system has steadily increased (Vitart, 2014; Vitart et al., 2014). This improvement in the ECMWF system is most likely attributable to numerous changes in model physics since 2002. In addition, enhancements in model resolution need to be taken into account to explain the improvements of the forecasts. The variable resolution approach in the subseasonal prediction system described in (Vitart et al., 2008) was introduced to capture small-scale (30 km), severe weather events in the early forecast range up to day 10 and to provide accurate large-scale (60 km) forecast guidance for lead times up to 32 days. This new model system increased the skill for 2-m temperature forecasts at all time ranges over the extratropics compared to the previous model version. The same conclusions are valid for other variables such as precipitation anomalies and mean sea level pressure. More recently, Vitart (2014) showed a steady increase of forecast skill during the last 10 years of the North Atlantic Oscillation, the MJO, and 2-m temperature and precipitation over the northern extratropics, which is associated with improvements in the model physics of the subseasonal prediction system. Again, the forecasts were verified against reanalysis data sets. Although the changes in model physics parametrizations were designed to reduce systematic errors and improve skill for the medium-range time scale, it seems that these changes have also led to better subseasonal forecasts.

Many different applications are dependent on meteorological predictions for lead times beyond 10 days. Nowadays, subseasonal predictions are being used to run hydrological models for flood and drought management (Addor et al., 2011; Crochemore et al., 2017; Fundel et al., 2013), to provide a planning baseline for agricultural decisions (Calanca et al., 2011; Hirschi et al., 2012), or increasingly for planning and optimization purposes for renewable energy resources such as hydropower and wind (Alemu et al., 2011; Anghileri et al., 2013; Barros et al., 2003; Reyers et al., 2014). In addition, several hydrological analyses have been performed to assess the benefit of using such meteorological predictions. For example, Orth and Seneviratne (2013) and Bogner et al. (2018) showed the potential benefit of using the direct model output from the subseasonal meteorological predictions to drive a hydrological model in the Alpine area. Common to all these applications is the requirement of best possible forecasts of meteorological surface variables in an absolute sense. Despite the recent improvements of subseasonal forecasts mentioned above, the forecast models still exhibit systematic biases and are not capable to capture all local characteristics, in particular the presence of complex topography or in the vicinity of large water bodies. To correct such systematic biases, many different statistical techniques have been developed (Fowler et al., 2007; Themeßl et al., 2011). Such statistical techniques make use of the historical forecasts (so-called hindcasts or reforecasts) provided by the forecast model. This means that in addition to the actual forecasts, the model is run for the same date in the past 20 to 30 years with initial conditions from reanalysis data sets. The correction can then be derived by comparing these reforecasts of the system with the corresponding past observations. If the corresponding observations are at higher spatial resolution or even at the point scale, the bias correction implies a downscaling step. To determine the parameters of a bias correction or downscaling approach, the sample of reforecasts and corresponding observations should be large enough (Shi et al., 2015). At the subseasonal time scale the bias

correction of operational weather forecasts has two major constraints. The model drift causes the bias to be lead time dependent. Therefore, a lead time-dependent daily correction is favored. In addition, the biases depend on the season. In an operational context, the reforecasts are produced along the forecasts. The reforecasts of a particular forecast date are available just a few days before the respective forecasts. Hence, the correction cannot be estimated from a large symmetric window because the reforecasts of a date in the future (e.g., the reforecast of the forecast that will be calculated in 3 weeks from now) are not provided in advance. These constraints have to be balanced with the requirements of having a large data set for the correction estimation.

Although applications often require bias corrected and downscaled information, only few studies exist that verified ensemble forecasts for surface parameters at point locations for medium-range forecasts (e.g., Hagedorn et al., 2007; Hamill et al., 2008). To our knowledge, such a verification of subseasonal forecasts against ground observations does not yet exist, only verifications against reanalyses or gridded data sets have been published. But often, the reanalyses inherit biases from their driving model that leads to an optimistic verification if forecasts are verified against their own analysis (e.g., Park et al., 2008). Furthermore, the quality and representativeness of gridded data sets strongly depend on the methodology and the density of the network. And in contrast to ground observations at single sites each grid point in a gridded data set describes an areal average. In this study we aim at analyzing the performance of ECMWF subseasonal forecasts at the local scale and examine different postprocessing variants to improve local scale predictions. We verify subseasonal forecasts against an extensive ground observation data set for Europe. To evaluate ensemble forecast systems and the effectiveness of postprocessing techniques, different probabilistic verification metrics are applied.

This manuscript is structured as follows: We first introduce the characteristics of the forecasting system, the observational data sets (both in section 2) and the bias correction and verification methods used (section 3). The results section (section 4) starts with a focus on the performance of the raw reforecasts across Europe (section 4.1), followed by the effect of the bias correction techniques (section 4.2), the spatial characteristics of the reforecast performance (section 4.3), and a comparison with the skill of the operational forecasts within the analysis period. A closer insight in the reforecast performance for stations located around the Alps is given in the following subsections (sections 4.4–4.6) before discussing the results in relation to existing literature (section 5).

2. Data

2.1. Forecast and Reforecast Data

We analyze the subseasonal reforecasts and operational forecasts from the ECMWF IFS version CY40r1. This version was operational from 19 November 2013 to 12 May 2015. This is a unique data set because no system change took place for nearly one and a half year (http://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model/cycle-40r1/cycle-40r1 for the documentation of IFS CY40r1). Routinely, the model is updated more frequently within 1 year, and therefore, changes in the system, for example, changes to the horizontal and vertical resolution, or changes in the parametrization of physical processes affect the homogeneity of systematic errors and the skill of the forecast.

Subseasonal forecasts with IFS Cy40r1 were computed twice a week (Mondays and Thursdays) with an ensemble of 51 members (1 control run and 50 perturbed runs) for the next 32 days. The model was run at two different resolutions to reduce computational load. The first 10 days were calculated at a resolution of T639 (~30 km), whereas day 11 to 32 were computed at a resolution of T319 (~60 km). This change in resolution needs to be considered in the station based bias correction scheme described below.

For forecasts issued on Thursdays the historical reforecasts (i.e., hindcasts) were run for the same date in the last 20 years using ERA-Interim analysis for the initialization (European Centre for Medium-Range Weather Forecasts, 2014). These reforecasts are essential for the postprocessing: the bias of the forecasting system can be estimated using the reforecasts; future forecasts (or as in this analysis the reforecasts itself) can then be corrected. The reforecasts are produced with the identical model version, but with five members only. General descriptions of the model system can be found in Vitart (2004) and Vitart (2014). The daily mean temperature and precipitation data have been retrieved and bilinearly interpolated on a regular grid with 50-km resolution and covering the area −40.5° to 75.5° east and 25° to 75° north; hence, the entire European land

surface is covered in the analysis. The forecasts analyzed in this study were initialized between Thursday, 3 April 2014 and Monday 30 March 2015; the corresponding set of reforecast thus covers 3 April 1993 to 27 March 2014.

2.2. Observational Data

We combined data from three observational data sets to achieve an optimum coverage over Europe with ground based observations. As the primary data set, we used the data provided by the European Climate Assessment and Dataset project (ECA&D). This data set consists of long time series of quality controlled observational data for climate assessment studies (Klein Tank et al., 2002, http://www.ecad.eu/). We used all the time series providing daily temperature and precipitation records for the entire reforecast period with no more than 5% missing values. These criteria resulted in a reduction of stations from 6,936 to 1,023. For several countries, no observations remained in the data set after this filtering. As a supplement, we therefore added the Global Surface Summary of the Day (GSOD) data set (https://data.noaa.gov/dataset/dataset/global-surface-summary-of-the-day-gsod), provided by the National Centers for Environment Information. The GSOD is the largest data set providing observational data worldwide (Kilibarda et al., 2015) and is used for example to assess heat waves (e.g., Ceccherini et al., 2017). As the GSOD quality control is less strict than that of ECA&D, only observation sites not already included in the ECA&D data set were appended, and the same missing value threshold of 5% was applied. Finally, we added 75 quality controlled stations over Switzerland from the Swiss national observing system SwissMetNet operated by the Federal Office of Meteorology and Climatology MeteoSwiss (more information can be found at http://www.meteoschweiz.admin.ch/home/mess-und-prognosesysteme/bodenstationen/automatisches-messnetz.html). The resulting combined data set includes 1,637 stations covering all of Europe.

3. Methods

3.1. Postprocessing

Two different postprocessing methods were applied: a mean debiasing or scaling (MD) approach and a quantile mapping (QM) approach. The mean debiasing method is a correction technique in which the lead time-dependent ensemble mean bias between the forecasts and the observations is estimated using a local polynomial regression (LOESS) (Cleveland et al., 1992; Cleveland & Devlin, 1988). Mahlstein et al. (2015) have shown that daily fluctuations in the climatology can be smoothed by applying such a LOESS fit and correction factors can be estimated more robustly than with alternative methods. For temperature we applied an additive correction and for precipitation a multiplicative correction to avoid artificial generation of negative precipitation values. Many previous studies follow the same approach when applying bias correction techniques (e.g., Addor et al., 2014; Teutschbein & Seibert, 2012; Werner & Cannon, 2015).

