Probabilistic flood forecasting on the Rhone River: evaluation with ensemble and analogue-based precipitation forecasts

Joseph Bellier[1,a], Isabella Zin[1], Stanislas Siblot[2] and Guillaume Bontron[2]

[1]*Université Grenoble Alpes, Laboratoire d'étude des Transferts en Hydrologie et Environnement, 70 rue de la Physique, 38400 St-Martin-d'Hères, France*
[2]*Compagnie Nationale du Rhône, 2 rue André Bonin, 69004 Lyon, France*

Abstract. Hydrological ensemble forecasting performances are analysed over 5 basins up to 2000 km² in the French Upper Rhone region. Streamflow forecasts are issued at an hourly time step from lumped ARX rainfall-runoff models forced by different precipitation forecasts. Ensemble meteorological forecasts from ECMWF and NCEP are considered, as well as analogue-based forecasts fed by their corresponding control forecast. Analogue forecasts are rearranged using an adaptation of the Schaake-Shuffle method in order to ensure the temporal coherence. A new evaluation approach is proposed, separating forecasting performances on peak amplitudes and peak timings for high flow events. Evaluation is conducted against both simulated and observed streamflow (so that relative meteorological and hydrological uncertainties can be assessed), by means of CRPS and rank histograms, over the 2007-2014 period. Results show a general agreement of the forecasting performances when averaged over the 5 basins. However, ensemble-based and analogue-based streamflow forecasts produce a different signature on peak events in terms of bias, spread and reliability. Strengths and weaknesses of both approaches are discussed as well as potential improvements, notably towards their merging.

1 Introduction

Flood events may cause severe damages in terms of both human lives and economic losses, especially when the anticipation has failed. Flood forecasts, in a form of streamflow forecasts, help decision makers to reduce these potential damages. However, streamflow forecasts are inherently uncertain. While single (or deterministic) forecasts makes end-users over-confident, probabilistic forecasts, which provide an estimate of the uncertainty, enable rational decision making [1, 2] and offer additional economic value [3]. Although difficulties still remain in the use of probabilities [4, 5], several operational flood forecasting services now routinely run Hydrological Ensemble Prediction Systems (HEPS) which provide probabilistic streamflow forecasts based on a number of possible hydrological scenarios (among others: EFAS [6], NOAA's HEFS [7], see [8] for a review).

The predictive uncertainty in streamflow forecasting is classically considered as coming from two main sources. Hydrological uncertainty, which represents errors made by the rainfall-runoff model, dominates for short lead times (i.e. up to the concentration time of the basin). Meteorological uncertainty, which represents the uncertain knowledge of future precipitation, is later assumed to represent the largest source of uncertainty [8]. This paper focuses on meteorological uncertainty.

To take the meteorological uncertainty into account, the flood forecasting community has started to widely use ensemble forecasts issued by Ensemble Prediction Systems (EPS). Based on the idea that the atmosphere is chaotic [9], ensemble forecasts consist of multiple runs of a Numerical Weather Prediction (NWP) model which simulate atmospheric physical processes with slightly altered initial conditions and model assumptions [10]. It therefore provides an ensemble of traces (as a form of precipitation time series) considered as equiprobable. These are further used as forcing (i.e. input data) of rainfall-runoff models. However, despite progress achieved in NWP in the last decades [11], ensemble forecasts for precipitation still suffer from biases in their mean and spread essentially due to the grid resolution which is often coarser than the scale of physical processes responsible for precipitation. Several previous studies showed that their correction is difficult and generally results in modest improvements [12, 13]. As an alternative, the statistical approach based on the search of analogues for producing precipitation forcing [14] is relevant for streamflow forecasting since it corrects a part of the bias. This method is operationally used at Compagnie Nationale du Rhône (CNR), Électricité de France (EDF) as well as in different French regional flood forecasting services. In a nutshell, ensemble forecasts represent mainly the uncertainty due to the dynamical evolution of the synoptic situation, while

[a] Corresponding author: joseph.bellier@univ-grenoble-alpes.fr

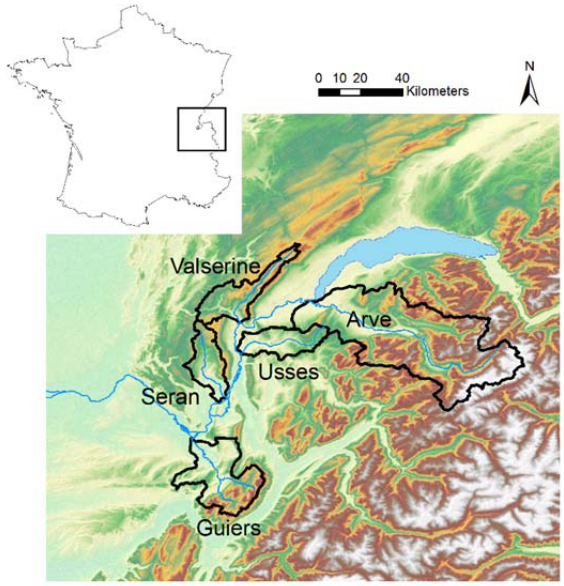

Figure 1. Location of the studied basins in the Upper Rhone region in France.

analogue-based forecasts represent the uncertainty of the thermodynamics related to the precipitation generation processes from a given synoptic situation. To date, very few studies [15] have evaluated and compared these two approaches in a flood forecasting context.

Evaluation aims to assess (by means of scores) the quality of forecasts given the observations. This step can assist forecasting system designers to choose which meteorological forcing suits their needs. Moreover, it contributes to improve the system by helping in the assessment of strengths and weaknesses regarding specific attributes [16], or as a communication tool with non-scientists concerning resource needs [17].

Evaluation of probabilistic precipitation forecasts can be conducted against precipitation observations. Nevertheless, several studies have shown that improvements in precipitation forecasts do not translate proportionally into streamflow forecasts [12, 13, 18]. Evaluation after the hydrological modelling step is therefore a major concern. Most of HEPS have thus implemented a probabilistic evaluation framework for streamflow forecasts [12, 19, 20].

It therefore raises the question about the variable being evaluated. In a flood management context, it is worth noting that binary decisions such as issuing a warning or opening a gate are generally taken according to the forecast peak flow values. Very few studies [21], though, have evaluated flood forecasts within a peak-based framework.

Besides, evaluations for severe flood events will always face the lack of data since these events are inherently rare. In addition, the non-stationarity of river basin conditions (changes in land cover, modification of river beds, hydroelectricity) and the evolution of NWP models make the use of long archives improper [8]. To address this shortcoming, two approaches can be undertaken: to strive for statistically robust evaluations by lowering thresholds defining events being evaluated, or to conduct case by case analyses for a better understanding of forecast behaviours.

Basin	Area (km²)	Mean elevation (m)	95th percentage of obs. streamflow (m³/s)
Arve	2082	1333	165
Valserine	361	990	50
Usses	309	603	17
Séran	290	717	22
Guiers	609	778	54

Table 1. Characteristics of the five study basins.

