Bluecat: A Local Uncertainty Estimator for Deterministic Simulations and Predictions

D. Koutsoyiannis[1], and A. Montanari[2]

[1]National Technical University of Athens, Zographou, Greece
[2]Department DICAM, University of Bologna, Bologna, Italy

Key Points:

- We propose a new method to frame a deterministic prediction model into a stochastic setting with probability based uncertainty assessment.
- We theoretically and empirically prove the optimal performance of the method for operational applications.
- We provide an open source computer code to apply the method and perform diagnostic checking.

Corresponding author: Alberto Montanari, `alberto.montanari@unibo.it`

Abstract

We present a new method for simulating and predicting hydrologic variables with uncertainty assessment and provide example applications to river flows. The method is identified with the acronym "Bluecat" and is based on the use of a deterministic model which is subsequently converted to a stochastic formulation. The latter provides an adjustment on statistical basis of the deterministic prediction along with its confidence limits. The distinguishing features of the proposed approach are the ability to infer the probability distribution of the prediction without requiring strong hypotheses on the statistical characterization of the prediction error (e.g. normality, homoscedasticity) and its transparent and intuitive use of the observations. Bluecat makes use of a rigorous theory to estimate the probability distribution of the predictand conditioned by the deterministic model output, by inferring the conditional statistics of observations. Therefore Bluecat bridges the gaps between deterministic (possibly physically-based, or deep learning-based) and stochastic models as well as between rigorous theory and transparent use of data with an innovative and user oriented approach. We present two examples of application to the case studies of the Arno river at Subbiano and Sieve river at Fornacina. The results confirm the distinguishing features of the method along with its technical soundness. We provide an open software working in the R environment, along with help facilities and detailed instructions to reproduce the case studies presented here.

Plain Language Summary

We present a new method for simulating and predicting hydrologic variables and in particular river flows, which is rooted in the probability theory and conceived in order to provide a reliable quantification of its uncertainty for operational applications. In fact, recent practical experience during extreme events has shown that simulation and prediction uncertainty is essential information for decision makers and the public. A reliable and transparent uncertainty assessment has also been shown to be essential to gain public and institutional trust in real science. Our approach, which we term with the acronym "Bluecat", results from a theoretical and numerical development, and is conceived to make a transparent and intuitive use of the observations which in turn mirror the observed reality. Therefore, Bluecat makes use of a rigorous theory while at the same time proofing the concept that environmental resources should be managed by making the best use of empirical evidence and experience. We provide an open and user friendly software to apply the method to the simulation and prediction of river flows and test Bluecat's reliability for operational applications.

1 Introduction

Recent extreme events like the flood that occurred in central Europe in 2021 have shown that reliable hydrological predictions are essential to issue early warnings to institutions and population. Indeed, effective warnings require people to be informed on the magnitude of a forthcoming event and the likelihood of that happening. Namely, a prediction along with its uncertainty needs to be timely developed and communicated. The time factor is in fact essential and therefore the whole warning system needs to be fast and reliable, in the estimation of both prediction and uncertainty (see, for instance, Ramos et al. (2013) and Pagano et al. (2014)). An additional key element for the success of a warning system is its credibility, which is usually evaluated by end users by confronting the prediction method with their expert judgment and empirical evaluation (Blöschl, 2008). This is precisely the reason why the prediction and its uncertainty should be elaborated with a transparent approach by making a perceptional use of the available information and data, which in turn mirror the observed reality of previous and likely future events.

In particular, the uncertainty inherent in scientific information is one of the reasons for failing to act on disaster warnings. Forecasts are often elaborated with methodologies that are not easily understood by those who need it. Such lack of understanding of uncertainty estimation may lead people to interpret the predictions as unreliable, and to believe that estimations should no longer be trusted.

Prediction and forecasting have been the focus of an intensive research activity in hydrology (see, for instance, Blöschl et al. (2013)). Here, we concentrate on uncertainty assessment which has been the subject of relevant efforts since the early works of Spear and Hornberger (1980) and Beven and Binley (1992). The literature is branched in several subtopics ranging from data uncertainty, parameter fitting, model structural uncertainty, operational uncertainty and so forth (Montanari, 2011).

To date, the most used method for estimating the uncertainty of hydrological simulations and predictions is the Generalized Likelihood Uncertainty Estimator (GLUE) (see Beven and Binley (1992) and Beven (2006)). GLUE rejects the concept of one single optimal model and adopts the notion of equifinality of modeling solutions (Beven & Lane, 2019). It makes use of an informal likelihood function that has been the subject of an interesting debate (see, for instance, Montanari (2005); Vrugt et al. (2009); Beven (2009)). Bayesian methods are widely used and include, among the others, Bayesian model averaging (see the recent work by Reggiani et al. (2021)), Bayesian estimation of model errors (Tajiki et al., 2020) and Bayesian data assimilation (Bulygina & Gupta, 2009), and signature domain calibration (Kavetski et al., 2018). In a Bayesian framework, identifying a suitable likelihood function for hydrological models is a challenging task which requires the introduction of assumptions that need to be carefully checked as sometimes the related approximations are not easily understandable by end users.

Another relevant example of Bayesian method is the Bayesian Forecasting System introduced by Krzysztofowicz (1999), which produces a probabilistic river stage or flow forecast based on a probabilistic quantitative precipitation forecast as an input to a hydrological model. The BFS assumes that the dominant source of uncertainty derives from the imperfect knowledge of the future precipitation, so that it can be assumed that all other sources of uncertainty play a minor role. While it may be justified for operational forecasting, this assumption looks restrictive for hydrologic simulations where model structural uncertainty may also be substantial.

The literature presented several approaches to uncertainty assessment based on the statistical analysis of the probability distribution of model errors or, analogously, the joint probability distribution of observed and simulated data. These methods belong to the category of the post-processing approaches, which have been proved to outperform analyses that consider all the sources of uncertainty (see, for instance, the recent contribution by Valdez et al. (2021)). This class of methods can be further subdivided in likelihood based and likelihood-free approaches. The use of likelihood is considered by Tajiki et al. (2020) and previously by Schoups and Vrugt (2010), while likelihood-free methods include the works by Montanari and Brath (2004), Montanari and Grossi (2008) and Montanari and Koutsoyiannis (2012). The statistical analysis of model errors to estimate simulation and prediction uncertainty with a likelihood-free approach presents the advantage of being transparent to end users and computationally fast.

In particular, Montanari and Koutsoyiannis (2012) proposed a theoretically based method to convert a deterministic hydrologic model into a stochastic approach by fitting the model error with a meta-Gaussian probability distribution. A similar approach was applied by Quilty and Adamowski (2020) and several other works. Notably, Sikorska et al. (2015) proposed a nearest neighbour approach to represent the probability distribution of the model error which makes the method flexible and fast. Similar approaches were applied by Papacharalampous, Tyralis, and Koutsoyiannis (2019), Papacharalampous et al. (2020), Papacharalampous, Tyralis, Langousis, et al. (2019b), Tyralis, Papacharalampous, and Langousis (2019),Tyralis, Papacharalampous, Burnetas, and Langousis (2019) and Papacharalampous, Tyralis, Langousis, et al. (2019a). Notwithstanding the above research efforts, the statistical representation of the model error remains difficult in some applications and thus there is still the need for end users to further simplify the procedure.

In view of the above previous works and the requirement for effective predictions, we present here an innovative and transparent approach that builds on the concept proposed by Montanari and Koutsoyiannis (2012) to transform a generic deterministic model into a stochastic predictor. A distinguishing feature of the proposed method is its ability to infer the probability distribution of the prediction without running multiple simulations and without requiring strong hypotheses on the statistical characterization of the prediction itself or its error, therefore resolving critical issues that affect the previously proposed methods. Although intuitive, the method is supported by a rigorous theoretical development that ensures the best use of the information content of the observed data. The method can be applied to either physically-based, process-based and data-based deterministic prediction/simulation models. It can also be applied in conjunction with prediction models based on deep learning, which are gaining increasing popularity for hydrological predictions (see, for instance, Frame et al. (2021)).

We make available an open software in the public domain, working in the R environment (R Core Team, 2013), along with instructions and examples of applications, to support applications by end users. The software also provides goodness of fit procedures that are based on the best practices of engineering and applied forecasting.

We propose for our approach the acronym Bluecat, from "**B**risk **l**ocal **u**ncertainty **e**stimator for generi**c** simul**a**tions and predic**t**ions". In this paper we focus on river flow and therefore assume that the deterministic model is a rainfall-runoff model. However, the procedure can be generalized to any type of deterministic prediction model. In what follows, we use the term "prediction" to encompass simulation, prediction and forecasting.

2 Concept of Bluecat

Bluecat is a simple and transparent tool to transform point predictions obtained by any deterministic model in stochastic predictions, therefore deriving the probability distribution of the predictand. In what follows, we will use the terms "D-model" and "S-model" to denote the deterministic model and its stochastic counterpart, respectively.

The information that is needed to perform the above transformation is obtained in Bluecat by building on the well established concept of comparing the D-model output with observed data; namely, the same concept that we commonly use for parameter estimation. Basing on such comparison, Bluecat estimates the probability distribution of observed data conditioned on the D-model output and therefore obtains the corresponding S-model output, along with its mean (or median) value and confidence band. It is important to make clear that the S-model prediction may be markedly different from the D-model one. In fact, the latter is not necessarily included into the confidence band of the S-model, which are displaced around the mean prediction of the S-model itself. Such possible outcome is schematically represented in Figure 1, where the concept of Bluecat is depicted.

Being based on the comparison between the D-model output and the observations, Bluecat is therefore transparent and easily understandable, while the theoretical development that we present in Section 3 ensures that such interpretation of uncertainty is rigorous and asymptotically consistent in estimating global uncertainty.

Bluecat is based on the following main assumptions:

1. A single D-model is considered, with a single parameter set. Section 6 will present a discussion on the possible extension of the Bluecat concept to multimodel applications.
2. The stochastic processes describing the modelled variables are stationary during the calibration and application period. Non-stationarity can be accounted for by using non-stationary D-models (Koutsoyiannis & Montanari, 2015; Montanari & Koutsoyiannis, 2014a). Such extension is not considered in the present contribution but a discussion is provided in Section 6.
3. The calibration data set is extended enough to ensure that sufficient information is available to upgrade the D-model into the S-model.

