

Seasonal streamflow forecasting by conditioning climatology with precipitation indices

Louise Crochemore^{1,3}, Maria-Helena Ramos¹, Florian Pappenberger², Charles Perrin¹

¹IRSTEA, Catchment Hydrology Research Group, UR HBAN, Antony, France

5 ²ECMWF, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, UK

³Now at: Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden

Correspondence to: Louise Crochemore (louise.crochemore@smhi.se)

Abstract. Many fields such as drought risk assessment or reservoir management can benefit from long-range streamflow forecasts. Climatology has long been used in long-range streamflow forecasting. Conditioning methods have been proposed to select or weight relevant historical time series from climatology. They are often based on general circulation model (GCM) outputs that are specific to the forecast date due to the initialisation of GCMs on current conditions. This study investigates the impact of conditioning methods on the performance of seasonal streamflow forecasts. Four conditioning statistics based on seasonal forecasts of cumulative precipitation and the Standardized Precipitation Index were used to select relevant traces within historical streamflows and precipitations respectively. This resulted in eight conditioned streamflow forecast scenarios. These scenarios were compared to the climatology of historical streamflows, the Ensemble Streamflow Prediction approach, and the streamflow forecasts obtained from ECMWF System 4 precipitation forecasts. The impact of conditioning was assessed in terms of forecast sharpness (spread), reliability, overall performance and low-flow event detection. Results showed that conditioning past observations on seasonal precipitation indices generally improves forecast sharpness, but may reduce reliability, with respect to climatology. Conversely, conditioned ensembles were more reliable but less sharp than streamflow forecasts derived from System 4 precipitations. Forecast attributes from conditioned and unconditioned ensembles are illustrated for a case of drought risk forecasting: the 2003 drought in France. In the case of low-flow forecasting, conditioning results in ensembles that can better assess weekly deficit volumes and durations over a wider range of lead times.

1 Introduction

1.1 Approaches to seasonal streamflow forecasting

5 Numerical prediction is valuable to proactively manage risks in areas such as hydropower, drinking water production and drought preparedness (Wilhite et al., 2000). Regardless of the application, probabilistic forecasts are preferred over deterministic ones to convey uncertainties (Krzysztofowicz, 2001; Ramos et al., 2013). The main sources of uncertainty in informing decision-making depend on the variable being forecast, the forecast horizon, but also on the location. For instance, region-specific tools have been developed in the world to predict and anticipate drought events weeks, months or even years in advance (Anderson et al., 2000; Ceppi et al., 2014; Hao et al., 2014; Sheffield et al., 2013; Shukla et al., 2014). Nevertheless, anticipating river runoff events at long lead times remains a challenge (Yuan et al., 2015).

10 The predictability of streamflow at long lead times lies in the initial hydrological conditions and the meteorological forcing. Research has shown that the relative role of each source of predictability mainly depends on the “inertia” or “memory” of the studied basin, the forecast season and the forecast lead time (Wood and Lettenmaier, 2008; Shukla et al., 2013; Yossef et al., 2013; Wood et al., 2016). Yossef et al. (2013) showed that in Western Europe, from July to October, streamflow forecasts are more dependent on meteorological forcing than they are on initial conditions, even one month ahead. The conclusions of Shukla et al. (2013) are quite consistent with these findings. They found that the predictability of a forecast issued in July in France lies in the meteorological forcing for horizons longer than three months. However, at shorter lead times, their results show that predictability can be led by either initial conditions or meteorological forcing, depending on the geographical location in France.

In practice, two approaches are often used to forecast streamflow at the seasonal scale (Easey et al., 2006). Statistical approaches rely on past observations and statistical relationships between a predictor and a predictand. For instance, climatology (past observations) is considered a good indicator of the range of possible outcomes for a given time of the year. Dynamical approaches rely on coupled general circulation model (GCM) outputs or past observations to feed a hydrological rainfall-runoff model. For example, Day (1985) introduced the Ensemble Streamflow Prediction (ESP), which uses the climatology of meteorological forcings as input to a hydrological model previously initialised for the forecast date. This approach has been extensively used, for research purposes and operationally in seasonal streamflow forecasting (Wang et al., 2011) and reservoir operations (Faber and Stedinger, 2001), among other fields. An alternative to climatological forcings is the seasonal forecasts issued by GCMs (Yuan et al., 2015). While these are initialized and forced for a specific forecast day, precipitation climatology additionally provides a range of what has been previously observed on that forecast day, regardless of the current atmospheric situation and latest observations. The choice of one approach over the other will depend on the purpose of the forecast, the region of interest and the available data. More importantly, some studies have shown that dynamical and statistical approaches can complement and benefit from each other (Block and Rajagopalan, 2009; Seibert and Trambauer, 2015).

1.2 Selecting ensembles for long-range forecasting

More recently, research has focused on fine-tuning the traditional ESP method by selecting relevant years within the climatology of precipitation. Many studies have proposed to condition or weight past observations based on climate signals. In Northern America, for instance, several studies have taken advantage of the influence of the El Niño Southern Oscillation (ENSO) and the Pacific Decadal Oscillation (PDO) to improve the overall skill of seasonal forecasts. Werner et al. (2004) selected and weighted traces based on the ENSO and showed that some of the proposed methods yielded improvements in forecast overall performance. Gobena and Gan (2010) used the PDO in several resampling strategies, including an approach benefiting from monthly precipitation and temperature statistically derived from climate model outputs. Their study showed that the method yielded moderate improvements to overall forecast skill. At the scale of the globe, van Dijk et al. (2013) selected traces within precipitation climatology based on climate indicators that were proven influential for the region and time period. They showed that using climate information improved forecast skill in Southeast Asia and South America. Bierkens and van Beek (2009) exploited the teleconnection found between winter precipitations and the Northern Atlantic Oscillation (NAO) to select traces within the precipitation climatology and forecast seasonal streamflows in Europe. Their work highlighted the challenges encountered in Europe to using climate indices for seasonal streamflow forecasting. In Europe, teleconnections show complex patterns and a strong seasonal dependence (Ionita et al., 2015). Some studies have thus proposed to condition past precipitation or streamflow scenarios based on previous amounts of precipitation or on previous streamflow anomalies (Sauquet et al., 2008; Svensson, 2016).

In other studies, such as Carpenter and Georgakakos (2001), historical precipitations are conditioned on the precipitation anomaly forecast by a GCM, based on the hypothesis that *“it is not necessary [...] that low skill in reproducing regional precipitation is an index of the utility of GCM information for systems acting as low-pass filters, such as the hydrological and reservoir systems are”*. They found that this conditioning was particularly efficient to forecast the 30-day low inflow volumes to the Folsom lake, and that the GCM-conditioned ensemble outperformed ESP (Yao and Georgakakos, 2001)..

While most studies focus on overall skill, some studies propose to look more closely at specific attributes of the skill, notably forecast sharpness (i.e. the width of forecast members), reliability (i.e. the statistical consistency between observed frequencies and forecast probabilities) and the capacity of ensemble predictions to detect critical events. In Czech Republic, Šípek et Daňhelka (2015) ran a hydrological model with synthetic series of precipitation and temperature generated from climate forecasts and historical meteorological series. The advantage of this modified ESP approach for forecasting was the gain in sharpness, as well as a better capacity to detect high- and low-flow events. Also, Trambauer et al. (2015) recently applied the method proposed by Werner et al. (2004) to forecast drought conditions in Southern Africa. They found that the skill of the conditioned ensemble was lower than that of GCM-based seasonal forecasts but higher than that of ESP forecasts. Some studies have investigated sharpness and reliability simultaneously. For instance, Hamlet and Lettenmaier (1999) selected past precipitations based on categories of ENSO and PDO to feed a hydrological model for streamflow forecasting, and, later on, for reservoir operation (Hamlet et al., 2002). They noted that the conditioning improved forecast sharpness.

However, in six months of the year, climatology was more reliable than the conditioned ensembles in terms of observed streamflow falling within the forecast range. Yao and Georgakakos (2001) compared the method proposed by Carpenter and Georgakakos (2001) with the ESP approach and with a conditioned forecast ensemble based on historical streamflows and the latest observed reservoir inflows. An in-depth evaluation of the latter showed a gain in sharpness but also a loss in reliability as compared to historical streamflow data. Nevertheless, decisions based on the conditioned ensemble were able to eliminate flood damage and generate more energy than decisions based on the other two ensemble approaches.

1.3 Scope of the study

This study investigates the impact of conditioning methods on the performance of seasonal streamflow forecasts. It proposes an insight into how conditioning approaches impact forecast attributes such as reliability, sharpness and the detection of low-flow events. The aim is not to provide an overall better ensemble but to shed light on which forecast attributes can be expected to improve or deteriorate after conditioning. For that purpose, we used conditioning statistics based on precipitation indices derived from the System 4 seasonal precipitation forecasts issued by the European Centre for Medium-range Weather Forecasts (ECMWF) to select traces of past observed precipitation and streamflow. Eight streamflow forecast scenarios were built and analysed.

Section 2 presents the data and the methodology used to build streamflow forecasts. In Section 3, we analyse the impact of the conditioning on the overall performance, sharpness and reliability of seasonal streamflow forecasts over the whole year. Then, we investigate the ability of the ensemble prediction systems to forecast low-flow events. We also illustrate the differences in forecast attributes with a drought risk assessment graph for the case of the 2003 severe drought in France. In Section 4, we discuss the main outcomes and perspectives of the study.

20 2 Data and methods

2.1 Observed and forecast hydrometeorological data

Observed precipitation data come from the SAFRAN reanalysis of Météo-France (Quintana-Seguí et al., 2008; Vidal et al., 2010). Daily values are available from August 1958 until July 2010 (i.e. 51 complete years) at an 8x8 km grid resolution covering France. Data were aggregated at the catchment scale. Mean areal potential evapotranspiration was computed for each catchment using a temperature-based formula (Oudin et al., 2005) and observed temperatures from the SAFRAN reanalysis. Daily streamflow data at the outlet of each catchment come from the French national archive (Banque Hydro, www.hydro.eaufrance.fr).

Seasonal precipitation forecasts were collected from ECMWF GCM, System 4. Once a month, ECMWF provides a 51-member forecast ensemble for the next seven months at a T_{L255} ($\sim 0.7^\circ$) spatial resolution (Molteni et al., 2011). ECMWF issued hindcasts for the 1st of each month from 1981 to 2010. These hindcasts are composed of 51 members when issued in

February, May, August and November, and 15 members in other months. For the purpose of this study, System 4 forecasts were aggregated at the catchment scale with a weighted mean based on the catchment area covered by each forecast grid cell (two to ten grid cells per catchment). Only forecasts with lead times up to 90-days were considered. In a previous study, several bias corrections were applied to System 4 precipitation forecasts and compared based on their impacts on seasonal streamflow forecasts (Crochemore et al., 2016). The study showed that the empirical distribution mapping of daily values improved the reliability of both precipitation and streamflow forecasts. Following these results, System 4 precipitation forecasts used here were previously bias corrected with the empirical distribution mapping of daily values.

2.2 Catchments and hydrological model

Sixteen catchments spread over France were selected from the database used by Nicolle et al. (2014). Using a set of catchments helps getting more general conclusions (see e.g. Andréassian et al., 2009; Gupta et al., 2014). However, it should be noted that identifying relations between performances and catchment characteristics is outside the scope of this study. These catchments are dominated by a pluvial regime and the quality of their streamflow data during low flows is good. Additionally, the selected catchments have an average solid fraction of precipitation below 10%. Their location, hydrological regimes and main characteristics are presented in Figure 1 and Table 1, respectively. In these catchments, low flows are observed between May and October, i.e. from late spring to early autumn. Major drought events include the droughts of 1976, 1989, 2003 and 2005. Among these, the 2003 drought was estimated to have caused 15,000 deaths and cost over a billion euros just in France (UNEP, 2004; Poumadère et al., 2005). Here, this particular event is used to illustrate the impact of conditioning methods on drought risk assessment.