In this study, both approaches were conducted towards the assessment of strengths and weaknesses of ensemble-based and analogue-based streamflow forecasts on high flow events. The evaluation was made against both observed and simulated streamflow (output of the rainfall-runoff model fed with observed precipitation) in order to separate the effect of the meteorological uncertainty. We proposed a new evaluation approach focused on peak flows, by evaluating separately the amplitude and the timing of the peak using a probabilistic score. The paper is organized as follows. Study basins are briefly described in section 2. Section 3 presents the rainfall-runoff model and the observed and forecasted data. A detailed setup of the experiment is then given in section 4. Results, subsequent discussions and perspectives are presented in section 5 and 6.

2 Study basins

Streamflow forecasts were studied on five tributary basins of the Rhone River in its French upper part. Their locations are shown in Figure 1 while their characteristics are given in Table 1. These basins vary depending on their size and as a consequence on their high flow event amplitude.

3 Available data

3.1. Model and observed data

For rainfall-runoff modelling, lumped models of ARX type (i.e. Auto-Regressive model with eXogenous input [22]) were used. They are part of the integrated forecasting chain developed at CNR over the whole Rhone basin [23]. The output variable at a given time step t, here the streamflow Q^t, is calculated with an equation of the form:

$$Q^t = a + b_1 Q^{t-1} + b_2 Q^{t-2} + c_1 P^{t-1} + c_2 P^{t-2} \quad (1)$$

where precipitation P (averaged over the whole basin) is the exogenous input. Coefficients a, b_1, b_2, c_1, c_2 vary according to streamflow ranges, soil moisture conditions and seasons. Equation (1) shows a second-order autoregressive process for both Q and P but the order may be higher depending on basins. As any data-driven model, the calibration of ARX models requires a rather long archive of observed data. However, its operational use is easy, computationally inexpensive and low data

Name	Met. centre	NWP model	Resolution	members
ECMWF-Ens	ECMWF	IFS-EPS	0.25°	51
NCEP-Ens	NCEP	GEFS	1°	21

Table 2. Characteristics of ECMWF-ens and NCEP-ens forecasts.

demanding. The initialization, before a forecasting use, consists in calculating soil moisture conditions using antecedent precipitation observations.

Precipitation used for calibration and initialization came from a 6-hour dataset covering the 1992-2014 period. Calculated and furnished by Meteo-France, this dataset represents total precipitation (liquid and solid) spatially averaged over each basin. These precipitation data were also used as input to calculate simulated streamflow against which the evaluation of streamflow forecasts were made (see section 4 and 5).

Streamflow observations used for calibration and initialization came from CNR's gauging stations at the outlet of each basin. The dataset is available at an hourly time step, covering a 18 to 26 year period depending on basins.

One can note that temperature is not involved in equation (1), despite its influence on mountainous basins. In CNR's forecasting chain, a snow module is run before the rainfall-runoff model. Values for *P* are modified to take into account snow melt (more effective water for runoff) and solid precipitation (less effective water). However, this aspect has not been included in this study and will be discussed in section 5 and 6.

3.2. Precipitation forecasts

In this study, focus was made on ensemble forecasts and analogue-based precipitation forecasts. Since the latter are less commonly used in HEPS, a more detailed description is given in this section.

3.2.1 Ensemble forecasts from EPS

We used ensemble forecasts from two different NWP models: the Ensemble Prediction System of the Integrated Forecast System (IFS-EPS) from the European Centre for Medium-range Weather Forecasts (ECMWF) and the Global Ensemble Prediction System (GEFS) from the National Centers for Environmental Prediction (NCEP). More information about ensemble forecasting can be found in [24] and [25]. The variable of interest in this study is the 6-hour accumulated total precipitation, with a lead time from 6 to 120 hours. Forecasts are taken from the 00 UTC cycle. Characteristics of these forecasts are given in Table 2. Both datasets have been extracted from the TIGGE archive [25] over the 2007-2014 period. As raw forecasts are grid-based, Thiessen-based spatial averaging has been made in order to fulfil the requirement of the lumped rainfall-runoff model (a single value par basin). Thereafter, they will be referred as ECMWF-Ens and NCEP-Ens forecasts.

Lev.	Predictors	Analogy criterion	Domain	Number of analogues
L0	T850 (0h) & T500 (6h)	RMSE (0.5°)	Closest grid point	9000
L1	Z1000 (6h) & Z500 (0h)	S1 (2.5°)	Optimized	175
L2	RH850 × TCW (0-6h)	RMSE (0.5°)	Optimized	40

Table 3. Three levels of analogy applied for the search of analogues. *Optimized* means that the domain size and location has been optimized specifically for each basin.

3.2.2 Analogue-based forecasts

It is generally accepted that NWP models better simulate synoptic (i.e. large scale) variables that control the atmosphere dynamic, such as geopotential heights, than local variables such as precipitation. Taking this into consideration, the principle of analogue methods is to bypass the thermodynamics part in NWP modelling. The main idea postulates that analogue synoptic situations should lead to similar local effects, the term *analogue* standing for states of the atmosphere which resemble each other closely [9].

The different steps can be described as follow. A *target* situation is given by a deterministic forecast from a NWP model. This situation is characterized by means of large scale predictor fields. Then, analogue situations are selected among *candidate* situations coming from an archive of reanalyses. This selection is made according to an analogy criterion applied over a given spatial domain. Finally, local precipitation having been recorded on these analogue dates are selected and provide an empirical probability distribution of the forecasted precipitation.

All analogue methods are based on the search of analogues, but they may vary depending on predictands, predictors and analogy criteria [26-29]. In this study, we applied the method operationally used at CNR and developed successively by [14, 26, 27, 30]. The predictand is the 6-hour precipitation amounts. Predictors are given by a NWP deterministic forecast at 0.5° spatial resolution. Three selection levels, based on the analogy of different predictors, are applied for the search of past analogue situations:

0. Selection of 9000 analogues based on temperature fields (T): this level aims to constrain the analogues dates to respect the seasonality. This step is considered as a preselection and therefore labelled as "level 0".
1. Selection of 175 analogues, among the 9000 previous, based on geopotential height fields (Z): such variables characterize the general circulation pattern.
2. Selection of 40 analogues, among the 175 previous, based on the precipitable water content, calculated as the product of the relative humidity (RH) and the total column water (TCW): it is a predictor describing the local air mass humidity.

More specifically, each selection level is defined by:

- The predictor(s): meteorological variable(s) taken at a given pressure level and a given lead time. For instance, if the predictand is the precipitation amount over the 12-18 UTC time step, "T500 (6h)" means the temperature taken at the 500 hPa pressure level at 18 UTC.
- The analogy criterion: a statistical score representing the "distance" between two gridded fields, and the spatial resolution at which it is computed. Depending on the analogy level, two different scores are used in this study: the RMSE and the Teweless-Wobus or S1 score [16].
- The spatial domain over which the analogy criterion is computed.
- The number of analogues being selected at the end of the selection level.

