multi-Risk sciEnce for resilienT commUnities undeR a changiNgclimate
Codice progetto MUR: **PE00000005 – I33C22006910006**

Deliverable title: A method for converting deterministic projections into a stochastic prediction.

Deliverable ID: 8.5.3

Due date: 31 March 2026

Submission date: 25 March 2026

AUTHORS
Alberto Montanari (UNIBO) and part of the RETURN Uncertainty Task Force

1. Technical references

Project Acronym	RETURN
Project Title	multi-Risk sciEnce for resilienT commUnities undeR a changiNg climate
Project Coordinator	Andrea Prota UNIVERSITA DEGLI STUDI DI NAPOLI FEDERICO II domcalca@unina.it
Project Duration	December 2022 – November 2025 (36 months)

Deliverable No.	DV8.5.3
Dissemination level*	CO
Work Package	WP5 - Uncertainty assessment for climate and weather scenarios
Task	T 8.5.2 - Definition of a statistical method to assess uncertainty of hazard indicators for the historical period
Lead beneficiary	UNIBO
Contributing beneficiary/ies	UNIPD

* PU = Public

PP = Restricted to other programme participants (including the Commission Services)

RE = Restricted to a group specified by the consortium (including the Commission Services)

CO = Confidential, only for members of the consortium (including the Commission Services)

Document history

Version	Date	Lead contributor	Description
0.1	10.03.2026	Alberto Montanari (UNIBO)	First draft
0.2	18.03.2026	Alberto Montanari (UNIBO) and Uncertainty Task Force	Critical review and proofreading

2. Abstract

Based on the activities carried out in Task 8.5.2, this deliverable introduces a methodology for estimating uncertainty of environmental prediction, by converting a deterministic prediction model into a stochastic model.

The present deliverables includes two papers published in scientific journals describing the procedure and presenting the software for the application, that is available in the public domain.

The methodology is an innovative product of the RETURN project, allowing a seamless estimation of the uncertainty of predictions.

Table of contents

3. Introduction

The RETURN methodology presented here is identified with the acronym Bluecat, which stands for “Brisk local uncertainty estimator for generic simulations and predictions”.

In essence, Bluecat is a method to transform a deterministic prediction model into a stochastic prediction model, therefore turning from a point prediction to the probability distribution of the predictand. From the latter distribution, a mean (or median) prediction is obtained along with the confidence bands. In September 2024, Bluecat was extended to allow uncertainty assessment of predictions obtained with multiple alternative models (ensemble prediction), by using minimum uncertainty as a criterion for model selection.

Therefore Bluecat allows the user to performs three tasks:

- updating the deterministic prediction, therefore providing a new point prediction;
- computing confidence limits for the prediction for an assigned confidence level;
- combining predictions obtained with several alternative models by using a criterion of minimum uncertainty.

Therefore, Bluecat is not only an uncertainty assessment method: it is rather a new prediction (multi)model with a stochastic structure, and eventually a physical basis that is rooted into the deterministic model. Bluecat can be applied in conjunction with any deterministic prediction model. To provide an example of application, we refer here to rainfall-runoff modeling .

The next section of the deliverable presents a copy of a paper published to a scientific journal describing Bluecat and its application to the Arno and Sieve rivers. The final section presents a second paper providing an extension of Bluecat to multimodel deterministic predictions and presenting the software available in the open domain for the reproduction of the results and the application of Bluecat to any case study.

4. Bluecat: A Local Uncertainty Estimator for Deterministic Simulations and Predictions

AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2021WR031215

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 performances of the method for operational applications
- We provide an open source computer code to apply the method and perform diagnostic checking

Correspondence to:
A. Montanari,
alberto.montanari@unibo.it

Citation:
Koutsoyiannis, D., & Montanari, A. (2022). Bluecat: A local uncertainty estimator for deterministic simulations and predictions. *Water Resources Research*, *58*, e2021WR031215. https://doi.org/10.1029/2021WR031215

Received 11 SEP 2021
Accepted 31 DEC 2021

Author Contributions:
Conceptualization: D. Koutsoyiannis, A. Montanari
Data curation: D. Koutsoyiannis, A. Montanari
Formal analysis: D. Koutsoyiannis, A. Montanari
Investigation: D. Koutsoyiannis, A. Montanari
Methodology: D. Koutsoyiannis, A. Montanari
Resources: D. Koutsoyiannis, A. Montanari
Software: D. Koutsoyiannis, A. Montanari
Supervision: D. Koutsoyiannis, A. Montanari
Validation: D. Koutsoyiannis, A. Montanari
Visualization: D. Koutsoyiannis, A. Montanari

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

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.

Writing – original draft: D. Koutsoyiannis, A. Montanari
Writing – review & editing: D. Koutsoyiannis, A. Montanari

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 neighbor approach to represent the probability distribution of the model error which makes the method flexible and fast. Similar approaches were applied by Papacharalampous et al. (2019), Papacharalampous et al. (2020), Papacharalampous, et al. (2019b), Tyralis, Papacharalampous, and Langousis (2019), Tyralis, Papacharalampous, Burnetas, and Langousis (2019) and Papacharalampous, 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

10.1029/2021WR031215

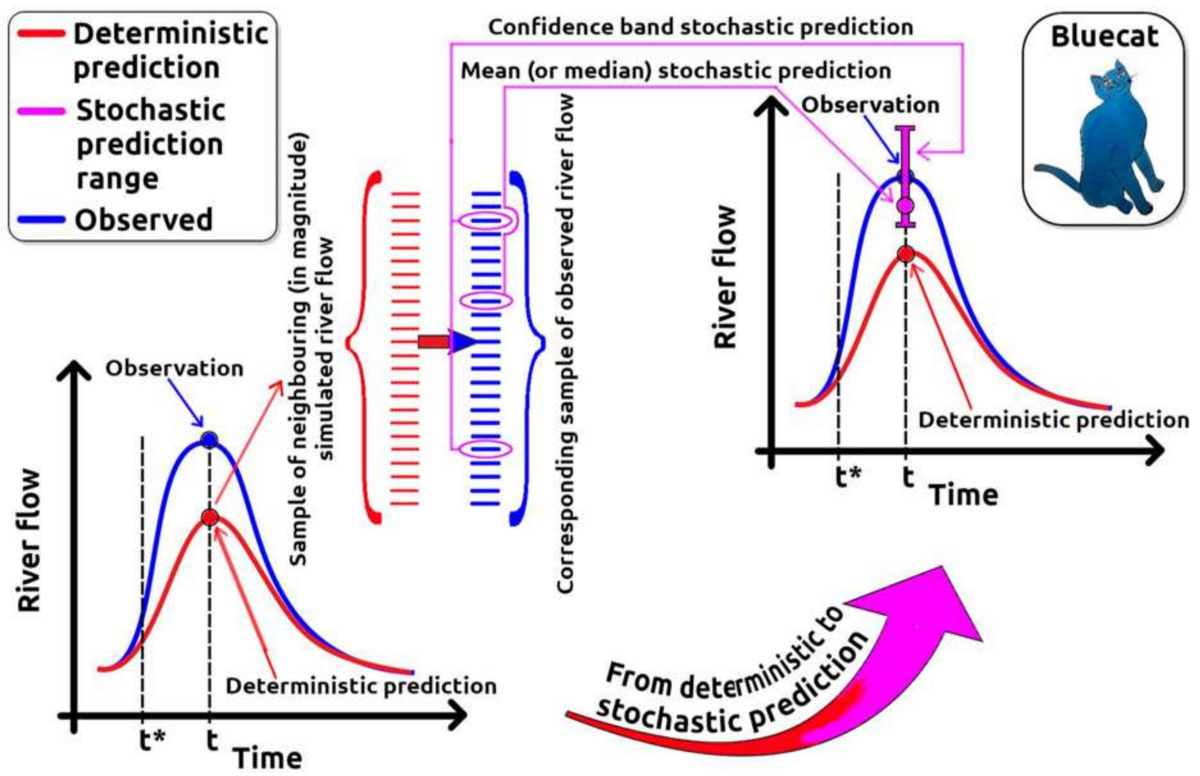

Figure 1. Schematic representation of the Bluecat concept underlying the transformation of the deterministic model (D-model) to a stochastic model (S-model). The painting in the upper right corner is cropped from the picture available at https://www.flickr.com/photos/cizauskas/36142084534/ of the Andy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

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** simula**t**ions 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 modeled 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, $i = 1, \ldots, m_1 + m_2 + 1$ 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 performs 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

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

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 five 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, considering the uncertainty in model parameters and input variables. This work has been discussed (Montanari & Koutsoyiannis, 2014b; Nearing, 2014) and advanced in other studies (Papacharalampous et al., 2019; Quilty & Adamowski, 2020; Sikorska et al., 2015). 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 Equation 1 is an estimate thereof. By considering $\mathbf{x}_\tau$ in Equation 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:

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

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

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

Equation 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

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

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 Equation 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 Equation 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

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

While a parametric approach like the above is attractive from many aspects, here we propose a fully data based approach, that is, 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 $\bar{q}_i, i = 1, \ldots, (m_1 + m_2 + 1)$, of Q-neighbors based on:

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

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 $\bar{q}_i$. 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, $m_1 = m_2 = m = 20$, that is, $m_1 + m_2 + 1 = 41$, 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 $\bar{q}_i$ 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 $\bar{q}_i$. 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; Koutsoyiannis, 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)

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

with $p \ge q$ 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

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

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

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

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

$$\Lambda_\infty = \begin{cases} \Gamma(1-\xi)^{\frac{1}{\xi}}, & \xi \neq 0 \\ e^{\gamma}, & \xi = 0 \end{cases} \tag{23}$$

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, substituting in Equation 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 Equation 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 PBF 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 $F_w(w) = P\{\underline{w} \le w\} = P\left\{F_Q(\underline{q}) \le w\right\} = P\left\{\underline{q} \le F_Q^{-1}(w)\right\}$ 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, that is, 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" 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, that is, 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 $\tau = 1, 2, ...$, the sequence of variables $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_\tau$, whose values are:

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

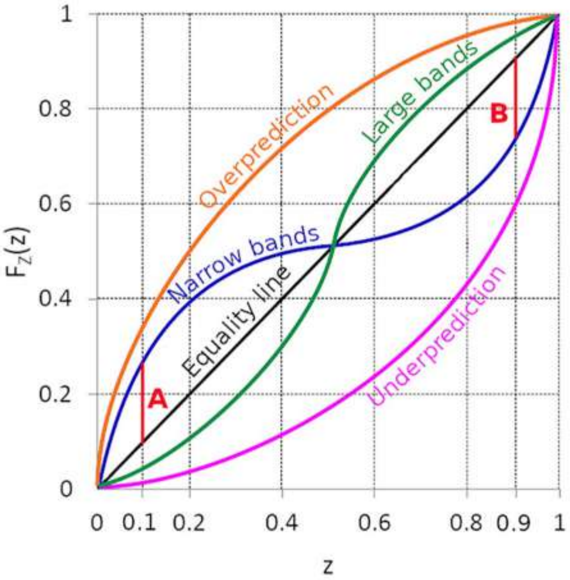

Figure 2. Information conveyed by the predictive probability-probability plot.

are independent and identically distributed with uniform distribution in [0, 1]. Note that here we used the vector notation $\mathbf{x}_{\tau-1} := [x_{\tau-1}, \ldots, x_1]^T$ to represent all values of the process earlier than τ. One may see an analogy of $\underline{z}_Q$ defined in Equation 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 Equation 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 γ, $0 \leq \gamma \leq 1$, so that, for any Q, $F_{q|Q}(q) = \gamma$ then the latter relationship specifies a confidence curve for q, which is a function $q = h(Q)$, given that γ is constant. The probability

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

is constant, independent of Q. Moreover, given the definition of z and its property not to depend on Q, one obtains $z = \gamma$. If the distribution of z is uniform in 0,1, that is, $F_z(z) = z$, the value of $F_z(z)$ at the point $z = \gamma$ will be equal to γ. Therefore any deviation from uniformity is a relative measure of the number of observations exceeding the value γ 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 Equation 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.

