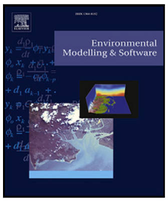
Contents lists available at ScienceDirect

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

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 (Hermann 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:

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

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 $F_{y|Y}(y|Y)$. 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:

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

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

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

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 $2m+1$.

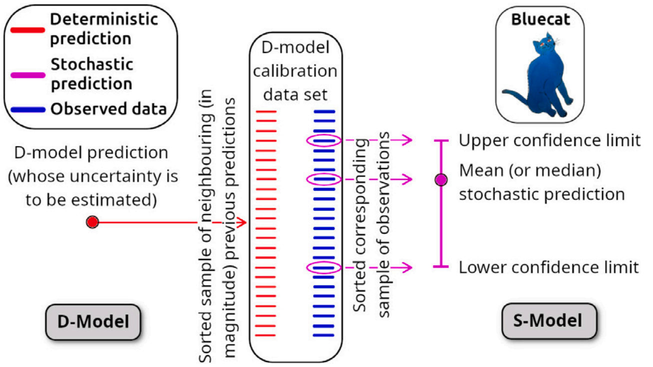

Deterministic prediction
Stochastic prediction
Observed data
D-model prediction (whose uncertainty is to be estimated)
D-Model
Sorted sample of neighbouring (in magnitude) previous predictions
D-model calibration data set
Sorted corresponding sample of observations
Bluecat
Upper confidence limit
Mean (or median) stochastic prediction
Lower confidence limit
S-Model

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 $\min(\bar{y}_i)$ and $\max(\bar{y}_i)$, 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 $\bar{y}_i$. In the BLUECAT software we use the approximation $F^{-1}_{\underline{y}|Y}(y|Y) = \min(\bar{y}_i)$, $\forall Y \leq \min(\bar{Y}_i)$ and $F^{-1}_{\underline{y}|Y}(y|Y) = \max(\bar{y}_i)$, $\forall Y \geq \max(\bar{Y}_i)$, where $\bar{Y}_i$ is the sample of predicted data corresponding to $\bar{y}_i$.

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.

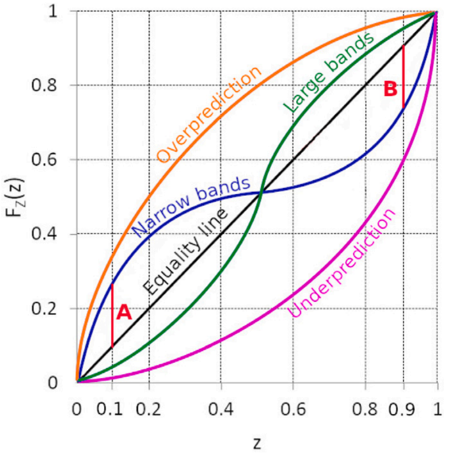


Fig. 2. Predictive probability-probability plot. Segments A and B provide a measure of the excess percentage of observations lying below the lower and above the upper confidence limit, respectively, for significance level of 20%..
Source: Reproduced from Koutsoyiannis and Montanari (2022a).

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

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

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 $\bar{y}_i$ 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, … , 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;

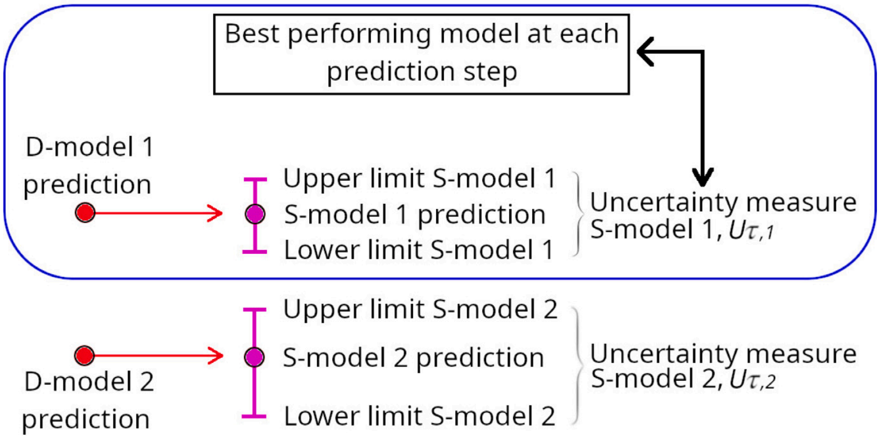


Fig. 3. The multimodel BLUECAT uncertainty based, step-by-step model selection.