Further assumptions will be introduced and discussed in Section 3.

The third assumption above highlights that the S-model, like the D-model, needs a proper calibration, which implies that a sufficiently long record of observed data, referring to a variety of hydrologic conditions, is available for model training. Such requirement may be difficult to satisfy in real world applications, which often refer to poorly gauged or ungauged conditions. We will discuss in Section 6 the implications of running Bluecat with a limited training.

The flow chart of the procedure for applying Bluecat is as follows (see Figure 1):

1. The D-model is calibrated by using observed data;
2. At the prediction time t^* the D-model is run to produce an estimated river flow $Q(t)$ at time t;
3. A set of size m_1+m_2+1 (see Section 3.1 for details) of predicted river flows from the calibration data set, including the one with the smallest difference from $Q(t)$ plus m_1 lower and m_2 greater in magnitude of it, is extracted and the corresponding simulated river flows $q_i, i = 1, ..., 2m + 1$ are identified;
4. From the obtained sample of q_i the mean (or median) prediction and the confidence band for assigned confidence level from the S-model are estimated by using one of the methods described in Section 3.

Thus, the S-model operates an adjustment of the D-model to compensate its inability to fully reproduce the observed reality. We develop and present in the following section a theory to prove the rigorousness of the concept and the ability of the S-model to asymptotically represent the desired probability distribution of the predictand.

3 Theory of Bluecat

We consider a hydrologic D-model transforming inputs $\mathbf{x}_\tau$ (e.g. rainfall) at discrete time τ to deterministic outputs Q_τ (e.g. river discharge) by means of a relationship that takes the form

$$Q_\tau = G\left(\mathbf{x}_\tau\right), \tag{1}$$

where $\mathbf{x}_\tau$ is a vector containing a number of consecutive input variables, or even a matrix consisting of several input variables (such as rainfall, evapotranspiration, perhaps river discharge in an upstream basin, and possibly others). The transformation function is generally complicated, also involving additional state variables (e.g. soil moisture). A model is never identical to reality and the observed output (the predictand) q_τ will be different from the model prediction Q_τ. In the present work we consider the HyMod rainfall-runoff model (Boyle, 2000) as D-model, which involves 5 parameters.

As mentioned above, Montanari and Koutsoyiannis (2012) proposed a framework to upgrade a deterministic model into a stochastic one, which provides the probability distribution of the predictand given the inputs and the deterministic model output, con-

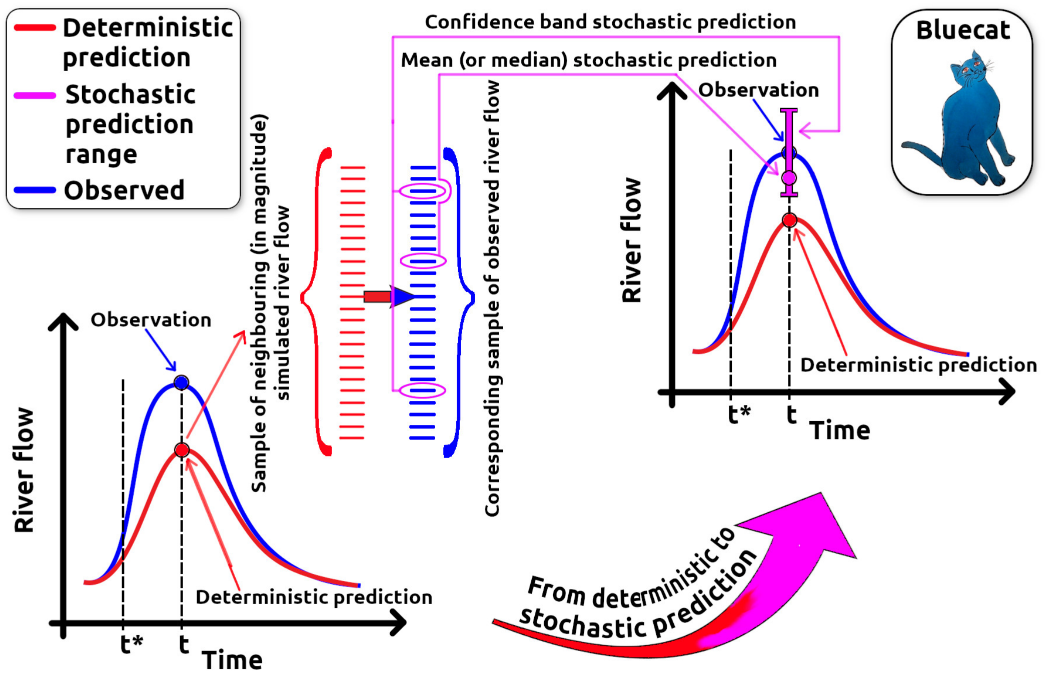

Figure 1. Schematic representation of the Bluecat concept underlying the transformation of the deterministic model (D-model) to a stochastic model (S-model).
The painting in the upper right corner is cropped from the picture available at https://www.flickr.com/photos/cizauskas/36142084534/ of the Andy Warhol exhibition at the High Museum, Atlanta, Georgia, USA (CC BY-NC-ND 4.0).

sidering the uncertainty in model parameters and input variables. This work has been discussed (Nearing, 2014; Montanari & Koutsoyiannis, 2014b) and advanced in other studies (Sikorska et al., 2015; Quilty & Adamowski, 2020; Papacharalampous, Tyralis, & Koutsoyiannis, 2019). Here we pursue the same aim but in a different setting, with the purpose of upgrading the D-model into the S-model by using the simplest approach based on data analysis.

As anticipated in Section 2 we assume that the information contained in the true outputs q_τ and concurrent predictions by the D-model Q_τ is sufficient to support the above upgrade. This implies that the upgrade is properly trained over a sufficiently long calibration period. Transparency and ease of understanding of the procedure is a principal objective and therefore we do not involve multiple simulations, but rather focus on a single model for which we aim to estimate the global prediction uncertainty. As a consequence, we do not consider parameter uncertainty in the D-model on the basis that another parameter set is in fact another model. This assumption is further discussed in Section 6.

Second, we do not subdivide uncertainty in different components as Bluecat automatically incorporate all types, including the uncertainty in input data and parameters, for which no particular provision is necessary. As already mentioned, the framework also assumes stationarity. If different subperiods are characterized by different model parameters or different input uncertainty, then one can split the entire simulated period in subperiods in which stationarity can be safely assumed. In alternative, the assumption of stationarity may be relaxed by considering a non-stationary D-model, as discussed in Section 6.

For advancing the D-model into its corresponding S-model we regard all related quantities as stochastic (random) variables and their sequences as stochastic processes. For notational clarity we underline stochastic variables, stochastic processes and stochastic functions. We use non-underlined symbols for non stochastic variables and deterministic functions, as well as for realizations of stochastic variables and stochastic processes, where the latter realizations are also known as time series.

We assume that the inputs $\underline{\mathbf{x}}_\tau$, at discrete times τ, have a stationary probability density function $f_{\mathbf{X}}(\mathbf{x})$ and distribution function $F_{\mathbf{X}}(\mathbf{x})$. The output $\underline{q}_\tau$ depends on the inputs $\underline{\mathbf{x}}_\tau$ and is given through some stochastic function (S-model) as

$$\underline{q}_\tau = \underline{g}\left(\underline{\mathbf{x}}_\tau\right). \tag{2}$$

The stochastic process $\underline{q}_\tau$ is assumed to correspond to the real process, while the outcome of the deterministic model (D-model) of eq. (1) is an estimate thereof. By considering $\mathbf{x}_\tau$ in eq. (1) as a stochastic process, retaining however the function $G(\neq \underline{g})$ as a deterministic function, we obtain the estimator $\underline{Q}_\tau$ of the output $\underline{q}_\tau$ as:

$$\underline{Q}_\tau = G\left(\underline{\mathbf{x}}_\tau\right). \tag{3}$$

To advance from the D-model, in its form (3), to the S-model in (2) we just need to specify the conditional distribution:

$$F_{q|Q}(q|Q) = P\left\{\underline{q} \leq q | \underline{Q} = Q\right\}, \tag{4}$$

with q and Q assumed concurrent and referring to discrete time τ. In other words, here conditioning is meant in scalar setting. An extension where Q is a vector containing the current and earlier predictions by the D-model and possibly other variables is straightforward but not considered here (see also the discussion in Section 6).

It is relatively easy to infer from data the marginal distribution and density functions of the S-variable $\underline{q}$ and D-predicted variable $\underline{Q}$. Therefore we may assume that $f_q(q)$

and $f_Q(Q)$ are known. Then the conditional density sought should obey

$$\int_{-\infty}^{\infty} f_{q|Q}(q|Q)dq = 1 \tag{5}$$

and

$$\int_{-\infty}^{\infty} f_{q|Q}(q|Q)f_Q(Q)dQ = f_q(q). \tag{6}$$

Eq. (5) is trivial. If we set $z = F_Q(Q)$ in (6), with $Q = F_Q^{-1}(z)$, so that $f_Q(Q)dQ = dz$, we obtain

$$\int_0^1 f_{q|Q}\left(q|F_Q^{-1}(z)\right)dz = f_q(q). \tag{7}$$

By integration one finds

$$\int_0^q \int_0^1 f_{q|Q}\left(a|F_Q^{-1}(z)\right)dzda = F_q(q), \tag{8}$$

and changing the order of the integrals we finally find

$$\int_0^1 F_{q|Q}\left(q|F_Q^{-1}(z)\right)dz = F_q(q). \tag{9}$$

At this stage, if one has time series of concurrent Q and q, each of size n, and if $Q_{(i:n)}$ is the ith smallest value in the time series of Q and $q_{(j:n)}$ is the jth smallest value in the time series of q, then the approximations $F_Q(Q_i) \approx i/n$ and $F_q(q_j) \approx j/n$ can be used and thus one approximates $F_q(q)$ in (9) as

$$\frac{1}{n}\sum_{i=1}^{n} F_{q|Q}\left(q|Q_{(i:n)}\right) \approx F_q(q), \tag{10}$$

and, for $q = q_j$,

$$\frac{1}{n}\sum_{i=1}^{n} F_{q|Q}\left(q_{(j:n)}|Q_{(i:n)}\right) \approx \frac{j}{n}. \tag{11}$$

Hence,

$$B_j := \sum_{i=1}^{n} F_{q|Q}\left(q_{(j:n)}|Q_{(i:n)}\right) = j. \tag{12}$$

We can thus attempt to determine $F_{q|Q}$ by minimizing the quantity

$$A := \sum_{j=1}^{n}(B_j - j)^2 = \sum_{j=1}^{n}\left(\sum_{i=1}^{n} F_{q|Q}\left(q_{(j:n)}|Q_{(i:n)}\right) - j\right)^2, \tag{13}$$

therefore obtaining the desired conditional distribution which leads to the formulation of the S-model corresponding to the D-model.