The hydrological model used in this study is the GR6J model, a daily conceptual model with six free parameters specifically proposed for low-flow simulation by Pushpalatha et al. (2011). 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. Its inputs are daily precipitation and potential evapotranspiration at the catchment scale, and its output is the streamflow at the catchment outlet. In this study, the mean interannual potential evapotranspiration was used as input to the GR6J model, regardless of the forecast year, i.e. for a given day of the year, the estimated potential evapotranspiration on this day is assumed to be the mean of all potential evapotranspiration computed for this day of the year, from 1958 to 2010. Regardless of the precipitation scenario fed to the model, the same interannual potential evapotranspiration scenario is used as input to the model. This allows us to focus solely on the influence of precipitation inputs on streamflow forecasts. In addition, when the model is applied to forecast streamflows, the last observed streamflow at the time of forecast is used to update the levels of the routing reservoirs before issuing the forecasts.

The GR6J model was calibrated in each catchment with the one-year-leave-out method (Arlot and Celisse, 2010) and with the Kling-Gupta Efficiency (Gupta et al., 2009) applied to inverse flows to focus on the lowest flows (Pushpalatha et al., 2012). We obtained an average KGE of 0.78 in calibration (ranging from 0.46 to 0.94) and 0.76 in validation (ranging from

0.41 to 0.94) over the sixteen catchments. An average KGE applied to root-squared flows of 0.86 was obtained in validation (ranging from 0.54 to 0.94), showing that the model also performs well for median to high flows. The distance of the bias from 1 (i.e. 1-bias, with bias defined as the ratio between observed and simulated streamflows) is moderate in simulation, with values ranging from -0.1 to 0.1 in all catchments but three. In these three catchments, values of 0.12, -0.14 and -0.94 are
5 obtained.

2.3 Forecast scenario building method

Eight ensemble forecast scenarios were built to investigate the impact of conditioning on forecast performance. The eight scenarios are based on four conditioning statistics and three methods that are commonly used in seasonal streamflow forecasting. These three methods (named “base ensembles” in the following) and the conditioned scenarios are introduced
10 below. Table 2 summarizes the different ensemble forecast scenarios analysed in this study.

2.3.1 Description of the base ensembles

The simplest ensemble forecast scenario uses the long-term statistical variability of historical streamflows. It is assumed that the streamflow at a given day of the year is likely to fall within the range of streamflows observed in other years, on that same day. This is a “poorman’s approach” that can serve as a naïve benchmark, where no hydrological model but only a long
15 streamflow time series of records is available. It is named **HistQ** hereafter.

Another base ensemble is the traditional **ESP** method. It requires a hydrological model and a long time series of precipitation records. This ensemble is based on the assumption that the precipitation of a given day is likely to fall within the range of past precipitations observed on that same day in previous years. For a given forecast day, a precipitation ensemble is thus
20 built by using precipitations observed in other years. The precipitation ensemble has as many members as the number of years different from the forecast year available in the precipitation record. The states of the GR6J hydrological model are first initialized with a one year run up to the forecast date. The precipitation ensemble and interannual potential evapotranspiration are then used as input to the model.

The third base ensemble is similar to ESP but uses the bias corrected ECMWF System 4 seasonal precipitation forecasts as input to the GR6J hydrological model. Both the System 4 GCM and the hydrological model are initialized for the forecast
25 day. This ensemble can be considered the most costly in terms of implementation and computational needs. Hereafter, this ensemble is named **Sys4**.

2.3.2 Description of the conditioned scenarios

From the base ensembles, we built eight scenarios by selecting traces within the HistQ and ESP ensembles. The conditioning was based on four statistics derived at each forecast date and from each ensemble member of the System 4 precipitation
30 forecasts. Two of these statistics are based on cumulative rainfalls, and two on the standardized precipitation index (SPI).

The SPI transforms the distribution fitted to a long precipitation record into a normal distribution (McKee et al., 1993; WMO, 2012). An SPI value of 0 corresponds to conditions close to the long-term average of precipitations. Negative (positive) SPI values correspond to drier (wetter) conditions. The four conditioning statistics are:

- the cumulative precipitation forecast over the first three months of lead time altogether (**Sum3**);
 - 5 - the series of cumulative precipitation forecast over the first, second and third months separately (i.e. one value per lead time, **Sum1**, decomposed into Sum1-1, Sum1-2 and Sum1-3, depending on the lead month);
 - the SPI over the first three months altogether (**SPI3**);
 - the SPI over the first, second and third months separately (i.e. one value per lead time, **SPI1**, decomposed into SPI1-1, SPI1-2 and SPI1-3, depending on the lead month).
- 10 The statistics (SPI or Sum for the precipitation volumes) derived from System 4 forecasts are then used to select traces within HistQ and ESP. For that purpose, statistics are also computed for sequences of historical precipitations. For a given forecast member, the sequence in the historical precipitations that is the closest, in terms of Euclidian distance, to this member with respect to the considered statistics is selected. When searching for the closest historical sequence, we only consider sequences that start within a 31-day window centred on the forecast date and in years different from the forecast
- 15 year. Note that the same “closest” historical sequence can be associated to several forecast members. This procedure leads to a conditioned ensemble with the same size as the System4 forecast.

Once the historical sequences are selected, two cases can then lead to a streamflow forecast ensemble: (a) the selected precipitation sequences are used as input to the hydrological model to generate a streamflow forecast ensemble (this is the case for: **ESP_Sum3**, **ESP_Sum1**, **ESP_SPI3**, **ESP_SPI1**), or (b) the historical streamflows corresponding to the selected

20 precipitation sequences are directly used as ensemble members to build a streamflow forecast ensemble (this is the case for: **HistQ_Sum3**, **HistQ_Sum1**, **HistQ_SPI3**, **HistQ_SPI1**). In the latter case, the streamflow sequences obtained may result in unrealistic forecast scenarios due to an initial hydrologic condition on the forecast date that is far from what was historically observed for a selected sequence. Therefore, when directly selecting scenarios from past streamflow observations, we have also added the streamflow observed on the day of issuing the forecast as a conditioning criterion in the computation of the

25 Euclidian distance.

2.4 Forecast verification methods

Many criteria exist to assess the performance of probabilistic forecasts. Here, we assessed the overall performance of the forecasts, their capacity of discrimination, their sharpness and reliability. For these last two attributes, we consider the paradigm that better forecasts are those that maximize sharpness while guaranteeing reliability (Gneiting et al., 2007).

2.4.1 Evaluation of forecast attributes

The overall performance of the forecast systems was evaluated using the Continuous Rank Probability Score (CRPS, Hersbach, 2000). The CRPS averages the area between the cumulative forecast distribution and the step function corresponding to the observation over the evaluation period.

5 Sharpness is an intrinsic attribute of the forecast ensemble. It indicates how spread the members of an ensemble forecast are. Here, sharpness was computed as the average difference between the 95th and the 5th percentiles of the forecast distribution over the evaluation period (Gneiting et al., 2007). It thus corresponds to the average 90% interquantile range (IQR).

Reliability refers to the statistical consistency between observed frequencies and forecast probabilities. Reliability was evaluated with the Probability Integral Transform diagram (PIT, Gneiting et al., 2007; Laio and Tamea, 2007). The PIT

10 diagram represents the cumulative distribution of the positions of the observation within the distribution of forecast values. The PIT diagram of a perfectly reliable forecast is superposed with the 1:1 diagonal, meaning that the observation uniformly falls within the forecast distribution. To numerically compare results for large datasets, Renard et al. (2010) proposed to compute the area between the PIT diagram and the 1:1 diagonal. The smaller the PIT area, the more reliable the ensemble.

Note that the CRPS is sensitive to both the reliability and the sharpness of the forecasts. Each attribute influences two
15 independent terms of the decomposition of the CRPS. A decrease in one can thus be compensated by an increase in the other, which would remain unnoticed in the CRPS value.

The discrimination of a system is its capacity to detect an event defined by a threshold. The Relative Operating Characteristics diagram (ROC, Mason and Graham, 1999) is used to assess the discrimination of the forecasting systems. In

20 this study, the threshold used to define events is the 80th exceedance percentile of observed streamflow (i.e. 80% of the observed values are above this threshold). To build the diagram, the proportion of ensemble members below the threshold necessary to trigger an alert varies from none to all ensemble members. For each of these proportions, the probability of detection is plotted against the false alarm ratio. The ROC diagram is plotted for a given threshold, catchment and forecast lead time. The Area Under the Curve (AUC) summarizes the ROC diagram into one numerical value and allows for an easier comparison of forecast systems. The closer the AUC is to 1, the better the system is at discriminating between events (i.e.
25 threshold exceedances) and non-events.

2.4.2 Skill scores

The skill of forecast systems is computed as follows for a given lead time i :

$$Skill\ score_i = \frac{Score_i^{ref} - Score_i^{syst}}{Score_i^{ref} + Score_i^{syst}} \quad (1)$$

This normalized skill ranges within [-1,1]. A skill superior to 0 (inferior to 0) indicates that the forecast system performs
30 better (worse) than the reference. Here, we evaluated the conditioned forecast scenarios against the base ensembles they were based on (i.e. Sys4, ESP or HistQ). The skill score was computed based on the CRPS, the IQR and the PIT area. These

scores are abbreviated CRPSS, IQRSS and PITSS. Since we compared ensembles with different ensemble sizes (see Table 2), which is known to induce bias when computing skill scores, the correction proposed by Ferro et al. (2008) was applied to remove such bias in the computation of the CRPSS.

3 Analysis of the quality of the streamflow forecasting systems

5 3.1 Skill of System 4 in forecasting conditioning statistics

We first evaluated the skill of System 4 in forecasting the conditioning statistics (cumulative precipitations Sum and SPI). Figure 2 shows their skill in overall performance (CRPSS) and in sharpness (IQRSS), and Figure 3 shows their reliability (PIT diagram). The reference forecast used to compute the skill scores is historical precipitations (i.e. climatology). Regardless of the considered statistic, System 4 performs as well as climatology, while being sharper (Figure 2). In addition, SPI forecasts issued from System 4 are reliable overall and in standard precipitation conditions (Figure 3). In dry conditions (i.e. SPI values smaller than -1), however, forecasts tend to overestimate SPI values, while in wet conditions (i.e. SPI values greater than 1) forecasts tend to underestimate SPI values. Similar PIT diagrams are observed with SPI forecasts from historical precipitations (not shown). Dutra et al. (2014) did a similar comparison and showed that SPI forecasts from System 4 always had skill as compared to historical precipitations, with respect to discrimination, accuracy and anomaly correlation, in South Africa.

3.2 Statistical evaluation of overall performance, sharpness and reliability

3.2.1 Forecast attributes of the conditioned scenarios with respect to HistQ and ESP base ensembles

First, we evaluated the gain and loss in skill of daily streamflow forecasts due to the four types of conditioning applied to the HistQ base ensemble. Figure 4 shows the CRPSS, IQRSS and PITSS for lead times up to 90 days, and the PIT diagram for a lead time of 45 days. The reference for the computation of the skill is HistQ, i.e. historical streamflows with all available years. Each line corresponds to one of the 16 catchments.

The first conclusion from this figure is that all four conditionings lead to similar results. Their impact on forecasts reliability (PIT) and sharpness (IQR) is uniform over the lead times, while their impact on overall performance (CRPS) is greater at shorter lead times. Conditioning HistQ improves sharpness at most lead times (IQRSS above zero) and for all conditioning statistics (Sum or SPI). However, as a result of narrower ensembles, there is a decrease in the PIT values (reliability) at most lead times (PITSS below zero). Nevertheless, the PIT diagrams at 45 days show that the conditioned ensembles remain quite reliable as a whole (PIT values close to the diagonal line), especially when the conditioning is based on SPI statistics. Regarding overall performance (CRPS), the conditioning increases performance up to 5 to 15 ahead in most catchments, and up to 30 days in some catchments. Improvement is greater when traces are selected based on cumulative precipitations

(Sum3 or Sum1) or SPI3 than when they are selected based on the series of SPI1 values. The improvement in overall performance in the first lead times can be attributed to the fact that the conditioning of historical streamflow takes also into account the last observed streamflow to better match current initial conditions (cf. Section 2.3.2). At longer lead times, the overall performance of conditioned scenarios is, in the majority of catchments, equivalent or worse than that of HistQ. In one of the catchments, however, we observed improvements up to 90 days ahead. This catchment corresponds to catchment 1, in which interannual streamflow variability dominates over seasonality (cf. Section 2.2) due to a high base flow index.