QM, also referred as distribution mapping or quantile-quantile transformation, is a statistical bias correction method that showed good performance for postprocessing of forecast and climate model data (e.g., Crochemore et al., 2016; Rajczak et al., 2016; Verkade et al., 2013). QM does not only correct the mean but the full marginal distribution and thereby the frequency and intensity of the target variable (Themeßl et al., 2011). The basic principle behind QM is to correct the forecasts such that the statistical distribution of the full forecast data set matches the one of the observations. The correction can be estimated based on the empirical distribution function (Bennett et al., 2014; Wood et al., 2002, 2004) or based on parametric distributions fitted to the data (Gudmundsson et al., 2012; Piani et al., 2010). In this study we use the empirical QM. The correction factors are estimated for each percentile separately from the 1st to the 99th percentile. Forecasts outside that interval are corrected using the correction factors for the 1st and 99th percentile, respectively. In analogy to the MD method, an additive correction is implemented for temperature and a multiplicative correction for precipitation forecasts. Although the term quantile mapping implies a direct mapping of the quantiles of the forecast to the one of the observations, in most technical implementations correction factors are calculated based on the quantiles of the forecast and observation pairs and applied to the forecast to be corrected.

We performed the daily bias correction using the methods described above in a leave-one-year-out fashion (e.g., Themeßl et al., 2012). The correction of a reforecast for a given year only considers the remaining 19 years for the estimation of the correction function. This procedure is adopted to mimic the correction

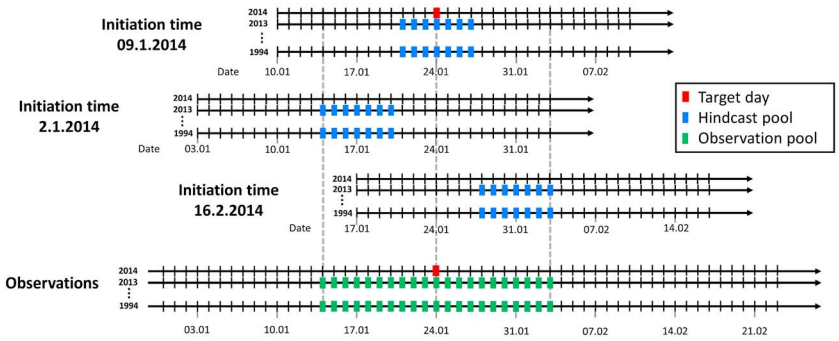

Figure 1. Illustration of the data used for the calibration of daily forecast values. To correct the forecast for a given target date (red mark), all reforecast dates valid within a 1-week window around the target date are used to form the model climate distribution (blue marks in the top row) and contrasted with the observations of the corresponding date range. In addition, the reforecasts with the same lead time initiated in the previous week (blue marks in the second row) and in the subsequent week (blue marks in the third row) are considered in order to enhance the sample size. The corresponding observations (green marks) cover the entire 21-day window around the target day. This results in a sample size of 399 ($3 \times 7 \times 19$) daily values for the observations and 1,995 ($3 \times 7 \times 19 \times 5$) daily values for the reforecasts to estimate the correction factors for the quantile mapping (QM) technique.

of an actual, out-of-sample forecast and to prevent artificial skill from being generated by introducing information about the year that is reforecast.

The correction for a given lead time was estimated based on a sample of reforecasts and corresponding observations within a 7-day window centered on the lead time of interest. Hereby, the corresponding values of the preceding and the subsequent reforecast dates were also used to enlarge the sample size (see Figure 1). The resulting sample size for estimating the daily corrections thus consists of 399 daily values *(19 years, 7-day window, 3 initialization dates)* for observations and 1995 values *(19 years, 7-day window, 3 initialization dates, 5 ensemble members)* for the reforecasts, respectively. In principle, the bias correction of reforecasts could have been performed based on an even larger sample of reforecast dates, but limiting the sample to three reforecast dates ensures that the identical procedure can be applied for the correction of operational forecasts, as reforecasts are only available 1 week ahead of the corresponding forecast. In addition, given that the bias depends on the season a larger window (e.g., covering more than 1 month) can be problematic especially in the transition seasons spring and autumn.

We also accounted for the resolution change between day 10 and 11 in the forecasts in the bias correction by estimating the daily correction factors separately for the two legs with different resolution. The separate treatment of the two legs proved to be necessary to satisfactorily correct the abrupt jumps in the biases that were found at some stations between days 10 and 11 as a consequence of the resolution change (see section 4.6).

3.2. Verification

A variety of different verification metrics exists to characterize the quality of a forecast. As single scores only provide insights in specific aspects of forecast quality, different scores should be considered to comprehensively assess the forecast performance. In this section we provide an overview of the metrics used for the verification of categorical and continuous forecasts in this study. Detailed descriptions of the mathematical formulations can be found for example in Wilks (2011) and Jolliffe and Stephenson (2012).

The bias of a probabilistic forecast is defined as the difference between the ensemble mean of the forecast and the corresponding observation. In case of precipitation the bias is scaled with the climatology for the given site, corresponding to a relative bias. Otherwise, dry locations tend to have small biases, whereas wet stations generally have larger biases.

Moreover, we use the spread to error relationship—an indicator for the dispersion characteristics of an ensemble forecast. A good spread to error relationship is a mirror of reliability. Reliability is the attribute related to the conditional bias of an ensemble forecast, which compares the forecast probability to the observation frequency (for all observations recorded in correspondence to the given forecast probability). As an

example, if a probability for rain of 30% is forecasted 180 times, rain should be observed for 60 of these forecasts. For reliable ensembles the spread to error relationship equals 1. Larger ratios indicate overdispersive ensembles and smaller ratios indicate overconfident ensembles. A more detailed, visual evaluation of the ensemble dispersion characteristics is provided by the rank histogram (Anderson, 1996; Hamill & Colucci, 1997; Talagrand et al., 1997). In a rank histogram the ensembles of a given forecast are ranked to determine the bins and the observations are assigned to the bin they fall into. U-shaped rank histograms indicate underdispersion (overconfidence), whereas a concave shape indicates overdispersion (underconfidence) and a uniform shape of the histogram is expected for reliable ensemble forecasts. Additional measures such as probabilistic integral transform diagrams and reliability diagrams can inform about the reliability of ensemble forecasts. However, as the reforecasts analyzed in this study have a small ensemble size of five members, we focus on the spread to error ratio and the rank histograms as these are more robust descriptors in case of small ensemble sizes.

The relative or receiver operation characteristic diagram (ROC diagram) originates from signal detection theory in electrical engineering and was first introduced by in the meteorological literature by (Mason, 1982). The ROC diagram relates the number of correct forecasts of events (hit rates) to the number of false alarms. Thus, this measure is applied to categorical forecasts where the categories are usually defined as the terciles. The area under the resulting curve (ROC area A_z) indicates how well a forecast can discriminate between different outcomes. This ROC area corresponds to the ROC Score with values of 0.5 indicating no skill and values of 1 indicating perfect forecast. The ROC skill score (ROCSS) is then defined as ROCSS = $2 \times A_z - 1$ to meet the same range as the other skill scores. Hence, the ROCSS measures the discriminative power of a forecast, is insensitive to systematic biases of the forecasting system, and can therefore be interpreted as a measure of the potential predictability of a system, that is, the predictability (or skill) irrespective of its systematic biases.

To assess the accuracy of the forecast, we use a scoring rule for ensemble forecasts—the CRPS. The CRPS corresponds to the integral of the squared difference between the cumulative distribution function of the probabilistic forecast for a given outcome (F (y)) and the cumulative distribution function of the observed outcome (F_O (y)) (formula adapted from Wilks, 2011).

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} [F(y) - F_o(y)]^2 \mathrm{d}y$$

where

$$F_o(y) = \begin{cases} 0, & y < \text{observed value} \\ 1, & y > \text{observed value} \end{cases}$$

A low CRPS characterizing an accurate forecast is achieved, if the forecasted distribution is narrow and centered around the observed value and hence has the correct statistical distribution (Hersbach, 2000). The CRPS is therefore sensitive to both, to the bias in the ensemble mean and to the overdispersion or underdispersion, that is, both reliability and resolution. Deviation in either resolution or reliability will lead to higher CRPS values (worse score). Note that the CRPS is negatively biased for small ensemble sizes of the forecasts, which can be corrected with an analytical solution. Hence, to convert the findings based on the reforecast with small ensemble size to the forecasts, we apply the correction proposed by Ferro et al. (2008) to account for the finite ensemble size.

A score measures an attribute of a specific forecast itself. To evaluate the performance of a forecast with respect to a reference forecasts the generic description of a skill score can be used.

$$\mathrm{SKILL\ SCORE} = \frac{\mathrm{SCORE_{forecast}} - \mathrm{SCORE_{ref}}}{\mathrm{SCORE_{perfect}} - \mathrm{SCORE_{ref}}}$$

where $\mathrm{SCORE_{forecast}}$ is the score of the forecast to evaluate, $\mathrm{SCORE_{ref}}$ the score of the reference, and $\mathrm{SCORE_{perfect}}$ the score of a perfect forecast. For negatively oriented scores, the score of a perfect forecast is 0. Hence, a skill score spans over the range between 1 and $-\infty$. Positive skill scores indicate better performance of the forecast compared to the reference, negative values the opposite. In this study we use the

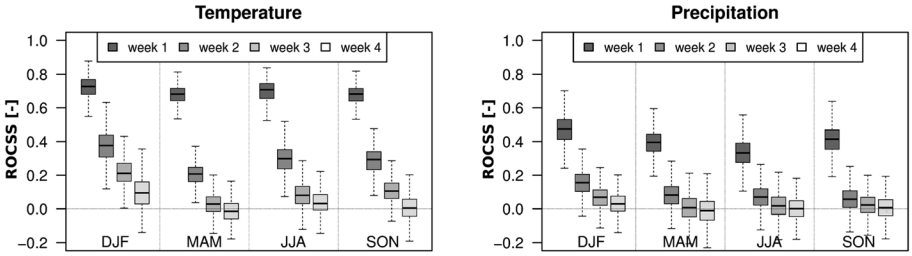

Figure 2. Upper tercile ROC skill score for 1994–2015 reforecasts of weekly mean temperature (left) and precipitation sums (right) grouped by season (DJF, MAM, JJA, and SON). The shading of the boxes denotes lead time, whereas week 1 corresponds to days 5–11, week 2 to days 12–18, and so on. An individual box shows the distribution of the ROC skill scores for all observation sites averaged over all 13 reforecast initialization dates within one season. ROC = receiver operation characteristic; DJF = December–February; MAM = March–May; JJA = June–August; SON = September–November.

climatological forecasts as the benchmark, that is, reference forecast. The reference forecasts are constructed in analogy to the calibration illustrated in Figure 1. We use the observations of each year in the corresponding reforecast period including the previous and subsequent weeks in a leave-one-year-out manner. The resulting reference forecast contains 57 members (3 initiation dates times 19 years) and takes into account $3 \times 19 \times 7$ (= 399) daily observations per week that are verified.