These characteristics are summarized in Table 3. When two predictors are specified for a single level, the analogy criterion is computed for both and the mean is considered.

The most the candidate situations found in the reanalyses archive are analogue to the target ones, the most the precipitation forecasts issued from these situations are expected to be reliable. Analogue methods therefore require a long common archive of reanalyses and precipitation observations. This is especially true for extreme precipitation forecasting, where driven atmospheric conditions leading to severe events are supposed to be rare. If only "mediocre" analogues (not so similar) are found in the archive, precipitation might be under-forecasted. The CNR's 6-hour precipitation observation dataset (cf. 3.1.), which covers a 23 year period, is here the limiting factor.

Another requirement is the consistency of target and candidate situations. At least, grid resolutions must match in order to calculate analogy criteria. Better, the NWP models must be the same for both forecasts and reanalyses, in order to find best analogies possible. However, the constant evolution of NWP models makes this constraint too restricting. In this study, we relaxed it considering the NWP model for forecasting and for reanalyses coming from the same meteorological centre. Differences in truncations, parametrization may exist, but the model core is assumed to be similar.

The two same meteorological centres than for ensemble forecasts were used: ECMWF and NCEP. Two different analogue-based forecast datasets were therefore produced. Thereafter, they will be referred as ECMWF-Ana and NCEP-Ana. As shown in Table 4, the respective reanalyse datasets used were ERA-Interim [31] and CFSR [32]. The respective deterministic forecasts used were the control forecasts of the aforementioned IFS-EPS and GEFS models. Reanalyses and control forecasts were all interpolated at the 0.5° spatial resolution. Spatial domains for levels 1 and 2 were optimized separately for each reanalyse dataset, in perfect-prognosis conditions (both target and candidate situations come from reanalyses).

Name	Reanalyses	Deterministic forecasts	members
ECMWF-Ana	ERA-Interim	IFS-EPS control	40
NCEP-Ana	CFSR	GEFS control	40

Table 4. Characteristics of analogue-based forecasts. The term *member* corresponds to the number of analogues finally obtained after the 3 selection levels.

Basin	Number of peak events in the evaluation sample	Amplitude in m^3/s (min-median-max)
Arve	123	165 - 229 - 633
Valserine	132	50 - 86 - 179
Usses	87	18 - 45 - 179
Séran	105	22 - 35 - 125
Guiers	122	55 - 83 - 291

Table 5. Characteristics of the evaluation sample for each basin.

4 Experimental plan

4.1. Temporal coherence of Analogue-based forecasts

Ensemble and analogue-based precipitation forecasts are both considered as probabilistic forecasts since the forecast for each lead time is made of an ensemble of *M* discrete values representing the uncertainty. However, in the ensemble approach, each value is associated to a specific run of the NWP model, and can consequently be linked to values for other lead times, leading to an ensemble of *M* temporal traces which are coherent. This is not the case with the analogue method, as the search for analogues is made independently for different lead times. In other words, there is no temporal traces issued from the analogue method. For hydrological application, such an ensemble of traces is required because each trace will correspond to a specific run of the rainfall-runoff model, leading to an ensemble of streamflow traces. A reordering of analogue-based forecasts is therefore necessary, the term *reordering* standing for the association of forecasted values for different lead times.

If no temporal correlation (or autocorrelation) would be observed, a random reordering could be sufficient. However, an autocorrelation pattern does exist on 6-hour precipitation. Spearman's rank correlation coefficient (*Rho*) for autocorrelation has been calculated on the observed precipitation observation dataset, after having taken away zero precipitation values. *Rho* coefficients for lag 1 (i.e. between two 6-hour time steps) were found between 0.3 and 0.5 depending on basins, and negative for higher lags. It means that positive precipitation values are correlated with the preceding and the following ones. With a random reordering of the forecasts, the probability for a high value (in the discrete distribution) at time step t to be associated with another high value at time step t+1 is low. This is nevertheless highly probable in reality, where successive high precipitation time steps may cause flood events.

As a consequence, the reordering has to respect the temporal correlation observed in precipitation observations. To address this issue, [33] proposed the *Schaake shuffle* reordering method, whose successive steps are briefly described below.

- For a forecast of *M* members to reorder, *M* precipitation time-series lasting 120 hours each are randomly selected among the observation dataset.
- For each 6-hour time step, these *M* observations are sorted and their rank noted. For example, one randomly selected time serie might be ranked 29th for time step 1, 34th for time step 2, etc. Equal values are treated randomly.
- Forecast values are likewise sorted, and the rank structure obtained above from observations is used to link forecast values: the forecast value being ranked 29th for time step 1 will be linked with the one being ranked 34th for time step 2, etc.

This way, an ensemble of traces is formed and the so-obtained precipitation forecasts respect the temporal rank dependence structure that has been observed in the past climatology. However, autocorrelation patterns may change for different atmospheric situations. With such method, observed time series are randomly selected, no matters whether their atmospheric situation is similar to the target one or not.

To address this shortcoming, we propose an adaptation of the Schaake shuffle where observed time series are no more randomly selected in the past archive but start with the dates having been selected with the analogue method for the first lead time. This method thereby ensures that observed time series are similar to the atmospheric initial situation. This *Analogue Schaake shuffle* was applied on ECMWF-Ana and NCEP-Ana forecasts over the 2007-2014 period.

4.2. A peak-based approach for the selection of events

The objective of this study was to assess the contribution of ensemble and analogue-based precipitation forecast on flood forecasting. As mentioned in the introduction, such evaluation involves two antagonistic statements. On the one hand, the evaluation of probabilistic forecasts requires large samples to be statistically robust. On the other hand, flood events are inherently rare. Decision was made to consider here all the events with a hydrological response exceeding a given streamflow threshold, taken for each basin as 95th percentile of observed streamflow. The choice of the percentile has been made arbitrarily, while nevertheless ensuring a sufficiently large sample of pairs forecast/reference. The evaluation period covered the period from 2007-03-25 to 2014-12-31. Peak occurrence dates were noted so as to set up a list of peak events on which the evaluation was made. Table 5 gives some characteristics of this sample for the different basins.

Hydrologic simulations were run at 06 UTC. The ARX rainfall-runoff model was fed with precipitation forecasts from the previous 00 UTC cycle in order to take into account the dissemination time of forecasts in an operational context. Each event was evaluated for 5 different prediction horizons: *D*, *D-1*, *D-2*, *D-3* and *D-4* i.e. when peak events were forecasted 0-24, 24-48, 48-72, 72-96 and 96-120 hours ahead, respectively. Figure 2 shows an example of a *D-2* streamflow forecast obtained with ECMWF-Ens forecasts as input.

