# Combining uncertainty quantification and entropy-inspired concepts into a single objective function for rainfall-runoff model calibration

Alonso Pizarro[1], Demetris Koutsoyiannis[2], Alberto Montanari[3]

[1]Escuela de Ingeniería en Obras Civiles, Universidad Diego Portales, Santiago, 8370109, Chile
[2]National Technical University of Athens, Zographou, Athens, 15772, Greece
[3]Department DICAM, University of Bologna, Via del Risorgimento 2, Bologna, 40136, Italy

*Correspondence to*: Alonso Pizarro (alonso.pizarro@mail.udp.cl)

**Abstract.** A novel metric for rainfall-runoff model calibration and performance assessment is proposed. By integrating entropy and mutual information concepts as well as uncertainty quantification through BLUECAT (likelihood-free approach), RUMI (Ratio of Uncertainty to Mutual Information) offers a robust framework for quantifying the shared information between observed and simulated stream flows. RUMI's capabilities to calibrate rainfall-runoff models is demonstrated using the GR4J rainfall-runoff model over 99 catchments from various macroclimatic zones, ensuring a comprehensive evaluation. Four additional performance metrics and 50 hydrological signatures were also used for performance assessment. Key findings indicate that RUMI-based simulations provide more consistent and reliable results compared to the traditional Kling-Gupta Efficiency (KGE), with improved performance across multiple metrics and reduced variability. Additionally, RUMI includes uncertainty quantification as a core computation step, offering a more holistic view of model performance. This study highlights the potential of RUMI to enhance hydrological modelling through better performance metrics and uncertainty assessment, contributing to more accurate and reliable hydrological predictions.

## 1 Introduction

### 1.1 Motivation

Rainfall-runoff models are valuable tools for studying catchment responses to different hydrometeorological inputs and variations in catchment characteristics. Rainfall-runoff modelling considers various modelling choices that can significantly affect modelling results (see, e.g., Alexander et al., 2023; Knoben et al., 2019; Melsen et al., 2019; Mendoza et al., 2016; Thirel et al., 2024; Trotter et al., 2022). Among these, it is worth mentioning the model structure, spatial and temporal discretisation, input data, and calibration strategies. The latter refers not only to the selection period for warm-up, calibration, and validation, but also to one or more hydrological variable(s) considered for calibration purposes. The adopted objective function, which quantifies the similarity between observations and simulations, is also a critical step. Previous studies have highlighted the need for particular objective functions to reproduce case-specific parts of the streamflow time series (see, e.g., Acuña and Pizarro, 2023; Garcia et al., 2017; Mizukami et al., 2019). For instance, if the modeller is intended to

reproduce high flows (without caring too much about low flows), specific objective functions for high flows are recommended (Hundecha and Bárdossy, 2004; Mizukami et al., 2019). The same can be said for low or middle flows (Garcia et al., 2017).

The Nash-Sutcliffe efficiency (NSE, Nash and Sutcliffe, 1970) and the Kling-Gupta efficiency (KGE, Gupta et al., 2009) are two widely used objective functions for calibration purposes in rainfall-runoff modelling. Despite their popularity, alternatives are available in the literature (see, e.g., without intending to provide a comprehensive list, Kling et al., 2012; Koutsoyiannis, 2025; Onyutha, 2022; Pechlivanidis et al., 2014; Pizarro and Jorquera, 2024; Pool et al., 2018; Tang et al., 2021; Yilmaz et al., 2008). The reader is also referred to the following studies: Bai et al., 2021; Barber et al., 2020; Clark et al., 2021; Jackson et al., 2019; Lamontagne et al., 2020; Lin et al., 2017; Liu, 2020; Melsen et al., n.d.; Pushpalatha et al., 2012; Vrugt and de Oliveira, 2022; Ye et al., 2021. However, and to the best of our knowledge, only a small number of objective functions consider uncertainty quantification explicitly as a core step in their computation (even though hydrology has witnessed a growing emphasis on uncertainty quantification, driven by the need to enhance our understanding of catchments and to provide decision-makers with accurate model predictions). Advancements in the direction of proposing a novel and easy-to-use objective function that considers uncertainty quantification in its formulation is the primary goal of this paper.