3.1 Determining the conditional distribution

In real world applications the D-model will provide an uncertain and possibly biased prediction. In such cases the S-model is applied by sampling from the conditional distribution $F_{q|Q}(q|Q)$ which incorporates both a shift of the prediction Q toward the real value q (bias correction) and the probabilistic assessment of the stochastic error (uncertainty assessment). A necessary preliminary step is the definition of the above conditional distribution as defined by eq. (4).

One strategy to tackle the problem is to use a parametric relationship for the function $F_{q|Q}(q|Q)$ and determine its parameters by minimizing the quantity A in eq. (12).

A possibility would be to assume $F_{q|Q}(q|Q)$ to be a Pareto-Burr-Feller (PBF) distribution (see Koutsoyiannis (2021)) with constant tail indices ξ and ζ and scale parameter varying with Q. A similar approach would be to assume a copula $C\left(F_q(q), F_Q(Q)\right)$ and determine $F_{q|Q}(q|Q)$ as

$$F_{q|Q}(q|Q) = \frac{F_{qQ}(q,Q)}{f_Q(Q)}, \tag{14}$$

with

$$F_{q|Q}(q,Q) = C\left(F_q(q), F_Q(Q)\right). \tag{15}$$

While a parametric approach like the above is attractive from many aspects, here we propose a fully data based approach, i.e. we try to determine $F_{q|Q}(q|Q)$ from the data alone (see figure 1). As the variables of interest in hydrology are of continuous type, we may expect that each value Q_τ in the available time series appears only once. Thus we cannot form a sample of observed data for a particular value of Q. However, as a simple approximation of $F_{q|Q}(q|Q)$, we can form a sample $\overline{q}_i, i = 1, ..., (m_1 + m_2 + 1)$, of Q-neighbours based on:

$$\begin{aligned} F_{q|Q}(q|Q) = P\left\{\underline{q} \le q|\underline{Q} = Q\right\} \approx P\left\{\underline{q} \le q|Q - \Delta Q_1 \le \underline{Q} \le Q + \Delta Q_2\right\} \approx \\ \approx P\left\{\underline{q} \le q|F_Q(Q) - \Delta F_1 \le F_Q(\underline{Q}) \le F_Q(Q) + \Delta F_2\right\} =: F_{q|Q}\left(q|Q, \Delta F_1, \Delta F_2\right), \end{aligned} \tag{16}$$

where the increments ΔQ_i and ΔF_i can be chosen based on the requirement that the intervals below and above the value Q (or $F_Q(Q)$) contain appropriate numbers of data values, $m_1 := \Delta F_1 n$ and $m_2 := \Delta F_2 n$, respectively. The numbers m_1 and m_2 should not be too large, so that $F_Q(Q) \pm \Delta F_{1,2}$ be close to $F_Q(Q)$, nor too small, so that the probability

$$P\left\{\underline{q} \le q|(F_Q(Q) - m_1/n \le F_Q(\underline{Q}) \le F_Q(Q) + m_2/n)\right\} \tag{17}$$

can be estimated from the sample of $\overline{q}_i$. From the above probability distribution one can easily estimate the mean value, or alternatively the median which may be more robust against outliers, which gives the S-model prediction. As for the confidence limits one possibility is to compute empirical quantiles through order statistics. For example, one may choose $\Delta F_1 = \Delta F_2 = \Delta F$ and $m_1 = m_2 = m$. If one sets, say, $m_1 = m_2 = m = 20$, i.e. $m_1 + m_2 + 1 = 41$, the lowest and highest quantiles that can be empirically estimated would correspond to $1/41 \approx 2.5\%$ and $1 - 1/40 \approx 97.5\%$, respectively. Conversely, for probabilities 2.5% and 97.5%, which correspond to a confidence level of 95%, we can empirically estimate the corresponding quantiles of q as the minimum and the maximum observed value, respectively, in the sample $\overline{q}_i$ of $m_1 + m_2 + 1$ values.

One should note that a sample size of $m_1 + m_2 + 1$ may not be obtained for the extreme values of the simulation, for which a number m_1 of lower predictions and a number m_2 of higher ones may not be available. In such cases the sample size need to be reduced accordingly.

We point out that order statistics deliver quantile estimation for a limited set of probabilities that correspond to the frequency of data in the sample $\overline{q}_i$. Therefore the above approach cannot be used for estimating quantiles for arbitrary probabilities of the conditional distribution $F_{q|Q}(q|Q)$. When such need arises, for instance when performing large ensemble simulations, a parametric relationship for $F_{q|Q}(q|Q)$ should be adopted and fitted as suggested above. Since here we do not use a parametric approach, we will handle this problem by the concept of K-moments discussed in section 3.2, noting though that even this cannot exceed some limits imposed by the subsample length (m_1+m_2+1).

3.2 Robust Estimation of Empirical Quantiles

The above empirical estimation of quantiles through order statistics is based on one data point only, as it identifies the single observation that is closer in frequency to the probability that corresponds to the desired confidence level. A possible solution to increase the robustness of the estimation is offered by the recently introduced concept of knowable moments (K-moments, see Koutsoyiannis (2019, 2021)) which gives an alternative for empirical quantile evaluation that is more reliable than order statistics as it combines many data points in each estimate. Furthermore, K-moments offer unbiased estimates of distribution quantiles, while the order statistics enable unbiased estimates of the distribution function. The two estimates may differ substantially for heavy-tailed distributions.

The noncentral knowable moment (or noncentral K-moment) of order (p, q) of the random variable $\underline{x}$ is defined as (Koutsoyiannis, 2019)

$$K'_{pq} := (p - q + 1)\mathrm{E}\left[(F(\underline{x}))^{p-q}\, \underline{x}^q\right], \tag{18}$$

with $p \geq q$ and E indicating the expected value. A most interesting special case is $q = 1$. In fact, the noncentral knowable moment of order $(p, 1)$ is given by

$$K'_p = p\mathrm{E}\left[(F(\underline{x}))^{p-1}\, \underline{x}\right], \tag{19}$$

with $p \geq 1$. A basic property that connects the K-moments with expectations of maxima is

$$K'_p = \mathrm{E}\left[\underline{x}_{(p)}\right] = \mathrm{E}\left[\max\left(\underline{x}_1, \underline{x}_2, ..., \underline{x}_p\right)\right]. \tag{20}$$

For expectations of minima another type of K-moments is defined, as described in Koutsoyiannis (2021). Therefore, by definition K'_p represents the expected value of the maximum of p copies of $\underline{x}$ and thus it is an estimate for the empirical quantile, which is computed by considering the whole data sample.

A key step in the above procedure is the estimation of two K-moment orders p_h and p_l, corresponding to the desired confidence level, for the upper and lower confidence limit, respectively. We illustrate here below the procedure for computing p_h and refer to Koutsoyiannis (2021) for details on the computation of p_l.

First, let us introduce the Λ-coefficient of order p_h as

$$\Lambda_{p_h} := \frac{1}{p_h\left(1 - F(K'_{p_h})\right)}. \tag{21}$$

Λ_{p_h} varies only slightly with p_h. Any symmetric distribution will give exactly $\Lambda_1 = 2$ because K'_1 is the mean, which in a symmetric distribution coincides with the median and thus yields $F(K'_{p_h}) = 1/2$. The exact value Λ_1 is easy to determine, as it is directly related to the mean, namely,

$$\Lambda_1 := \frac{1}{1 - F(\mu)}, \tag{22}$$

while the exact value of Λ_∞ depends only on the tail index ξ of the distribution according to

$$\Lambda_\infty = \begin{cases} \Gamma(1-\xi)^{\frac{1}{\xi}} \\ e^\gamma \end{cases} \tag{23}$$

where $\gamma = 0.577$ is the Euler's constant.

Basing on the above estimates for Λ_1 and Λ_∞ the following approximation may be used for estimating Λ_{p_h}, which is satisfactory for several distributions:

$$\Lambda_{p_h} \approx \Lambda_\infty + \frac{\Lambda_1 - \Lambda_\infty}{p_h}, \tag{24}$$

and, substituing in eq. (21)

$$F\left(K'_{p_h}\right) \approx 1 - \frac{1}{\Lambda_\infty p_h + (\Lambda_1 - \Lambda_\infty)}. \tag{25}$$

Conversely, for a given non-exceedance probability F, we can calculate the quantile x as the K'_{p_h} that corresponds to:

$$p_h \approx \frac{1}{\Lambda_\infty (1-F)} + 1 - \frac{\Lambda_1}{\Lambda_\infty} \tag{26}$$

where, in our case, $F = 1-\alpha/2$, being α the significance level of the confidence band.

For estimating Λ_1 an expression for the probability distribution of F is to be selected and plugged into eq. (22). Koutsoyiannis (2021) provides ready-to-use relationship for Λ_1 for several probability distributions. The distribution F can be assumed to be invariant over the range of the simulated river flows. Therefore, estimates for the tail index can be obtained by fitting the whole observed data sample (or the mean prediction sample obtained with the S-model) with a suitable probability distribution (we use the Pareto-Burr-Feller distribution for the case studies presented in Section 5). Note that the above distributional assumption on the whole data set has the only purpose of providing estimates for the tail index ($F(\mu)$ is also required but this can readily be estimated from data even without fitting a distribution) and therefore we do not make any assumption on the distribution of each individual sample that is used for the estimation of the empirical quantiles at each time step.