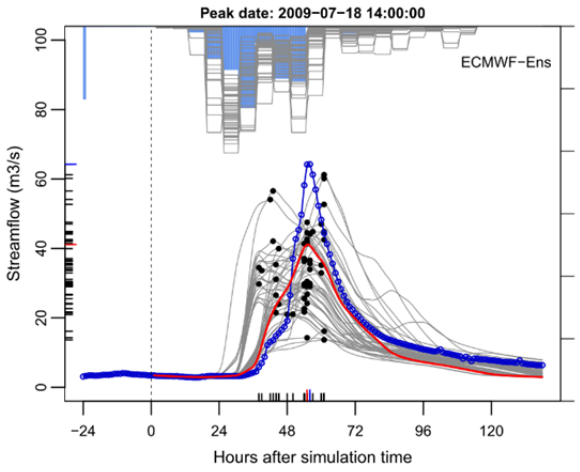

Figure 2. Example of a *D-2* streamflow forecast obtained with ECMWF-Ens precipitation forecasts. The peak event occurs the 2009-07-18 at 14 UTC (top title). The simulation date is the 2009-07-16 at 06 UTC (represented with a vertical dashed line). On top, observed (blue) and forecasted (grey) precipitation are plotted. At the bottom, observed (blue), simulated (red) and forecasted (grey) streamflow. Forecasted peaks are displayed with black points.

For each event and each prediction horizon, forecasted streamflow peaks were found and their amplitude (in m^3/s) and timing (in hours) noted for the evaluation. This operation worked as follow. For each trace of the forecast, the algorithm searched for all peaks in the simulation and kept the closest (in time) to the observed peak. As many peaks as members were therefore found for each streamflow forecast: 51, 21, 40 and 40 for ECMWF-Ens, NCEP-Ens, ECMWF-Ana and NCEP-Ana streamflow forecasts, respectively. These peaks are displayed with black points on the hydrographs as in Figure 2 and thereafter.

4.3. Evaluation framework

Evaluation aims to assess the quality of forecasts given a certain reference. In our experiment, simulated streamflow were first taken as reference. They correspond to the output of the model fed with observed precipitation. By doing so, rainfall-runoff model errors are cancelled out, enabling the assessment of the contribution of precipitation forecasts on streamflow forecasts. Then, observed streamflow were set up as reference so as to estimate the importance of the hydrological uncertainty.

Different attributes of the forecasts may be evaluated, depending on user's needs. In this study, no assumptions were made on potential user's needs, and the overall quality of forecasts was evaluated, for both the amplitude and the timing of peak events. The Continuous Ranked Probability Score (*CRPS*) was used:

$$CRPS = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{+\infty}\left[F_i(x) - H_{x_i^0}\right]^2 dx \qquad (2)$$

where N is the number of pairs forecast/observation in the evaluation sample, $F_i(x)$ is the cumulative distribution function of the i^{th} forecast, x_i^0 is the i^{th} reference value, expressed for calculations as the Heaviside function $H_{x_i^0}$ (equal to 0 for values below x_i^0 and to 1 for values above x_i^0). This score is a measure of the overall quality of forecasts in a sample. It is negatively oriented, meaning that smaller values correspond to better forecasts. For a given forecast (before being averaged over all forecasts), its value can be represented as shown on Figure 3 by the shaded grey area. The smaller this area, the smaller the *CRPS* and the better the forecast.

The *CRPS* has the same unit than the forecasted variable. It is thus expressed in m^3/s for the amplitude and in hours for the timing. However, it is sensitive to the order of magnitude of the variable being forecasted. For the amplitude of the peaks, this order of magnitude vary depending on basins (cf. Table 1), making spatial comparisons across basins difficult. As proposed by [35], the *CRPS* for the amplitude is normalized by σ, the standard deviation of the reference value for each basin over the evaluation period. Since there is a finite number of forecast/reference pairs in the evaluation sample, an uncertainty exists in the computation of the *CRPS*. If the sample is too short, the uncertainty is too large and makes any conclusions from the evaluation statistically non robust. The bootstrap method [36] was therefore applied in order to compute 5%-95% confidence intervals around each *CRPS* value. Such an interval corresponds to a range where the "true" value of the *CRPS* has 90% of chance to belong. The *CRPS* score, as a measure of the overall quality of forecasts, can be used as a criteria by

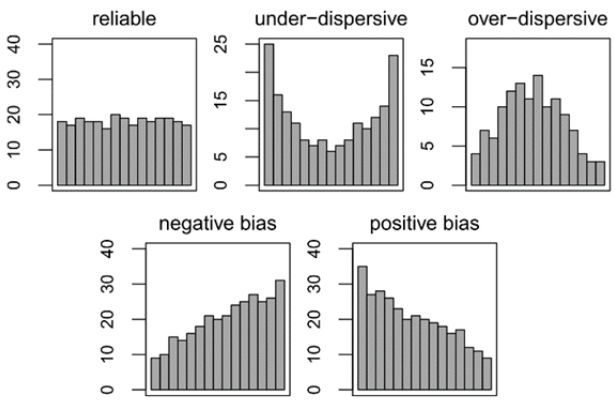

Figure 3. Graphical interpretation of rank histograms.

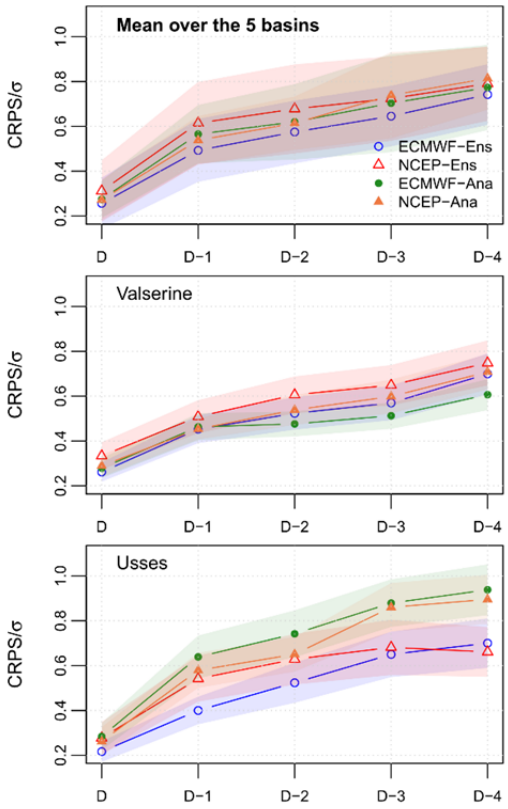

Figure 4. *CRPS/σ* (unitless) for the amplitude of peak events at different prediction horizons. Shaded areas represent 90% confidence intervals of *CRPS* values (computed with bootstrap, cf. 4.3).

forecasters when choosing the best forecasting system. However, such a single value doesn't show up the behaviour of the probabilistic forecasting system.