### 1.2 Uncertainty quantification methods

Various methodologies are available aimed at better treating uncertainty, each differing in underlying assumptions, mathematical rigour, and the treatment of error sources (see, e.g., Beven, 2018; Blazkova and Beven, 2002, 2004; Krzysztofowicz, 2002). Among these approaches (see Gupta and Govindaraju 2023 for a recent review), we can mention the additive Gaussian and generalised-Gaussian process, the inference in the spectral domain, the time-varying model parameters, and multi-model ensemble methods. Additionally, two philosophies for uncertainty analysis are widely recognised, following formal and informal Bayesian methods (Kennedy and O'Hagan, 2001; Kuczera et al., 2006).

Formal Bayesian methods offer robust frameworks for uncertainty estimation, but they come with their own challenges. Identifying a suitable likelihood function for hydrological models involves careful assumptions that must be transparent and understandable to end users (Beven, 2024; Vrugt et al., 2022). Statistical analysis of model errors and likelihood-free approaches have also been proposed. For example, Montanari and Koutsoyiannis (2012) proposed converting deterministic models into stochastic predictors by fitting model errors with meta-Gaussian probability distributions. Similarly, Sikorska, Montanari, and Koutsoyiannis (2015) proposed the nearest neighbouring method to estimate the conditional probability distribution of the error. More recently, Koutsoyiannis and Montanari (2022) introduced a simple method to simulate stochastic runoff responses called Brisk Local Uncertainty Estimator for Hydrological Simulations and Predictions (BLUECAT). BLUECAT is a likelihood-free approach as relies on data only. BLUECAT has recently been applied coupled with climate extrapolations (Koutsoyiannis and Montanari 2022), rainfall-runoff modelling in a variety of different

hydroclimatic conditions (Jorquera and Pizarro, 2023), and comparisons with machine-learning methods (Auer et al., 2024; Rozos et al., 2022).

Informal Bayesian methods are more flexible, but they lack statistical rigour. A notable example of a relatively simple approach is the Generalised Likelihood Uncertainty Estimation (GLUE) method introduced by Beven and Binley (1992). GLUE operates within the framework of Monte Carlo analysis coupled with Bayesian or fuzzy uncertainty estimation and propagation. Since its introduction, GLUE has seen widespread application across various fields, including rainfall-runoff modelling (among others). Its popularity is mainly due to its conceptual simplicity and ease of implementation. It can account for all causes of uncertainty, either explicitly or implicitly, and allows for evaluating multiple competing modelling approaches, embracing the concept of equifinality (Beven, 1993). However, GLUE has faced criticism in terms of the subjective decisions required in its application and how these affect prediction limits (informal likelihood function, lack of maximum likelihood parameter estimation, and omission of explicit model error consideration). This subjectivity might lead to not being formally Bayesian (for that reason, GLUE includes the term "generalised" in its name). Proponents of GLUE argue that it is a practical methodology for assessing uncertainty in non-ideal cases (Beven, 2006), while critics advocate for coherent probabilistic approaches. This ongoing debate underscores the need to establish common ground between these perspectives. Under various conditions, both Bayesian and informal Bayesian methods can yield similar estimates of predictive uncertainty. Building on previous work (see, e.g., Blasone et al. 2008), researchers have compared GLUE with formal Bayesian approaches. In this regard, both formal Bayesian approaches as well as GLUE can be used with advanced Monte Carlo Markov Chain (MCMC) schemes such as the Differential Evolution Adaptive Metropolis (DREAM, Vrugt et al. 2008). It is important to note that defining likelihood functions and searching the solution space during calibration are two independent issues. One way to get around these problems relies on the limits of acceptability which are typically used (but not mandatory) with GLUE (see, e.g., Beven et al., 2024; Beven and Lane, 2022; Freer et al., 2004; Page et al., 2023; Vrugt and Beven, 2018), involving more thoughtful decisions about the data (even though still with subjectivity). Additionally, studies have addressed these questions by assessing the uncertainty in synthetic river flow data using GLUE (see, e.g., Montanari 2005) and introducing open-source software packages such as the CREDIBLE uncertainty estimation toolbox (CURE, Page et al. (2023)), coded in Matlab (https://www.lancaster.ac.uk/lec/sites/qnfm/credible/default.htm, last access: 03/12/2024). CURE includes several methods, among them the Forward Uncertainty Estimation; GLUE; and, Bayesian Statistical Methods.