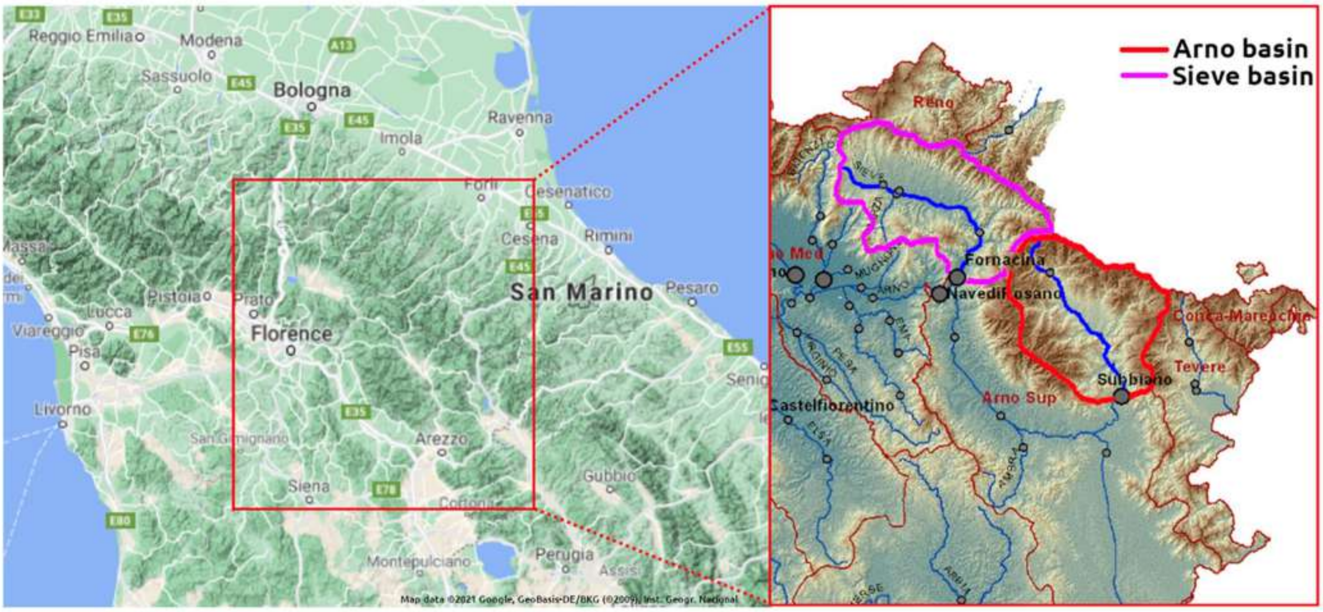

Figure 3. Basins of the Arno river at Subbiano and the Sieve river at Fornacina.

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 (Barber et al., 2020; Clark et al., 2021). 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):

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

where x and y are original and transformed data, respectively, and λ is a parameter. For $\lambda \to 0$ and $\lambda \to \infty$ Equation 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 overparameterization, 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 Equation 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.

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

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 1,657 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 Equation 33 with $\lambda = 0.000$ 1, 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 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 Figures 5 and 6 and Table 2. 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 summarized by a Pearson correlation coefficient of 0.80 and a Nash efficiency of 0.57. Slightly 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 Equation 4 to compute the S-model prediction. It also suggests an overestimation and underestimation of low and high flows, respectively.

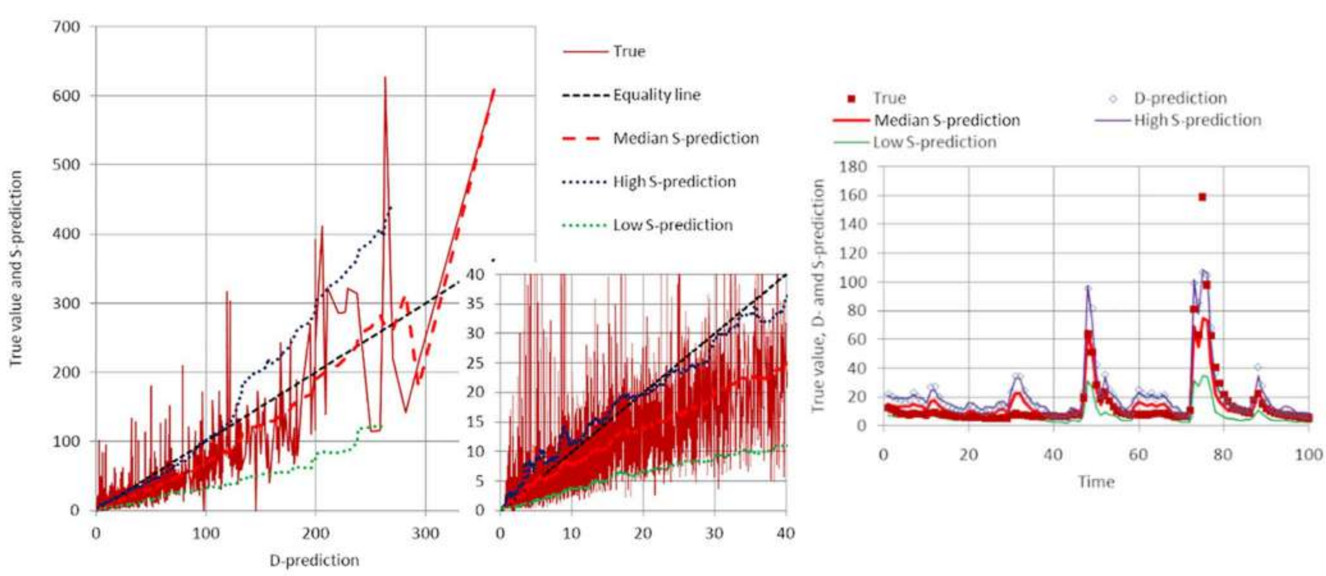

Figure 4. D-model and S-model predictions, along with confidence limits, for the calibration period of the Arno river at Subbiano. The right panel depicts 100 days of the calibration period, where the first day is January 1st, 2011.

Table 2
Percentage of Observations Lying Outside the 80% Confidence Limits for the Considered Case Studies

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%

Note. Band was estimated with both order statistics and robust estimation. Subscripts *h* and *l* refer to upper and lower limit, 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 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 1,637 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 hr of the calibration period starting from September 16th, 1992 at 5 a.m.

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

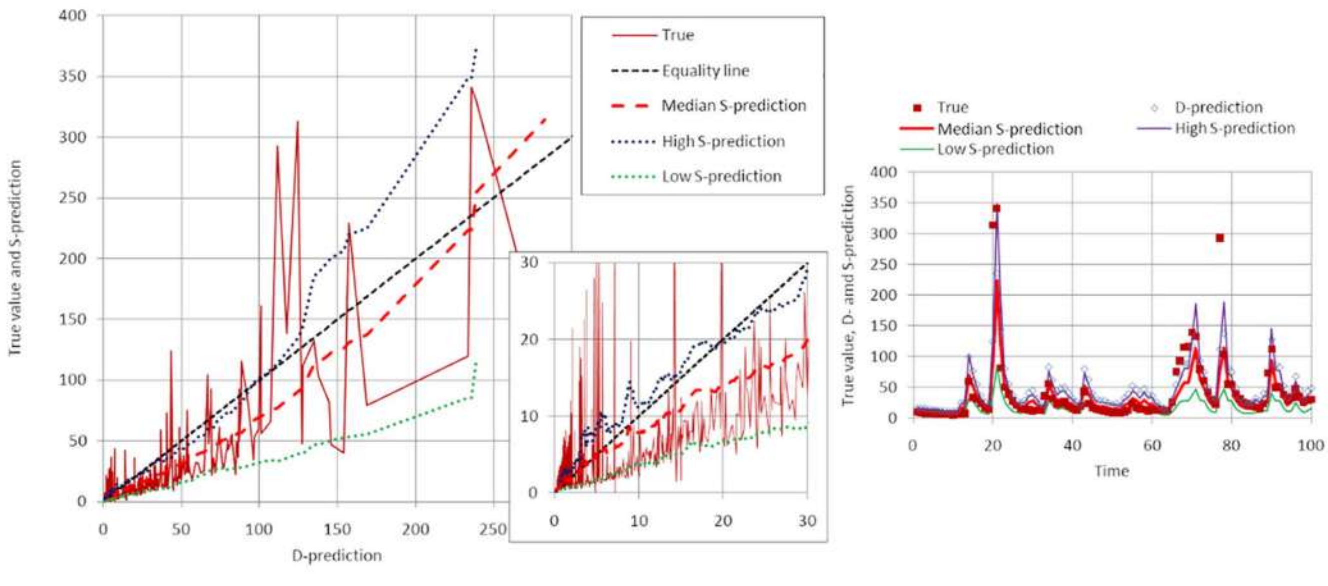

Figure 5. D-model and S-model predictions, along with confidence limits, for the validation period of the Arno river at Subbiano. The right panel depicts 100 days of the validation period, where the first day is January 1st, 2013.

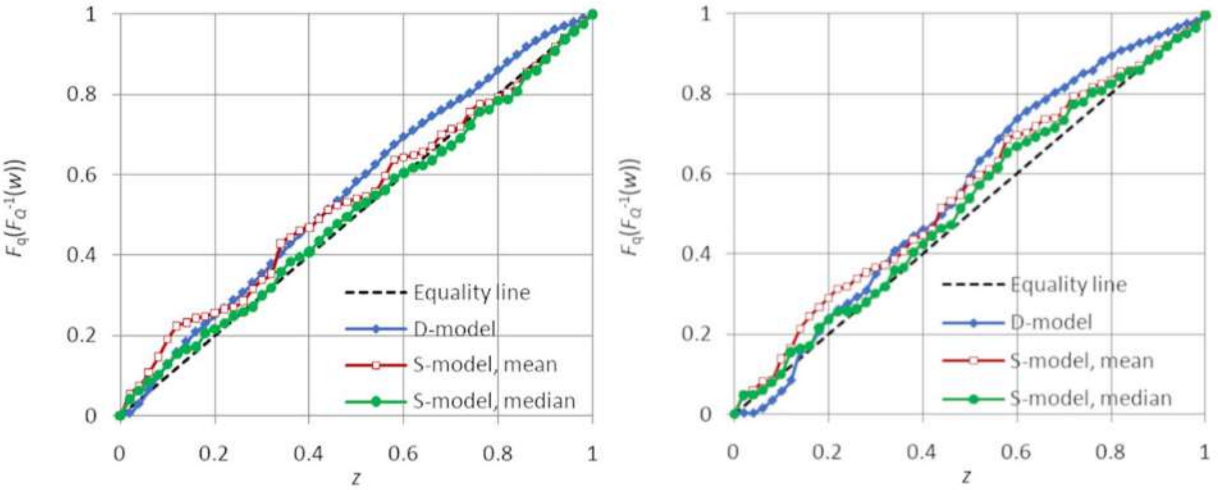

Figure 6. Combined probability-probability plots for the predictions of the river flows of the Arno river at Subbiano in calibration (left) and validation (right).

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.

Validation results are shown in Figure 8, where the right panel depicts 150 hr of the validation period starting from January 5th, 1996 at 12 a.m., 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.

