Comparing hydrological post-processors including ensembles predictions into full predictive probability distribution of streamflow

D. Biondi[1] **and E. Todini**[2]

[1] Università della Calabria, DIMES, Arcavacata di Rende (CS), Italy

[2] Italian Hydrological Society, Bologna (BO), Italy.

Corresponding author: Daniela Biondi (daniela.biondi@unical.it)

Key Points:

- Full predictive probability distributions are essential for taking informed decisions thus improving decision making
- Hydro-meteorological models provide ensemble predictions which do not meet the definition of the predictive probability distribution
- Several uncertainty post-processors, incorporating ensemble spread information, are here introduced to properly incorporating ensemble predictions into predictive distributions

Abstract

Although not matching the formal definition of the predictive probability distribution, meteorological as well as hydrological ensembles have been frequently interpreted and directly used to assess flood-forecasting predictive uncertainty. With the objective of correctly assessing the predictive probability of floods, this paper introduces ways of taking into account the measures of uncertainty provided in the form of ensemble forecasts by modifying a number of well established uncertainty post processors, such as for instance BMA (Bayesian Model Averaging) and MCP (Model Conditional Processor). The uncertainty post-processors, were developed on the assumption that the future unknown quantity (predictand) is uncertain while model forecasts (predictors) are given, which imply they are perfectly known. With this in mind, we propose to relax this assumption by considering ensemble predictions, in analogy to measurement errors, as expressions of errors in model predictions to be integrated in the post-processors coefficients estimation process. The analyses of the methodologies proposed in this work are conducted on a real case study based on meteorological ensemble predictions for the Po River at Pontelagoscuro in Italy. After showing how improper can be the direct use of ensemble predictions to describe the predictive probability distribution, results from the modified post-processors are compared and discussed.

1 Introduction

It is now more than two decades that “deterministic” meteorological forecasts have given way to “probabilistic forecasts” essentially based on meteorological ensembles predictions accounting for the uncertainty on model forecasts due to parameterizations as well as to assumptions made on parameter values and on initial and boundary conditions. Similarly, there has been a tendency to convert hydrological forecasts into probabilistic hydrological ensemble forecasts, by forcing the hydrological models with the above mentioned meteorological ensemble members as inputs (see for example the HEPEX web site https://hepex.irstea.fr/; Schaake et al., 2007). The demand for probabilistic forecasts directly descends from the need of robust, correct and reliable decisions under the uncertainty of future events, and it is only through probabilistic forecasts that this objective can be met. Following the Bayesian decision approaches, rational decisions can be reached through the maximization of “expected utility functions”, which can be estimated if a correct predictive probability density of the future unknown quantities, such as discharges, water stages or water volumes, is available.

State of the art operational (or pre-operational) medium range flood forecasting systems, use ensembles of Numerical Weather Predictions (NWPs) as input to a hydrological and/or hydraulic model to produce river discharge predictions, which are often referred to as Ensemble Streamflow Predictions (ESP) (Cloke & Pappenberger, 2009). Current ensemble-based forecasting systems are mostly designed to represent only uncertainty in NWPs that is then cascaded through the forecasting systems to produce an uncertain prediction of flooding. Nevertheless, this format may not be adequate to represent the predictive probability distribution required to reach informed decisions. Although, unknown future precipitation has been often demonstrated to be the dominant source of uncertainty for many streamflow forecasts (Bartholmes & Todini 2005; Pappenberger et al., 2005; Cuo et al., 2011), becoming much more important with increasing lead time (Nester et al., 2012), other uncertainties arise further down in the flood forecasting cascade which could also be significant according to forecast lead time, magnitude of the event and catchments characteristics. The major sources contributing to the hydrologic uncertainty include for example: spatial and temporal downscaling of the average amount of precipitation over the basin; geometry of the system;

observational errors; model structural uncertainties representing the limitations of the models available to fully represent the true processes; errors in model parameter value and uncertainty in the model boundary or initial conditions. Moreover, there is a number of many recent studies suggesting that uncertainty is unavoidable in hydrology originating from natural variability and related to inherent unpredictability in deterministic terms (Montanari et al., 2009; Koutsoyiannis et al., 2009). Therefore, it is possible that the assumptions of equal probability are violated and the total uncertainty is underestimated (Golding, 2000).

Explicitly addressing the individual aspects of uncertainty along the forecast chain, of course, is a priority involving efforts of researchers and forecasters, in order to identify the dominant sources of uncertainty and reduce them in a meaningful way, for example by prioritizing improvements in each of the different sources (e.g. model states data assimilation, real-time parameter updating, use of different model structures). Conversely, not only this assessment is time consuming and there are many situations, particularly in real-time forecasting, where time constraints do not allow for computationally demanding efforts required to account for the full range of uncertainty, but also, it becomes less relevant for the prediction/decision chain, as will be discussed in Section 3.

It is also worthy to note that the predictive probability distribution is defined as the probability distribution of a future occurrence given (conditional to) all the available knowledge, usually encapsulated in one or more model forecasts (Krzysztofowicz, 1999; Todini, 2008), which can then be used to estimate the expected value of benefits or losses. This is why extensive work was done in the last decades to develop a variety of "uncertainty post-processors", i.e. statistical techniques that are applied consequently to one or multiple model run in order to reduce or estimate uncertainty, that more effectively improve the forecast quality and allow to assess the full range of predictive probabilities (Benninga et al., 2017; Li & Duan, 2018). A recent comprehensive review of the commonly used statistical post-processing methods for both meteorological (e.g. precipitation) and hydrological (e.g. streamflow) forecasts, is provided in Li et al. (2017). Statistical post-processing of model outputs characterizes the frequency distribution of past prediction errors and use this information to condition forecasts at a future time. Former post-processors relied on parametric univariate techniques considering deterministic single model forecasts to generate probabilistic forecasts (e.g. Krzysztofowicz, 1999; Montanari & Brath, 2004; Seo et al., 2006; Montanari & Grossi, 2008; Regonda et al., 2013), but many of them have been subsequently extended to handle multiple hydrological forecasts.

Reggiani et al. (2009), for instance, extended the HUP (Hydrological Uncertainty Processor; Krzysztofowicz, 1999) into BEUP (Bayesian Ensemble Uncertainty Processor) to treat meteorological ensembles and estimate the predictive uncertainty for operational River Rhine forecasting system. This approach somehow takes into account the ensemble spread but, in the Gaussian space, implies homoscedasticity of the error variance, which is assumed independent from the magnitude of the observed or forecasted values as well as from the ensemble spread.

In order to overcome this limitation, Weerts et al. (2011) used the Quantile Regression (QR; Koenker, 2005) to consider heteroscedasticity in the error variance. The Quantile Regression approach tries to represent the error heteroscedasticity identifying a linear (or non-linear) variation of each quantile of the predictive density as a function of the model forecast. Nevertheless, the approach requires the estimation of at least two parameters per quantile leading to a non-parsimonious problem and in addition, as shown by Coccia and Todini (2011) poor results may be obtained.

Introduced by Raftery (1993), Bayesian Model Averaging (BMA) has been frequently applied in hydrology to combine hydrological forecasts from individual models and characterize predictive uncertainty induced by model structure. The predictive uncertainty is in this case estimated as an approximation of the formal conditional density, via a Bayesian mixture of densities i.e. as a weighted mean of the predictive distributions of individual models. In the original form of BMA, constant variance is assumed for the predictive distributions of individual models, irrespective of magnitude of the variable. Recently, Madadgar and Moradkhani (2014) introduced a copula-embedded BMA, a combination of copulas and BMA that relaxes any assumption on the shape of conditional probability density functions and enables modelling the dependency structures between variates independently from their marginal distributions.

Gneiting et al. (2005), to make use of the ensemble spread information, introduced a variant of Model Output Statistics (MOS) for handling ensembles, which they called EMOS (Ensemble Model Output Statistics). To post-process hydro-meteorological ensemble forecasts, EMOS uses a heteroschedastic regression taking into account the ensemble spread variance.

Based on the properties of the Multi-Normal distribution, Todini (2008) developed the Model Conditional Processor (MCP) that conveniently handles multivariate situations, making it suitable for multi-model, multi-site, and multi-lead time problems. Similarly to HUP, the basic MCP implies homoscedasticity. In order to account for heteroscedasticity and to improve adaptation to low and high flows, the MCP approach was thus extended from the Normal to the Truncated Normal distributions (Coccia & Todini, 2011).

The assumption of a parametric joint distribution is central to all aforementioned statistical post-processors and (transformed) forecasts and observations are often considered joint normally distributed. Brown and Seo (2010) proposed a nonparametric approach for the post-processing of hydrological ensemble forecasts similar to indicator co-kriging in geo-statistics.

Post-processing of ensembles is, indeed, one of HEPEX six major themes (Verkade et al., 2013; Ramos et al., 2013;) and a significant work has been done within the HEPEX community for improving and diffusing the use of ensembles to enhance operational forecasting and decision-making by sharing accomplishments and lessons learned.

Unfortunately, a large effort is still needed in terms of diffusion of innovation to pass from the innovation to the implementation and confirmation stages (Rogers, 2003). First, although ensemble prediction systems have been accepted to add value to flood incident management by increasing the capability to issue warnings, Demerit et al. (2010) have shown how ESP are intended and used in a variety of different and sometimes contradictory ways by operational flood forecasters. How to evaluate and select the best strategy among the proliferation of techniques and the necessity of a thorough understanding and training by operational forecasters are among the challenges related to the regular operational use of statistical post-processors in hydrological ensemble prediction, identified in a blog post by Voisin et al. (2013). Moreover, Hamill (2018) emphasizes the problem related to the lengthy and high quality datasets and the need for post-processors suitable for environment where training data are limited.

De facto, there are only a few examples to date where the uncertainty post-processors for hydrological model outputs, used alone or in combination with preprocessing techniques applied to meteorological forecasts, have been operationally implemented and there is still a lot work to do to persuade decision makers to adopt more formally correct approaches as an alternative to the today common practice of directly using hydro-meteorological ensembles. The NOAA's National Weather Service (NWS), for example, is implementing a short-to long-range Hydrologic Ensemble Forecast Service (HEFS) (Demargne, 2014) that, based on separate modeling of the input and hydrologic uncertainties, includes a pre-processing of weather forecasts from multiple NWP models and a hydrologic ensemble post-processor, which aims to correct for systematic biases in streamflow. Bennett et al. (2014) describe the System for Continuous Hydrological Ensemble Forecasting (SCHEF), in Australia, that processes deterministic NWPs and use the ensemble rainfall forecast to force a hydrological model and produce streamflow forecast that are in turn post-processed through a hydrological error correction. Other existing operational or pre-operational systems for large-scale flood forecasting, like for example the European Flood Alert System (EFAS) of the European Commission Joint Research Centre (Thielen et al., 2009; Bartholmes et al., 2009) and the Global Flood Awareness System (GloFAS) are based on raw ensemble for delivering both the flood forecast and also uncertainty information to the stakeholders.

The aim of this paper is to improve “knowledge” on the proper use of uncertainty post-processors in the presence of ensembles and favor “persuasion”, two of the key elements in the Innovation-Decision process as defined by Rogers (2003) and adopted, for the case of seasonal forecasting, by Whateley et al. (2015).

This work corroborates and reinforces the perception that the usual meteorological and hydrological ensembles addressing only input forcing uncertainty do not provide a correct representation of the required predictive probability density, and an under-dispersion of the streamflow forecasts may be expected. This does not mean that ensemble approaches should not be used, but ensemble predictions should be drawn from the appropriate predictive densities (Herr & Krzystofowicz, 2015) and not, as usually done, generated by perturbing model parameters, initial and boundary conditions, etc.

In this paper, we propose to modify existing post-processing methodologies to produce predictive probability distributions of streamflow, which properly account for the uncertainty described by the spread of ensemble forecasts. The considered approaches, mainly based on Bayesian inference, have already been devised with the aim to improve the reliability and to correct for deficiencies in probabilistic forecasts calibration, but have been here adapted to incorporate ensemble spread information and to allow for an adequate assessment of the time heteroscedasticity of the estimated predictive distribution’s variance.

The remainder of the paper is organized as follows. Section 2 settles the necessity for probabilistic forecasts within a rational decision making process. Section 3 discusses the reasons for the improper use of the ensemble predictions to assess predictive probabilities and introduces how to modify existing uncertainty post-processors in order to account more appropriately for the ensemble prediction uncertainty. Section 4 describes the experimental setup: the Po River case study, the data and the verification tools. Section 5 presents and discusses the results obtained from a direct use of ensemble members as compared to those descending from the proposed alternative post-processing approaches. Conclusions are finally drawn in Section 6.

2 The need for probabilistic forecasts: decision making under uncertainty

In general terms, decision-making requires selecting an appropriate action $a \in A$ among a set A of possible actions to be taken. From a Bayesian perspective, robust decision-making is obtained by maximizing a utility function U, which is in general a subjective function $U[a, y]$ of the chosen action a as well as of a decision triggering variable y, as for instance the water level that may overtop an alert threshold or the dykes. The utility function U, usually expressing the decision maker’s propensity at risk, can also be a function of more than one triggering variable.

If the future value of the triggering variable y is known, the most appropriate action can be derived as:

$$a^* = \underset{a \in A}{arg\,max}\, U[a, y] \tag{1}$$

In reality most decisions have to be taken in advance, when y is still unknown, because the time to implement the decisions is too short, which requires predicting a future value for y. Therefore, instead of y one must produce the best estimate of y conditional to all the available information I, including prior observations and knowledge about the predictand, namely $f\{y|I\}$ and, being uncertain, the decision equation must be modified as:

$$a^* = \underset{a \in A}{arg\,max}\, E\{\, U[a, y]|I\} = \underset{a \in A}{arg\,max} \int_{\Omega_y} U[a, y]\, f\{y|I\}\, dy \tag{2}$$

As pointed out by Krzysztofowicz (1999), the available information is usually encapsulated in one or more model predictions, so that the scope of a probabilistic forecast becomes the “predictive probability”:

$$f\{y|I\} = f\{y|\hat{y}\,\} \tag{3}$$

namely the probability of the future occurrence y conditional upon $\hat{y}$ a single or multiple model forecasts.