- $siglev$, 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

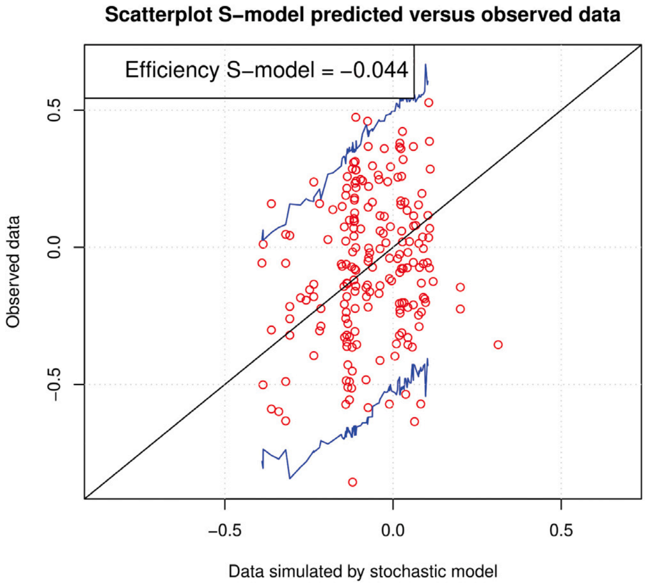


Fig. 4. Case study of tree ring widths. Scatterplot of observed versus predicted data and confidence band at the 80% confidence level, in validation, as provided by the BLUECAT software. Data are standardised and temperature sensitive tree-ring chronologies averaged over the grid boxes considered by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

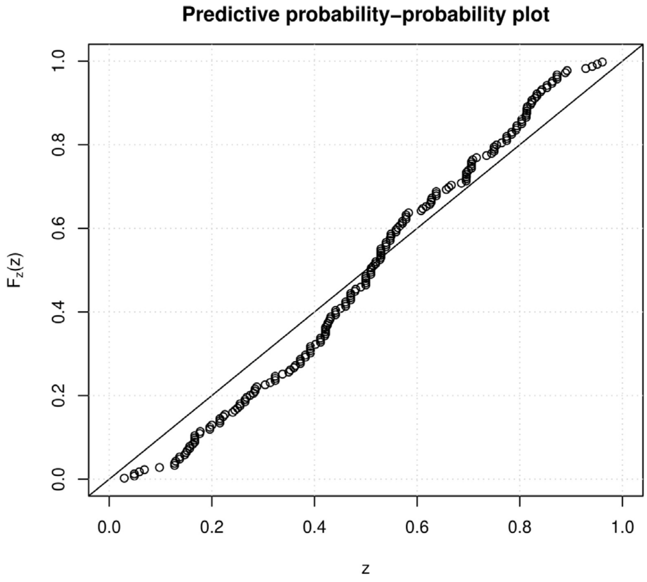


Fig. 5. Case study of tree ring widths. Predictive probability-probability plot (PPP) as provided by the BLUECAT software. The validation curve is displaced below the equality line in the first part of the diagram and above the same line in the second part, thus indicating that the band is slightly large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

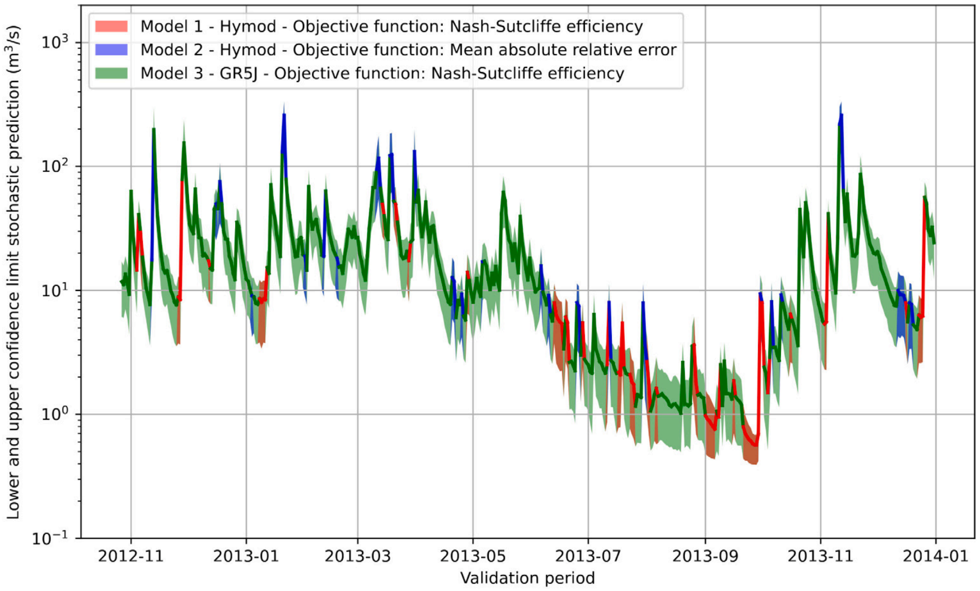


Fig. 6. Case study of Arno River at Subbiano. Results of the step-by-step model selection for the time window October 26, 2012-December 31, 2013 of the validation period and uncertainty measure given by Eq. (6). Lower and upper confidence limits, and the confidence band between them, are marked at each 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.