For that purpose, the rank histogram is a useful complementary tool, by helping the user to detect biases in probabilistic forecasts in a form of an ensemble of discrete values. Forecasts are said to be unbiased (or reliable) if all members are equally likely and statistically undistinguishable from the reference. In other words, the reference is expected to fall, over a long period, as many times in each interval between two member's values. A rank histogram computes, over the complete evaluation sample, the number of times (Y-axis) the reference falls in each interval (X-axis). As shown on Figure 3, reliable forecasts should show a flat histogram. Two types of bias can affect the reliability. If forecasts are under-dispersive (over-dispersive), reference values will be more likely to fall in the tails (centre) of the forecasted distributions and the histogram will show a U-shape (∩-shape). Moreover, if forecasts are negatively (positively) biased, reference values will be more likely to fall in the upper (lower) part of the forecasted distributions and the histogram will show a rising (decreasing) shape.

5 Results

Evaluation results for both peak amplitude and peak timing forecasts are plotted in a series of figures showing

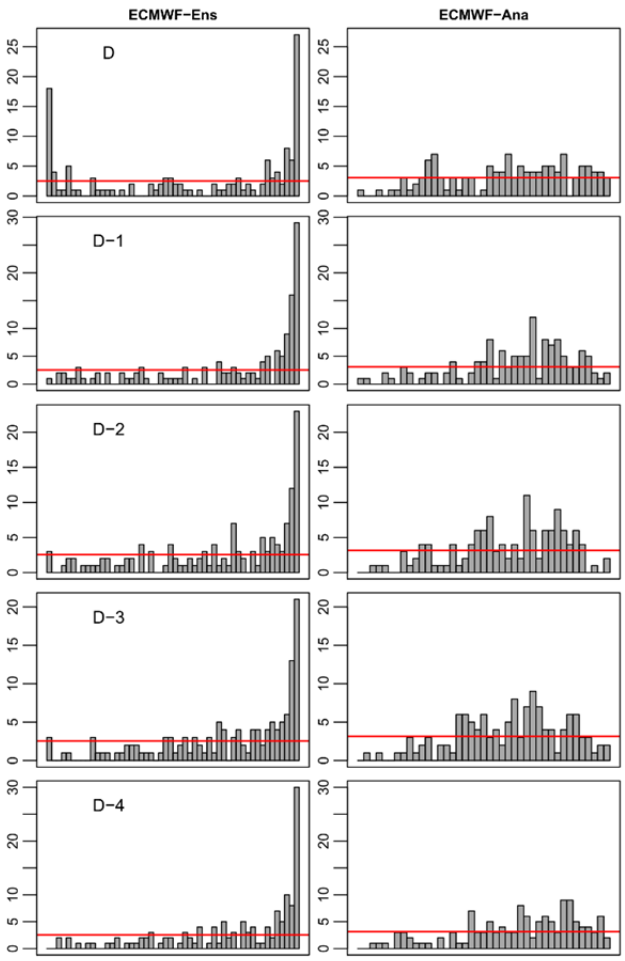

Figure 5. Rank histograms for the amplitude of peak events, for the Valserine basin. Left column corresponds to ECMWF-Ens streamflow forecasts, right to EMWF-Ana. Rows show different prediction horizons. The horizontal red line indicates a flat histogram (unbiased forecasts).

CRPS as a function of the prediction horizon, for the 4 considered precipitation forecasts: ECMWF-Ens, NCEP-Ens, ECMWF-Ana and NCEP-Ana. It is worth reminding that low *CRPS* values indicate better forecasting performances. Each *CRPS* value is surrounded by its 90% confidence interval. For peak amplitude forecasts, *CRPS* is normalized by σ, allowing comparisons across basins.

5.1. Evaluation against simulated streamflow

Results for peak amplitude are given in Figure 4. Upper panel shows the *CRPS*/σ averaged over the 5 study basins. Streamflow forecasts obtained with the 4 different precipitation forecasts exhibit very similar performances. By taking into account the uncertainty on *CRPS* values, it is not possible to state that one precipitation forecast dataset is clearly better than another for all basins. However, different patterns exist between basins. Middle and lower panel of Figure 4

5 show performances for the Valserine and the Usses basins. They were chosen as they exhibit the most different patterns among all basins. The Usses basin shows significantly lower performances for analogue-based streamflow forecasts compared to ensemble-based forecasts. On the Valserine basin, it appears that ECMWF-forced forecasts are better than NCEP-forced ones, regardless the type of method (ensemble or analogue) the forecasts come from. Although ECMWF-Ens and ECMWF-Ana present similar performances on the Valserine basin, notable differences exists when looking at the behaviour of the two forecast datasets. Figure 5 shows rank histograms relative to the forecasted Valserine streamflow peak amplitude. ECMWF-Ens forecasts lead to under-dispersive streamflow forecasts for all prediction horizons. Besides, except for short prediction horizons (*D*), they globally show a negative bias, with a very large number of under-estimated peaks. On the contrary, ECMWF-Ana tend to be over-dispersive for all prediction horizons. They also show a negative bias, but it is less visible due to the large spread of the forecasts. Same findings were found for other basins (not shown), as well as using NCEP-ENS and NCEP-Ana forecasts.

These behaviours can be illustrated by Figure 6 and Figure 7, which display respectively *D* and *D-2* streamflow forecasts for the highest peak event of the Valserine basin, for ECMWF-Ens and ECMWF-Ana forecasts. One can guess the negative bias of both types of forecast since peak amplitudes are under-estimated. Due to it large spread, though, ECMWF-Ana forecasts have several traces above the simulated peak, unlike ECMWF-Ens forecasts whose traces remain all below. It is worth reminding that such an interpretation on a single event aims to help in the understanding of forecasts behaviour, and cannot replace the evaluation of probabilistic forecasts over a long period, as it has been done with the *CRPS* or the rank histogram.

All forecasts see their performances deteriorating with increasing prediction horizons. A significant decrease of performance, though, exists between prediction horizons *D* and *D-1* because of the concentration time of the basins. Indeed, some *D* forecasts (but not all) are based on observed precipitation, meaning that their ensemble of traces are all very close to the simulated streamflow.

Results for peak timing are given in Figure 8. As for the peak amplitude, when considering all basins, the 4 streamflow forecast datasets show very close performances despite differences across basins. It appears, however, that for a given meteorological centre, ensemble-based forecasts are slightly better than analogue-based forecasts, especially for longer prediction horizons (*D-3* and *D-4*).