Therefore, in order to appropriately use the probabilistic forecasts within a rational decision making process, forecasts must be provided either in terms of a continuous predictive density, as discussed above, or in the form of ensembles, drawn as to represent the mentioned predictive density in discrete form. And it is only by providing probabilistic forecasts correctly representing the shape of the predictive density, in continuous or in discrete form, that decisions can be selected by maximizing (or minimizing) the expected value of the utility function (Draper & Krnjajić, 2013) as in Eq. 2.

3 Using ensemble predictions to assess predictive probabilities

3.1 Why the direct use of ensemble predictions should be avoided

As shown in Figure 1, most of the currently available post-processors derive the predictive density as a measure of the uncertainty on a future event given (conditional upon) all the available information, which is usually encapsulated in one or more forecasts considered as known (namely certain and not affected by uncertainty) at the time of forecast. In reality, the predictor, the model forecast $\hat{y}$, is not perfectly known due to the influence of a wide number of uncertainties to be quantified, such as models structures, parameters, initial and boundary conditions, chaining of meteorological, hydrological and hydraulic forecasts, etc. To describe this uncertainty, perturbing input, parameters, initial and boundary conditions, etc. ensembles can be generated, which will describe our lack of knowledge. Moreover, when using several models, ensembles can additionally represent the variability of predictions due to our uncertainty on which is (if any) the correct model to be used.

But to understand why an ensemble generated by perturbing parameters and initial or boundary conditions cannot be directly considered as a representation of the predictive distribution let us discuss, without loss of generality the case of parameter perturbations, disregarding for the sake of clarity the input or initial and boundary conditions. Reasoning in the Normal space and in the case of a single forecasting model helps at clarifying concepts. If the joint predictand-predictor distribution is not a bivariate Normal, one can still transform the variables into the Normal space using either probability matching or the Normal Quantile Transform as proposed by several authors, thus re-conducting the problem to the linear regression case.

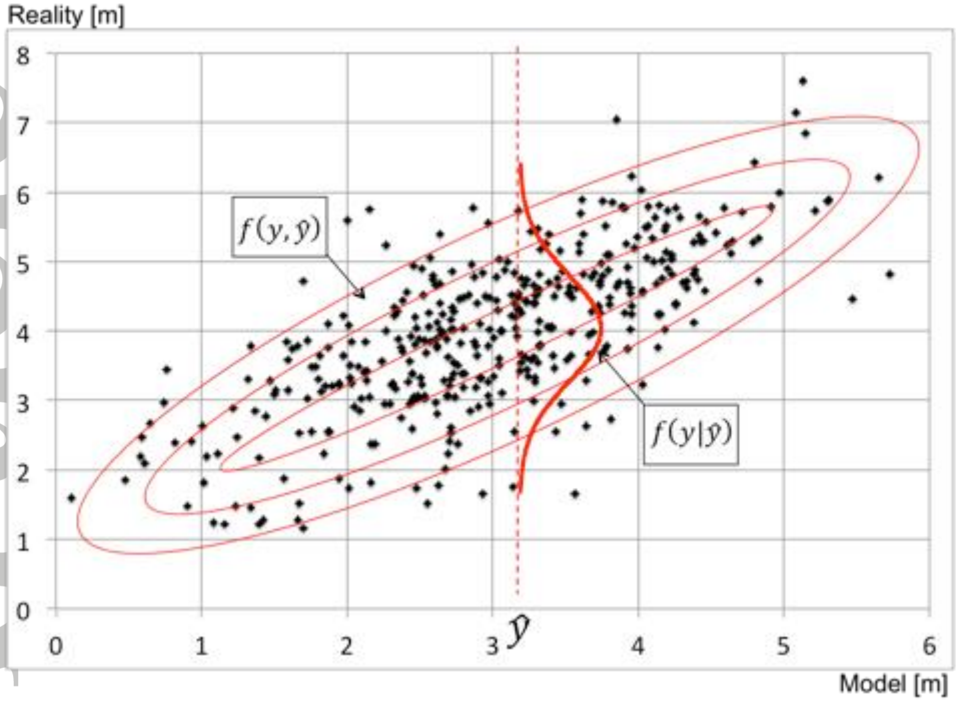


Figure 1. The predictive density defined as the probability density of the real quantity y conditional upon model forecasts $\hat{y}$, where $\hat{y}$ is considered as known (namely certain and not affected by uncertainty) at the time of forecast.

We represent with y the future true unknown quantity, and with $\hat{y}(\alpha) = \hat{y}|\alpha$ a prediction $\hat{y}$, which is here conveniently represented as a function depending on a set of parameter values α. If a specific set of parameter values $\alpha = \alpha^*$ is known then $\hat{y}(\alpha^*)$ becomes a deterministic quantity, usually addressed as "deterministic prediction"; if on the contrary we are not certain on a specific value for α to chose, then we must assume α as an uncertain quantity and provide a probability density function $f\{\alpha\}$ to describe it in order to produce a "probabilistic prediction". In this case $\hat{y}(\alpha)$ is no more a deterministic quantity but rather a random quantity, which probability obviously equals the probability of α, being a function of α, since its uncertainty only depends on the lack of knowledge we have on α. This can be written as (Benjamin & Cornell, 1970): $F\{\hat{y}(\alpha)\} = F\{\alpha\}$, or as $f\{\hat{y}(\alpha)\}\, d\hat{y}(\alpha) = f\{\alpha\}\, d\alpha$. The parameter density $f\{\alpha\}$ can either be assumed to be a known prior density or can be derived through a Bayesian inferential process by combining an assumed prior to the Likelihood of observations to obtain a posterior density.

Let us assume, for the sake of simplicity and without losing generality, the probability density and the probability distribution functions of parameters $f\{\alpha\}$ and $F\{\alpha\}$ to be known. By inverting the probability distribution function $F\{\alpha\}$ one can generate ensembles of equiprobable parameter values.

To approximate the probability density of the forecast $f\{\hat{y}(\alpha)\}$ one can then generate an ensemble of m parameter values $\alpha_i \quad \forall\, i \in 1, m$ and the corresponding forecast ensemble members $\hat{y}(\alpha_i) \quad \forall\, i \in 1, m$.

The ranked ensemble members $\hat{y}(\alpha_i)$ can be viewed as estimators of the m quantiles of $f\{\hat{y}(\alpha)\}$ but they cannot be assumed as representative of the predictive density, which in the deterministic prediction case, when the parameter values are preliminary estimated from a historical record, is defined as:

$$f\{y|\hat{y}\} = f\{y|\hat{y}(\alpha^{*})\} \tag{4}$$

or, in the more general case of parameters values considered known on the basis of a given prior (or posterior) probability density $f\{\alpha\}$, as:

$$f\{y|\hat{y}\} = \int_{\Omega_\alpha} f\{y|\hat{y}(\alpha)\}\ f\{\hat{y}(\alpha)\}\ d\hat{y}(\alpha) = \int_{\Omega_\alpha} f\{y|\hat{y}(\alpha)\}\ f\{\alpha\}\ d\alpha \tag{5}$$

In Eq. 5, Ω_α represents the domain of existence of parameters, and $f\{\hat{y}(\alpha)\}\ d\hat{y}(\alpha) = f\{\alpha\}\ d\alpha$ because $dF\{\hat{y}(\alpha)\} = dF\{\alpha\}$.

$f\{\hat{y}(\alpha)\}$ describes the spread of the forecast as a function of the spread of perturbed parameter values regardless to the future unknown quantity of interest y. On the contrary, Eq. 5 describes the "expected" uncertainty of a future value y conditional upon a prediction $\hat{y}(\alpha)$, after "marginalizing" the effect of our lack of knowledge on the parameters.

For instance, when using an ensemble, on the assumption that the m values α_i are equiprobably distributed, namely $Prob\{\hat{y}(\alpha_i)\} = Prob\{\alpha_i\} = 1/m$, Eq. 5 becomes:

$$f\{y|\hat{y}\} = \frac{1}{m}\sum_{i=1}^{m} f\{y|\hat{y}(\alpha_i)\} \tag{6}$$

which shows that the predictive density $f\{y|\hat{y}\}$ largely differs from the density $f\{\hat{y}(\alpha)\}$ described by the ensemble spread.

To summarize, whereas the $f\{\hat{y}(\alpha)\}$, based on the perturbation of the parameters, is mainly focused at describing the spread of model forecasts $\hat{y}(\alpha)$, the predictive probability focuses on the uncertainty of a future unknown quantity y conditional upon the knowledge of the forecasts $\hat{y}$.

What is shown for parameter uncertainty can be extended to all the other sources of uncertainty such as input, initial and boundary conditions uncertainty, with the result that what is commonly known as "meteorological" or "hydrological" ensembles obtained by perturbing parameters, initial and boundary conditions, do not represent the quantiles of the predictive density, but rather the spread of model forecasts as a function of the mentioned perturbations.

It is also true, as shown in Figure 2a, that the real situation would imply uncertainty increasing at each stage in the models chain and even more if we would also question the validity of model structures, but the point that we intend to clarify in the next section, is to show how the final assessed predictor uncertainty (whatever it is, and in our case it will look as in Figure 2b) can be incorporated in the derivation of the predictive uncertainty, to be approached using the modified uncertainty post processors.

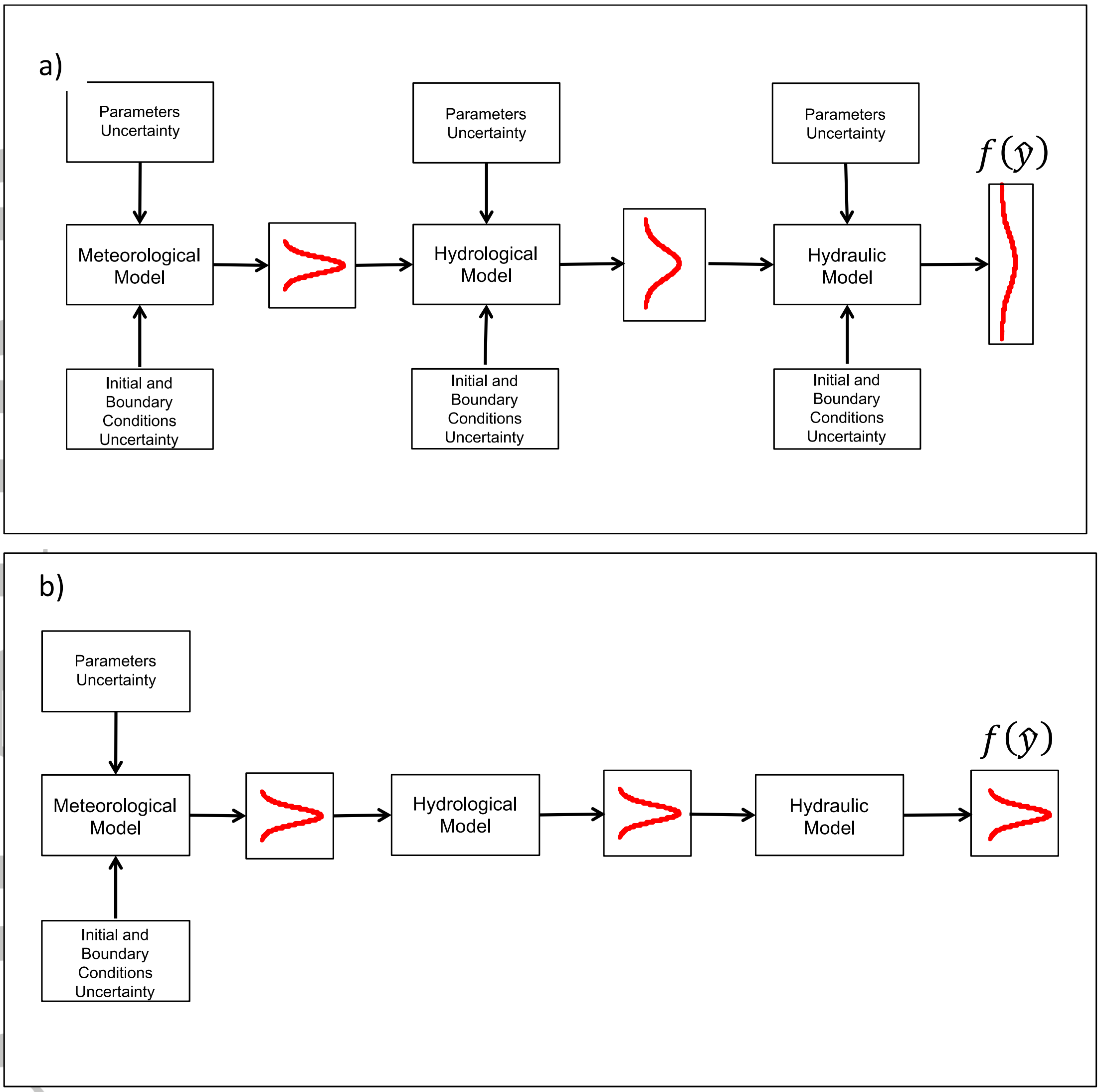


Figure 2. a) Assessed predictor's uncertainty comprehensively taking into account the meteorological-hydrological-hydraulic models parameters, initial and boundary conditions uncertainty; b) Assessed predictor's uncertainty only taking into account uncertainty due to meteorological model parameters, initial and boundary conditions

3.2 Alternative ways to incorporating ensemble information

A number of alternative ways have been implemented and analyzed in this work as a more correct approach to the direct use of ensembles to assess the predictive density, including a univariate approach not requiring additional assumptions and three multivarate approaches, requiring an additional assumption known as "exchangeability".