4 Assessment of Goodness of Fit

Assessment of performance is essential to provide end users with an indication of the reliability of the S-model and its confidence limits. Besides providing values of the Pearson correlation coefficient between observed and simulated data and the Nash efficiency for both the D-model and S-model, we also draw the diagnostic plots described below and report the percentage of observations lying outside the confidence limits, estimated by using both order statistics and robust estimation.

4.1 Combined Probability-Probability (CPP) Plot

A simple graphical test is introduced here to assess the performances of the S-model. It is based on the comparison of the marginal distributions of observed and predicted variables. Here we refer to it as “Combined Probability-Probability” (CPP) plot. CPP is a plot of the empirical distribution function $F_w(w)$ of a stochastic variable $\underline{w}$ against its value w. The variable is defined as the non-exceedance probability:

$$\underline{w} := F_Q(\underline{q}). \tag{27}$$

Its distribution function is $F_w(w) = P\left\{\underline{w} \leq w\right\} = P\left\{F_Q(\underline{q}) \leq w\right\} = P\left\{\underline{q} \leq F_Q^{-1}(w)\right\}$ and hence:

$$F_w(w) = F_q\left(F_Q^{-1}(w)\right). \tag{28}$$

In other words, $F_w(w)$ combines the distribution functions of predictions Q and real quantities q. The predictions are regarded as good if the plot $F_w(w)$ versus w is the equality line, i.e., if $F_w(w) = w$, which means that the distribution of w is uniform. In this case $F_q^{-1}(w)) = F_Q^{-1}(w)$. This is possible only if $F_Q(x)$ is identical to $F_q(x)$, which is what we would like to check. Note that a CPP plot lying above (below) the equality line indicates overprediction (underprediction) while a S-shaped CPP plot with the initial part above (below) the equality line and the second part below (above) the equality line indicates overestimation of low (high) flows and underestimation of high (low) flows.

In essence, the plot tests whether the two distributions, estimated from the data, are identical. We note that the CPP plot, except for assessing the proximity of the two marginal distributions, does not give any other indication if the predictions are good.

For example if $\underline{Q}$ is completely independent from $\underline{q}$ (as it may happen if an obviously irrelevant model is used) but the two distributions are identical, again the distribution of w will be uniform.

4.2 Predictive Probability-Probability Plot

A second check is herein used to verify the reliability of the estimated confidence band. Laio and Tamea (2007) have introduced a diagnostic plot combining probability distributions of predictions and true values, which has become later popular in similar studies, having been termed "predictive quantile-quantile" (PQQ) plot (Eslamian, 2014), even though in the original paper it has been called simply probability plot. Here we refer to it as "predictive probability-probability" (PPP) plot because the plot actually represents probabilities. PPP is a plot of the empirical distribution function $F_z(z)$, of a stochastic variable $\underline{z}$, where the latter also represents probability, i.e., a conditional non-exceedance probability, namely

$$\underline{z}_Q := F_{q|Q}(\underline{q}). \tag{29}$$

In other words, $\underline{z}$ is the distribution function of the prediction evaluated for the observed value of the predictand. The idea of PPP comes from the Rosenblatt's result that for any stochastic process $\underline{x}_\tau$ in discrete time $\tau = 1, 2, \ldots$, the sequence of variables $\underline{z}_1, \underline{z}_2, ..., \underline{z}_\tau$, whose values are:

$$z_\tau := P\left\{\underline{x}_\tau \leq x_\tau | \underline{x}_{\tau-1} = x_{\tau-1}, ..., \underline{x}_1 = x_1\right\} = F_{x_\tau|x_{\tau-1}}(x_\tau|\mathbf{x}_{\tau-1}) \tag{30}$$

are independent and identically distributed with uniform distribution in $[0, 1]$. Note that here we used the vector notation $\mathbf{x}_{\tau-1} := [x_{\tau-1}, ..., x_1]^T$ to represent all values of the process earlier than τ. One may see an analogy of $\underline{z}_Q$ defined in eq. (29) with z_τ defined in (30) as they both are predictive distributions. Extending this analogy, one would expect that different $\underline{z}$ defined by eq. (29) would also be independent and identically distributed, which allows considering the different values as a sample of a single variable $\underline{z}$. In turn, this enables estimating the distribution function of $\underline{z}$ from the sample.

The information conveyed by the PPP plot is useful as it provides an overview of the reliability of the estimated confidence band for any confidence level, by showing departures of the calibrated predictive distribution from the optimal one. Specifically, a shape of the validation curve above or below the equality line indicates overprediction and underprediction, respectively, while a shape above (below) the equality line in the first part of the diagram and below (above) the same line in the second part means that the forecast is narrow (large). Figure 2 provides a graphical overview of the above features, while more details are given by Laio and Tamea (2007). Furthermore, the departure of the PPP plot from the equality line is a relative (with respect to the sample size) measure of the number of points lying below the lower and above the upper confidence limit. For example, coverage probabilities for confidence level of 0.8 are related to segments A and B in figure 2.

In fact, the percentage of observations lying below a confidence limit is such that for a given Q the probability that the true discharge is not greater than q is

$$P\left\{\underline{q} \leq q | \underline{Q} = Q\right\} = F_{q|Q}(q). \tag{31}$$

If we choose a non-exceedance probability α, $0 \leq \alpha \leq 1$, so that, for any Q, $F_{q|Q}(q) = \alpha$ then the latter relationship specifies a confidence curve for q, which is a function $q = h(Q)$, given that α is constant. The probability

$$P\left\{\underline{q} \leq h(Q) | \underline{Q} = Q\right\} = F_{q|Q}(h(Q)) = \alpha \tag{32}$$

is constant, independent of Q. Moreover, given the definition of z and its property not to depend on Q, one obtains $z = \alpha$. If the distribution of z is uniform in (0,1), i.e. $F_z(z) =$

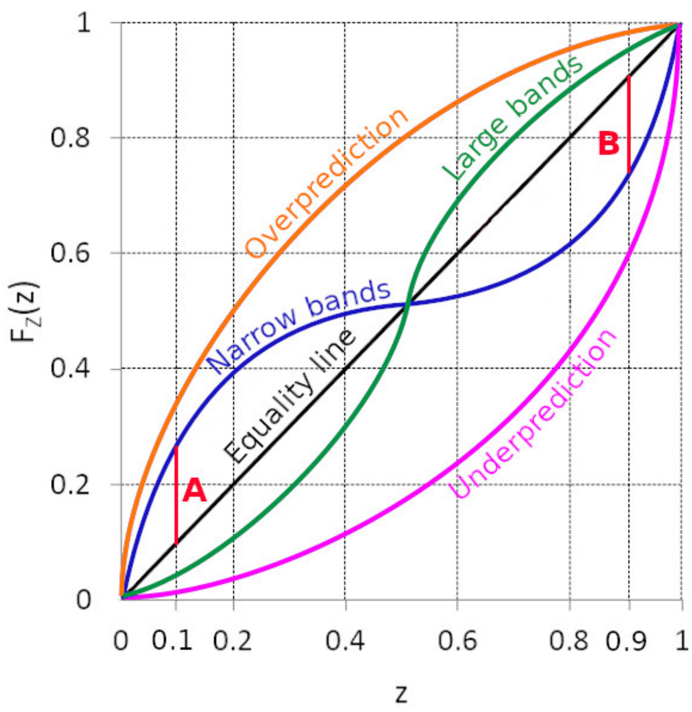

Figure 2. Information conveyed by the PPP plot.

z, the value of $F_z(z)$ at the point $z = \alpha$ will be equal to α. Therefore any deviation from uniformity is a relative measure of the number of observations exceeding the value α that would be expected that fall outside the confidence limit.

Note that the non-parametric fully data based approach of Bluecat infers $F_{q|Q}(\underline{q})$ in calibration from eq. (16), basing on subranges of Q. Therefore, if one estimates the z_τ sample for the same values of Q the empirical distribution of z will be clearly uniform, regardless of the D-model performance or any other feature of the processes q_τ and Q_τ. Therefore, the PPP plot for the calibration period of Bluecat will always be a straight line (equality line) by definition, because the data to be predicted are those that have been used to estimate the predictive distribution.

5 Case Studies

Bluecat was first tested with control experiments that have been presented by Koutsoyiannis and Montanari (2020). These confirmed the capability of the method to estimate reliably stochastic predictions and coverage probabilities in controlled conditions.

Here we present two case studies to test the performances of Bluecat in real world applications. They refer to the cases of the Arno river at Subbiano and the Siver river at Fornacina, for which a rainfall-runoff model is used to elaborate river flow predictions. The Sieve river is a tributary of the Arno river. They flow in the Tuscany Region, in Italy. Figure 3 presents a schematic map of the river basins. Climate is continental with low flows during Summer and high flows in the Fall and Spring seasons. Occasionally high flow events may occur during the winter.

We apply to both case studies the rainfall-runoff model HyMod (Boyle, 2000; Montanari, 2005) with 5 parameters. These are C_m[length], the maximum storage capacity within the basin), β[dimensionless], the degree of spatial variability of the soil moisture capacity within the basin, α[dimensionless], a factor for partitioning the flow between two routing procedures, k_1 [time] and k_2 [time], characteristic times for the two routing components.

For both case studies we calibrated the HyMod model by minimizing the Nash-Sutcliffe efficiency. It is well known that performance metrics are affected by significant sampling uncertainty (Clark et al., 2021; Barber et al., 2020). Lamontagne et al. (2020) have shown that estimation robustness may be improved by performing a preliminary logarithmic transformation of observed and simulated river flow data. Therefore, we considered the following transformation, which can be applied also to intermittent river flows (Koutsoyiannis, 2021):

$$y = \lambda \log(1 + \frac{x}{\lambda}) \tag{33}$$

where x and y are original and transformed data, respectively, and λ is a parameter. For $\lambda \rightarrow 0$ and $\lambda \rightarrow \infty$ eq. (33) becomes equivalent to the logarithmic and the identity ($y = x$) transform, respectively.

