# Strategies for hydrologic ensemble generation and calibration: On the merits of using model-based predictors

Anne-Laure Tiberi-Wadier, Nicole Goutal, Sophie Ricci, Philippe Sergent, Maxime Taillardat, François Bouttier, Céline Monteil

# Strategies for hydrologic ensemble generation and calibration: on the merits of using model-based predictors

Anne-Laure Tiberi-Wadier[a,*], Nicole Goutal[b], Sophie Ricci[c], Philippe Sergent[d], Maxime Taillardat[e], François Bouttier[f], Céline Monteil[g]

[a] *Cerema Eau Mer et Fleuves, Plouzané, France*
[b] *Saint-Venant Hydraulics laboratory, EDF R&D, Chatou, France*
[c] *CECI, CNRS UMR 5318/CERFACS, Toulouse, France*
[d] *Cerema Eau Mer et Fleuves, Margny-Lès-Compiègne, France*
[e] *Météo-France, Toulouse, France*
[f] *CNRM, Toulouse University, Météo-France and CNRS, Toulouse, France*
[g] *EDF R&D LNHE - Laboratoire National d'Hydraulique et Environnement, Chatou, France*

**Abstract**

This paper investigates the hydrometeorological chain with an ensemble approach. The objective is the generation of Hydrologic Ensemble Forecasts (HEF) on the Odet catchment (France, Brittany), using the Quantile Regression Forest (QRF) method usually applied for the ensemble calibration of meteorological forecats. First, a Global Sensitivity Analysis (GSA) in the distributed MORDOR-TS model is carried out taking into account uncertainty in forecasted rain with AromeEPS-RR1 and in model parameters. GSA highlights the role and importance of the different hydrologic model parameters during rain events and allows to only take into account the most influent parameters for the generation of an Hydrologic Ensemble Forecast (HEF).

Three strategies for the generation of HEF are then compared. First (i), a raw ensemble is built with a model-based only approach using the deterministic forecast rainfall Expert-RR3 and perturbed model parameters, without further statistical calibration. Then, the QRF calibration method is used to generate two ensembles of quantiles: (ii) the observation-based approach uses only predic-

*Corresponding author
*Email address:* Anne-Laure.Tiberi-Wadier@cerema.fr (Anne-Laure Tiberi-Wadier)

tors that are independent from hydrology, whereas (iii) the combined model and observation approach combines these predictors with statistics of the raw hydrologic ensemble (mean, standard deviation). This latter approach was shown to outperfom the previous ones, enhancing the importance of the choice of the predictors in the QRF method. In the prospect of using the hydrologic ensemble as input for hydraulic simulation, the Ensemble Copula Coupling method (ECC) and a trajectory smoothing procedure is then applied on (iii). This step slightly deteriorates the reliability of hourly streamflows, yet Continuous Ranked Probability Score (CRPS) and forecast skills on the cumulated or maximum streamflows are improved.

*Keywords:* Ensemble Prediction System, Global Sensitivity Analysis, Hydrologic Ensemble Forecast, Quantile Regression Forest, MORDOR-TS, Odet catchment

*2010 MSC:* 00-01, 99-00

---

## 1. Introduction

In spite of the great efforts that have been put into land surface models coupled to hydrologic models, deterministic simulation and forecast of streamflow remains limited mostly due to the stochastic nature of precipitation and the complexity of meteorological and hydrologic processes. Thus an ensemble approach is favored; it provides a probabilistic hydrologic forecast needed by decision support systems dealing with risk-based stakes in real-time (low and high flow) as well as by water resources management actors (hydropower production, irrigation, navigation, tourism). Forcing hydraulic models with forecasted hydrologic inflow allows to extend forecast lead time at stations where security and production are at stake. Before use for application, the performance of the ensemble forecast should be assessed (Jolliffe and Stephenson, 2003) in terms of reliability and resolution, with respect to observations. The reliability refers to the statistical consistency between the ensemble and the observations. The resolution describes the ability of the ensemble to discriminate situations leading or not to an event.

### *1.1. Hydrometeorological uncertainties*

The objective of ensemble forecasts is to cover and represent the uncertainties existing in the prediction chain. Numerous and various sources of uncertainties along the hydrometeorological simulation chain lead to uncertainty in discharge simulation and forecast. Three main sources of hydrometeorological uncertainties are acknowledged in the literature (Bourgin, 2014; Zappa et al., 2011; Thiboult et al., 2016; Demargne et al., 2014): atmospheric forcing observation and prediction, hydrologic model initial condition, hydrologic model structure and parameters. According to the classification proposed by Krzysztofowicz (1999), atmospheric forcing is referred to as input uncertainty, whereas the other sources are referred to as hydrologic uncertainties.

On the one hand, input uncertainty, i.e meteorological uncertainty, can be accounted for by an Ensemble Prediction System (EPS), which accounts for

uncertainties in initial conditions and model physics in Numerical Weather Prediction (NWP) models. The most straightforward technique to issue Hydrologic Ensemble Forecasts (HEF) is to use EPS as input for a hydrologic model, thus producing an ensemble of discharges. Numerous studies and operational applications have been conducted on this topic (Cloke and Pappenberger (2009),Pappenberger et al. (2016)). On the other hand, hydrologic uncertainty is also taken into account to issue HEF considering model parameters uncertainties Dietrich et al. (2009) or using a multi-model approach (Hopson and Webster, 2010; Velázquez et al., 2011; Thiboult et al., 2016; Bellier, 2018). Since 2004, research, operational and user communities gathered around the HEPEX initiative (Hydrologic Ensemble Prediction EXperiment, www.hepex.org), which aims at advancing the science and practice of hydrological ensemble prediction and demonstrating their utiliy in decision making (Thielen et al., 2008; Schaake et al., 2010).

### *1.2. Uncertainty quantification for Hydrologic Ensemble Forecasts (HEF) generation*

The performance of the raw hydrologic ensemble is linked to how the different sources of uncertainties are accounted for in the ensemble generation. The HEF system should be built taking into account major sources of uncertainties (both atmospheric and hydrologic) with associated ranges of uncertainty. To that end, identifying and ranking sources of uncertainty is necessary. Given assumptions on uncertain parameters, this can be achieved with a sensitivity analysis (SA). A review of SA methods is available in Iooss and Lemaître (2015).

Global Sensitivity Analysis (GSA) (Saltelli, 2008) allows to quantify the contribution of model inputs to its outputs. It implies the integration of an ensemble of simulations from which sensitivity indices are computed; for instance Sobol indices based on variance-based methods (Efron and Stein, 1981). GSA is widely applied in hydrology. In Emery et al. (2016), a GSA is achieved in order to highlight the key parameters impacting the river-routing scheme Total Runoff Integrating Pathways (TRIP) that simulates river water height and discharge on

the Amazon catchment. Garambois et al. (2013) analyzed MARINE hydrologic model sensitivity during flash-floods. In Michon and Castaings (2017), a GSA on the Kling-Gupta Efficiency (KGE, Gupta et al. (2009)) of the MORDOR-TS hydrologic model was carried out on several French catchments.

### *1.3. Statistical calibration of ensembles*

Ensemble approaches aim at representing the uncertainty along a simulation chain. However, the total uncertainty is rarely fully covered and raw EPS are known to be underdispersive and biased (Hamill and Colucci, 1997; Schaake et al., 2010). A post-processing step on the ensemble precipitation forecast (post-processing with respect to meteorology, pre-processing with respect to hydrology) should thus be achieved before using them as input for rainfall-runoff models and issuing HEF. Similarly, HEF should be post-processed to account for uncertainty in EPS as well as in hydrologic model parameters. The improvement of EPSs' and HEFs' reliability and resolution relies on a statistical post-processing, named ensemble calibration.

Statistical ensemble calibration is an active field of research in meteorology and hydrology. It exploits the relation between the past previsions and their corresponding observations (Wilks, 2018) to correct the forecast. Parametric and non-parametric calibration methods are reported in the literature. Parametric methods rely on an a priori assumption for the output data distribution which parameters are identified by the calibration algorithm. In contrast, non-parametric methods are data-based only. Wilks (2018) describes the state of the art of statistical postprocessing of meteorological ensemble forecasts. A review of statistical ensemble calibration methods used in the field of hydrology is available in Li et al. (2017).

Two widely used parametric methods are Ensemble Model Output Statistics (EMOS that is a regression method) (Gneiting et al., 2005) and the Bayesian Model Averaging method (BMA that is an ensemble dressing method) (Raftery et al., 2005). Both EMOS and BMA provide the entire predictive distribution for the output variable. In the field of EPS calibration, EMOS technique was used

in Taillardat et al. (2016, 2019); Bellier (2018) for calibration of temperature and precipitation ensemble forecast. In the field of ensemble streamflow calibration, EMOS was used in Bellier (2018); Hemri et al. (2015), and BMA was used in Duan et al. (2007).

Popular non-parametric methods are: rank histogram recalibration (Hamill and Colucci, 1997), quantile regression (Bremnes, 2004), individual ensemble-member adjustments (Van Schaeybroeck and Vannitsem, 2015) or statistical learning methods, also called machine learning methods (Hastie et al., 2009). While non-parametric methods require very large training data sets, they are flexible, data-adaptive, and adapted to non-linearities in the input-output relation (Wilks, 2018). The non-parametric Quantile Regression Forest (QRF) method proposed by Meinshausen (2006) is a statistical learning approach. QRF provides an estimation of desired quantiles for the output data, but not the entire distribution, as opposed to EMOS or BMA.

### *1.4. Scope of the paper*

The objective of this paper is the implementation and assessment of an HEF system for small to medium size catchments taking into account hydrologic model parameters' uncertainty. The first part of the study is dedicated to the analysis and classification of uncertainties in the distributed MORDOR-TS model (Garçon, 1996; Garavaglia et al., 2017; Rouhier et al., 2017) with GSA in order to identify the most significant sources of uncertainties to take into account into the ensemble generation. The GSA is carried out with respect to uncertainty in hydrologic model parameters and EPS using forecasted precipitation provided by Arome Ensemble Prediction System [1] (Seity et al., 2011; Bouttier et al., 2012; Raynaud and Bouttier, 2016; Bouttier et al., 2016), when available.

The second part of the study is dedicated to the HEF generation and en-

[1] This product, denoted by AromeEPS-RR1 in the following, provides a forecast of 1-hour rainfall cumul, with a maximum lead-time of 45 hours, updated every day.

semble calibration. Different strategies are compared : a model-based approach where the raw ensemble comes from ensemble hydrologic simulation is implemented, then two approaches are implemented with QRF ensemble calibration method, with different predictors.

Figure 1 summarizes the general process flow of the study. The two main objectives of the study correspond to the grey boxes.

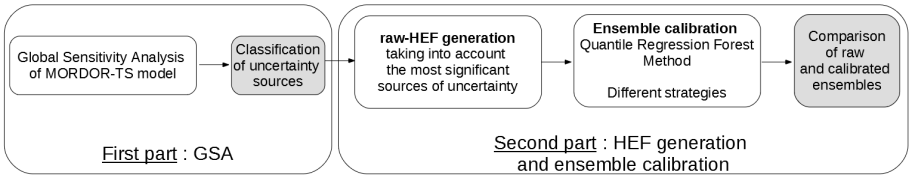



Figure 1: General process flow of the study.

The paper is organized as follows. Section 2 presents the Odet catchment on which the study is carried out, the hydrologic model as well as the atmospheric and hydrologic data sets. Methods for GSA and QRF are detailed in section 3. Section 4 presents the experimental settings for the GSA and the framework for the generation and ensemble calibration of the HEF. Section 5.1 discusses the results of the GSA for the MORDOR-TS hydrologic model and section 5.2 presents the comparative skills of the raw and calibrated hydrologic ensembles. Conclusion and perspectives are given in Section 6.

## 2. Material: models and data

### *2.1. Modeling the hydrology of the Odet catchment with MORDOR-TS*

#### *2.1.1. The Odet catchment*

The Odet river shown in Fig. 2 is a French coastal river located in Western Brittany. It flows through the city of Quimper, then South to the sea. Its mouth is located at Plaisance where astronomical tide ranges between 1.40 m and 5.55 m. The Jet and Steir rivers are two tributaries of the Odet river. The Odet catchment area is 720 km$^2$ and the Odet river is of about 60 km long. Quimper is often subject to flooding in urban areas, resulting from the combination of two phenomena: excessive rainfall that contribute to the rivers, and high tides increased by storm surges impacting the flow up to Quimper. The distributed MORDOR-TS rainfall-runoff model is built on three upstream sub-catchments (Fig. 2, Tab. 1), with outlets at respectively Tréodet (Odet river), Kerjean (Jet river) and Ty-Planche (Steir river).