On the grounds that meteorological ensemble members lack individually distinguishable physical features, Fraley et al. (2010) addressed the concept of exchangeability, implicitly considered in other ensemble post-processing techniques (e.g. the best member dressing proposed by Roulston and Smith (2003)). In their work, BMA (Raftery, 1993; Raftery et al., 2005) and EMOS (Gneiting et al., 2005) model parameters have been constrained to be equal within each exchangeable group to account for exchangeability. Similarly, in this work, exchangeability is also exploited within the MCP approach (Todini, 2008; Coccia, 2011; Coccia & Todini, 2011), by ranking the ensemble predictions in ascending order at each time step and using, as predictors, the quantiles instead of the actual ensemble member. Figures 3a and 3b show an example on how the ensemble members are converted into the corresponding quantiles. In practice, the ensemble forecasts are resampled according to rank.

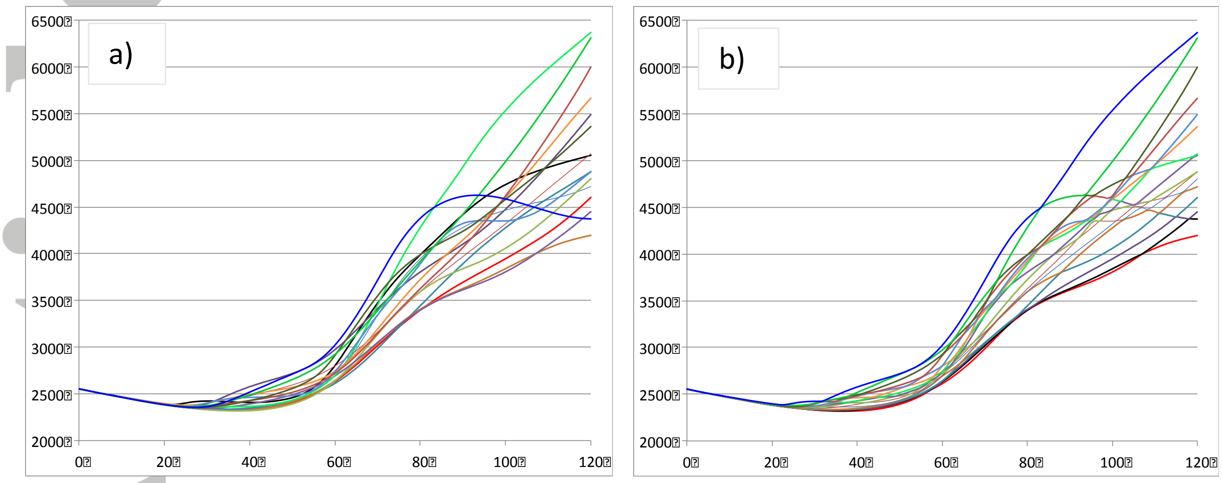


Figure 3. The 16 member forecasts shown in Figure 3a are re-sampled in Figure 3b in order to provide, at each step in time and at each forecasting horizon, 16 ascending order quantiles, each of which is now considered as a predictor.

For instance, for a given forecasting time, the largest forecasts at each forecasting horizon will be assumed as a representation of the largest quantile, the same applies to the second largest, the third largest and so on till the smallest quantile. In this way, a succession is created, so that the i^{th} quantile will remain consistently the i^{th} quantile, independently from the time of forecast and the forecasting horizon. Because the i^{th}quantile will remain consistently the i^{th}quantile at each time step it is now possible to use multivarate processors such as BMA or MCP based on m predictors, where m is the number of quantiles.

The different predicting densities were obtained:

1) by using the ensemble mean within the MCP univariate model framework, without needing to invoke exchangeability, but assuming the ensemble mean being affected, at each point in time, by a different estimation error, which variance is derived from the ensemble spread;

2) by deriving a predictive probability distribution for each ensemble member and then taking the expected predictive density on the assumption of equiprobable ensemble members; or

3) by deriving a predictive probability distribution for each ensemble member and then using BMA to obtain the predictive density conditional to all the ensemble members; or

4) by using all the ensemble members in the MCP multi-model framework (each member of the ensemble is considered as a different model prediction) to directly derive the predictive density conditional on all the ensemble members.

All the approaches were developed in the Gaussian space as described below.

3.3 Transformation into the Gaussian space

The observations y_t $\forall\, t \in 1, T$ where T is the length of the observations record are ranked in ascending order and converted into the Gaussian space either via the Normal

Quantile Transform (NQT) or by fitting appropriate parametric probability functions, to obtain η_t the Normal image of the observations. Similarly, for any given forecasting lead time the $m \times T$ ensemble predictions, $\hat{y}_{i,t}\ \forall\ i \in 1, m;\ \forall\ t \in 1, T$ with m the number of ensemble members are ranked in ascending order and converted into the Gaussian space to obtain $\hat{\eta}_{i,t}$, the Normal image of the m members of the ensemble forecasts, as done by Reggiani and Weerts (2008). The reason for ranking all the ensemble members together is to underline that no a-priori preference is set on the single ensemble members, which are assumed equi-probable. In the Gaussian space, when using the NQT the converted variables will result into standard Normal variables, namely $\eta_t \equiv N(0,1)$ and $\hat{\eta}_{i,t} \equiv N(0,1)$.

3.4 Univariate Model Conditional Processor

For any given forecasting lead-time, by computing the expected value of the ensemble predictions in the Gaussian space, one obtains a new variable $\bar{\eta}_t$ defined as:

$$\bar{\eta}_t = \frac{1}{m}\sum_{i=1}^{m}\hat{\eta}_{i,t} \tag{7}$$

which will be Normally distributed with zero mean $\mu_{\bar{\eta}_t} = 0$ and variance $\sigma^2_{\bar{\eta}_t} = Var\{\bar{\eta}_t\}$.

Based on the properties of the multivariate Normal distribution (Mardia et al., 1979), by considering $\bar{\eta}_t$ as a deterministic predictor, the MCP univariate estimator of the conditional mean and variance of the predictive density of η_t given $\bar{\eta}_t$ would give:

$$\begin{cases}\mu_{\eta_t|\bar{\eta}_t} = w\,\bar{\eta}_t \\ \sigma^2_{\eta_t|\bar{\eta}_t} = 1 - w^2\,\sigma^2_{\bar{\eta}_t}\end{cases} \tag{8}$$

with the weights $w = \frac{\gamma_{\eta\bar{\eta}}}{\sigma^2_{\bar{\eta}_t}}$ and where $\gamma_{\eta\bar{\eta}} = \frac{1}{T-1}\sum_{t=1}^{T}\eta_t\bar{\eta}_t$ is the covariance of the observations and the ensemble mean prediction.

As can be seen from Eqs. 8, in analogy to the linear regression, the original MCP algorithm considers $\bar{\eta}_t$ as perfectly known. In order to account for the uncertainty on $\bar{\eta}_t$, now represented by the ensemble spread, we may proceed as in the case of a predictor affected by measurement errors and modify accordingly the MCP algorithm from what it essentially is, a linear regression in the Gaussian space, into a Deming (1943) type regression. This can be done by evaluating $S_{\hat{\eta}_t}$, the variance of the ensemble, defined at each step as:

$$S_{\hat{\eta}_t} = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\eta}_{i,t} - \bar{\eta}_t\right)^2 \tag{9}$$

to be then converted into the variance of the mean $S_{\bar{\eta}_t} = S_{\hat{\eta}_t}/m$. This is necessary because our predictor is the ensemble mean and the variance must be estimated accordingly.

When assuming that the predictor is uncertain, the MCP algorithm, formally maintains the expression of Eqs. 8, the weights now becoming $w_t = \frac{\gamma_{\eta\bar{\eta}}}{\sigma^2_{\bar{\eta}_t} + S_{\bar{\eta}_t}}$ to give:

$$\begin{cases}\mu_{\eta_t|\bar{\eta}_t} = w_t\,\bar{\eta}_t \\ \sigma^2_{\eta_t|\bar{\eta}_t} = 1 - w_t^2\,\sigma^2_{\bar{\eta}_t}\end{cases} \tag{10}$$

As can be seen from Eqs.10, this algorithm, which only differs from Eqs. 8 because the weights w_t now vary in time, is rather simple and can be easily extended to the multi-model case, when each model is producing a prediction ensemble.

Please note that the univariate approach is the only one considered in this work not requiring re-ordering of the ensemble members. All the additional approaches described below require reordering of the ensemble members on the assumption of exchangeability (Gneiting et al., 2008).

3.5 Uniform weighting

For any given lead time and for each time $t \in 1, T$, $\hat{\eta}_{i,t}\ \forall\ i \in 1, m; \forall\ t \in 1, T$ is ranked in ascending order, namely $\hat{\eta}^o_{1,t} \leq \hat{\eta}^o_{2,t} \leq \cdots \leq \hat{\eta}^o_{m,t}\ \forall t$. The ith ranked member can now be interpreted as the ith quantile describing the model prediction uncertainty. The T elements of each quantile are not necessarily zero mean and unit variance, a property only holding when all the $m \times T$ observations are considered. Therefore we assume $\mu_{\hat{\eta}^o_i} \neq 0$ and $\sigma^2_{\hat{\eta}^o_i} \neq 1\ \forall\ i \in 1, m$

Using the m predictors, m univariate MCP models can be set up leading to:

$$\begin{cases} \mu_{\eta_t|\hat{\eta}^o_{i,t}} = w_t \left(\hat{\eta}^o_{i,t} - \mu_{\hat{\eta}^o_i} \right) \\ \sigma^2_{\eta_t|\hat{\eta}^o_{i,t}} = 1 - w\,^2_t\, \sigma^2_{\hat{\eta}^o_i} \end{cases} \tag{11}$$

In Eq. 11, $w_t = \frac{\gamma_{\eta\hat{\eta}^o_i}}{\sigma^2_{\hat{\eta}^o_i} + S_{\hat{\eta}^o_{i,t}}}$; $\mu_{\hat{\eta}^o_i} = \frac{1}{T}\sum_{t=1}^{T} \hat{\eta}^o_{i,t}$; $\sigma^2_{\hat{\eta}^o_i} = \frac{1}{T-1}\sum_{t=1}^{T} \left(\hat{\eta}^o_{i,t} - \mu_{\hat{\eta}^o_i} \right)^2$; $\gamma_{\eta\hat{\eta}^o_i} = \frac{1}{T-1}\sum_{t=1}^{T} (\eta_t - \mu_\eta) \left(\hat{\eta}^o_{i,t} - \mu_{\hat{\eta}^o_i} \right)$ and $S_{\hat{\eta}^o_{i,t}} = \alpha_i\, S_{\hat{\eta}_t} = \frac{\alpha_i}{m-1}\sum_{j=1}^{m} \left(\hat{\eta}_{j,t} - \bar{\eta}_t \right)^2$, with α_i a coefficient arising from the ordering of the ensemble members, which distributes a portion of the variance of the ensemble to each ranked individual member as described in Appendix A.

In Uniform Weighting (UW), one estimates the "expected predictive density" on the basis of the m univariate conditional predictions assuming uniform weighting $w_i = 1/m$, underlining the fact that all the ensemble members are considered to have the same likelihood to occur. This leads to:

$$f(\eta_t|\boldsymbol{\hat{\eta}^o_t}) = \frac{1}{m}\sum_{i=1}^{m} f\left(\eta_t|\hat{\eta}^o_{i,t}\right) \tag{12}$$

with $f(\eta_t|\boldsymbol{\hat{\eta}}^o_t)$ an approximately Normal distribution with mean $\mu_{\eta_t|\boldsymbol{\hat{\eta}}^o_t}$ and variance $\sigma^2_{\eta_t|\boldsymbol{\hat{\eta}}^o_t}$, defined as:

$$\begin{cases} \mu_{\eta_t|\boldsymbol{\hat{\eta}}^o_t} = \frac{1}{m}\sum_{i=1}^{m}\ \mu_{\eta_t|\hat{\eta}^o_{i,t}} \\ \sigma^2_{\eta_t|\boldsymbol{\hat{\eta}}^o_t} = \frac{1}{m}\sum_{i=1}^{m} \left(\sigma^2_{\eta_t|\hat{\eta}^o_{i,t}} + \mu^2_{\eta_t|\hat{\eta}^o_{i,t}} \right) - \mu^2_{\eta_t|\boldsymbol{\hat{\eta}}^o_t} \end{cases} \tag{13}$$

The expression of the variance in Eq. 13 is here derived as the variance of the probability distribution resulting from the linear combination of m probability distributions.

3.6 Bayesian Model Averaging

After performing the conversion of observations and ordered predictions into the Gaussian space, as described for the Uniform Weighting, m univariate conditional probability densities can be set up using the m predictors as per Eq. (12).

Given the univariate conditional densities, BMA can be applied by estimating posterior weights w_i via a modified EM algorithm (Dempster et al., 1977; code available at Vrugt, 2016) to account for the time heteroscedasticity of the predictive variance. Given the weights, the BMA estimator leads to:

$$f(\eta_t|\widehat{\boldsymbol{\eta}}_{\boldsymbol{t}}^{\boldsymbol{o}}) = \sum_{i=1}^{m} w_i\, f\left(\eta_t|\hat{\eta}_{i,t}^{o}\right) \tag{14}$$

with $f(\eta_t|\widehat{\boldsymbol{\eta}}_t^o)$ an approximately Normal distribution with mean $\mu_{\eta_t|\widehat{\boldsymbol{\eta}}_t^o}$ and variance $\sigma^2_{\eta_t|\widehat{\boldsymbol{\eta}}_t^o}$, defined by:

$$\begin{cases} \mu_{\eta_t|\widehat{\boldsymbol{\eta}}_t^o} = \sum_{i=1}^{m} w_i\, \mu_{\eta_t|\widehat{\eta}_{i,t}^o} \\ \sigma^2_{\eta_t|\widehat{\boldsymbol{\eta}}_t^o} = \sum_{i=1}^{m} w_i \left(\sigma^2_{\eta_t|\widehat{\eta}_{i,t}^o} + \mu^2_{\eta_t|\widehat{\eta}_{i,t}^o} \right) - \mu^2_{\eta_t|\widehat{\boldsymbol{\eta}}_{\boldsymbol{t}}^{\boldsymbol{o}}} \end{cases} \tag{15}$$

The expression of the variance in Eq. 15 is here derived as the variance of the probability distribution resulting from the linear combination of m probability distributions; although it appears as formally different from the one proposed by Raftery, 1993 and by Raftery et al., 2005, the two expressions coincide when, as in BMA, $\sum_{i=1}^{m} w_i = 1$.