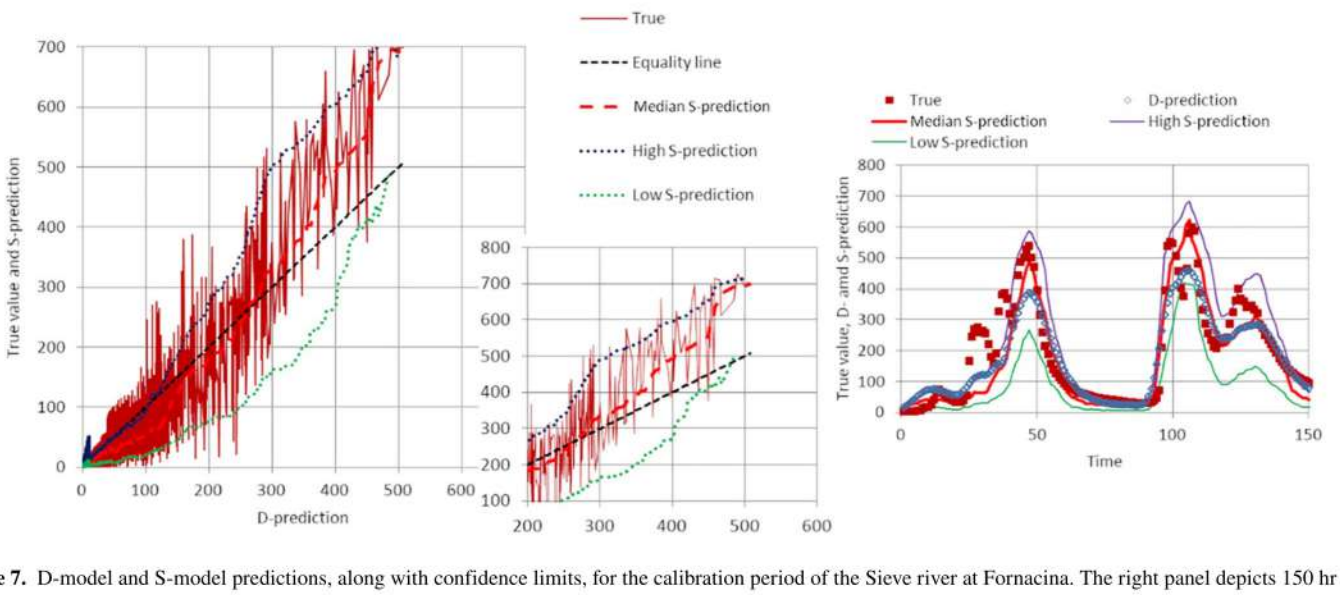

Figure 7. D-model and S-model predictions, along with confidence limits, for the calibration period of the Sieve river at Fornacina. The right panel depicts 150 hr of the calibration period starting from September 16th, 1992 at 5 a.m.

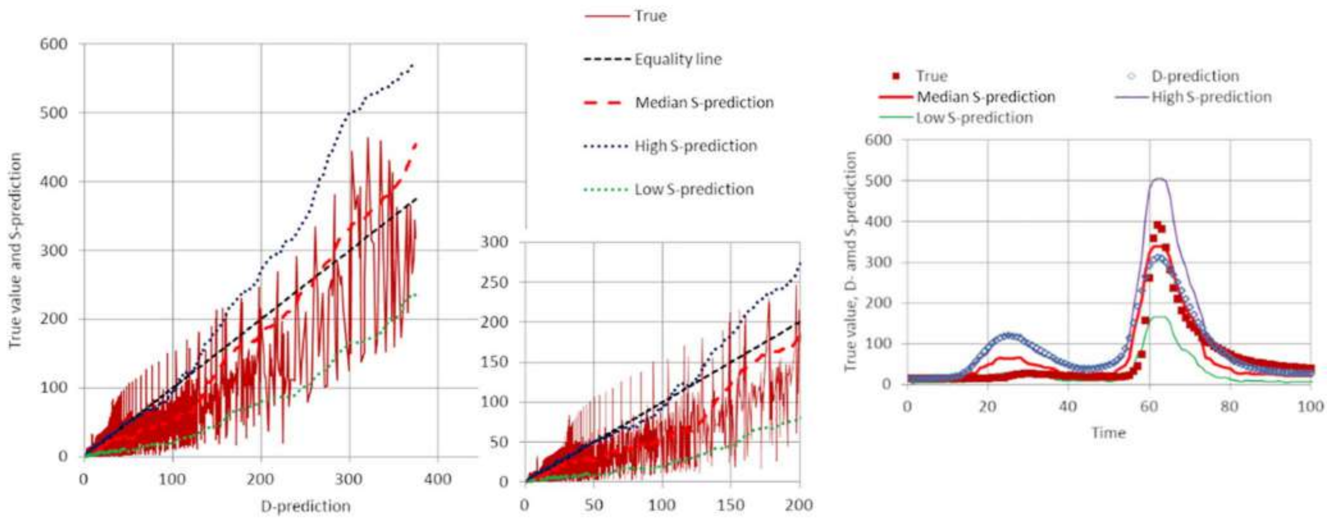

Figure 8. D-model and S-model predictions, along with confidence limits, for the validation period of the Sieve river at Fornacina. The right panel depicts 150 hr of the validation period starting from January 5th, 1996.

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 Equation 4 also holds asymptotically.

When the calibration data set is not extended enough one may experience overparameterization, which implies that the calibrated model exhibits satisfactory performances 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 obtained 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.

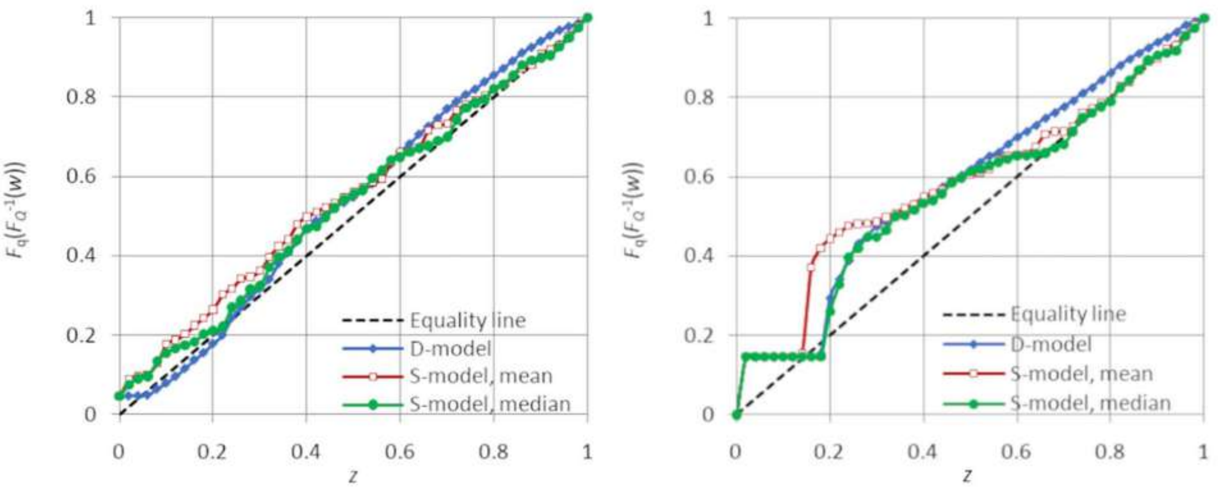

Figure 9. Combined probability-probability plots for the predictions of the river flows of the Sieve river at Fornacina in calibration (left) and validation (right).

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 Equation 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; Montanari and Koutsoyiannis, 2014a). We also note that the conditional distribution given by Equation 4 might be seasonal, although part of the seasonality features are already incorporated in the D-model (e.g., 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 Equation 4. If one would like to consider a non-stationary uncertainty, then a parametric and non-stationary distribution (perhaps a PBF 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 modeling 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.

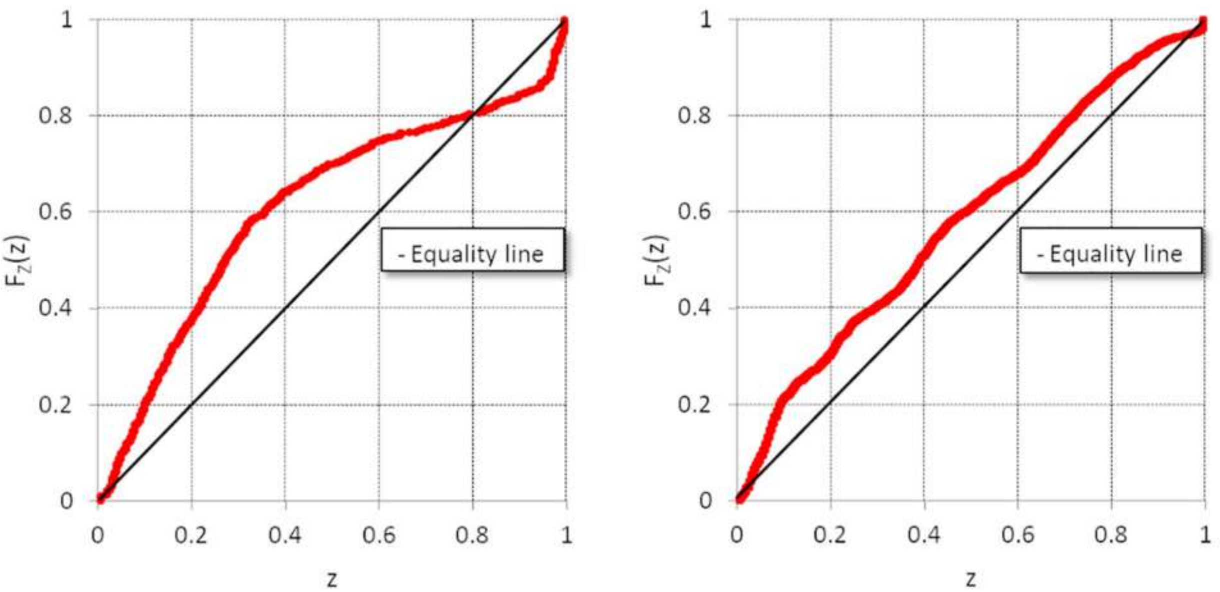

Figure 10. Predictive probability-probability plots for the validation of the river flows predictions for the Arno river (left) and the Sieve river (right).

One may wonder what the distinguishing behavior of Bluecat is 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 neighboring 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.

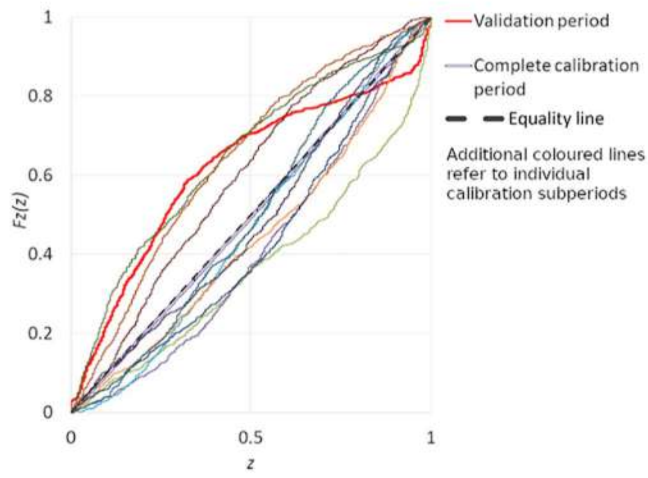

Figure 11. Sampling variability for the Predictive probability-probability (PPP) plot of the Arno river in calibration and comparison with the PPP plot in validation.

Although Bluecat has been conceived to be applied to one single model, a multimodel 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-environment (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 Sections 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.

Data Availability Statement

The software and data that have been used to develop this work are included in a package working under the R environment, that is open source and available for download at: https://github.com/albertomontanari/hymodbluecat.

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. 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. DK is grateful to the University of Bologna and its Institute of Advanced Studies for hosting him as visitor in fall 2019 and providing the opportunity for this collaboration. Open Access Funding provided by Universita degli Studi di Bologna within the CRUI-CARE Agreement.