5.2. Evaluation against observed streamflow

As a reminder, the *CRPS* represents the error on streamflow coming from the meteorological part when simulated streamflow are taken as reference, while it represents to total error (from both meteorological and hydrological parts) when observed streamflow are taken as reference. Results of the evaluation against observed streamflow are shown in Figure 9 for the peak amplitude. Performances are significantly lowered, as expected since

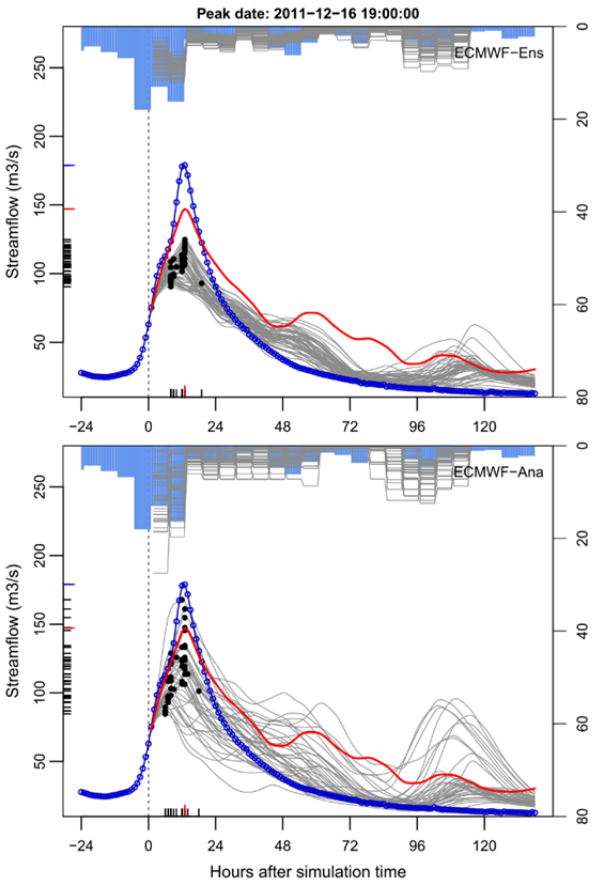

Figure 6. *D* forecast for the 2011-12-16 peak event. Upper panel: with ECMWF-Ens precipitation forecasts. Lower panel: with ECMWF-Ana.

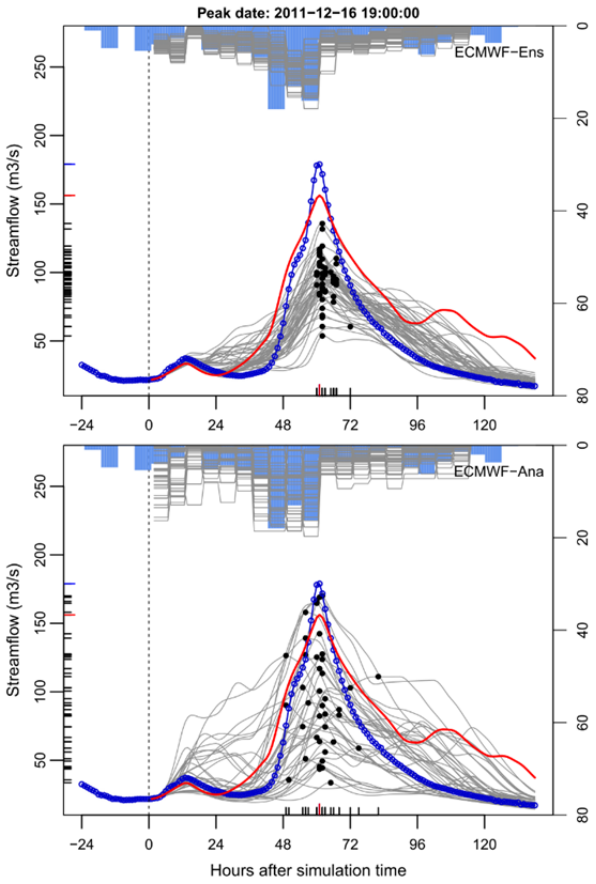

Figure 7. *D-2* forecast for the same peak event.

hydrological uncertainty is not taken into account. In the upper panel corresponding to the performances averaged over all basins, the error coming from the meteorological part is equal to 38-44 % (depending on the different forecasts considered) of the total error for the prediction horizon *D*. This percentage rises to 70-78 % for prediction horizons *D-4*, meaning that the uncertainty from precipitation become more and more important with increasing prediction horizons, regarding the total uncertainty.

It appears that the part of the error coming from the meteorological part, and consequently the one coming from the hydrological part, depends on basins as shown by the middle and the lower panel for the Valserine and the Usses basin, respectively. This can be partly explained as follow. Skill evaluation of the ARX rainfall-runoff models showed that simulated peak amplitude were in average equal to 69% of observed peak amplitude for the Valserine, and to 83% for the Usses basin. Both models thus tend to under-estimate peak events, but this bias is stronger for the Valserine model. As a consequence, since streamflow forecasts tend to also underestimate peak events, the less the ARX model underestimates, the smaller the error coming from the hydrological part. Another explanation is related to snow aspects, as will be discussed in next section.

Evaluation of peak timing forecasts against observed streamflow indicates a very small performance loss (not shown). It means that ARX rainfall-runoff models play a minor role in the timing of peak events compared to the role played by precipitation forecasts.

6 Discussion and perspectives

Questions arisen from these results were: what are the strengths and weaknesses of ensemble-based and analogue-based streamflow forecasts in a flood forecasting context? How can these findings help to improve precipitation forecasts for flood forecasting? If one manage to better estimate the uncertainty on precipitation forecasts, will it finally improve streamflow forecasts?

Forecasting performances of peak amplitude and peak timing were found very similar between ensemble-based and analogue-based streamflow forecasts, according to *CRPS* averaged over all basins. Behind this result, however, their behaviour appeared to be very different. Ensemble-based streamflow forecasts exhibited a smaller spread than analogue-based ones. This would be a strength if it was not to the detriment of reliability. Ensemble-based forecasts were indeed found to be under-dispersive for all prediction horizons, and negatively biased after *D-1*. This reflects the lack of variability of precipitation amounts in ensemble forecasts, partly due to the grid resolution of NWP models, which is sometimes

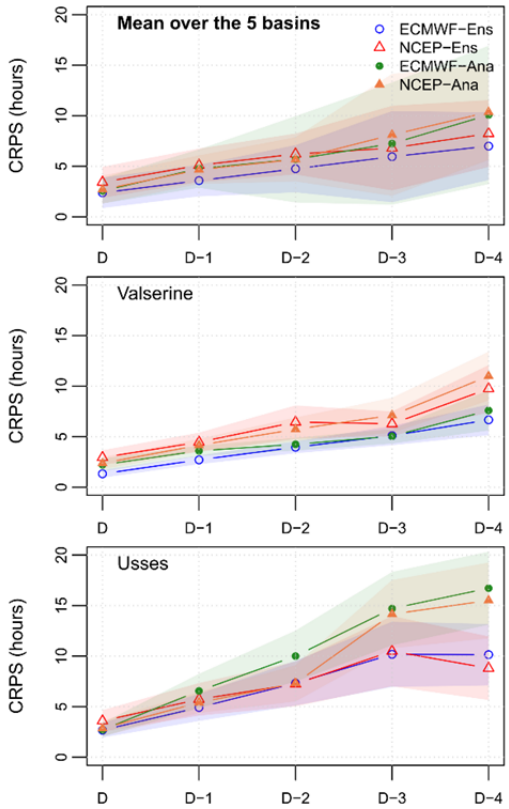

Figure 8. *CRPS* (in hours) for the difference of timing of peak events at different prediction horizons.