3.7 Multivariate Model Conditional Processor

After performing the conversion of observations and ordered predictions into the Gaussian space, as described for the Uniform Weighting, the following multivariate MCP approach which considers uncertain predictors can be set-up in the form of:

$$\begin{cases} \mu_{\eta_t|\widehat{\boldsymbol{\eta}}_t^o} = \mathbf{w}_t^T(\widehat{\boldsymbol{\eta}}_t^o - \boldsymbol{\mu}^{\boldsymbol{o}}) \\ \sigma^2_{\eta_t|\widehat{\boldsymbol{\eta}}_t^o} = 1 - \mathbf{w}_t^T\, \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\eta}}_t^o\widehat{\boldsymbol{\eta}}_t^o}\, \mathbf{w}_{\boldsymbol{t}} \end{cases} \tag{16}$$

In Eq. (16) the weights are estimated as $\mathbf{w}_t^T = \boldsymbol{\Sigma}_{\eta_t\widehat{\boldsymbol{\eta}}_t^o}\left(\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\eta}}_t^o\widehat{\boldsymbol{\eta}}_t^o} + \mathbf{S}_{\widehat{\boldsymbol{\eta}}_t^o}\right)^{-1}$ with: $\widehat{\boldsymbol{\eta}}_t^o(i) = \hat{\eta}_{i,t}^o$; $\mathbf{S}_{\widehat{\boldsymbol{\eta}}_t^o}(i,i) = \frac{\alpha_i}{m-1}\sum_{i=1}^{m}\left(\hat{\eta}_{i,t}^o - \bar{\eta}_t^o\right)^2$; $\mathbf{S}_{\widehat{\boldsymbol{\eta}}_t^o}(i,j) = 0\ \forall\ i \neq j$; $\boldsymbol{\mu}^{\boldsymbol{o}}(i) = \frac{1}{T}\sum_{t=1}^{T}\hat{\eta}_{i,t}^o$; $\boldsymbol{\Sigma}_{\eta_t\widehat{\boldsymbol{\eta}}_t^o}(i) = \frac{1}{T-1}\sum_{t=1}^{T}\eta_t\hat{\eta}_{i,t}^o$; $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\eta}}_t^o\widehat{\boldsymbol{\eta}}_t^o}(i,j) = \frac{1}{T-1}\sum_{t=1}^{T}\hat{\eta}_{i,t}^o\hat{\eta}_{j,t}^o$ and α_i defined in Appendix A in analogy to what mentioned in section 3.5.

Please note that the BMA weights w_i do not vary in time while the multivariate MCP weights do vary in time since they are updated every time step.

4 Testing alternative post-processing approaches: experimental setup

The four post-processors and the variants introduced to account for the ensemble spread described in the previous section were tested on the basis of a streamflow ensemble data set available on a real case study of streamflow forecast for one section of the Po River in Italy. Specifically, we analyzed the univariate version of the MCP, that do not require reordering of the ensemble members, and three approaches that rely on the assumption of exchangeability, namely the Uniform Weighting, the BMA and the multivariate MCP. As opposed to the direct use of unprocessed, streamflow ensembles, the four investigated post-processors identify acceptable predictive probability density functions as per the applied statistical tests.

In this section, the case study and the data set are firstly described and then the verification tools used to evaluate the alternative approaches are presented.

4.1 The Po River case study

The case study used in this work relates to operational flood forecasts at the gauged section of Pontelagoscuro that is conventionally recognized as the closure cross section of the Po River (see Figure 4), the longest river in Italy. The Po River, with its 141 main tributaries, flows ~650 km eastward across northern Italy, from the northern-eastern Alps to a delta projecting into the Adriatic Sea near Venice. Its drainage basin extends over an area of about 74’000 km^2 at the delta; approximately 71’000 km^2 are in Italy, nearly one-third of which constitutes the alluvial plain of the river, known as the Po Valley, where more than 16 million people live. The remaining parts are mostly located in Switzerland and, for very a small portion, in France.

At Pontelagoscuro, 96 km from the sea covering around 70’000 km^2, although the average Po River discharge is about 1’500 m^3/s, much larger discharges occur during floods, reaching 10’300 m^3/s in the great flood of 1951, and 9’600 m^3/s in a more recent flood event occurred in October 2000.

ARPA-SIM, the Hydro-Meteorological Service of the Emilia-Romagna Regional Agency for Environmental Protection, which operationally runs the real-time flood forecasting systm fro the Po river, provided hydrological prediction ensembles based on COSMO LEPS (Limited Area Ensemble Prediction System) ensemble members routed through a cascade of a hydrological distributed model, the TOPKAPI (Ciarapica & Todini, 2002; Liu & Todini, 2002) and a full dynamic hydraulic model, the SOBEK (Stelling & Verwey, 2005; Deltares, 2014), to the cross section of Pontelagoscuro. COSMO-LEPS is the Limited Area Ensemble Prediction System developed within COSMO consortium (Consortium for Small-Scale Modelling) in order to improve the short-to-medium range forecast of extreme and localized weather events and it is made up of 16 integrations of COSMO-Model with 7 km resolution starting on initial and boundary conditions from 16 representative members of an ECMWF-EPS super-ensemble.

No ensembles describing the uncertainty of the hydrological and hydraulic models were available for this example. Therefore, the only additional uncertainty we coupled with the post-processor is provided by the spread of the COSMO-LEPS meteorological ensemble. Nevertheless, the developed approach doesn’t prevent to correctly include, whenever available, the additional hydrological and hydraulic uncertainty in the form of ensemble.

Observations and forecasts are sampled in time every 6 hours $(\Delta t = 6\ h)$, the latter consisting in an array of hourly streamflow forecasts for the next 120 hours. The available dataset is split into a training set that is used to calibrate the post-processors (calibration period: 9 November 2012 – 2 July 2013) and a validation set that is used for testing (validation period: 24 July 2013 – 28 February 2015).

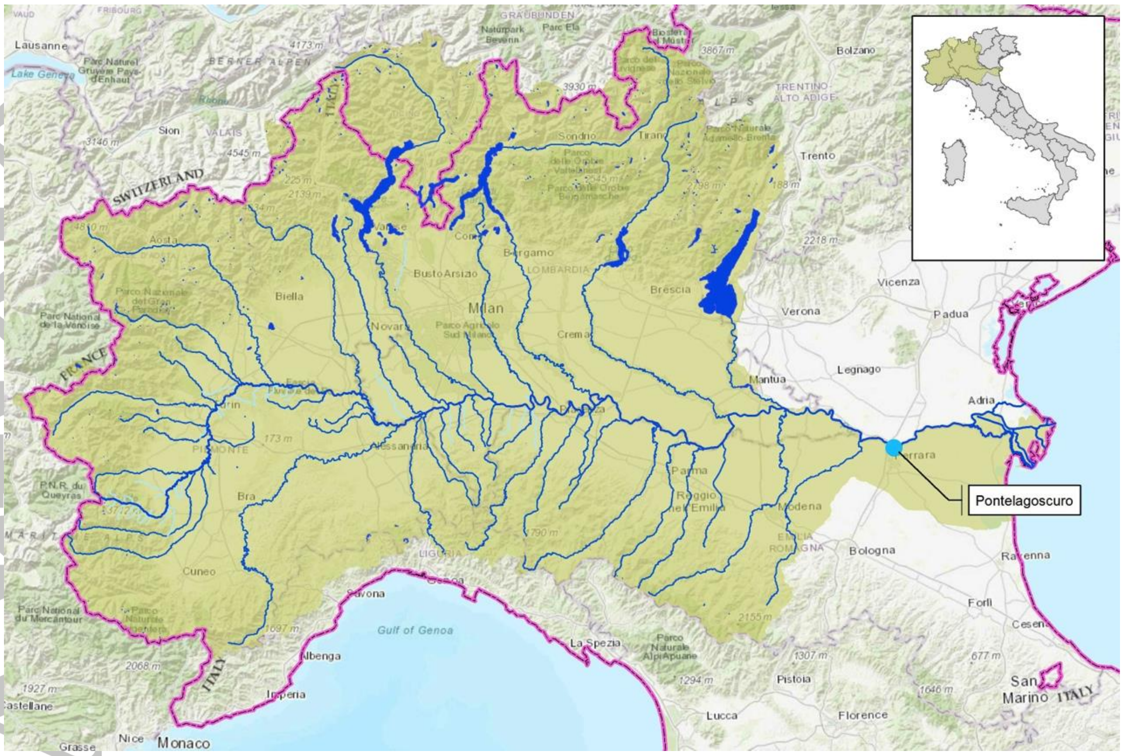


Figure 4. Map of the Po River catchment. Morphology and location of the Pontelagoscuro hydrometric station.

4.2 Evaluation Criteria

The assessment of the "quality" of a forecast is commonly regarded as the degree to which the forecast corresponds to what actually happened (Murphy, 1993). When dealing with probabilistic forecast of continuous predictands, standard goodness-of-fit metrics used for deterministic forecasts, do not allow a complete and fair evaluation of the forecast performance and the use of appropriate diagnostic approaches is recommended. This is because the focus of a probabilistic forecast is not limited to correctly predict the expected future value, but rather to assess the full predictive density, which is the essential tool for rational decision making, such as for instance the Bayesian decision approach.

In this paper, we consider the "accuracy" and the "calibration" of the forecasts as diagnostic approaches and primary requirements that we verified in the assessment of its probabilistic correctness. The former attribute, accuracy, defines the level of agreement between individual pairs of forecasts and observations, considering a single deterministic value as representative of the entire predictive distribution. The calibration attribute, often referred as reliability, is here intended as the statistical consistency between the obtained predictive distribution and the observed streamflow time series, embracing the paradigm that the main goal of post-processing is to achieve well calibrated and yet sharp probabilistic predictions (Gneiting et al., 2007).

For accuracy evaluation, we rely on the mean value of the ensemble or of predictive distribution as representative of the best deterministic forecast and make use of traditional deterministic performance measure such as the Nash-Sutcliffe coefficient, NS, the Root Mean Square Error, RMSE, and the mean of the absolute errors on forecast, MAE.

As in Verkade et al., (2013), conditional accuracy was determined by calculating the above-mentioned verification metrics for different levels of the non-exceedance climatological probability P in the sample of observations, nominally ranging from 0 to 1.

When P = 0 all the pairs are considered and essentially corresponds to an unconditional verification; when, for example, P=0.90 only the data pairs with observations included in the top 10% of the sample are considered, thus providing an evaluation that is conditional on the magnitude of the verifying observations.

Reliability assessment, with the aim of verifying if the forecast is correct under a statistical viewpoint, is performed through verification rank histograms or Talagrand diagram, (Hamill, 2001; Talagrand et al., 1997) for the raw ensemble data set, while the approach recently proposed by Laio and Tamea (2007) based on the use of the Probability Integral Transform (PIT), is used for the probabilistic forecasts from post-processors. In the latter approach, the PIT values z_i for each forecasting occasions and lead time is defined as the predictive cumulative distribution function (CDF) evaluated at the observation ($z_i = F\{y_i\}$), and plotted against their empirical cumulative distribution function, R_i/n. The probabilities z_i should have a uniform distribution, i.e., the observations should look like random samples from the predictive distribution, in which case the points lie close to the bisector of the diagram. Moreover, the shape of the resulting curve reveals insights into the properties of the predictive distribution: problems related to under or over dispersion, positive or negative bias can be detected.

Additional information about the inclusion of the ensemble spread into investigated statistical post-processing approaches will be provided in terms of effects induced on the resulting conditional variances and heteroscedasticity. Finally, the raw ensembles and the probability forecasts bands derived from post-processing will be compared and discussed for two exemplifying events, one in the calibration and one in the validation period.

5 Results and discussion

Specifically, this section: 1) presents the outcomes from the direct use of the ensembles and 2) describes the main results of ensemble streamflow post-processing according to the proposed novel variants, highlighting the benefit derived by the predictive distribution estimate.

5.1 Raw Ensemble Forecasts

Accuracy verification metrics for the raw ensembles are shown in Figure 5 for lead times from 6 to 96 hours; the corresponding performance metrics evaluated for the investigated post-processors are also presented in the same figure. All the performance measures are essentially based on the expected value of the prediction and test how this approaches the observations. In practice, they all measure the quality of a deterministic model prediction.

The ensemble streamflow mean, considered as representative of the best deterministic forecast, is found to be skillful, particularly for shorter lead times, but indicates a clear quality reduction with increasing lead time. This behavior, which is evident also for post-processors, is attributable to the fact that moving beyond the concentration time of the basin the effect of observed precipitation becomes smaller while the influence of precipitation forecasts is more and more important which explains the deterioration of the performance.

Regarding the unconditional sample (i.e. all the dataset corresponding to non-exceedance climatological probability P=0) and the calibration period (Figures 5a, 5c, 5e), the NS value decreases from 0.998, for a forecasting horizon of 6 hours, to 0.915 for a lead-time of 96 hours; the RMSE goes from to 46.84 m^3/s to 337.21 m^3/s, and the mean of the absolute errors varies from 29.55 m^3/s to 229.39 m^3/s. The same patterns are substantially confirmed in the validation period, as shown in Figures 5b, 5d and 5f and Table 1.