In the following section the skill scores will be presented for each season separately. The scores are calculated for each reforecast initiation date, lead time week, and station separately. These scores are then averaged over the 13 forecast dates per season for both reforecasts and reference. The skill score is than derived from these average scores. The computation is performed with the open source R-packages SpecsVerification (Siegert, 2015) and easyVerification (MeteoSwiss, 2016), which provide routines for ensemble forecast verification for large data sets.

4. Results

4.1. Forecast Skill of Direct Model Output

We start with a skill analysis of tercile probability reforecasts for weekly means of temperature and weekly precipitation sums using the ROCSS. The ROCSS for the upper tercile of weekly mean temperatures and weekly precipitation sums as computed for the 1993–2014 reforecasts is shown in Figure 2.

The reforecasts for both variables show a large jump in skill between 5- to 11-day lead time and 12- to 18-day lead time with substantial variations across the season at longer lead times. Best skill is found for weekly mean temperatures in the winter season (December–February, DJF), where positive skill remains as far as 26- to 32-day lead time for most of the observation sites. Skill is worst in spring (March–May, MAM) at all lead times, and only about half of the observation sites remain at positive skill levels at 19- to 25-day lead time. Summer (June–August, JJA) and autumn (September–November, SON) show similar skill, with positive values at 26- to 32-day lead time for more than half of the observation sites.

Precipitation skill is generally worse. As in case of temperature, skill in DJF is better than in other seasons, with positive values for more than half of the observation sites as far as 26- to 32-day lead time. Skill in spring, summer, and autumn is similar, with scores in week 3 (19–25 days) dropping below those of DJF week 4.

The ROCSS is insensitive to biases; therefore, the biases are presented separately in the following Figure. Seasonal differences are also found in the ensemble mean bias of the reforecasts. The map in Figure 3 shows the biases for 12- to 18-day lead time in all four seasons and for both weekly mean temperatures and weekly sums of precipitation.

Overall, temperature reforecasts tend to exhibit negative biases. Largest biases are found over the Alpine area, the Iberian Peninsula, and in Norway. This general spatial pattern is found in all four seasons, whereas biases are larger in the winter season. Biases are smallest in summer with a tendency of positive biases in Eastern Europe. The increase of the cold bias with altitude (e.g., in the Alps) is consistent with climate model biases (Kotlarski et al., 2012, 2014). It is worth mentioning that individual stations with very warm biases stand out in these maps, which is most probably due to large differences between model topography and station heights (e.g., observation sites on mountain tops).

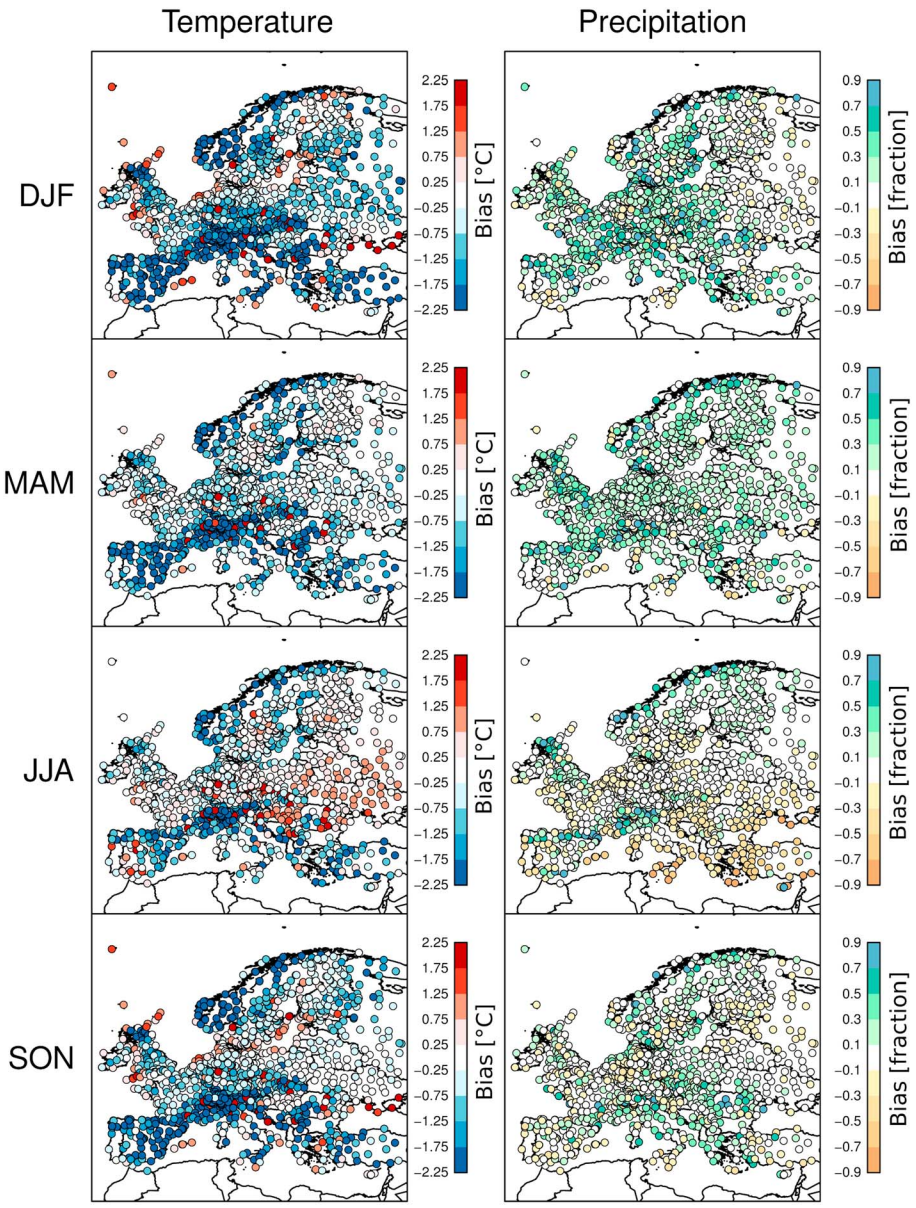

Figure 3. Weekly mean biases (ensemble mean errors) of forecast days 12–18 for temperature (left) and relative bias for precipitation (right) for the different seasons (winter on top, autumn on the bottom). DJF = December–February; MAM = March–May; JJA = June–August; SON = September–November.

In case of precipitation, the dominating sign of the bias is positive, indicating a general wet bias of the reforecasts (especially in DJF and MAM). On the other hand in JJA and SON the bias is wet for half of the stations and dry for the other half, meaning that precipitation is underpredicted in these cases. The precipitation bias did not show a systematic spatial pattern. To further evaluate the performance of the forecasting system, the effect of the bias correction techniques will be presented in the following section.

4.2. Effect of Bias Correction

The overview plot in Figure 4 shows the biases before and after the bias correction. In analogy to the maps presented above, raw temperature reforecasts tended to exhibit negative biases in all seasons and for all lead times with smallest biases in JJA. Precipitation reforecasts generally overpredict the precipitation but in JJA and SON the reforecasts show negative biases for approximately half of the stations depending on the lead time. After the bias correction, the bias of both temperature and precipitation reforecasts is substantially

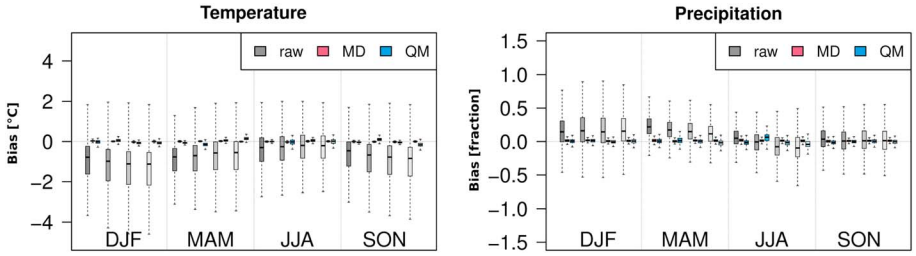

Figure 4. Overview of weekly mean temperature biases (left) and weekly precipitation sum relative biases (right) in 1994–2015 reforecasts. The box shading corresponds to different lead times as in Figure 2, gray shading for raw reforecasts, red for bias-corrected reforecasts using mean debiasing (MD), and blue for quantile mapping (QM) correction. DJF = December–February; MAM = March–May; JJA = June–August; SON = September–November.

reduced. In case of precipitation the bias is scaled with the climatological precipitation for the given season and lead time. Hence, a ratio of 1 indicates a bias as large as the climatologically expected precipitation amount.

To provide a deeper insight in the effect of the bias correction techniques, the continuous ranked probability skill score (CRPSS) is presented in Figure 5 for raw and corrected temperature and precipitation reforecasts through all seasons and lead times. The CRPSS measures both reliability and resolution of a forecast and is sensitive to the ensemble mean bias and the spread of the forecasts. Direct model output temperature reforecasts do not show any skill beyond 12- to 18-day lead time for the majority of observation sites, and differences among observation sites are large as indicated by the large spread. Both correction techniques increase skill overall with marginally better skill in case of QM. The skill spread between sites is largely reduced. Seasonal differences of bias corrected temperature reforecasts essentially confirmed the picture found in the potential skill analysis (Figure 2): Best skill in DJF, followed by SON, JJA, and MAM.

