Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts

Louise Crochemore[1], Maria-Helena Ramos[1], and Florian Pappenberger[2,3]

[1]Irstea, Hydrosystems and Bioprocesses Research Unit, 1 rue Pierre Gilles de Gennes, F- 92 761, Antony, France.
[2]ECMWF, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, UK.
[3]School of Geographical Sciences, University of Bristol, University Road, Bristol, BS8 1SS, UK.

Correspondence to: Louise Crochemore (louise.crochemore@irstea.fr)

Abstract. Meteorological centres make sustained efforts to provide seasonal forecasts that are increasingly skilful, which has the potential to benefit streamflow forecasting. Seasonal streamflow forecasts can help to take anticipatory measures for a range of applications, such as water supply or hydropower reservoir operation and drought risk management. This study assesses the skill of seasonal precipitation and streamflow forecasts in France to provide insights into the way bias correcting precipitation forecasts can improve the skill of streamflow forecasts at extended lead times. We apply eight variants of bias correction approaches to the precipitation forecasts prior to generating the streamflow forecasts. The approaches are based on the linear scaling and the distribution mapping methods. A daily hydrological model is applied at the catchment scale to transform precipitation into streamflow. We then evaluate the skill of raw (without bias correction) and bias corrected precipitation and streamflow ensemble forecasts in sixteen catchments in France. The skill of the ensemble forecasts is assessed in reliability, sharpness, accuracy, and overall performance. A reference prediction system, based on historical observed precipitation and catchment initial conditions at the time of forecast (i.e., ESP method), is used as benchmark in the computation of the skill. The results show that, in most catchments, raw seasonal precipitation and streamflow forecasts are often more skilful than the conventional ESP method in terms of sharpness. However, they are not significantly better in terms of reliability. Forecast skill is generally improved when applying bias correction. Two bias correction methods show the best performance for the studied catchments, each method being more successful in improving specific attributes of the forecasts: the simple linear scaling of monthly values contributes mainly to increasing forecast sharpness and accuracy, while the empirical distribution mapping of daily values is successful in improving forecast reliability.

1 Introduction

Numerous activities with economic, environmental and political stakes benefit from knowing and anticipating future streamflow conditions at different lead times. Streamflow forecasting systems are frequently developed to take the latest useful information content into account (e.g. last observed discharges, soil moisture or snow cover) and to make use of numerical weather model outputs to extend the range of skilful predictions.

Seasonal forecasts have shown to perfectly fall within a context of proactive risk management, for example, for drought management (e.g. Wilhite et al., 2000; Dutra et al., 2014; Mwangi et al., 2014; Wetterhall et al., 2015). Extended-range fore-

casting systems can be valuable to help decision-makers in planning long-term strategies for water storage (Crochemore et al., 2016) and to support adaptation to climate change (Winsemius et al., 2014). Nevertheless, several users still remain doubtful whether seasonal forecasts can be trustworthy or skilful enough to enhance decision-making (Rayner et al., 2005). Lemos et al. (2002) list the performance of seasonal forecasts, the misuse of seasonal forecasts by end-users and the lack of consideration of end-users' needs in the development of products as major obstacles to the widespread use of seasonal forecasting in North-East Brazil. It is therefore crucial to assess the potential of available seasonal forecasting products and communicate on the assets and shortcomings of the different approaches for the water sector (Hartmann et al., 2002).

Seasonal forecasting methods in hydrology can be broadly divided into two categories: statistical methods, which use a statistical relationship between a predictor and a predictand (e.g. Jenicek et al., 2016, and references therein), and dynamical methods, which use seasonal meteorological forecasts as input to a hydrological model. More recently, mixed approaches have been investigated to take advantage of initial land surface conditions, seasonal predictions of atmospheric variables and the predictability information contained in large-scale climate features (see Robertson et al., 2013; Yuan et al., 2015, and references therein). Ensemble Streamflow Prediction (ESP; Day, 1985) is a dynamical method that is widely used to forecast low flows and reservoir inflows at long lead times (Faber and Stedinger, 2001; Nicolle et al., 2014; Demirel et al., 2015). It consists in using historical weather data as input to a hydrological model whose states were initialized for the time of the forecast. The ESP method is also used along with the Reverse-ESP method to determine the relative impacts of meteorological forcings and hydrological initial conditions on the skill of streamflow predictions (Wood and Lettenmaier, 2008; Shukla et al., 2013; Yossef et al., 2013). An alternative dynamical method consists in using seasonal forecasts from regional climate models (RCMs) (Wood et al., 2005). This approach yields better results when seasonal predictability is enhanced by meteorological forcings. Climate model outputs may also be more suitable to capture the specific climate conditions at the time of the forecast, whereas ESP-based methods will be limited to the range of past observations and challenged by climate non-stationarity.

The use of climate model outputs in hydrology has however some methodological implications. For instance, outputs are produced for coarse grid scales, which can lead to errors in capturing forecast uncertainty and induce biases. Post-processing (including bias correction and downscaling) is usually a necessary first step prior to using climate model outputs to model streamflow. A range of methods has been proposed in the literature, with performance varying depending on on the modelling chain and the studied area (Christensen et al., 2008; Gudmundsson et al., 2012). Weather forecasting has performed bias correction of numerical model outputs through model output statistics (MOS) for decades. In hydrologic ensemble prediction systems, post-processing has become more and more popular in the last decade, particularly for medium-range ensemble forecasting (e.g. Weerts et al., 2011; Zalachori et al., 2012; Verkade et al., 2013; Madadgar et al., 2014; Roulin and Vannitsem, 2015). In seasonal forecasting, two popular bias correction methods are linear scaling and distribution mapping (Yuan et al., 2015). Linear scaling corrects the mean of the forecasts based on the difference between observed and forecast means, whereas distribution mapping matches the statistical distribution of forecasts to the distribution of observations. These approaches focus on increasing forecast skill and reliability, by reducing errors in the forecast mean and improving forecast spread.

Studies comparing different bias correction methods in seasonal hydrological forecasting are still rare in the literature. However, we can find studies reviewing and comparing methods to bias correct RCM outputs and quantify climate change impacts,

although their efficiency in this context is still a topic of discussion (Ehret et al., 2012; Muerth et al., 2013; Teutschbein and Seibert, 2013). Teutschbein and Seibert (2012) compared six methods, among which linear scaling and parametric distribution mapping, to bias correct RCM simulations of precipitation and temperature in Sweden. The authors recommended using the distribution mapping method for current climate conditions. They also highlighted the need to assume that bias correction procedures are stationary to correct future climate projections and evaluate changes in flow regimes. In Norway, Gudmundsson et al. (2012) proposed a comparison of eleven methods to bias correct RCM precipitation, including distribution mapping based on fitted theoretical or empirical distributions and linear scaling. Their study highlighted the differences between the bias corrections and the necessity to test methods prior to their application. The authors recommended using nonparametric methods since these methods were the most effective to reduce the bias and did not require any approximations of the empirical distributions.

The European Centre for Medium-range Weather Forecasts (ECMWF) produces seasonal forecasts from GCM simulations (Molteni et al., 2011). Weisheimer and Palmer (2014) evaluated the reliability of the precipitation forecasts issued by ECMWF System 4 on a scale ranging from "dangerous" to "perfect". Over the world, forecasts often fell within the "marginally useful" category. In France, they were ranked as "marginally useful" during wet winters and summers, "not useful" in dry winters, and "dangerous" in dry summers. Kim et al. (2012) also evaluated the skill of System 4 precipitation and temperature forecasts at the global scale. Despite good overall performances, they identified systematic biases, e.g. a warm bias in the North Atlantic. Several studies have proposed to bias correct ECMWF System 4 forecasts in different contexts. Di Giuseppe et al. (2013) applied a spatially-based precipitation bias correction to improve malaria forecasts. Trambauer et al. (2015) applied a linear scaling method to forecast hydrological droughts in Southern Africa. In the same context, Wetterhall et al. (2015) applied a quantile mapping method to daily precipitation values, and showed that bias correction was able to improve the skill of dry spell forecasts.

Despite these recent works, and to the knowledge of the authors, no previous study has compared bias correction methods and their impact on streamflow forecasting in a systematic way, with a focus on understanding how the main attributes of forecast performance are impacted by bias correction. This paper aims to provide insights into the way bias correcting seasonal precipitation forecasts can contribute to the skill of seasonal streamflow predictions, notably in terms of overall performance, reliability, sharpness and skilful lead time. It investigates the potential of bias corrected ECMWF System 4 forecasts to improve streamflow forecasts at extended lead times over 16 catchments in France. An in-depth comparison of eight variants of linear scaling and distribution mapping methods applied over the 1981-2010 period is presented. Section 2 presents the catchment set, the forecast and observed data, as well as the hydrological model used. Section 3 presents the bias correction methods investigated, as well as the calibration and evaluation frameworks adopted. Results are presented in Sections 4 to 6 for the quality of the raw (uncorrected) and the bias corrected forecasts. In Section 7, conclusions and limitations are discussed.

2 Data and hydrological model

2.1 Seasonal forecasts and observed data

Daily seasonal precipitation forecasts come from ECMWF System 4, which provides ensemble forecasts for the next seven months at a TL255 (about 0.7°) spatial resolution, for the period running from 1981 to 2010 (Molteni et al., 2011). Forecasts are composed of 51 ensemble members for February, May, August and November, and 15 members for the other months. In this study, areal precipitations were computed for each catchment, and only the first 90 days of the forecast horizon were considered.

Daily observed precipitations used for the calibration and evaluation of the bias correction methods come from the 8x8 km grid resolution SAFRAN reanalysis of Météo-France (Quintana-Seguí et al., 2008; Vidal et al., 2010). They were also aggregated at the catchment scale. Mean areal potential evapotranspiration was computed for each catchment based on daily observed temperatures from the SAFRAN reanalysis (Oudin et al., 2005). The interannual potential evapotranspiration was then computed in each catchment, i.e. for a given day of the year, we computed the average potential evapotranspiration for this day over all available years (1958 to 2010). Daily streamflow data at the outlet of each catchment come from the French national archive (*Banque Hydro*).

2.2 Studied catchments and hydrological model

The catchment set was selected from the database in Nicolle et al. (2014). It comprises 16 catchments in France (Fig. 1) with a dominant pluvial regime. Catchments show an average solid fraction of precipitation below 10% and are thus not heavily influenced by snow. Their main characteristics are shown in Table 1.