It is well known that a limited training for hydrologic models may cause overparameterisation, which in turn implies that model performances in calibration may not deliver a useful information on the reliability of model predictions in validation. This issue will be further discussed in Section 6.

We estimated confidence limits by applying both robust estimation and order statistics by adopting a confidence level of 80%. We selected $m_1 = m_2 = 100$ which means that each prediction distribution is estimated over a sample of $m_1 + m_2 + 1 = 201$ observations. For the extreme values of the prediction the sample size was reduced when enough lower/higher predictions were not available (see the note at the bottom of Section 3.1). The S-model predictions were obtained by estimating the median value of the conditional probability distribution given by eq. (4), although CPP plots were drawn for the mean stochastic prediction as well.

Median prediction and confidence band for the S-model were estimated for both the calibration and validation period. Of course we expect better performances of the S-model for the calibration period while the validation exercise is expected to provide an indication of the Bluecat performances for out of sample prediction. Goodness of fit is estimated by the performance indicators discussed in Section 4.

5.1 Arno River at Subbiano

The catchment of the Arno river at Subbiano is located within the mountain belt of the Northern Apennines, with mean, minimum and maximum elevation of 750, 250 and 1657 m above sea level, respectively. The catchment area is about 752 km^2 and the average catchment slope is about 14%. The data of mean areal daily rainfall (estimated from raingauge observations) and evapotranspiration (estimated from temperature data) span the 22-year period 1992-2013. We use the first 20 years for model calibration and the last two years for model validation. Optimization was performed after transforming data as in eq. (33) with $\lambda = 0.0001$, a value that was selected by looking at the S-model performances in calibration. Calibrated model parameters are given in Table 1. For the calibration period the Pearson correlation coefficient between the D-model outputs Q and the observed values q is 0.84, which means that the model is able to explain $0.84^2 = 71\%$ of the total variance. The Nash efficiency is 0.63.

Figure 4 shows the results of the application of Bluecat in calibration mode with robust estimation. In the left panel a scatterplot of D-model predictions versus observed values and S-model predictions is shown, along with the related confidence limits. The

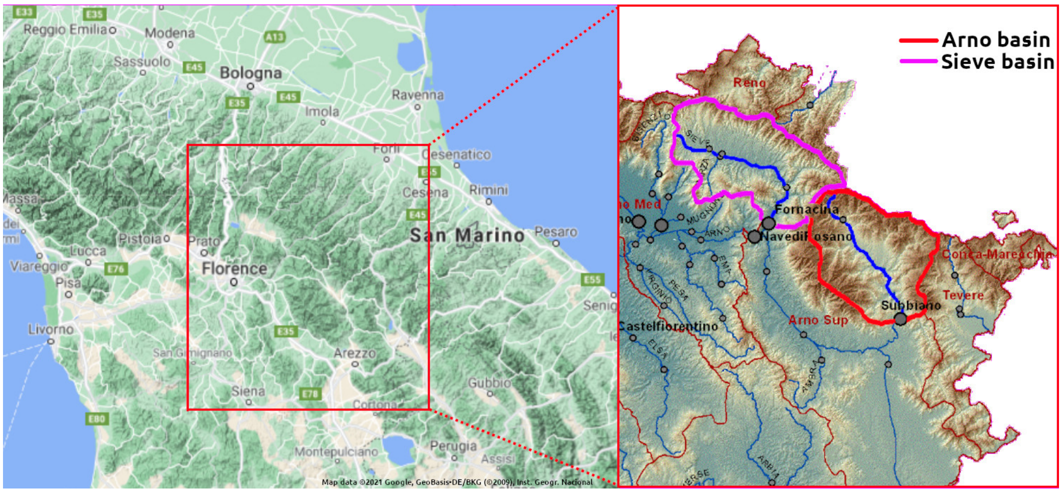

Figure 3. Basins of the Arno river at Subbiano and the Sieve river at Fornacina.

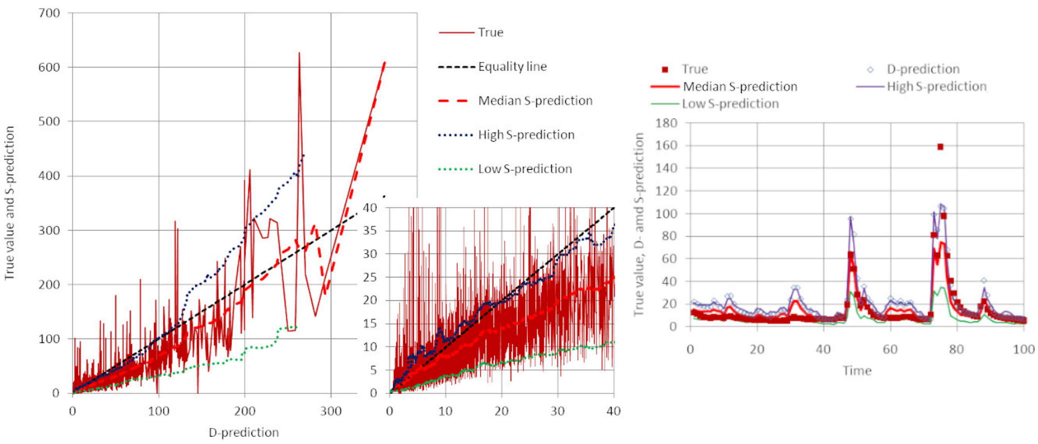

Figure 4. D-model and S-model predictions, along with confidence limits, for the calibration period of the Arno river at Subbiano. The right panel depicts 100 days of the calibration period, where the first day is January 1st, 2011.

inset shows a detailed representation of the low flow range. The right panel depicts 100 days of the calibration period, where the first day is January 1st, 2011.

The S-model displayed improved predicting performances, with a Pearson correlation coefficient of 0.88 and a Nash efficiency of 0.77 (median prediction). Figure 4, particularly in the inset, also shows that the D-model overpredicts low discharges and underpredicts high ones. The bias is reduced by the S-model. Coverage probabilities are reported in Table 2, for confidence band estimated with both order statistics and robust estimation. The CPP plot, shown in Figure 6, confirms the prediction bias of the D-model and the improved performances of the S-model which, however, still overpredicts the low flows as Figure 4 anticipated.

The results of the validation are shown in Figure 5, Table 2 and Figure 6. The right panel in Figure 5 depicts 100 days of the validation period, where the first day is January 1st, 2013. The D-model performance in validation is summarised by a Pearson cor-

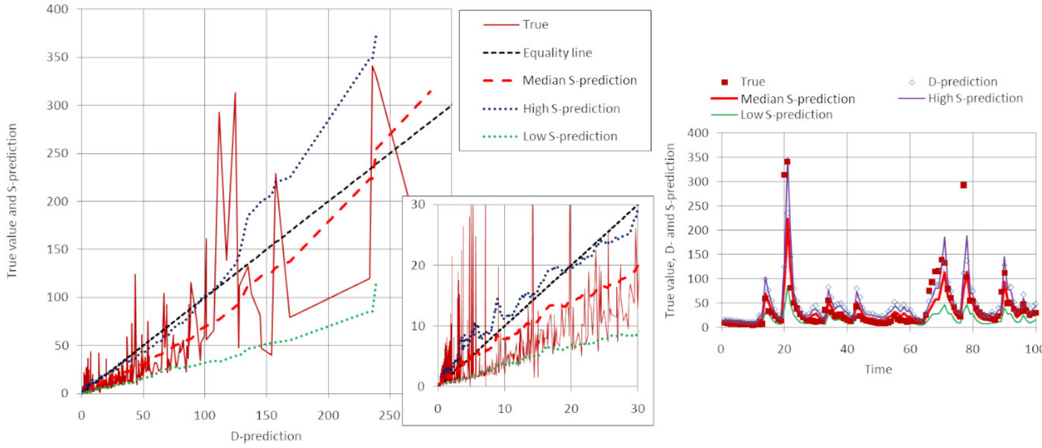

Figure 5. D-model and S-model predictions, along with confidence limits, for the validation period of the Arno river at Subbiano. The right panel depicts 100 days of the validation period, where the first day is January 1st, 2013.

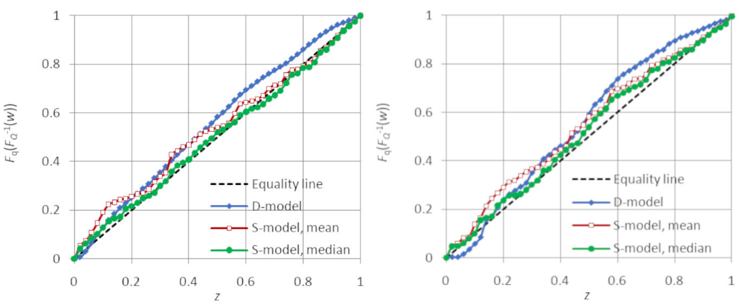

Figure 6. Combined probability-probability (CPP) plots for the predictions of the river flows of the Arno river at Subbiano in calibration (left) and validation (right).

Table 1. HyMod model calibrated parameters for the considered case studies

Basin	$C_m[mm]$	$\beta[-]$	$\alpha[-]$	k_1[days]	k_2[days]
Arno	336	0.10	0.61	24.34	1.25
Sieve	323	0.20	0.55	4.61	357.53

relation coefficient of 0.80 and a Nash efficiency of 0.57. Sligthly better performances are given by the S-model prediction, with Pearson coefficient of 0.81 and a Nash efficiency of 0.62. The CPP plot confirms that the S-model improves the performances in terms of probability distribution of the predictions and proves the slightly better performances of the median with respect to the mean of the probability distribution given by eq. (4) to compute the S-model prediction. It also suggests an overestimation and underestimation of low and high flows, respectively.

The PPP plot is reported in Figure 10 (left) and shows that in validation the confidence limits are narrow. This outcome is confirmed by the percentage of observations lying outside the confidence limits, which are reported in Table 2, which are higher than the values of 10% for each band that one would expect for a confidence level of 80%. Further consideration on the PPP plot results for the Arno River are found in the Section 6.