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. https://doi.org/10.1080/02626667.2019.1686639

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1007/978-3-319-70766-2_6

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

Bulygina, N., & Gupta, H. (2009). Estimating the uncertain mathematical structure of a water balance model via bayesian data assimilation. *Water Resources Research*, *45*(12), W00B13. https://doi.org/10.1029/2007wr006749

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J., Tang, G., et al. (2021). The abuse of popular performance metrics in hydrologic modeling. *Water Resources Research*, *57*, e2020WR029001. https://doi.org/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. https://doi.org/10.1016/j.advwatres.2019.05.003

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*. https://doi.org/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. https://doi.org/10.1002/2017wr020528

Koutsoyiannis, D. (2019). Knowable moments for high-order stochastic characterization and modelling of hydrological processes. *Hydrological Sciences Journal*, *64*(1), 19–33. https://doi.org/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. https://doi.org/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* (pp. 10125). https://doi.org/10.5194/egusphere-egu2020-10125

Krzysztofowicz, R. (1999). Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research*, *35*(9), 2739–2750. https://doi.org/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. https://doi.org/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. https://doi.org/10.1029/2020wr027101

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 united s tates. *Water Resources Research*, *53*(7), 5469–5494. https://doi.org/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). https://doi.org/10.1029/2004wr003826

Montanari, A. (2011). Uncertainty of hydrological predictions. In P. Wilderer (Ed.), *Treatise on water science* (Vol. 2, pp. 459–478). Elsevier. https://doi.org/10.1016/b978-0-444-53199-5.00045-2

Montanari, A., & Brath, A. (2004). A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research*, *40*(1), W01106. https://doi.org/10.1029/2003WR002540

Montanari, A., & Grossi, G. (2008). Estimating the uncertainty of hydrological forecasts: A statistical approach. *Water Resources Research*, *44*(12), W00B08. https://doi.org/10.1029/2008WR006897

Montanari, A., & Koutsoyiannis, D. (2012). A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research*, *48*(9), W09555. https://doi.org/10.1029/2011WR011412

Montanari, A., & Koutsoyiannis, D. (2014a). Modeling and mitigating natural hazards: Stationarity is immortal. *Water Resources Research*, *50*(12), 9748–9756. https://doi.org/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 Resources Research*, *50*, 6264–6268. https://doi.org/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. https://doi.org/10.1002/2013WR014812

Pagano, T. C., Wood, A. W., Ramos, M.-H., Cloke, H. L., Pappenberger, F., Clark, M. P., et al. (2014). Challenges of operational river forecasting. *Journal of Hydrometeorology*, *15*(4), 1692–1707. https://doi.org/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. https://doi.org/10.1007/s00477-018-1638-6

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. https://doi.org/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). https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/j.envsoft.2020.104718

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. https://doi.org/10.5194/hess-17-2219-2013,2013

R Core Team. (2013). *R: A language and environment for statistical computing [computer software manual]*. Retrieved from http://www.R-project.org/

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. https://doi.org/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), W10531. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1007/s00477-008-0274-y

5. Uncertainty estimation for environmental multimodel predictions: The BLUECAT approach and software

Environmental Modelling and Software 188 (2025) 106419

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft

Position Paper

Uncertainty estimation for environmental multimodel predictions: The BLUECAT approach and software

Alberto Montanari [a,*], Demetris Koutsoyiannis [b]

[a] Deparment DICAM, University of Bologna, Via del Risorgimento 2, Bologna, 40136, Italy
[b] National Technical University of Athens, Heroon Polytechneiou 9, Zographou, Athens, 15772, Greece

ARTICLE INFO

Keywords:
BLUECAT
Uncertainty
Prediction confidence band
Multimodel prediction
Stochastics

ABSTRACT

An extension of the BLUECAT approach and software for uncertainty assessment of environmental predictions is presented, allowing the application to multimodel outputs. BLUECAT operates by transforming a point prediction provided by deterministic models to a corresponding stochastic formulation, thereby allowing the estimation of a bias corrected expected value along with confidence limits. In this paper we also propose to use BLUECAT for model selection in the context of multimodel predictions, by using a measure of uncertainty as selection criterion. We emphasise here the value of BLUECAT for gaining an improved understanding of the underlying environmental systems and multimodel combination. Two examples of applications are presented, highlighting the benefits attainable through uncertainty driven integration of several prediction models. These case studies can be reproduced through the BLUECAT software, that is available in the public domain along with help facilities and instructions.

1. Software and data availability

- Name of software: Bluecat-R and Bluecat-Python (R and Python versions, respectively)
- Developers: Alberto Montanari and Demetris Koutsoyiannis
- Contact: alberto.montanari@unibo.it
- Date first available: August 8, 2024
- Software required: R statistical environment, Python3 environment
- Program language: R and Python
- Source code at: https://github.com/albertomontanari/Bluecat-R and https://github.com/albertomontanari/Bluecat-Python (R and Python versions, respectively)
- Documentation: Detailed documentation for application installation, testing, and deployment can be found at https://github.com/albertomontanari/Bluecat-R/blob/main/README.md and https://github.com/albertomontanari/Bluecat-Python/blob/main/README.md (R and Python versions, respectively). Further information is provided by the R help (included in the R version)
- Data required for reproducing the case studies presented in the paper are included in the repository of the source codes as application examples

2. Introduction

Uncertainty means lack of deterministic predictability (Anderson et al., 2001). It is the real reason why managing environmental issues and emergencies has continuously been an essential and difficult task for humans during their history and evolution (Hughes, 2016). Uncertainty is due to the complexity, chaotic behaviours and our limited understanding of several involved processes (Dewulf and Biesbroek, 2018). Understanding uncertainty is the key to gain a better comprehension of the involved environmental systems.

In fact, uncertainty of predictions is today recognised as an essential information for elaborating reliable environmental risk mitigation and adaptation strategies (White et al., 2021; Sheikholeslami et al., 2024). Indeed, humans are used to take decisions under uncertainty in everyday life. However, we also recognise the value of a rigorous and quantitative approach to uncertainty estimation and communication, in particular when the risk associated to the decision becomes relevant (Vose, 2008).

Uncertainty assessment in environmental modelling has been long investigated and discussed. See, for instance, Koutsoyiannis (2023), Beven (2018), Refsgaard et al. (2007), Burke et al. (2015), Kim et al. (2024), Hughes and Lawrence (2024), Liang et al. (2024), Lin et al. (2024), Plunge et al. (2024) and Auer et al. (2024), to cite only a few.

* Corresponding author.
E-mail address: alberto.montanari@unibo.it (A. Montanari).

https://doi.org/10.1016/j.envsoft.2025.106419
Received 24 September 2024; Received in revised form 27 February 2025; Accepted 3 March 2025
Available online 12 March 2025
1364-8152/

The problem is multifaceted, for the diversity of applications, contexts and available information.

Here, we focus on the general case where environmental variables are predicted by using one or more calibrated models (multimodel) that produce one or more point estimations for which uncertainty assessment is needed. In most of those cases, models are deterministic and process based or data-driven (see, for example, the recent applications by Gomes Jr. et al. (2024), Imhoff et al. (2024), Zou et al. (2024) and Jonsson et al. (2024)), but uncertainty assessment may also be required for the statistics or parameters of stochastic models (see, for instance, Cappelli et al. (2024)). We also refer to the case where a sufficiently long record of past outputs from each considered model is available that can be compared with the corresponding true values, that are typically derived from observations. Under such circumstances, uncertainty of model predictions can be assessed by comparing the predictions themselves with the corresponding reality. Several past studies (see, e.g., Beven (2016)) have demonstrated that drawing conclusions basing on such comparison is not an easy task. Prediction errors show a diversity of statistical behaviours, arising from several sources of uncertainty depending on the state of the considered system and therefore change in time and space.

Accordingly, a variety of approaches to uncertainty assessment have been proposed by the literature, including (1) data analysis methods, comprising analytical and statistical procedures for evaluating the accuracy of data, (2) derived distribution methods to compute the probability distribution function of the model output, (3) simulation and sampling-based methods, estimating the full distribution of the model output via simulation with different models and/or parameters. The category of the data analysis methods includes, among the others, statistical approaches (Honti et al., 2013), artificial intelligence (Kabir et al., 2018) and in particular machine learning (Shrestha and Solomatine, 2008). Simulation and sampling methods include multimodel approaches that are widely applied in environmental sciences (Herrmann and Marzocchi, 2023; Slater et al., 2019). In general, methods for assessing uncertainty are formulated for a single model, but can be converted to the case of multimodel prediction.

Several data analysis methods are based on the analysis of model prediction errors, which in most of the cases is performed by using statistical procedures (see, for instance, Montanari and Brath (2004), Montanari and Grossi (2008), Montanari and Koutsoyiannis (2012), Sikorska et al. (2015) and Liang et al. (2024)). Several contributions have pointed out that these methods are based on assumptions, like independence and homoscedasticity of model errors, which may be not satisfied and thus result in wrong uncertainty estimates (Beven, 2019). Therefore, the use of approaches that extract information directly from data rather than their statistics may be preferable.

Building on the above considerations, Koutsoyiannis and Montanari (2022a) proposed the BLUECAT approach, a simple, easy-to-use and transparent methodology to upgrade a deterministic model into a stochastic one, thereby producing an estimate of the probability distribution of the true value to be predicted. Therefore, BLUECAT first upgrades the deterministic prediction into the stochastic expected value, by essentially operating a bias correction, and then produces an estimate of the confidence band for the considered variable. A software working in R-environment for the application of BLUECAT to predictions given by the HyMod rainfall-runoff model (Boyle, 2000) is available at https://github.com/albertomontanari/hymodbluecat. The method has been applied to a number of case studie in the realm of hydrology (Jorquera and Pizarro, 2023; Rozos et al., 2022; Koutsoyiannis and Montanari, 2022b).

Here, we present an updated and more general version of the BLUECAT approach and software, to allow the application to any environmental prediction obtained with a single model or a set of models. In fact, multimodels are increasingly used in environmental modelling to investigate the possible range of environmental predictions and simulations (see the recent contributions by Mangukiya et al. (2024), Wang et al. (2024) and Tu et al. (2024)). Thus, estimating their uncertainty is emerging as a key issue in environmental modelling that motivates the effort to remove any assumption on the nature and number of predictive models (Giustolisi et al., 2007). The present work discusses the whole set of hypotheses conditioning the application of BLUECAT to multimodels as well as an extended set of procedures for testing the validity of the estimated uncertainty measures. The updated BLUECAT software is provided in two languages - R (R Core Team, 2021) and Python3 (Van Rossum and Drake, 2009).

3. The BLUECAT approach

We discuss here BLUECAT by referring first to the case of a single model prediction as in Koutsoyiannis and Montanari (2022a). We will discuss application to multimodel prediction in Section 6.

Let us denote with the symbol Y_τ the output from a generic deterministic environmental model at discrete prediction step τ, with $Y_\tau \in \Re$. Step τ indicates any allocation index of the individual model output into a set of predictions. We take for given that the true value of the predicted variable is available, which we denote with the symbol y_τ. The first assumption of BLUECAT is that the information contained in the available samples of Y_τ and y_τ is sufficient to support the transformation from the deterministic to the stochastic output, therefore allowing to estimate uncertainty of the output itself.

The above first assumption does not imply severe limitations in practical applications. Indeed, most environmental models are calibrated and/or can produce hindcasts of the relevant variables. In both cases, a record of predictions along with corresponding observations is produced, so that uncertainty can be assessed by comparing the model output with the corresponding true value.

The target of BLUECAT is to efficiently extract such information in order to produce a reliable estimate of uncertainty, with the simplest approach possible, by avoiding sophisticated assumptions. In what follows, we underline stochastic entities (variables, processes and functions). Variable values, deterministic functions and realisation of stochastic processes are indicated with non-underlined symbols. Stochastic processes correspond to the real processes, while the outcome of the deterministic model (D-model) is an estimate thereof.

To update the deterministic prediction Y_τ to its stochastic form (S-model), we need to specify the conditional probability distribution:

$$F_{\underline{y}|\underline{Y}}(y|Y) = P\{\underline{y} \leq y | \underline{Y} = Y\} \tag{1}$$

where y and Y correspond to the same discrete step τ and P indicates probability. Let us note that Y is a scalar, i.e., a model output for a single prediction step.

Koutsoyiannis and Montanari (2022a) suggested a fully data based approach to estimate the conditional distribution $F_{\underline{y}|\underline{Y}}(y|Y)$. First, a sample $\bar{y}_i$, $i = 1, \ldots, m_l + m_u + 1$ of true values is assembled that correspond to the sample $\bar{Y}_i$ of the D-model outputs that are closest in value to Y, according to:

$$F_{\underline{y}|\underline{Y}}(y|Y) \approx P\{\underline{y} \leq y | Y - \Delta Y_1 \leq \underline{Y} \leq Y + \Delta Y_2\} \tag{2}$$

where ΔY_1 and ΔY_2 are chosen to include a number of lower and upper neighbours to Y equal to $m_l := \Delta F_1 n$ and $m_u := \Delta F_2 n$, respectively; n is the sample size of the available y and Y values. Numbers m_l and m_u should not be too large, in order to ensure that $F_Y(Y) \pm \Delta F_{1,2}$ is close to $F_Y(Y)$, nor too small, to ensure that the probability

$$P\left\{\underline{y} \leq y | F_Y(Y) - m_l/n \leq F_Y(\underline{Y}) \leq F_Y(Y) + m_u/n)\right\} \tag{3}$$

can be estimated from the sample $\bar{y}_i$. Note that it may not be possible to collect the desired sample size of model output for the extreme values of the prediction, for which enough lower or higher model outputs may not be available, so that the numbers m_l and m_u should be ad-hoc reduced. This solution is adopted in the BLUECAT software. Here, we adopt $m_l = m_u = m$ and therefore the resulting sample size of $\bar{y}_i$ is $2m+1$.