We applied the conceptual, reservoir-based GR6J hydrological model (Pushpalatha et al., 2011) at the daily time step. The model has three reservoirs (one for the production function and two for the routing function), and one unit hydrograph to account for flow delays. The model inputs are daily precipitation and potential evapotranspiration at the catchment scale. The model output is the daily streamflow at the catchment outlet. Here, the series of interannual potential evapotranspiration corresponding to the forecast period was systematically used as input to the hydrological model. With this setup, we aimed to isolate the influence of precipitation forecast inputs on the quality of streamflow forecasts. This setup is also consistent with the fact that our catchment set is dominated by a pluvial regime. The model was calibrated in each catchment with the Kling-Gupta Efficiency (Gupta et al., 2009) applied to root-squared flows. We obtained an average KGE of 0.95 in calibration and 0.94 in validation over the sixteen catchments. The bias obtained in simulation ranges from 0.95 to 1.02. When the model is applied to forecast streamflow, the model states are initialized by running the model in simulation mode for the year preceding the forecast date. The last observed streamflow is then used to update the levels of the routing reservoirs before issuing the forecast.

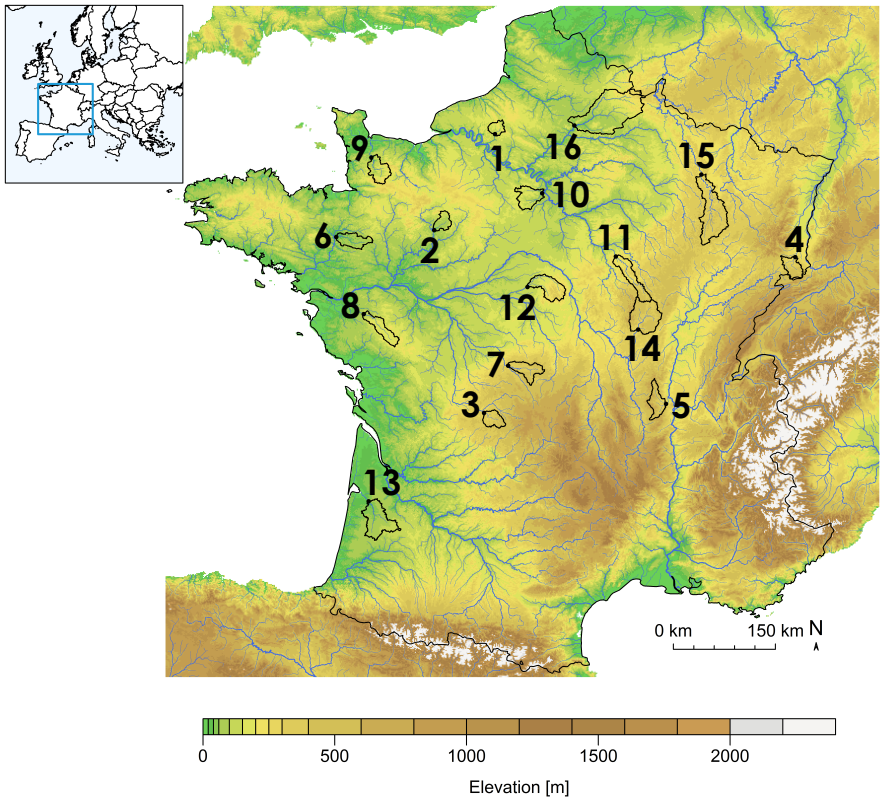

Figure 1. Location of the studied catchments in France, identified by their numbers (see Table 1).

3 Methods

3.1 Overview of the calibration approach

The leave-one-year-out cross-validation method (Arlot and Celisse, 2010) was applied to calibrate the bias correction methods in each catchment over independent periods within the 1981-2010 period. Given a target application year, all available years but the target year are used in the calibration process. Results of the calibration are then applied to the target year and bias corrected forecasts are evaluated against observations.

In the calibration step, we considered two approaches: (1) all days of the years within the calibration dataset are used, (2) the bias correction methods are calibrated for each calendar month. Additionally, since we are dealing with forecasts issued up to 90 days ahead, and since forecast performance varies with lead time, calibration also takes the lead time into account. Lead times were grouped from 1 to 30 days, 31 to 60 days and 61 to 90 days ahead. The calibrated bias correction factors are then applied to the daily values of the ensemble precipitation forecasts in the target year. The hydrological model is forced by raw and bias corrected precipitation forecasts, which results in streamflow ensemble forecasts.

3.2 Bias correction methods

We applied the linear scaling (LS) and the distribution mapping (DM) methods to the raw System 4 precipitation forecasts. The DM method was applied considering the empirical distribution of monthly values (EDM), a fitted gamma distribution of monthly values (GDM), and the empirical distribution of daily values (EDMD). Each method was applied on a monthly (-m) or a yearly (-y) basis (Table 2).

3.2.1 Linear scaling of precipitations

LS consists in correcting the monthly mean values of the forecasts to match the monthly mean values of the observations. A scaling factor (or bias) is calculated considering the ratio between the observed and the forecast (ensemble mean) values. A scaling factor higher (lower) than 1 indicates that the mean ensemble forecast underpredicts (overpredicts) the mean observed value. A value of 1 indicates no bias in the forecasts. The scaling factor obtained through calibration is then applied as a multiplicative factor to correct raw daily precipitation forecasts.

3.2.2 Distribution mapping of precipitations

DM consists in correcting the precipitation forecasts so that their statistical distribution matches that of the observations. There are several ways to match forecast and observed distributions or quantiles, and existing techniques mainly differ on how the cumulative distribution functions (CDF) are considered. In some techniques, a parametric distribution is fitted to the datasets, while in others the empirical distributions and linear interpolations between data points or estimated quantiles are considered.

In this study, the calibration of the DM method was first carried out considering empirical (EDM) and gamma-fitted (GDM) distributions of observed and forecast (ensemble mean) precipitation values averaged monthly. A third variant considered directly the empirical distribution of the daily values of the ensemble members (EDMD). These variants are listed in Table 2. After calibration, bias correction is applied to the daily precipitation forecasts of each target period. In the case of EDM and GDM, the monthly values are first corrected based on the distribution mapping procedure. Then, for a given month, the ratio of the corrected monthly value and the non-corrected one is used to correct all daily values within this month. In the case of EDMD, each daily precipitation value of each forecast member is corrected individually.

3.3 Evaluation framework

The quality of the forecasts was evaluated as a function of lead time and for the winter (December-January-February), the spring (March-April-May), the summer (June-July-August) and the autumn (September-October-November) seasons. Four criteria were used to assess reliability, sharpness, accuracy and overall performance of the forecasts (Gneiting et al., 2007; Eslamian, 2015; Musy et al., 2015).

3.3.1 Evaluation criteria

Reliability is a forecast attribute that refers to the statistical consistency between observed frequencies and forecast probabilities. In this study, it was evaluated with the Probability Integral Transform (PIT) diagram (Gneiting et al., 2007; Laio and Tamea, 2007). The PIT diagram is the cumulative distribution of the PIT values, which are defined by the values of the predictive distribution function at the observations, computed at each time step. In the case of a reliable forecast, the observations uniformly fall within the predictive distribution and the PIT diagram coincides with the 1:1 diagonal. If the PIT diagram is systematically above (below) the diagonal, the observed values are too frequently located in the lower (upper) parts of the forecast distribution, suggesting a systematic bias of the forecasts towards overprediction (underprediction). If the PIT diagram tends to resemble a horizontal line, observations fall too frequently in the tails of the forecast distribution, indicating that forecasts are too narrow. On the contrary, if the PIT diagram is closer to a vertical line, too many observations fall in the midrange of the forecast distribution, indicating that forecasts are too wide. We also represented the Kolmogorov significance bands at +0.1 and -0.1 from the bisector, which ensure a 5% significance. In order to numerically compare results among catchments, we also computed the area between the curve of the PIT diagram and the 1:1 diagonal, as proposed by Renard et al. (2010). The smaller this area, the more reliable the ensemble.

Sharpness is a property of the forecasts only. It refers to the concentration of the predictive distribution and indicates how spread the members of an ensemble forecast are. In this study, sharpness was evaluated with the 90% interquantile range (IQR), i.e. the difference between the 95^{th} and the 5^{th} percentiles of the forecast distribution. The final IQR score is the average of the interquantile range at each time step of the evaluation period. The narrower the IQR, the sharper the ensemble. In this study, we considered that, given two reliable systems, the sharpest one is the best (Gneiting et al., 2007).

The accuracy of the forecasts is assessed with the mean absolute error (MAE). The MAE computes the average (over the evaluation period) of the absolute difference between the forecast ensemble mean and the observed value. Smaller MAE values correspond to more accurate forecasts.

Lastly, the Continuous Ranked Probability Score (CRPS) evaluates the overall performance of the forecasts. It is defined as the integral of the squared distance between the cumulative distribution of the forecast members and a step function for the observation (Hersbach, 2000). The CRPS score is the average of this integral computed at each time step of the evaluation period. The lower the CRPS, the better the overall performance of the forecasts.

3.3.2 Skill scores

Forecast skill is evaluated by comparing the performance of a given forecast system with the performance of a reference forecast. The skill score is computed for a given lead time *i*.

$$SkillScore_i = 1 - \frac{Score_i^{Syst}}{Score_i^{Ref}} \quad (1)$$

When the skill score is superior (inferior) to zero, the forecast system is more (less) skilful than the reference. When it is equal to zero, the system and the reference have equivalent skill.

The skill scores were computed for the probabilistic scores. They are noted PITSS, IQRSS and CRPSS. The reference precipitation forecast is based on past observations and is representative of the catchment climatology: for a given day and year, it is the ensemble of precipitation values observed on that same Gregorian day in other years of the observation period (1958 to 2010). Two reference streamflow forecasts are used. The first is the Ensemble Streamflow Prediction (ESP), which corresponds to the streamflow ensemble obtained when the reference precipitation ensemble is used as input to the hydrological model. The ESP is a commonly used method in seasonal forecasting. It allows applying the same hydrological modelling setup to both the precipitation forecasts and the reference precipitation ensemble. Therefore, differences in performance are mainly due to differences between the precipitation inputs to the model. The second reference is based on past streamflow observations (on the same day as the given forecast day, in a 36- to 52-year period running up to 2010) to evaluate performance. This reference ensemble does not use any precipitation forecasts or hydrological model.

Finally, several studies have shown that the ensemble size induces a bias when computing skill scores with ensembles of different sizes. This bias usually leads to an underestimation of the skill of the forecast system when the system has fewer members than the reference. Ferro et al. (2008) provide a synthesis of previous studies on the influence of ensemble size on probability scores and propose a correction factor to remove the bias in the computation of CRPS skill scores. This correction was applied to compute the CRPSS in this study. Since the ensemble size of System 4 precipitation forecasts varies with the month, we used the ensemble size averaged over one year.

3.3.3 Gain in lead time from bias correcting seasonal forecasts

To investigate the gain in performance brought by bias correction methods, we use the raw (uncorrected) forecasts as reference in the computation of the skill scores. An indicator of forecast performance can be derived: the lead time up to which bias corrected forecasts have more skill than raw forecasts. Nicolle et al. (2014) defined an indicator named UFL (Useful Forecasting Lead time) as the first "lead time beyond which model performance is not at least 20% better than benchmark performance". Here, we considered the lead time beyond which the seven-day moving average of the skill score becomes negative. UFL values were then grouped in four categories: (1) None: no improvement over the forecast reference, (2) <30: gain up to 30 days, (3) <60: gain greater than 30 days and up to 60 days and (4) >60: gain greater than 60 days.