5.2 Sieve River at Fornacina

The Sieve river is a tributary of the Arno river that is also located in the Northern Apennines, with mean, minimum and maximum elevation of 488, 96 and 1637 m above sea level, respectively. The catchment area is about 846 km^2 and the average catchment slope is about 12%. The data of mean areal hourly rainfall (estimated from raingauge observations) and evapotranspiration (estimated from temperature observations) span the 5-year period 1992-1996. The flow regime of the Sieve river is intermittent with the presence of about 4% of zero values in the available record.

We use the data from June 1st, 1992 to December 31st, 1994 for model calibration and the data from June 2nd, 1995 to December 31st, 1996 for model validation. Note that we discarded the January-May period for both calibration and validation because high flows typically occur in that season that are not satisfactorily reproduced by HyMod for the limited duration of the model warm up. We maximized the Nash-Sutcliffe efficiency to calibrate the parameters without applying any transformation to the data, as this led to the best S-model performances in terms of median prediction and coverage probabilities.

Calibrated model parameters are given in Table 1. For the calibration period the correlation coefficient between the D-model outputs Q and the observed values q is 0.91, which means that the model is able to explain 82% of the total variance. The Nash efficiency is 0.81. Figure 7 confirms the good fit of the model in calibration. The right panel depicts 150 hours of the calibration period starting from September 16th, 1992 at 5 AM.

The calibration results confirm the improved performances of the S-model, whose mean prediction has a Pearson correlation coefficient of 0.94 and Nash efficiency of 0.88. Figure 7, particularly in the inset, shows that the S-model corrects the prediction bias of the D-model. The percentage of points lying above and below the confidence limits is reported in Table 2. The CPP plot, shown in Figure 9, confirms the improved performances of the S-model and in particular its effectiveness in correcting the D-model bias in the high flow domain.

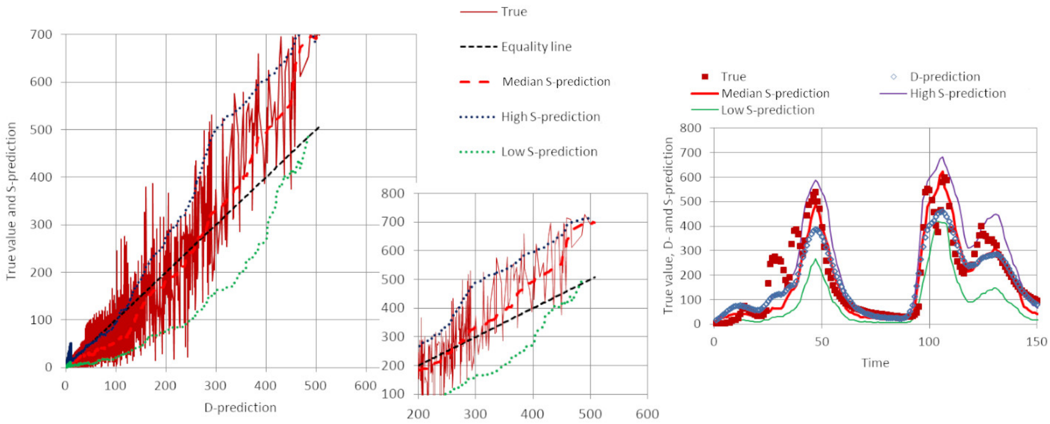

Figure 7. D-model and S-model predictions, along with confidence limits, for the calibration period of the Sieve river at Fornacina. The right panel depicts 150 hours of the calibration period starting from September 16th, 1992 at 5 AM.

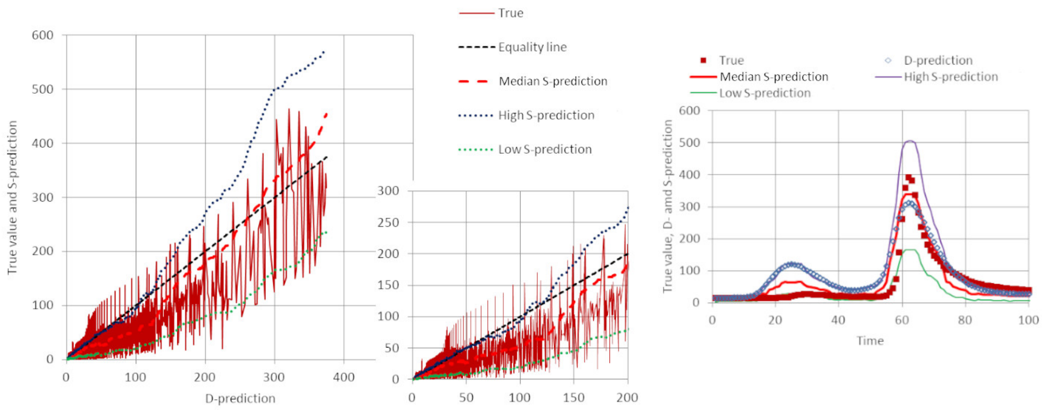

Figure 8. D-model and S-model predictions, along with confidence limits, for the validation period of the Sieve river at Fornacina. The right panel depicts 150 hours of the validation period starting from January 5th, 1996.

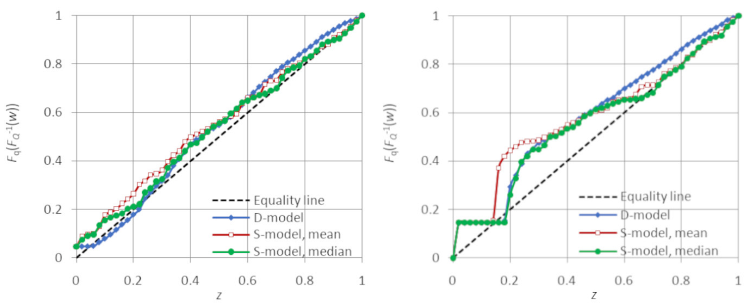

Figure 9. Combined probability-probability (CPP) plots for the predictions of the river flows of the Sieve river at Fornacina in calibration (left) and validation (right).

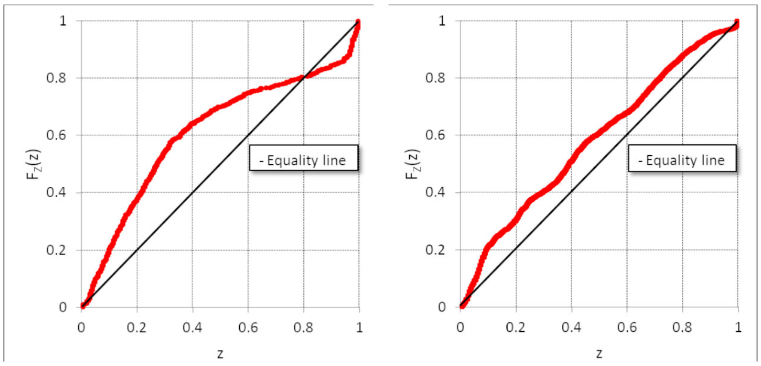

Figure 10. Predictive probability-probability (PPP) plots for the validation of the river flows predictions for the Arno river (left) and the Sieve river (right).

Table 2. Percentage of observations lying outside the 80% confidence limits for the considered case studies. Band was estimated with both order statistics and robust estimation. Subscripts h and l refer to upper and lower limit, respectively.

Arno calibration		Arno validation		Sieve calibration		Sieve validation	
$\%_h$	$\%_l$	$\%_h$	$\%_l$	$\%_h$	$\%_l$	$\%_h$	$\%_l$
Robust estimation							
10%	8%	17%	16%	17%	7%	13%	14%
Estimation with order statistics							
9%	10%	17%	22%	8%	9%	6%	16%

Validation results are shown in Figure 8, where the right panel depicts 150 hours of the validation period starting from January 5th, 1996 at 12 AM, and Figure 9. The D-model performance in validation is summarized by a Pearson correlation coefficient of 0.87 and a Nash efficiency of 0.53. The low value of the Nash efficiency is due to the significant overestimation of the low flows by the D-model. It is interesting to note that the S-model prediction exhibits a better fit with a Pearson coefficient of 0.88 and a Nash efficiency of 0.66. The latter is markedly improved thanks to the capability of Bluecat to correct the prediction bias. As for the confidence band, the PPP plot shows overall a good fit with a slight overprediction especially with regard to the lower limit (see also Table 2).

The results of the two case studies will be further discussed in Section 6.

6 Discussion

In introducing Bluecat we assumed that the probability distribution of the observed data, conditioned to the D-model simulation, can be reliably inferred from a calibration exercise (see Sections 2 and 3). Actually, this assumption holds asymptotically, namely, when the size of the calibration data sample is large. Furthermore, the assumption that we made that input and parameter uncertainty are satisfactorily resembled by the probability distribution given by eq. (4) also holds asymptotically.

When the calibration data set is not extended enough one may experience overparameterisation, which implies that the calibrated model exhibits satisfactory perfor-

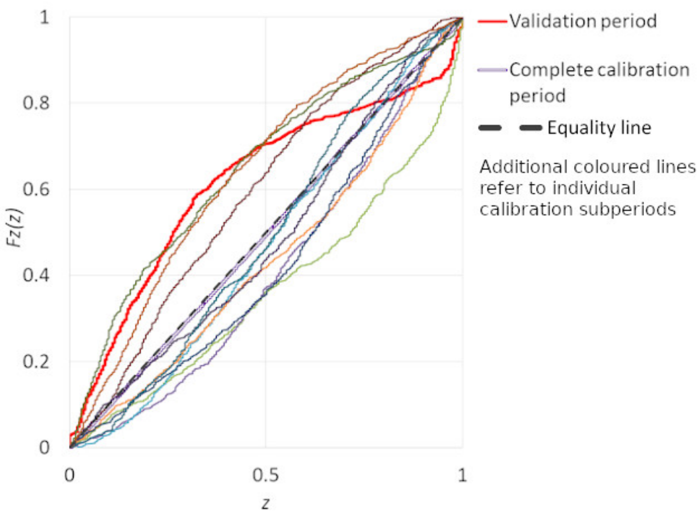

Figure 11. Sampling variability for the PPP plot of the Arno river in calibration and comparison with the PPP plot in validation.

mances in calibration that are not confirmed in validation. Therefore, in such cases the D-model errors in calibration may be much smaller than those in validation, which implies that the S-model generated by Bluecat may underestimate prediction uncertainty. That is, confidence band may be narrow, which means that the PPP plot would be S-shaped with the first and second part displaced above and below the equality line, respectively.