In addition to these methods, information theory offers valuable tools for quantifying information in hydrological models. Shannon's (1948) seminal work on information theory introduced measures such as Shannon entropy, which quantifies the expected surprise (or information) in a sample from a distribution of states. Shannon entropy can be extended to joint distributions of multiple variables, including conditional dependencies. In hydrology, Shannon entropy and mutual information have been used to assess the uncertainty in discharge predictions, as demonstrated by Amorocho and Espildora (1973) and Chapman (1986). More recently, Weijs, Schoups, and van de Giesen (2010); Weijs, Van Nooijen, and Van De Giesen (2010); Gong et al. (2013, 2014); Pechlivanidis et al. (2014); Pechlivanidis et al. (2016); Ruddell, Drewry, and

Nearing (2019) used information-theoretic objective functions for model evaluation. Despite the challenges associated with accounting for uncertainties and statistical dependencies in time series data, information-theoretic objective functions have proven valuable for streamflow simulations, complementing traditional measures such as the Nash-Sutcliffe efficiency (NSE; Nash and Sutcliffe 1970) and the Kling-Gupta efficiency (KGE; Gupta et al. 2009; Kling, Fuchs, and Paulin 2012).

### 1.3 Manuscript's goals

In this work, we study the combination of likelihood-free (BLUECAT) and information theory approaches for rainfall-runoff modelling over 99 catchments having different hydroclimatic contexts. The latter with the intention to quantify and reduce uncertainty in hydrological predictions. The Ratio of Uncertainty to Mutual Information (RUMI) is proposed as a dimensionless metric to be adopted as objective function for calibration purposes. The target aligns with the twentieth of the twenty-three unsolved problems in hydrology (*20. How can we disentangle and reduce model structural/parameter/input uncertainty in hydrological prediction?*, Blöschl et al. 2019). In detail, the following questions are herein addressed:

a) How can the calibration of deterministic model parameters be improved by using a stochastic formulation of the deterministic model?
b) How can uncertainty resulting from the final stochastic model be incorporated into the calibration process of the deterministic model?

This paper is organised as follows: Section 2 presents the used database (catchments properties and data availability), rainfall-runoff model description, and calibration strategies. Section 3 shows the calibration's and validation's results of RUMI-based simulations (as well as KGE-based ones). Daily runoff simulations as well as hydrological signatures' are considered. Strengths and limitations are discussed in Section 4, and conclusions are at the end.

## 2 Methods

### 2.1 Data

99 catchments were selected from the CAMELS-CL database (Alvarez-Garreton et al., 2018) to ensure that only catchments with near-natural hydrological regimes were included (see Figure 1 for location and chosen catchment characteristics. Five macroclimatic zones are covered). The latter was achieved through eight specific criteria: first, the daily discharge time series, though possibly non-consecutive, had to have less than 25% missing data for the period 1990–2018. Additionally, catchments with large dams were excluded (big_dam = 0). Additionally, catchments with more than 10% of discharge allocated to consumptive uses were excluded (i.e., interv_degree < 0.1 to be considered). Catchments with glacier cover higher than 5% were also excluded (i.e., lc_glacier < 5% to be considered). Furthermore, the selected catchments had less than 5% of their area classified as urban (imp_frac < 5%), and irrigation abstractions did not exceed 20% (crop_frac < 20%). Areas with forest plantations covering more than 20% of the catchment area were also excluded (fp_frac < 20%). Finally, catchments showing signs of artificial regulation in their hydrographs were removed. Worth mentioning is that after each

criterion mentioned above there is a parenthesis which followed the CAMELS-CL nomenclature. For instance, glaciar cover is catalogued as “lc_glacier” and large dams as “big_dam”.

The chosen catchments have diverse characteristics, reflecting significant variability. For instance, the smallest catchment has a size of 35 $km^2$, whereas the largest one has a size of 11,137 $km^2$ (median catchment size is 672 $km^2$). In terms of mean annual precipitation, it ranges from 94 to 3,660 mm/year (median value of 1,393 mm/year). The aridity index also covers a wide spectrum of values, ranging from 0.3 (Southern Chile) to 31.6 (Northern Chile). Its median is 0.69. In terms of mean elevations, they range between 118 (western, Pacific Ocean) and 4,270 (eastern, Andes Mountains) meters above sea level (m.a.s.l.). They have a median elevation of 1,052 m.a.s.l.. In terms of seasonality, winter rainfall predominates with a few exceptions in Northern catchments where precipitation is concentrated during the summer (Garreaud, 2009). Additionally, precipitation usually increases from north to south while temperatures decrease (Sarricolea et al., 2017). Daily precipitation and potential evapotranspiration data from the CAMELS-CL database were used, with the primary output being simulated daily streamflow. The analysis focuses on the period from 1990 to 2018, with a warm-up phase from 1990 to 1992, a calibration phase from 1992 to 2005, and a validation phase from 2005 to 2018.