4 Quality of the raw seasonal forecasts

4.1 Performance of raw precipitation forecasts

Figure 2 presents the evolution of IQRSS and CRPSS with lead time, for winter (DJF) and summer (JJA). Each line corresponds to a catchment. Skill in sharpness and overall performance is very similar in winter and in summer (as well as in spring and autumn, not shown). Precipitation forecasts are overall sharper than historical precipitations in the large majority of catchments and up to long lead times. Some exceptions appear for lead times longer than three weeks, and especially in winter (wetter season in the majority of catchments). In terms of overall performance, precipitation forecasts clearly have skill up to two to

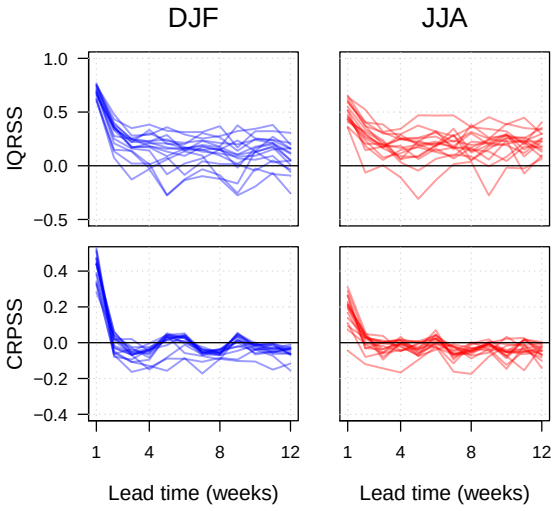

Figure 2. Skill of raw weekly precipitation forecasts as a function of the lead time for all catchments and for the winter (DJF) and summer (JJA) seasons. The skill is computed based on the IQR (top) and the CRPS (bottom) and the reference is historical precipitations. Each column corresponds to a target season. Each line represents the skill score in a catchment for forecast horizons within the target season.

three weeks ahead for 7-day averaged areal precipitation. At longer lead times, they are equivalent or perform slightly worse than historical precipitations.

Figure 3 shows the PIT diagrams for lead times of 30 and 90 days, for winter and summer. Grey lines represent the reliability of historical precipitations and coloured lines represent the reliability of System 4 precipitation forecasts in each catchment. Dotted lines represent the Kolmogorov significance bands to ensure a 5% significance test. The two seasons yield very similar results (also observed in spring and autumn, not shown). In all catchments and for both lead times, historical precipitations are reliable, as expected. Seasonal precipitation forecasts also show some reliability, but tend to overpredict precipitations in both seasons and at both lead times. The concentration of points in the zero end points in most of the curves shows that low values of the observations are too often falling in the lower tail of the forecast distribution. This effect tends to decrease with increasing lead time. This is an indication that forecasts are too narrow and overpredict the lowest observations. It can also indicate a difficulty of the system to forecast null precipitation.

4.2 Performance of raw streamflow forecasts

Streamflow forecasts are generated by using raw precipitation forecasts as input to the hydrological model. Forecast skill is evaluated using the ESP method as reference (Fig. 4). Differences in forecast skill between the winter and summer seasons are more noticeable when evaluating streamflow forecasts rather than precipitation forecasts. Streamflow forecasts generated from raw precipitation forecasts are sharper than ESP up to twelve weeks ahead in most catchments (IQRSS above zero in Fig. 4). Approximately, only four catchments stand out in both seasons with lower skill than ESP (six in spring and one in autumn, not

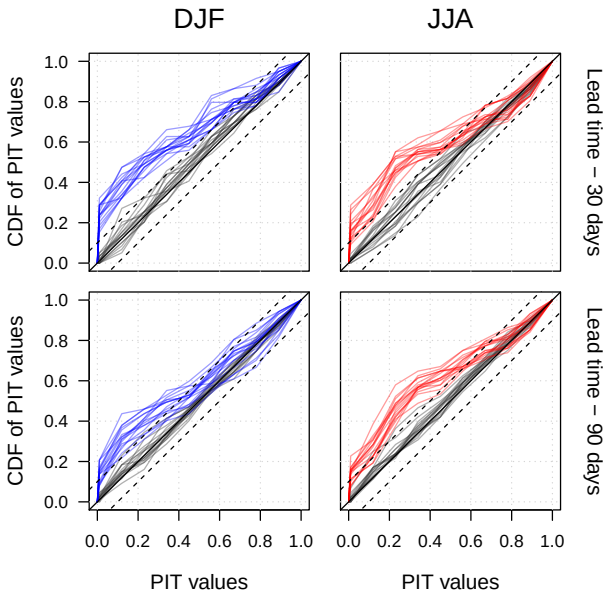

Figure 3. PIT diagram of raw precipitation forecasts (coloured lines) and historical precipitations (grey lines) for lead times of 30 days (top) and 90 days (bottom). Each column corresponds to a target season. Each line represents the PIT diagram in a catchment for forecast horizons within the target season. Dotted lines represent the Kolmogorov significance bands for a 5% significance test.

shown). However, even in these catchments, sharpness can be improved using seasonal precipitation forecasts for lead times up to three weeks in winter (as well as in spring and autumn, not shown). Concerning overall performance (CRPSS in Fig. 4), skill can be observed for lead times up to four weeks in some catchments. At longer lead times, ESP and raw streamflow forecasts are equivalent in most catchments for the winter season. In summer, as well as in spring and autumn (not shown), the difference in skill at longer lead times is more pronounced and most catchments have a negative skill in terms of overall performance.

PIT diagrams are shown for each catchment, for the winter and summer seasons, and for lead times of 30 and 90 days (Fig. 5). In winter and spring (not shown), ESP and raw streamflow forecasts show good reliability, although the curves above the diagonal indicate that forecasts are slightly overpredicting streamflow. Streamflow forecasts for the autumn season (not shown) also show good reliability, but with a tendency to underpredict streamflow. In summer (Fig. 5, right), streamflow forecasts from both ESP and raw forecasts, show problems in forecast reliability. PIT curves clearly indicate a concentration of points at the end points of the diagram and, consequently, narrow ensemble forecasts. In most catchments, 20% to 60% of observed values fall in the lowest interval of the forecast distribution or below it. Although reliability is slightly improved with lead time, streamflow forecasts remain under-dispersive at 90 days of lead time. This could be the result of at least two factors acting alone or jointly: a difficulty of the hydrological model to reach the lowest streamflow values in the simulations of the recession periods, and the influence of not considering uncertainties in the hydrological initial conditions at the time of forecasting.

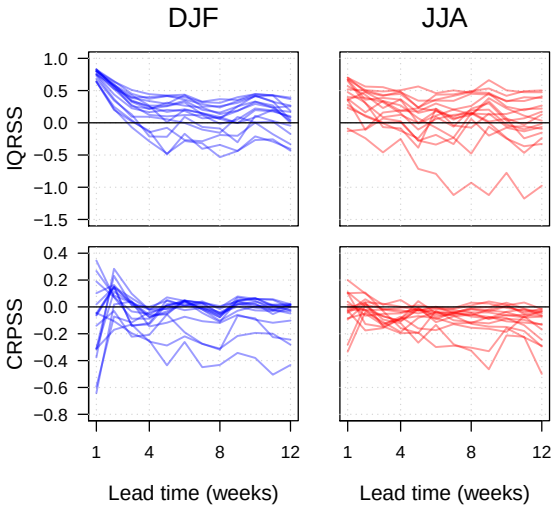

Figure 4. Skill of weekly streamflow forecasts from raw precipitation forecasts as a function of the lead time for all catchments and for the winter (DJF) and summer (JJA) seasons. The skill is computed based on the IQR (top) and the CRPS (bottom) and the reference is Ensemble Streamflow Prediction. Each column corresponds to a target season. Each line represents the skill score in a catchment for forecast horizons within the target season.

4.3 Summary of the quality of raw seasonal forecasts

Skill in the overall performance of System 4 raw precipitation forecasts, at the catchment scale and over a reference forecast based on past observed precipitations, was observed up to two to three weeks in the studied catchments. When looking at streamflow forecasts generated from raw precipitation forecasts, skill over the traditional ESP method was observed up to four weeks, but only in few catchments. The asset of System 4 raw precipitation forecasts and related streamflow forecasts over historical precipitations and ESP, respectively, resides mainly in their sharpness. However, the evaluation of forecast quality shows also that forecasts are often too narrow and suffer from underprediction or overprediction. Improving forecast reliability, while maintaining forecast sharpness is clearly a challenge.

5 Bias correction of seasonal precipitation forecasts

5.1 Overview of the effectiveness of the bias correction methods

Forecast bias, i.e. the ratio between the mean observation and the average forecast ensemble mean, was computed for each catchment over the 1981-2010 period. The bias was computed for each calendar month, but also considering the whole year. Figure 6 shows the biases expressed as deviations from 1 (i.e., $1-Bias$), before and after applying the bias correction methods. It illustrates the results obtained in four catchments at the month-2 lead time (i.e., forecasts issued for day 31 to day 60). The effectiveness of each bias correction method can be observed: unbiased forecasts have a deviation equal to 0 (white); positive

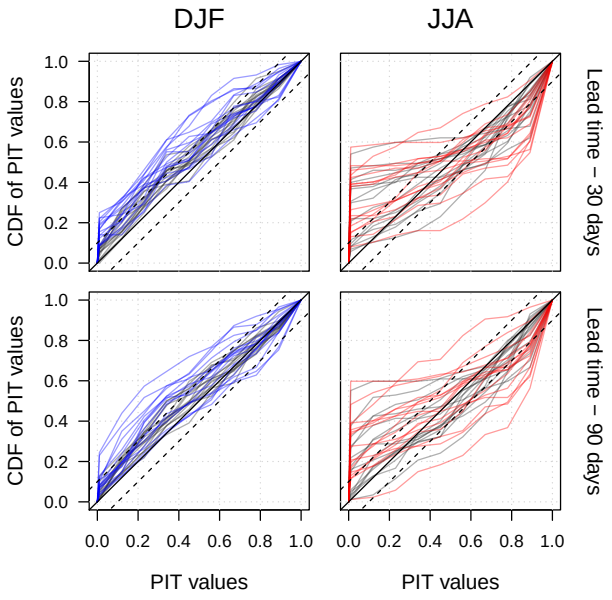

Figure 5. PIT diagram of streamflow forecasts from raw precipitation forecasts (coloured lines) and Ensemble Streamflow Prediction (grey lines) for lead times of 30 days (top) and 90 days (bottom). Each column corresponds to a target season. Each line represents the PIT diagram in a catchment for forecast horizons within the target season. Dotted lines represent the Kolmogorov significance bands for a 5% significance test.

deviations (red) and negative deviations (blue) indicate overprediction and underprediction, respectively. A deviation equal to 0.75 (-3) can be interpreted as the mean forecast being four times larger (smaller) than the mean observation. Overall, when computing the deviations for all monthly lead times of the forecast range, we observed that the biases vary more with the calendar month of the forecast horizon than with lead time. For this reason, we only show the month-2 lead time.