Furthermore, a limited extension of the validation period may imply uncertainty due to sampling variability. Namely, even if the confidence limits are statistically correct they may still provide a poor assessment of uncertainty when referring to specific and short prediction periods.

To inspect this issue, we performed an additional experiment for the Arno river by referring to the calibration period. We first computed the PPP plot in calibration, therefore obtaining an equality line as expected (see figure 11). Then, we redrew the PPP plot for 10 non-overlapping subperiods including 731 observations, which is precisely the length of the validation period. As expected, figure 11 shows that sampling variability causes a dispersion of the obtained PPP plots. For the sake of comparison, figure 11 also shows the PPP plot for the validation period, which is almost entirely included within the envelop of the calibration PPP plots obtained for the same sample size. Therefore, figure 11 shows that the deviation from the equality line that we obtaned for the Arno river in validation may be explained by sampling variability.

About the CPP plot, one should always take into account that the marginal distributions of predicted and observed data may be incidentally similar even if the prediction is poor. In particular, this may happen when the model performances in terms of correlation and Nash efficiency are far from satisfactory.

Regarding the case studies presented here it is intersting to note that for both Arno and Sieve rivers the stochastic prediction outperformed the deterministic model by correcting its bias for the various flow regimes. This outcome confirms that the additional value provided by the S-model is technically useful.

With regard to the confidence band, for the cases presented here, we indeed found that the observations lying outside the higher and lower confidence limits in validation are often higher than the value of 10% for each band that one would expect for a confidence level of 80% (see Table 2). Such deviations are expected when the simulation period is short, even due merely to sampling variability, as illustrated with the Arno river case study.

In technical applications it is important for the user to recognize the cases of "huge uncertainty in uncertainty assessment". First, we conclude that an accurate selection of the model calibration period is particularly important for Bluecat, which is calibrated at each local flow range. It is not possible to provide a general rule for assessing if a calibration period is long enough, as the answer depends on the type of model, the variability of the modeled processes, data seasonality and many others. It may be useful to split the available data sample in non-overlapping pieces and perform repeated validation tests to assess whether model performances are stable. The split sample exercise also allows to infer sampling variability. Second, we suggest that the final model training before application is carried out by using the largest possible data set and paying particular attention to detect possible model deficiencies that may not be resembled by the estimated conditional probability distribution of eq. (4).

We would like to discuss further the assumption of stationarity, which may be regarded as a limitation if one believes that the impact due to a possibly changing climate may be better predicted with a non-stationary approach (for an extended discussion on this subject see, e.g., Luke et al. (2017) and Montanari and Koutsoyiannis (2014a)). We also note that the conditional distribution given by eq. (4) might be seasonal, although part of the seasonality features are already incorporated in the D-model (for example, a large prediction of discharge would appear during the rainy, rather than the dry, period). There are many possible solutions for applying Bluecat in a non-stationary context. We may suggest to first consider a D-model with time varying (perhaps seasonal) parameters under the assumption that the uncertainty of the non-stationary model is described by a stationary distribution as given by eq. (4). If one would like to consider a non-stationary uncertainty, then a parametric and non-stationary distribution (perhaps a Pareto-Burr-Feller distribution with time varying seasonal parameters) may be adopted to describe uncertainty as described in Section 3.1, by paying particular attention to the increased risk of overparameterization that non-stationary models imply. Indeed, exploring the above dependencies in a stochastic framework would require an extended calibration data set to compensate the uncertainty introduced by additional model complexity. Overall, such modelling choices will unavoidably increase uncertainty and therefore would hardly be advisable for copying with real-word problems. If the extent of the data set is large enough, in cases justifying a seasonal approach, partioning the whole data set into seasons is a possible solution to ensure that both the D-model and Bluecat provide a good fit of seasonality. If a permanent change of the process statistics is detected (e.g. due to urbanization) it would recommendable to "stationarize" the data, adapting them to the current conditions and perform similar adaptations to the D-model. This is similar (albeit opposite) to "naturalization" of data series that is typically made in cases of river modifications due to dams and so forth.

One may wonder what is the distinguishing behavior of Bluecat with respect to the approaches that we previously proposed (Montanari & Koutsoyiannis, 2012; Sikorska et al., 2015). We first note that Bluecat relies on different assumptions and procedures. In Montanari and Koutsoyiannis (2012) we adopted a meta-Gaussian distribution to describe uncertainty of model predictions which were preliminarily transformed to stabilize their variance. Bluecat, in a similar manner as Sikorska et al. (2015)), avoids data transformation as the conditional probability distribution is automatically defined by the data. Furthermore, in Montanari and Koutsoyiannis (2012) and Sikorska et al. (2015) we accounted for parameter uncertainty at the expense of a more demanding approach for model calibration and application, which is a concern as in a data assimilation context calibration is to be frequently repeated. In fact, by avoiding any data transformation and offering a fast calibration, Bluecat allows technical applications with limited computational requirements and time.

Bluecat indeed shares some similarities with the nearest neighbouring method by Sikorska et al. (2015), which may be also used to correct the D-model bias (see, for instance, Ehlers et al. (2019)). However, we note that Bluecat infers the conditional probability distribution of the true data, while Sikorska et al. (2015) estimate the conditional probability distribution of the simulation error. Thus, they estimate the prediction uncertainty of the D-model rather than updating the D-model to the S-model. Therefore Bluecat provides a more comprehensive perspective. In view of the above differences, the user may select the most appropriate approach for the considered case study, with the awareness that model selection should be tailored to the underlying assumptions and operational needs.

Although Bluecat has been conceived to be applied to one single model, a multi-model application would be straightforward. It was already mentioned in Section 3 that an extension where Q is a vector containing the current and earlier predictions by the D-model is possible, yet here we study the simpler scalar version of the model. Likewise, the multi-model case is another possible vector version of Bluecat, where the vector Q contains the outcomes of the various D-models.

We believe that the application of Bluecat to the considered case studies offers encouraging performances for technical applications. Indeed, Bluecat does, under a rigorous statistical interpretation and clear assumptions, what the intuition of a technician would suggest: to correct model predictions and estimate their uncertainty by looking at model performances in the simulation of known data. It is a straightforward and extremely simple concept.

Finally, the end users should be informed that hydrologic modeling, including uncertainty assessment, is always uncertain and therefore the information provided by the confidence band should be interpreted critically. Nevertheless, this information is tremendously useful: by selecting an appropriate confidence level Bluecat provides the desired information for an assigned safety level of the prediction.

7 The Bluecat package

In order to facilitate the application of Bluecat we make available a software working under the R-enviroment (R Core Team, 2013) to fit the HyMod rainfall-runoff model and estimate its prediction uncertainty. Model fitting can be performed by maximizing the Nash-Sutcliffe efficiency using either untransformed data or transformed with equation (33), with the option of selecting two different optimization algorithms. Confidence limits can be defined by estimating empirical quantiles through order statistics or robust estimation (see Section 3.1 and 3.2). Assessment of the goodness of fit is performed by plotting the CPP and PPP plots and estimating the Nash-Sutcliffe efficiency. The software is accompanied by instructions (to be displayed with the R help function) and data bases of rainfall and potential evapotranspiration for the Arno and Sieve case studies that have been presented here. We also include instructions to be used within R to reproduce the case studies and the results we presented above.

While the package focuses on river flow prediction with HyMod, it can be easily adapted by substituting HyMod with any deterministic model. In fact, the model routine is isolated into a subroutine, currently written in the Fortran 95 programming language, that can be quickly replaced.

The software is available for download at the web address:
https://github.com/albertomontanari/hymodbluecat
along with instructions to compile it in R.

8 Conclusions

We introduce here a new method identified with the acronym "Bluecat" for simulating and predicting hydrologic processes, which is based on the use of a generic deterministic model that is subsequently converted to a stochastic formulation. The latter provides an update of the deterministic prediction along with uncertainty assessment with a transparent data based approach.

The results of the presented case studies confirm the distinguishing features of Bluecat, its reliability and robustness. In fact, for both case studies the stochastic version of the deterministic model provided an improvement of the performances of the deterministic model alone, both in calibration and validation. Furthermore, the estimated confidence band turned out to be informative: even if some uncertainty affected the estimation of coverage probabilities, we provided quantitative tests to verify their reliability. In fact, for both case studies Bluecat improved the prediction and provided confidence limits with an innovative and rigorous information content for technical applications.

In our opinion, for its computational efficiency and transparency Bluecat is a step forward for hydrogic modeling with uncertainty assessment. It is also flexible, as it can work in conjunction with any type of deterministic model and can be extended to multimodel applications or multiple predictor variables.

In view of technical applications, particular care is to be payed to the reliability and extension of the calibration data set. In fact, it is usual in hydrology to work in poorly gauged conditions, which may lead to overparametrisation, sampling variability and consequent inflation of uncertainty. Although Bluecat has been proven to be robust, the reliability of the deterministic model calibration should be carefully considered in order to avoid a "huge uncertainty in uncertainty assessment". We discussed potential solutions to support operational assessment of calibration reliability, which should ultimately rely on a careful assessment by end users.

When developing Bluecat and preparing this paper we decided to give high priority to simplicity, transparency, openness and reproducibility. For this reason we make available a software to support Bluecat operational applications and reproduction of the case studies presented here. We are looking forward to interacting with users for improving the software in an open access and open source context.

Acknowledgments

Preparation of this paper was slowed down by the outbreak of COVID19, which took all of the working time of AM for managing the emergency. We would like to address a special thought to those who suffered and lost their lives from the pandemic, with the hope that their sacrifice may help humanity to more effectively recognize that life and health are first priorities.

The software that has been used to develop this work, a package working under the R environment, is open source and available for download at:
https://github.com/albertomontanari/hymodbluecat.
We thank Emanuele Baratti and Elena Toth who helped to collect the data that were used for developing the case studies, which are included into the above software, along with instruction to reproduce the results presented here.

We are grateful to Richard Vogel, John Quilty, an anonymous reviewer and an anonymous Associate Editor for their encouraging and constructive comments, which greatly helped us to improve the paper. We are also thankful to Keith Beven and Alberto Viglione for providing very useful suggestions.