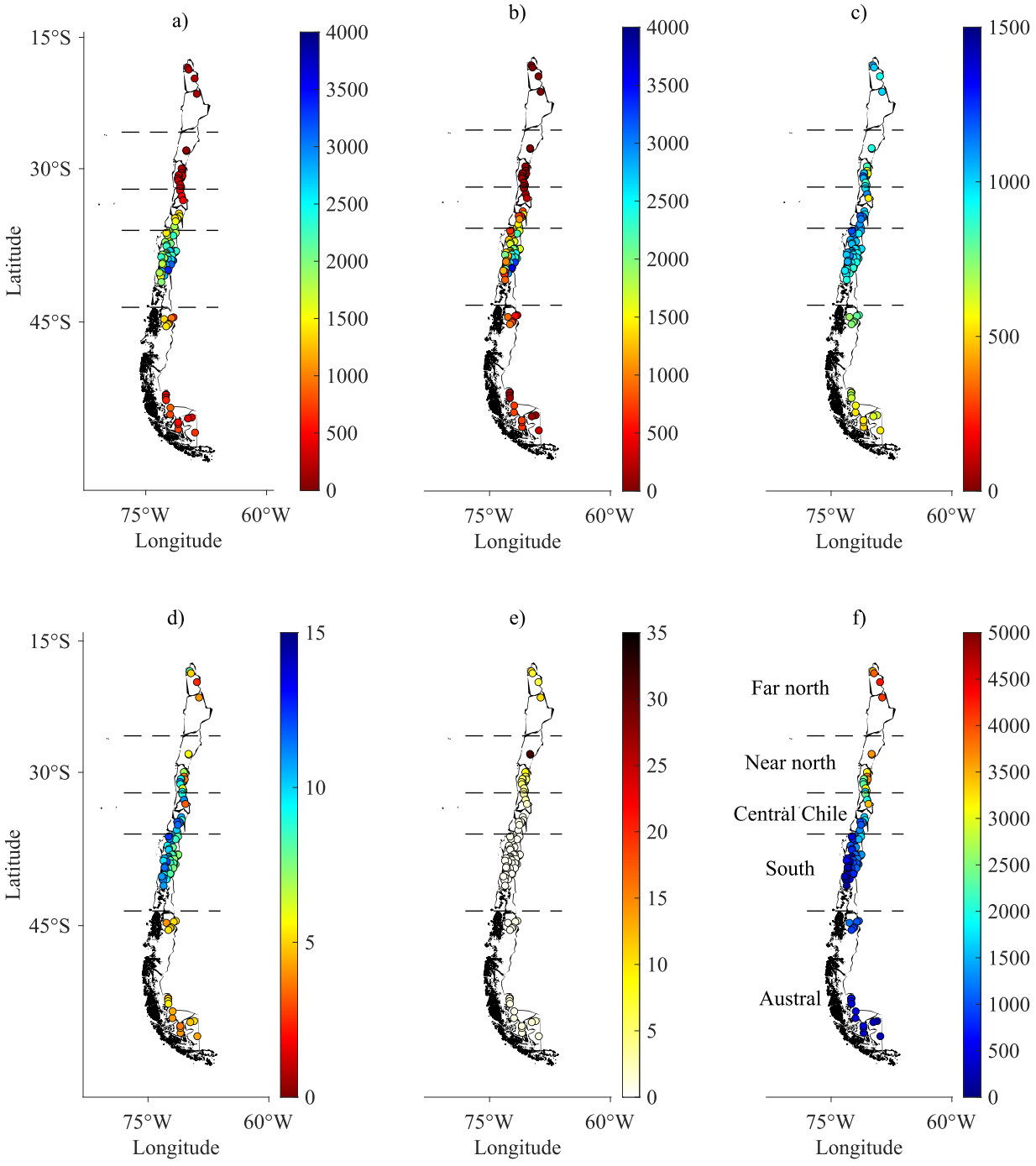


**Figure 1: Locations and characteristics of analysed catchments. Coloured dots represent the catchment outlet locations. Five zones are explicitly presented on the right to highlight differences of catchment climatic characteristics. From a) to c), mean annual**

**precipitation, runoff, and potential evapotranspiration (all of them in [mm]). d) Mean annual temperature in [º C], e) Aridity index (dimensionless), and, f) Catchment outlet elevations in [m].**

### 2.2 Rainfall-Runoff Model

The Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT – Knoben, Freer, Fowler, et al., 2019; Trotter et al., 2022) was selected due to its open-source feature and modular structure. Implemented in MATLAB, MARRMoT offers a suite of 47 lumped models for simulating rainfall-runoff processes.