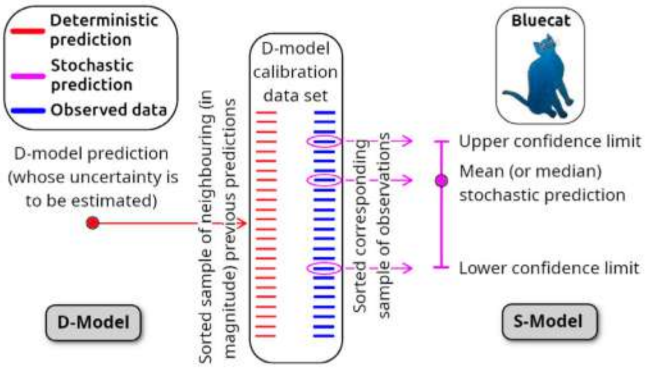

Fig. 1. Schematic of the BLUECAT workflow and software. The deterministic model (D-Model) is transformed to the stochastic model (S-model) by a stochastic analysis of the D-model predicted data in calibration versus the corresponding observations. The painting in the upper right is cropped from the picture 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).

From the probability distribution given by (3) one can easily estimate the mean value (alternatively the median which may be more robust against outliers) which gives the S-model prediction, and quantiles corresponding to assigned probabilities, which may be used to define the confidence band of the S-model prediction for given confidence level.

In the BLUECAT software we estimate quantiles through order statistics or, in alternative, a robust approach based on the concept of knowable moments (K-moments, see Koutsoyiannis (2019, 2023)). The approach is presented with full details in Koutsoyiannis and Montanari (2022a), to which the interested reader is referred to. Note that for both order statistics and K-moments the minimum and maximum quantiles that can be estimated are $\min(\tilde{y}_i)$ and $\max(\tilde{y}_i)$, respectively. If one needs to extrapolate quantiles for arbitrary probabilities, a parametric relationship for $F_{y|Y}(y|Y)$ should be adopted. For instance, Koutsoyiannis and Montanari (2022a) fitted a local linear regression between Y and y to extrapolate quantiles beyond the lowest and highest values in $\tilde{y}_i$. In the BLUECAT software we use the approximation $F^{-1}_{y|Y}(y|Y) = \min(\tilde{y}_i)$, $\forall Y \leq \min(\tilde{Y}_i)$ and $F^{-1}_{y|Y}(y|Y) = \max(\tilde{y}_i)$, $\forall Y \geq \max(\tilde{Y}_i)$, where $\tilde{Y}_i$ is the sample of predicted data corresponding to $\tilde{y}_i$.

A schematic of the BLUECAT workflow is presented in Fig. 1.

4. BLUECAT testing

The BLUECAT software includes procedures for testing the reliability of the estimated confidence bands against observed values of the variable to be predicted. Let us point out that the true values y_τ should necessarily fall with probability $1-\alpha$ within the confidence bands estimated for significance level α. It follows that a first opportunity to check the BLUECAT output is simply to count the percentage of true values lying within (or outside) the confidence band. This check is automatically performed by the BLUECAT software.

Moreover, the above percentage should necessarily be independent of the value of y_τ, namely, the number of true values within the confidence bands should not change for different values of y_τ. Koutsoyiannis and Montanari (2022a) propose two graphical methods to check the reliability of the BLUECAT output at local scale along the whole range of predicted variables: the "Combined Probability-Probability" (CPP) plot and the "Predictive Probability-Probability" (PPP) plot, which are drawn by the BLUECAT software provided the testing flag is set to "true" (see Section 7).

The CPP plot essentially compares the probability distributions of a set of observed and predicted data. It is described in detail by Koutsoyiannis and Montanari (2022a) to which the interested reader is referred to.

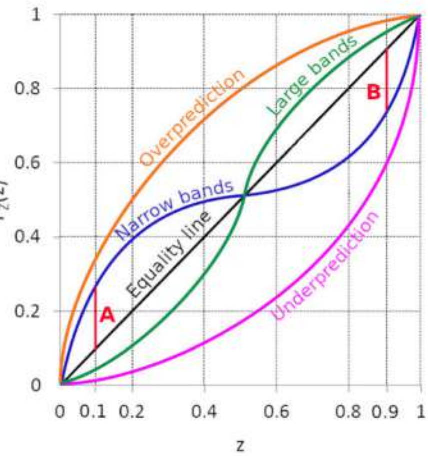

Fig. 2. Predictive probability-probability plot. Segments A and B provide a measure of the excess percentage of observations lying below the lower and above the upper confidence limit, respectively, for significance level of 20%..
Source: Reproduced from Koutsoyiannis and Montanari (2022a).

4.1. Predictive Probability-Probability plot (PPP)

We discuss here the PPP plot to further support, with additional considerations, its use for testing the reliability of the uncertainty assessment provided by BLUECAT. The PPP plot was first introduced by Laio and Tamea (2007) and then discussed by several authors, including Eslamian (2014). The plot was referred to with different terms in previous studies. PPP is a plot of the empirical distribution function $F_z(z)$ of a stochastic variable $\underline{z}$, where the latter also is a conditional non-exceedance probability, namely

$$\underline{z}_Y := F_{y|Y}(\underline{y}). \quad (4)$$

One notes that $\underline{z}$ is the distribution function of the observed values evaluated for any prediction. Such probabilities are regarded as independent and identically distributed with uniform distribution in [0,1]. To check such condition, for each prediction Y_τ we look at the corresponding sample $\tilde{y}_i$ and compute the sample frequency of the value y_τ, that is evaluated by using the Weibull plotting position in the BLUECAT software. To check whether $\underline{z}_Y$ is uniformly distributed the PPP plot displays its values against the corresponding sample frequency. If the plot lies over the identity line then we can conclude that the confidence band is reliably estimated for any value of Y_τ.

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 band is narrow (large). Fig. 2 provides a graphical overview of the above features. Furthermore, the departure of the PPP plot from the equality line provides a measure of the reliability of the confidence band for a given confidence level. For instance, for a confidence level of 0.8 one would expect about 10% of the observed points lying below and 10% lying above a reliably estimated confidence band. Then, if we refer to the blue line in Fig. 2 that is an example of a narrow band, segments A and B provide a measure of the excess percentage of observations lying below the lower and above the upper confidence limit.

5. Summary of the BLUECAT assumptions, limitations, and options

Model building is inevitably based on assumptions. They do not undermine the efficiency and credibility of approaches, but rather allow their application to rigorously defined contexts. More than that, assumptions allow researchers to gain an improved understanding of natural processes. They are unavoidably needed to set up and test models, but they need to be discussed transparently and, when possible, checked through rigorous testing. For these reasons, we believe it is appropriate the summarise in the following 4 items the main assumptions of BLUECAT:

1. The statistical behaviours of the stochastic processes describing the modelled variables do not change in the application phase with respect to calibration. This assumption can be relaxed by using D-models accounting for changes, for instance non-stationary models (Koutsoyiannis and Montanari, 2022a,b).
2. The calibration data set is extended enough to ensure that sufficient information is available to upgrade the D-model into the S-model.
3. The difference between the model output and the corresponding true value quantifies in an aggregated form all types of uncertainty, including uncertainty due to input data and parameters, model structure and so forth.
4. The information needed to assess predictive uncertainty at each prediction step is synthesised by the value of the model prediction.

The implications of the above assumptions determine the limitations of the approach. Regarding the first assumption, in the presence of changes in the stochastics processes the uncertainties estimated in calibration may differ with respect to application. Particular attention should be paid to the sample size of the calibration data set. Uncertainties estimated over a limited amount of information may not be reliable. The information requirements depend on the local application and context, and in particular the statistical behaviours of data and predictive uncertainty. Therefore, the minimal sample size required for reliably assessing uncertainty should be evaluated through expert opinion, case by case. Particular care should be paid when estimating uncertainty for predictions outside the range of calibration data. The information supporting BLUECAT testing should also be carefully evaluated. The software automatically includes warnings when testing is performed against a data set including less than 20 points.

The implication of the last assumption is that BLUECAT provides the same estimate for identical values of model predictions, regardless of other conditions, for instance related to the state of the system, which may impact model reliability. Eventually, the assumption may be removed by conditioning the probability distribution $F_{y|Y}(y|Y)$ at the left hand side of Eq. (1) to additional variables besides Y at the right hand side (for instance, see Koutsoyiannis and Montanari (2022b)). Such potential for further research and extension of BLUECAT is an interesting opportunity to further increase the information content of environmental predictions.

6. BLUECAT application to multimodel prediction

While any uncertainty assessment method for single models can potentially be extended to the multimodel case, actually such extension introduces additional research questions related to (a) how to combine the predictions of different deterministic models and (b) how to estimate uncertainty for the obtained combination.

Question (a) is addressed by a diverse set of approaches in environmental sciences. Examples are ensemble averaging (Marmion et al., 2009; Grenouillet et al., 2011) and Bayesian algorithms (Tebaldi and Knutti, 2007). Unweighted averaging of multimodel predictions is frequently used, thus loosing part of the information conveyed by singular models that may significantly diverge and implying a smoothing effect, that reduces the internal variability of the signal.

In BLUECAT, we propose to use uncertainty of the considered model estimated at each prediction step as a criteria to select the optimal ensemble member. Accordingly, a single model prediction corresponding to the least uncertain ensemble member, that is identified through a proper measure, is picked up at each prediction step, thereby allowing the identification and use of the supposedly best performing model in the specific context and system state.

A key step of the above procedure is the identification of the proper uncertainty measure $U_{\tau,k}$ for S-model k at prediction step τ, which depends on model type and intended use. We tested several different options, that were identified in coherence with the BLUECAT aim of seeking flexibility and simplicity, and finally included in the BLUECAT software the following 6 measures:

$$U_{\tau,k} = |Y_{\tau,k,u} - Y_{\tau,k,l}| \tag{5}$$

$$U_{\tau,k} = |(Y_{\tau,k,u} - Y_{\tau,k,l})/Y_{\tau,k,S}| \tag{6}$$

$$U_{\tau,k} = |Y_{\tau,k,D} - Y_{\tau,k,S}| \tag{7}$$

$$U_{\tau,k} = |(Y_{\tau,k,D} - Y_{\tau,k,S})/Y_{\tau,k,S}| \tag{8}$$

$$U_{\tau,k} = NE_{\tau,k,D} \tag{9}$$

$$U_{\tau,k} = SAE_{\tau,k,D} \tag{10}$$

where $Y_{\tau,k,u} - Y_{\tau,k,l}$ are the upper and lower confidence limits for the prediction $Y_{\tau,k,S}$ by S-model k, $Y_{\tau,k,D}$ is the prediction by D-model k, and $NE_{\tau,k,D}$ and $SAE_{\tau,k,D}$ are the Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970) and sum of absolute errors, respectively, of the prediction by D-model k of the sample of true values $\bar{y}_i$ identified at each prediction step τ (see Section 3).

We found that Eq. (5) may be indicated for applications where the variability of Y is relatively limited. The BLUECAT software reports back which model has been used at each step τ thus allowing to test the different options. Fig. 3 shows a sketch of the uncertainty based step-by-step model selection procedure.

More rigorous methods for estimating $U_{\tau,k}$ may be applied based on the estimate provided by BLUECAT of the conditioned probability distribution $F_{y|Y}(y|Y)$. Interested users are welcome to update the BLUECAT software (see Section 7) with additional options.