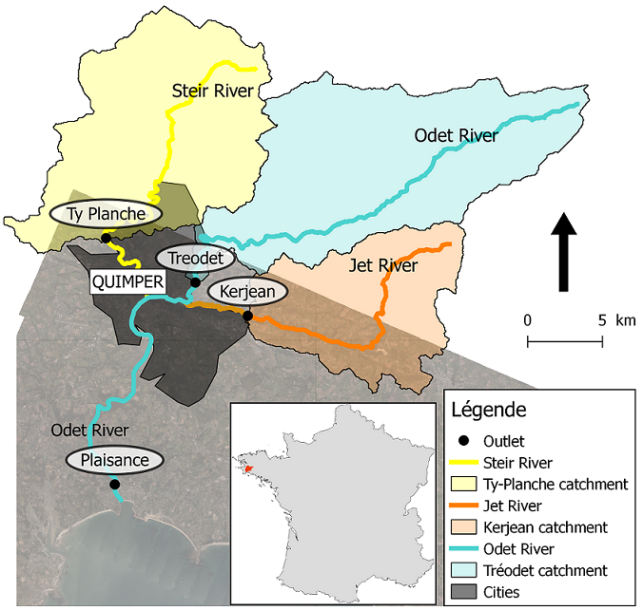



Figure 2: Odet catchment

Table 1: Characteristics of the Tréodet, Kerjean, Ty-Planche sub-catchments

| Sub-catchment | Tréodet | Kerjean | Ty-Planche |
|---|---|---|---|
| Elevation of the source | 175 m | 200 m | 100 m |
| Total river length (km) | 37 | 21 | 23 |
| Catchment area ($km^2$) | 205 | 107 | 179 |
| Mean flow ($m^3/s$) | 4.8 | 2.27 | 3.79 |
| 10-year flow ($m^3/s$) | 55 | 19 | 39 |
| 50-year flow ($m^3/s$) | 75 | 25 | 53 |
| Highest flow recorded 12-2000 ($m^3/s$) | 110 | 46.6 | 81 |
| Second highest flow recorded 12-2013 ($m^3/s$) | 91.5 | 17.6 | 42.7 |
| Mean rainfall (mm/yr) | 743 | 672 | 671 |

*2.1.2. The MORDOR-TS distributed conceptual rainfall-runoff model*

The MORDOR-TS model (Garçon, 1996; Garavaglia et al., 2017; Rouhier et al., 2017) dedicated to water resource management is implemented on each sub-catchment of the Odet catchment. MORDOR-TS is a spatialized and continuous conceptual rainfall-runoff hydrologic model that connects the mesh cells according to the hydrographic network. At each time step, the production is calculated for each cell and then routed to simulation points on the mesh. The structure of the production module is presented in Fig. 3 and MORDOR-TS's hydrologic parameters are described in Tab. 2. The production module takes spatially distributed precipitation (P) and temperature (T) as input data and adjusts the water balance through two coefficients $c_p$ and $k_{min}$. This latter parameter is involved in the calculation of the actual evapostranspiration $AET$. The production module is then composed of six conceptual reservoirs; two for ice and snow (not active here), and four others: a superficial reservoir U of capacity $U_{max}$, an intermediate reservoir L of capacity $L_{max}$ of which filling level is driven by the parameter $ev_L$, an evaporating reservoir Z of capacity $Z_{max}$ and a deep reservoir N of which filling level is driven by the parameter $lk_N$. Three fluxes components are transfered from the production module to the routing module: area runoff $Q_s$, subarea runoff $Q_v$ and base runoff $Q_b$. The parameter

$k_r$ determines the ratio of the water feeding reservoir N and the subarea runoff. The routing module propagates the water production of each cell into the hydrographic network. The transfer function is based on the 1D diffusive wave model, with celerity $Cel$ and diffusion $Diff$ coefficients independent from the runoff (Hayami, 1951).

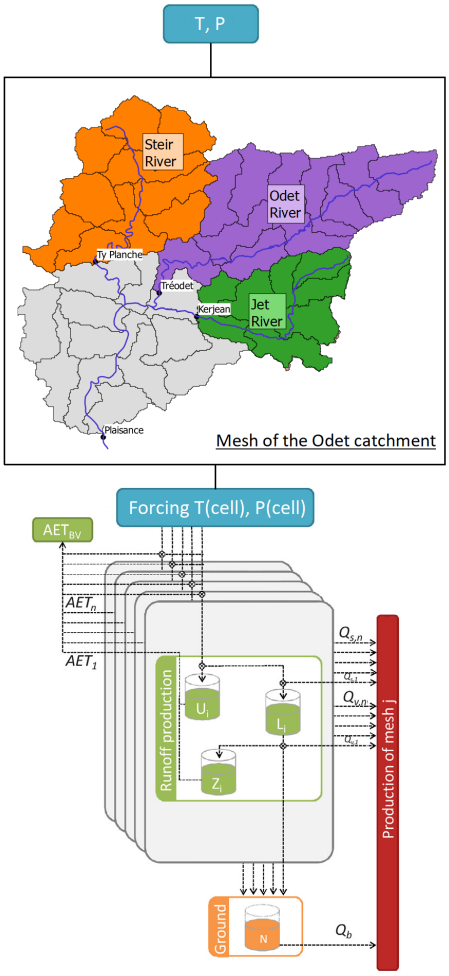



Figure 3: Structure of the MORDOR-TS model over the Odet catchment. Adapted from Rouhier et al. (2017)

Table 2: Description of the parameters of MORDOR-TS hydrologic model

| Parameter | Unit | Description | Module |
|---|---|---|---|
| $c_p$ | - | Precipitation multiplicative correction factor | Water balance |
| $k_{min}$ | - | Maximum seasonal crop coefficient | |
| $U_{max}$ | mm | Maximum capacity of the root zone U | Runoff production |
| $L_{max}$ | mm | Maximum capacity of the hillslope zone L | |
| $ev_L$ | - | Outflow exponent of storage L | |
| $Z_{max}$ | mm | Maximum capacity of the capillarity storage Z | |
| $k_r$ | - | Runoff coefficient | |
| $lk_N$ | $mm.h^{-1}$ | Outflow coefficient of storage N | |
| $Cel$ | $m.s^{-1}$ | Wave celerity | Routing module |
| $Diff$ | $m^2.s^{-1}$ | Wave diffusion | |

The hydrologic calibration [2] of the 10 previously described parameters of MORDOR-TS (Tab. 2) is achieved after a one-year spinup, using Banque Hydro streamflow observations, ANTILOPE rainfall and SAFRAN surface temperature forcing described in Sect. 2.2. The hydrologic calibration is carried out with respect to a multi-objective function using the caRamel genetic algorithm (Le Moine et al., 2015; Monteil et al., 2019). The multi-objective function gathers three scores : (i) Nash-Sutcliffe Efficiency (NSE, Nash and Sutcliffe (1970)) over the entire time serie, (ii) NSE over the inter-annual daily regime and (iii) NSE over the empirical cumulative distribution.

The hydrologic calibration is first achieved over a 6-year period (01/01/2007-12/31/2013) and validated over the period 01/01/2014-05/31/2017. Over the validation period, scores values are of the same order of magnitude as over the calibration period, which validates the hydrologic calibration. In the following, the deterministic simulation with calibrated hydrologic parameters is referred to

[2] In this paper, the term *calibration* refers either to the calibration of the parameters of the hydrologic model MORDOR-TS or to the statistical ensemble calibration with the Quantile Regression Forest (QRf) method. To avoid confusion, the calibration of the hydrologic parameters is always denoted as *hydrologic calibration*. The statistical calibration with QRF is denoted as *ensemble calibration* when necessary.

as the *reference simulation*. In a second time, a set of hydrologic calibrations is achieved over 2-year periods from 01/01/2007 to 05/31/2017 in order to estimate the uncertainty in each parameter, required to perform the GSA and to generate the raw HEF (methodology fully described in Sect. 4.2.1).

### *2.2. Data sets*

Depending on their availability, different data sets are used in the study.

#### *2.2.1. Observed data*

Spatially distributed observed rainfall, temperature data and flow discharge are available at hourly time step from January 2006 to May 2017 on the Odet catchment.

- Rain data: The ANTILOPE rainfall product is a combination of radar and gauge rainfall data from Météo-France (Champeaux et al., 2009). The grid resolution is 1 km.
- Temperature data: Surface temperature is extracted from SAFRAN reanalysis (Vidal et al., 2010). The grid resolution is 8 km.
- Discharge data: River flow discharge time series at Tréodet, Kerjean and Ty-Planche are extracted from the French national archive (Banque hydro, http://www.hydro.eaufrance.fr, Leleu, Isabelle et al. (2014)).

#### *2.2.2. Forecasted data*

AromeEPS-RR1 ensemble rain forecast product is used to describe uncertainty in rainfall for the GSA. As this product is available over a limited period only, the deterministic Expert-RR3 rain forecast product is used by debault, in particular for the generation and the ensemble-calibration of the HEF.

*AromeEPS-RR1.* AromeEPS-RR1 uses the regional atmospheric model AROME described in Seity et al. (2011) with ensemble perturbations documented in Bouttier et al. (2012); Raynaud and Bouttier (2016); Bouttier et al. (2016), accounting for uncertainties in initial conditions, boundary conditions, surface

conditions and the model physics. Its large scale boundary conditions are provided by the global PEARP ensemble (Descamps et al., 2015). AromeEPS-RR1 ensemble is composed of 12 equiprobable members and covers a 1800x1700km $^2$ Western European domain that encompasses the Odet catchment with a horizontal grid at 2.5 km resolution. AromeEPS-RR1 is operational since the end of 2016. A limited period of 112 days in early 2016 was made available *a posteriori* for this study in order to cover former rain events on the Odet catchment. It provides a daily forecast at 21:00 UTC, with hourly output over a 45-hour forecast range.

*Expert-RR3.* Expert-RR3[3] is a 3-hours deterministic rainfall accumulation forecast specified by human experts on the basis of numerical forecasts from atmospheric models, with a 72 hours forecast range. These data are used operationally for flood forecasting in the French governmental services and available over 2011-2014 for this study.

[3]RR3 provide a forecast of 3-hour rainfall cumul, with a maximum lead-time of 72 hours, updated every 15 minutes

## 3. Methods

### *3.1. Global Sensitivity Analysis (GSA)*

In the present study, the GSA stands in the computation of Sobol' indices (Sobol, 2001), with the assumption that the input aleatory variables are independant. Sobol' indices apportion the variance of the output $Y = f(X)$ with $X = (X_1, X_2, ..., X_k)$, to the variation of different inputs $X_1, ..., X_k$ on their uncertainty domain. With the assumption that the variance of Y is finite and the input variables are independant, the Hoeffing decomposition (Hoeffding, 1948) leads to the definition of the Sobol'indices :

$$1 = \sum_i S_i + \sum_i \sum_{j>i} S_{i,j} + ... + S_{1,2,3,...,k} \tag{1}$$

where

- $S_i = \frac{V_i}{V(Y)}$ is the first order Sobol index of $X_i$ and represents the normalized elementary contribution of $X_i$ to V(Y),
- $S_{i,j} = \frac{V_{i,j}}{V(Y)}$ is the second order Sobol index of $X_i$ and $X_j$ and represent the normalized contribution due to interactions between $X_i$ et $X_j$ to V(Y),

and so on.

The total Sobol' index $S_{T_i}$, gathering all contributions related to $X_i$ is then defined as

$$S_{T_i} = S_i + \sum_{j \neq i} S_{i,j} + \sum_{j \neq i, k \neq i, j<k} S_{i,j,k} + ... + S_{1,2,3,...,k} = \sum_{l \in \sharp i} S_l \tag{2}$$

where $\sharp i$ are all the subsets of $\{1, ..., k\}$ including i.

Sobol' indices thus measure the influence of the different independant imputs $X_1, X_2, ..., X_k$ and their interactions on the output $Y = f(X)$. The more sensitive the model response is to an input parameter, the larger its associated Sobol index. Usually only first and total Sobol indices are computed. The first order indices are useful for factor prioritization (FP) and provides the parameter(s) that most significantly control(s) the most the output variance, whereas

the total order indices are hepfull in factor fixing (FF), to determine which parameters can be fixed without consequences on the output (Saltelli, 2008). In practice, the estimation of Sobol' indices is generally achieved with a stochastic estimation using an ensemble of model output realizations (Saltelli and Annoni, 2010). The computation of Sobol' indices is here achieved using the python modules OpenTURNS (http://openturns.org/) and OT-Batman (T.Roy et al., 2018).

### *3.2. Quantile Regression Forest (QRF)*

*Non parametric regression with QRF method.* The Quantile Regression Forest (QRF) method (Meinshausen (2006)) technique is detailed in Zamo et al. (2014); Taillardat et al. (2019). The principle of the method is the aggregation of meteorological or hydrological situations according to their forecasts, with the assumption that close forecasts predictors lead to close observations. In that way, this method can be linked to the analog method (Hamill and Whitaker, 2006; Zalachori et al., 2012; Delle Monache et al., 2013). QRF is a non-parameric and non-linear regression, which consists in building random forests from binary decision trees given a set of predictors (Breiman, 2001). Contrary to random forests that approximate the conditional mean, QRF estimates the full conditional distribution of the response variable.

*Chosing the predictors.* A wide range of predictors are available for the QRF method. They can obviously stem from the ensemble forecast : statistics (ensemble mean, variance or percentiles) of the variable to calibrate or other variables within the ensemble, as well as other characteristics of the forecast, such as the time or day (providing a modality instead of a value). Predictors can also be described from real observation of the forecasted variable. The choice of the predictors represents a key element in the implementation of the QRF calibration method. In the following, a learning sample is formed by chosing a set of predictors and picking the associated observations.

*Building a binary decision tree.* A binary decision tree is built by iteratively splitting the learning sample into two groups. For a quantitative predictor, the split is done according to a threshold, while for a qualitative predictor, the split is done according to the modality. The predictor and the splitting criteria are chosen to minimize the variability of the associated observations in the resulting two groups. Each resulting group is then itself split in two following the same algorithm, until a stopping criterion is reached (for example a minimum number of data in the sub-groups, or an insufficient decrease of variance). Each final group is called leaf and contains a set of observations, also called predictand. The splitting algorithm is illustrated in Fig. 4 with two predictors p1 and p2 ranging between 0 to 1. In this example, the tree has three leaves with associated observations (rain or discharge for our purpose).

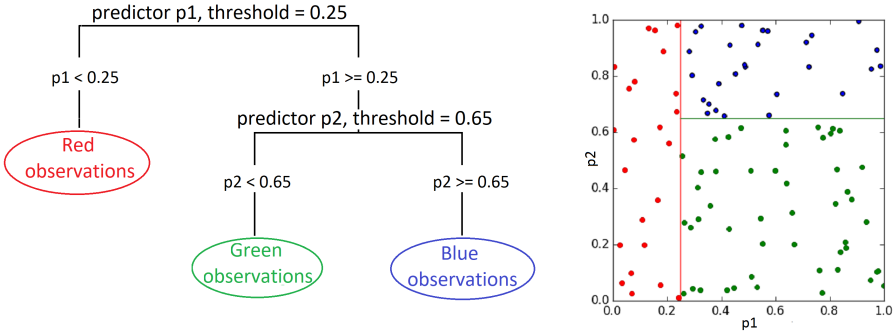



Figure 4: Illustration of binary decision tree with 2 predictors p1 and p2.

*Building a forest.* Breiman (1996) proposes to improve the robustness of the prediction issued from a decision tree by selecting different learning samples to build several trees and form a forest. Since this approach would require a large amount of data, usually not available in practice, bootstrap sampling is often used. Randomly choosing a subset of predictors for each split of each tree, as suggested by Breiman (2001), enhances the independence of the trees and consequently forms a *random* forest.

*How to use QRF to calibrate a new ensemble forecast ?.* In operational mode, when a new ensemble forecast is available for calibration, its associated predictors are computed and run through each tree of the previously constructed forest. The terminal leaf predictand values are used to estimate the predictive CDF and quantiles of the calibrated forecast.