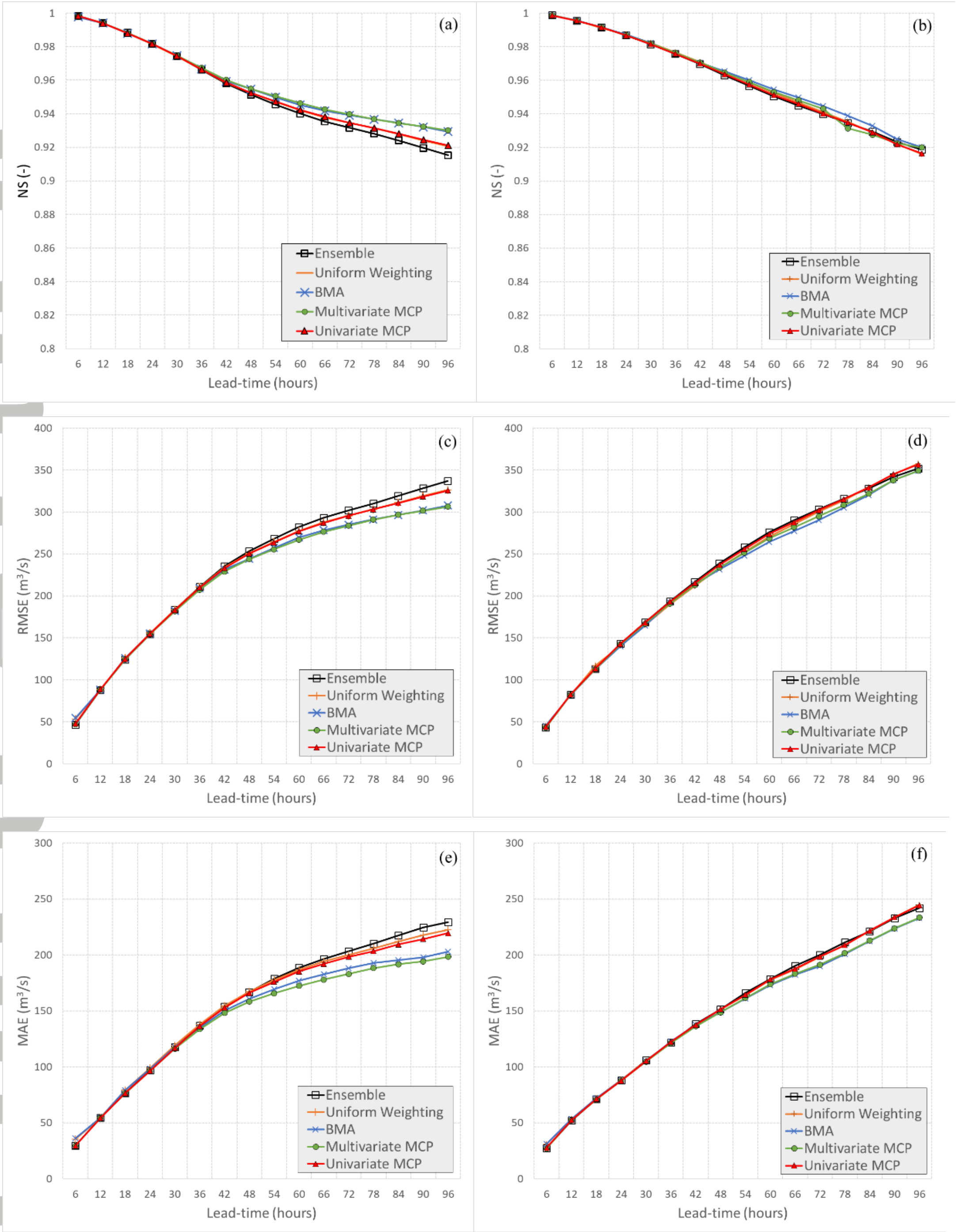


Figure 5. Performance measures, Nash–Sutcliffe coefficient (NS), Root Mean Square Error (RMSE), Mean of the Absolute Error (MAE), for the mean of the raw ensemble forecasts and the expected values resulting from the 4 post-processors (Uniform Weighting, BMA, multivariate MCP, univariate MCP) for predictive probability estimation (lead-time varying from 6 to 96 hours): (**a**), (**c**), (**e**) calibration period; (**b**), (**d**), (**f**) validation period.

Accuracy performances are seen to be strongly conditional on the magnitude of the observed discharge, and highlight a limited ability to correctly predict large discharge values for longer lead times. As shown in Table 1, at climatological probability equal 90% (P=0.90) in the calibration period, which corresponds to 79 pairs in our experimental setup, verification metrics reveal a heavier deterioration for increasing lead times: NSE varies from 0.992 to 0.574 for lead times of 6h and 96h respectively; RMSE equals 90.41 m^3/s and

690.01 m^3/s, while MAE ranges from 71.74 to 596.12 m^3/s within the same interval of forecasting lead times. Conversely, conditional performance metrics for the validation period at P=0.90 (82 pairs) show a less evident dependence on the lead-time: NSE varies from 0.996 to 0.801; RMSE equals 85.37 m^3/s and 637.34 m^3/s respectively, while MAE ranges from 63.09 to 549.96 m^3/s.

		Calibration period				**Validation period**			
		6 hours		**96 hours**		**6 hours**		**96 hours**	
		P = 0	**P = 90**	**P = 0**	**P = 90**	**P = 0**	**P = 90**	**P = 0**	**P = 90**
	NS (-)	0.998	0.992	0.915	0.574	0.998	0.996	0.918	0.801
Ensembles	**RMSE (m^3/s)**	46.84	90.41	337.21	690.01	43.2	85.37	351.94	637.34
	MAE (m^3/s)	29.55	71.74	229.39	596.12	27.34	63.09	241.89	549.96

Table 1. Accuracy metrics, Nash–Sutcliffe coefficient (NS), Root Mean Square Error (RMSE), Mean of the Absolute Error (MAE), for the mean of the raw ensemble forecasts at different levels of the non-exceedance climatological probability P in the sample of observations.

In terms of probabilistic forecasts, raw ensembles, as already shown in several other studies both in weather and hydrologic forecasting applications (Raftery et al., 2005; Gneiting et al., 2008; Bougeault et al., 2010; Hemry et al., 2015; Park et al., 2008), are often miscalibrated and fail to meet the reliability requirement. Figure 6 represents the analysis in terms of verification rank histograms, performed at $t + 8\,\Delta t$, namely 48 hours in advance. Other lead-times showed overall similar behavior, but are not presented here for the sake of brevity. On the assumption that the ensemble members do represent the different quantiles at each step in time, the relative frequency of the observations that lie within the corresponding forecast intervals are evaluated. If the predictive density is correctly represented, then one would expect these frequencies to be uniformly distributed in all the quantiles. Conversely, as can be seen from Figure 6, in both calibration (left) and validation (right) periods, typical U-shaped rank histograms are obtained, indicating under-dispersion and suggesting that the estimated variance of the distribution is too low. Most of observations falls in the extreme quantiles (~86% in calibration and ~92% in verification), indicating that a large quantity of predictions underestimating or overestimating the flood values. Moreover, it should be noticed that the majority of frequency values, shown as a histogram for the $m + 1$ coverages generated by the m quantiles $\left(0 \leftrightarrow \frac{1}{m+1}, \dots, \frac{i-1}{m+1} \leftrightarrow \frac{i}{m+1}, \dots, \frac{m}{m+1} \leftrightarrow 1\right)$, lie far from the expected value (solid red horizontal line) and outside the confidence limits estimated with the Wilson (1927) bounds (dashed red horizontal lines).

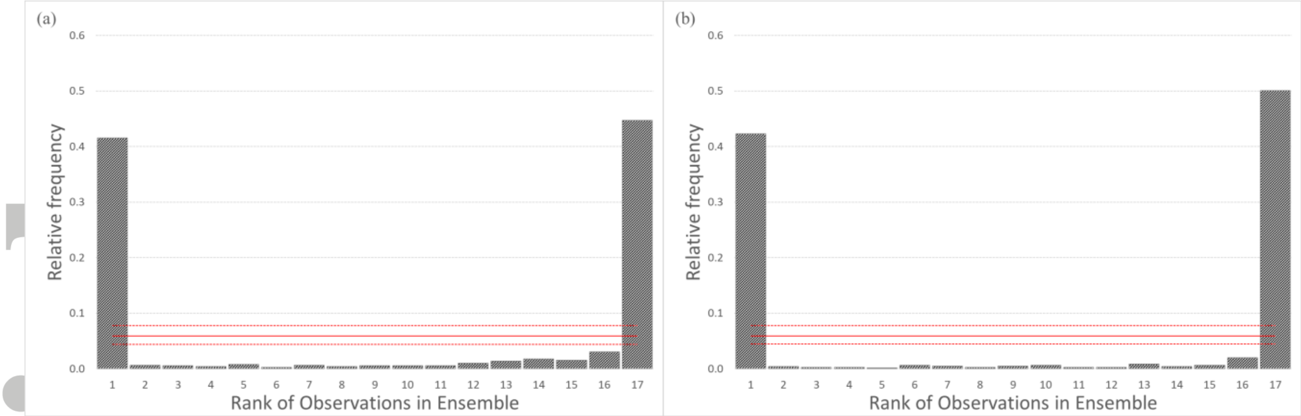


Figure 6. Verification rank histograms along with the Wilson (1927) 95% probability confidence intervals, for the 48-h streamflow ensemble forecasts based on COSMO-LEPS: (**a**) calibration period (*T* = 794) and (**b**) validation period (*T* = 825). Note that the observed relative frequency is by far outside the range of the estimated confidence intervals.

5.2 Modified uncertainty post-processors

The four modified approaches described above are here analyzed and compared to their original version. Specifically, the quality, in terms of accuracy and calibration/reliability, of the estimated predictive distributions obtained post-processing streamflow ensembles are shown and discussed.

In the Gaussian space the conditional predictive distributions deriving from post-processing approaches, such as HUP (Krzysztofowicz, 1999) or MCP (Todini, 2008), resulting from the linear combination of model predictions, generate constant predictive variance if the ensemble spread is not explicitly accounted for in the derivation of weights. On the contrary, approaches such as UW or BMA, which estimate the expected conditional predictive distribution, already generate a predictive variance variable in time as a function of the ensemble spread (Eq.13 and Eq. 15 respectively). Nonetheless, consistently with the modifications proposed for the univariate and multivarate MCP, one should also account in UW and BMA for the single model prediction uncertainty, as measured by the ensemble spread, in the weights station phase. Therefore, each of the approaches, as a consequence of the modifications described in section 3, allows the predictive variance in the Gaussian space to vary at each time step by accounting for the information on the uncertain predictions provided by the ensemble spread. The effect introduced by this modeling setup has a major impact for larger lead times where the predictability is lower and thus the variability within the ensemble is more pronounced. In Figure 7, focusing on the 96 h lead-time, where the ensemble spread is larger, both the spread of the raw ensemble, $S_{\hat{\eta}_t}$, and the conditional variance obtained through different post-processors, indicated as $\sigma^2_{\eta_t|\bar{\eta}_t}$ for the univariate MCP and $\sigma^2_{\eta_t|\hat{\boldsymbol{\eta}}^o_t}$ for the Uniform Weighting, the BMA and the multivariate MCP, are shown in the Gaussian space for comparison purposes.

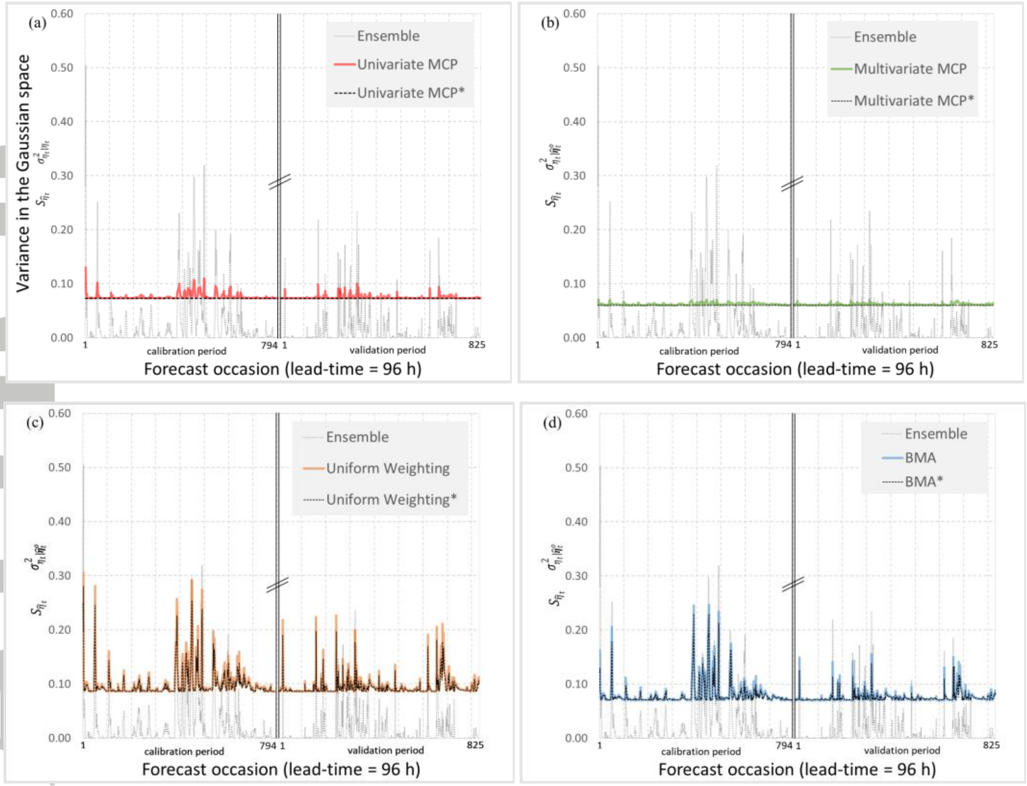


Figure 7. Predictive variances $\sigma^2_{\eta_t|\bar{\eta}_t}$ and $\sigma^2_{\eta_t|\hat{\boldsymbol{\eta}}^o_t}$ of the four post-processing approaches (7a – univariate MCP; 7b multivariate MCP; 7c Uniform Weighting; 7d BMA) in the Gaussian space for a 96 h lead-time. The variance of the ensemble spread $S_{\hat{\eta}_t}$ is also shown in light grey. Calibration period (*T*= 794) is shown on the left side of figures, and validation period (*T* = 825) on the right. The asterisk (*) indicates the approach applied without taking into account the ensemble spread.