We also examined the loss and gain in skill due to conditioning applied to the ESP base ensemble (Figure 5). This time, the reference used in the computation of the skill is ESP. Here again, the four conditionings seem to have a similar impact on performance. Conditioned streamflow forecasts appear to be equivalent or worse than ESP in terms of overall performance (CRPSS). When conditioning ESP with SPI, this often translates in a gain in sharpness (IQRSS) and a loss in reliability (PITSS). When conditioning with the Sum statistics, forecasts lose sharpness in some catchments and reliability in most catchments. Results are also less homogeneous between catchments. The loss in overall performance is also greater with the conditionings based on cumulative precipitations. The PIT diagrams show that not all ensembles are perfectly reliable, with observations too often falling below the forecast range in most catchments. This tendency may be caused by the precipitation inputs but also by the hydrological model, since in ESP-based approaches it also plays a role.

In summary, Figures 4 and 5 show that, in general, conditioning has the same impact on forecast attributes regardless of the conditioning statistics. It tends to increase sharpness and decrease reliability. However, conditioning with cumulative precipitations can also decrease both attributes, sharpness and reliability, which is not satisfying. In addition, conditioning based on SPI provides more consistent results between catchments and minimizes the losses in reliability and overall performance, comparatively to base ensembles. In the following paragraphs, only HistQ_SPI3 and ESP_SPI3 were retained in order to further explore the quality of conditioned ensembles.

3.2.2 Forecast attributes of the conditioned scenarios with respect to Sys4 base ensemble

In a previous study (Crochemore et al., 2016), we assessed the performance of Sys4 precipitation forecasts for seasonal streamflow forecasting in the studied catchments. We observed a good overall performance of the streamflow forecasts after bias correction, but also a general lack of reliability during summer (June-July-August). In Figure 6, we evaluate the quality of the conditioned scenarios ESP_SPI3 and HistQ_SPI3 with respect to Sys4, from which the conditioning statistics are derived.

ESP_SPI3 conditioned ensembles show better overall performance than Sys4 for lead times shorter than 5 days, worse performance for lead times from 5 to 15 days, and equivalent performance at longer lead times. In terms of reliability and sharpness, ESP_SPI3 is overall more reliable but less sharp than Sys4 for lead times shorter than 45 days. At longer lead times, ESP_SPI3 becomes equivalent to Sys4. If we now look at HistQ_SPI3, it has lower overall performance than Sys4,

especially for lead times shorter than 15 days. HistQ_SPI3 provides forecasts that are more reliable than Sys4, except at long lead times (more than 45 days) in some catchments.

In summary, Figure 6 illustrates how conditioning the base ensembles from the ESP method or from historical streamflows on SPI3 statistics derived from GCM-based seasonal forecasts can be beneficial for several catchments at lead times longer than 15 to 30 days since it allows the conditioned ensembles to be at least as sharp as the GCM-based forecasts while being also, in most cases, more reliable than or as reliable as GCM-based seasonal forecasts.

3.2.3 Overall influence of conditioning on streamflow forecasts reliability and sharpness

Figure 7 proposes a simultaneous evaluation of the reliability (PIT area) and sharpness (IQR) of ESP_SPI3 and HistQ_SPI3. For a given catchment, lead time and reference, the skill in reliability is plotted against the skill in sharpness. Each point corresponds to a catchment, each column corresponds to a lead time and each row corresponds to a forecast ensemble. Two references are chosen for each ensemble: ESP_SPI3 is evaluated against ESP and Sys4, and HistQ_SPI3, against HistQ and Sys4. Each reference is identified by its colour and shape (cf. legend). If a point is located in the upper left part of the graph, the conditioned ensemble is more reliable but less sharp than the reference (indicated by the colour of the point) in the corresponding catchment. Conversely, if a point is located in the lower right part, the conditioned ensemble is sharper but less reliable than the reference. At best, both reliability and sharpness are improved, and points are located in the upper right part of the graph. At worst, both reliability and sharpness are deteriorated with respect to the reference, and points are located in the bottom left part of the graph.

Overall, the conditioning tends to have more impact on reliability than on sharpness (y-axes extend further than x-axes). The main conclusion from this graph is that conditioned ensembles are generally more reliable but less sharp than Sys4, and they are sharper but less reliable than ESP or HistQ. More specifically, we observe that:

- For a lead time of 10 days, ESP_SPI3 and HistQ_SPI3 can be more reliable and sharper than the ensembles they are selected from. This applies to nine catchments with ESP_SPI3, and three catchments with HistQ_SPI3;
- For a lead time of 30 days, fewer catchments benefit from a gain in both reliability and sharpness. The loss in sharpness and the gain in reliability with respect to Sys4 are less pronounced than for a lead time of 10 days. For instance, the maximum PITSS values for ESP_SPI3 move from 0.45 (for a lead time of 10 days) to 0.2 (for a lead time of 30 days) and those for HistQ_SPI3 move from 0.7 to 0.4. The gain in sharpness and the loss in reliability with respect to ESP and HistQ remain in the same ranges as observed for a lead time of 10 days;
- For a lead time of 90 days, the gain of ESP_SPI3 over Sys4 is further reduced and varies with the catchment. The same is observed to a lesser extent for HistQ_SPI3, even though a positive impact of the conditioning on reliability can still be observed in several catchments. At this lead time, both ESP_SPI3 and HistQ_SPI3 provide forecasts that are still sharper, yet less reliable, than the climatology they are selected from.

Figure 7 can also be interpreted in terms of similarities in forecast attributes between approaches. Indeed, the (0,0) point corresponds to the location of the references used for the skill scores. From this perspective, ESP_SPI3 is closer to ESP than to Sys4 for a lead time of 10 days. However, as the lead time increases, ESP_SPI3 becomes closer to Sys4 and further apart from ESP. The proximity between ESP_SPI3 and Sys4 at longer lead times can be attributed to the conditioning itself. The proximity between ESP_SPI3 and ESP and their distance to Sys4 at shorter lead times may be attributed to the initialization of the climate model. Since initial hydrological conditions are the same for the three forecast ensembles, differences are caused by the meteorological forcing only. The main difference between System 4 precipitation and climatology at such lead times is the initialization of the GCM, which leads to sharper System 4 forecasts for the first lead times. Similarly, we observe that HistQ_SPI3 becomes closer to Sys4 as the lead time increases due to conditioning. However, its distance to HistQ remains the same at all lead times. This distance is probably due to the use of previous streamflow conditions as a conditioning criterion within HistQ.

As a summary guideline, Table 3 ranks the different ensembles investigated based on the analyses of overall performance, reliability and sharpness, and for different lead time ranges: from 10 to 30 days, from 30 to 60 days and from 60 to 90 days. The rankings are based on averaged skill score values. Two ensembles are considered equivalent (and thus receive the same rank) if the difference in the averaged skill scores is smaller than 0.01. This table serves as a guideline. For instance, in operational conditions, a practitioner could choose to emphasize one of the three attributes of forecast quality over the others, and could choose the forecasting approach to be implemented based on this table. From Table 3, we can say that, if one seeks an overall performing ensemble with 10 to 30 days lead, one would use Sys4. For horizons longer than 30 days, ESP and ESP_SPI3 offer good alternatives. If one seeks, above all, a reliable ensemble, one could simply use HistQ, ESP, or even HistQ_SPI3 for lead times shorter than 30 days. For ensembles that are both good in terms of reliability and sharpness, and for horizons longer than 30 days, the HistQ_SPI3 ensembles seem to offer the best trade-off.

3.3 Statistical evaluation of low flows

We now investigate the impact of conditioning on the performance of the ensemble forecast scenarios to forecast summer low flows and drought risks. Many ways of characterizing severe low flows and droughts exist in the literature (Mishra and Singh, 2010; Smakhtin, 2001; Tallaksen et al., 1997; WMO, 2008). In the following, the low-flow variables considered are the low-flow duration and deficit volume, both computed for the 80th exceedance percentile. In this section, only forecast horizons falling within the May to October period are considered.

3.3.1 Impact of the conditioning on forecast discrimination

The capacity of the different systems to discriminate between low-flow events and non-events is assessed. Figure 8 presents the ranges of the Area Under the Curve (AUC) of the ROC diagram obtained from five ensemble forecast scenarios, namely Sys4, ESP_SPI3, ESP, HistQ_SPI3 and HistQ. AUC values were assessed for the 80th exceedance percentile and for lead

times of 10 days, 30 days and 90 days. Each boxplot gathers the AUC values obtained in the 16 catchments. The letters below the boxplots result from the Friedman test (Lowry, 2008). This test consists in considering catchments as judges of the five methods. The test, which is based on rankings as evaluated by the catchments, assesses whether the methods are significantly different by evaluating if their rankings resemble a random shuffling. Based on this test, two boxplots that share
5 a letter at a given lead time are not significantly different.

Results show that all ensembles but HistQ are very close in terms of discrimination. As expected, their performance decreases as the lead time increases, except for HistQ, whose discrimination does not vary with the lead time. For all lead times, ESP significantly provides the best discrimination, with most AUC values superior to 0.88. ESP_SPI3 and Sys4 have equivalent performance in terms of discrimination and appear as second best, with most AUC values greater than 0.82.
10 HistQ_SPI3 is also very close to the performances of Sys4 and ESP_SPI3, but does not score as high as they do, especially for longer lead times. Overall, the discrimination of the conditioned ensembles is always between that of Sys4 and that of the base ensemble they are selected from (i.e. ESP for ESP-SPI3 and HistQ for HistQ_SPI3). For HistQ_SPI3, this translates into a gain in discrimination. However, for ESP_SPI3, this translates into a loss as ESP is already superior to Sys4. Finally, we also note that ensembles based on hydrological modelling (Sys4, ESP and ESP_SPI3) provide the best skills in terms of
15 forecast discrimination, at least for lead times shorter than 90 days, probably because they take into account initial hydrological conditions. All these conclusions are also valid when the 60th exceedance percentile is used as threshold (not shown).

3.3.2 Impact of the conditioning on forecasting low-flow variables

We now compare the forecast systems based on variables of interest for water management during low flows, namely the
20 weekly deficit duration and the weekly deficit volume. The weekly deficit duration corresponds to the number of days per week during which the daily streamflow is below a given threshold. The weekly deficit volume corresponds to the flow volume per week below this threshold. Figure 9 presents the PIT areas obtained with Sys4, ESP_SPI3, ESP, HistQ_SPI3 and HistQ when forecasting the weekly number of days below the 80th exceedance percentile. Boxplots represent the range of PIT areas obtained over the catchment set. Results are presented for lead times of two weeks, five weeks and twelve weeks
25 (columns). Again, letters represent the results of the Friedman test. Two boxplots that share a letter are not significantly different. Figure 10 proposes the same evaluation for the weekly streamflow deficit volume below the 80th exceedance percentile.

Figure 9 shows that the difference between the five ensembles is very tenuous when forecasting the deficit duration. For instance, all lower and upper quartiles of Sys4, ESP_SPI3, ESP and HistQ_SPI3 are included in the [0.01, 0.08] interval of
30 PIT area values, regardless of the lead time. Overall, ESP, ESP_SPI3 or HistQ_SPI3 perform best to forecast the deficit duration. The analysis of the corresponding PIT diagrams (not presented) showed that all ensembles are equivalently reliable, except for HistQ, which systematically overestimates the deficit duration. Here again, the reliability of the

conditioned ensembles in forecasting low-flow duration is located between that of Sys4 and that of the base ensemble they are selected from. An exception is that HistQ_SPI3 is significantly the best performing ensemble for a lead time of two weeks. In that case, conditioning has managed to improve over both Sys4 and HistQ base ensembles.