The difference between the potential (ROCSS, Figure 2) and actual (CRPSS, Figure 5) performance is more pronounced for uncorrected reforecasts. The spread among observation sites is generally much larger in case of temperature with respect to precipitation for all lead times and seasons. Uncorrected precipitation reforecasts only partly exhibit this large spread mainly at early lead times in DJF and MAM. Both potential and actual performance show a large drop in skill between lead times 5–11 and 12–18. In general, the actual skill in terms of the CRPSS of both temperature and precipitation reforecasts comes close to what can potentially be expected from the forecast system based on the ROCSS. In terms of the CRPSS, precipitation reforecasts postprocessed with QM generally exhibit better performance compared to the MD corrected reforecasts.

These differences between the two postprocessing methods are more pronounced in the ratio between the spread and the error of the reforecasts (Figure 6). Direct model output temperature reforecasts differ from the ideal spread to error ratio indicating overconfident ensembles over the whole forecast range, with lowest values at early lead times. Both bias correction techniques improve the spread to error ratio for temperature, and the QM technique shows a better performance compared to the MD technique, what is expected because QM corrects all moments of the distribution, including the spread. For precipitation, only the bias correction by QM is able to improve the spread to error ratio. For some seasons, the MD correction reduces the spread to error ratio even further, indicating that the MD method deteriorates the precipitation reforecasts.

This inability of the MD method to provide reliable forecast is confirmed by the rank histograms shown in Figure 7. The direct model output exhibits a negative bias for temperature and a positive bias for precipitation, and the U shape of the rank histogram for both parameters indicates overconfidence and thus not fully reliable reforecasts. For temperature the bias can be eliminated and the overconfidence especially at longer lead times is reduced with both techniques. In contrast, postprocessed precipitation reforecasts using the MD technique still exhibit a positive bias and a U shape. The QM processing performs better than MD as indicated by the more uniform shapes. The superiority of QM is found for all lead times and in all seasons, but QM is not able either to entirely remove the overconfidence at shorter lead times.

Given the clear superiority of the QM bias correction for both temperature and precipitation, the following spatial analyses will only present QM-corrected results.

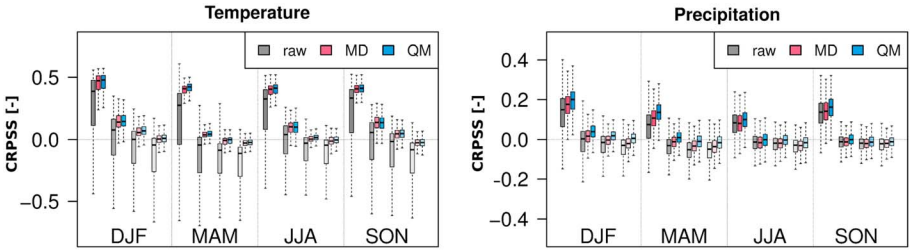

Figure 5. As in Figure 4 but CRPSS of weekly mean temperature (left) and weekly precipitation sums (right). CRPSS = continuous ranked probability skill score; MD = mean debiasing; QM = quantile mapping; DJF = December–February; MAM = March–May; JJA = June–August; SON = September–November.

4.3. Spatial Characteristics of Forecast Skill

Forecast skill depends on both season and lead time and also exhibits regional differences. We present the spatial distribution of the CRPSS because in this skill score shows pronounced regional differences and because it is a good indicator of the overall skill (measuring both resolution and reliability). Figure 8 shows the CRPSS for all individual stations within this analysis at lead time days 12–18. Except for MAM, temperature reforecasts show quite pronounced spatial skill patterns. DJF temperature reforecasts show higher skill toward northern and eastern Europe. Best skill is found in DJF with positive CRPSS in the whole area except for the Iberian Peninsula and the northernmost part of Scandinavia, and particularly high values were found for locations situated around and east of the Baltic Sea. A similar pattern albeit a generally lower skill can be identified in JJA and SON, with a pronounced area of higher skill over Central Europe.

For precipitation, skill beyond lead times of 11 days as measured by the CRPSS is limited to DJF and areas on the Iberian Peninsula and western Norway. Interestingly, the positive skill in DJF precipitation reforecasts on the Iberian Peninsula contrasts with temperature skill. The other seasons (MAM, JJA, and SON) only show marginal skill without any evident spatial pattern.

However, on a European scale there is promising skill (for temperature reforecasts) in all seasons except MAM. The relationship of these results to other studies is discussed in section 5.

4.4. Forecast and Reforecast Skill

The analysis of the skill of the reforecasts indicates promising results on how much can be improved when the reforecasts are postprocessed using simple debiasing techniques. However, these results might not directly be transferred to the performance of the actual forecasts itself. In contrast to the reforecasts that are initialized from the ERA-Interim reanalysis, the forecasts are initialized from an operational analysis. Hence, the postprocessing might be affected by this difference in the statistical characteristics of both analyses, because it relies on the reforecasts to correct the forecasts. Figure 9 shows the CRPSS of the raw and postprocessed forecasts within the period April 2014 to March 2015 using the operational forecast ensemble with 51 members. The result agrees well with the results from the reforecast data set. Overall, both postprocessing techniques are able to enhance the overall performance in terms of the CRPSS for both variables. For temperature, the operational forecast exhibit even higher CRPSS values on average compared to the reforecasts for the lead times 5–11 and 12–18 days. Especially in the autumn season (SON) most of the stations still

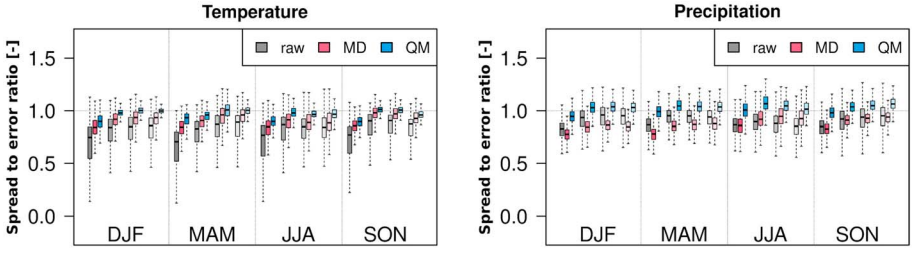

Figure 6. Same presentation as in Figures 4 and 5 for the spread to error ratio. MD = mean debiasing; QM = quantile mapping; DJF = December–February; MAM = March–May; JJA = June–August; SON = September–November.

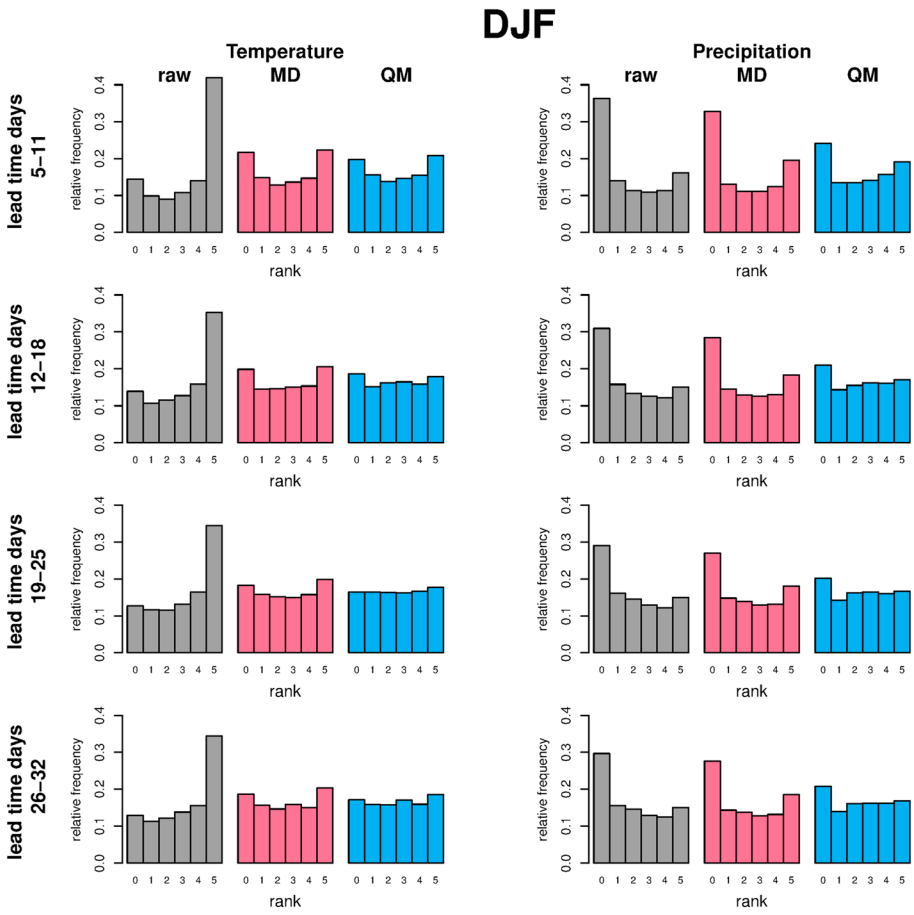

Figure 7. Rank histograms of reforecasts initialized in winter for lead times of 1 to 4 weeks (from top to bottom), raw reforecasts in gray, MD-corrected reforecasts in red, and QM-corrected reforecasts in blue. MD = mean debiasing; QM = quantile mapping; DJF = December–February.

show a positive CRPSS for all lead times. On the other hand, the skill spread over all 1637 stations is larger for the operational forecasts.