*Merits and limitation of the QRF method.* It should be noted that no a priori assumption is made on the distribution of the variable to calibrate neither before nor after calibration. Also, since the terminal leaves are composed of observations, the ensemble calibration is bound to output physically consistent values (for example, no negative amount of precipitation). The associated drawback is that by construction, the predicted CDF is unable to predict values outside of the observation range within the learning sample. This may be limiting when dealing with extreme values but may be overcome when working with anomalies or fitting a parametric function to the CDF (Taillardat et al., 2019). It should finally be noted that one of the main drawback of QRF method stands in the need of a large data set.

*Reconstuction of the calibrated members.* In the following, the Ensemble Copula Coupling (ECC, Schefzik et al. (2013)) method is used to reorder the post-processed quantiles to recreate time-series, but this method yields unrealistic jumps when applied to HEF streamflows. The Trajectory Smoothing (TS) procedure proposed by Bellier (2018) was thus applied to preserve temporal correlation consistency, in the perspective of using the ensemble of reconstructed streamflow time series as forcing to hydraulic ensemble simulations. It should be noted that, as opposed to Ensemble Copula Coupling, the Trajectory Smoothing procedure modifies the post-processed ensemble.

*Ensemble evaluation metrics.* Various tools are available to evaluate probabilistic forecasts and are well described in the literature. In this study, two widely used verification measures for ensemble forecast are used. The Continuous Ranked Probability Score (Matheson and Winkler (1976), Hersbach (2000),

Gneiting and Raftery (2007)) assesses reliability and resolution simultaneously. It is negatively oriented: the lower the better. The Rank Histogram (Talagrand et al. (1997), Hamill and Colucci (1997), Anderson (1996)) is useful to assess reliability only. A reliable ensemble implies that each rank is filled with the same probability, so the rank histogram is flat. Candille and Talagrand (2005), then Delle Monache et al. (2006) and Taillardat et al. (2016) introduced the notion of norms of a rank histogram with K+1 ranks. In particular, $\|.\|_2$ can be defined as :

$$\|.\|_2 = \sqrt{\sum_{i=1}^{K+1}(f_i - \frac{1}{K+1})^2} \tag{3}$$

where $f_i$ represents the frequency of observations in the $i$th rank. For a perfectly reliable ensemble system, $\|.\|_2$ is equal to zero.

## 4. Experimental settings

### *4.1. General ensemble workflow*

The first part of the study consists in a Global Sensitivity Analysis (GSA) applied to the distributed MORDOR-TS model, taking into account (i) uncertainties in rainfall and hydrologic model parameters (GSA-Arome), or (ii) only uncertainty in hydrologic model parameters (GSA-hydro). This last GSA without uncertainty in rainfall aims at assessing the consistency of the results when a deterministic rainfall (ANTILOPE) is taken into account in place of an EPS (AromeEPS-RR1). The implementations of the two GSAs are described in Sect. 4.2. It should be noted that the use of AromeEPS-RR1 for GSA-Arome is preceded by a QRF ensemble-calibration step. These two GSAs studies validate the methodology used for HEF generation in the second part of the study: hypothesis on the choice of uncertain variables, associated statistical distributions and related hyper parameters.

The second part of the study is dedicated to the HEF generation and calibration over a three-year (2011-2014) period over which there was significant hydrological events. Only uncertainty in hydrologic parameters is considered since the AromeEPS-RR1 product is not available over this period. The raw HEF ensemble is then calibrated with the QRF method.

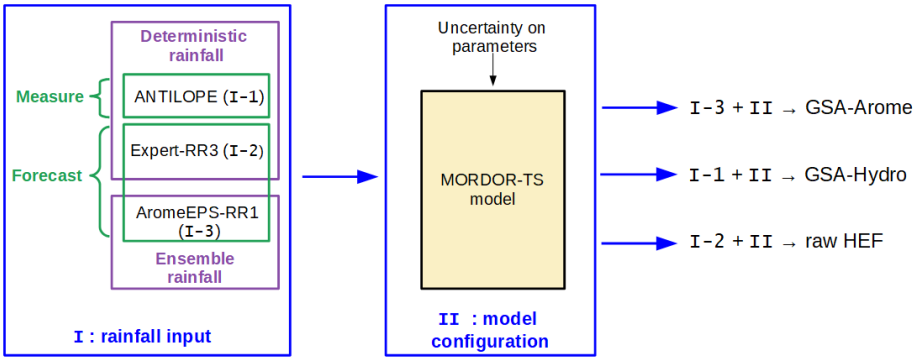



Figure 5: Configurations for GSA-Arome, GSA-hydro and raw HEF

Fig. 5 presents the different model settings for the three hydrologic raw ensembles built in the study: GSA-Arome, GSA-hydro and raw HEF. Tab 3 presents the required data needed for each step of the study :

- The GSA-Arome requires the integration of an ensemble of hydrology simulations over $P_{GSA-Arome}$ from 12/24/2015 to 03/15/2016. The atmospheric forcing is provided by AromeEPS-RR1 ensemble calibrated with QRF against ANTILOPE rainfall observed data.

- The GSA-hydro is carried out over two periods. It requires the integration of an ensemble of hydrology simulations over the period $P_{GSA-hydro}$ that is either $P_{GSA-Arome}$ or a sub period of $P_{HEF}$ (12/23/2013-12/26/2013). The deterministic atmospheric forcing is provided by ANTILOPE for both periods.

- The raw HEF generation is achieved using atmopsheric forcing from the Météo-France Expert-RR3 deterministic forecast (as AromeEPS-RR1 is not available) over $P_{HEF}$ (09/01/2011-06/01/2014). HEF ensemble streamflows are calibrated with QRF against Banque Hydro streamflow observations, using ANTILOPE observations, Expert-RR3 rain forecast and Banque Hydro streamflow observations as predictors.

In both GSAs and HEF, hydrology simulations start from a spin-up forced by ANTILOPE and SAFRAN observed data.

### *4.2. GSAs for MORDOR-TS model*

The GSA-Arome and GSA-hydro for streamflow on the Odet catchment are carried out with respect to the forecasted runoff at the outlet of each of the subcatchments Tréodet, Kerjean and Ty-Planche. Both GSA take into account uncertainties that relate to a set of 10 parameters for MORDOR-TS (Sect. 4.2.1). GSA-Arome (Sect. 4.2.2) is performed for each lead-time with a cycled procedure and takes also into account uncertainties that relate to rain forcing, considering an integer that represents the index within the 12-member

Table 3: Data for the GSA and QRF calibration of rainfall and streamflow

| | Type | Name | QRF AromeEPS-RR1 | GSA-Arome | GSA-hydro | QRF HEF |
|---|---|---|---|---|---|---|
| | | Period | $P_{GSA-Arome}$ | $P_{GSA-Arome}$ | $P_{GSA-hydro}$ | $P_{HEF}$ |
| Observed data | rainfall | ANTILOPE | X | X | X | X |
| | temperature | SAFRAN | | X | X | X |
| | streamflow | BanqueHydro | | | | X |
| Forecasted data | rainfall | AromeEPS-RR1 | X | X | | |
| | rainfall | Expert-RR3 | | | | X |

calibrated AromeEPS-RR1 ensemble, drawn from a uniform distribution $U_{[1,12]}$. In GSA-hydro (Sect. 4.2.3), ANTILOPE deterministic rain is used to force all model runs within the ensemble while hydrologic parameters vary.

### *4.2.1. Hydrologic parameters' distributions*

The 10 MORDOR-TS parameters are drawn from uniform distributions with $V_{min}$ and $V_{max}$ extreme values, shown in Tab 4 and determined from a set of 2-year of hydrologic calibrations. The lower bound (respectively upper bound) of the uniform distribution is chosen as the minimum (respectively maximum) of the values reached in the different hydrologic calibrations (Sect. 2.1.2).

Table 4: Extreme values for MORDOR-TS model parameters' uniform PDFs

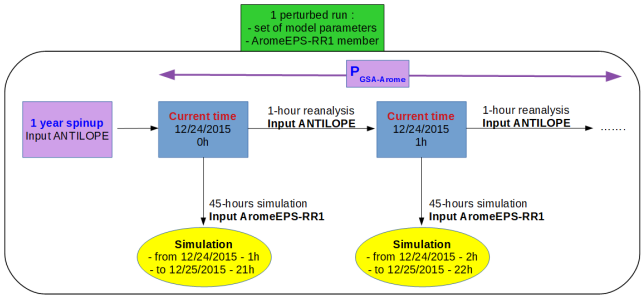
| Paramètre | Tréodet min | Tréodet max | Kerjean min | Kerjean max | Ty-Planche min | Ty-Planche max |
|---|---|---|---|---|---|---|
| $c_p$ (-) | 0.98 | 1.21 | 0.97 | 1.13 | 0.99 | 1.18 |
| $k_{min}$ (-) | 0.18 | 1.11 | 0.14 | 1.03 | 0.32 | 0.79 |
| $U_{max}$ (mm) | 75 | 196 | 30 | 111 | 39 | 116 |
| $L_{max}$ (mm) | 64 | 153 | 30 | 493 | 74 | 298 |
| $ev_L$ (-) | 3.95 | 4.00 | 1.50 | 3.98 | 3.38 | 4.00 |
| $Z_max$ (mm) | 47 | 138 | 77 | 256 | 72 | 482 |
| $k_r$ (-) | 0.10 | 0.28 | 0.10 | 0.41 | 0.10 | 0.30 |
| $lk_N$ ($mm.h^{-1}$) | -5.8 | -5.2 | -6.2 | -5.5 | -5.7 | -5.4 |
| $Cel$ ($km.h^{-1}$) | 0.44 | 0.56 | 0.45 | 1.10 | 0.50 | 0.55 |
| $Diff$ ($km^2.h^{-1}$) | 1 | 659 | 1 | 5000 | 329 | 731 |

#### *4.2.2. GSA-Arome implementation*

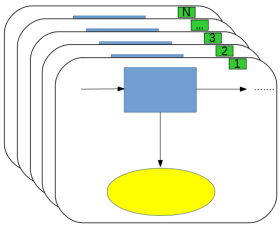
*Ensemble calibration of AromeEPS-RR1.* AromeEPS-RR1 ensemble is calibrated with the QRF method before its use in GSA-Arome. In order to increase the volume of data available for AromeEPS-RR1 ensemble calibration over the limited period $P_{GSA-Arome}$ in 2016, the calibration is carried out without discriminating the lead-times. The ensemble gathers all AROME cells within a subcatchment (63, 36 and 51 for Tréodet, Kerjean and Ty-Planche respectively); thus forming a new ensemble with (12*number of cells) members. The size of the learning sample is 45 (lead-times) $\times$112 (days). The ensemble is calibrated against the ANTILOPE observations, averaged over the whole subcatchment. For each subcatchment, the predictors are the percentiles 10, 50 and 90 of the raw AromeEPS-RR1 hourly rainfall, surface humidity, surface temperature and the moment of the day of the lead-time (morning, afternoon, evening, night). The evaluation of the calibration is achieved with a leave-one out method: each element of the training sample is alternatively used for validation, while the other are used for learning. This AromeEPS-RR1 calibration strategy relies on the fact that the error in rainfall intensity is homogeneous over a small catchment, and that it only depends on the moment of the day. The loss of predictability as the lead time increases is assumed to be negligible. It is also assumed that the rainfall intensity data are non correlated over time. This assumption would not be valid for temperature or streamflow. For rainfall, a 6h correlation is suspected and could be taken into account to improve the robustness of the calibration, especially if used to calibrate new data.

*GSA-Arome.* The GSA-Arome study is carried out in an operational framework with hourly updated AromeEPS-RR1 forecast. For that purpose, the ensemble is built as shown in Fig. 6(a) over a spin-up, a re-analysis and a forecast period. Each ensemble member is associated to a realisation of the set of hydrologic parameters and of AromeEPS-RR1. Each hydrologic simulation is run over a 1-year spin-up and a 1-hour re-analysis with ANTILOPE rainfall, then run over a 45-hour forecast AromeEPS-RR1 rainfall. This is cycled hourly as the rain

product is updated hourly. The GSA-Arome indices are computed hourly for the 45 lead times over the $N$ cycled simulations (Fig. 6(b)). Here, to ensure the convergence of the Sobol' indices estimation, the GSA-Arome study is achieved with an ensemble of 192000 MORDOR-TS simulations.




(a) 1 cycled run




(b) N cycled runs

Figure 6: Cycled runs for GSA-Arome

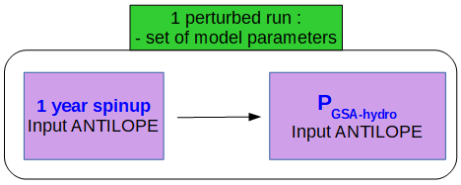



Figure 7: Continuous runs for GSA-hydro

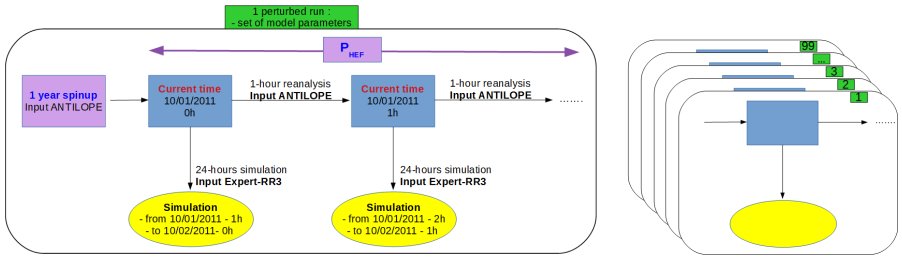
#### *4.2.3. GSA-hydro implementation*

The GSA-hydro study uses the observed rainfall ANTILOPE as input for every member, thus the notion of lead-time is no longer relevant. Each member is initiated with a 1-year spin-up, then integrated over $P_{GSA-hydro}$ using its associated set of hydrologic parameters as illustrated in Fig. 7. To ensure the convergence of the Sobol' indices estimation, the GSA-hydro Sobol indices are computed hourly over 154000 MORDOR-TS simulations.