The gap between ensembles widens when looking at the deficit volume (Figure 10). For lead times of two and five weeks, ESP and ESP_SPI3 provide consistently reliable ensembles, and lower PIT areas than the others. For a lead time of twelve weeks, ESP_SPI3 along with Sys4 and HistQ_SPI3 provide the most reliable ensembles. The corresponding PIT diagrams (not presented) showed that HistQ_SPI3 tends to underestimate deficit volumes at all lead times. Ensembles issued with hydrological modelling also slightly underestimate the deficit volume at long lead times. Again, overall, conditioned ensembles are located between the two ensembles they are based on (Sys4 for the conditioning statistics and their respective base ensemble for the application of the conditioning). Here, the case of ESP_SPI3 is particularly interesting. Indeed, at short lead times, ESP_SPI3 benefits from ESP, which is more reliable than Sys4. At long lead times, it benefits from Sys4, as it becomes more reliable. ESP_SPI3 is thus consistently one of the best options to forecast deficit volumes for all three lead times. This shows that a conditioning approach can be of great interest when the ensembles used to build the conditioned scenarios show good performance but at different lead time ranges.

15 **3.4 Using the conditioned ensembles in drought risk forecasting**

Figure 11 illustrates the case of the 2003 drought with the streamflow forecasts issued on July 1st 2003 for the 90 days ahead. The figure focuses on catchment 5, the Azergues at Lozanne, in which the 2003 drought was hydrologically more severe than the reference 1976 drought. Each column represents the graphs obtained with one of the five ensemble forecasts (Sys4, ESP_SPI3, ESP, HistQ_SPI3 and HistQ). The upper row presents the graphical representation we propose to assess drought risks based on the ensemble forecasts. The graphs represent the deficit duration against the deficit volume, both computed based on the 80th exceedance percentile. The graph is divided into 49 boxes corresponding to possible combinations and ranges of deficit volumes and durations. The colour within each of these boxes indicates the percentage of ensemble members that falls within each box. The darker the boxes, the more ensemble members are indicating the associated drought risk in terms of deficit duration (y-axis) and volume (x-axis). Darker boxes may also reflect a sharper ensemble, and, if the darker boxes are around the observation, an ensemble with good discrimination (at least for the event considered). Coloured dots represent the observation (indicated as “observed”) and two references: the 1976 drought (indicated as “drought”) and the historical mean duration and deficit volume over the forecast period (indicated as “climatology”). The lower row in the figure presents the corresponding hydrographs over the 90-day forecast period. The black line represents the observed streamflow, the red line represents the 80th exceedance percentile and the blue lines represent the members of the ensemble forecast.

All ensembles produce similar patterns, but with different probabilities. The maximum probability is obtained with HistQ_SPI3 with 60 % of the ensemble members falling in the same cell. Ensembles based on hydrological modelling reach

maximum probabilities of 20 to 30 %, and HistQ does not exceed a probability of 14 %. These colours reflect how sharp the ensemble forecasts are for this forecast. The objective is to have a maximum of darker cells close to the observation. We observe that the graph obtained with HistQ puts equivalent weights to a wide range of scenarios, ranging from no risk to high risk of a drought situation, which remind us of its good reliability but poor sharpness. This ensemble conveys little information to assess drought risks. HistQ_SPI3, as opposed to HistQ, offers a more confident risk assessment with the highest forecast probabilities and only three coloured cells. Eighty percent of the forecast members indicate a drought equivalent or more severe than that of 1976. The high probability may be explained by the fact that SPI forecast members and initial hydrological conditions were often best represented by the same driest year (as suggested by the hydrographs), namely 1976.

10 The ESP forecast provides a wider view of the risk of drought, with higher probabilities located in the upper right part of the graph and small probabilities of having moderate low flow conditions. ESP is able to forecast a more severe event than the one observed during the 1976 drought. This good performance can only be attributed to the initial hydrological conditions since ESP does not have any information on future precipitations apart from climatology. Conditioning ESP (ESP_SPI3) slightly reduces the number of coloured cells with slightly higher probabilities in some of the upper right cells. The difference between ESP and ESP_SPI3 is clear when looking at the hydrographs. With ESP_SPI3, the number of high-flow peaks is reduced. The SPI3 conditioning seems to prevent the selection of some wet sequences from the climatology.

15 Sys4 also provides a quite good risk assessment since only the upper right cells are coloured. For this event, there seems to be an added value from the use of GCM-based forecasts (directly as forcing to a streamflow forecasting model or through a conditioning statistics) to better assess the risk of drought. Notably, in the specific case illustrated here, the conditioned ensembles (ESP_SPI3 and HistQ_SPI3) indicated a (small) probability of drought in the box corresponding to the observation, while their base ensembles (ESP and HistQ, respectively) indicated none.

20 In summary, while ensembles based on hydrological modelling, i.e. ESP, ESP_SPI3 and Sys4, are limited by the capacity of the model to reproduce small low-flow variations and thus to slightly underestimate the deficit volume, ensembles based on historical streamflows are limited within the range of past precipitation and streamflow scenarios. This highlights the fact that the studied methods, and here specifically Sys4, ESP_SPI3 and HistQ_SPI3, have different limitations but also different assets. Note that different contexts might penalize or favour different methods.

4 Conclusion

We investigated the impact on forecast attributes from conditioning precipitation climatology and historical streamflows on precipitation indices derived from ECMWF System 4 (GCM) seasonal forecasts. In a first step, the attributes of overall performance, sharpness and reliability of the conditioned ensembles were analysed with respect to the performance of the ensembles they were based on. Lead times up to 90 days and 16 catchments in France were considered. The main conclusions from this analysis are:

- The use of Sys4 forecasts to derive conditioned ensembles generally did not improve the overall performance of seasonal streamflow forecasts. Overall performance criteria typically give equal weight to complementary features of forecast quality. This is the case of the CRPS, which, unless specified otherwise, gives equal weights to its components of reliability, resolution and uncertainty (Pappenberger et al., 2015).
- 5 - Selecting traces within precipitation climatology or historical streamflow generally improved sharpness and decreased reliability. Conditioning based on the SPI provided more consistent results between catchments than conditioning based on cumulative precipitations.
- Particularly, conditioning based on SPI3 statistics derived from GCM-based seasonal forecasts proved to be beneficial for several catchments at lead times longer than 15 to 30 days. Indeed, the conditioned ensembles could
10 be as sharp as the GCM-based forecasts while being also, in most cases, more reliable than or as reliable as GCM-based seasonal forecasts.
- A simultaneous evaluation of the attributes of sharpness and reliability of the conditioned ensembles showed that conditioning led to ensembles that were more reliable and less sharp than the streamflow forecasts generated from System 4 precipitations. The conditioned ensembles were however less reliable and sharper than the ensembles they
15 were selected from (here, ESP and historical streamflows). Also, the conditioned ensembles seemed to take advantage of the information of either precipitation climatology or historical streamflows at the shorter lead times and of the information of GCM-based forecasts at the longer lead times.
- Conditioning could, in some cases, improve reliability and sharpness simultaneously, especially for lead times shorter than a month ahead. Nevertheless, this was seen in a few cases and, more often, a trade-off between
20 reliability and sharpness was highlighted. This is in accordance with other studies (Hamlet and Lettenmaier, 1999; Yao and Georgakakos, 2001).

The performance of the ensembles in forecasting low-flow events and low-flow variables was evaluated. Their capacity to discriminate between low-flow events and non-events and their capacity to forecast streamflow deficit volume and duration, as defined by the 80th exceedance percentile, were assessed. The main conclusion from this evaluation is that building
25 conditioned scenarios in seasonal low flow forecasting can be especially valuable when the forecasts that provide information for the conditioning approach (either by providing a conditioning statistics or by serving as a base ensemble to which the conditioning will be applied) perform well for different lead times. Conditioned ensembles can benefit from the good performance of different ensembles at different lead times. They can thus provide more consistent performances throughout a wider range of lead times.

30 Lastly, a drought-risk graphic representation was proposed to illustrate how different conditioned ensembles, with different performance in terms of the main forecast attributes evaluated in this study, could detect a drought event that occurred in 2003 in France. In this particular case, a 3-month forecast with conditioned ensembles based on SPI3 showed better results

in terms of indicating higher probabilities closer to the observed deficits in duration and volume of streamflows below the 80% percentile.

In this paper, we evaluated eight streamflow forecast scenarios with the aim of investigating the impact of conditioning on forecast attributes. Further investigations could be done with other conditioning methods of interest for operational use. For instance, the conditioning based on the forecast SPI or on the cumulative precipitation for the three coming months puts an equivalent weight on all three lead times to select past precipitations. However, seasonal forecasts issued by GCMs usually have more skill for the coming month than for the second and third months. Therefore, we could explore a weighting of these three forecast lead times in order to put more weight on the first lead month in the selection of past precipitations. In addition, one important parameter to forecast low flows and droughts is the temperature. A more advanced approach would consist in selecting past scenarios based on the SPEI (Standardized Precipitation-Evapotranspiration Index) calculated from seasonal precipitation and temperature forecasts.

Other types of conditionings can be found in the literature and could also be investigated. As an example, Werner et al. (2005) and Shukla et al. (2012) have investigated the use of medium-range weather forecasts to improve long-range forecasting. These approaches rely on the fact that short-term events are well forecast by short-term to medium-term forecasts issued by GCMs and that the benefit from medium-range forecasts can be extended to longer lead times through the inertia of a catchment. One could also apply a multi-model averaging method to merge the forecasts from the different ensembles investigated in this paper (see, for instance, Raftery et al., 2005; Duan et al., 2007; Najafi and Moradkhani, 2016). The influence of such a method on the evaluation of forecast attributes could be compared to the findings of this study with the conditioning approaches, especially towards a better assessment of the trade-off between reliability and sharpness.

Finally, we investigated conditionings within climatology solely based on past precipitations, past streamflows and catchment conditions. SPI values were computed after an aggregation of System 4 precipitation forecasts at the catchment scale, and, therefore, the conditioning and the spatial aggregation were independent. Further investigations could assess the potential of the conditioning methods for the spatial downscaling of System 4 seasonal precipitation forecasts before their application to hydrologic forecasting.

25 **5 Data availability**

ECMWF data are available under a range of licences. For more information please visit <http://www.ecmwf.int>. SAFRAN meteorological reanalysis data come from Météo-France and are available under licence conditions. Streamflow data are made available by the French Ministère de l'Écologie et du Développement Durable, and can be downloaded from <http://www.hydro.eaufrance.fr>.

Acknowledgements

The first author was partly funded by the DROP project (Benefit of governance in DROught adaptation) of the Interreg IVB NWE programme of the European Union. The second and the third authors were partly funded by the IMPREX project supported by the European Commission under the Horizon 2020 Framework programme, with grant nr 641811. The authors
5 thank Météo-France and SCHAPI for providing climate and hydrological data, respectively, and ECMWF for providing seasonal forecast data and for hosting the first author for two weeks.