7. The BLUECAT software

The BLUECAT software is available in the R and Python3 versions (see section 'Software and data availability'). The R software installs a R function requiring the arguments listed below. The Python3 software runs as a stand alone code, reading the same arguments, settings and input data from text files. Additional details on file format, installation and running the software are given in the R help, README files of R and Python3 versions and examples of application provided for both versions, which allow to reproduce the case studies presented in Section 8.

The user needs to specify the following arguments (acronyms and variable names are those used in the software):

- *resultcalib*, real values, list in the R version, matrix in the Python3 version, providing the predicted and observed data, for each considered model, to be used for calibrating BLUECAT;
- *modelsim*, matrix of real values, providing the D-model output for which uncertainty is to be assessed, for each considered model;
- *nmodels*, integer value, the number of models considered in the multimodel approach. Default is 1;
- *uncmeas*, integer, $1, \ldots, 6$ for using the uncertainty measure given by Eqs. (5), (6), (7), (8), (9) and (10), respectively, in the multimodel approach. Default is 2;
- *predsmodel*, character, "avg" (default) or "mdn" for adopting the average value or the median of the sample as S-model prediction;
- *empquant*, logical value, T or F (default) for estimating empirical quantiles with sample statistics or K-moments;

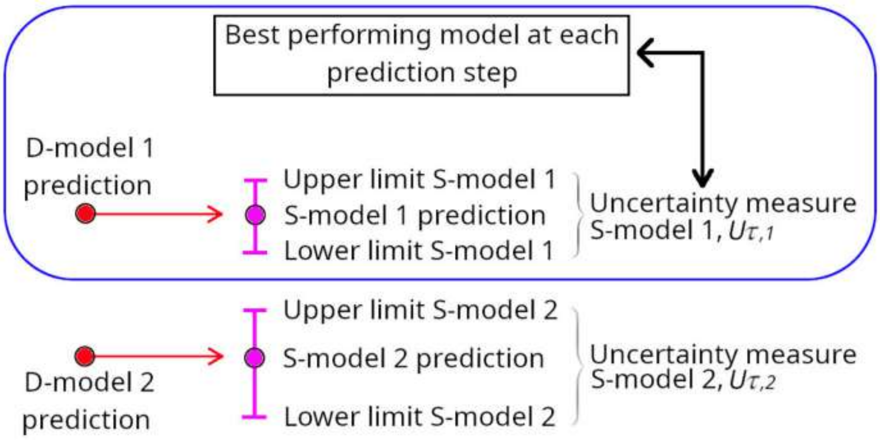

Fig. 3. The multimodel BLUECAT uncertainty based, step-by-step model selection.

- *siglev*, real value, significance level of the estimated confidence bands; default value is 0.2;
- *m*, integer, the number of predicted (and corresponding observed) data points lower and higher than each model output that are used to build the sample of observations to support uncertainty estimation. Default is 100;
- m_1, integer, the number of K-moments used for the robust estimation of quantiles. Default is 80;
- *paramd*, vector of 4 real numbers, the initial values for the parameters of the PBF distribution, default is (0.1,1,10,NA);
- *lowparamd*, vector of 4 real numbers, the lower values for the parameters of the PBF distribution, default is (0.001,0.5,0.001,0);
- *upparamd*, vector of 4 real numbers, the upper values for the parameters of the PBF distribution, default is (1,5,20,NA);
- *qoss*, vector of real values, observed data corresponding to modelsim. Default values is NULL for no data available. In this case confidence bands are computed, basing on the calibration data, but results are not tested;
- *plotflag*, logical value, *T* (default) or F for performing or not the goodness of fit testing with plots;
- *cpptresh*, real value, threshold level indicating the minimum value of observed data to be used for the goodness of fit tests. Default is the minimum of observed and predicted data.

The above options are specified in the R command line invoking the *bluecat.sim* R function, or the file *settings.txt* in the Python3 version. The interested user may refer to the R help and README files of R and Python3 versions for a step-by-step guidance to the installation of the software and reproduction of the case studies.

Computational time for the applications presented here is few seconds with an Intel Core i7-9700 CPU at 3.00 GHz and 16 GB RAM under the Linux operating system.

8. Examples of application

We present here two examples of uncertainty estimation with BLUECAT, that refer to a single model and a multimodel prediction, respectively.

8.1. Single model prediction of tree ring width

Franke et al. (2021) considered predictions by a single model of temperature-sensitive chronologies of standardised tree ring width (TRW) for the period 1401–2000. The predicted series were gridded and averaged over the Northern Hemisphere to reduce local noise, thus obtaining one simulation average at annual resolution, including 600 data points. Corresponding observed data were gridded and averaged in the same way. See Franke et al. (2021) for more details on data and standardisation, in particular Fig. 5 in their contribution. Their work supported a detection and attribution study of climate variations due to volcanic forcing. Prediction of TRW was obtained by applying the Vaganov–Shashkin Lite (VSL) sensor model, which estimates standardised tree-ring width (TRW) annual chronologies based on monthly mean temperature, precipitation and latitude. More details and the data herein used are given by Franke et al. (2021) to which the interested reader is referred to. Here, we limit our analysis to uncertainty assessment of the predicted and normalised TRW by comparing them with the corresponding observed values. Franke et al. (2021) report a correlation coefficient between observed and predicted series of 0.23, so the width of the estimated confidence band is expected to be large, in order to include the expected percentage of observed points. We point out that the target of Franke et al. (2021) was not to reproduce the observed data with the highest accuracy, but rather to filter the observed series to eliminate noise. Thus, uncertainty is indeed expected to be large, with the width of the confidence band indicating the magnitude of the noise that was removed.

We applied BLUECAT and goodness of fit tests by computing quantiles with K-moments (*empquant* = F), significance level (*siglev*) 0.2, $m = 50$, $m_1 = 40$ and default values for the remaining options. BLUECAT was calibrated against the first 400 data points (1401–1800) and validated over the last 200 (1801–2000).

Results are summarised in Figs. 4 and 5. The percentage of validation points lying above and below the confidence bands are 3.1% and 2.5%, respectively, namely, lower than the expected values (10%). This means that the width of the confidence band is slightly overestimated, probably in view of the approximations introduced by the low value of *m* and the reduced sample size of the calibration data set. Overestimation of the confidence band width is confirmed by the PPP plot. According to our experience and testing, the above results look satisfactory in a validation experiment over a limited sample size.

8.2. Multimodel prediction of daily river flows

Koutsoyiannis and Montanari (2022a) presented a calibration and validation experiment for BLUECAT that considered the 1-step ahead prediction of daily river flows for the Arno River at Subbiano. The D-model is HyMod (Boyle, 2000), which was calibrated by maximising Nash–Sutcliffe efficiency. The data of mean areal daily rainfall (estimated from raingauge observations), evapotranspiration (estimated from temperature data) and river flow span the 22-year period 1992–2013. We used the first 20 years for HyMod and BLUECAT calibration and the last two years for validation. The Nash–Sutcliffe efficiency of the HyMod D-model is 0.63 in calibration and 0.57 in validation. The S-model efficiency in validation is 0.62, with an overestimation of low

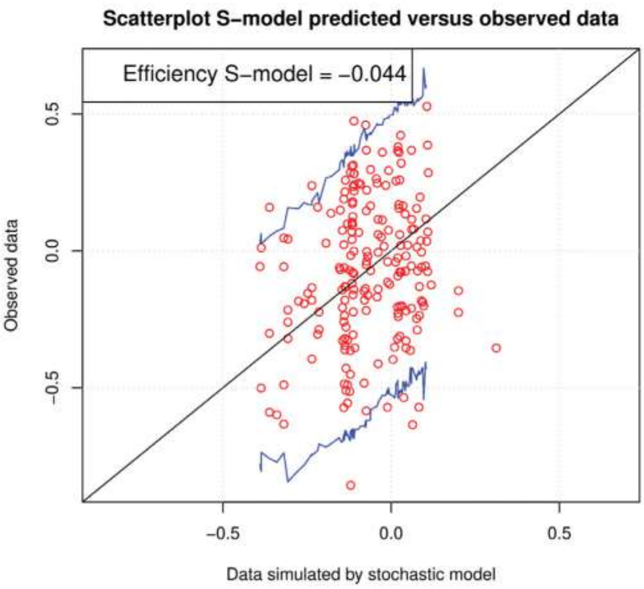

Fig. 4. Case study of tree ring widths. Scatterplot of observed versus predicted data and confidence band at the 80% confidence level, in validation, as provided by the BLUECAT software. Data are standardised and temperature sensitive tree-ring chronologies averaged over the grid boxes considered by Franke et al. (2021) (nondimensional).

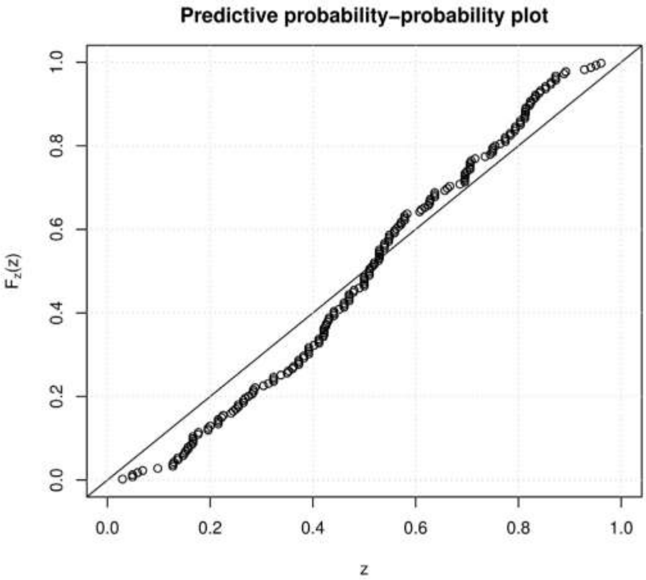

Fig. 5. Case study of tree ring widths. Predictive probability-probability plot (PPP) as provided by the BLUECAT software. The validation curve is displaced below the equality line in the first part of the diagram and above the same line in the second part, thus indicating that the band is slightly large (see also Fig. 2).

flows and confidence bands that are slightly narrow. Table 1 reports the percentage of observations lying outside the confidence limits, which are in fact higher than the value of 10% – for each limit – that one would expect for the confidence level of 80% that was used here. Extended details on model and catchment, as well as calibration and validation results, are provided by Koutsoyiannis and Montanari (2022a). In what follows, we term the above application of the HyMod model as "D-model 1".

Here, we consider an additional version of HyMod with a different parameter set ("D-model 2"), that is obtained by using the mean absolute relative error instead of the Nash–Sutcliffe efficiency as objective function for HyMod calibration. D-model 2 efficiency is 0.68 in calibration and 0.61 in validation, namely, sligthly better performances with respect to D-model 1. S-model 2 efficiency in validation is 0.63, again a slightly improvement with respect to S-model 1.

We also consider a third model for the Arno River at Subbiano ("D-model 3"), namely, the GR5J model (Perrin et al., 2003; Le Moine, 2008; Coron et al., 2017), counting 5 parameters and calibrated by maximising Nash–Sutcliffe efficiency. D-model 3 efficiency is 0.82 in calibration and 0.73 in validation, while S-model 3 efficiency in validation is 0.75. These performances mark an improvement with respect to S-model 1 and 2. Percentage of observations lying outside the confidence limits for the 3 S-models in validation are reported in Table 1.

We applied the multimodel BLUECAT and goodness of fit tests by computing quantiles with K-moments (*empquant* = F), significance level (*siglev*) 0.2, 5 different combinations of parameters m, ranging from 20 to 100, and m_1, ranging from 10 to 80. We also used the two uncertainty measures given by Eqs. (6) and (9), and default values for the remaining options. Different combinations of parameters m and m_1 allow us to test the sensitivity of BLUECAT output.

Multimodel efficiency in validation is 0.65 and 0.74, for the uncertainty measures by Eqs. (6) and (9), respectively, and the combination of parameters $m = 100$, $m_1 = 80$. Table 1 reports the percentage of observations lying outside the confidence limits for both multimodel solutions. S-model 1, S-model 2, and S-model 3 were selected as least uncertain model in 21%, 7% and 72% of the validation steps, respectively, for uncertainty measure by Eq. (6), and 37%, 7% and 56% of the validation steps, respectively, for uncertainty measure by Eq. (9).