4.5. Forecasts Skill in the Alpine Region

Bias correction techniques account for systematic biases in the forecasts. Such biases can depend to a large degree on the topographic characteristics. We present a closer insight into the forecast performance for observation sites situated in the Alpine region. Figure 10 shows the CRPSS of all available sites within the Alpine region. Direct model output temperature reforecasts have large negative CRPSS values, and skill is clearly lower than in other regions. After applying QM for bias correction, skill can be increased to almost the same level as in other regions and shows the same seasonal variation: best skill in DJF and worst results in MAM. Overall, the relative effect of the correction is more pronounced for the site in the Alpine area (compare Figure 5 and Figure 10). Precipitation reforecast skill shows a similar behavior, that is, lower skill of direct model output compared to other regions and bias correction by QM yielding similar skill as in other regions. We also note that the advantage of QM over MD bias correction is more pronounced for both temperature and precipitation in this subset of data.

To illustrate the spatial skill characteristics, Figure 11 shows the direct model output and bias corrected DJF reforecasts for each station at lead time 12–18 days. Most of the observation sites exhibit pronounced negative skill for both raw temperature and raw precipitation reforecasts with relatively strong differences between the sites. After applying the QM correction, the skill becomes more uniform among sites with positive skill for temperature reforecasts at all sites and slight positive skill for precipitation reforecasts.

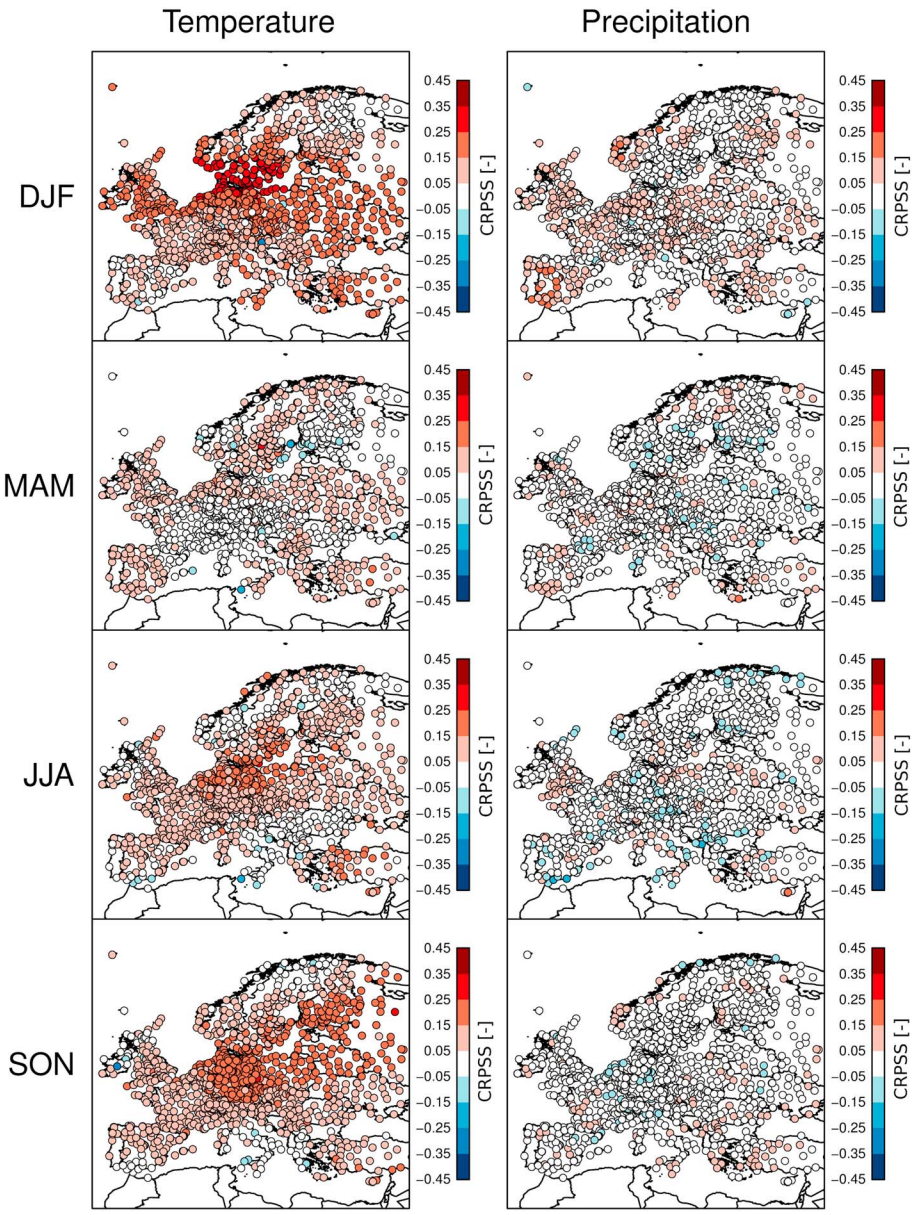

Figure 8. CRPSS of weekly mean temperature (left) and precipitation sum (right) reforecasts after bias correction with quantile mapping. The maps show the results for 12–18 days lead time for the four different seasons (from top to bottom). CRPSS = continuous ranked probability skill score; DJF = December–February; MAM = March–May; JJA = June–August; SON = September–November.

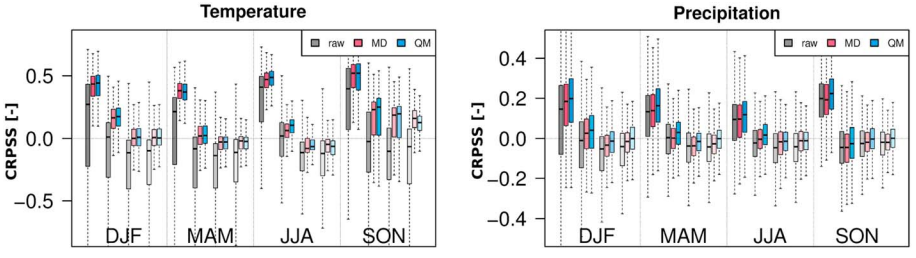

Figure 9. CRPSS as presented in Figure 5 but for the actual forecasts with 51 members. CRPSS = continuous ranked probability skill score; DJF = December–February; MAM = March–May; JJA = June–August; SON = September–November.

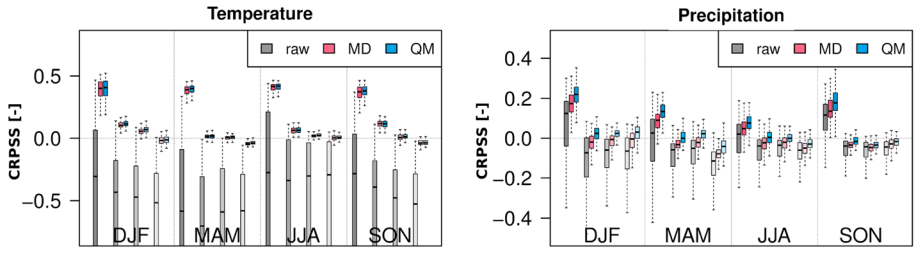

Figure 10. CRPSS as presented in Figure 5, but only for a subsample of stations (n = 174) located in the Alpine region. CRPSS = continuous ranked probability skill score; DJF = December–February; MAM = March–May; JJA = June–August; SON = September–November.

4.6. Example of Selected Station

Several site-specific features of the forecast influencing the bias correction cannot be observed in the aggregated skill analysis presented above. In Figure 12 two exemplary stations are selected to illustrate the lead time dependent bias and the influence of the resolution change at single locations. The observation site Sion is located in an Alpine valley, that is, an area characterized by highly complex topography. The temperature bias of the direct model output shows a pronounced decrease between lead days 10 and 11, in accordance with the horizontal resolution change of the subseasonal prediction model. Assuming the bias being dominated by altitude differences between model topography and observation sites, such a pronounced change of the bias with the resolution change is not surprising. After the QM bias correction, this jump and the overall bias for all lead times is clearly reduced. At the lowland observation site (Zurich Fluntern) the bias of the direct model output temperature reforecasts showed a rather steady dependence on lead time. For early lead times, the bias is smaller than for lead times at the end of the reforecasts. After the correction of the reforecasts this lead time dependence is eliminated.

The bias of precipitation reforecasts shows larger variations. To illustrate this, the daily bias of JJA precipitation reforecasts is shown. For Sion, early lead times show a large positive bias in the direct model output data, which is slightly decreased with lead time. The biases of the lowland observation site (Zurich Fluntern) do not show a strong dependence on lead time. Nevertheless, a minimal decrease of the bias with lead time can be identified. This is in accordance with the analysis of the whole data set.

5. Discussion and Conclusion

This study investigates the skill of subseasonal ECMWF forecasts over Europe for temperature and precipitation. The focus is on the prediction of station observations and to what extent bias correction and downscaling techniques can maintain or recover the predictability. The study makes extensive use of a forecast data

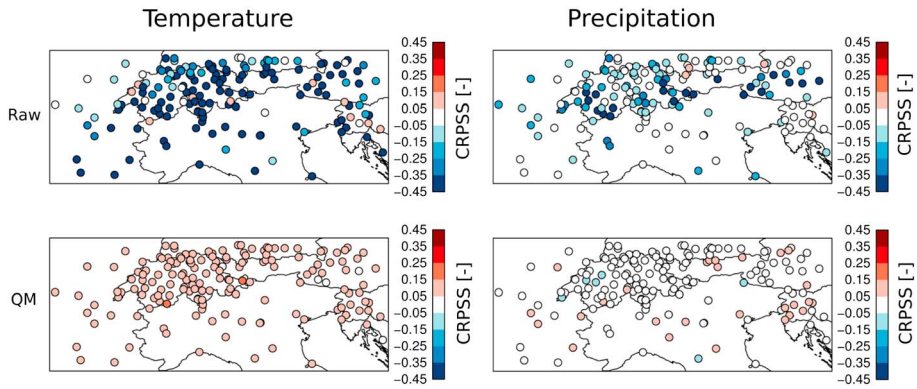

Figure 11. CRPSS for observation sites in the Alpine region for the winter season, weekly mean temperatures (left) and precipitation sums (right) for lead times 12–18 days. Raw reforecasts (top) and postprocessed reforecasts using quantile mapping (bottom). CRPSS = continuous ranked probability skill score; QM = quantile mapping.