### *4.3. Generation and ensemble-calibration of the HEF*

It was not possible to apply the QRF calibration on the ensemble of streamflows from the MORDOR-TS simulations that were used for the GSA-Arome over $P_{GSA-Arome}$. Indeed, this period is too short, the learning sample for the QRF method is thus too small, especially given that it should be further reduced to account for temporal correlation of streamflow. Another ensemble was thus generated, over a longer period $P_{HEF}$ (2011-2014) over which significant hydrological events occured, but without considering uncertainty related to the rainfall since the AromeEPS-RR1 product was not available over this period. Expert-RR3 deterministic forecast rain was used to generate the raw HEF. The HEF generation and calibration is achieved with a hydrologic ensemble of 99 members. The raw HEF is generated in an operational framework with hourly updated RR3 forecast. For that purpose, the ensemble is built as shown in Fig. 8(a), similarly to that for GSA-Arome over a spin-up, a re-analysis and a forecast period, except that the deterministic forecast Expert-RR3 is used in place of AromeEPS-RR1. This is cycled hourly as the RR3 forecast rain product is updated several times per hour. Only the predominant parameters that were previously identified by the GSA are taken into account for the ensemble generation of raw HEF. The forecasts are calibrated and evaluated against observations over 27 months, from October 2011 to June 2014, excluding summer months (July, August and September). A cross-validation method is used: each month of the calibration period is alternatively used for validation, while the 26 other months are used for learning. It should be noted that data over a 10-day

period before and after the validation month was removed from the learning sample to avoid auto-correlation between learning and validation data.




(a) 1 cycled run (b) 99 cycled runs

Figure 8: Cycled runs for HEF

*Ensemble calibration strategies.* Different strategies to generate hydrologic ensembles are implemented as shown in Fig. 9. The model-based-only approach relies on the integration of MORDOR-TS forced by observed ANTILOPE and forecast RR3 rain product with perturbed hydrologic parameters; this leads to the raw ensemble denoted as raw HEF. Then, two approaches are implemented with QRF calibration method. QRF aims at constructing a relation between chosen predictors and corresponding observations, which further allows to estimate desired quantiles of the observed quantity. All available predictors for QRF are presented in Tab 5. In the observation-based approach, predictors for the QRF method use no information from the hydrologic simulations. Strictly speaking, this is not a calibration strategy as the raw ensemble is not used; QRF is used to generate calibrated quantiles of an observed variable. This leads to the ensemble denoted as QRF-nothydro quantiles. In the combined model-observation approach, predictors for QRF calibration uses the mean and standard deviation of the raw ensemble. The resulting ensemble is denoted as QRF-hydro quantiles. After applying the Ensemble Copula Coupling (ECC) method to reorder the calibrated QRF-hydro quantiles, the Trajectory Smoothing (TS) procedure was finally applied, leading to QRF-hydro-TS ensemble.

The impact of smoothing the reconstructed hydrologic time-series from QRF quantiles is here investigated in the prospect of using the hydrologic ensemble as input for hydraulic simulation. It should be noted that the ECC method can not be applied on QRF-nothydro quantiles, since it requires the availability of the raw ensemble as dependence template.

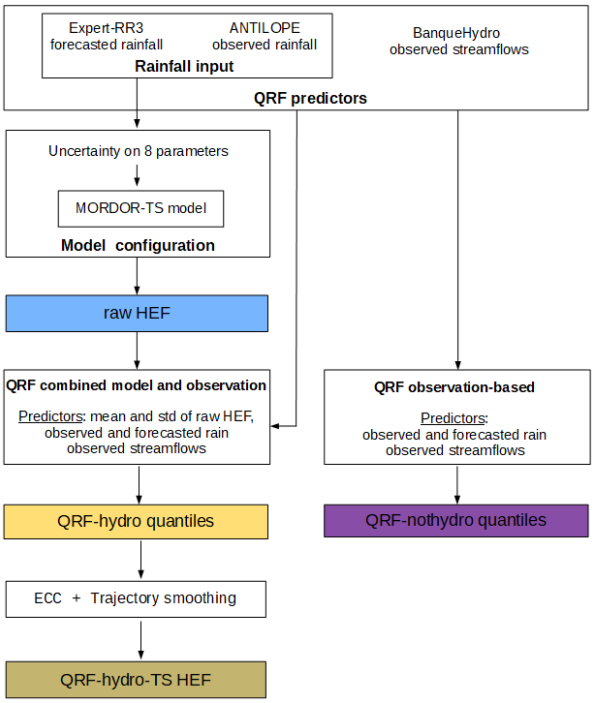



Figure 9: Strategies for Hydrologic Ensemble Forecast generation. Data used for hydrologic model integration as well as QRF predictors are shown. The resulting four HEFs or ensemble of quantiles are indicated in colored boxes.

*Computation of CRPS and RH for HEF assessment.* HEF assessment is first achieved with univariate criteria with CRPS and RH computed for streamflow over each catchment and for each lead-time. As suggested by Bellier (2018), cumulated and maximum quantities are also considered, and such quantities are denoted as global variables in the following. CRPS and RH are thus computed for the cumulated and maximum streamflow that respectively relate to

Table 5: Choice of predictors for QRF-hydro and QRF-nothydro HEF calibration strategies.

| Name | Unit | Description | QRF-hydro | QRF-nothydro |
|---|---|---|---|---|
| *Mean* | $m^3/s$ | mean of raw ensemble streamflows | X | |
| *Sigma* | $m^3/s$ | standard deviation of raw ensemble streamflows | X | |
| *Month* | | month of the validation time | X | X |
| *Period* | | period of validation time (0 am - 6 am, 6 am - 12 am, 0 pm - 6 pm, 6 pm - 12 pm) | X | X |
| $Q_0$ | $m^3/s$ | measured streamflow at the current time | X | X |
| $GradQ_0$ | $m^3$ | gradient of the measured streamflow at the current time | X | X |
| $M_{rain}$ | mm | measured rain over the catchment during the $N_b$ past hours of re-analysis ($N_b$ = 24 hours - lead-time) | X | X |
| $F_{rain}$ | mm | forecasted rain over the catchment between the current time and the lead-time | X | X |

volume and peak flow (Hemri et al., 2015). Following conclusions from Bellier et al. (2017) and Gneiting and Ranjan (2011), the HEF is assessed with a forecast-based stratification for CRPS and RH computation, focusing on the $(forecast, observation)$ pairs for which the maximum of the ensemble is beyond the $90^{\text{th}}$ percentile of the associated observation (hourly, cumulated or maximum measured streamflow), computed for heavy rainfall.

## 5. Results

### 5.1. GSAs for MORDOR-TS model

#### 5.1.1. Calibration of AromeEPS-RR1

As shown in Fig. 10, ensemble calibration improves AromeEPS-RR1 with smaller mean CRPS value and flatter rank histograms than that of the raw ensemble (0.119 and 0.1351 respectively for the CRPS). The $\|.\|_2$ norm for the QRF rank histogram ($1.11 * 10^{-2}$) is smaller than that of the raw ensemble ($3.58 * 10^{-2}$). The calibrated AromeEPS-RR1 data are thus more consistent with observations than before calibration, and thus can be used as input of GSA-Arome.

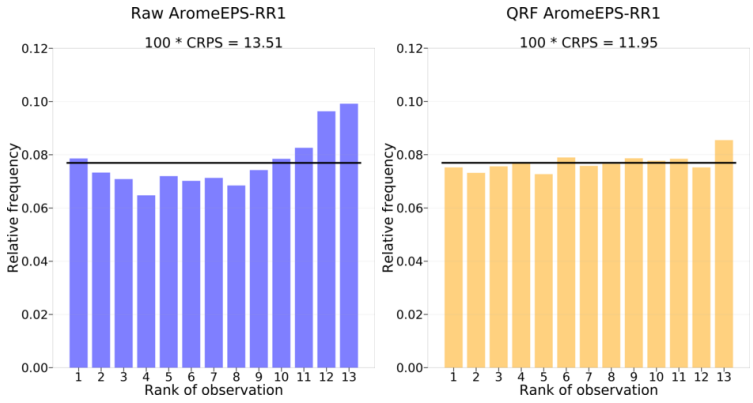
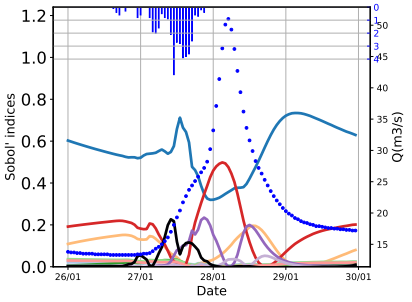


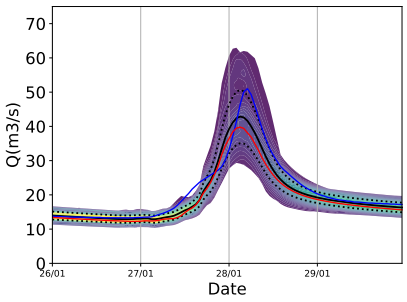

Figure 10: Rank histogram and CRPS for the raw and calibrated AromeEPS-RR1 ensemble.

#### 5.1.2. GSA-Arome Results

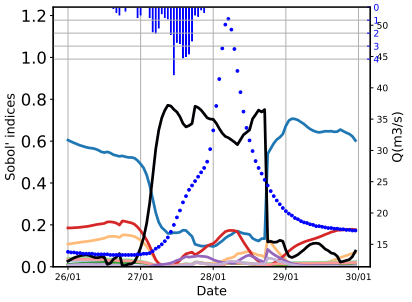
Results for the hydrologic GSA-Arome on forecasted streamflow are shown here over the subperiod 01/26/2016-01/30/2016 of $P_{GSA-Arome}$ for 6-hour and 21-hour lead-times. Time-varying Sobol' indices and streamflow hydrographs at Tréodet are shown in Fig. 11. A vertical section of Fig. 11(b) and Fig. 11(d) displays the probability density function at the given date. Similar results for Kerjean and Ty-Planche can be found as supplementary material in Section 7 (Fig. 17 and 18).



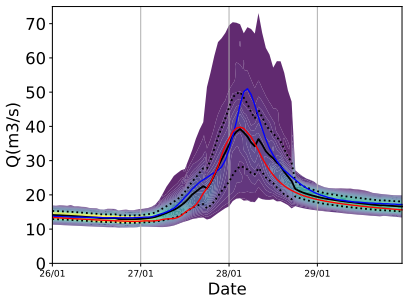

(a) Sobol'indices - 6h lead-time



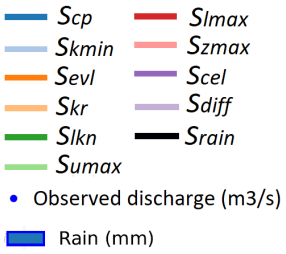

(b) Streamflow hydrographs - 6h lead-time



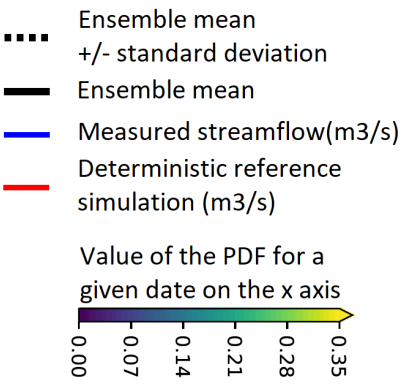

(c) Sobol'indices - 21h lead-time




(d) Streamflow hydrographs - 21h lead-time




(e) Legend for MORDOR-TS Sobol'indices




(f) Legend for streamflow graphs

Figure 11: First order Sobol' indices and associated hydrographs for the event between 01/26/2016 and 01/30/2016 at Tréodet catchment for GSA-Arome. A vertical section of Fig(b) and Fig(d) displays the probability density function at the given date, according to the legend shown in (f). The results are shown for 6-hour ((a) and (b)) and 21-hour ((c) and (d)) lead-times.

*Sobol' indices.* The Sobol' indices time-series for GSA-Arome are presented in Fig. 11(a) and 11(c) for 6-hour and 21-hour lead times with time on the x-axis. The observed discharge is represented by the blue dotted curve (right y-axis) and the observed rainfall histogram is represented at the top of the panel. It shows that, for both lead times, the simulated streamflow is dominated by the precipitation correcting factor $c_p$ (dark blue curve). This predominance decreases in favor of the choice of the AromeEPS-RR1 rain scenario when heavy precipitations occur, especially for longer lead-times, since this choice strongly influences the total amount of water in the MORDOR-TS reservoirs. Streamflow is thus mostly dominated by $c_p$ and rain that both control the water balance of the model. Rainfall mostly matters for lead-times that are greater than the concentration time of the catchment. This implies that the construction of HEF for short lead times can be achieved only taking into account uncertainties that relate to hydrology. To a lesser extent, at the beginning of the event, the streamflow is influenced by the capacity of the reservoir $L_{max}$ (red curve) that directly feeds the runoff. The importance of $L_{max}$ decreases before the end of the event; indeed, when the reservoir is full, the overflow directly feeds the runoff and its capacity no longer has impact on the simulated streamflow. For 6-hour lead-time, the wave celerity $Cel$ (dark purple curve) has noticeable importance when the streamflow gradient is strong (both in increase and decrease phases). The runoff coefficient $k_r$ (light orange curve) has noticeable importance at the peak of the event (and beyond). Indeed, when reservoir $Z$ is full, part of the excess amount of water $k_r \times out_Z$ directly feeds the runoff. Finally, simulated streamflow is not sensitive to the two parameters $k_{min}$ and $Diff$ over the three catchments.

*Temporal PDF.* The streamflow probability density functions represented over time for GSA-Arome are displayed in 11(b) and 11(d) for 6-hour and 21-hour lead times respectively, with time on the x-axis. The observed discharge is plotted in blue, the determinisitic reference simulation with parameters issued from the hydrologic calibration presented in Section 2.1.2 is plotted in red and the

MORDOR-TS ensemble mean and mean plus or minus standard deviation are plotted in black (solid and dotted lines). The reference simulation clearly underestimates the flood peak (except for the Kerjean catchment). The ensemble probability density function underestimates the flood peak since the measured streamflow is closer to the upper dotted black line in Fig. 11 (ensemble mean + standard deviation) than to the mean curve; neither the perturbation of the hydrologic model parameters nor that of the rain scenario allows to overcome this effect.