It is interesting to note that the performances of the multimodel solution, in terms of Nash–Sutcliffe efficiency and percentage of observations lying outside the confidence limits, are not necessarily improved with respect to the best performing single model, that is, GR5J. This result is expected, as the multimodel is the composition of the least uncertain model at each prediction step, according to a given uncertainty measure. Such composition does not necessarily lead to an improvement with respect to the best performing model in terms of Nash–Sutcliffe efficiency and number of observations encompassed by the confidence bands for the overall simulation. In fact, these are two different performance indexes with respect to the uncertainty measure that has been used to compose the multimodel. Not surprisingly, when the uncertainty measure given by Eq. (9) is used, that is, the Nash–Sutcliffe efficiency in the prediction of the sample of true values $\bar{y}_i$ identified at each prediction step τ, we obtain an overall efficiency for the multimodel that is close to the best efficiency obtained by the GR5J model.

The above reasoning highlights the fundamental role of the uncertainty measure in multimodel selection, that should be chosen by bearing in mind the purpose of the application, to make sure that models are mixed with an optimal solution from a technical point of view.

Furthermore, we note from Table 1 that the performances of BLUECAT, in terms of percentage of observations lying outside the confidence band, are not much sensitive to parameters m and m_1. This means that uncertainty assessment looks reliable even when the probability distribution given by Eq. (3) is estimated over a limited data sample. This result is particularly relevant when assessing uncertainty in regions of the prediction domain with few observed data points (like, for instance, the region of peak flows). However, we noticed that the confidence band looks less regular and more fluctuating when working with low m values, for the obvious reason that estimation variance is larger. For this reason we would like to reiterate that particular care should be paid when working with limited data samples and in particular when estimating uncertainty for predictions outside the range of calibration data (see also Section 5).

Figs. 6 and 7 show the estimated confidence band for the time window October 26, 2012–December 31, 2013 of the validation period and the two selected uncertainty measures. Lower and upper confidence

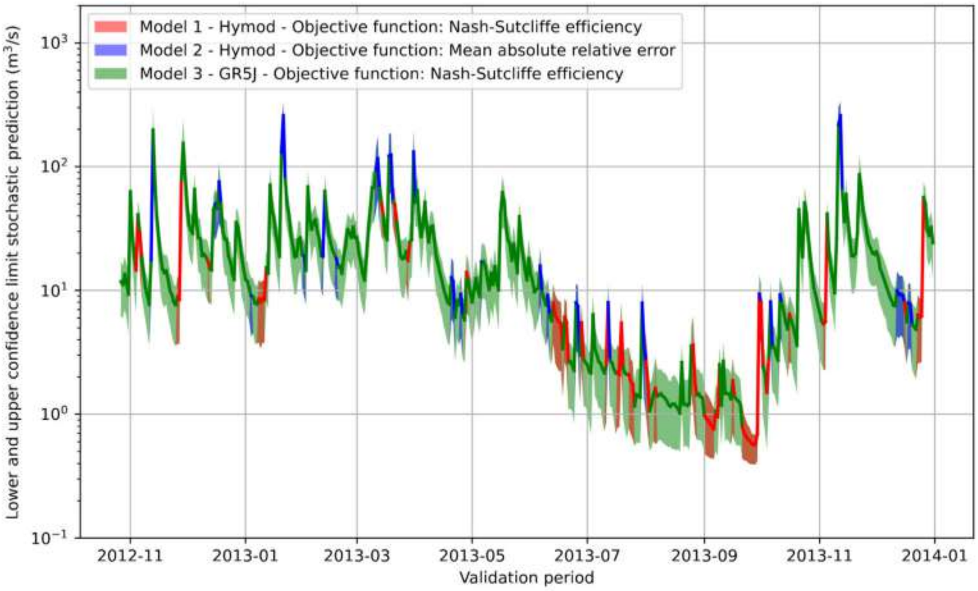

Fig. 6. Case study of Arno River at Subbiano. Results of the step-by-step model selection for the time window October 26, 2012-December 31, 2013 of the validation period and uncertainty measure given by Eq. (6). Lower and upper confidence limits, and the confidence band between them, are marked at each prediction step with the same colour: red, blue and green when band is estimated by S-model 1, S-model 2 and S-model 3, respectively.

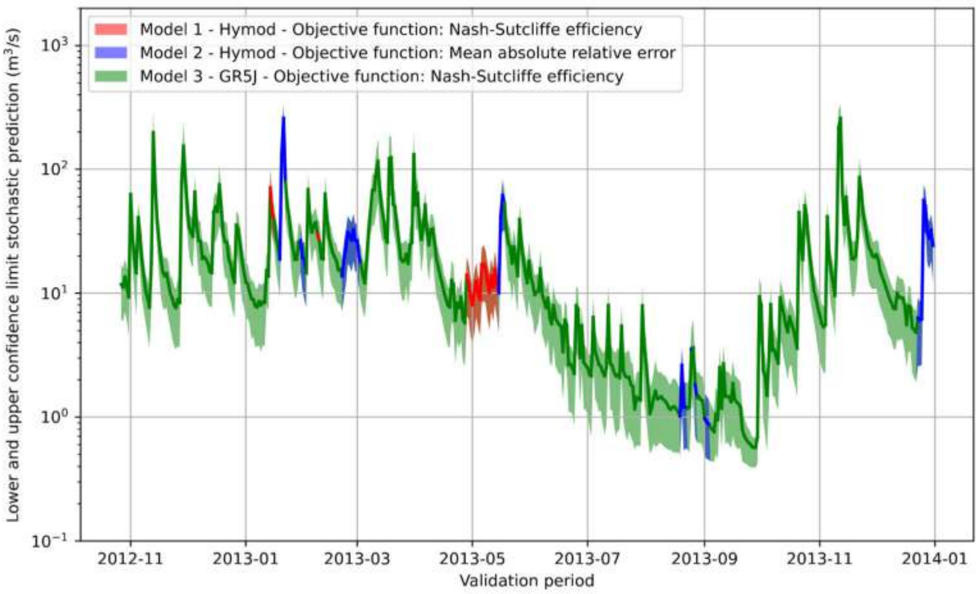

Fig. 7. Case study of Arno River at Subbiano. Results of the step-by-step model selection for the time window October 26, 2012-December 31, 2013 of the validation period and uncertainty measure given by Eq. (9). Lower and upper confidence limits, and the confidence band between them, are marked at each prediction step with the same colour: red, blue and green when band is estimated by S-model 1, S-model 2 and S-model 3, respectively.

limits, and the confidence band between them, are marked at each prediction step with the same colour: red when band is estimated by S-model 1, blue for S-model 2 and green for S-model 3. It is confirmed that S-model 3 is selected for most of the prediction steps.

Fig. 8 displays the PPP plots for S-model 1, S-model 2 , S-model 3 and S-multimodel with uncertainty measure given by Eq. (9). Fig. 9 shows the scatterplot of observed versus predicted by the same S-multimodel river flows, along with confidence limits at the 80% confidence level, in validation.

The results further confirm that the uncertainty measure plays a relevant role for model selection. However, a careful inspection of the simulation results presented in Figs. 6 and 7 revealed that there is no large difference between the two multimodels in the magnitude of the prediction at each time step τ.

Table 1
Percentage of observations lying outside the 80% confidence limits for the multimodel case study of the Arno River at Subbiano. Band was estimated with K-moments. Subscripts u and l refer to upper and lower limit, respectively.

m, m_1	S1		S2		S3		SM1		SM2	
	$\%_u$	$\%_l$	$\%_u$	$\%_l$	$\%_u$	$\%_l$	$\%_u$	$\%_l$	$\%_u$	$\%_l$
100, 80	16%	15%	19%	11%	11%	14%	11%	17%	15%	17%
80, 60	16%	15%	19%	10%	12%	16%	11%	19%	15%	17%
60, 40	16%	15%	19%	10%	11%	15%	12%	19%	14%	17%
40, 20	16%	16%	20%	11%	11%	15%	12%	19%	14%	17%
20, 10	16%	15%	20%	11%	12%	17%	14%	19%	13%	17%

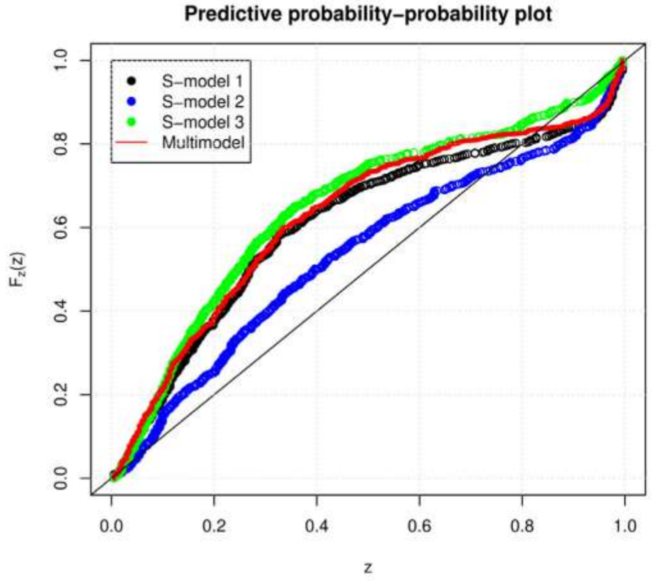

Fig. 8. Case study of Arno River at Subbiano. Predictive probability-probability plots for S-model 1, S-model 2 and S-multimodel with uncertainty measure given by Eq. (9) in validation provided by the BLUECAT software.

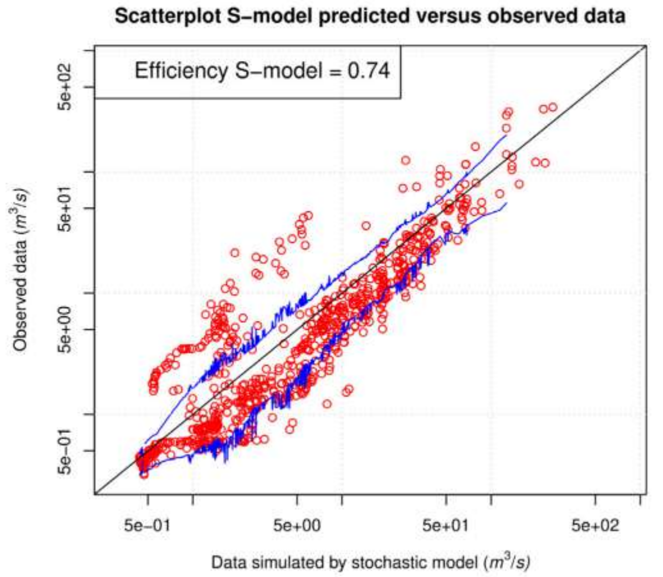

Fig. 9. Case study of Arno River at Subbiano. Scatterplot in logarithmic scale of observed versus predicted data and confidence band for the multimodel with uncertainty measure given by Eq. (9).

9. Conclusions

We present here an extension of the BLUECAT method allowing uncertainty assessment for the output from a single or multiple calibrated deterministic models. The new version of the method is suited for assessing reliability of environmental predictions and quantifying their uncertainty, which is particularly important for providing vital information to decision makers and managing environmental emergencies.

BLUECAT transforms the deterministic model – or multimodel – into a stochastic formulation, basing on assumptions that are not particularly restrictive which are discussed in Section 5. If a multimodel application is considered, BLUECAT selects at each prediction step the optimal ensemble member by identifying the solution corresponding to the minimum of a suitable uncertainty measure. Therefore, BLUECAT can be applied to combine different models to ensure that uncertainty is minimised in dependence of the state of the system.

A software is made available in the public domain for the swift application of BLUECAT in the R and Python environments (see Sections 'Software and data availability' and 7). The software comes with data and help facilities to allow reproduction of the case studies herein presented. We note that calibration of the deterministic models and BLUECAT implies optimisation procedures and therefore the reproduced results may slightly differ with respect to what is presented here.