MARRMoT version 2.1.2, with the GR4J model, was employed for this study. The GR4J model has four parameters and two storage components. Its primary purpose is to represent processes such as vegetation interception, time delays within the catchment, and water exchange with neighbouring catchments (for detailed information of the GR4J model, see Perrin et al., 2003; and the official website of the developers: https://webgr.inrae.fr/eng/tools/hydrological-models). MARRMoT's nomenclature for rainfall-runoff models is "m_XX_YY_ZZp_KKs", where XX is the number of the model within MARRMoT, YY is the model name, ZZ is the number of parameters, and KK is the number of storages. As a consequence, the GR4J model following MARRMoT nomenclature is: "m_07_gr4j_4p_2s". For a comprehensive description, readers are directed to the MARRMoT user manual, available at: https://github.com/wknoben/MARRMoT/blob/master/MARRMoT/User%20manual/v2.-%20User%20manual%20-%20Appendices.pdf (last accessed: 03/12/2024).

### 2.3 Ratio of Uncertainty to Mutual Information (RUMI) objective function

The primary goal of this paper is to introduce a new objective function that considers uncertainty quantification in its formulation and, therefore, it is expected to minimise this quantified uncertainty in calibration. As a consequence, the Ratio of Uncertainty to Mutual Information (RUMI) is proposed (see Eq. 4 for the mathematical expression and Figure 3 for RUMI computation flowchart). RUMI relies on BLUECAT and mutual information (entropy-based computation) which are briefly introduced as follows.

Koutsoyiannis and Montanari (2022) proposed BLUECAT with the intention to transform a deterministic prediction model into a stochastic one. BLUECAT's predecessor was introduced by Montanari and Koutsoyiannis (2012). BLUECAT transforms deterministic simulations into stochastic simulations (with confidence bands). Unlike deterministic predictions, the confidence band represents a range of possible outcomes, allowing to consider the stochastic result as a representative value of the sample (such as the mean or median). It is worth mentioning that uncertainty can be quantified as well. We use BLUECAT to transform deterministic rainfall-runoff simulations to stochastic ones to consider uncertainty quantification in model calibration.

BLUECAT's flowchart starts with a deterministic simulation and identifies the simulated variable (streamflow in our case) at each time point (see Figure 2 for a conceptual illustration of BLUECAT methodology). For each point, a sample is established comprising neighbouring simulated river flows (in magnitude), defined by $m_1$ flows smaller and $m_2$ flows larger

than the point's discharge, both with the smallest differences. The observed data corresponding to these simulated flows forms a sample of streamflow values. The latter is happening at each time point. An empirical distribution function of this sample is then used to estimate uncertainty for a given confidence level, using the mean or median as representative results of the stochastic simulation. Alternative methods, such as the ones using a theoretical probability distribution can also manage the sample (e.g., Pareto-Burr-Feller with knowable moments).