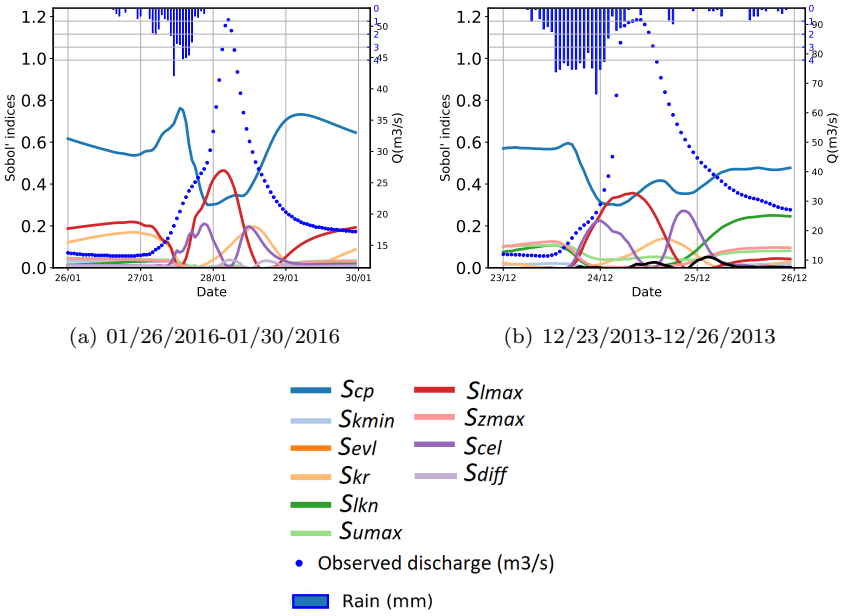
#### *5.1.3. Consistency between GSA-Arome and GSA-hydro results*

GSA-hydro was carried out over the subperiod of $P_{GSA-Arome}$ (01/26/2016-01/30/2016) and over a subperiod of $P_{HEF}$ (12/23/2013-12/26/2013) and associated Sobol' indices are shown in Fig. 12 at Tréodet. Results at Kerjean and Ty-Planche can be found in Fig. 19 and 20 as supplementary material in section 7.

For the 2016 event, GSA-hydro (Fig. 12(a)) and GSA-Arome (Fig. 11(a)) at short lead time, for which uncertainty in rainfall is not significant, show similar results. This allows to rely on GSA analysis when rainfall uncertainty is neglected for HEF construction at short forecast lead time. Moreover, GSA-hydro over $P_{GSA-Arome}$ (Fig. 12(a)) and $P_{HEF}$ (Fig. 12(b)) show similar results since these two events correspond to similar weather conditions. This allows to further the study with HEF construction over $P_{HEF}$. Finally, as previously stated over the three catchments, the two parameters $k_{min}$ and $Diff$ have very low Sobol' indices meaning that these two parameters have thus no impact on the simulated runoff. These parameters were thus fixed to their hydrologic calibration value for HEF generation in the following while the other parameters are drawn from uniform distribution described in Tab. 4.

### *5.2. Calibration of the HEF*

Ensemble reliability and resolution are assessed with univariate and global RH and CRPS metrics computed over streamflow. Fig. 13 shows rank his-



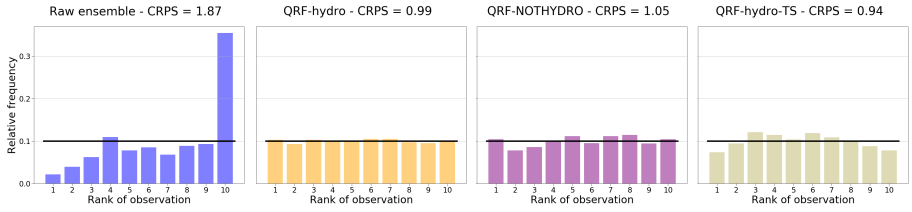

(a) 01/26/2016-01/30/2016

(b) 12/23/2013-12/26/2013

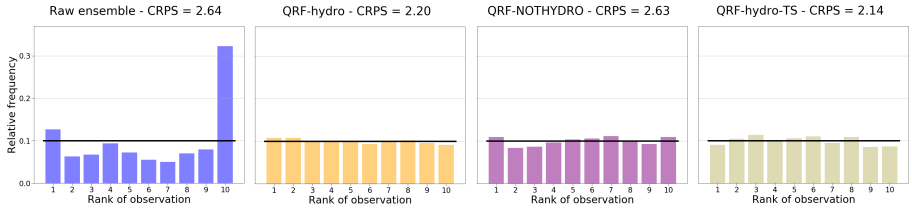
(c) Legend

Figure 12: First order Sobol' indices without uncertainty on rain input for the events 01/26/2016-01/30/2016 (12(a)) and 12/23/2013-12/26/2013 (12(b)) at Tréodet catchment

tograms for raw HEF, QRF-hydro quantiles, QRF-nothydro quantiles and QRF-hydro-TS HEF for 6-hour and 21-hour lead-times with associated CRPS, for Tréodet catchment. Similar results for Kerjean (respectively Ty-Planche) are given in Fig. 21 (respectively Fig. 22) as supplementary material in Section 7. Tab. 6 presents the time-averaged $\|.\|_2$ norm for RHs and Fig. 14 shows the time-varying CRPS for the four ensembles on all catchments.




(a) 6-hours lead-time




(b) 21-hours lead-time

Figure 13: Rank histograms for raw HEF, QRF-hydro quantiles, QRF-nothydro quantiles and QRF-hydro-TS HEF for 6-hours and 21-hours lead-times for Treodet. The associated CRPS is given at the top of the panel.

Table 6: Time-averaged $\|.\|_2$ RHs norm over lead-times ranging from 1 to 24 hours for raw HEF, QRF-hydro quantiles, QRF-nothydro quantiles and QRF-hydro-TS HEF, for Tréodet, Kerjean and Ty-Planche.

| Catchment | $10^{-2}\|.\|_2$ raw | $10^{-2}\|.\|_2$ QRF-hydro | $10^{-2}\|.\|_2$ QRF-nothydro | $10^{-2}\|.\|_2$ QRF-hydro-TS |
|---|---|---|---|---|
| Tréodet | 25.8 | 1.8 | 3.3 | 4.5 |
| Kerjean | 15.8 | 1.8 | 2.8 | 4.0 |
| Ty-Planche | 9.3 | 1.9 | 3.6 | 3.1 |

The raw ensemble is biased and underdispersive over all three catchments,

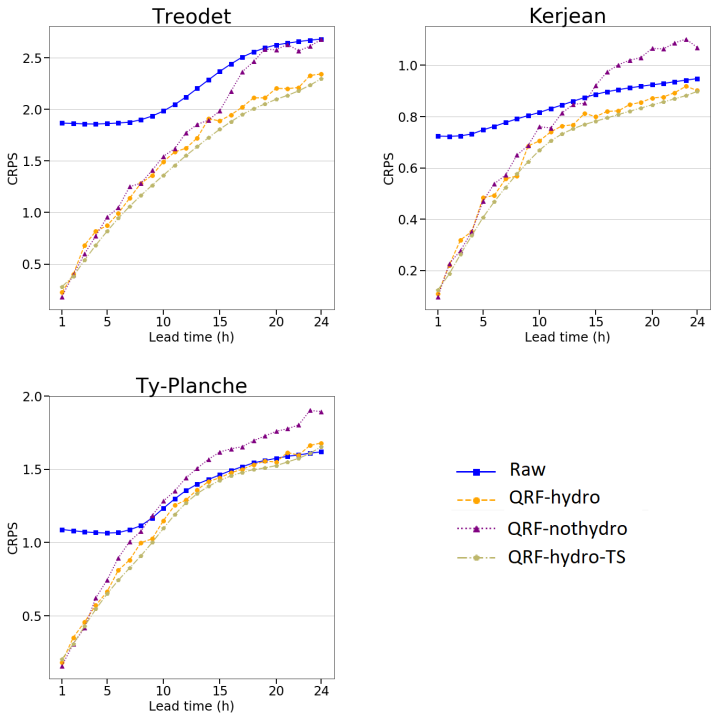
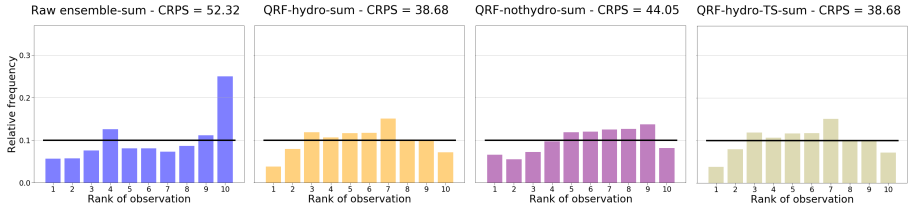



Figure 14: Time-varying CRPS for raw HEF, QRF-hydro quantiles, QRF-nothydro quantiles and QRF-hydro-TS HEF over lead-time for Tréodet, Kerjean and Ty-Planche.

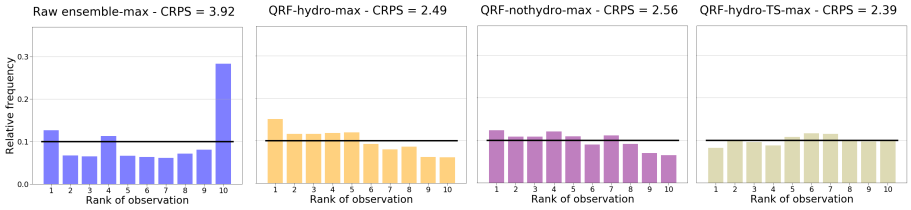
and tends to underestimate the flood peak values. QRF-hydro and QRF-nothydro calibrated ensembles display flatter RH with significantly smaller $\|.\|_2$ values. The use of hydrology-related predictors for calibration leads to the most satisfying results in terms of reliability. The TS smoothing procedure slightly deteriorates reliability. This is confirmed with statistical tests for the detection of slope, convexity or waves in the rank histograms (Jolliffe and Primo, 2008; Zamo, 2016) that lead to the rejection of the flatness hypothesis for the raw and the QRF-hydro-TS ensembles. It should be noted that reliability improvement remains significant as lead-time increases as shown in Fig. 13. Ensemble calibration considerably improves CRPS values for QRF-hydro and QRF-nothydro as shown in Fig. 14. Yet, the improvement decreases as the lead-time increases, especially when no hydrology-related predictors are used. The TS smoothing procedure further improves the CRPS values as this strategy provides less dispersive trajectories, a priori centered on observations after calibration step. In conclusion, QRF-hydro calibration leads to the best reliability. It should be noted that TS slightly degradates reliability, yet provides far better results than the raw ensemble. The CRPS simultaneously assesses for reliability and resolution. Since the reliability of QRF-hydro-TS is degraded and the CRPS is improved against QRF-hydro, that means that the TS procedure improves the resolution of the ensemble.

The cumulated and maximum streamflow over the 24 lead-times are computed for the four ensembles. For both quantities, the rank histogram, $\|.\|_2$ associated norm and CRPS are computed and shown in Fig. 15 and in Tab. 7. Similar results for Kerjean (respectively Ty-Planche) are given in Fig. 23 (respectively Fig. 24) as supplementary material in Section 7.3.

Ensemble calibration improves the rank histogram for both cumulated and maximum streamflow. The rank histogram for cumulated and maximum streamflow shows that the raw ensemble is underdispersive and biased as it tends to underestimate high cumulated and maximum streamflows for all three catchments. For Tréodet, all calibrated ensembles show overdispersive rank histogram for cumulated streamflow while it is rather flat for Kerjean and Ty-Planche when




(a) Cumulated streamflow




(b) Maximum streamflow

Figure 15: Rank histograms for the cumulated and the maximum streamflows over the 24 hours lead-time for the four ensembles for Treodet catchment. The associated CRPS is given at the top of the panel.

Table 7: Time-averaged $\|.\|_2$ RHs norm for cumulated and maximum streamflow cumulated over 24 lead-times for the four ensemble for Tréodet, Kerjean and Ty-Planche catchments.

| | Catchment | $10^{-2}\|.\|_2$ raw | $10^{-2}\|.\|_2$ QRF-hydro | $10^{-2}\|.\|_2$ QRF-nothydro | $10^{-2}\|.\|_2$ QRF-hydro-TS |
|---|---|---|---|---|---|
| Cumulated Stramflow | Tréodet | 17.1 | 9.4 | 8.8 | 9.4 |
| | Kerjean | 13.3 | 6.9 | 11.1 | 6.9 |
| | Ty-Planche | 9.1 | 6.8 | 9.5 | 6.8 |
| Maximum streamflow | Tréodet | 20.5 | 8.6 | 6.0 | 3.3 |
| | Kerjean | 16.8 | 6.6 | 4.8 | 3.9 |
| | Ty-Planche | 20.1 | 4.0 | 3.6 | 6.7 |

hydrology-related predictors are used. It should be noted that TS has no impact on rank histogram for cumulated streamflow, meaning that it preserves the volume of water within each simulated member. The rank histogram for maximum streamflow is further improved when trajectory smoothing is used at Tréodet and Kerjean but not at Ty-Planche. The merits of ensemble calibration on reliability is confirmed by the $\|.\|_2$ norm values that are significantly reduced for cumulated and maximum streamflow ensembles. Ensemble calibration also significantly improves CRPS values, especially when hydrology-related predictors are used for the three catchments. For the three catchments, the CRPS computed for maximum streamflow is minimized when trajectory smoothing is applied. To conclude, ensemble calibration with hydrology related predictors globally improves reliability and CRPS values for both maximum and cumulated streamflow and TS procedure brings a slight additional improvement for the maximum streamflow.

#### *5.2.1. Importance of the QRF predictors*

The a priori choice of the predictors is a key element in the QRF calibration. The a posteriori usefulness of the predictors is assessed as the loss in the mean-squared error of the whole forest if the predictor is randomly permuted without replacement: the values of the given predictor is a random sample of the original values. The predictor is of great (little) importance if the mean-squared error does (not) significantly increase when the predictor is randomly permuted.