References

- Anderson, M. L., Mierzwa, M. D. and Kavvas, M. L.: Probabilistic seasonal forecasts of droughts with a simplified coupled hydrologic-atmospheric model for water resources planning, *Stoch. Environ. Res. Risk Assess.*, 14, 263–274,
10 doi:10.1007/s004770000049, 2000.
- Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M.-H. and Valéry, A.: HESS Opinions “Crash tests for a standardized evaluation of hydrological models,” *Hydrol. Earth Syst. Sci.*, 13(10), 1757–1764, doi:10.5194/hess-13-1757-2009, 2009.
- Arlot, S. and Celisse, A.: A survey of cross-validation procedures for model selection, *Stat. Surv.*, 40–79, doi:10.1214/09-SS054, 2010.
15
- Bierkens, M. F. P. and van Beek, L. P. H.: Seasonal Predictability of European Discharge: NAO and Hydrological Response Time, *J. Hydrometeorol.*, 10(4), 953–968, doi:10.1175/2009JHM1034.1, 2009.
- Block, P. and Rajagopalan, B.: Statistical–Dynamical Approach for Streamflow Modeling at Malakal, Sudan, on the White Nile River, *J. Hydrol. Eng.*, 14(2), 185–196, doi:10.1061/(ASCE)1084-0699(2009)14:2(185), 2009.
- 20 Carpenter, T. M. and Georgakakos, K. P.: Assessment of Folsom lake response to historical and potential future climate scenarios: 1. Forecasting, *J. Hydrol.*, 249(1–4), 148–175, doi:10.1016/S0022-1694(01)00417-6, 2001.
- Ceppi, A., Ravazzani, G., Corbari, C., Salerno, R., Meucci, S. and Mancini, M.: Real-time drought forecasting system for irrigation management, *Hydrol. Earth Syst. Sci.*, 18(9), 3353–3366, doi:10.5194/hess-18-3353-2014, 2014.
- Crochemore, L., Ramos, M.-H. and Pappenberger, F.: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, *Hydrol. Earth Syst. Sci. Discuss.*, 2016, 1–32, doi:10.5194/hess-2016-78, 2016.
25
- Day, G.: Extended Streamflow Forecasting Using NWSRFS, *J. Water Resour. Plan. Manag.*, 111(2), 157–170, 1985.
- 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 Resour. Res.*, 49(5), 2729–2746, doi:10.1002/wrcr.20251, 2013.
- 30 Duan, Q., Ajami, N. K., Gao, X. and Sorooshian, S.: Multi-model ensemble hydrologic prediction using Bayesian model averaging, *Adv. Water Resour.*, 30(5), 1371–1386, doi:http://dx.doi.org/10.1016/j.advwatres.2006.11.014, 2007.

- Dutra, E., Pozzi, W., Wetterhall, F., Di Giuseppe, F., Magnusson, L., Naumann, G., Barbosa, P., Vogt, J. and Pappenberger, F.: Global meteorological drought – Part 2: Seasonal forecasts, *Hydrol. Earth Syst. Sci.*, 18(7), 2669–2678, doi:10.5194/hess-18-2669-2014, 2014.
- 5 Easey, J., Prudhomme, C. and Hannah, D. M.: Seasonal forecasting of river flows: a review of the state-of-the-art, in *Proceedings of the fifth FRIEND World Conference*, vol. 308, pp. 158–162, IAHS Publ., Havana, Cuba., 2006.
- Faber, B. A. and Stedinger, J. R.: Reservoir optimization using sampling SDP with ensemble streamflow prediction (ESP) forecasts, *J. Hydrol.*, 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, *Meteorol. Appl.*, 15(1), 19–24, doi:10.1002/met.45, 2008.
- 10 Gneiting, T., Balabdaoui, F. and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(2), 243–268, 2007.
- Gobena, A. K. and Gan, T. Y.: Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system, *J. Hydrol.*, 385(1–4), 336–352, doi:10.1016/j.jhydrol.2010.03.002, 2010.
- 15 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, *J. Hydrol.*, 377(1–2), 80–91, 2009.
- Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M. and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, *Hydrol. Earth Syst. Sci.*, 18(2), 463–477, doi:10.5194/hess-18-463-2014, 2014.
- Hamlet, A. F. and Lettenmaier, D. P.: Columbia River Streamflow Forecasting Based on ENSO and PDO Climate Signals, *J. Water Resour. Plan. Manag.*, 125(6), 333–341, doi:10.1061/(ASCE)0733-9496(1999)125:6(333), 1999.
- 20 Hamlet, A. F., Huppert, D. and Lettenmaier, D. P.: Economic value of long-lead streamflow forecasts for Columbia River hydropower, *J. Water Resour. Plan. Manag.*, 128(2), 91–101, doi:10.1061/(ASCE)0733-9496(2002)128:2(91), 2002.
- Hao, Z., AghaKouchak, A., Nakhjiri, N. and Farahmand, A.: Global integrated drought monitoring and prediction system, *Sci. Data*, 1, 140001, 2014.
- 25 Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather Forecast.*, 15(5), 559–570, 2000.
- Ionita, M., Boroneant, C. and Chelcea, S.: Seasonal modes of dryness and wetness variability over Europe and their connections with large scale atmospheric circulation and global sea surface temperature, *Clim. Dyn.*, 45(9), 2803–2829, doi:10.1007/s00382-015-2508-2, 2015.
- Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, *J. Hydrol.*, 249(1–4), 2–9, doi:10.1016/S0022-1694(01)00420-6, 2001.
- 30 Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrol. Earth Syst. Sci.*, 11(4), 1267–1277, doi:10.5194/hess-11-1267-2007, 2007.
- Lowry, R.: *Concepts and Applications of Inferential Statistics*, Vassar College., 2008.

- Mason, S. J. and Graham, N. E.: Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels, *Weather Forecast.*, 14(5), 713–725, doi:10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2, 1999.
- McKee, T., Doeskin, N. and Kleist, J.: The relationship of drought frequency and duration to time scales, pp. 179–184., 1993.
- 5 Mishra, A. K. and Singh, V. P.: A review of drought concepts, *J. Hydrol.*, 391(1–2), 202–216, doi:10.1016/j.jhydrol.2010.07.012, 2010.
- Molteni, F., Stockdale, T., Balsaseda, 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., 2011.
- 10 Najafi, M. R. and Moradkhani, H.: Ensemble Combination of Seasonal Streamflow Forecasts, *J. Hydrol. Eng.*, 21(1), 04015043, doi:10.1061/(ASCE)HE.1943-5584.0001250, 2016.
- 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.
- 15 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, *J. Hydrol.*, 303(1–4), 290–306, 2005.
- Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A. and Salamon, P.: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction, *J. Hydrol.*, 522, 697–713, doi:http://dx.doi.org/10.1016/j.jhydrol.2015.01.024, 2015.
- Poumadère, M., Mays, C., Le Mer, S. and Blong, R.: The 2003 Heat Wave in France: Dangerous Climate Change Here and Now, *Risk Anal.*, 25(6), 1483–1494, doi:10.1111/j.1539-6924.2005.00694.x, 2005.
- Pushpalatha, R., Perrin, C., Mathevet, T. and Andreassian, V.: A downward structural sensitivity analysis of hydrological models to improve low-flow simulation, *J. Hydrol.*, 411(1–2), 66–76, doi:10.1016/j.jhydrol.2011.09.034, 2011.
- 25 Pushpalatha, R., Perrin, C., Moine, N. L. and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, *J. Hydrol.*, 420–421(0), 171–182, doi:10.1016/j.jhydrol.2011.11.055, 2012.
- 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, *J. Appl. Meteorol. Climatol.*, 47(1), 92–107, doi:10.1175/2007JAMC1636.1, 2008.
- 30 Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, *Mon. Weather Rev.*, 133(5), 1155–1174, doi:10.1175/MWR2906.1, 2005.
- Ramos, M. H., van Andel, S. J. and Pappenberger, F.: Do probabilistic forecasts lead to better decisions?, *Hydrol. Earth Syst. Sci.*, 17(6), 2219–2232, doi:10.5194/hess-17-2219-2013, 2013.
- 35 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 Resour. Res.*, 46(5), W05521, doi:10.1029/2009WR008328, 2010.

- Sauquet, E., Lerat, J. and Prudhomme, C.: La prévision hydro-météorologique à 3-6 mois. Etat des connaissances et applications, *Houille Blanche*, (6), 77–84, doi:10.1051/lhb:2008075, 2008.
- Seibert, M. and Trambauer, P.: Seasonal forecasts of hydrological drought in the Limpopo basin: Getting the most out of a bouquet of methods, in *Drought: Research and Science-Policy Interfacing*, pp. 307–313, CRC Press. [online] Available from: doi: 10.1201/b18077-52, 2015.
- Sheffield, J., Wood, E. F., Chaney, N., Guan, K., Sadri, S., Yuan, X., Olang, L., Amani, A., Ali, A., Demuth, S. and Ogallo, L.: A Drought Monitoring and Forecasting System for Sub-Sahara African Water Resources and Food Security, *Bull. Am. Meteorol. Soc.*, 95(6), 861–882, doi:10.1175/BAMS-D-12-00124.1, 2013.
- Shukla, S., Voisin, N. and Lettenmaier, D. P.: Value of medium range weather forecasts in the improvement of seasonal hydrologic prediction skill, *Hydrol. Earth Syst. Sci.*, 16(8), 2825–2838, doi:10.5194/hess-16-2825-2012, 2012.
- Shukla, S., Sheffield, J., Wood, E. F. and Lettenmaier, D. P.: On the sources of global land surface hydrologic predictability, *Hydrol. Earth Syst. Sci.*, 17(7), 2781–2796, doi:10.5194/hess-17-2781-2013, 2013.
- Shukla, S., McNally, A., Husak, G. and Funk, C.: A seasonal agricultural drought forecast system for food-insecure regions of East Africa, *Hydrol. Earth Syst. Sci.*, 18(10), 3907–3921, doi:10.5194/hess-18-3907-2014, 2014.
- Šípek, V. and Daňhelka, J.: Modification of input datasets for the Ensemble Streamflow Prediction based on large-scale climatic indices and weather generator, *J. Hydrol.*, 528, 720–733, doi:10.1016/j.jhydrol.2015.07.008, 2015.
- Smakhtin, V. U.: Low flow hydrology: a review, *J. Hydrol.*, 240(3–4), 147–186, doi:10.1016/S0022-1694(00)00340-1, 2001.
- Svensson, C.: Seasonal river flow forecasts for the United Kingdom using persistence and historical analogues, *Hydrol. Sci. J.*, 61(1), 19–35, doi:10.1080/02626667.2014.992788, 2016.
- Tallaksen, L. M., Madsen, H. and Clausen, B.: On the definition and modelling of streamflow drought duration and deficit volume, *Hydrol. Sci. J.*, 42(1), 15–33, doi:10.1080/02626669709492003, 1997.
- 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, *Hydrol. Earth Syst. Sci.*, 19(4), 1695–1711, doi:10.5194/hess-19-1695-2015, 2015.
- UNEP: Impacts of summer 2003 heat wave in Europe, *Environment Alert Bulletin*, United Nations Environment Programme, Nairobi. [online] Available from: http://www.unisdr.org/files/1145_ewheatwave.en.pdf, 2004.
- 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, *Int. J. Climatol.*, 30(11), 1627–1644, doi:10.1002/joc.2003, 2010.
- Wang, E., Zhang, Y., Luo, J., Chiew, F. H. S. and Wang, Q. J.: Monthly and seasonal streamflow forecasts using rainfall-runoff modeling and historical weather data, *Water Resour. Res.*, 47(5), doi:10.1029/2010WR009922, 2011.
- Werner, K., Brandon, D., Clark, M. and Gangopadhyay, S.: Climate index weighting schemes for NWS ESP-based seasonal volume forecasts., *J. Hydrometeorol.*, 5(6), 1076–1090, 2004.
- Werner, K., Brandon, D., Clark, M. and Gangopadhyay, S.: Incorporating medium-range numerical weather model output into the Ensemble Streamflow Prediction system of the National Weather Service, *J. Hydrometeorol.*, 6(2), 101–114, 2005.

Wilhite, D. A., Hayes, M. J., Knutson, C. and Smith, K. H.: Planning for drought: Moving from crisis to risk management, *JAWRA J. Am. Water Resour. Assoc.*, 36(4), 697–710, doi:10.1111/j.1752-1688.2000.tb04299.x, 2000.

WMO: Manual on Low-flow Estimation and Prediction, World Meteorological Organization., 2008.

WMO: Standardized Precipitation Index User Guide, World Meteorological Organization., 2012.

- 5 Wood, A. W. and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, *Geophys. Res. Lett.*, 35(14), L14401, doi:10.1029/2008GL034648, 2008.