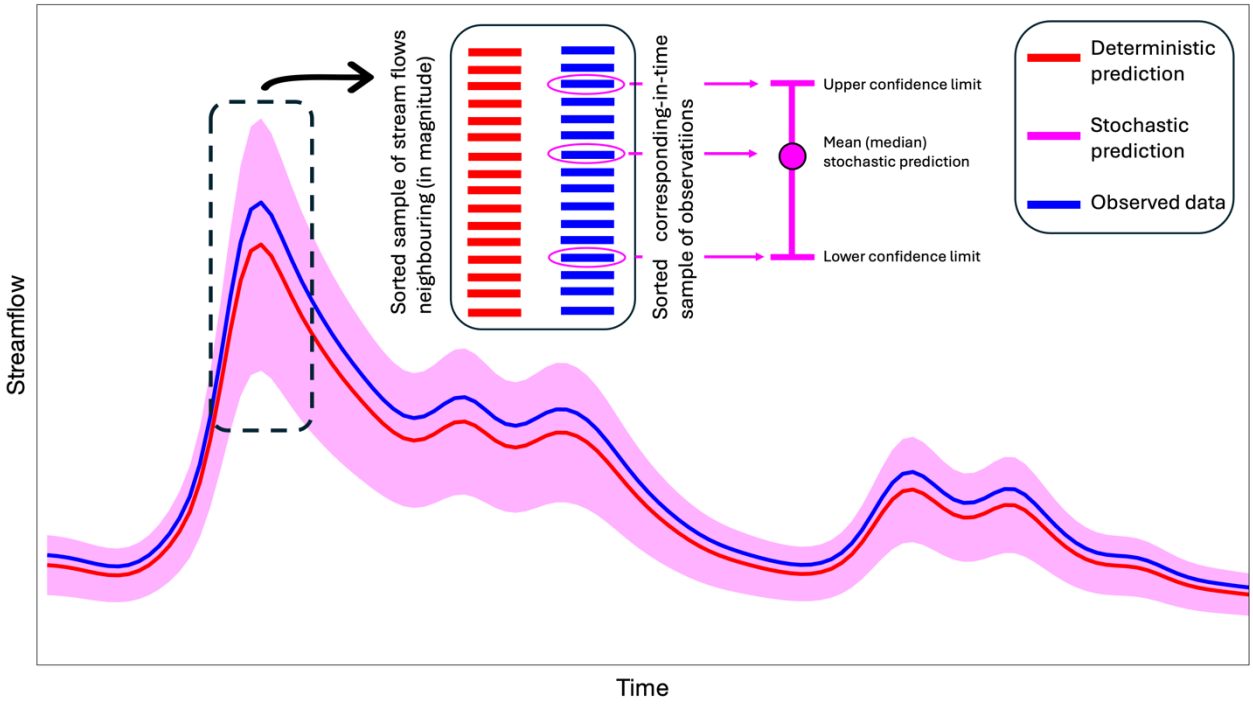


**Figure 2: Conceptual illustration of BLUECAT methodology. Blue colour represents observed (streamflow) data, whereas red and pink colours are deterministic and stochastic predictions respectively.**

In this work, BLUECAT is used with empirical computations with the intention to avoid any additional assumption. It is worth mentioning that BLUECAT allows uncertainty quantification through an uncertainty measure. Montanari and Koutsoyiannis (2025) proposed 4 measures basing on the distance between the confidence bands, for a given significance level, and the mean value of the prediction. BLUECAT was originally implemented in R (coupled with the HyMod rainfall-runoff model, Koutsoyiannis and Montanari 2022) and recently, Montanari and Koutsoyiannis (2025) made available BLUECAT with multimodel usage in R and Python. Codes in Matlab are also available (see Jorquera and Pizarro 2023).

In information theory, the entropy of a random variable is a measure of its uncertainty or the measure of the information amount required, on average, to describe the random variable itself (Thomas and Joy, 2006). The amount of information one random variable contains about another random variable is usually defined as mutual information (MI). MI is, indeed, the

reduction of one random variable uncertainty due to the knowledge of the other. MI can be defined as a function of marginal $\left(H\left(\underline{Y}\right)\right)$ and conditional entropies $\left(H(\underline{Y}/\underline{X})\right)$:

$$\mathrm{MI}(\underline{Y},\underline{X}) = H\left(\underline{Y}\right) - H(\underline{Y}/\underline{X}), \quad (1)$$

where $H\left(\underline{Y}\right) = -E\left[\log\left(p(Y)\right)\right]$, $H\left(\underline{Y}/\underline{X}\right) = -E\left[\log\left(p(Y/X)\right)\right]$, $p(\alpha)$ is the probability mass function of a random variable $\underline{\alpha}$ (or the probability density if the variable is of continuous type), and $E[\ \ ]$ denotes expectation. Note that random variables are underlined, following the Dutch convention (Hemelrijk, 1966).