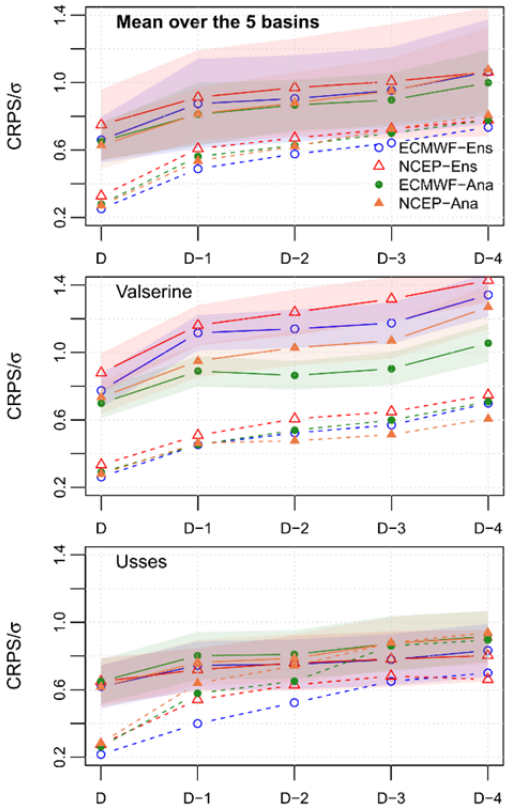

Figure 9. *CRPS/σ* for the amplitude of peak events. Solid lines (with 90% confidence interval) show the evaluation against observed streamflow, dashed lines show the evaluation against simulated streamflow (same as Figure 5).

smaller than the scale of physical processes responsible for precipitation. On the other hand, analogue-based forecasts were found more reliable thanks to a larger spread, despite their tendency to be over-dispersive. These findings are in accordance with results of [15], although these were obtained at daily time step. A negative bias was also observed, weaker though than the one of ensemble-based forecasts. The large spread of analogues improves their reliability, but as a consequence decision making may sometimes be difficult. Such an example is illustrated in the lower panel of Figure 8 where the forecast spread for the peak amplitude is very large. Perspectives towards the improvement of the analogue method (applied in this study) exist. First, the negative bias could be reduced by optimizing the parameters of the analogue method (analogy domains and numbers of selected analogues) over a number of high precipitation events instead of over the entire calibration period, as suggested by [28]. Another possible improvement is to consider the number of selected analogues no more as fixed, but depending on the "quality" of the analogy. It is worth noting that the optimal number of analogues, here 40, has been chosen according to a calibration over a long period but, taking flood events separately, a different number would have probably led to better results. In case of severe events, as in Figure 8 it is possible that "most analogue" members are closer to the simulated peak while "least analogues" correspond to low peaks. Such an assumption needs however to be studied in future works. This also requires to develop a "global" analogy criteria representing the quality of the analogy over all successive selection levels, which is not trivial. Nevertheless, this approach would be an interesting way for generating sharper and hence more helpful forecasts with the analogue method.

Another reason explaining the large spread of analogue-based streamflow forecasts is the application of the Schaake shuffle reordering method. Such reordering method tends to associate high precipitation values together for consecutive time steps, leading to a certain number of extreme streamflow traces after the rainfall-runoff modelling. These traces contribute to enlarge the spread on streamflow forecasts, and therefore may affect the quality of forecasts in terms of *CRPS*, but they are nevertheless highly informative for the forecaster. This raises the question whether a probabilistic score such as the *CRPS* is the most suitable metric for the evaluation of the quality of a forecast in a flood forecasting context.

Discussion of strengths and weaknesses of ensemble-based and analogue-based streamflow forecasts rises the possibility to combine both types of forecast. This approach has been already tackled by two previous studies. Authors in [15] disaggregated daily analogue-based precipitation forecasts with the chronology of 6-hour ensemble forecasts, with the idea that the analogue method provides more reliable forecasts of precipitation amounts, while ensemble forecasts provide a useful information about the temporal dynamic of rainfall events. This approach, however, uses ensemble forecasts only for their temporal dynamic. Precipitation forecasts

from ensemble forecasts have been found to be under-dispersive, but the dynamical evolution of the synoptic situation remains informative. The merging described by [13] follows this idea: the search for M_{ana} analogues is done on the M_{ens} synoptic situations given by the ensemble traces, with the objective to increase the spread of the ensemble forecast. This approach looks particularly interesting, as it benefits from the strengths of the analogue method for the generation of precipitation amounts given a synoptic situation, as well as the strengths of the ensemble forecasts for the uncertainty related to the synoptic situation. It therefore reduces the risk of using the analogue method over a synoptic situation which has been poorly forecasted by a deterministic NWP model. However, care must be taken not to generate over-dispersive forecasts. To address this, a non-fixed number of selected analogues, as introduced above, could be an interesting improvement.

As explained in paragraph 5.2., the percentage of the total error of peak amplitude forecasting coming from the meteorological part is important and increases with prediction horizons. Logic would therefore encourage to improve precipitation forecasts, by correcting their biases, in order to improve streamflow forecasts. However, as mentioned in the introduction, several studies have shown that this approach doesn't improve significantly streamflow forecasts [12, 13, 18]. By doing so, the percentage of the total error coming from the meteorological part would probably decrease, but it may introduce new errors in the hydrological modelling, leading to finally poor improvements on streamflow forecasts. They found that the bias-correction of streamflow forecasts is more effective. Such studies, though, used conceptual rainfall-runoff models. Further works are therefore necessary in order to determine if same findings can be found with statistical models such as ARX models. For flood forecasting, bias-corrected precipitation forecasts remain anyway necessary. One could merely bias-correct streamflow forecasts, but this single bias-correction step will necessarily compensate precipitation errors with hydrological errors. In case of a very rare and severe precipitation event, the forecasting system will not be able to properly correct the bias, and streamflow will be more likely to be under-forecasted.

In this study, aspects related to snow were not treated. In mountainous basins, snow melt generates more effective water (for runoff) and solid phase of precipitation cuts off part of the effective water. As a consequence, performance losses when evaluating forecasts against observed streamflow (cf. 5.2) were observed to be higher for the Arve and the Valserine basins, which are the two most mountainous basins. This highlights the importance of taking into account temperature forecasts for a better rainfall-runoff modelling.

The proposed methodology for the evaluation of different streamflow forecasts regarding peak amplitude and peak timing should be applied with different rainfall-runoff models. This would allow to verify whether findings about ensemble-based and analogue-based streamflow forecasts remain valid with rainfall-runoff models using conceptual or physical equations.

Moreover, evaluation should be extended to a larger number of basins. In this study, considering five different basins lets us observe that relative performances of ensemble-based and analogue-based streamflow forecasts vary depending on basins. However, there was not enough basins to establish robust links between these relative performances and the physical characteristics of the basins.