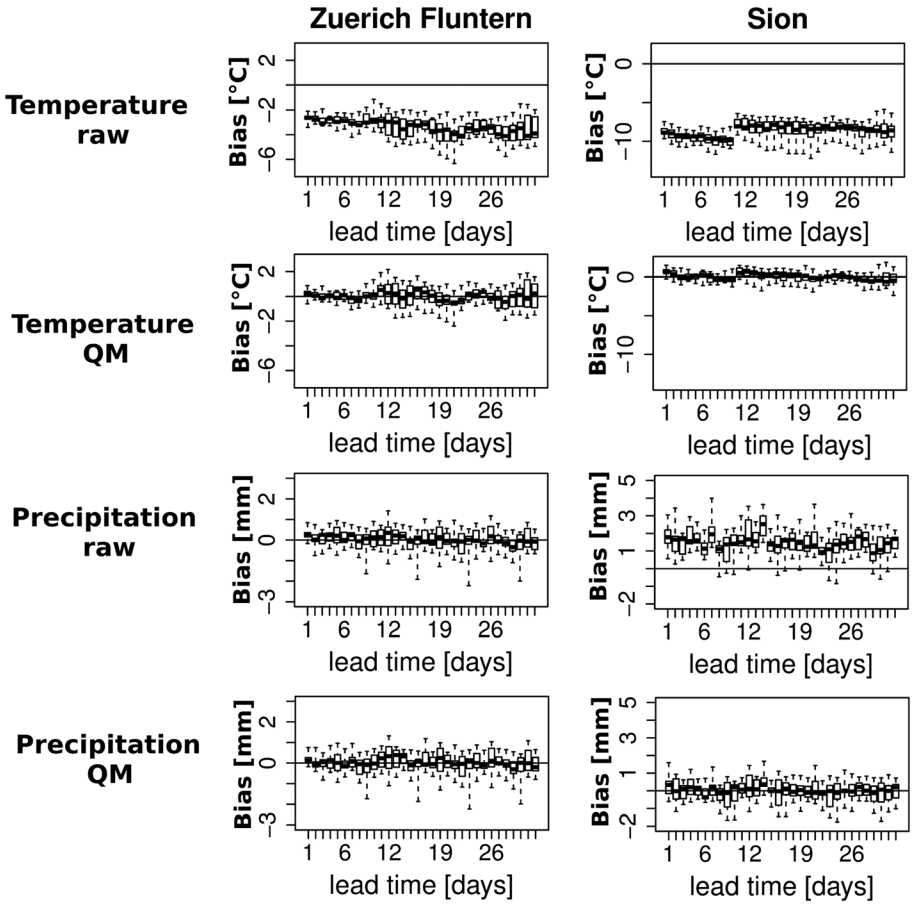

Figure 12. Lead time-dependent daily biases of temperature and precipitation for direct model output and corrected reforecasts at a site located in an Alpine valley (Sion) and at a location in the lowland (Zurich/Fluntern), respectively. QM = quantile mapping.

set based on the same model version for more than a year and the full associated reforecast data set of 20 years. The direct model output of the temperature reforecasts shows potential predictability using a skill score not being sensitive to biases for up to 3 weeks depending on the observation site and season. Less but still positive skill for the first week (lead days 5–11) is observed for the precipitation reforecasts. For both variables reforecasts initiated in DJF obtained highest scores. The direct model output exhibits large spread in skill among the different weather stations. Similarly, large spread in skill between different locations is reported in a skill analysis of the ECMWF seasonal prediction system 4 by Crochemore et al. (2016). The authors found an overall skill of up to 2–3 weeks ahead for raw precipitation forecasts. In contrast to our site-specific verification the authors verified the precipitation aggregated over entire catchments what can explain the difference in skill. Nevertheless, a similar behavior of the bias is found with a tendency to overpredict precipitation in winter and spring and underpredict precipitation in summer.

The correction methods applied here substantially reduce the bias and enhance the skill of both temperature and precipitation reforecasts with slight superior performance of the QM compared to MD corrected reforecasts. For most of the skill metrics considered, the MD and the QM method show similar performance with slightly superior performance of the QM technique. More pronounced differences could be found for the spread to error ratio and rank histograms. The mean debiasing technique results even in worse spread to error ratios compared to direct model output reforecasts (Figure 6). This indicates that the spread is not accurately corrected if only the mean correction is applied. In contrast, the spread to error ratios and rank histograms improved substantially after applying QM correction indicating an improvement in reliability. To some extent, this has to be expected because QM corrects all moments of the distribution, including the (model) climatology spread, not in a conditional manner though.

Numerous studies in the context of climate change adaption showed that QM can effectively be used to correct model output data (e.g., Rajczak et al., 2016; Themeßl et al., 2012) and in hydrological forecasting at long time scales (e.g., Crochemore et al., 2016; Jörg-Hess et al., 2015). Similarly, Crochemore et al. (2016) found that the QM clearly improves the reliability of the forecasts, whereas the sharpness and accuracy is better improved by the linear scaling method (corresponding to what we defined as mean debiasing method). Nevertheless, it must be stressed that QM does have limitations. Maraun (2013) highlighted limitations of QM when applied to bias correct and downscale precipitation from climate model simulations to the local scale. In particular he mentions the inability of QM to correctly represent the spatial and temporal structure of the corrected time series and problems with the area-mean extremes. In the context of subseasonal predictions these limitations cannot be neglected although our results tend to be very promising. In a recent study Zhao et al. (2017) emphasized the inability of QM to provide fully reliable and coherent forecasts in case of seasonal predictions. Our results, however, show a clear improvement in reliability at lead times beyond 2 weeks after QM is applied and precipitation forecasts are generally found to be coherent, that is, corrected forecasts are not worse than climatology. This discrepancy to the findings by Zhao et al. (2017) might be associated with the different forecast system and QM setup used. In contrast to the daily correction for individual stations performed in the present study, they corrected monthly aggregated data and used gridded observation at the same resolution as their predictions. Our analysis suggests that despite its known limitations, QM is a valuable technique to correct and downscale ECMWF extended-range predictions to the local scale at least for longer lead times. The limitation of not taking into account conditional biases is problematic at shorter lead times (where correspondence between forecast and observations is strongest) and manifested itself in QM's inability to entirely correct overconfidence. However, especially when applied in an operational setup the benefit from the computational efficiency of the QM technique may overweight its limitations. It is noted here that more advanced postprocessing techniques that do ensure the correspondence between forecasts and observations could be a valuable option for further improvements, especially for shorter lead times where the rank histograms still indicate overconfidence in the QM corrected reforecasts. Different methods could account for such relationships, for example, ensemble model output statistics (EMOS) combined with ensemble copula coupling (Schefzik, 2017) or an extension of the QM technique to a more advanced quantile regression neural network approach (Cannon, 2011).

Better skill for ECMWF subseasonal temperature forecasts in the winter season have been previously reported in the literature (Vitart, 2004; Weigel et al., 2008). Interestingly, temperature forecasts for different long-range prediction systems stress better performance of forecast for the winter season (Johansson et al., 1998; Scaife et al., 2014). In addition to seasonal differences in skill of the reforecasts we found geographical differences. The spatial pattern in forecast performance found in section 4.3 resembles the pattern described in earlier studies. Johansson et al. (1998) found such a similar pattern for a seasonal forecasting model based on an empirical methodology, using a canonical correlation analysis. As well Scaife et al. (2014) reported higher skill toward northern Europe in terms of correlation of winter mean temperature from seasonal forecasts of the Met Office Global Seasonal Forecast System 5 (GloSea5). Furthermore, similar patterns were found for correlations of surface air temperature from the ECMWF Seasonal prediction system 3 (Doblas-Reyes, 2010) which coincide with stronger warming trends over the past decades in Western Europe (van Oldenborgh et al., 2009). Especially in the region of the Baltic Sea these warming trends seem to be enhanced for the 30 year period of the late 20th century (Barkhordarian et al., 2016) which overlaps with the reforecast period used in the present study to estimate the correction functions. The influence of a trend on the skill of temperature forecasts have already been reported for seasonal and decadal predictions (Doblas-Reyes et al., 2006; Lienert & Doblas-Reyes, 2017; Liniger et al., 2007). At these forecast lead times a realistic greenhouse gas forcing leads to increased predictability of temperatures. The temperature trend could also be an explanation for the somewhat higher skill found in the operational forecasts compared to the reforecasts for temperature, but not for precipitation.

Another regional difference in forecast performance can be related to the complexity of the topography around the observation sites. Forecast performance clearly differs depending on the complexity of the terrain. The performance for sites located in the Alpine area with complex terrain is generally lower compared to the rest of Europe. But the bias correction, in particular the QM, largely compensates this effect and can recover the skill and can as expected eliminate outliers. This indicates that using local climatologies to correct the forecasts is crucial to improve forecast performance. Hence, a postprocessing setup that accounts for

local effects that are most probably on a subgrid scale clearly improves the forecasts. It is not surprising that this effect is more pronounced for temperature reforecasts, as the biases are more sensitive to the mismatch between the model topography and station height and due to local effects that cannot be resolved with a coarse model resolution. Also, the European scale patterns of skill and the lack of local noise in the skill maps of postprocessed forecasts are an indication that the predictability is rather driven by large-scale processes, such as synoptic dynamics, than local phenomena.