In general, seasonal forecasts tend to overpredict precipitations over the year in most catchments. Overprediction tends to occur near the end of the winter (rainy) season and throughout the spring season. Conversely, precipitations tend to be underpredicted from the end of the summer (dry) season and until the beginning, and sometimes throughout, the autumn season. The four selected catchments illustrate the variety of conditions we encountered in the bias correction analysis. In catchment 2, precipitations could be considered unbiased when carrying the analysis over the year. However, this result hides monthly underpredicting and overpredicting biases which compensate over the year. In this catchment, forecasts tend to overpredict from February to June and underpredict from July to October. The yearly result may also be a reflection of the lack of important biases in the months of December and January, which are, climatologically, the rainiest months. This type of variation in bias was also observed in catchments 6, 11, 12 and 13. In catchment 4, precipitation forecasts are strongly overpredicting observations in all calendar months and thus over the year. This catchment stands out because in no other catchment do we observe a similarly strong and systematic bias. This catchment is the one located at the easternmost part of France. Its main river (l'Ill) is a tributary of the Rhine river. It has its sources in the Jura mountains and receives several tributaries from

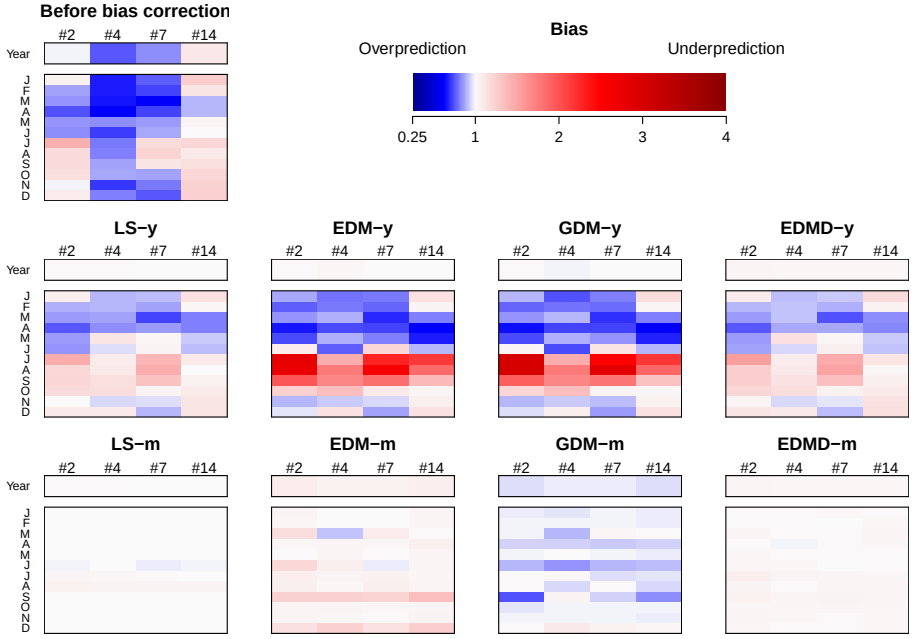

Figure 6. Bias in precipitation for catchments 2, 4, 7 and 14, over the 1981-2010 period. The bias is shown for the whole year (top line) and for each calendar month. The bias is only shown for lead times between 31 and 60 days. Blue-shaded areas represent a tendency of overpredicting precipitations and red-shaded areas represent a tendency of underpredicting precipitations. The top left graph represents the bias of raw precipitation forecasts, and each of the other graphs represents the bias after applying one of the bias correction methods.

the Vosges mountains. In catchment 7, precipitations are overpredicted over the year, with the strongest positive deviations concentrated during the rainy season, basically from November to April. The same behaviour is found in catchments 5, 10 and 15. Interestingly, catchments with a clear overprediction, i.e. catchments following the patterns depicted in Fig. 6 for catchments 4 and 7, correspond to the catchments in which System 4 raw precipitation and streamflow forecasts showed low skill in sharpness and/or overall performance. Lastly, catchment 14 is representative of catchments 1, 3, 8, 9 and 16 in the database. Forecasts slightly underpredict precipitations over the year, with a tendency to underpredict precipitations in all seasons but the spring season, when precipitations are slightly overpredicted.

Figure 6 also presents the remaining biases after the application of the eight bias correction methods. All correction methods are effective to correct biases of precipitation forecasts over the year. However, results for the methods calibrated on a yearly basis (LS-y, EDM-y, GDM-y, EDMD-y) show that the absence of bias over the year is mainly achieved through an effect of compensation between over and underprediction among the calendar months. Particularly EDM-y and GDM-y methods show a strong pattern of monthly biases, even after bias correction, towards overprediction of precipitations in winter and spring, and underprediction in summer and autumn. When looking at monthly biases, monthly calibrated methods perform much better by construction. LS-m and EDMD-m are particularly effective in all catchments. Forecasts corrected with EDM-m tend to slightly underpredict precipitations, while forecasts corrected with GDM-m tend to overpredict precipitations.

5.2 Impact of bias correction on the useful forecasting lead time

Skill scores of bias corrected precipitations and related streamflow forecasts were computed using raw forecasts as reference. For each variable (precipitation and streamflow), each evaluation criterion, each bias correction method, catchment and season, we obtained the corresponding UFL (Useful Forecasting Lead time) and evaluated the proportion of catchments falling in each UFL group (as defined in Section 3.3.3). Results are shown in Fig. 7 and Fig. 8, for precipitation and streamflow forecasts, respectively.

In Fig. 7, the two bias correction methods that stand out regarding overall performance (CRPS), in all seasons, are LS and EDMD. When looking more closely at improvements in the PIT criterion, as measured by the UFL, EDMD clearly stands out from the other methods. The proportion of catchments with skill improvement over raw precipitation forecasts is almost always 100%, and skill is often extended up to 60 days and more. The other methods are quite equivalent to each other, although LS performs slightly better, with greater improvements in larger proportions of catchments, especially in winter and spring, for reliability (PIT), accuracy (MAE) and overall performance (CRPS). In terms of sharpness (IQR), the best performing method varies with the season. Precipitation forecasts in spring are sharper when corrected with methods calibrated monthly, while forecasts in summer and autumn are sharper with methods calibrated yearly.

Figure 8 shows that LS and EDMD methods are able to extend the lead time of bias corrected streamflow forecasts further than other methods, and for a higher proportion of catchments in the large majority of seasons and criteria. Again, EDMD methods yield the best improvements in reliability. LS yields results slightly better than EDMD in sharpness and accuracy. EDM and GDM clearly have lower performance, except in some cases in sharpness and for spring and summer.

5.3 Summary of the comparison of bias correction methods

In general, LS and EDMD bias correction methods show good performance for precipitation and streamflow forecasts, although in a distinct way. While EDMD clearly improves forecast reliability, LS shows better performance in improving sharpness and accuracy. Since streamflow forecasts generated from raw System 4 precipitation forecasts are already, in most of the studied catchments, sharper than the ESP reference, but lack reliability (as shown in Fig. 4 and Fig. 5), it seems appropriate to give priority to a correction method that improves reliability, while providing good overall performance. Therefore, in the following, we will only consider the monthly calibrated version of EDMD (EDMD-m) to further investigate the skill of bias corrected seasonal forecasts. The monthly version is chosen to ensure that monthly biases are removed and that the correction will perform relatively equally in all seasons, while avoiding the "mis-estimation" of forecast skill (Hamill and Juras, 2006).

6 Skill scores of bias corrected seasonal forecasts

6.1 Performance of bias corrected precipitation forecasts

Figure 9 (for sharpness and overall performance) and Fig. 10 (for reliability) present the skill of seasonal precipitation forecasts bias corrected with EDMD-m. Skill scores are computed with historical precipitation as the reference. In order to better evaluate

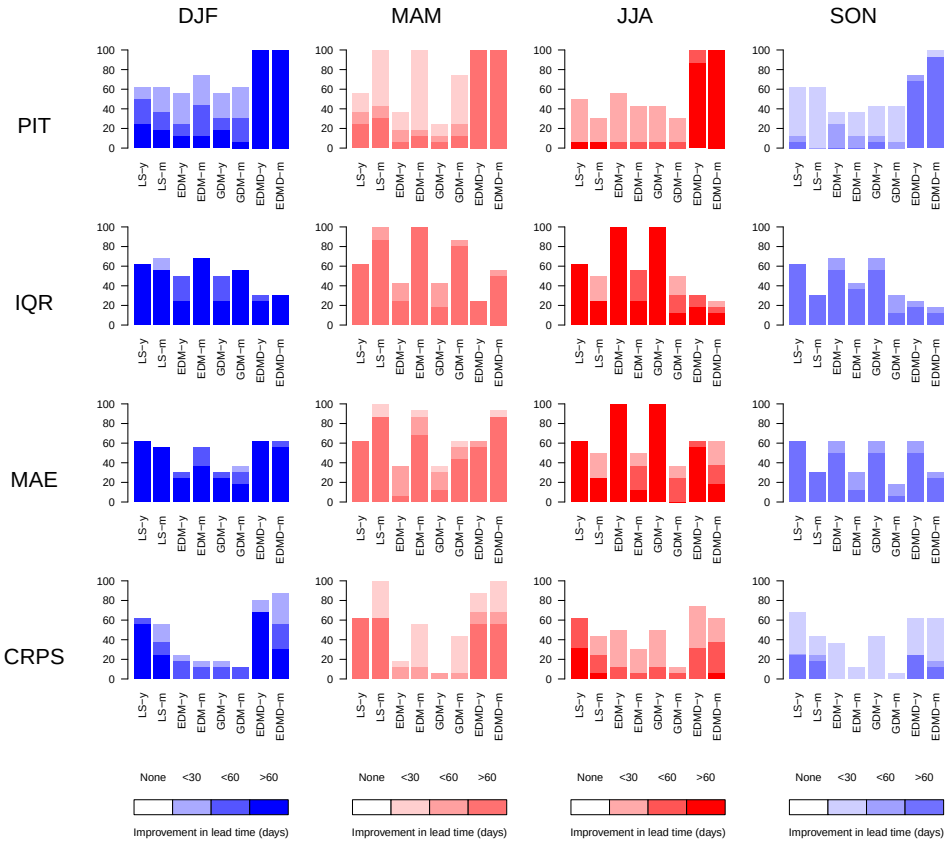

Figure 7. Fraction of catchments (%) in each UFL value category, i.e. fraction of catchments in which bias corrections increase the lead time up to which seasonal precipitation forecasts have skill with respect to raw seasonal precipitation forecasts. Each row corresponds to an evaluation criterion and each column corresponds to a season. Colour shades indicate the UFL category, i.e. the lead time up to which precipitation forecasts are improved.

the impact of bias correction on forecast skill, the y-axes in Fig. 9 are the same as in Fig. 2. The comparison of these two figures shows that bias correcting the raw forecasts reduces the differences in skill between catchments. After bias correction, catchments present very similar evolutions of the skill with the lead time. In some catchments, the values of IQR are lower, but bias corrected forecasts remain sharper than the reference (i.e., skill scores are mostly greater than zero). In the catchments where the raw forecasts performed worse than historical precipitations (i.e., skill scores lower than zero in Fig. 2), bias corrected forecasts become sharper and gain skill. Forecast skill in overall performance (CRPSS) is observed up to two to three weeks ahead. Skill is improved in catchments that performed worse than the reference prior to bias correction (i.e., skill scores lower than zero in Fig. 2). Figure 9 illustrates these findings for winter (DJF) and summer (JJA), but results are similar for spring and autumn (not shown).