References

Barber, C., Lamontagne, J. R., & Vogel, R. M. (2020). Improved estimators of correlation and r2 for skewed hydrologic data. *Hydrological Sciences Journal*, *65*(1), 87–101.

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of hydrology*, *320*(1-2), 18–36. doi: 10.1016/j.jhydrol.2005.07.007

Beven, K. (2009). Comment on "equifinality of formal (dream) and informal (glue) bayesian approaches in hydrologic modeling?" by jasper a. vrugt, cajo jf ter braak, hoshin v. gupta and bruce a. robinson. *Stochastic environmental research and risk assessment*, *23*(7), 1059–1060. doi: 10.1007/s00477-008-0283-x

Beven, K., & Binley, A. (1992). The future of distributed models: model calibration and uncertainty prediction. *Hydrological processes*, *6*(3), 279–298. doi: 10.1002/hyp.3360060305

Beven, K., & Lane, S. (2019). Invalidation of models and fitness-for-purpose: a rejectionist approach. In *Computer simulation validation* (pp. 145–171). Springer.

Blöschl, G. (2008). Flood warning-on the value of local information. *International Journal of River Basin Management*, *6*(1), 41–50. doi: 10.1080/15715124.2008.9635336

Blöschl, G., Bloschl, G., Sivapalan, M., Wagener, T., Savenije, H., & Viglione, A. (2013). *Runoff prediction in ungauged basins: synthesis across processes, places and scales*. Cambridge University Press.

Boyle, D. (2000). *Multicriteria calibration of hydrological models* (Unpublished doctoral dissertation). Univ of Arizona, Tucson.

Bulygina, N., & Gupta, H. (2009). Estimating the uncertain mathematical structure of a water balance model via bayesian data assimilation. *Water Resources Research*, *45*(12). doi: 0.1029/2007WR006749

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J., Tang, G., . . . others (2021). The abuse of popular performance metrics in hydrologic modeling. *Water Resources Research*, e2020WR029001. doi: 10.1029/2020WR029001

Ehlers, L., Wani, O., Koch, J., Sonnenborg, T., & Refsgaard, J. (2019). Using a simple post-processor to predict residual uncertainty for multiple hydrological model outputs. *Advances in Water Resources*, *129*, 16–30.

Eslamian, S. (2014). *Handbook of engineering hydrology: modeling, climate change, and variability*. CRC Press.

Frame, J., Kratzert, F., Klots, D., Gauch, M., Shelev, G., Gilon, O., . . . Nearing, G. S. (2021). Deep learning rainfall-runoff predictions of extreme events. *Hydrology and Earth System Sciences Discussions*. doi: 10.5194/hess-2021-423

Kavetski, D., Fenicia, F., Reichert, P., & Albert, C. (2018). Signature-domain calibration of hydrological models using approximate bayesian computation: Theory and comparison to existing applications. *Water Resources Research*, *54*(6), 4059–4083.

Koutsoyiannis, D. (2019). Knowable moments for high-order stochastic characterization and modelling of hydrological processes. *Hydrological sciences journal*, *64*(1), 19–33. doi: 10.1080/02626667.2018.1556794

Koutsoyiannis, D. (2021). *Stochastics of hydroclimatic extremes – a cool look at risk*. Hellenic Academic Ebooks. Retrieved from `http://hdl.handle.net/11419/6522`

Koutsoyiannis, D., & Montanari, A. (2015). Negligent killing of scientific concepts: the stationarity case. *Hydrological Sciences Journal*, *60*(7-8), 1174–1183. doi: 10.1080/02626667.2014.959959

Koutsoyiannis, D., & Montanari, A. (2020). A brisk local uncertainty estimator for hydrologic simulations and predictions (blue cat). In *Egu general assembly conference abstracts* (p. 10125). doi: 10.5194/egusphere-egu2020-10125

Krzysztofowicz, R. (1999). Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research*, *35*(9), 2739–2750. doi: 10.1029/1999WR900099

Laio, F., & Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, *11*(4), 1267–1277. doi: 10.5194/hess-11-1267-2007

Lamontagne, J. R., Barber, C. A., & Vogel, R. M. (2020). Improved estimators of model performance efficiency for skewed hydrologic data. *Water Resources Research*, *56*(9), e2020WR027101.

Luke, A., Vrugt, J. A., AghaKouchak, A., Matthew, R., & Sanders, B. F. (2017). Predicting nonstationary flood frequencies: Evidence supports an updated stationarity thesis in the u nited s tates. *Water Resources Research*, *53*(7), 5469–5494. doi: 10.1002/2016WR019676

Montanari, A. (2005). Large sample behaviors of the generalized likelihood uncertainty estimation (glue) in assessing the uncertainty of rainfall-runoff simulations. *Water resources research*, *41*(8).

Montanari, A. (2011). Uncertainty of hydrological predictions. In P. Wilderer (Ed.), *Treatise on water science, vol. 2* (p. 459-478). Elsevier.

Montanari, A., & Brath, A. (2004). A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research*, *40*(1). doi: 10.1029/2003WR002540

Montanari, A., & Grossi, G. (2008). Estimating the uncertainty of hydrological forecasts: A statistical approach. *Water Resources Research*, *44*(12). doi: 10.1029/2008WR006897

Montanari, A., & Koutsoyiannis, D. (2012). A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research*, *48*(9). doi: 10.1029/2011WR011412

Montanari, A., & Koutsoyiannis, D. (2014a). Modeling and mitigating natural hazards: Stationarity is immortal! *Water resources research*, *50*(12), 9748–9756. doi: 10.1002/2014WR016092

Montanari, A., & Koutsoyiannis, D. (2014b). Reply to comment by grey nearing on "a blueprint for process-based modeling of uncertain hydrological systems". *Water Resour. Res*, *50*, 6264–6268. doi: 10.1002/2013WR014987

Nearing, G. S. (2014). Comment on "a blueprint for process-based modeling of uncertain hydrological systems" by alberto montanari and demetris kousoyiannis. *Water Resources Research*, *50*, 6260–6263. doi: 10.1002/2013WR014812

Pagano, T. C., Wood, A. W., Ramos, M.-H., Cloke, H. L., Pappenberger, F., Clark, M. P., ... others (2014). Challenges of operational river forecasting. *Journal of Hydrometeorology*, *15*(4), 1692–1707. doi: 10.1175/JHM-D-13-0188.1

Papacharalampous, G., Tyralis, H., & Koutsoyiannis, D. (2019). Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. *Stochastic Environmental Research and Risk Assessment*, *33*(2), 481–514. doi: 10.1007/978-3-540-79881-1{_}3

Papacharalampous, G., Tyralis, H., Koutsoyiannis, D., & Montanari, A. (2020). Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale. *Advances in Water Resources*, *136*, 103470. doi: 10.1016/j.advwatres.2019.103470

Papacharalampous, G., Tyralis, H., Langousis, A., Jayawardena, A. W., Sivakumar, B., Mamassis, N., ... Koutsoyiannis, D. (2019a). Large-scale comparison of machine learning regression algorithms for probabilistic hydrological modelling via post-processing of point predictions. In *Geophysical research abstracts* (Vol. 21). doi: 10.1007/s00521-020-05172-3

Papacharalampous, G., Tyralis, H., Langousis, A., Jayawardena, A. W., Sivakumar, B., Mamassis, N., ... Koutsoyiannis, D. (2019b). Probabilistic hydrological post-processing at scale: Why and how to apply machine-learning quantile regression algorithms. *Water*, *11*(10), 2126. doi: 10.3390/w11102126

Quilty, J., & Adamowski, J. (2020). A stochastic wavelet-based data-driven framework for forecasting uncertain multiscale hydrological and water resources processes. *Environmental Modelling & Software*, *130*, 104718.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Ramos, M. H., Van Andel, S. J., & Pappenberger, F. (2013). Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, *17*(6), 2219–2232. doi: 10.5194/hess-17-2219-2013,2013

Reggiani, P., Todini, E., Boyko, O., & Buizza, R. (2021). Assessing uncertainty for decision-making in climate adaptation and risk mitigation. *International Journal of Climatology*, *41*(5), 2891–2912. doi: 10.1002/joc.6996

Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-gaussian errors. *Water Resources Research*, *46*(10). doi: 10.1029/2009WR008933

Sikorska, A. E., Montanari, A., & Koutsoyiannis, D. (2015). Estimating the uncertainty of hydrological predictions through data-driven resampling techniques. *Journal of Hydrologic Engineering*, *20*(1), A4014009. doi: 10.1061/(ASCE)HE.1943-5584.0000926

Spear, R., & Hornberger, G. (1980). Eutrophication in peel inlet—ii. identification of critical uncertainties via generalized sensitivity analysis. *Water research*, *14*(1), 43–49. doi: 10.1016/0043-1354(80)90040-8

Tajiki, M., Schoups, G., Hendricks Franssen, H., Najafinejad, A., & Bahremand, A. (2020). Recursive bayesian estimation of conceptual rainfall-runoff model errors in real-time prediction of streamflow. *Water Resources Research*, *56*(2), e2019WR025237. doi: 10.1029/2019WR025237

Tyralis, H., Papacharalampous, G., Burnetas, A., & Langousis, A. (2019). Hydrological post-processing using stacked generalization of quantile regression algorithms: Large-scale application over conus. *Journal of Hydrology*, *577*, 123957. doi: 10.1016/j.jhydrol.2019.123957

Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water*, *11*(5), 910. doi: 10.3390/w11050910

Valdez, E. S., Anctil, F., & Ramos, M.-H. (2021). Choosing between post-processing precipitation forecasts or chaining several uncertainty quantification tools in hydrological forecasting systems. *Hydrology and Earth System Sciences Discussions*, 1–40. doi: 10.5194/hess-2021-391

Vrugt, J. A., Ter Braak, C. J., Gupta, H. V., & Robinson, B. A. (2009). Equifinality of formal (dream) and informal (glue) bayesian approaches in hydrologic modeling? *Stochastic environmental research and risk assessment*, *23*(7), 1011–1026. doi: 10.1007/s00477-008-0274-y