Wood, A. W., Hopson, T., Newman, A., Brekke, L., Arnold, J. and Clark, M.: Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill, *J. Hydrometeorol.*, 17(2), 651–668, doi:10.1175/JHM-D-14-0213.1, 2016.
- 10 Yao, H. and Georgakakos, A.: Assessment of Folsom Lake response to historical and potential future climate scenarios: 2. Reservoir management, *J. Hydrol.*, 249(1), 176–196, 2001.

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 Resour. Res.*, 49(8), 4687–4699, doi:10.1002/wrcr.20350, 2013.
- 15 Yuan, X., Wood, E. F. and Ma, Z.: A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, *Wiley Interdiscip. Rev. Water*, 523–536, doi:10.1002/wat2.1088, 2015.

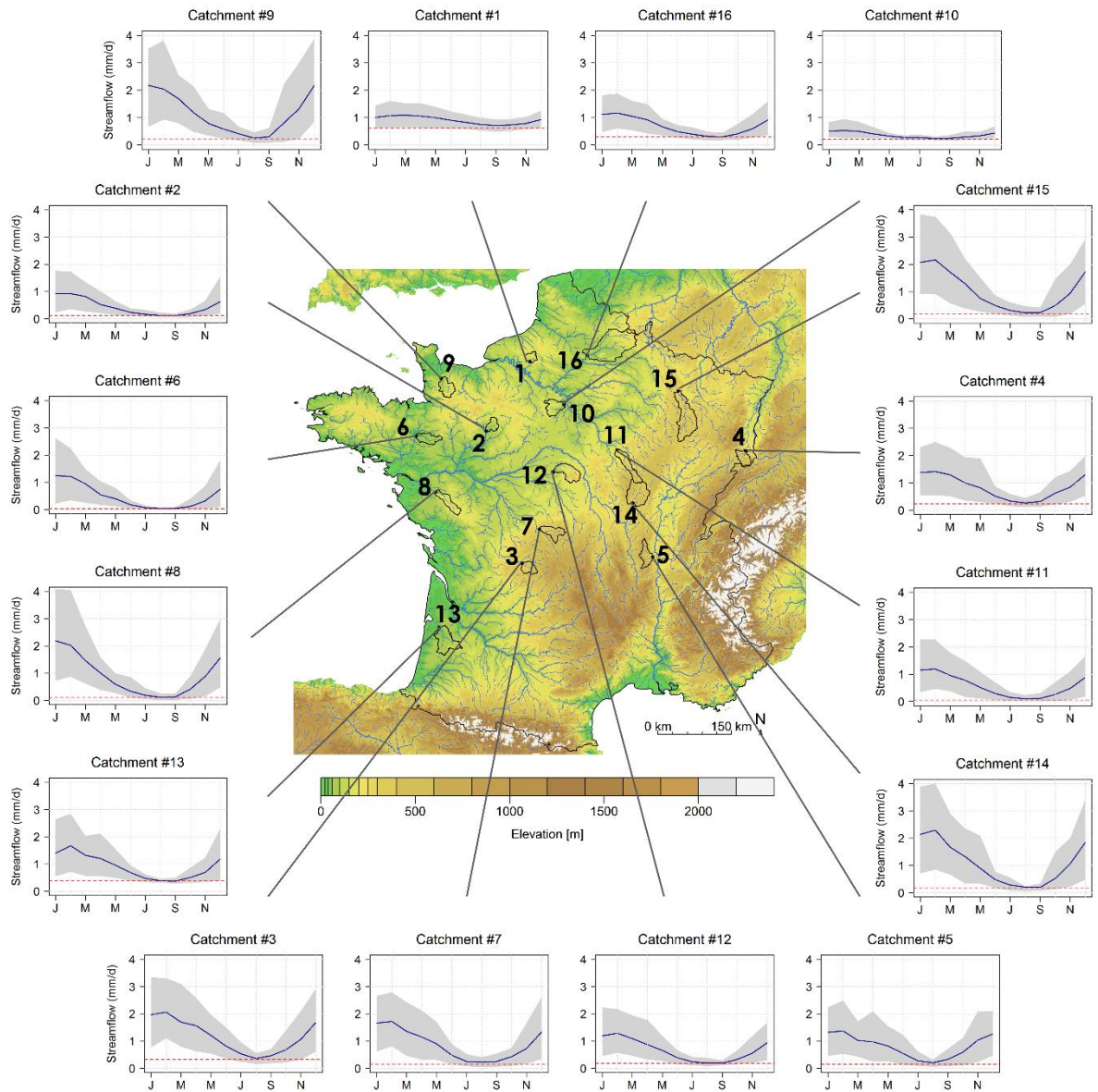


Figure 1 Location in France and hydrological regime of the 16 catchments. Solid lines represent mean interannual monthly flows. Grey-shaded areas represent the 10th and 90th percentiles of interannual monthly flows. Dotted red lines represent the 80th exceedance percentile (i.e. the daily flow exceeded by 80 % of the data). The catchments are numbered from the smallest to the largest. Statistics are computed

5 over the streamflow record available for each catchment, i.e. 36 to 52 years (cf. Table 1).



Figure 2 CRPSS and IQRSS of SPI forecasts and forecasts of cumulative precipitations produced from bias corrected System 4 precipitation forecasts. The reference for the skill scores is climatology. Skill scores are presented for statistics calculated for each month of the three-month lead time (SUM1 and SPI1) and over the three months altogether (SUM3 and SPI3). Columns correspond to scores computed for sums, SPI values, SPI values smaller than -1 (dry), SPI values within -1 and 1 (normal) and SPI values greater than 1 (wet).

5

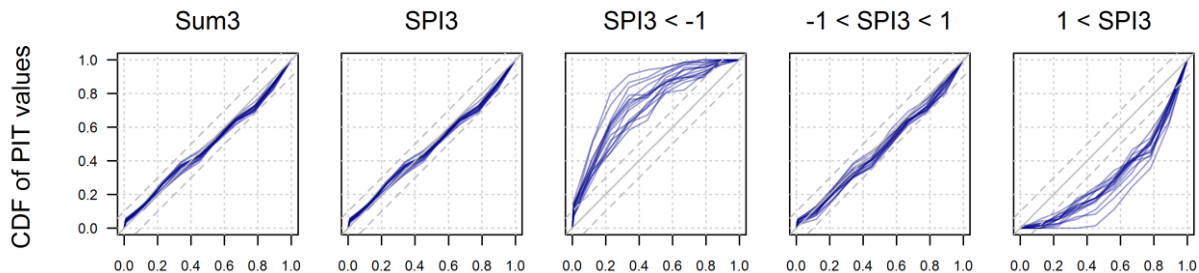


Figure 3 Reliability of SPI forecasts and forecasts of cumulative precipitations produced from bias corrected System 4 precipitation forecasts. PIT diagrams are presented for statistics calculated over the first three months altogether (Sum3 and SPI3). Columns correspond to scores computed for sums, SPI values, SPI values smaller than -1 (dry), SPI values within -1 and 1 (normal) and SPI values greater than 1 (wet).

15

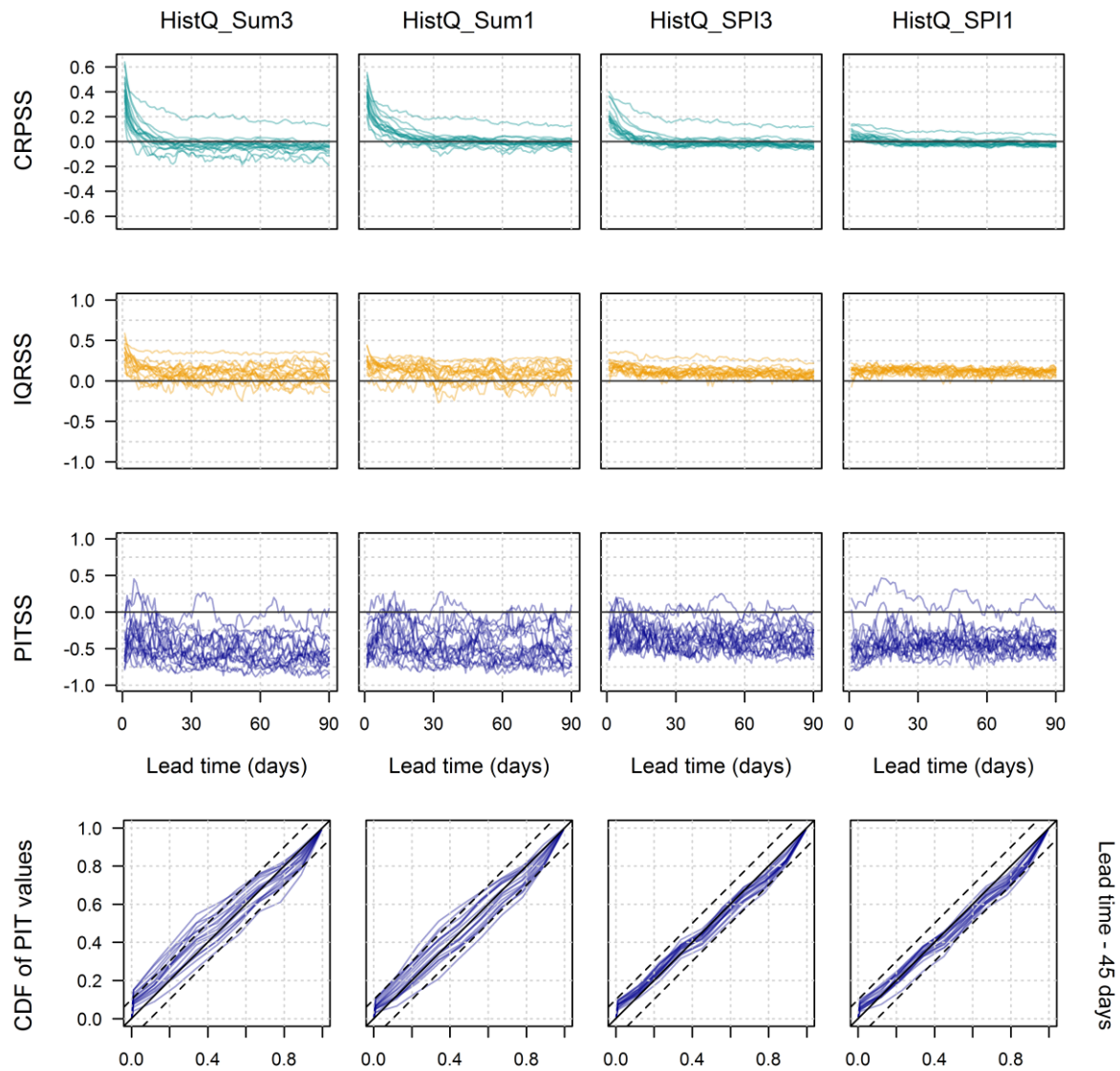


Figure 4 Skill scores (CRPSS, IQRSS, PITSS; first three rows) and PIT diagrams for a lead time of 45 days (last row) of the conditioned ensemble forecast scenarios: HistQ_Sum3, HistQ_Sum1, HistQ_SPI3 and HistQ_SPI1. In the skill scores, the reference forecast is the base ensemble HistQ. Each line represents one of the 16 catchments investigated.