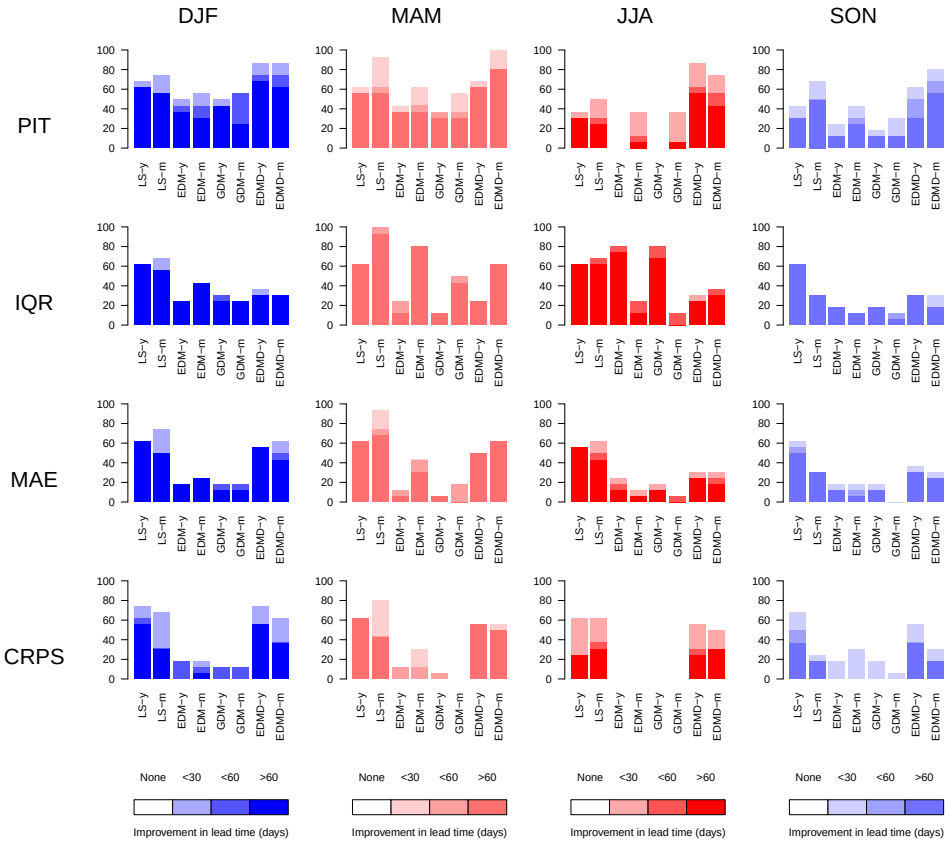

Figure 8. Fraction of catchments (%) in each UFL value category, i.e. fraction of catchments in which bias corrections increase the lead time up to which seasonal streamflow forecasts have skill with respect to seasonal streamflow forecasts generated from raw seasonal precipitation forecasts. Each row corresponds to an evaluation criterion and each column corresponds to a season. Colour shades indicate the UFL category, i.e. the lead time up to which streamflow forecasts are improved.

Figure 10 shows that the most remarkable improvement in performance due to bias correction is achieved in reliability. While precipitation forecasts had a tendency to overpredict prior to bias correction, bias corrected precipitations are reliable in all catchments. Figure 10 shows the results for winter and summer, and for lead times of 30 and 90 days, but conclusions are similar in the other seasons and lead times (not shown). Even though a slight tendency to overpredict precipitation remains in winter for short lead times, the improvements are noticeable. The EDMD-m bias correction was able to address the concentration of points in the zero end point observed in Fig. 3 for the raw forecasts.

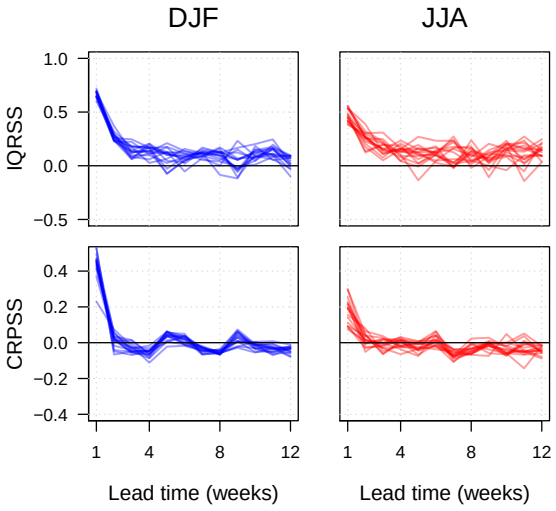

Figure 9. Skill of weekly precipitation forecasts corrected with EDMD-m as a function of the lead time for all catchments and for the winter (DJF) and summer (JJA) seasons. The skill is computed based on the IQR (top) and the CRPS (bottom) and the reference is historical precipitations. Each column corresponds to a target season. Each line represents the skill score in a catchment for forecast horizons within the target season.

6.2 Performance of bias corrected streamflow forecasts

The quality of the streamflow forecasts generated from the precipitation forecasts corrected with EDMD-m is investigated in Fig. 11 and Fig. 12 (IQRSS and CRPSS) and in Fig. 13 (PIT diagrams). These figures can be compared to Fig. 4 and Fig. 5 for raw streamflow forecasts. As seen with precipitation forecasts, bias correction also reduces the differences in streamflow forecast skill between catchments and seasons (Fig. 11). Again, this translates into a loss in skill in catchments with the sharpest ensemble forecasts before bias correction, but also in a gain in skill in catchments where raw streamflow forecasts had negative skill. Overall, after bias correction, streamflow forecasts are sharper than ESP in most catchments and for most lead times. In terms of overall performance (CRPSS), the skill of streamflow forecasts was largely improved, especially in catchments that had very low skill prior to bias correction (i.e., CRPSS values well below zero in Fig. 4). In winter, autumn and spring, skill over the ESP reference is observed up to four weeks ahead in several catchments (even up to five weeks ahead in spring and autumn), while in summer, it is observed up to two to three weeks. At longer lead times, streamflow forecasts show an overall performance equivalent or slightly lower than the performance of the ESP method. Some studies use past streamflow observations (referred to as streamflow climatology) as the reference forecast to assess the skill of streamflow forecasts (e.g. Trambauer et al., 2015; Wetterhall et al., 2015). Figure 12 shows the skill in overall performance and sharpness when streamflow climatology is used as reference. Streamflow forecasts generated from bias corrected precipitation forecasts are sharper and present better overall performance than streamflow climatology, even for lead times of up to twelve weeks in some catchments. This was expected because ensembles based on hydrological modelling benefit from knowledge of initial

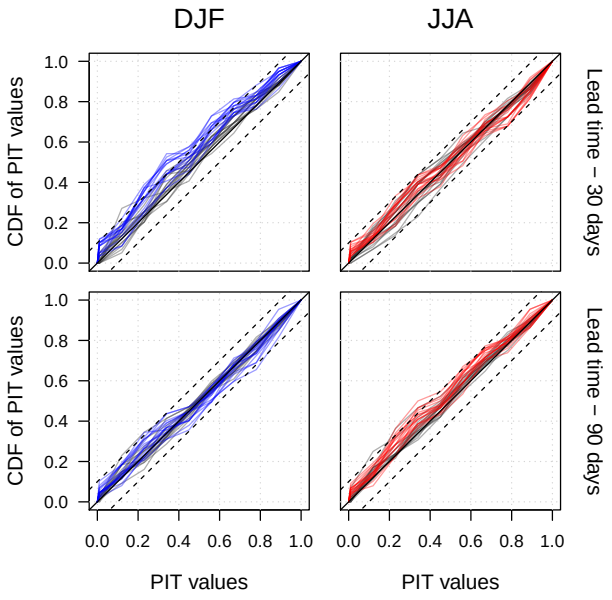

Figure 10. PIT diagram of precipitation forecasts corrected with EDMD-m (coloured lines) and historical precipitations (grey lines) for lead times of 30 days (top) and 90 days (bottom). Each column corresponds to a target season. Each line represents the PIT diagram in a catchment for forecast horizons within the target season. Dotted lines represent the Kolmogorov significance bands for a 5% significance test.

hydrologic conditions. In one catchment (catchment 1), skill scores are systematically higher than the scores of the other catchments. In this catchment, streamflow climatology is very wide, with interannual variability of the same order of magnitude as interseasonal variability.

The PIT diagrams in Fig. 13 show that the reliability of streamflow forecasts is also improved after bias correcting precipitation forecasts. In winter (DJF) and spring (not shown), streamflow forecasts are now reliable and equivalent to ESP, although forecasts still show a slight tendency to overpredict streamflows. In autumn (not shown), streamflow forecasts are also reliable in most catchments, but with a tendency to underpredict streamflows. Summer (JJA) streamflow forecasts are also more reliable after bias correction, but they still depict poor reliability and show that there is room for improvements. As shown by other studies in ensemble forecasting (Zalachori et al., 2012; Verkade et al., 2013; Roulin and Vannitsem, 2015), a simple bias correction of meteorological inputs is obviously not enough to achieve streamflow forecast reliability. In our case, the difficulties of the hydrological model in reaching lower streamflow values remain. This highlights the need for taking into account other sources of hydrological modelling uncertainties and including additional post-processing, targeting directly streamflow forecasts.