Additionally, the normalised mutual information (also called as uncertainty coefficient, entropy coefficient, or Theil's U) can be computed as:

$$U\left(\underline{Y},\underline{X}\right) = \frac{\mathrm{MI}(\underline{Y},\underline{X})}{H\left(\underline{Y}\right)} = \frac{H(\underline{Y}) - H(\underline{Y}/\underline{X})}{H\left(\underline{Y}\right)}. \quad (2)$$

Taking $\underline{Y}$ as the observed streamflow $\left(\underline{Q_{\mathrm{obs}}}\right)$ and $\underline{X}$ as the simulated one with BLUECAT ($\underline{Q_{\mathrm{sim}}}$, given by the mean value of the distribution of the predictand), $U\left(\underline{Y},\underline{X}\right) = U\left(\underline{Q_{\mathrm{obs}}},\underline{Q_{\mathrm{sim}}}\right)$ represents the normalised amount of information that $\underline{Q_{\mathrm{sim}}}$ contains about $\underline{Q_{\mathrm{obs}}}$. Note that $\underline{Q_{\mathrm{sim}}}$ can also be estimated by the median value of the distribution of the predictand (or another quantile). The decision of using the mean value relies on Jorquera and Pizarro (2023) results that showed higher KGE values using the mean than the median value for all analysed catchments. Additionally, and with the intention to avoid any additional assumption, marginal and conditional entropies are computed empirically with bins.

Furthermore, an uncertainty measure (in line with Jorquera and Pizarro (2023) and Montanari and Koutsoyiannis (2025) uncertainty quantification proposal) of the stochastic model computed with BLUECAT can be defined as the width of the confidence limits divided by its mean value and averaged through the whole simulation period, i.e.:

$$u = \sum_{\tau=1}^{n} \frac{1}{n} \left| \frac{Q_{\tau,u} - Q_{\tau,l}}{Q_{\tau,\mathrm{sim}}} \right|, \quad (3)$$

where $Q_{\tau,u} - Q_{\tau,l}$ are the upper and lower confidence limits for the streamflow stochastic prediction at time step $\tau$, $Q_{\tau,\mathrm{sim}}$ is its mean value at time step $\tau$, and $n$ is the total number of time steps.

Notice that both $u$ and $U\left(\underline{Q_{\mathrm{obs}}},\underline{Q_{\mathrm{sim}}}\right)$ are dimensionless quantities and, in ideal conditions, it is desirable that $u$ is minimised (i.e., low uncertainty), whereas $U\left(\underline{Q_{\mathrm{obs}}},\underline{Q_{\mathrm{sim}}}\right)$ is maximised (i.e., high mutual information between simulated and observed stream flows). Therefore, the ratio between $u$ and $U\left(\underline{Q_{\mathrm{obs}}},\underline{Q_{\mathrm{sim}}}\right)$ gives a measure of the simulation performance. It is worth to mention that the advantage of taking this ratio does not only rely on a mathematical desire (i.e., the ratio should be minimised in calibration) but on the fact that it is possible to have narrow confidence limits (i.e., low uncertainty) with a bad performance between the stochastic model predictand and observed values (i.e., low mutual information. See Fig. 3a). Additionally, it is also possible to have high mutual information (stochastic model predictand close to observed values) but

with high uncertainty as shown in Fig. 3b. Therefore, taking the ratio is twofold: i) mathematical desire (i.e., optimisation); and, ii) deductive conceptual reasoning. As a consequence, and with the intention to provide a metric ranging between 0 and 1, the **R**atio of **U**ncertainty to **M**utual **I**nformation (**RUMI**) is presented:

$$\mathrm{RUMI} = \frac{1}{1+\phi} = \frac{1}{1+\frac{u}{U(Q_{\mathrm{obs}},Q_{\mathrm{sim}})}}. \quad (4)$$

Notice that RUMI follows common-efficiency notions (i.e., perfect simulation means the highest metric value). Figure 3d shows the core steps of RUMI computation, whereas codes for RUMI are also available within this manuscript in Matlab and R (see Code and Results Availability statement).