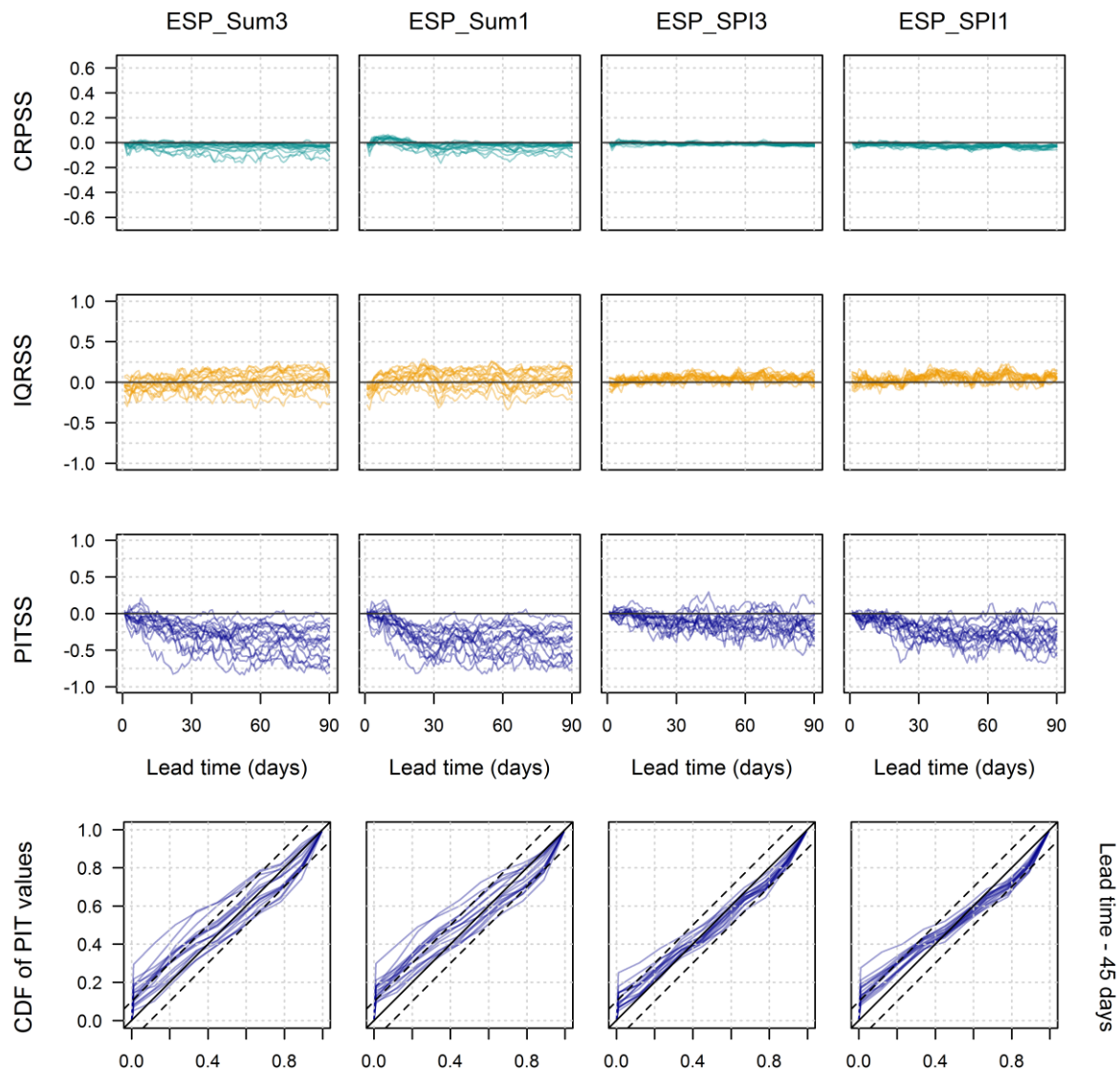


Figure 5 Same as Figure 4 but the forecast ensembles are ESP_Sum3, ESP_Sum1, ESP_SPI3 and ESP_SPI1 and the reference for the computation of the skill is ESP.

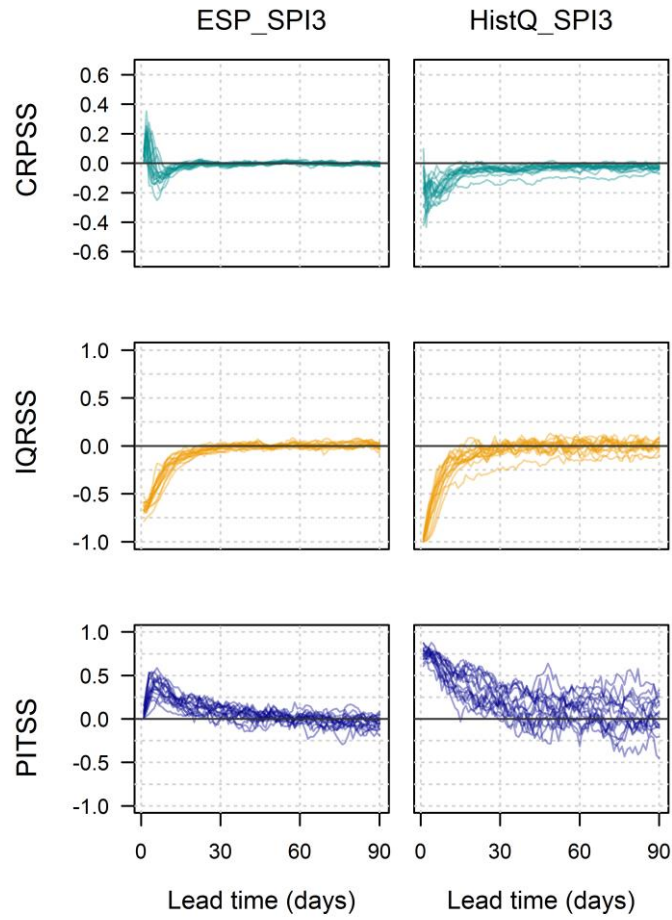


Figure 6 Skill scores (CRPSS, IQRSS, PITSS) of the conditioned ensemble forecast scenarios: ESP_SPI3 and HistQ_SPI3. In the skill scores, the reference forecast is the base ensemble Sys4. Each line represents one of the 16 catchments investigated.

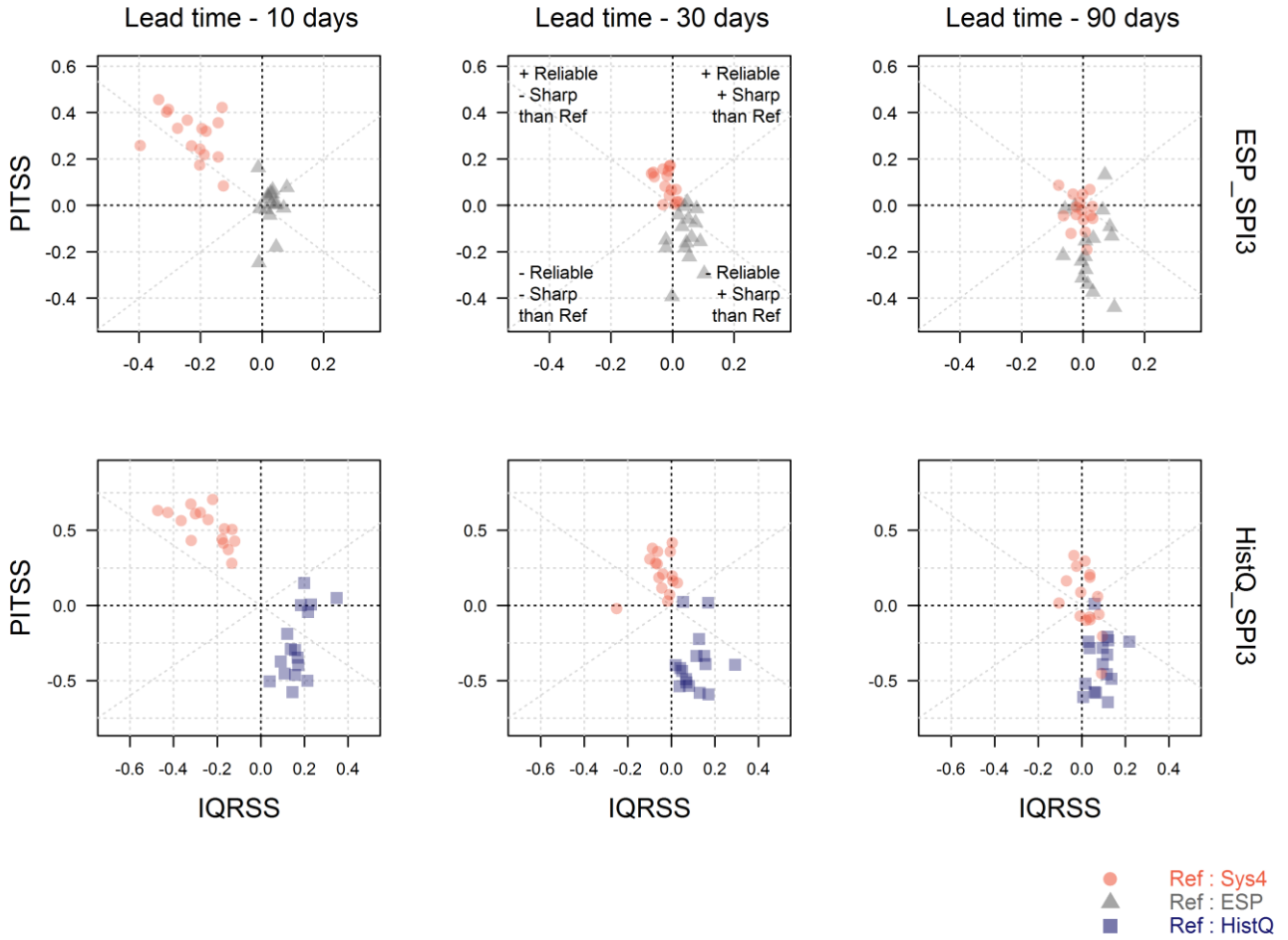


Figure 7 PITSS (reliability) versus IQRSS (sharpness) for ESP_SPI3 (upper row) and HistQ_SPI3 (lower row), and lead times of 10, 30 and 90 days (columns). ESP_SPI3 is compared to Sys4 (red) and ESP (grey), while HistQ_SPI3 is compared to Sys4 (red) and HistQ (blue). Each point represents one of the 16 catchments.

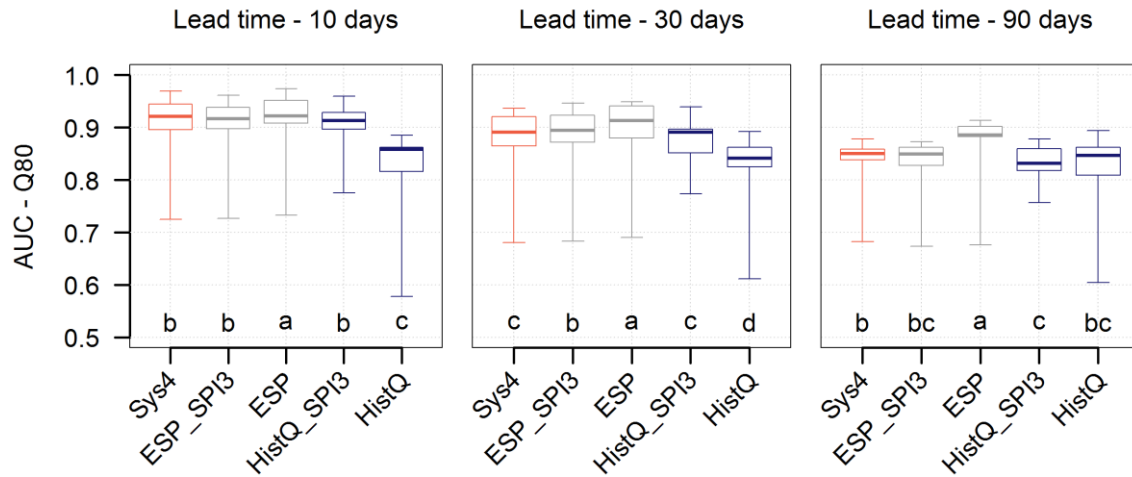


Figure 8 Ranges of the Area Under the Curve (AUC) of the ROC diagram based on the 80th exceedance percentile for each of the five selected ensemble forecasts (Sys4, ESP, HistQ, ESP_SPI3, HistQ_SPI3). Boxplots gather the AUC values for the 16 catchments. The boxes extend to the 25th and 75th percentiles and the whiskers, to the data extremes. Graphs are presented for 10-day, 30-days and 90-day lead times (columns). The letters below the boxplots result from the Friedman test. For a given lead time, two boxplots sharing a letter are not significantly different.