Fig. 16 displays the log-importance of the QRF-hydro predictors for 2-hour and 16-hour lead-times for Tréodet. Since the QRF calibration is achieved with a cross-validation approach, a forest is built for each of the 27 months of $P_{HEF}$, and the importance of each predictor is thus computed over each of the 27 forests. For short lead-times, the most important predictor is the measured streamflow $Q_0$; that is consistent with the fact that hourly streamflows are strongly correlated in time. The second most important predictor is the gradient of the measured streamflow $GradQ_0$ that accounts for the flow dynamics. When the lead-time increases, rain-related predictors become predominant. When

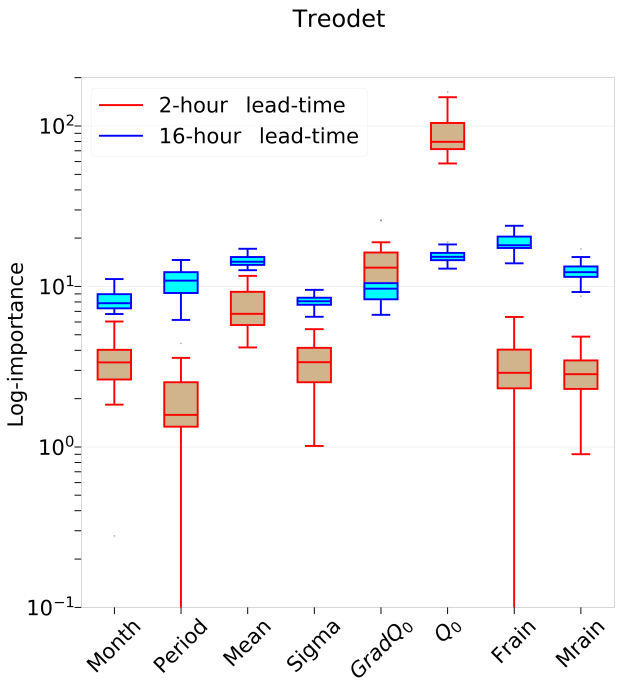



Figure 16: Log-importance of QRF-hydro predictors for the 2-hour and 16-hour lead-times at Tréodet. The box-plot is built for the measure of importance over 27 forests. The box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend from the box to show the range of the data.

the lead-time exceeds the catchment concentration time (about 15 hours for Tréodet), the cumulated forecasted rain since the beginning of the run $F_{rain}$, is the most important predictor, before $Q_0$. It also predominates the impact of the measured rain $M_{rain}$, meaning that the measured rainfall before the run has less importance than forecasted rainfall between the current time and the lead-time.

For the short and long lead-times, the raw ensemble mean and the standard deviation ($Mean$ and $Sigma$) have moderate influence (with a sligth predominance of $Mean$ over $Sigma$). This is consistent with the fact that the QRF-nothydro ensemble, which has not been built with these predictors, presents comparatively less favourable performances than the QRF-hydro ensemble. It should finally be noted that the importance of the different predictors tend to homogenise as the lead-time increases, and the interquartile range tends to decrease.

## 6. Discussion and perspectives

In this study, the hydrometeorological chain was investigated with an ensemble approach at a catchment scale, with the aim of issuing a statistically coherent Hydrological Ensemble Forecast (HEF) for up to 24 hours for the Odet cathment in France. The predominance of uncertainty sources is assessed with a global sensitivity analysis, first taking into account uncertainty in rain using AromeEPS-RR1, then only focusing on hydrological parameters from MORDOR-TS model. Predominant parameters were identified and taken into account in the HEF generation.

Because of the lack of AromeEPS-RR1 availability, the HEF is generated without taking into account uncertainty in rain forecast. However, it should be pointed out that for smaller lead-times than the concentration time of the catchments, the conclusions of this study are expected to be similar, as demonstrated by the comparative GSA without uncertainty in rainfall: the Sobol index associated to the rain is negligible at 6-hour lead-time. It was shown that the raw hydrological ensemble is underdispersive and underestimates observed streamflows, especially large values. The model-based-only approach is thus not sufficient to generate well-calibrated ensembles. The QRF calibration strategy is also applied for quantiles generation, with two different approaches: an observation-based-only approach, where the hydrologic model is not used and a combined model-observation approach where both observed and simulated information from the raw ensemble are used as predictors. The merits of the QRF calibration are assessed with reliability and resolution metrics computed for time-varying, cumulated and maximal streamflow. This article demonstrates, on the Odet catchment, that forest-based techniques, often used for the calibration of EPS, are also suitable for hydrologic ensemble calibration. QRF calibration provides an improvement of the reliability and CRPS values over the three studied sub-catchments. It was shown that better forecast skills are obtained when hydrology-related predictors are used. This highlights that the choice of the predictors is of great importance for the QRF calibration strategy

to be successful. Moreover, it also shows that statistically learning from observation fails to substitute for solving partial derivative equations when it comes to representing and forecasting the dynamics of the catchment. It should also be kept in mind that the reconstruction step following QRF-calibration requires the availability of the raw HEF in the Ensemble Copula Coupling method. The reconstructed HEF should finally be smoothed out before use as input for a hydraulic simulation with a hydrometeorological chain. The reliability of the smoothed ensemble is slightly deteriorated for hourly streamflow, but CRPS and global skills are mainly improved.

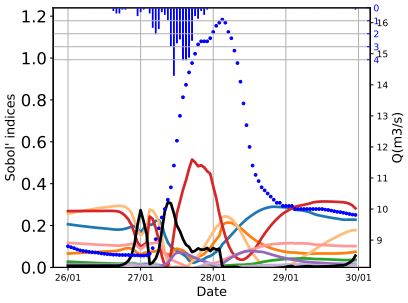
This strategy paves the way for an operational HEF system. A straightforward perspective is to apply this study over a period for which AromeEPS-RR1 is available, hydrological events occured, and that is long enough for QRF calibration to be applied on streamflows. This would allow to fully consider sources of uncertainty, especially those that relate to precipitation. Taking into account a larger variation of the corrective rain factor in MORDOR-TS jointly to using EPS could also be considered. This methodology could *a priori* be applied and assessed to other catchments for consistency check. The choice of predictors may be catchment dependant. For instance, the predictors $M_{rain}$ and $F_{rain}$ are closely related to the concentration time of the catchment: the period (in hours) over which the mean of measured or forecasted rain is computed may be adjusted, hence reduced for catchments that are subject to flash floods. The predictor *Month* may also be more significant as it allows to discriminate autumn flash floods from other events.

Finally, the hydrometeorological simulation chain could be extended to hydraulics, using forecasted calibrated streamflows as inputs to a hydraulic model to provide Hydraulics Ensemble Forecast. GSA on discretized water level and discharge in the river would be carried out with respect to rain-, hydrology- and hydraulic- related sources of uncertainty, now also considering uncertain friction and river bathymetry. Major sources uncertainties in the meteo-hydro-hydraulic chain could then be corrected in real time with assimilation of observed water level relying on the multi-physics ensemble simulation system.

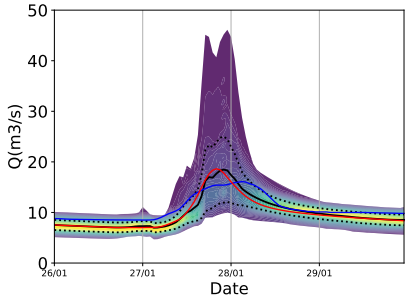
*Acknowledgements.* The authors greatfuly thank the French national service for flood forecasting (SCHAPI) for supporting this study, EDF-DTG for providing MORDOR-TS hydrologic model as well as Rémy Garçon and Matthtieu Le Lay for their active support on the model, Fabrice Zaoui (EDF R&D) for developping the Python API interface to MORDOR-TS and Charles Perrin and Julie Viatgé (INRAE) for fruitful discusions on the GRP hydrologic model. The authors also acknowledge the Pôle de Calcul et de Données Marines (PCDM) for providing DATARMOR storage and computational resources (http://www.ifremer.fr/pcdm).

## 7. Supplementary material

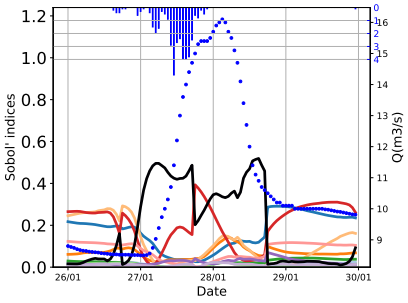
### *7.1. Sobol' indices at Kerjean and Ty-Planche*



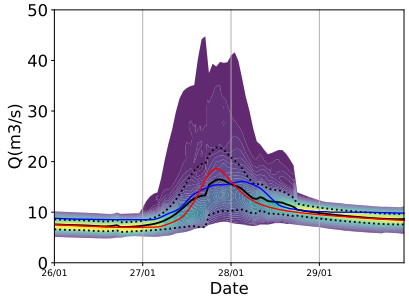

(a) Sobol'indices - 6h lead-time



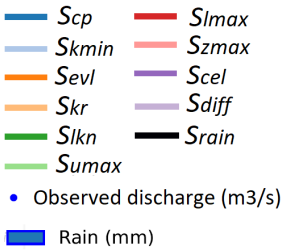

(b) Streamflow hydrographs - 6h lead-time



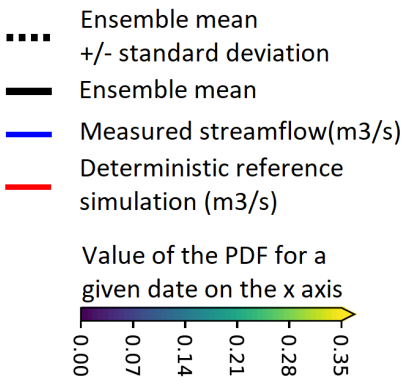

(c) Sobol'indices - 21h lead-time




(d) Streamflow hydrographs - 21h lead-time

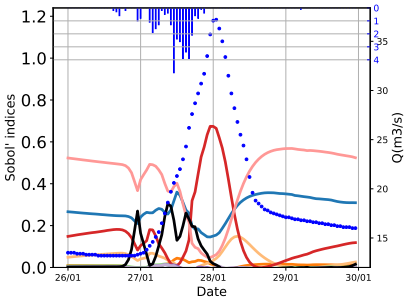



(e) Legend for MORDOR-TS Sobol'indices

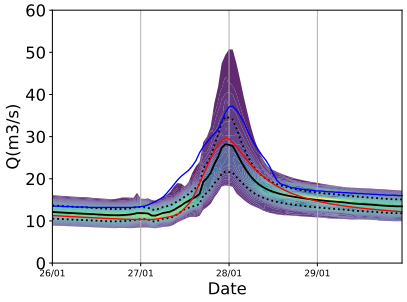



(f) Legend for streamflow graphs

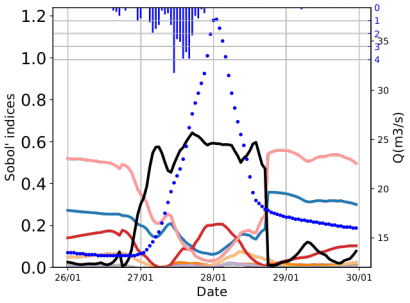
Figure 17: First order Sobol' indices and associated hydrographs for the event between 01/26/2016 and 01/30/2016 at Kerjean catchment for GSA-Arome. A vertical section of Fig(b) and Fig(d) displays the probability density function at the given date, according to the legend shown in (f). The results are shown for 6-hour ((a) and (b)) and 21-hour ((c) and (d)) lead-times.



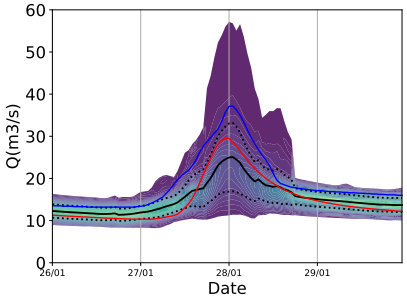

(a) Sobol'indices - 6h lead-time



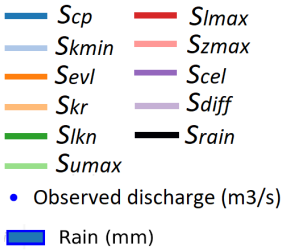

(b) Streamflow hydrographs - 6h lead-time



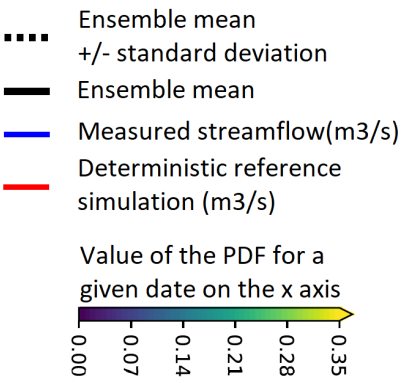

(c) Sobol'indices - 21h lead-time




(d) Streamflow hydrographs - 21h lead-time

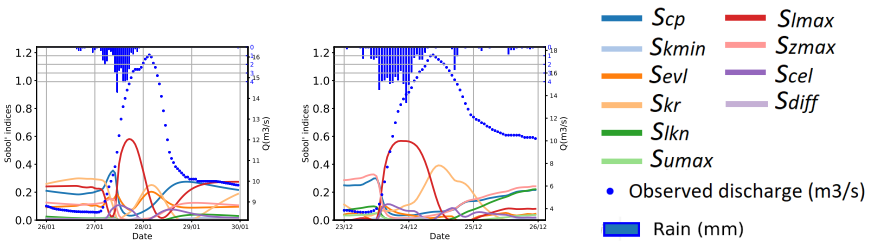



(e) Legend for MORDOR-TS Sobol'indices



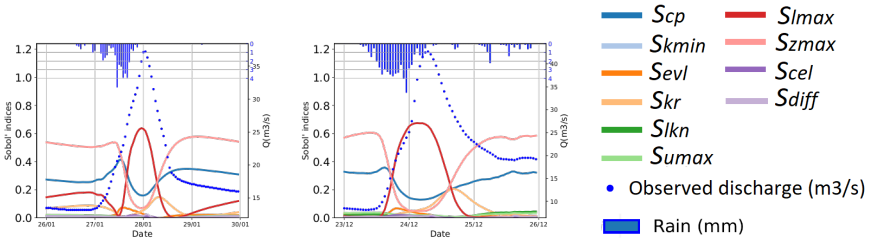

(f) Legend for streamflow graphs

Figure 18: First order Sobol' indices and associated hydrographs for the event between 01/26/2016 and 01/30/2016 at Ty-Planche catchment for GSA-Arome. A vertical section of Fig(b) and Fig(d) displays the probability density function at the given date, according to the legend shown in (f). The results are shown for 6-hour ((a) and (b)) and 21-hour ((c) and (d)) lead-times.