Acknowledgements

This work has been supported by a grant from Labex OSUG@2020 (Investissements d'avenir – ANR10 LABX56) and CNR. We also thank Sebastien Legrand whose comments on ARX models were very helpful.

References

1. Krzysztofowicz R. (2001). The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, **249(1–4)**, 2–9.
2. Ramos M. H., Van Andel S. J. and Pappenberger, F. (2013). Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, **17**, 2219–2232.
3. Roulin E. (2007). Skill and relative economic value of medium-range hydrological ensemble predictions. *Hydrology and Earth System Sciences*, **11(2)**, 725–737.
4. Houdant B. (2004). Contribution à l'amélioration de la prévision hydrométéorologique opérationnelle. Pour l'usage des probabilités dans la communication entre acteurs. *PhD thesis*, ENGREF (AgroParisTech).
5. Demeritt D., Nobert S., Cloke H. and Pappenberger F. (2010). Challenges in communicating and using ensembles in operational flood forecasting. *Meteorological Applications*, **17(2)**, 209–222
6. Thielen J., Bartholmes J., Ramos M.-H. and de Roo A. (2009). The European Flood Alert System – Part 1: Concept and development. *Hydrology and Earth System Sciences*, **13(2)**, 125–140.
7. Demargne J., Wu L., Regonda S. K., Brown J. D., Lee H., He M., et al. (2014). The Science of NOAA's Operational Hydrologic Ensemble Forecast Service. *Bulletin of the American Meteorological Society*, **95(1)**, 79–98.
8. Cloke H. L. and Pappenberger F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, **375(3–4)**, 613–626.
9. Lorenz E. N. (1969). Atmospheric Predictability as Revealed by Naturally Occurring Analogues. *Journal of the Atmospheric Sciences*, **26(4)**, 636–646.
10. Buizza R., Milleer M. and Palmer T. N. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **125(560)**, 2887–2908.

11. Bauer P., Thorpe A. and Brunet G. (2015). The quiet revolution of numerical weather prediction. *Nature*, **525(7567),** 47–55.
12. Verkade J. S., Brown J. D., Reggiani P. and Weerts A. H. (2013). Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology*, **501**, 73–91.
13. Zalachori I., Ramos M.-H., Garçon R., Mathevet T. and Gailhard J. (2012). Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. *Advances in Science and Research*, **8**, 135–141.
14. Obled C., Bontron G. and Garçon R. (2002). Quantitative precipitation forecasts: a statistical adaptation of model outputs through an analogues sorting approach. *Atmospheric Research*, **63(3–4)**, 303–324.
15. Marty R., Zin I. and Obled C. (2013). Sensitivity of hydrological ensemble forecasts to different sources and temporal resolutions of probabilistic quantitative precipitation forecasts: flash flood case studies in the Cévennes-Vivarais region (Southern France). *Hydrological Processes*, **27(1)**, 33–44.
16. Jolliffe I. T. and Stephenson D. B. (2011). *Forecast Verification: A Practitioner's Guide in Atmospheric Science (2nd Edition)*, Wiley.
17. Welles E., Sorooshian S., Carter G. and Olsen B. (2007). Hydrologic Verification: A Call for Action and Collaboration. *Bulletin of the American Meteorological Society*, **88(4)**, 503–511.
18. Kang T.-H., Kim Y.-O. and Hong I.-P. (2010). Comparison of pre- and post-processors for ensemble streamflow prediction. *Atmospheric Science Letters*, **11(2)**, 153–159.
19. Bartholomes J. C., Thielen J., Ramos M. H. and Gentilini S. (2009). The European Flood Alert System EFAS – Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrology and Earth System Sciences*, **13(2),** 141–53
20. Brown J. D., Demargne J., Seo D.-J. and Liu Y. (2010). The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environmental Modelling & Software*, **25(7)**, 854–872.
21. Zappa M., Fundel F. and Jaun, S. (2013). A 'Peak-Box' approach for supporting interpretation and verification of operational ensemble peak-flow forecasts. *Hydrological Processes*, **27(1)**, 117–131.
22. Box G. E. P., Jenkins G. M., Reinsel G. C. and Ljung G. M. (2015). *Time Series Analysis: Forecasting and Control*, John Wiley & Sons.
23. Bompart P., Bontron G., Celie S. and Haond M. (2009). Une chaîne opérationnelle de prévision hydrométéorologique pour les besoins de la production hydroélectrique de la CNR. *La Houille Blanche*, **5**, 54–60.
24. Leutbecher M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, **227(7)**, 3515–3539.
25. Park Y.-Y., Buizza R. and Leutbecher M. (2008). TIGGE: Preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society*, **134(637)**, 2029–2050.
26. Bontron G. and Obled C. (2005). A probabilistic adaptation of meteorological model outputs to hydrological forecasting. *La Houille Blanche*, **1**, 23-28.
27. Ben Daoud A., Sauquet E., Lang M., Bontron G. and Obled C. (2011). Precipitation forecasting through an analog sorting technique: a comparative study. *Advances in Geosciences*, 29, 103–107.
28. Marty R., Zin I., Obled C., Bontron G. and Djerboua A. (2012). Toward Real-Time Daily PQPF by an analog sorting approach: application to flash-flood catchments. *Journal of Applied Meteorology and Climatology*, 51, 505-520.
29. Chardon J., Hingray B., Favre A. C., Autin P., Gailhard J., Zin I., Obled C. (2014). Spatial Similarity and Transferability of Analog Dates for Precipitation Downscaling over France. *Journal of Climate*, **27(13)**, 5056-5074.
30. Ben Daoud A., Sauquet E., Bontron G., Obled C. and Lang M. (2016). Daily quantitative precipitation forecasts based on the analogue method: Improvements and application to a French large river basin. *Atmospheric Research*, **169**, 147–59.
31. Dee D. P., Uppala S. M., Simmons A. J., Berrisford P., Poli P., Kobayashi S., et al. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137(656)**, 553–597.
32. Saha S., Moorthi S., Pan H.-L., Wu X., Wang J., Nadiga, S. et al. (2010). The NCEP Climate Forecast System Reanalysis. *Bulletin of the American Meteorological Society*, **91(8)**, 1015–1057.
33. Clark M., Gangopadhyay S., Hay L., Rajagopalan B. and Wilby, R. (2004). The Schaake Shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, **5(1)**, 243–262
34. Bontron G. (2004). Prévision quantitative des précipitations : adaptation probabiliste par recherche d'analogues. Utilisation des Réanalyses NCEP/NCAR et application aux précipitations du Sud-Est de la France. *PhD thesis*, Institut National Polytechnique de Grenoble.
35. Trihn B. N., Thielen J., Thirel G. (2013). The reduction continuous rank probability score for evaluating discharge forecasts from hydrological ensemble prediction systems. *Atmospheric Science Letters*, **14(2)**, 61-65.
36. Efron B. and Tibshirani R. J. (1994). *An Introduction to the Bootstrap*, CRC Press.