We recommend that BLUECAT application is carried out by bearing in mind the underlying assumptions and limitations presented in Sections 5 and 6. In particular, the sample size of the data set that is used to assess uncertainty and the uncertainty measure that is used for model selection in multimodel applications should be carefully evaluated by taking into account the behaviours of the predicted variables and the target of the analysis. These issues necessarily have to be evaluated on the basis of expert knowledge and dialogue between researchers, policy makers and end users.

We would like to emphasise that BLUECAT delivers insights on the performances and weaknesses of the underlying deterministic models, therefore providing valuable support for improving our understanding of environmental systems and model accuracy. Recognising and assessing uncertainty is not only providing an essential support to policy makers and agencies in charge of civil protection: it also delivers key information towards the improvement of environmental knowledge and predictions.

CRediT authorship contribution statement

Alberto Montanari: Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Demetris Koutsoyiannis:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Alberto Montanari reports financial support was provided by Fondazione RETURN. Alberto Montanari reports financial support was provided by Ministero dell'Universitá e della Ricerca. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are extremely grateful to four reviewers, three anonymous and one eponymous, Prof. Salvatore Grimaldi, who provided critical reviews of a previous version of the manuscript. Their comments greatly helped us to improve the analysis. We are also grateful to the Editor and Associate Editor for the professional handling of the review process.

A.M. was partially supported by (1) the RETURN Extended Partnership which received funding from the European Union NextGenerationEU (National Recovery and Resilience Plan – NRRP, Mission 4, Component 2, Investment 1.3 – D.D. 1243 2/8/2022, PE0000005) and (2) the Italian Science Fund through the project "Stochastic amplification of climate change into floods and droughts change ($CO_2$2 Water)", grant number J53C23003860001. Post Scriptum: after presenting BLUECAT, we were asked several times what is the meaning of the name we decided for the method. BLUECAT is an acronym from "Brisk local uncertainty estimator for generic simulations and

predictions". However, the name was actually inspired by the pop art on blue cats by Andy Warhol (https://en.wikipedia.org/wiki/Andy_Warhol), a success creation stimulated by a simple idea that gives a feeling of positive thinking and optimism, that inspired us during our work.

Data availability

We have included in the paper a statement mentioning that software and data to fully reproduce the results presented here are already openly available in GitHub.

References

Anderson, C.J., Glassman, M., McAfee, R.B., Pinelli, T., 2001. An investigation of factors affecting how engineers and scientists seek information. J. Eng. Technol. Manage. 18 (2), 131–155.

Auer, A., Gauch, M., Kratzert, F., Nearing, G., Hochreiter, S., Klotz, D., 2024. A data-centric perspective on the information needed for hydrological uncertainty predictions. Hydrol. Earth Syst. Sci. Discuss. 2024, 1–37.

Beven, K., 2016. Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. Hydrol. Sci. J. 61 (9), 1652–1665.

Beven, K., 2018. Environmental Modelling: an Uncertain Future? CRC Press.

Beven, K., 2019. How to make advances in hydrological modelling. Hydrol. Res. 50 (6), 1481–1494.

Boyle, D., 2000. Multicriteria Calibration of Hydrological Models (Ph.D. thesis). Univ of Arizona, Tucson.

Burke, M., Dykema, J., Lobell, D.B., Miguel, E., Satyanath, S., 2015. Incorporating climate uncertainty into estimates of climate change impacts. Rev. Econ. Stat. 97 (2), 461–471.

Cappelli, F., Papalexiou, S.M., Markonis, Y., Grimaldi, S., 2024. PyCoSMoS: An advanced toolbox for simulating real-world hydroclimatic data. Environ. Model. Softw. 178, 106076.

Coron, L., Thirel, G., Delaigue, O., Perrin, C., Andréassian, V., 2017. The suite of lumped GR hydrological models in an R package. Environ. Model. Softw. 94, 166–171.

Dewulf, A., Biesbroek, R., 2018. Nine lives of uncertainty in decision-making: strategies for dealing with uncertainty in environmental governance. Policy Soc. 37 (4), 441–458.

Eslamian, S., 2014. Handbook of Engineering Hydrology: Modeling, Climate Change, and Variability. CRC Press.

Franke, J., Evans, M.N., Schurer, A., Hegerl, G.C., 2021. Climate change detection and attribution using observed and simulated tree-ring width. Clim. Past Discuss. 2021, 1–13.

Giustolisi, O., Doglioni, A., Savic, D.A., Webb, B., 2007. A multi-model approach to analysis of environmental phenomena. Environ. Model. Softw. 22 (5), 674–682.

Gomes Jr., M.N., Giacomoni, M.H., Navarro, F.A.R., Mendiondo, E.M., 2024. Global optimization-based calibration algorithm for a 2D distributed hydrologic-hydrodynamic and water quality model. Environ. Model. Softw. 179, 106128.

Grenouillet, G., Buisson, L., Casajus, N., Lek, S., 2011. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. Ecography 34 (1), 9–17.

Herrmann, M., Marzocchi, W., 2023. Maximizing the forecasting skill of an ensemble model. Geophys. J. Int. 234 (1), 73–87.

Honti, M., Stamm, C., Reichert, P., 2013. Integrated uncertainty assessment of discharge predictions with a statistical error model. Water Resour. Res. 49 (8), 4866–4884.

Hughes, J.D., 2016. What is Environmental History? John Wiley & Sons.

Hughes, D., Lawrence, D., 2024. Quantifying 'realistic'uncertainty bounds as a part of sound hydrological modelling practice in data scarce regions of Southern Africa. Environ. Model. Softw. 106112.

Imhoff, R.O., Buitink, J., van Verseveld, W.J., Weerts, A.H., 2024. A fast high resolution distributed hydrological model for forecasting, climate scenarios and digital twin applications using wflow_sbm. Environ. Model. Softw. 106099.

Jonsson, E., Todorović, A., Blicharska, M., Francisco, A., Grabs, T., Sušnik, J., Teutschbein, C., 2024. An introduction to data-driven modelling of the water-energy-food-ecosystem nexus. Environ. Model. Softw. 181, 106182.

Jorquera, J., Pizarro, A., 2023. Unlocking the potential of stochastic simulation through Bluecat: Enhancing runoff predictions in arid and high-altitude regions. Hydrol. Process. 37 (12), e15046.

Kabir, H.D., Khosravi, A., Hosen, M.A., Nahavandi, S., 2018. Neural network-based uncertainty quantification: A survey of methodologies and applications. IEEE Access 6, 36218–36234.

Kim, S.S., Marshall, L.A., Hughes, J.D., Seo, L., Lerat, J., Sharma, A., Vaze, J., 2024. Improving the statistical reliability of river model predictions via simple state adjustments. Environ. Model. Softw. 171, 105858.

Koutsoyiannis, D., 2019. Knowable moments for high-order stochastic characterization and modelling of hydrological processes. Hydrol. Sci. J. 64 (1), 19–33. http://dx.doi.org/10.1080/02626667.2018.1556794.

Koutsoyiannis, D., 2023. Stochastics of Hydroclimatic Extremes – A Cool Look at Risk, Edition 3. Kallipos Open Academic Editions, p. 391. http://dx.doi.org/10.57713/kallipos-1, URL: http://hdl.handle.net/11419/6522.

Koutsoyiannis, D., Montanari, A., 2022a. Bluecat: A local uncertainty estimator for deterministic simulations and predictions. Water Resour. Res. 58 (1), e2021WR031215.

Koutsoyiannis, D., Montanari, A., 2022b. Climate extrapolations in hydrology: the expanded BlueCat methodology. Hydrology 9 (5), 86.

Laio, F., Tamea, S., 2007. Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrol. Earth Syst. Sci. 11 (4), 1267–1277. http://dx.doi.org/10.5194/hess-11-1267-2007.

Le Moine, N., 2008. Le Bassin Versant de Surface vu Par le Souterrain: Une Voie d'Amélioration des Performances et du Réalisme des Modèles Pluie-Débit? (Ph.D. thesis). Doctorat Géosciences et Ressources Naturelles, Université Pierre et Marie ....

Liang, J., Liu, S., Zhou, Z., Zhong, G., Zhen, Y., 2024. Improving probabilistic streamflow predictions through a nonparametric residual error model. Environ. Model. Softw. 175, 105981.

Lin, Q., Zhang, D., Wu, J., Chen, X., Fang, Y., Lin, B., 2024. PASS4SWAT: Orchestration of containerized SWAT for facilitating computational reproducibility of model calibration and uncertainty analysis. Environ. Model. Softw. 178, 106085.

Mangukiya, N.K., Kushwaha, S., Sharma, A., 2024. A novel multi-model ensemble framework for fluvial flood inundation mapping. Environ. Model. Softw. 180, 106163.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K., Thuiller, W., 2009. Evaluation of consensus methods in predictive species distribution modelling. Diversity and Distributions 15 (1), 59–69.

Montanari, A., Brath, A., 2004. A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. Water Resour. Res. 40 (1), http://dx.doi.org/10.1029/2003WR002540.

Montanari, A., Grossi, G., 2008. Estimating the uncertainty of hydrological forecasts: A statistical approach. Water Resour. Res. 44 (12).

Montanari, A., Koutsoyiannis, D., 2012. A blueprint for process-based modeling of uncertain hydrological systems. Water Resour. Res. 48 (9), http://dx.doi.org/10.1029/2011WR011412.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. J. Hydrol. 10 (3), 282–290.

Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. J. Hydrol. 279 (1–4), 275–289.

Plunge, S., Schürz, C., Čerkasova, N., Strauch, M., Piniewski, M., 2024. SWAT+ model setup verification tool: SWATdoctR. Environ. Model. Softw. 171, 105878.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL: https://www.R-project.org/.

Refsgaard, J.C., van der Sluijs, J.P., Højberg, A.L., Vanrolleghem, P.A., 2007. Uncertainty in the environmental modelling process–a framework and guidance. Environ. Model. Softw. 22 (11), 1543–1556.

Rozos, E., Koutsoyiannis, D., Montanari, A., 2022. KNN vs. Bluecat—Machine learning vs. classical statistics. Hydrology 9 (6), 101.

Sheikholeslami, R., Golkar, M.K., Hall, J.W., 2024. Large uncertainty in global estimates of manure phosphorus runoff. Environ. Model. Softw. 177, 106067.

Shrestha, D.L., Solomatine, D.P., 2008. Data-driven approaches for estimating uncertainty in rainfall-runoff modelling. Int. J. River Basin Manag. 6 (2), 109–122.

Sikorska, A.E., Montanari, A., Koutsoyiannis, D., 2015. Estimating the uncertainty of hydrological predictions through data-driven resampling techniques. J. Hydrol. Eng. 20 (1), A4014009.

Slater, L.J., Villarini, G., Bradley, A.A., 2019. Evaluation of the skill of north-American multi-model ensemble (NMME) global climate models in predicting average and extreme precipitation and temperature over the continental USA. Clim. Dyn. 53, 7381–7396.

Tebaldi, C., Knutti, R., 2007. The use of the multi-model ensemble in probabilistic climate projections. Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 365 (1857), 2053–2075.

Tu, T., Wang, J., Wang, C., Liang, Z., Duan, K., 2024. Reconstructing long-term natural flows by ensemble machine learning. Environ. Model. Softw. 177, 106069.

Van Rossum, G., Drake, F.L., 2009. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.

Vose, D., 2008. Risk Analysis: a Quantitative Guide. John Wiley & Sons.

Wang, S., Zhang, K., Chao, L., Chen, G., Xia, Y., Bao, H., Zhang, C., 2024. A new approach for gridded risk assessment of rainfall-triggered flood and landslide hazards over a large region based on coupled flood-landslide modelling and ensemble simulation. Environ. Model. Softw. 172, 105917.

White, J.T., Hemmings, B., Fienen, M.N., Knowling, M.J., 2021. Towards improved environmental modeling outcomes: Enabling low-cost access to high-dimensional, geostatistical-based decision-support analyses. Environ. Model. Softw. 139, 105022.

Zou, H., Marshall, L., Sharma, A., Jian, J., Stephens, C., Higgins, P., 2024. Modelling vegetation dynamics for future climates in Australian catchments: Comparison of a conceptual eco-hydrological modelling approach with a deep learning alternative. Environ. Model. Softw. 181, 106179.