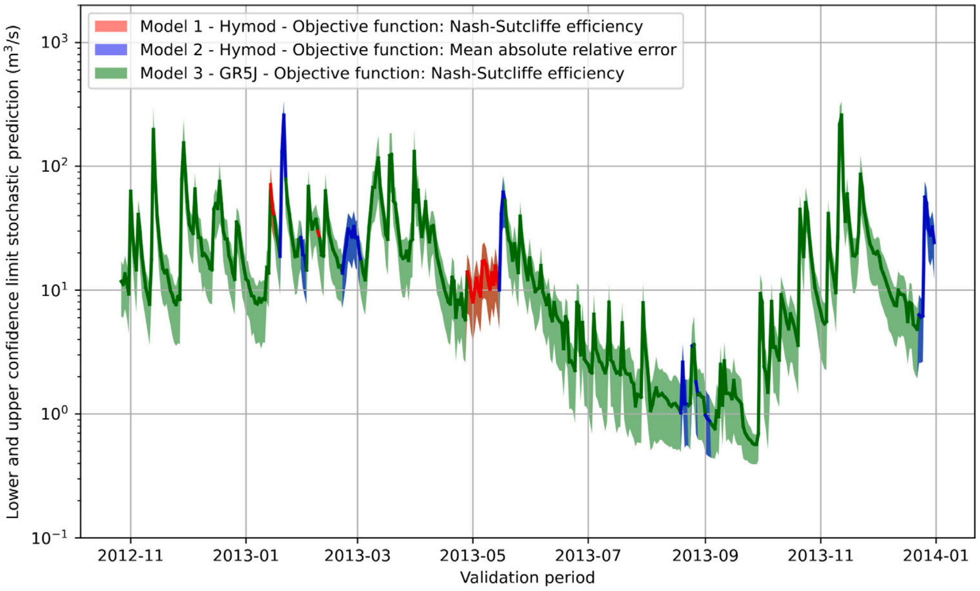


Fig. 7. Case study of Arno River at Subbiano. Results of the step-by-step model selection for the time window October 26, 2012-December 31, 2013 of the validation period and uncertainty measure given by Eq. (9). Lower and upper confidence limits, and the confidence band between them, are marked at each 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%

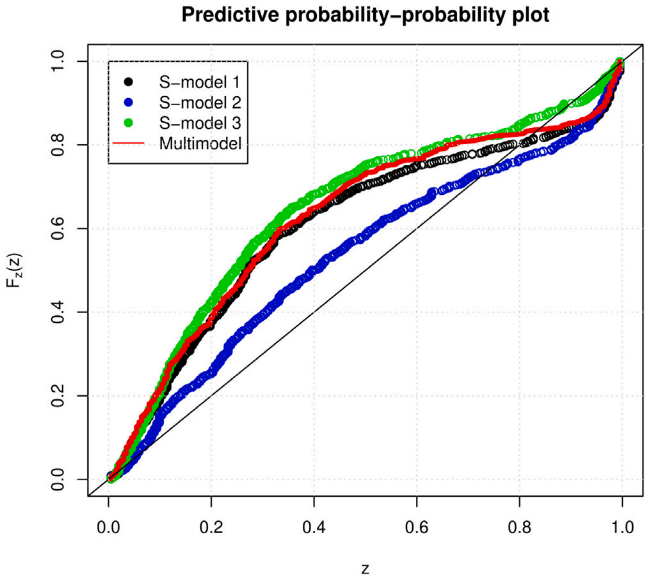


Fig. 8. Case study of Arno River at Subbiano. Predictive probability-probability plots for S-model 1, S-model 2 and S-multimodel with uncertainty measure given by Eq. (9) in validation provided by the BLUECAT software.

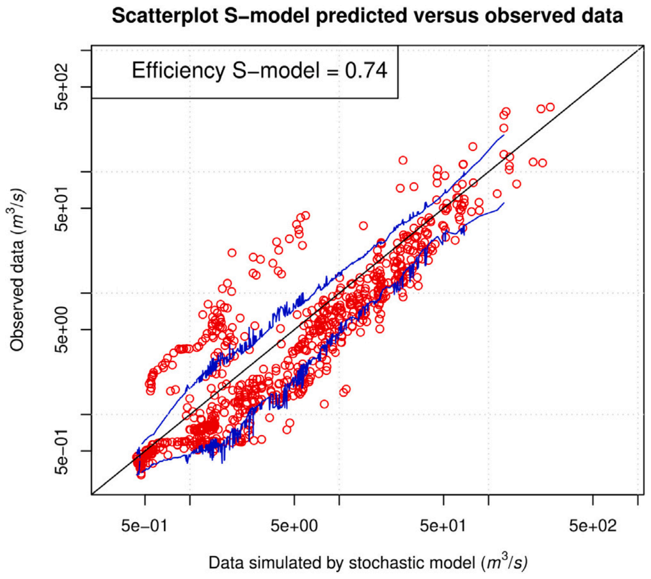


Fig. 9. Case study of Arno River at Subbiano. Scatterplot in logarithmic scale of observed versus predicted data and confidence band for the multimodel with uncertainty measure given by Eq. (9).

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_22 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.