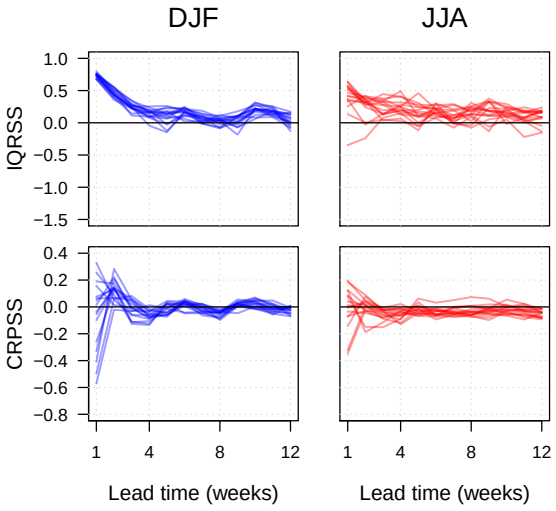

Figure 11. Skill of streamflow forecasts obtained from precipitation forecasts corrected with EDMD-m as a function of the lead time for all catchments and for the winter (DJF) and summer (JJA) seasons. The skill is computed based on the IQR (top) and the CRPS (bottom) and the reference is Ensemble Streamflow Prediction. Each column corresponds to a target season. Each line represents the skill score in a catchment for forecast horizons within the target season.

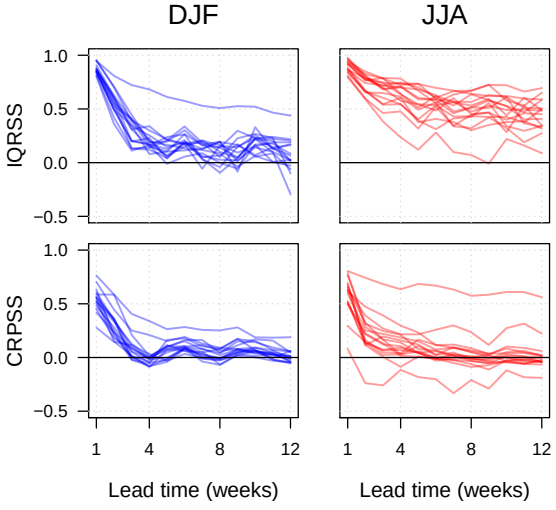

Figure 12. Skill of EDMD-m debiased streamflow forecasts as a function of the lead time for all catchments and for the winter (DJF) and summer (JJA) seasons. The skill is computed based on the IQR (top) and based on the CRPS (right) and the reference is historical streamflow. Each column corresponds to the target season of forecast lead times. Each plotted line represents the performance of a catchment.

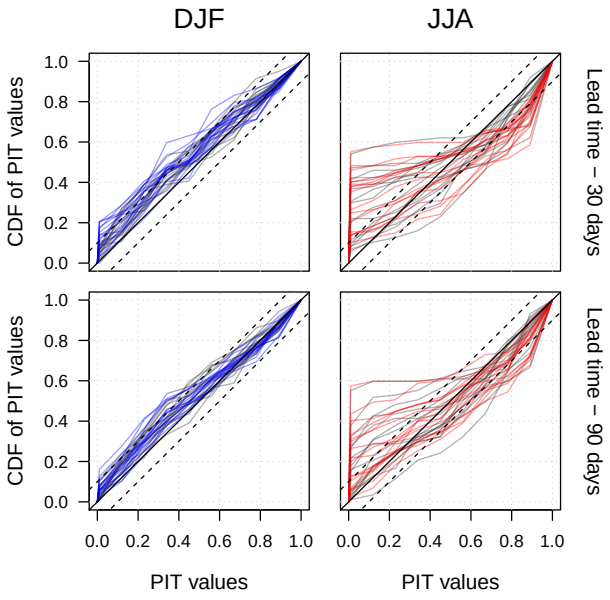

Figure 13. PIT diagram of streamflow forecasts obtained from precipitation forecasts bias corrected with EDMD-m (coloured lines) and Ensemble Streamflow Prediction (grey lines) for lead times of 30 days (top) and 90 days (bottom). Each column corresponds to a target season. Each line represents the PIT diagram in a catchment for forecast horizons within the target season. Dotted lines represent the Kolmogorov significance bands for a 5% significance test.

6.3 How improvements in precipitation forecasts propagate to streamflow forecasts

We have seen that the use of reliable precipitation forecasts as input to a hydrological model does not automatically generate reliable streamflow forecasts. In order to further understand how improvements in precipitation forecasts propagate to streamflow forecasts, we compared the skill scores of EDMD-m bias corrected precipitation forecasts with the skill scores of the streamflow forecasts generated from these bias corrected precipitations. We focused the analysis on the four catchments previously selected as representative of the database, i.e. catchments 2, 4, 7 and 14.

Figure 14 presents the CRPSS, IQRSS and the PITSS (PIT area) in these four catchments, when raw forecasts are used as reference. The reference forecast for the computation of the skill scores of the bias corrected forecasts is the raw forecast. The skill thus represents a measure of the improvement due to bias correction. Skill scores were averaged over lead times of 10 to 90 days.

In overall performance (CRPSS), bias correcting precipitation forecasts either led to a gain in skill in both precipitation and streamflow forecasts, as in catchments 4 and 7 and in some seasons in catchment 2, or to a skill equivalent to the skill prior to bias correction, as in catchment 14. Since catchments 4 and 7 were the ones with the most biased forecasts (cf. Fig. 6), there was more room for improvement in these catchments. Catchment 14 had the smallest bias of the four catchments. Bias correction had thus little impact on precipitation forecasts, and therefore also on streamflow forecasts. Interestingly, the improvement

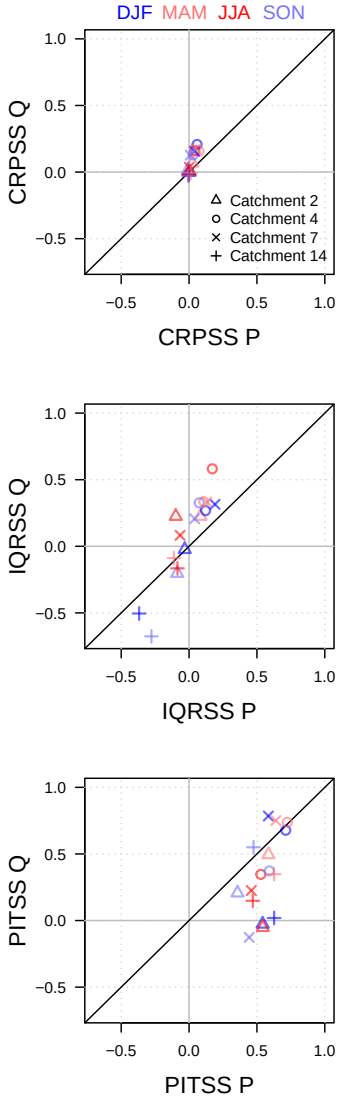

Figure 14. Skill scores of streamflow forecasts after correction with EDMD-m against skill scores of precipitation forecasts after correction with EDMD-m. The skill score of forecasts corrected with EDMD-m is computed using raw forecasts as reference. It is then averaged over lead times 10 to 90 days to obtain a single value. Results are shown for all four seasons in four selected catchments (Catchments 2, 4, 7 and 14). Skill scores were obtained based on the CRPS (top), the IQR (middle) and the PIT diagram area (bottom). The 1:1 diagonal corresponds to an equivalent performance increase in precipitation and streamflow.

achieved in streamflow is always superior to the improvement achieved in precipitation, or equivalent when there was no gain in skill. It seems therefore that a small improvement in the overall performance of precipitation inputs (as measured by the CRPS) can translate in a greater improvement in streamflow forecasts.

If we look at the skill in sharpness (IQRSS) and in reliability (PITSS), we observe different behaviours. In sharpness, a loss in skill was observed in catchments 2 and 14, while a gain was observed in catchments 4 and 7. When a gain was achieved, the gain is superior in streamflow forecasts than in precipitation forecasts. In reliability, skill was always improved by bias correcting the precipitation forecasts, with skill scores always superior to 0.3. The gain in streamflow is mainly positive, but not always, as in the case of precipitation forecasts. Although the majority of skill scores are superior to 0.1, some values are below zero. The gain in reliability from the application of bias correction to precipitation forecasts is, in general, superior in precipitation forecasts than in streamflow forecasts.

Based on our results, we can say that in catchments with small biases, here represented by catchments 2 and 14, overall performance was mainly stable from precipitation to streamflow forecasts. However, in these catchments, a gain in reliability was generally associated with a loss in sharpness. In catchments with greater biases, here represented by catchments 4 and 7, overall performance, sharpness and reliability were improved for both precipitation and streamflow forecasts by simply bias correcting the precipitation forecasts.

6.4 Example of forecast hydrographs in a selected catchment

Figure 15 presents the hydrographs of the forecasts obtained from historical streamflow (HistQ), ESP, and seasonal forecasts bias corrected with LS-m and EDMD-m, from April 2004 to April 2007 in catchment 7. We show forecasts for lead times from 31 days to 60 days. Ensemble forecasts are represented by the median forecasts and two prediction intervals: the 50% interval (between the 25^{th} and 75^{th} percentiles; dark grey zone), and the 90% interval (between the 5^{th} and 95^{th} percentiles; light grey zone). Observed streamflow is also shown. In this catchment, seasonal forecasts had a strong bias and bias correction methods performed well.

The hydrograph for historical streamflow (HistQ plot) represents the interannual variability in the catchment, except that the forecast year is excluded for cross-validation. Visually, the observations fall within the forecast ranges in most cases. The actual coverage of the 90% and 50% prediction intervals is 97% and 66% respectively, which indicates forecast overdispersion and poor sharpness. Accuracy of the median forecast (50^{th} percentile) is, in general, good with a mean absolute error (MAE) of 3.8 m^3/s, although, visually, we observe that too high and low peak flows are not well reproduced.

The forecasts obtained with the ESP method use past observations of precipitation as input to the hydrological model rather than seasonal meteorological forecasts. They show visible improvements in sharpness, notably during low-flow periods. The 90% and 50% prediction intervals actually cover 92% and 60% of the observations, respectively. Accuracy of the median forecasts seems equal or lower than observed with HistQ, which is consistent with an MAE of 4.1 m^3/s.