The verification of the operational forecasts in comparison to that of the reforecasts also provides some insight on postprocessing of subseasonal forecasts. Although the skill scores show larger variability for the operational forecasts, the general features of the verification are in good agreement with the analysis of the reforecasts. This indicates that the applied postprocessing techniques also provide more accurate and reliable forecasts in the real out-of-sample setup. In addition, we note that cross validation does consistently lead neither to an overestimation in the reforecast skill in our setup as it is described in Shabbar and Kharin (2007) nor to a degeneracy in skill as found in in other studies (Barnston and van den Dool, 1993; Gangsto et al., 2013). Furthermore, the agreement of the skill score between the operational forecasts with its 51-member ensemble and the reforecasts with only 5 members also illustrates the validity of the fair scores concept (Ferro et al., 2008).

Our analysis shows that subseasonal forecasts exhibit skill even at the local scale after bias correcting. Nevertheless, the ECMWF subseasonal forecasts provide only a limited number of 20 years of reforecasts; the sampling uncertainty in estimating the correction factors is still high (Shi et al., 2015; Weisheimer & Palmer, 2014). A second challenge is the choice of an optimal window size to estimate the correction factors. To make robust estimates, the data sample used should be large. On the other hand, the dependence of the bias on lead time and season at the subseasonal timescale, and the aim to apply a method that can be used in an operational setting, requires a short window. Hence, the calibration framework proposed in this study is a compromise to meet these demands. It is noteworthy that subseasonal forecasts act at the interface between weather and climate forecasts in the sense that for predictions of daily values a direct relationship between forecasted and observed values cannot be exploited efficiently anymore through techniques such as model output statistics. This study shows clearly that a climatological approach such as correcting the distribution as a whole indeed is capable to improve the forecasts nevertheless and might help to work toward a seamless approach with seasonal and decadal forecasts or even climate change scenarios.

Acknowledgments

Konrad Bogner's contribution is part of the Swiss Competence Centre for Energy Research- Supply of Electricity (SCCER-SoE) and is funded by the Commission for Technology and Innovation (CTI). Samuel Monhart's contribution is financed by the NRP 70-Energy Turnaround project (grant 407040153929). ECMWF forecast data are accessible through the MARS archive: http://apps.ecmwf.int/archive-catalogue/, observation records used in this study are available from the following websites: https://data.noaa.gov/dataset/dataset/global-surface-summary-of-the-day-gsod (NOAA GSOD), https://www.ecad.eu/dailydata/index.php (ECA&D), and https://gate.meteoswiss.ch/idaweb/more.do (SwissMetNet). The authors thank Ana Casanueva from MeteoSwiss for preparing the data from the GOSD data set for this analysis and three anonymous reviewers for insightful comments to the original manuscript.

References

Addor, N., Jaun, S., Fundel, F., & Zappa, M. (2011). An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): Skill, case studies and scenarios. *Hydrology and Earth System Sciences*, *15*(7), 2327–2347. https://doi.org/10.5194/hess-15-2327-2011

Addor, N., Rössler, O., Köplin, N., Huss, M., Weingartner, R., & Seibert, J. (2014). Robust changes and sources of uncertainty in the projected hydrological regimes of Swiss catchments. *Water Resources Research*, *50*, 7541–7562. https://doi.org/10.1002/2014WR015549

Alemu, E. T., Palmer, R. N., Polebitski, A., & Meaker, B. (2011). Decision support system for optimizing reservoir operations using ensemble streamflow predictions. *Journal of Water Resources Planning and Management*, *137*(1), 72–82. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000088

Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, *9*(7), 1518–1530. https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2

Anghileri, D., Castelletti, A., Pianosi, F., Soncini-Sessa, R., & Weber, E. (2013). Optimizing watershed management by coordinated operation of storing facilities. *Journal of Water Resources Planning and Management*, *139*(5), 492–500. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000313

Barkhordarian, A., Storch, H., & Zorita, E. (2016). An attempt to deconstruct recent climate change in the Baltic Sea basin. *Journal of Geophysical Research: Atmospheres*, *121*, 207–217. https://doi.org/10.1002/2015JD024648

Barnston, A. G., & van den Dool, H. M. (1993). A degeneracy in cross-validated skill in regression-based forecasts. *Journal of Climate*, *6*(5), 963–977. https://doi.org/10.1175/1520-0442(1993)006<0963:ADICVS>2.0.CO;2

Barros, M. T. L., Tsai, F. T.-C., Yang, S., Lopes, J. E. G., & Yeh, W. W.-G. (2003). Optimization of large-scale hydropower system operations. *Journal of Water Resources Planning and Management*, *129*(3), 178–188. https://doi.org/10.1061/(ASCE)0733-9496(2003)129:3(178)

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Bennett, J. C., Grose, M. R., Corney, S. P., White, C. J., Holz, G. K., Katzfey, J. J., et al. (2014). Performance of an empirical bias-correction of a high-resolution climate dataset. *International Journal of Climatology*, *34*(7), 2189–2204. https://doi.org/10.1002/joc.3830

Bogner, K., Liechti, K., Luzi, B., Monhart, S., & Zappa, M. (2018). Skill of hydrological extended range forecasts for water resources management in Switzerland. *Water Resources Management*, *32*, 969–984. https://doi.org/10.1007/s11269-017-1849-5

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y., & Wei, M. (2005). A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, *133*(5), 1076–1097. https://doi.org/10.1175/MWR2905.1

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

Calanca, P., Bolius, D., Weigel, A. P., & Liniger, M. A. (2011). Application of long-range weather forecasts to agricultural decision problems in Europe. *The Journal of Agricultural Science*, *149*(1), 15–22. https://doi.org/10.1017/S0021859610000729

Cannon, A. J. (2011). Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Computers & Geosciences*, *37*(9), 1277–1284. https://doi.org/10.1016/j.cageo.2010.07.005

Ceccherini, G., Russo, S., Ameztoy, I., Marchese, A. F., & Carmona-Moreno, C. (2017). Heat waves in Africa 1981–2015, observations and reanalysis. *Natural Hazards and Earth System Sciences Discussions*, *17*, 115–125. https://doi.org/10.5194/nhess-17-115-2017

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, *83*(403), 596.

Cleveland, W. S., Grosse, E., & Shyu, W. M. (1992). Local regression models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical models in S* (Chap. 8, pp. 309–376). Wadsworth & Brooks/Cole, Pacific Grove.

Crochemore, L., Ramos, M. H., & Pappenberger, F. (2016). Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrology and Earth System Sciences*, *20*(9), 3601–3618. https://doi.org/10.5194/hess-20-3601-2016

Crochemore, L., Ramos, M. H., Pappenberger, F., & Perrin, C. (2017). Seasonal streamflow forecasting by conditioning climatology with precipitation indices. *Hydrology and Earth System Sciences*, *21*(3), 1573–1591. https://doi.org/10.5194/hess-21-1573-2017

Doblas-Reyes, F.J. (2010). Seasonal prediction over Europe. *ECMWF Seminar on predictability in the European and Atlantic regions* (pp. 171–185).

Doblas-Reyes, F. J., Hagedorn, R., Palmer, T. N., & Morcrette, J. J. (2006). Impact of increasing greenhouse gas concentrations in seasonal ensemble forecasts. *Geophysical Research Letters*, *33*, L07708. https://doi.org/10.1029/2005GL025061

ECMWF (2014). IFS DOCUMENTATION—Cy40r1 operational implementation 22 November 2013 PART V: ENSEMBLE PREDICTION SYSTEM. ECMWF Documentation, 1–25.

Ferro, C. A. T., Richardson, D. S., & Weigel, A. P. (2008). On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, *15*(1), 19–24. https://doi.org/10.1002/met.45

Fowler, H. J., Blenkinsop, S., & Tebaldi, T. (2007). Review Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology*, *27*, 1547–1578.

Fundel, F., Jörg-Hess, S., & Zappa, M. (2013). Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices. *Hydrology and Earth System Sciences*, *17*(1), 395–407. https://doi.org/10.5194/hess-17-395-2013

Gangsto, R., Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2013). Methodological aspects of the validation of decadal predictions. *Climate Research*, *55*(3), 181–200. https://doi.org/10.3354/cr01135

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., & Engen-Skaugen, T. (2012). Technical note: Downscaling RCM precipitation to the station scale using statistical transformations—A comparison of methods. *Hydrology and Earth System Sciences*, *16*(9), 3383–3390. https://doi.org/10.5194/hess-16-3383-2012

Hagedorn, R., Hamill, T. M., & Whitaker, J. S. (2007). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Monthly Weather Review*, *136*, 2608–2619.

Hamill, T. M., & Colucci, S. J. (1997). Verification of eta – RSM short-range ensemble forecasts. *Monthly Weather Review*, *125*(6), 1312–1327. https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2

Hamill, T. M., Hagedorn, R., & Whitaker, J. S. (2008). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Monthly Weather Review*, *136*, 258–259.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Hirschi, M., Spirig, C., Weigel, A. P., Calanca, P., Samietz, J., & Rotach, M. W. (2012). Monthly weather forecasts in a pest forecasting context: Downscaling, recalibration, and skill improvement. *Journal of Applied Meteorology and Climatology*, *51*(9), 1633–1638. https://doi.org/10.1175/JAMC-D-12-082.1

Johansson, A., Barnston, A., Saha, S., & Van den Pool, H. (1998). On the level and origin of seasonal forecast skill in northern Europe. *American Meteorological Society*, 103–127.