Specifically, Figure 7 shows the original predictive variances not accounting for the ensemble spread in the derivation of the conditional distributions as well as the variances obtained with the modified algorithms. As can be seen in Figure 7a and Figure 7b, the conditional distributions for the univariate MCP formulation and the multivarate MCP respectively, show a variance which varies in time but the variations are limited if compared to the ones produced by the UW and the BMA approaches (Figures 7c and 7d), which compute the expected value of the univariate conditional distributions as an approximation to the actual conditional distribution. Nonetheless, UW and BMA, already sensitive to the ensemble spread even without the introduction of the ensemble uncertainty in the estimation phase, additionally increases the predictive variance, as can be seen from Figures 7c and 7d.

Figure 7 also reveals that the raw ensemble approach produces a very small bottom-line variance with large variations as a function of the ensemble spread. On the contrary, all the post-processors although following the variability of the raw ensemble, generate larger bottom-line variances but are less sensitive to the local variations of the ensemble spread. Among the post-processors the multivarate MCP approach provides the smallest base-line variances and limited sensitivity to the ensemble spread. The UW instead shows the largest base-line variance and sensitivity.

On the overall Figure 7 shows that the appropriate handling of the uncertainty expressed by the ensemble spread slightly increases the heteroscedasticity in the variance. This effect is not very large because in this work, as pointed out in the introduction, only the meteorological input uncertainty has been taken into account. The effect would be much stronger if considering all the sources of uncertainty in the forecasting chain, as discussed in section 3.1.

The unconditional performances of the four methods with respect to forecast accuracy considering the expected values of the corresponding predictive distributions are compared in Figure 5 for different forecasting horizons. The multivariate MCP and the BMA outperform other post-processors for longer lead-times; particularly in the calibration period, they are characterized by higher NS and by lower RMSE and MAE values (Figures 5a, 5c and 5e) that become more valuable for lead-times longer than 42 h. In the calibration period, all post-processors show improvements in accuracy compared to the raw ensemble mean. Uniform Weighting and univariate MCP show very similar values of the evaluation metrics, as well as BMA and multivariate MCP. Differences among post-processors as well as with the unprocessed ensemble mean are even less evident within the validation period. It is also noticed that the forecast accuracy, regardless of the metric used, as for the raw ensembles, sensibly decreases when the forecast lead-time increases particularly in the calibration set.

Moving from raw to post-processed streamflow ensemble has only a limited influence on performance conditional to the value of observations for the calibration period (Table 2). Considering a probability of 90% (P=0.90), conditional NS, RMSE and MAE for all the post-processors show similar values and negligible differences compared to the unconditional ones at short lead times that increase with increasing lead time. For example, the NS metric for the univariate MCP varies from 0.991 to 0.602 for the smallest to the largest lead time respectively. Nevertheless, it is worthy to note that post-processors have been calibrated on the whole sample and, inevitably, this assumption may lead to some conditional bias as a consequence of the focus on the optimality over the global dataset (Verkade et al., 2013). Similarly to what obtained for the ensemble mean, conditional accuracy metrics show a lower deterioration with the lead time in the validation period. Always considering the univariate MCP, NS goes from 0.996 for 6h lead time to 0.776 for 96h.

		Calibration period				**Validation period**			
		6 hours		**96 hours**		**6 hours**		**96 hours**	
		P = 0	**P = 90**	**P = 0**	**P = 90**	**P = 0**	**P = 90**	**P = 0**	**P = 90**
Univariate MCP	**NS (-)**	0.998	0.991	0.921	0.602	0.998	0.996	0.916	0.776
	RMSE (m^3/s)	47.73	95.57	325.53	667.63	44.04	131.25	356.73	1016.01
	MAE (m^3/s)	29.54	75.20	219.45	532.35	27.76	68.82	244.45	608.20
MCP	**NS (-)**	0.998	0.991	0.928	0.602	0.998	0.996	0.920	0.742
	RMSE (m^3/s)	47.01	92.02	311.11	667.61	43.10	125.00	349.28	1052.84
	MAE (m^3/s)	29.20	73.13	201.28	520.70	27.09	64.48	233.33	604.40
Uniform weighting	**NS (-)**	0.998	0.988	0.920	0.607	0.998	0.997	0.916	0.761
	RMSE (m^3/s)	53.44	107.31	327.25	662.90	45.25	114.16	357.65	1013.93
	MAE (m^3/s)	35.24	86.63	233.87	535.00	31.06	59.21	244.76	600.47
BMA	**NS (-)**	0.998	0.988	0.928	0.607	0.998	0.997	0.920	0.738
	RMSE (m^3/s)	54.65	110.83	309.98	663.12	45.37	114.73	348.99	1060.00
	MAE (m^3/s)	36.26	90.45	203.31	525.10	31.16	59.39	233.05	595.94

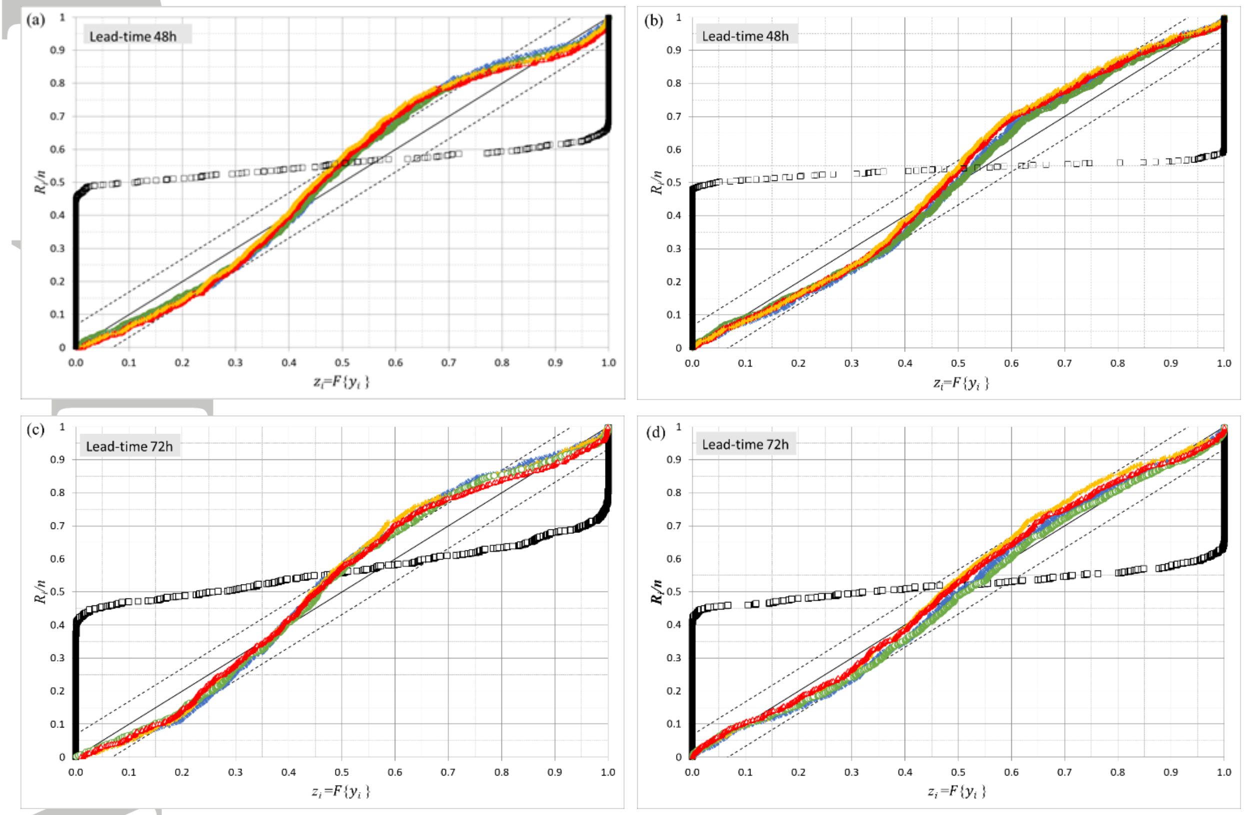
Table 2. Accuracy metrics, Nash–Sutcliffe coefficient (NS), Root Mean Square Error (RMSE), Mean of the Absolute Error (MAE), for the expected value of the post-processors at different levels of the non-exceedance climatological probability P in the sample of observations.

Inclusion of the ensemble spread in the parameter estimation phase has not generated significant variations in terms of accuracy metrics obtained with respect to the conditional mean of the predictive distributions.

The reliability of the four approaches to the estimation of the predictive distributions is assessed via the graphical representation based on a probability plot that does not require a subjective binning of the data as proposed in Laio and Tamea (2007).

Forecasting horizons from 48 to 96 hours are considered in Figure 8 where a comparison of the post-processors is shown. In this figure, also the points corresponding to unprocessed ensemble are displayed as a reference. Moreover, the Kolmogorov 5% significance bands are represented on the same graph in order to provide a more formal test of uniformity.

While in the case of the rank histograms of Figure 6 the ensemble members were ranked and considered as quantiles in discretized form, in order to use them in the PIT and compare the results the direct use of ensembles with the post-processing approaches, at each step in time a Normal distribution was assumed with mean and variance estimated from the ensemble. This does not substantially change the shape of the resulting curve: the outcomes obtained with the direct use of ensembles, as already emerged from the verification rank histograms, clearly show that they provide very narrow predictions and are unreliable, i.e. forecast probabilities do not match to frequency of observations.



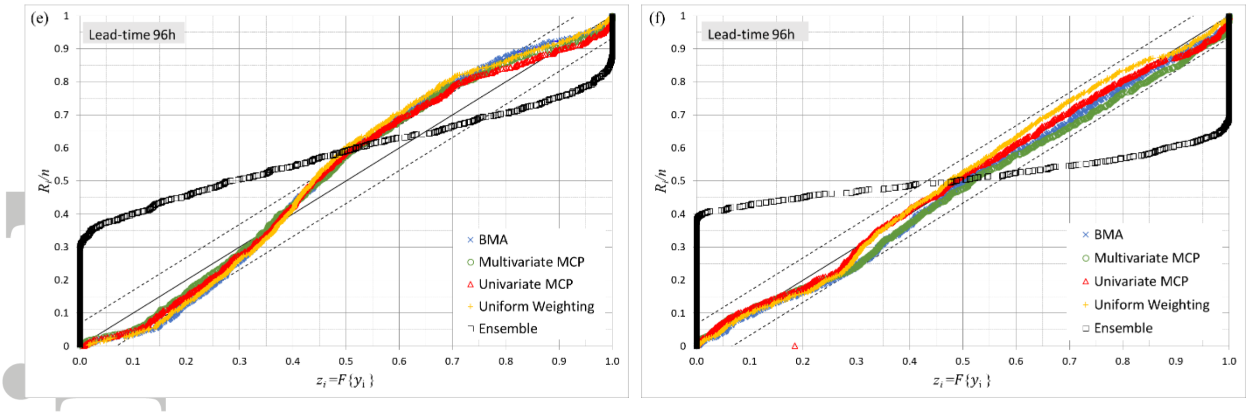


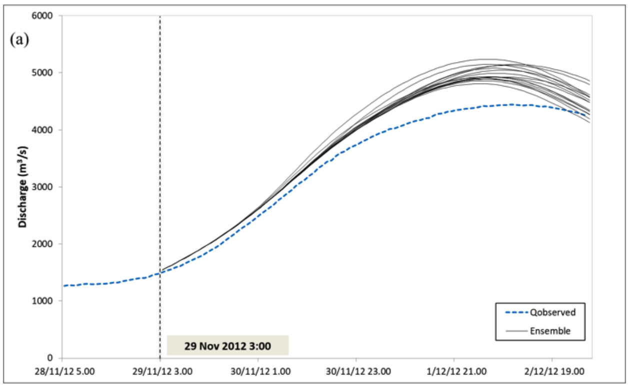
Figure 8. Probability plot of PIT of post-processed forecast for a 48, 72 and 96 h lead-times, where $z_i = F\{y_i\}$ is the predictive CDF evaluated at observation y_i, and R_i/n is the corresponding empirical cumulative distribution function: (**a**), (**c**), (**e**) calibration period and (**b**), (**d**), (**f**) validation period. Black squares represent ensemble forecasts; solid black line is the bisector and dashed black lines correspond to the Kolmogorov 5% significance bands.

What is also interesting to be noticed is that although the post processors show different predictive densities patterns, they are all more or less acceptable from the point of view of the probability plot and the relevant Kolmogorov-Smirnov test, as opposed to the approach which directly uses the ensemble spread as a measure of the predictive uncertainty.

Generally, post-processing leads to well-calibrated forecasts as shown in Figure 8: the points are very close to the Kolmogorov bands in the calibration period, while they are almost comprised within the bands for the validation period. Differences in the calibration between the different post-processed forecasts can hardly be detected.

5.3 Example forecasts on selected events

In order to illustrate the benefits descending from post-processing the hydrographs of an example prediction are presented and discussed using the univariate MCP, which makes no use of the exchangeability assumption and the multivariate MCP, which does. To this end, Figure 9 and Figure 10 show the forecasts for two high flow events covering a lead time of 96 hours issued on November 2012 (calibration period) and on November 2014 (validation period) respectively.