The hydrographs representing the streamflow forecasts obtained from bias corrected System 4 precipitation forecasts show forecasts that are sometimes even sharper than ESP forecasts, as seen, for instance, for the rising limb in 2005. In some situations, as in the peak event in August 2004, prediction intervals of bias corrected forecasts, particularly in the EDMD-m case, are closer to observations than ESP forecasts. In general, visual differences in quality between seasonal streamflow forecasts obtained from precipitation forecasts corrected with LS-m and EDMD-m are hardly noticeable. For instance, the accuracy of their median forecasts is identical with an MAE of 4.3 m^3/s. However, the 90% and 50% prediction intervals of

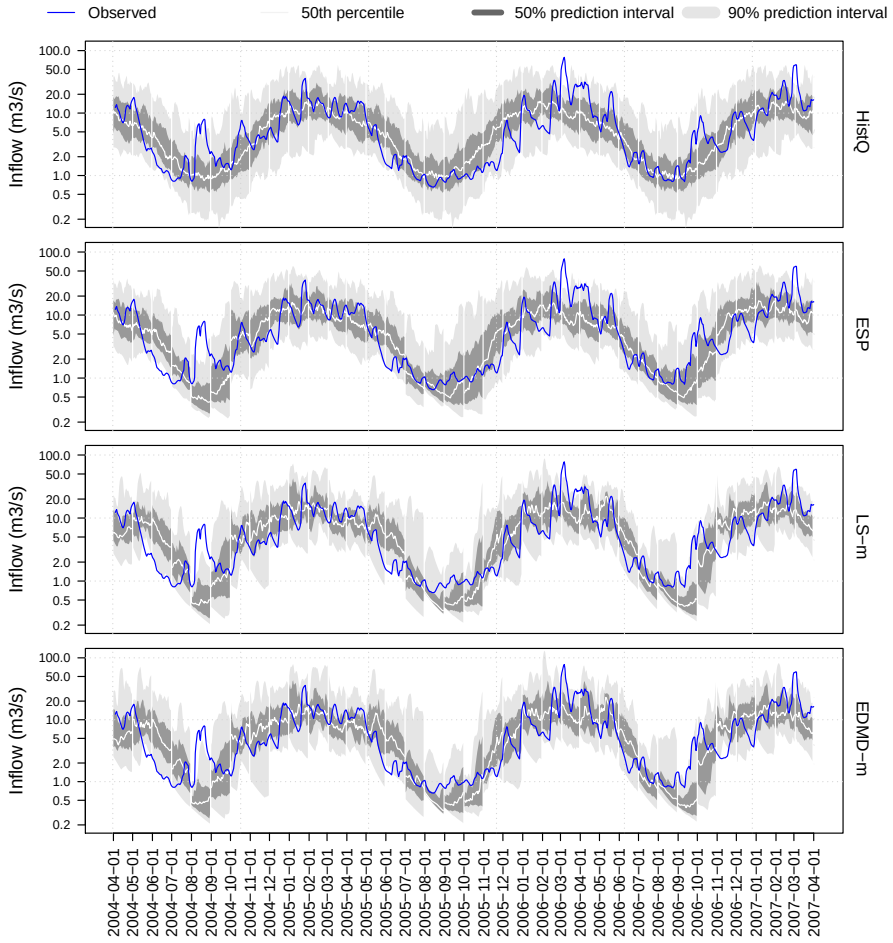

Figure 15. Hydrographs obtained with historical streamflow, ESP, seasonal forecasts corrected with LS-m and seasonal forecasts corrected with EDMD-m in catchment 7 from 1 April 2004 until 1 April 2007. The vertical axis is logarithmic. The blue line represents the observed streamflow. The grey shaded areas present the forecasts issued in the previous month, i.e. 31 to 60 days prior to the observations.

EDMD-m forecasts actually cover, respectively, 89% and 51% of the observations, which indicates better performance in terms of reliability comparatively to LS-m, for which the actual coverage of these prediction intervals is 85% and 46%, respectively.

7 Conclusions

We assessed the quality of ECMWF System 4 precipitation forecasts for seasonal streamflow forecasting in 16 catchments in France. We evaluated areal precipitation forecasts over the catchments and streamflow forecasts generated from inputting precipitation forecasts to a lumped hydrological model. Results show that, in most catchments, raw (uncorrected) System 4 precipitation forecasts are sharper than precipitation climatology (i.e., ensemble forecasts built from past observed precipita-

tions) in all seasons. However, raw precipitation forecasts show poor reliability and a tendency to overpredict precipitations. Likewise, streamflow forecasts generated from raw System 4 precipitations are sharper, but far less reliable than forecasts based on the ESP approach (i.e., ensemble forecasts obtained from running the hydrological model with current initial conditions and past observed precipitations). Yet, in overall performance, raw precipitation forecasts yield improvements up to two weeks in all catchments over precipitation climatology, and streamflow forecasts yield improvements up to three to four weeks over ESP in some catchments. In general, improving forecast reliability, while maintaining (or not diminishing too much) forecast sharpness, was clearly a challenge for bias correction methods.

An in-depth analysis of the biases of System 4 seasonal precipitation forecasts showed strong monthly biases sometimes hidden at the scale of the year, depending on the catchment. Bias correction methods calibrated over the whole year were therefore less efficient when evaluating forecasts over calendar months. In the majority of catchments, the empirical distribution mapping of daily values (EDMD) or the simple linear scaling method (LS) applied to raw precipitation forecasts showed more effectiveness in correcting the yearly but also the monthly biases. These methods also gave the highest increase in overall performance for streamflow forecasting. Empirical distribution mapping of daily values calibrated for each calendar month (EDMD-m) was particularly efficient to increase reliability of precipitation and streamflow forecasts, while linear scaling (LS-m) led to higher improvements in sharpness and accuracy.

The EDMD-m bias correction method was further investigated to better understand its impact on the skill of bias corrected seasonal forecasts. Overall, the application of bias correction reduced the differences in forecast performance between seasons and catchments for precipitation and streamflow forecasts. Also, bias correction ensured that precipitation and streamflow forecasts were at least equivalent in performance to the historical precipitations and the ESP method, respectively, up to three months ahead. In catchments with greater biases, overall performance, sharpness and reliability were improved for both precipitation and streamflow forecasts by simply bias correcting the precipitation forecasts. Overall performance was mainly stable in catchments with small biases. However, in these catchments, a gain in reliability was generally associated with a loss in sharpness. The evaluation of forecasts after bias correction, for the purposes of operational applications on water and risk management, may therefore involve a trade-off between sharpness and reliability. Furthermore, while precipitation forecast reliability is improved with bias correction, the evaluation of streamflow forecast reliability shows that there is still room for improvement. Notably, bias correction of precipitation inputs was not enough to achieve good reliability in summer streamflow forecasts. This highlighted the need for adding a step of streamflow post-processing to the forecasting system.

This study compared eight simple bias correction methods to correct precipitation seasonal forecasts and investigated how they impact the skill of streamflow forecasts. The catchments studied were not influenced by snowmelt flows and thus only precipitation was considered in the bias correction procedures. In other contexts, it may be interesting to also include bias correction of temperature forecasts, with appropriate methods to consider space-time interdependencies of the meteorological variables. The explicit consideration of temperature forecasts could also benefit the skill of low flow forecasts in summer, when evapotranspiration can play a crucial role.

Several other approaches for post-processing and bias correction exist, for instance, based on MOS techniques, space-time disaggregation schemes or Bayesian Model Averaging (Gneiting et al., 2005; Raftery et al., 2005; Liu et al., 2013; Hemri et al.,

2014). These could be investigated to contribute to the comprehensive comparison of options for bias correcting precipitation and temperature forecasts prior to seasonal streamflow forecasting.

Lastly, other forecasting methods selecting historical precipitations based on climate indicators have been investigated in the literature for seasonal hydrological forecasting in regions where strong correlations have been observed, e.g. in the United States or in Australia (Hamlet and Lettenmaier, 1999; Werner et al., 2004; van Dijk et al., 2013). In France, weak correlations have often shown that climate indicators may not be adapted to forecast precipitations at the seasonal scale. However, the use of indicators derived from seasonal forecasts could potentially improve the selection of past precipitation scenarios, which might enhance the skill of ESP methods to forecast streamflow.

Acknowledgements. This work was partly funded by the Interreg IVB NWE programme of the European Union, project DROP (Benefit of governance in DROught adaptation). The first author acknowledges Dr. Christopher A. T. Ferro for his insights on probabilistic scores.

References

Arlot, S. and Celisse, A.: A survey of cross-validation procedures for model selection, Statist. Surv., pp. 40–79, doi:10.1214/09-SS054, http://projecteuclid.org/euclid.ssu/1268143839, 2010.

Christensen, J. H., Boberg, F., Christensen, O. B., and Lucas-Picher, P.: On the need for bias correction of regional climate change projections of temperature and precipitation, Geophysical Research Letters, 35, L20 709, doi:10.1029/2008GL035694, 2008.

Crochemore, L., Ramos, M.-H., Pappenberger, F., van Andel, S. J., and Wood, A. W.: An experiment on risk-based decision-making in water management using monthly probabilistic forecasts, Bulletin of the American Meteorological Society, p. (in press), doi:10.1175/BAMS-D-14-00270.1, 2016.

Day, G.: Extended Streamflow Forecasting Using NWSRFS, Journal of Water Resources Planning and Management, 111, 157–170, 1985.

Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models, Hydrology and Earth System Sciences, 19, 275–291, doi:10.5194/hess-19-275-2015, 2015.

Di Giuseppe, F., Molteni, F., and Tompkins, A. M.: A rainfall calibration methodology for impacts modelling based on spatial mapping, Quarterly Journal of the Royal Meteorological Society, 139, 1389–1401, doi:10.1002/qj.2019, 2013.

Dutra, E., Wetterhall, F., Di Giuseppe, F., Naumann, G., Barbosa, P., Vogt, J., Pozzi, W., and Pappenberger, F.: Global meteorological drought – Part 1: Probabilistic monitoring, Hydrology and Earth System Sciences, 18, 2657–2667, doi:10.5194/hess-18-2657-2014, 2014.

Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., and Liebert, J.: HESS Opinions "Should we apply bias correction to global and regional climate model data?", Hydrology and Earth System Sciences, 16, 3391–3404, doi:10.5194/hess-16-3391-2012, 2012.

Eslamian, S.: Handbook of Engineering Hydrology: Modeling, Climate Change, and Variability, Handbook of Engineering Hydrology, CRC Press, 2015.

Faber, B. A. and Stedinger, J. R.: Reservoir optimization using sampling SDP with ensemble streamflow prediction (ESP) forecasts, Journal of Hydrology, 249, 113–133, doi:10.1016/S0022-1694(01)00419-X, 2001.

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

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Monthly Weather Review, 133, 1098–1118, doi:10.1175/MWR2904.1, 2005.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69, 243–268, doi:10.1111/j.1467-9868.2007.00587.x, 2007.

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., and Engen-Skaugen, T.: Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations - a comparison of methods, Hydrology and Earth System Sciences, 16, 3383–3390, doi:10.5194/hess-16-3383-2012, 2012.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, 2009.

Hamill, T. M. and Juras, J.: Measuring forecast skill: is it real skill or is it the varying climatology?, Quarterly Journal of the Royal Meteorological Society, 132, 2905–2923, doi:10.1256/qj.06.25, 2006.

Hamlet, A. F. and Lettenmaier, D. P.: Columbia River Streamflow Forecasting Based on ENSO and PDO Climate Signals, Journal of Water Resources Planning and Management, 125, 333–341, doi:10.1061/(ASCE)0733-9496(1999)125:6(333), 1999.