Jolliffe, I. T., & Stephenson, D. B. (Eds.) (2012). *Forecast verification: A practioner's guide in atmospheric science* (2nd ed.). Oxford: John Wiley.

Jörg-Hess, S., Kempf, S. B., Fundel, F., & Zappa, M. (2015). The benefit of climatological and calibrated reforecast data for simulating hydrological droughts in Switzerland. *Meteorological Applications*, *22*, 444–458.

Kilibarda, M., Tadić, M. P., Hengl, T., Luković, J., & Bajat, B. (2015). Global geographic and feature space coverage of temperature data in the context of spatio-temporal interpolation. *Spatial Statistics*, *14*, 22–38. https://doi.org/10.1016/j.spasta.2015.04.005

Klein Tank, A. M. G., Wijngaard, J. B., Können, G. P., Böhm, R., Demare, G., Gocheva, A., et al. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the European climate assessment. *International Journal of Climatology*, *1453*, 1441–1453.

Kotlarski, S., Bosshard, T., Lüthi, D., Pall, P., & Schär, C. (2012). Elevation gradients of European climate change in the regional climate model COSMO-CLM. *Climatic Change*, *112*(2), 189–215. https://doi.org/10.1007/s10584-011-0195-5

Kotlarski, S., Keuler, K., Christensen, O. B., Colette, A., Déqué, M., Gobiet, A., et al. (2014). Regional climate modeling on European scales: A joint standard evaluation of the EURO-CORDEX RCM ensemble. *Geoscientific Model Development*, *7*(4), 1297–1333. https://doi.org/10.5194/gmd-7-1297-2014

Li, S., & Robertson, A. W. (2015). Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. *Monthly Weather Review*, *143*(7), 2871–2889. https://doi.org/10.1175/MWR-D-14-00277.1

Lienert, F., & Doblas-Reyes, F. J. (2017). Prediction of interannual North Atlantic Sea surface temperature and its remote influence over land. *Climate Dynamics*, *48*(9–10), 3099–3114. https://doi.org/10.1007/s00382-016-3254-9

Liniger, M. A., Mathis, H., Appenzeller, C., & Doblas-Reyes, F. J. (2007). Realistic greenhouse gas forcing and seasonal forecasts. *Geophysical Research Letters*, *34*, L04705. https://doi.org/10.1029/2006GL028335

Mahlstein, I., Spirig, C., Liniger, M. A., & Appenzeller, C. (2015). Estimating daily climatologies for climate indices derived from climate model data and observations. *Journal of Geophysical Research: Atmospheres*, *120*, 2808–2818. https://doi.org/10.1002/2014JD022327

Maraun, D. (2013). Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *Journal of Climate*, *26*(6), 2137–2143. https://doi.org/10.1175/JCLI-D-12-00821.1

Mason, I. (1982). A model for assessment of weather forecasts. *Australian Meteorological Magazine*, *30*(4), 291–303.

MeteoSwiss (2016). EasyVerification: Ensemble forecast verification for large data sets. Retrieved from https://cran.r-project.org/web/package=easyVerification (accessed 02. September 2016).

Orth, R., & Seneviratne, S. I. (2013). Predictability of soil moisture and streamflow on subseasonal timescales: A case study. *Journal of Geophysical Research: Atmospheres*, *118*, 10,963–10,979. https://doi.org/10.1002/jgrd.50846

Park, Y., Buizza, R., & Leutbecher, M. (2008). TIGGE: Preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society*, *134*(637), 2029–2050. https://doi.org/10.1002/qj.334

Piani, C., Haerter, J. O., & Coppola, E. (2010). Statistical bias correction for daily precipitation in regional climate models over Europe. *Theoretical and Applied Climatology*, *99*(1-2), 187–192. https://doi.org/10.1007/s00704-009-0134-9

Rajczak, J., Kotlarski, S., & Schär, C. (2016). Does quantile mapping of simulated precipitation correct for biases in transition probabilities and spell lengths? *Journal of Climate*, *29*(5), 1605–1615. https://doi.org/10.1175/JCLI-D-15-0162.1

Reyers, M., Pinto, J. G., & Moemken, J. (2014). Statistical-dynamical downscaling for wind energy potentials: Evaluation and applications to decadal hindcasts and climate change projections. *International Journal of Climatology*, *35*(2), 229–244.

Robertson, A. W., Kumar, A., Peña, M., & Vitart, F. (2015). Improving and promoting subseasonal to seasonal prediction. *Bulletin of the American Meteorological Society*, *96*(3), ES49–ES53.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., et al. (2014). The NCEP climate forecast system version 2. *Journal of Climate*, *27*(6), 2185–2208. https://doi.org/10.1175/JCLI-D-12-00823.1

Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R. T., Dunstone, N., et al. (2014). Skillful long-range prediction of European and north American winters. *Geophysical Research Letters*, *5*, 2514–2519. https://doi.org/10.1002/2014GL059637

Schefzik, R. (2017). Ensemble calibration with preserved correlations: Unifying and comparing ensemble copula coupling and member-by-meber postprocessing. *Quarterly Journal of the Royal Meteorological Society*, *143*(703), 999–1008. https://doi.org/10.1002/qj.2984

Shabbar, A., & Kharin, V. V. (2007). An assessment of cross-validation for estimating skill of empirical seasonal forecasts using a global coupled model simulation. *CLIVAR Exchanges*, *12*(4), 10–12.

Shi, W., Schaller, N., Macleod, D., Palmer, T. N., & Weisheimer, A. (2015). Impact of hindcast length on estimates of seasonal climate predictability. *Geyophysical Research Letters*, *42*, 1554–1559. https://doi.org/10.1002/2014GL062829

Siegert, S. (2015). Specs verification: Forecast verification routines for the SPECS FP7 project. Retrieved from http://cran.r-project.org/web/packages/SpecsVerification (accessed 02. September 2016).

Talagrand, O., Vautard, R., & Strauss, B. (1997). Evaluation of probabilistic prediction systems. In *Proceedings of a workshop held at ECMWF on predictability* (pp. 1–25). Reading, UK: European Center for Medium-Range Weather Forecasts.

Teutschbein, C., & Seibert, J. (2012). Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology*, *456–457*, 12–29.

Themeßl, M. J., Gobiet, A., & Heinrich, G. (2012). Empirical-statistical downscaling and error correction of regional climate models and its impact on the climate change signal. *Climatic Change*, *112*(2), 449–468. https://doi.org/10.1007/s10584-011-0224-4

Themeßl, M. J., Gobiet, A., & Leuprecht, A. (2011). Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *International Journal of Climatology*, *31*(10), 1530–1544. https://doi.org/10.1002/joc.2168

van Oldenborgh, G. J., Drijfhout, S., van Ulden, A., Haarsma, R., Sterl, A., Severijns, C., & Hazeleger, W. (2009). Western Europe is warming much faster than expected. *Climate of the Past*, *5*(1), 1–12. https://doi.org/10.5194/cp-5-1-2009

Verkade, J. S., Brown, J. D., Reggiani, P., & Weerts, A. H. (2013). Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology*, *501*, 73–91. https://doi.org/10.1016/j.jhydrol.2013.07.039

Vitart, F. (2004). Monthly forecasting at ECMWF. *Monthly Weather Review*, *132*(12), 2761–2779. https://doi.org/10.1175/MWR2826.1

Vitart, F. (2014). Evolution of ECMWF sub-seasonal forecast skill scores. *Quarterly Journal of the Royal Meteorological Society*, *140*(683), 1889–1899. https://doi.org/10.1002/qj.2256

Vitart, F., Balsamo, G., Buizza, R., Ferranti, L., Keeley, S., Magnusson, L., et al. (2014). Sub-seasonal predictions. ECMWF Technical Memoranda, (October).

Vitart, F., Buizza, R., Balmaseda, M. A., Balsamo, G., Bidlot, J., Bonet, A., et al. (2008). The new VarEPS—Monthly forecasting system: A first step. *Quarterly Journal of the Royal Meteorological Society*, *1799*, 1789–1799.

Weigel, A. P., Baggenstos, D., Liniger, M. A., Vitart, F., & Appenzeller, C. (2008). Probabilistic verification of monthly temperature forecasts. *Monthly Weather Review*, *136*(12), 5162–5182. https://doi.org/10.1175/2008MWR2551.1

Weisheimer, A., & Palmer, T. N. (2014). On the reliability of seasonal climate forecasts. *Journal of the Royal Society Interface*, (96), 11, 20131162. https://doi.org/10.1098/rsif.2013.1162

Werner, A. T., & Cannon, A. J. (2015). Hydrologic extremes—An intercomparison of multiple gridded statistical downscaling methods. *Hydrology and Earth System Sciences Discussions*, *12*(6), 6179–6239. https://doi.org/10.5194/hessd-12-6179-2015

Wilks, D. S. (Ed.) (2011). *Statistical methods in the atmospheric sciences*, *International geophysics series* (3rd ed., Vol. 100). London: Academic Press Inc.

Wood, A. W., Leung, L. R., Sridhar, V., & Lettenmaier, D. P. (2004). Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, *62*(1–3), 189–216. https://doi.org/10.1023/B:CLIM.0000013685.99609.9e

Wood, A. W., Maurer, E. P., Kumar, A., & Lettenmaier, D. P. (2002). Long-range experimental hydrologic forecasting for the eastern United States. *Journal of Geophysical Research*, *107*(D20), 4429. https://doi.org/10.1029/2001JD000659

Zhao, T., Bennett, J., Wang, Q. J., Schepen, A., Wood, A., Robertson, D., & Ramos, M.-H. (2017). How suitable is quantile mapping for post-processing GCM precipitation forecasts. *Journal of Climate*, *30*(9), 3185–3196. https://doi.org/10.1175/JCLI-D-16-0652.1