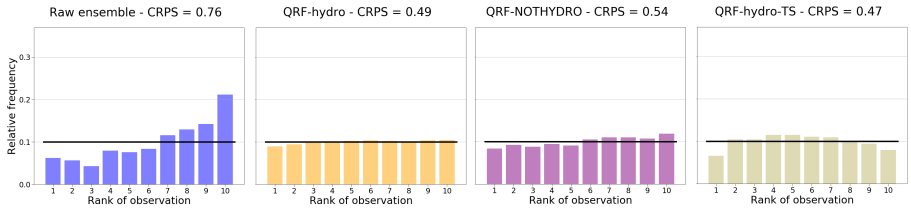

(a) 01/26/2016-01/30/2016 (b) 12/23/2013-12/26/2013 (c) Legend

Figure 19: First order Sobol' indices without uncertainty on rain input for the events 01/26/2016-01/30/2016 (19(a)) and 12/23/2013-12/26/2013 (19(b)) at Kerjean catchment

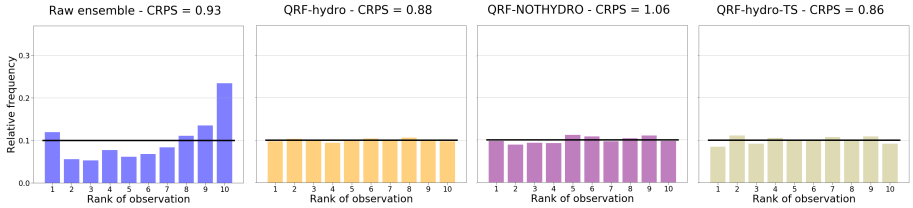



(a) 01/26/2016-01/30/2016 (b) 12/23/2013-12/26/2013 (c) Legend

Figure 20: First order Sobol' indices without uncertainty on rain input for the events 01/26/2016-01/30/2016 (20(a)) and 12/23/2013-12/26/2013 (20(b)) at Ty-Planche catchment

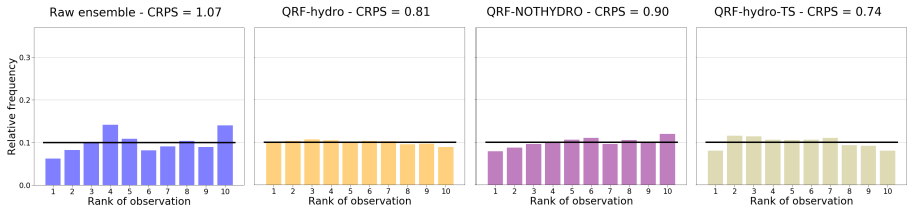



(a) 6-hours lead-time

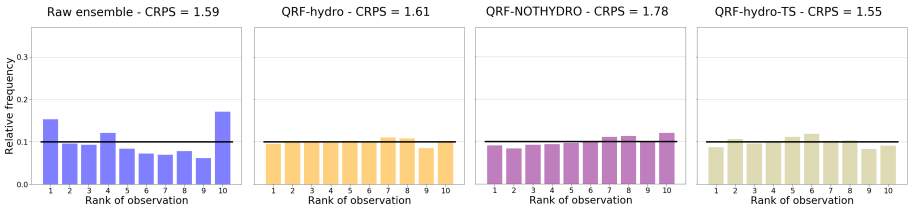



(b) 21-hours lead-time

Figure 21: Rank histograms for raw HEF, QRF-hydro quantiles, QRF-nothydro quantiles and QRF-hydro-TS HEF for 6-hours and 21-hours lead-times for Kerjean. The associated CRPS is given at the top of the panel.



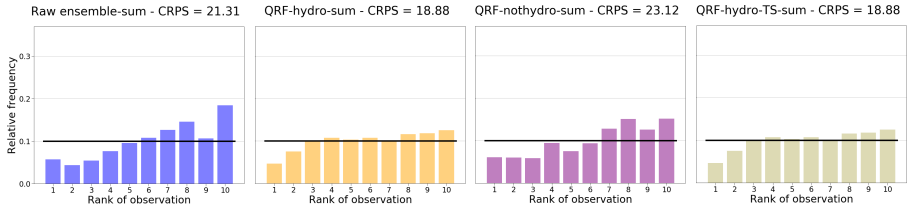

(a) 6-hours lead-time



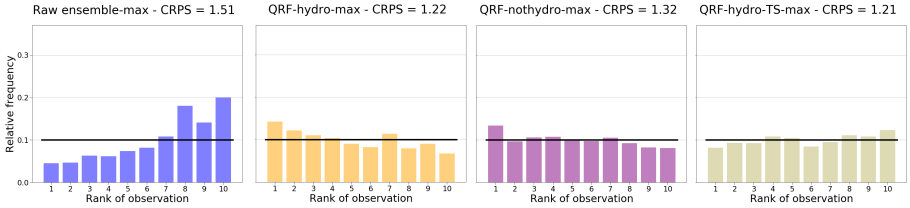

(b) 21-hours lead-time

Figure 22: Rank histograms for raw HEF, QRF-hydro quantiles, QRF-nothydro quantiles and QRF-hydro-TS HEF for 6-hours and 21-hours lead-times for Ty-Planche. The associated CRPS is given at the top of the panel.

### *7.3. Global skills for Kerjean and Ty-Planche catchments*

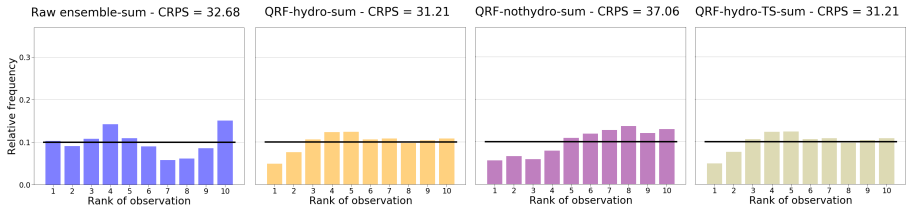



(a) Cumulated streamflow

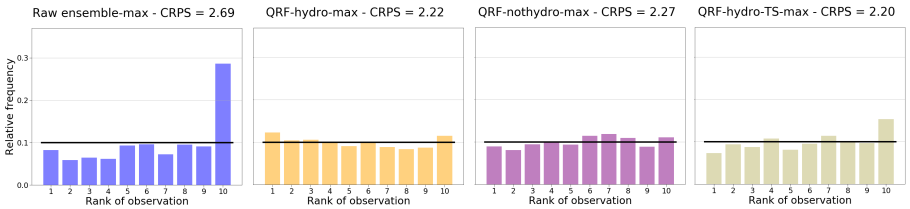



(b) Maximum streamflow

Figure 23: Rank histograms for the cumulated and the maximum streamflows over the 24 hours lead-time for the four ensembles for Kerjean catchment. The associated CRPS is given at the top of the panel.




(a) Cumulated streamflow




(b) Maximum streamflow

Figure 24: Rank histograms for the cumulated and the maximum streamflows over the 24 hours lead-time for the four ensembles for Ty-Planche catchment. The associated CRPS is given at the top of the panel.

## 8. Bibliographie

## References


Anderson, J.L., 1996. A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations. Journal of Climate 9, 1518–1530. URL: https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2, doi:10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2. number: 7.

Bellier, J., 2018. Prévisions hydrologiques probabilistes dans un cadre multivarié: quels outils pour assurer fiabilité et cohérence spatio-temporelle? Ph.D. thesis. Université Grenoble Alpes. URL: https://tel.archives-ouvertes.fr/tel-01950725.

Bellier, J., Zin, I., Bontron, G., 2017. Sample Stratification in Verification of Ensemble Forecasts of Continuous Scalar Variables: Potential Benefits and Pitfalls. Monthly Weather Review 145, 3529–3544. URL: http://journals.ametsoc.org/doi/10.1175/MWR-D-16-0487.1, doi:10.1175/MWR-D-16-0487.1. number: 9.

Bourgin, F., 2014. Comment quantifier l'incertitude prédictive en modélisation hydrologique ? : Travail exploratoire sur un grand échantillon de bassins versants. PhD Thesis. AgroParisTech. URL: http://www.theses.fr/2014AGPT0016/document.

Bouttier, F., Raynaud, L., Nuissier, O., Ménétrier, B., 2016. Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX. Quarterly Journal of the Royal Meteorological Society 142, 390–403. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2622, doi:10.1002/qj.2622. number: S1.

Bouttier, F., Vié, B., Nuissier, O., Raynaud, L., 2012. Impact of Stochastic Physics in a Convection-Permitting Ensemble. Monthly Weather Review

140, 3706–3721. URL: http://journals.ametsoc.org/doi/abs/10.1175/MWR-D-12-00031.1, doi:10.1175/MWR-D-12-00031.1. number: 11.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140. URL: http://link.springer.com/10.1007/BF00058655, doi:10.1007/BF00058655. number: 2.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. doi:10.1023/A:1010933404324.. number: 4.

Bremnes, J.B., 2004. Probabilistic wind power forecasts using local quantile regression. Wind Energy 7, 47–54. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/we.107, doi:10.1002/we.107. number: 1.

Candille, G., Talagrand, O., 2005. Evaluation of probabilistic prediction systems for a scalar variable. Quarterly Journal of the Royal Meteorological Society 131, 2131–2150. URL: http://doi.wiley.com/10.1256/qj.04.71, doi:10.1256/qj.04.71. number: 609.

Champeaux, J.L., Dupuy, P., Laurantin, O., Soulan, I., Tabary, P., Soubeyroux, J.M., 2009. Les mesures de précipitations et l'estimation des lames d'eau à Météo-France : état de l'art et perspectives. La Houille Blanche , 28–34URL: http://www.shf-lhb.org/10.1051/lhb/2009052, doi:10.1051/lhb/2009052. number: 5.

Cloke, H., Pappenberger, F., 2009. Ensemble flood forecasting: A review. Journal of Hydrology 375, 613–626. URL: https://linkinghub.elsevier.com/retrieve/pii/S0022169409003291, doi:10.1016/j.jhydrol.2009.06.005. number: 3-4.

Delle Monache, L., Eckel, F.A., Rife, D.L., Nagarajan, B., Searight, K., 2013. Probabilistic Weather Prediction with an Analog Ensemble. Monthly Weather Review 141, 3498–3516. URL: http://journals.ametsoc.org/doi/abs/10.1175/MWR-D-12-00281.1, doi:10.1175/MWR-D-12-00281.1. number: 10.

Delle Monache, L., Hacker, J.P., Zhou, Y., Deng, X., Stull, R.B., 2006. Probabilistic aspects of meteorological and ozone regional ensemble forecasts. Journal of Geophysical Research: Atmospheres 111. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JD006917, doi:10.1029/2005JD006917. number: D24.

Demargne, J., Wu, L., Regonda, S.K., Brown, J.D., Lee, H., He, M., Seo, D.J., Hartman, R., Herr, H.D., Fresch, M., Schaake, J., Zhu, Y., 2014. The Science of NOAA's Operational Hydrologic Ensemble Forecast Service. Bulletin of the American Meteorological Society 95, 79–98. URL: http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-12-00081.1, doi:10.1175/BAMS-D-12-00081.1. number: 1.

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P., Cébron, P., 2015. PEARP, the Météo-France short-range ensemble prediction system. Quarterly Journal of the Royal Meteorological Society 141, 1671–1685. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2469, doi:10.1002/qj.2469. number: 690.

Dietrich, J., Schumann, A.H., Redetzky, M., Walther, J., Denhard, M., Wang, Y., Pfützner, B., Büttner, U., 2009. Assessing uncertainties in flood forecasts for decision making: prototype of an operational flood management system integrating ensemble predictions. Natural Hazards and Earth System Science 9, 1529–1540. URL: http://www.nat-hazards-earth-syst-sci.net/9/1529/2009/, doi:10.5194/nhess-9-1529-2009. number: 4.

Duan, Q., Ajami, N.K., Gao, X., Sorooshian, S., 2007. Multi-model ensemble hydrologic prediction using Bayesian model averaging. Advances in Water Resources 30, 1371–1386. URL: https://linkinghub.elsevier.com/retrieve/pii/S030917080600220X, doi:10.1016/j.advwatres.2006.11.014. number: 5.

Efron, B., Stein, C., 1981. The Jackknife Estimate of Variance. The Annals

of Statistics 9, 586–596. URL: https://doi.org/10.1214/aos/1176345462, doi:10.1214/aos/1176345462. number: 3.

Emery, C.M., Biancamaria, S., Boone, A., Garambois, P.A., Ricci, S., Rochoux, M.C., Decharme, B., 2016. Temporal Variance-Based Sensitivity Analysis of the River-Routing Component of the Large-Scale Hydrological Model ISBA–TRIP: Application on the Amazon Basin. Journal of Hydrometeorology 17, 3007–3027. URL: http://journals.ametsoc.org/doi/10.1175/JHM-D-16-0050.1, doi:10.1175/JHM-D-16-0050.1. number: 12.

Garambois, P.A., Roux, H., Larnier, K., Castaings, W., Dartus, D., 2013. Characterization of process-oriented hydrologic model behavior with temporal sensitivity analysis for flash floods in Mediterranean catchments. Hydrology and Earth System Sciences 17, 2305–2322. URL: https://www.hydrol-earth-syst-sci.net/17/2305/2013/, doi:10.5194/hess-17-2305-2013. number: 6.

Garavaglia, F., Le Lay, M., Gottardi, F., Garçon, R., Gailhard, J., Paquet, E., Mathevet, T., 2017. Impact of model structure on flow simulation and hydrological realism: from a lumped to a semi-distributed approach. Hydrology and Earth System Sciences 21, 3937–3952. URL: https://www.hydrol-earth-syst-sci.net/21/3937/2017/, doi:10.5194/hess-21-3937-2017. number: 8.

Garçon, R., 1996. Prévision opérationnelle des apports de la Durance à Serre-Ponçon à l'aide du modèle MORDOR. Bilan de l'année 1994-1995. La Houille Blanche , 71–76URL: http://www.shf-lhb.org/10.1051/lhb/1996056, doi:10.1051/lhb/1996056. number: 5.

Gneiting, T., Raftery, A.E., 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association 102, 359–378. URL: http://www.tandfonline.com/doi/abs/10.1198/016214506000001437, doi:10.1198/016214506000001437. number: 477.

Gneiting, T., Raftery, A.E., Westveld, A.H., Goldman, T., 2005. Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. Monthly Weather Review 133, 1098–1118. URL: https://doi.org/10.1175/MWR2904.1, doi:10.1175/MWR2904.1. number: 5.