Hartmann, H. C., Pagano, T. C., Sorooshian, S., and Bales, R.: Confidence Builders: Evaluating Seasonal Climate Forecasts from User Perspectives, Bulletin of the American Meteorological Society, 83, 683–698, doi:10.1175/1520-0477(2002)083<0683:CBESCF>2.3.CO;2, 2002.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K., and Haiden, T.: Trends in the predictive performance of raw ensemble weather forecasts, Geophysical Research Letters, 41, 9197–9205, doi:10.1002/2014GL062472, 2014GL062472, 2014.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather and Forecasting, 15, 559–570, 2000.

Jenicek, M., Seibert, J., Zappa, M., Staudinger, M., and Jonas, T.: Importance of maximum snow accumulation for summer low flows in humid catchments, Hydrology and Earth System Sciences, 20, 859–874, http://www.hydrol-earth-syst-sci.net/20/859/2016/, 2016.

Kim, H.-M., Webster, P. J., and Curry, J. A.: Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter, Climate Dynamics, 39, 2957–2973, doi:10.1007/s00382-012-1364-6, 2012.

Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrology and Earth System Sciences, 11, 1267–1277, doi:10.5194/hess-11-1267-2007, 2007.

Lemos, M., Finan, T., Fox, R., Nelson, D., and Tucker, J.: The Use of Seasonal Climate Forecasting in Policymaking: Lessons from Northeast Brazil, Climatic Change, 55, 479–507, doi:10.1023/A:1020785826029, 2002.

Liu, Y., Duan, Q., Zhao, L., Ye, A., Tao, Y., Miao, C., Mu, X., and Schaake, J. C.: Evaluating the predictive skill of post-processed NCEP GFS ensemble precipitation forecasts in China's Huai river basin, Hydrological Processes, 27, 57–74, doi:10.1002/hyp.9496, 2013.

Madadgar, S., Moradkhani, H., and Garen, D.: Towards improved post-processing of hydrologic forecast ensembles, Hydrological Processes, 28, 104–122, doi:10.1002/hyp.9562, 2014.

Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T., and Vitart, F.: The new ECMWF seasonal forecast system (System 4), ECMWF Tech. Memo., 656, 49 pp., availableat:http://old.ecmwf.int/publications/library/ecpublications/_pdf/tm/601-700/tm656.pdf, 2011.

Muerth, M. J., Gauvin St-Denis, B., Ricard, S., Velázquez, J. A., Schmid, J., Minville, M., Caya, D., Chaumont, D., Ludwig, R., and Turcotte, R.: On the need for bias correction in regional climate scenarios to assess climate change impacts on river runoff, Hydrology and Earth System Sciences, 17, 1189–1204, doi:10.5194/hess-17-1189-2013, 2013.

Musy, A., Hingray, B., and Picouet, C.: Hydrology: A Science for Engineers, CRC Press, 2015.

Mwangi, E., Wetterhall, F., Dutra, E., Di Giuseppe, F., and Pappenberger, F.: Forecasting droughts in East Africa, Hydrology and Earth System Sciences, 18, 611–620, doi:10.5194/hess-18-611-2014, 2014.

Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., Le Lay, M., Besson, F., Soubeyroux, J. M., Viel, C., Regimbeau, F., Andréassian, V., Maugis, P., Augeard, B., and Morice, E.: Benchmarking hydrological models for low-flow simulation and forecasting on French catchments, Hydrol. Earth Syst. Sci., 18, 2829–2857, doi:10.5194/hessd-10-13979-2013, 2014.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall-runoff model ? Part 2 — Towards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling, Journal of Hydrology, 303, 290–306, 2005.

Pushpalatha, R., Perrin, C., Mathevet, T., and Andreassian, V.: A downward structural sensitivity analysis of hydrological models to improve low-flow simulation, Journal of Hydrology, 411, 66–76, doi:10.1016/j.jhydrol.2011.09.034, 2011.

Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., and Morel, S.: Analysis of Near-Surface Atmospheric Variables: Validation of the SAFRAN Analysis over France, Journal of Applied Meteorology and Climatology, 47, 92–107, doi:10.1175/2007JAMC1636.1, 2008.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Monthly Weather Review, 133, 1155–1174, doi:10.1175/MWR2906.1, 2005.

Rayner, S., Lach, D., and Ingram, H.: Weather Forecasts are for Wimps: Why Water Resource Managers Do Not Use Climate Forecasts, Climatic Change, 69, 197–227, doi:10.1007/s10584-005-3148-z, 2005.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resources Research, 46, W05 521, doi:10.1029/2009WR008328, 2010.

Robertson, D. E., Pokhrel, P., and Wang, Q. J.: Improving statistical forecasts of seasonal streamflows using hydrological model output, Hydrology and Earth System Sciences, 17, 579–593, doi:10.5194/hess-17-579-2013, 2013.

Roulin, E. and Vannitsem, S.: Post-processing of medium-range probabilistic hydrological forecasting: impact of forcing, initial conditions and model errors, Hydrological Processes, 29, 1434–1449, doi:10.1002/hyp.10259, 2015.

Shukla, S., Sheffield, J., Wood, E. F., and Lettenmaier, D. P.: On the sources of global land surface hydrologic predictability, Hydrology and Earth System Sciences, 17, 2781–2796, doi:10.5194/hess-17-2781-2013, 2013.

Teutschbein, C. and Seibert, J.: 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, doi:10.1016/j.jhydrol.2012.05.052, 2012.

Teutschbein, C. and Seibert, J.: Is bias correction of regional climate model (RCM) simulations possible for non-stationary conditions?, Hydrology and Earth System Sciences, 17, 5061–5077, doi:10.5194/hess-17-5061-2013, 2013.

Trambauer, P., Werner, M., Winsemius, H. C., Maskey, S., Dutra, E., and Uhlenbrook, S.: Hydrological drought forecasting and skill assessment for the Limpopo River basin, southern Africa, Hydrology and Earth System Sciences, 19, 1695–1711, doi:10.5194/hess-19-1695-2015, 2015.

van Dijk, A. I. J. M., Peña-Arancibia, J. L., Wood, E. F., Sheffield, J., and Beck, H. E.: Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide, Water Resources Research, 49, 2729–2746, doi:10.1002/wrcr.20251, 2013.

Verkade, J. S., Brown, J. D., Reggiani, P., and Weerts, A. H.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, Journal of Hydrology, 501, 73–91, http://www.sciencedirect.com/science/article/pii/S0022169413005660, 2013.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., and Soubeyroux, J.-M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, International Journal of Climatology, 30, 1627–1644, doi:10.1002/joc.2003, 2010.

Weerts, A. H., Winsemius, H. C., and Verkade, J. S.: Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales), Hydrology and Earth System Sciences, 15, 255–265, doi:10.5194/hess-15-255-2011, 2011.

Weisheimer, A. and Palmer, T. N.: On the reliability of seasonal climate forecasts, Journal of The Royal Society Interface, 11, 20131 162, doi:10.1098/rsif.2013.1162, 2014.

Werner, K., Brandon, D., Clark, M., and Gangopadhyay, S.: Climate index weighting schemes for NWS ESP-based seasonal volume forecasts., Journal of Hydrometeorology, 5, 1076–1090, http://dx.doi.org/10.1175/JHM-381.1, 2004.

Wetterhall, F., Winsemius, H. C., Dutra, E., Werner, M., and Pappenberger, E.: Seasonal predictions of agro-meteorological drought indicators for the Limpopo basin, Hydrology and Earth System Sciences, 19, 2577–2586, doi:10.5194/hess-19-2577-2015, 2015.

Wilhite, D. A., Hayes, M. J., Knutson, C., and Smith, K. H.: Planning for drought: Moving from crisis to risk management, JAWRA Journal of the American Water Resources Association, 36, 697–710, doi:10.1111/j.1752-1688.2000.tb04299.x, 2000.

Winsemius, H. C., Dutra, E., Engelbrecht, F. A., Archer Van Garderen, E., Wetterhall, F., Pappenberger, F., and Werner, M. G. F.: The potential value of seasonal forecasts in a changing climate in southern Africa, Hydrology and Earth System Sciences, 18, 1525–1538, doi:10.5194/hess-18-1525-2014, 2014.

Wood, A. W. and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, Geophysical Research Letters, 35, L14 401, doi:10.1029/2008GL034648, 2008.

Wood, A. W., Kumar, A., and Lettenmaier, D. P.: A retrospective assessment of National Centers for Environmental Prediction climate model–based ensemble hydrologic forecasting in the western United States, Journal of Geophysical Research: Atmospheres, 110, D04 105, doi:10.1029/2004JD004508, 2005.

Yossef, N. C., Winsemius, H., Weerts, A., van Beek, R., and Bierkens, M. F. P.: Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing, Water Resources Research, 49, 4687–4699, doi:10.1002/wrcr.20350, 2013.

Yuan, X., Wood, E. F., and Ma, Z.: A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, Wiley Interdisciplinary Reviews: Water, pp. 523–536, doi:10.1002/wat2.1088, 2015.

Zalachori, I., Ramos, M.-H., Garçon, R., Mathevet, T., and Gailhard, J.: Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies, Advances in Science and Research, 8, 135–141, doi:10.5194/asr-8-135-2012, 2012.

Table 1. Number, name, surface, and mean annual precipitation, potential evapotranspiration and streamflow for the studied catchments.

#	River	Gauging station	Surface (km^2)	Mean annual precipitation (mm/yr)	Mean annual evapotranspiration (mm/yr)	Mean annual flow (mm/yr)
1	Andelle	Vascoeuil	377	952	628	332
2	Orne Saosnoise	Montbizot [Moulin Neuf Cidrerie]	501	735	696	163
3	Briance	Condat-sur-Vienne [Chambon Veyrinas]	605	1100	706	427
4	Ill	Didenheim	668	956	664	309
5	Azergues	Lozanne	798	931	689	296
6	Seiche	Bruz [Carcé]	809	732	696	181
7	Petite Creuse	Fresselines [Puy Rageaud]	853	899	680	316
8	Sèvre Nantaise	Tiffauges [la Moulinette]	872	898	712	331
9	Vire	Saint-Lô [Moulin des Rondelles]	882	958	629	448
10	Orge	Morsang-sur-Orge	934	658	680	131
11	Serein	Chablis	1119	842	675	220
12	Sauldres	Salbris [Valaudran]	1220	803	684	240
13	Eyre	Salle	1678	1025	785	323
14	Arroux	Etang-sur-Arroux [Pont du Tacot]	1792	981	655	390
15	Meuse	Saint-Mihiel	2543	948	639	372
16	Oise	Sempigny	4320	805	639	250

Table 2. Bias corrections applied: corresponding abbreviations, method used for calibration and description.

Abbreviation	Calibration based on	Description
LS-y LS-m	the whole year calendar months	Linear scaling of monthly values
EDM-y EDM-m	the whole year calendar months	Empirical distribution mapping of monthly values
GDM-y GDM-m	the whole year calendar months	Gamma distribution mapping of monthly values
EDMD-y EDMD-m	the whole year calendar months	Empirical distribution mapping of daily values