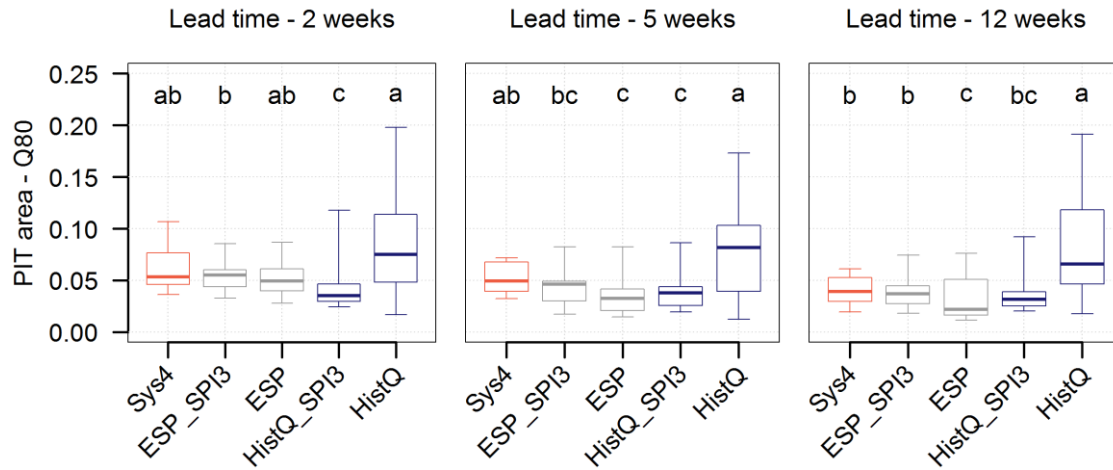


Figure 9 Same as Figure 7 but for PIT area ranges computed for deficit duration. Ranges are represented by boxplots which gather the PIT areas for the 16 catchments. Graphs are presented for lead times of two weeks, five weeks and twelve weeks (columns).

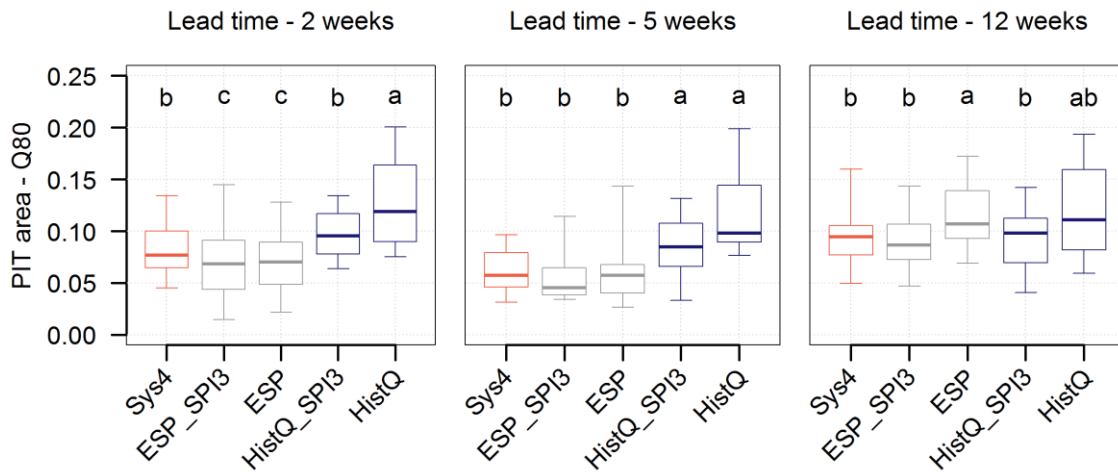


Figure 10 Same as Figure 8 for deficit volume.

5

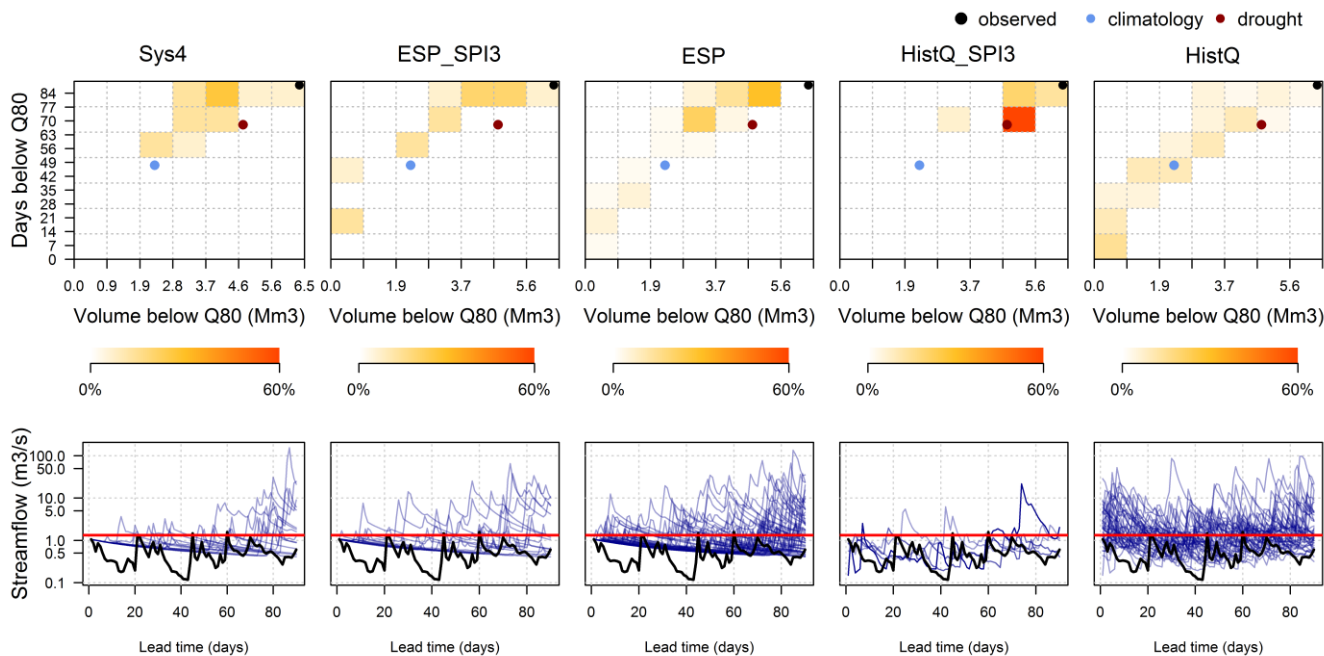


Figure 11 Risk graphs presenting the probabilities of deficit duration versus deficit volume based on the 80th exceedance percentile (upper row) and corresponding hydrographs (lower row). The maximum probability varies with the ensemble and the situation and is indicated in the colour scale. The black point corresponds to the observation, the dark red dot to the drought of 1976 and the blue dot to the mean duration and deficit volume observed in past streamflows. Each column corresponds to one of the five ensemble forecasts. Forecasts were issued for the Azergues at Lozanne (catchment 5) on 1 July 2003 for the next 90 days.

Table 1 River and gauging station, period with available streamflow observations, area and main hydroclimatic characteristics of the 16 catchments (ranked from the smallest to the largest). The mean annual streamflow is computed over the period of streamflow availability. The mean annual precipitation and evapotranspiration are computed over the 1958-2010 period.

#	River	Gauging station	Streamflow availability	Area (km ²)	Mean annual precipitation (mm/yr)	Mean annual potential evapotranspiration (mm/yr)	Mean annual streamflow (mm/yr)
1	Andelle	Vascoeuil	01/01/1973 - 27/02/2010	377	952	628	332
2	Orne Saosnoise	Montbizot [<i>Moulin Neuf Cidrerie</i>]	01/12/1967 - 04/03/2010	501	735	696	163
3	Briance	Condat-sur-Vienne [<i>Chambon Veyrinas</i>]	01/01/1966 - 28/03/2010	605	1100	706	427
4	Ill	Didenheim	01/11/1973 - 02/03/2010	668	956	664	309
5	Azergues	Lozanne	01/01/1965 - 28/03/2010	798	931	689	296
6	Seiche	Bruz [<i>Carcé</i>]	01/12/1967 - 11/03/2010	809	732	696	181
7	Petite Creuse	Fresselines [<i>Puy Rageaud</i>]	01/08/1958 - 28/03/2010	853	899	680	316
8	Sèvre Nantaise	Tiffauges [<i>la Moulinette</i>]	01/11/1967 - 04/03/2010	872	898	712	331
9	Vire	Saint-Lô [<i>Moulin des Rondelles</i>]	01/01/1971 - 03/02/2010	882	958	629	448
10	Orge	Morsang-sur-Orge	01/10/1967 - 07/03/2010	934	658	680	131
11	Serein	Chablis	01/08/1958 - 03/03/2010	1119	842	675	220
12	Sauldres	Salbris [<i>Valaudran</i>]	01/01/1971 - 28/03/2010	1220	803	684	240
13	Eyre	Salle	01/01/1967 - 19/03/2010	1678	1025	785	323
14	Arroux	Etang-sur-Arroux [<i>Pont du Tacot</i>]	01/11/1971 - 27/03/2010	1792	981	655	390
15	Meuse	Saint-Mihiel	01/07/1968 - 03/01/2010	2543	948	639	372
16	Oise	Sempigny	01/08/1958 - 02/03/2010	4320	805	639	250

Table 2 Summary of the methodology used to build the ensemble forecast scenarios.

	Name	Statistic on seasonal forecast used as condition	Additional condition	Size	Initial hydrological conditions	Hydrological model	Precipitation forecast
Base ensembles	HistQ	No condition	-	Between 35 and 51 depending on flow data availability (see Table 1)	no	no	no
	ESP	No condition	-	50	yes	yes	no
	Sys4	No condition	-	15 or 51	yes	yes	yes
	HistQ_Sum3	Precipitation volume			yes	no	no
Conditioned ensembles	HistQ_Sum1	Monthly precipitation volume	previous streamflow	15 or 51	yes	no	no
	HistQ_SPI3	SPI3			yes	no	no
	HistQ_SPI1	SPI1			yes	no	no
	ESP_Sum3	Precipitation volume			yes	yes	no
	ESP_Sum1	Monthly precipitation volume	-	15 or 51	yes	yes	no
	ESP_SPI3	SPI3			yes	yes	no
	ESP_SPI1	SPI1			yes	yes	no

Table 3 Rankings of the Sys4, ESP_SPI3, ESP, HistQ_SPI3 and HistQ streamflow ensembles, as evaluated by three evaluation criteria (in rows) and three lead time ranges (columns). The rankings are based on averaged skill scores for each ensemble, all catchments and for lead times 10 to 30, 31 to 60 and 61 to 90.

	10-30 days lead	30-60 days lead	60-90 days lead
Overall performance	1. Sys4	1. Sys4	1. Sys4
	2. ESP_SPI3	1. ESP_SPI3	1. ESP_SPI3
	2. ESP	1. ESP	1. ESP
	4. HistQ_SPI3	4. HistQ_SPI3	4. HistQ
	5. HistQ	4. HistQ	5. HistQ_SPI3
Sharpness	1. Sys4	1. Sys4	1. Sys4
	2. ESP_SPI3	1. ESP_SPI3	1. ESP_SPI3
	3. HistQ_SPI3	1. HistQ_SPI3	1. HistQ_SPI3
	4. ESP	4. ESP	4. ESP
	5. HistQ	5. HistQ	5. HistQ
Reliability	1. HistQ	1. HistQ	1. HistQ
	2. HistQ_SPI3	2. ESP	2. ESP
	3. ESP	3. HistQ_SPI3	3. HistQ_SPI3
	4. ESP_SPI3	4. ESP_SPI3	4. Sys4
	5. Sys4	5. Sys4	5. ESP_SPI3