Gneiting, T., Ranjan, R., 2011. Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. Journal of Business & Economic Statistics 29, 411–422. URL: http://www.tandfonline.com/doi/abs/10.1198/jbes.2010.08110, doi:10.1198/jbes.2010.08110. number: 3.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology 377, 80 – 91. URL: http://www.sciencedirect.com/science/article/pii/S0022169409004843, doi:https://doi.org/10.1016/j.jhydrol.2009.08.003. number: 1.

Hamill, T.M., Colucci, S.J., 1997. Verification of Eta–RSM Short-Range Ensemble Forecasts. Monthly Weather Review 125, 1312–1327. URL: http://journals.ametsoc.org/doi/abs/10.1175/1520-0493%281997%29125%3C1312%3AVOERSR%3E2.0.CO%3B2, doi:10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2. number: 6.

Hamill, T.M., Whitaker, J.S., 2006. Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application. Monthly Weather Review 134, 3209–3229. URL: http://journals.ametsoc.org/doi/abs/10.1175/MWR3237.1, doi:10.1175/MWR3237.1. number: 11.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning. Springer. URL: https://web.stanford.edu/~hastie/Papers/ESLII.pdf.

Hayami, S., 1951. On The Propagation Of Flood Waves. Technical Report.

Disaster prevention research institute. URL: https://repository.kulib.kyoto-u.ac.jp/dspace/bitstream/2433/123641/1/b01p0n000p01.pdf.

Hemri, S., Lisniak, D., Klein, B., 2015. Multivariate postprocessing techniques for probabilistic hydrological forecasting. Water Resources Research 51, 7436–7451. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014WR016473, doi:10.1002/2014WR016473. number: 9.

Hersbach, H., 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Weather and Forecasting 15, 559–570. URL: http://journals.ametsoc.org/doi/abs/10.1175/1520-0434%282000%29015%3C0559%3ADOTCRP%3E2.0.CO%3B2, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2. number: 5.

Hoeffding, W., 1948. A Class of Statistics with Asymptotically Normal Distribution. The Annals of Mathematical Statistics 19, 293–325. URL: https://doi.org/10.1214/aoms/1177730196, doi:10.1214/aoms/1177730196. number: 3.

Hopson, T.M., Webster, P.J., 2010. A 1–10-Day Ensemble Forecasting Scheme for the Major River Basins of Bangladesh: Forecasting Severe Floods of 2003–07*. Journal of Hydrometeorology 11, 618–641. URL: http://journals.ametsoc.org/doi/abs/10.1175/2009JHM1006.1, doi:10.1175/2009JHM1006.1. number: 3.

Iooss, B., Lemaître, P., 2015. A Review on Global Sensitivity Analysis Methods, in: Dellino, G., Meloni, C. (Eds.), Uncertainty Management in Simulation-Optimization of Complex Systems. Springer US, Boston, MA. volume 59, pp. 101–122. URL: http://link.springer.com/10.1007/978-1-4899-7547-8_5, doi:10.1007/978-1-4899-7547-8_5.

Jolliffe, I.T., Primo, C., 2008. Evaluating Rank Histograms Using Decompositions of the Chi-Square Test Statistic. Monthly Weather Review 136, 2133–2139. URL: https://doi.org/10.1175/2007MWR2219.1, doi:10.1175/2007MWR2219.1. number: 6.

Jolliffe, I.T., Stephenson, D.B., 2003. Forecast verification: a practitioner's guide in atmospheric science. Wiley, Chichester.

Krzysztofowicz, R., 1999. Bayesian theory of probabilistic forecasting via deterministic hydrologic model. Water Resources Research 35, 2739–2750. URL: http://doi.wiley.com/10.1029/1999WR900099, doi:10.1029/1999WR900099. number: 9.

Le Moine, N., Hendrickx, F., Gailhard, J., Garçon, R., Gottardi, F., 2015. Hydrologically Aided Interpolation of Daily Precipitation and Temperature Fields in a Mesoscale Alpine Catchment. Journal of Hydrometeorology 16, 2595–2618. URL: http://journals.ametsoc.org/doi/10.1175/JHM-D-14-0162.1, doi:10.1175/JHM-D-14-0162.1. number: 6.

Leleu, Isabelle, Tonnelier, Isabelle, Puechberty, Rachel, Gouin, Philippe, Viquendi, Isabelle, Cobos, Laurent, Foray, Anouck, Baillon, Martine, Ndima, Pierre-Olivier, 2014. La refonte du système d'information national pour la gestion et la mise à disposition des données hydrométriques. La Houille Blanche , 25–32URL: https://doi.org/10.1051/lhb/2014004, doi:10.1051/lhb/2014004. number: 1.

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., Di, Z., 2017. A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. WIREs Water 4, e1246. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wat2.1246, doi:10.1002/wat2.1246. number: 6 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wat2.1246.

Matheson, J.E., Winkler, R.L., 1976. Scoring Rules for Continuous Probability Distributions. Management Science 22, 1087–1096. URL: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.22.10.1087, doi:10.1287/mnsc.22.10.1087. number: 10.

Meinshausen, N., 2006. Quantile Regression Forests. Journal of Machine Learning Research 7, 17.

Michon, T., Castaings, W., 2017. Stratégie de calage du modèle hydrologique semi-distribué MORDOR-SD. Technical Report. TENEVIA.

Monteil, C., Zaoui, F., Le Moine, N., Hendrickx, F., 2019. Technical note: the caRamel R package for Automatic Calibration by Evolutionary Multi Objective Algorithm. Hydrology and Earth System Sciences Discussions 2019, 1–16. URL: https://www.hydrol-earth-syst-sci-discuss.net/hess-2019-259/, doi:10.5194/hess-2019-259.

Nash, J., Sutcliffe, J., 1970. River flow forecasting through conceptual models part I — A discussion of principles. Journal of Hydrology 10, 282–290. URL: https://linkinghub.elsevier.com/retrieve/pii/0022169470902556, doi:10.1016/0022-1694(70)90255-6. number: 3.

Pappenberger, F., Pagano, T.C., Brown, J.D., Alfieri, L., Lavers, D.A., Berthet, L., Bressand, F., Cloke, H.L., Cranston, M., Danhelka, J., Demargne, J., Demuth, N., de Saint-Aubin, C., Feikema, P.M., Fresch, M.A., Garçon, R., Gelfan, A., He, Y., Hu, Y.Z., Janet, B., Jurdy, N., Javelle, P., Kuchment, L., Laborda, Y., Langsholt, E., Le Lay, M., Li, Z.J., Mannessiez, F., Marchandise, A., Marty, R., Meißner, D., Manful, D., Organde, D., Pourret, V., Rademacher, S., Ramos, M.H., Reinbold, D., Tibaldi, S., Silvano, P., Salamon, P., Shin, D., Sorbet, C., Sprokkereef, E., Thiemig, V., Tuteja, N.K., van Andel, S.J., Verkade, J.S., Vehviläinen, B., Vogelbacher, A., Wetterhall, F., Zappa, M., Van der Zwan, R.E., Thielen-del Pozo, J., 2016. Hydrological Ensemble Prediction Systems Around the Globe, in: Duan, Q., Pappenberger, F., Thielen, J., Wood, A., Cloke, H.L., Schaake, J.C. (Eds.), Handbook of Hydrometeorological Ensemble Forecasting. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–35. URL: https://doi.org/10.1007/978-3-642-40457-3_47-1, doi:10.1007/978-3-642-40457-3_47-1.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Monthly

925 Weather Review 133, 1155–1174. URL: https://doi.org/10.1175/
926 MWR2906.1, doi:10.1175/MWR2906.1. number: 5.

927 Raynaud, L., Bouttier, F., 2016. Comparison of initial perturbation meth-
928 ods for ensemble prediction at convective scale. Quarterly Journal of
929 the Royal Meteorological Society 142, 854–866. URL: https://rmets.
930 onlinelibrary.wiley.com/doi/abs/10.1002/qj.2686, doi:10.1002/qj.
931 2686. number: 695.

932 Rouhier, L., Le Lay, M., Garavaglia, F., Le Moine, N., Hendrickx, F., Mon-
933 teil, C., Ribstein, P., 2017. Impact of mesoscale spatial variability of cli-
934 matic inputs and parameters on the hydrological response. Journal of Hy-
935 drology 553, 13–25. URL: http://linkinghub.elsevier.com/retrieve/
936 pii/S0022169417305012, doi:10.1016/j.jhydrol.2017.07.037.

937 Saltelli, A. (Ed.), 2008. Global sensitivity analysis: the primer. John Wiley,
938 Chichester, England ; Hoboken, NJ. OCLC: ocn180852094.

939 Saltelli, A., Annoni, P., 2010. How to avoid a perfunctory sensitivity anal-
940 ysis. Environmental Modelling & Software 25, 1508–1517. URL: https:
941 //linkinghub.elsevier.com/retrieve/pii/S1364815210001180, doi:10.
942 1016/j.envsoft.2010.04.012. number: 12.

943 Schaake, J., Pailleux, J., Thielen, J., Arritt, R., Hamill, T., Luo, L., Martin, E.,
944 McCollor, D., Pappenberger, F., 2010. Summary of recommendations of the
945 first workshop on Postprocessing and Downscaling Atmospheric Forecasts for
946 Hydrologic Applications held at Météo-France, Toulouse, France, 15-18 June
947 2009. Atmospheric Science Letters 11, 59–63. URL: http://doi.wiley.com/
948 10.1002/asl.267, doi:10.1002/asl.267. number: 2.

949 Schefzik, R., Thorarinsdottir, T.L., Gneiting, T., 2013. Uncertainty Quan-
950 tification in Complex Simulation Models Using Ensemble Copula Coupling.
951 Statistical Science 28, 616–640. URL: http://projecteuclid.org/euclid.
952 ss/1386078881, doi:10.1214/13-STS443. number: 4.

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., Masson, V., 2011. The AROME-France Convective-Scale Operational Model. Monthly Weather Review 139, 976–991. URL: http://journals.ametsoc.org/doi/abs/10.1175/2010MWR3425.1, doi:10.1175/2010MWR3425.1. number: 3.

Sobol, I.M., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation 55, 271 – 280. URL: http://www.sciencedirect.com/science/article/pii/S0378475400002706, doi:https://doi.org/10.1016/S0378-4754(00)00270-6. number: 1.

Taillardat, M., Fougères, A.L., Naveau, P., Mestre, O., 2019. Forest-Based and Semiparametric Methods for the Postprocessing of Rainfall Ensemble Forecasting. Weather and Forecasting 34, 617–634. URL: http://journals.ametsoc.org/doi/10.1175/WAF-D-18-0149.1, doi:10.1175/WAF-D-18-0149.1. number: 3.

Taillardat, M., Mestre, O., Zamo, M., Naveau, P., 2016. Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics. Monthly Weather Review 144, 2375–2393. URL: http://journals.ametsoc.org/doi/10.1175/MWR-D-15-0260.1, doi:10.1175/MWR-D-15-0260.1. number: 6.

Talagrand, O., Vautard, R., Strauss, B., 1997. Evaluation of probabilistic prediction systems, in: Workshop on Predictability, 20-22 October 1997, ECMWF, Shinfield Park, Reading. pp. 1–26. URL: https://www.ecmwf.int/node/12555.

Thiboult, A., Anctil, F., Boucher, M.A., 2016. Accounting for three sources of uncertainty in ensemble hydrological forecasting. Hydrology and Earth System Sciences 20, 1809–1825. URL: https://www.hydrol-earth-syst-sci.net/20/1809/2016/, doi:10.5194/hess-20-1809-2016. number: 5.

Thielen, J., Schaake, J., Hartman, R., Buizza, R., 2008. Aims, challenges and progress of the Hydrological Ensemble Prediction Experiment (HEPEX) following the third HEPEX workshop held in Stresa 27 to 29 June 2007. Atmospheric Science Letters 9, 29–35. URL: http://doi.wiley.com/10.1002/asl.168, doi:10.1002/asl.168. number: 2.

T.Roy, P., Ricci, S., Dupuis, R., Campet, R., Jouhaud, J.C., Fournier, C., 2018. BATMAN: Statistical analysis for expensive computer codes made easy. The Journal of Open Source Software 3, 493. URL: http://dx.doi.org/10.21105/joss.00493, doi:10.21105/joss.00493. number: 21.

Van Schaeybroeck, B., Vannitsem, S., 2015. Ensemble post-processing using member-by-member approaches: theoretical aspects. Quarterly Journal of the Royal Meteorological Society 141, 807–818. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2397, doi:https://doi.org/10.1002/qj.2397. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.2397.

Velázquez, J.A., Anctil, F., Ramos, M.H., Perrin, C., 2011. Can a multimodel approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. Advances in Geosciences 29, 33–42. URL: https://www.adv-geosci.net/29/33/2011/, doi:10.5194/adgeo-29-33-2011.

Vidal, J.P., Martin, E., Franchistéguy, L., Baillon, M., Soubeyroux, J.M., 2010. A 50-year high-resolution atmospheric reanalysis over France with the Safran system. International Journal of Climatology 30, 1627–1644. URL: http://doi.wiley.com/10.1002/joc.2003, doi:10.1002/joc.2003. number: 11.

Wilks, D.S., 2018. Univariate ensemble post-processing, in: Vannitsem, S., Wilks, D.S., Messner, J.W. (Eds.), Statistical postprocessing of ensemble forecasts. Amsterdam, Netherlands. Elsevier, pp. 49–89.

Zalachori, I., Ramos, M.H., Garçon, R., Mathevet, T., Gailhard, J., 2012. Statistical processing of forecasts for hydrological ensemble prediction: a com-

parative study of different bias correction strategies. Advances in Science and Research 8, 135–141. URL: http://www.adv-sci-res.net/8/135/2012/, doi:10.5194/asr-8-135-2012.

Zamo, M., 2016. Statistical post-processing of deterministic and ensemble wind speed forecasts on a grid.

Zamo, M., Mestre, O., Arbogast, P., Pannekoucke, O., 2014. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. Solar Energy 105, 804 – 816. URL: http://www.sciencedirect.com/science/article/pii/S0038092X14001601, doi:https://doi.org/10.1016/j.solener.2014.03.026.

Zappa, M., Jaun, S., Germann, U., Walser, A., Fundel, F., 2011. Superposition of three sources of uncertainties in operational flood forecasting chains. Atmospheric Research 100, 246–262. URL: https://linkinghub.elsevier.com/retrieve/pii/S016980951000342X, doi:10.1016/j.atmosres.2010.12.005. number: 2-3.