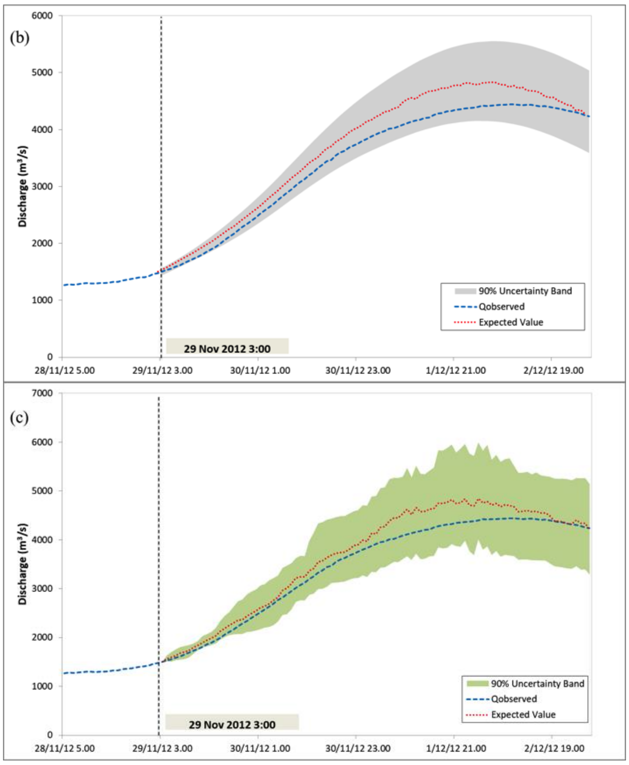
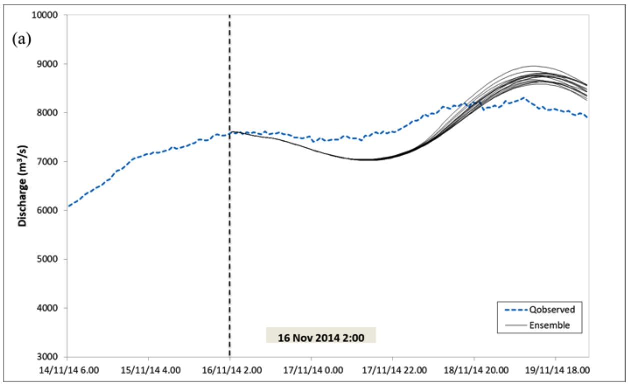


Figure 9. Po river at Pontelagoscuro: comparison between observed and forecast hydrographs for a high flow event issued on 29 November 2012, covering lead times 1–96 h. (**a**) Trajectories of the raw ensemble (**b**) 90% uncertainty band of the univariate MCP predictive distribution; (**c**) 90% uncertainty band of the multivariate MCP predictive distribution.

Regarding the flood event occurred in the calibration period, though the unprocessed ensemble is able to predict the magnitude of the event during the rising limb of the hydrograph, all members overestimate the observed peak flow as shown in Figure 9a. The post-processed probability forecasts shown in Figures 9b and 9c clearly improve the prediction compared to the raw ensemble. Specifically, the observed values (dashed blue line) are included in the 90% band of the probabilistic forecast for both the univariate and the multivariate version of the MCP.



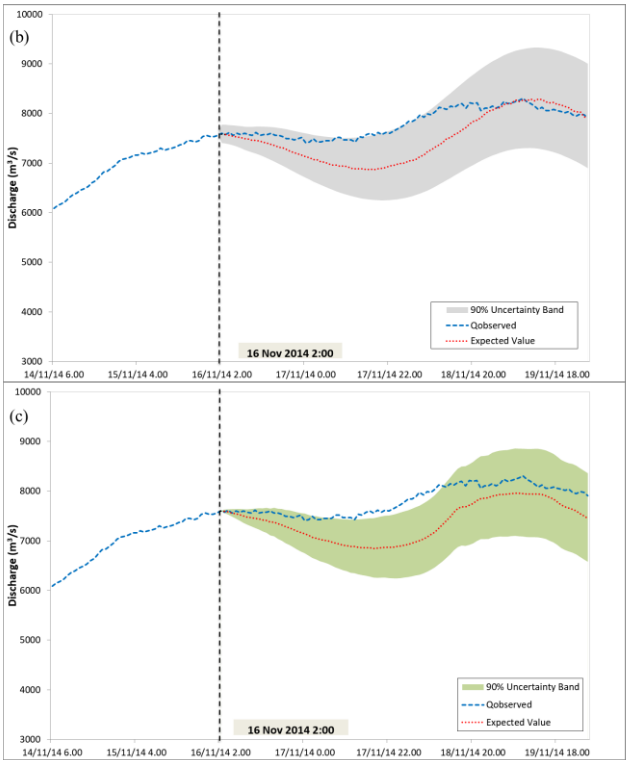


Figure 10. Po river at Pontelagoscuro: comparison between observed and forecast hydrographs for a high flow event issued on 16 November 2014, covering lead times 1–96 h. (**a**) Trajectories of the raw ensemble (**b**) 90% uncertainty band of the univariate MCP predictive distribution; (**c**) 90% uncertainty band of the multivariate MCP predictive distribution.

Similar considerations can be drawn for the flood event occurred on November 2014, the largest one, in terms of peak discharge, within the considered dataset. The flood peak, also in this case, is overestimated by the entire ensemble, while it is very close to the expected value of the predictive distributions for the two post-processing approaches. The univariate MCP, although simpler than the multivariate version of the MCP, almost comprises all the observations within the limits of the 90% uncertainty band.

6 Conclusions

Several conclusive remarks and prospective research themes can be drawn from the obtained results.

The first remark is that the raw meteorological (as well as the hydrological and hydraulic) ensembles should not be directly used as an alias of the predictive probability distribution. This is also true if they are “calibrated” in order to match the variance of prediction errors, because the real problem in probabilistic forecast, which is not limited at representing confidence limits but aimed at rational decision making, is not just the preservation of the variance of the predictive distribution, but rather the correct representation of the full distribution of prediction errors, as pointed out by several authors (Diebold et al., 1998; Laio & Tamea, 2007). From this point of view, the accuracy of the deterministic forecast represented by the ensemble mean does not necessarily imply a reliable prediction as the uncertainty range derived from the ensemble spread may not contain statistically consistent number of observations. Properly generated post-processing ensembles can in turn achieve this goal.

The second remark, which descends from the first one, is that whenever ensemble prediction is preferred or required as an alternative to a full predictive probability distribution, the correct ensemble members to be use in any successive decision making process are not the original model based ones, but rather the ones stochastically drawn from the resulting predictive distribution (Herr & Krzystofowicz, 2015), after verification of their reliability using the appropriate techniques.

The third issue is that several uncertainty post processors, incorporating the information provided by the spread of the ensemble members as models uncertainty, can be developed and used.

Provided that the mean of the ensemble is actually representative of the phenomenon showing sufficiently high correlation with the observed values, one of the analyzed approaches, univariate MCP, can be efficiently used, bearing in mind that it does not require the hypothesis of exchangeability and it is characterized by a reduced mathematics complexity. Bayesian mixture approaches, such as UW or BMA, have also shown satisfying results. UW assumes that all the ensemble members are probabilistically indistinguishable, which is close to the meteorological ensembles assumptions, while BMA produces a posterior estimate of the weights by maximizing a Likelihood function. Unfortunately, both UW and BMA do not take into account dependence among the ensemble members, which on the contrary is taken into account by the multivariate MCP. As shown in the Fig. 8, all these approaches, including the multivariate MCP, which produces the smallest predictive variance and is less sensitive to the ensemble spread, provide results that cannot be rejected by the used Kolmogorov-Smirnov test. Additional experiments, possibly also including hydrological and hydraulic uncertainties in the ensemble forecasts, are thus welcomed to further investigating the modified post processing properties.

In order to demonstrate to decision-makers the advantages of the proposed approaches, an interesting and important future theme of research, should be the comparison of the effects of using the derived predictive probability distribution, either in the form of statistical distribution, as opposed to directly using the traditional models derived ensemble members, on decision-making in areas such as such as flood warning, flood risk alleviation or reservoir management,.

The extension of the proposed approaches to the multi-model case (Coccia, 2011; Coccia & Todini, 2011) is straightforward, while additional research is needed to their extension to the multi-temporal problem (Coccia, 2011; Barbetta et al., 2017), which was recently shown to be of great interest to decision makers, given that it also accounts for time correlation of prediction errors.

Acknowledgments

The Authors would like expressing their acknowledgement and gratitude to those who made this research possible: Dr. Gabriele Coccia for providing the MCP, code developed as part of his Ph.D. Thesis, together with support and suggestions; Dr. Silvano Pecora from Agenzia Regionale per la Prevenzione, l’Ambiente e l’Energia dell´Emilia-Romagna, for providing the observations and the real time COSMO-LEPS based ensemble forecasts for the Po River at Pontelagoscuro.

Data used in this paper are originally obtained from Agenzia Regionale per la Prevenzione, l’Ambiente e l’Energia dell´Emilia-Romagna (https://www.arpae.it), courtesy of Dr. Silvano Pecora, and are included in the supporting information.

References

Barbetta, S., Coccia, G., Moramarco, T., Brocca, L., & Todini, E. (2017) The multi temporal/multi-model approach to predictive uncertainty assessment in real-time flood forecasting, *Journal of Hydrology* 551, 555–576.

Bartholmes, J. & Todini, E. (2005). Coupling meteorological and hydrological models for flood forecasting. *Hydrol. Heart Syst. Sci*., 9, 55-68.

Bartholmes, J., Thielen, J., Ramos, M.-H., & Gentilini, S. (2009). The European Flood Alert System EFAS– Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrology and Earth System Sciences*, 13, 141-153.

Benjamin, J.R., & Cornell, C.A. (1970). *Probability, Statistics and Decision for Civil Engineers*. McGraw Hill, NewYork.

Bennett, J.C., Robertson, D.E., Shrestha, D.L., Wang, Q.J., Enever, D., Hapuarachchi, P. & Tuteja, N.K. (2014). A System for Continuous Hydrological Ensemble Forecasting(SCHEF) to lead times of 9 days. *Journal of Hydrology*, 519, doi: 10.1016/j.jhydrol.2014.08.010.

Benninga, H.-J. F., Booij, M. J., Romanowicz, R. J. & Rientjes, T. H. M. (2017). Performance of ensemble streamflow forecasts under varied hydrometeorological conditions, *Hydrol. Earth Syst. Sci.*, 21, 5273–5291.

Bougeault, P., et al. (2010). The THORPEX interactive grand global ensemble, *Bull. Am. Meteorol. Soc.*, 91(8), 1059–1072. doi:10.1175/ 2010BAMS2853.1.

Ciarapica L., & Todini, E., (2002). TOPKAPI: a model for the representation of the rainfall-runoff process at different scales. *Hydrological Processes* 16(2), 207-229.

Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, 375, 613-629.

Coccia, G. (2011). *Analysis and developments of uncertainty processors for real time flood forecasting* (Doctoral Dissertation), Retrieved from AlmaDL – University of Bologna Digital Library. (DOI:10.6092/unibo/amsdottorato/3423). Alma Mater Studiorum University of Bologna.

Coccia, G., & Todini, E. (2011). Recent developments in predictive uncertainty assessment based on the model conditional processor approach. *Hydrology and Earth System Sciences*, 15(10), 3253-3274.

Cuo, L., Pagano, T. C., & Wang, Q. J. (2011). A review of quantitative precipitation forecasts and their use in short-to medium-range streamflow forecasting. *J. Hydrometeorol*., 12, 713–728.

Deltares, (2014). SOBEK – *Hydrodynamics, Rainfall Runoff and Real Time Control*, Version 1.00.34157. Deltares, Delft, The Netherlands.

Demargne, J. (2014). The Science of NOAA's Operational Hydrologic Ensemble Forecast Service. *Bulletin of the American Meteorological Society*. doi:10.1175/bams-d-12-00081.1

Demerit, D., Nobert, S., Cloke, H., & Pappenberger, F. (2010). Challenges in communicating and using ensembles in operational flood forecasting. *Meteorol. Appl.*, 17, 209–222.

Deming, W.E. (1943). *Statistical adjustment of data*. Wiley, NY (Dover Publications edition, 1985). ISBN 0-486-64685-8.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *J. Roy. Stat. Soc. Ser. B*, 39, 1–38.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management, *Int. Econ. Rev*., 39(4), 863–883.

Draper, D., & Krnjajic, M. (2013). Calibration Results for Bayesian Model Specification, Technical Report, Department of Applied Mathematics and Statistics, University of California, Santa Cruz, https://users.soe.ucsc.edu/~draper/draper-krnjajic-2013-draft.pdf. Last accessed July 09, 2018.

Fraley, C., Raftery, A. E., & Gneiting, T. (2010). Calibrating Multimodel Forecast Ensembles with Exchangeable and Missing Members Using Bayesian Model Averaging. *Monthly Weather Review*, 138, 190-201. doi: 10.1175/2009MWR3046.1

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.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc. Ser. B*, 69, 243–268.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., & Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. (Technical Report 537), University of Washington: Department of Statistics.

Golding, B. (2000). Quantitative precipitation forecasting in the UK. *Journal of Hydrology*, 239 (1–4), 286–305.

Hamill, T.M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129, 550–560.

Hamill, T.M. (2018). *Chapter 7: Practical Aspects of Statistical Postprocessing* in S. Vannitsem, D. Wilks and J. Messner (Eds.), Statistical Postprocessing of Ensemble Forecasts. Elsevier. ISBN: 9780128123720

Hemri, S., Lisniak, D., & Klein, B. (2015). Multivariate postprocessing techniques for probabilistic hydrological forecasting. *Water Resources Research*, 51, 7436–7451. doi:10.1002/ 2014WR016473.

Herr, H. D., & Krzystofowicz, R. (2015). Ensemble Bayesian forecasting system Part I: Theory and algorithms, *Journal of Hydrology*, 524, 789-802.

Koenker, R. (2005). Quantile Regression, Econometric Society Monographs, Cambridge University Press, New York, NY.

Koutsoyiannis, D., Makropoulos, C., Langousis, A., Baki, S., Efstratiadis, A., Christofides, A., Karavokiros, G. & Mamassis, N. (2009). HESS opinions: Climate, hydrology, energy, water: Recognizing uncertainty and seeking sustainability. *Hydrol. Earth Syst. Sci*., 13, 247–257.

Krzysztofowicz, R. (1999). Bayesian Theory of Probabilistic Forecasting via Deterministic Hydrologic Model. *Water Resources Research*, 35, 2739–2750.

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

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., & Di, Z., (2017). A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdisciplinary Reviews: Water*, 4(6), doi: 10.1002/wat2.1246.

Li, W., & Duan, Q., (2018) Recent development of post-processing methods in short-term hydrometeorological ensemble forecasting. (https://hepex.irstea.fr/recent-development-of-post-processing-methods-in-short-term-hydrometeorological-ensemble-forecasting/).

Liu, Z., & Todini E., (2002). Towards a comprehensive physically based rainfall-runoff model. *Hydrology and Earth System Sciences (HESS)*, 6(5), 859–881.

Madadgar, S., & Moradkhani, H., (2014). Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. *Water Resour. Res*., 50, 9586-9603, doi:10.1002/2014WR015965.

Mahler, H. C., & Dean, C.G. (2001). *Chapter 8: Credibility In Foundations of Casualty Actuarial Science* (4th ed.). Casualty Actuarial Society. pp. 525–526. ISBN 978–0–96247-622-8

Mardia, K.V., Kent, J.T., & Bibby, J. M. (1979). *Multivariate Analysis. Probability and Mathematical Statistics*. Academic Press, London.

Montanari, A., & Brath, A. (2004). A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resour. Res*., 40, W01106, doi: 10.1029/2003WR002540.

Montanari, A., & Grossi, G. (2008). Estimating the uncertainty of hydrological forecasts: A statistical approach. *Water Resour. Res*., 44, W00B08, doi:10.1029/2008WR006897.

Montanari, A., Shoemaker, C. A. & van de Giesen, N. (2009). Introduction to special section on uncertainty assessment in surface and subsurface hydrology: An overview of issues and challenges. *Water Resour. Res*.,45, W00B00, doi:10.1029/2009WR008471.

Murphy, A. H. (1993). What is a good forecast? An essay on nature of goodness in weather forecasting. *Weather and Forecasting*, 5(8), 281-293.

Nester, T., Komma, J., Viglione, A., & Blöschl, G. (2012). Flood forecast errors and ensemble spread – A case study. *Water Resources Research*, 48, W10502, doi:10.1029/2011WR011649

Pappenberger, F., Beven, K.J., Hunter, N., Bates, P., Gouweleeuw, B.T., Thielen, J. & De Roo, A.P.J. (2005). Cascading model uncertainty from medium range weather forecast (10 days) thorugh a rainfall-runoff model to flood inundation predictions within European Flood Forecasting System (EFFS). *Hydrol. Heart Syst. Sci*., 9, 103-115.

Park, Y.-Y., Buizza, R., & Leutbecher, M. (2008). TIGGE: Preliminary results on comparing and combining ensembles. *Quart. J. Roy. Meteor. Soc*., 134, 2029–2050.

Raftery, A. E. (1993). Bayesian model selection in structural equation models, in K.A. Bollen and J.S. Long (Eds.), *Testing Structural Equation Models*, pp. 163–180. Newbury Park, Calif. Sage.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5), 1155-1174.

Ramos, M-H., Voisin, N., & Verkade, J. (2013). HEPEX-SIP TOPIC: Post-Processing Part 2: Literature review on post-processing. (https://hepex.irstea.fr/hepex-sip-topic-post-processing-23/)

Reggiani, P., & Weerts, A. (2008). A Bayesian approach to decision-making under uncertainty: an application to real-time forecasting in the river Rhine. *Journal of Hydrology*, 356, 56–59, doi:10.1016/j.jhydrol.2008.03.027.

Reggiani, P., Renner, M., Weerts, A., & Van Gelder, P. (2009). Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system. *Water Resources Research*, 45, W02428. doi:10.1029/2007WR006758.

Regonda, S.K., Seo, D.-J., Lawrence, B., Brown, J.D. & Demargne, J. (2013). Short-term ensemble streamflow forecasting using operationallyproduced single-valued streamflow forecasts – A Hydrologic Model Output Statistics (HMOS) approach. *Journal of Hydrology*, 497, 80-96.

Rogers, E. (2003). *Diffusion of innovations,* Free Press, New York.

Roulston, M.S., & Smith, L.A. (2003). Combining dynamical and statistical ensembles. *Tellus*, 55A, 16–30.

Schaake, J.C., Hamill, T.H., Buizza, R., & Clark, M. (2007). HEPEX – The hydrological ensemble prediction experiment. *Bulletin of the American Meteorological Society,* 88(10), 1541–1547.

Seo, D.-J., Herr, H.D., & Schaake, J. C. (2006). A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrology and Earth System Sciences Discussions*, 3 (4), 1987-2035.

Stelling, G.S., & Verwey, A. (2005). Numerical flood simulation. In: *Encyclopedia of Hydrological Sciences*. John Wiley & Sons Ltd. DOI: 10.1002/0470848944.hsa025a.

Talagrand, O., Vautard, R., & Strauss, B. (1997). *Evaluation of probabilistic prediction systems.* Paper presented at Workshop on Predictability, European Centre for Medium-Range Weather Forecasts, Reading, UK.

Thielen, J., Bartholmes, J., Ramos, M.-H., & de Roo A. (2009). The European Flood Alert System – Part 1: Concept and development. *Hydrology and Earth System Sciences*, 13, 125-140.

Todini, E. (2008). A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management*, 6(2), 123-137.

Verkade, J.S., Brown, J.D., Reggiani, P., Weerts, A.H. (2013). Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology*, 501, 73-91.

Verkade, J., Ramos, M-H., & Voisin, N. (2013). HEPEX-SIP TOPIC: Post-Processing Part 1: What is hydrologic post processing. (https://hepex.irstea.fr/hepex-sip-topic-post-processing-13/)

Voisin, N., Verkade, J., & Ramos, M-H. (2013). HEPEX-SIP TOPIC: Post-Processing Part 3: Challenges and research needs. (https://hepex.irstea.fr/hepex-sip-topic-post-processing-33/)

Vrugt, J.A. (2016). MODELAVG: A MATLAB Toolbox for Postprocessing of Model Ensembles. http://faculty.sites.uci.edu/jasper/files/2016/04/manual_Model_averaging.pdf

Weerts, A.H., Winsemius, H.C., & Verkade, J.S. (2011). Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales). *Hydrology and Earth System Sciences*, 15(1), 255-265.

Weiss, N.A. (2006). *A course in probability*. Addison–Wesley, Boston, pp. 385–386.

Whateley, S., Palmer, R.N., & Brown, C. (2015). Seasonal Hydroclimatic Forecasts as innovations and the Challenge of Adoption by Water Managers, *J. of Water Resour. Plann. Manage.* 141(5): 04014071-1/13. DOI: 10.1061/(ASCE)WR.1943-5452.0000466.

Wilks, S.S. (1948). Order statistics. *Bulletin of the American Mathematical Society*, 54(1), Part 1, pp. 6-50.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference*. Journal of the American Statistical Association*, 22, 209-212. doi*:*10.1080/01621459.1927.10502953.

APPENDIX A – The variance of an ordered sample from a Normal Distribution

It is well known (Wilks, 1948) that the probability density of the probability $P\left(x_{k|N}\right)$ of the k_{th} member in a ranked sample of N members is the $Beta\{k, N-k+1\}$, namely:

$$f\left\{P\left(x_{k|N}\right)\right\} = \frac{N!}{(k-1)!\,(N-k)!} P\left(x_{k|N}\right)^{k-1} \left(1-P\left(x_{k|N}\right)\right)^{N-k}$$

with moments:

$$Mean\left\{P\left(x_{k|N}\right)\right\} = \frac{k}{N+1}$$

$$Var\left\{P\left(x_{k|N}\right)\right\} = \frac{k(N-k+1)}{(N+1)^2(N+2)}$$

$$St.Dev.\left\{P\left(x_{k|N}\right)\right\} = \frac{1}{N+1}\sqrt{\frac{k(N-k+1)}{N+2}}$$

If variable x is Normally distributed with mean μ_x and variance σ_x^2, then its density is

$$f\{x\} = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2}}{\sigma_x\sqrt{2\pi}}$$

The probability of $x_{k|N}$, namely $P\left(x_{k|N}\right)$ can be expressed as:

$$P\left(x_{k|N}\right) = \int_{-\infty}^{x_{k|N}} \frac{e^{-\frac{1}{2}\left(\frac{\xi-\mu_x}{\sigma_x}\right)^2}}{\sigma_x\sqrt{2\pi}} d\xi$$

The probability density of $P\left(x_{k|N}\right)$ can then be derived by probability matching (Benjamin & Cornell, 1970)

$$f\left\{P\left(x_{k|N}\right)\right\} dP\left(x_{k|N}\right) = f\left\{x_{k|N}\right\} dx_{k|N}$$

leading to:

$$f\left\{x_{k|N}\right\} = \; f\left\{P\left(x_{k|N}\right)\right\} \left|\frac{dP\left(x_{k|N}\right)}{dx_{k|N}}\right|$$

which, after substitution becomes:

$$f\left\{x_{k|N}\right\} = \frac{N!}{(k-1)!\,(N-k)!} \left(\int_{-\infty}^{x_{k|N}} \frac{e^{-\frac{1}{2}\left(\frac{\xi-\mu_x}{\sigma_x}\right)^2}}{\sigma_x\sqrt{2\pi}} d\xi\right)^{k-1} \left[1-\left(\int_{-\infty}^{x_{k|N}} \frac{e^{-\frac{1}{2}\left(\frac{\xi-\mu_x}{\sigma_x}\right)^2}}{\sigma_x\sqrt{2\pi}} d\xi\right)\right]^{N-k} \frac{e^{-\frac{1}{2}\left(\frac{x_{k|N}-\mu_x}{\sigma_x}\right)^2}}{\sigma_x\sqrt{2\pi}}$$

Accordingly

$$E\left\{x_{k|N}\right\} = \frac{N!}{(k-1)!\,(N-k)!} \int_{-\infty}^{+\infty} x_{k|N} \left(\int_{-\infty}^{x_{k|N}} \frac{e^{-\frac{1}{2}\left(\frac{\xi-\mu_x}{\sigma_x}\right)^2}}{\sigma_x\sqrt{2\pi}} d\xi\right)^{k-1} \left[1-\left(\int_{-\infty}^{x_{k|N}} \frac{e^{-\frac{1}{2}\left(\frac{\xi-\mu_x}{\sigma_x}\right)^2}}{\sigma_x\sqrt{2\pi}} d\xi\right)\right]^{N-k} \frac{e^{-\frac{1}{2}\left(\frac{x_{k|N}-\mu_x}{\sigma_x}\right)^2}}{\sigma_x\sqrt{2\pi}} dx_{k|N}$$

$$= \frac{N!}{(k-1)!\,(N-k)!} \int_{0}^{1} x_{k|N}\left[P\left(x_{k|N}\right)\right] P\left(x_{k|N}\right)^{k-1} \left(1-P\left(x_{k|N}\right)\right)^{N-k} dP\left(x_{k|N}\right)$$

$$E\{x_{k|N}^2\} = \frac{N!}{(k-1)!\,(N-k)!}\int_{-\infty}^{+\infty} x_{k|N}^2 \left(\int_{-\infty}^{x_{k|N}} \frac{e^{-\frac{1}{2}\left(\frac{\xi-\mu_x}{\sigma_x}\right)^2}}{\sigma_x\sqrt{2\pi}}d\xi\right)^{k-1}\left[1-\left(\int_{-\infty}^{x_{k|N}} \frac{e^{-\frac{1}{2}\left(\frac{\xi-\mu_x}{\sigma_x}\right)^2}}{\sigma_x\sqrt{2\pi}}d\xi\right)\right]^{N-k}\frac{e^{-\frac{1}{2}\left(\frac{x_{k|N}-\mu_x}{\sigma_x}\right)^2}}{\sigma_x\sqrt{2\pi}}dx_{k|N}$$

$$= \frac{N!}{(k-1)!\,(N-k)!}\int_0^1 x_{k|N}^2[P(x_{k|N})]\,P(x_{k|N})^{k-1}\left(1-P(x_{k|N})\right)^{N-k}dP(x_{k|N})$$

From which one can compute the variance as:

$$Var\{x_{k|N}^2\} = E\{x_{k|N}^2\} - E\{x_{k|N}\}^2$$

The computation can be greatly simplified by computing only once the distribution of variances for a $N(0,1)$ Standard Normal variable $x_{k|N}$ and successively by multiplying at each step in time the resulting values by the ensemble variance
In the present case under study with 16 models, the distribution of variances for the 16 ordered ensemble members will be σ_x^2 multiplied by the coefficients α_i in the following table.

Table A1 – Distribution of variances for the 16 ordered ensemble members. Coefficients α_i multiplying σ_x^2 for the 16 ranked members of the ensemble

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
α_i	0.295	0.174	0.136	0.118	0.107	0.101	0.097	0.096	0.096	0.097	0.101	0.107	0.118	0.136	0.174	0.295

Please note that:

1) the sum of weights adds up to 2.248 and not to 1 as one would expect in the case of the sum of independent variables. The point is that ordering introduces correlation, which is then reflected in the fact that the sum of dependent variables does not necessarily add up to 1 .
2) variables which are more distant from the mean (high and low rank variables) will be affected by higher variances, which is due to the fact that these variables are in the tail of the distribution with wider fields of existence.