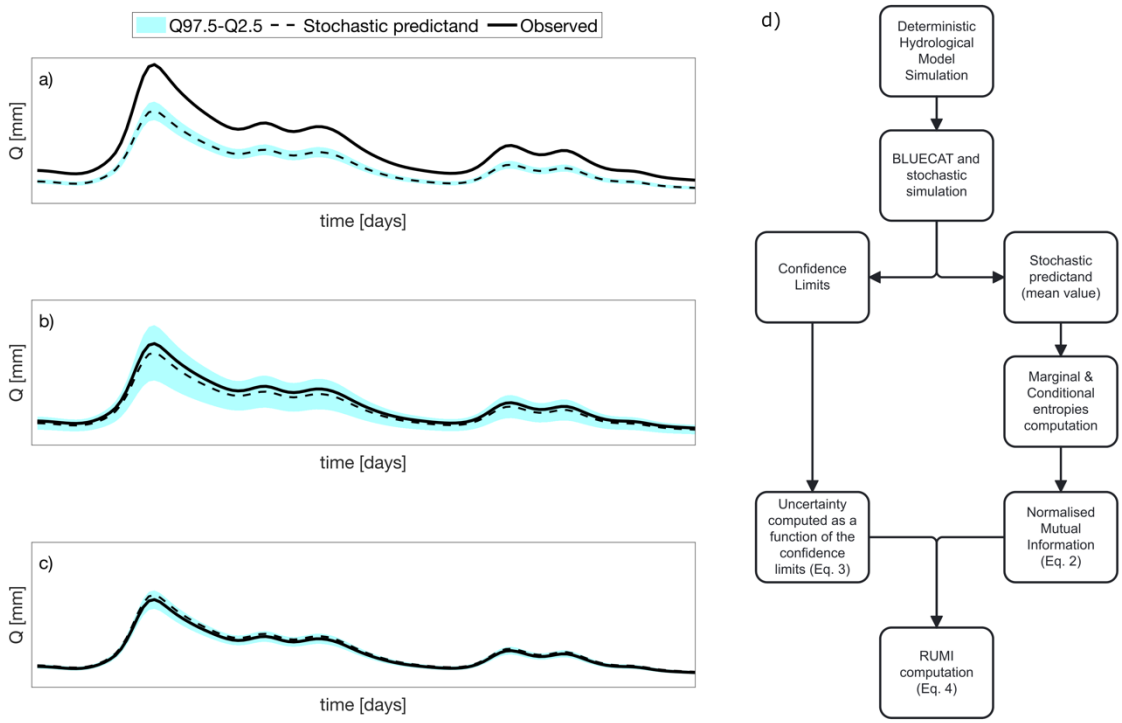


**Figure 3: Illustration of possible modelling scenarios: a) low uncertainty and low mutual information (i.e., low RUMI value); b) high uncertainty and high mutual information (i.e., low RUMI value); and, c) low uncertainty and high mutual information (i.e., high RUMI value). d) Flowchart of RUMI computation. Marginal and conditional entropies are computed empirically with bins. The filled cyan band is the area between the 97.5 and 2.5 percentiles of simulation estimated by BLUECAT.**

### 2.4 Calibration and validation strategies

The GR4J rainfall-runoff model calibration is conducted using the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) algorithm (Hansen et al., 2003; Hansen and Ostermeier, 1996). Catchments were calibrated with two different objective

functions: KGE and RUMI. KGE (Kling et al., 2012) – computed in this study with Eq. (5) – is the modified version of the KGE proposed initially by Gupta et al. (2009):

$$\mathrm{KGE} = 1 - \sqrt{\left(\frac{\mu_s}{\mu_o} - 1\right)^2 + \left(\frac{(\sigma_s/\mu_s)}{(\sigma_o/\mu_o)} - 1\right)^2 + (\rho - 1)^2}, \quad (5)$$

where, $\mu_s$ is the mean value of deterministic streamflow simulations; $\mu_o$ is the mean value of streamflow observations; $\sigma_s$ is the standard deviation of deterministic streamflow simulations; $\sigma_o$ is the standard deviation of streamflow observations; and, $\rho$ is the Pearson correlation coefficient between observed and deterministic simulation of streamflow.

Four additional metrics were used to assess performance of results: i) Nash-Sutcliffe Efficiency (NSE); ii) KGE knowable moments (KGEkm, Pizarro and Jorquera 2024); iii) Normalised Root Mean Squared Error (NRMSE); and, iv) Mean Absolute Relative Error (MARE). Equations for NSE, KGEkm, NRMSE, and MARE are presented from Eq. (6) to Eq. (9):

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(O_i - S_i)^2}{\sum_{i=1}^{n}(O_i - \mu_o)^2}, \quad (6)$$

$$\mathrm{KGEkm} = 1 - \sqrt{\left(\frac{\mathrm{K}_{1_S}}{\mathrm{K}_{1_O}} - 1\right)^2 + \left(\frac{\left(\sqrt{\mathrm{K}_{2_S}}/\mathrm{K}_{1_S}\right)}{\left(\sqrt{\mathrm{K}_{2_O}}/\mathrm{K}_{1_O}\right)} - 1\right)^2 + (\rho - 1)^2}, \quad (7)$$

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}(S_i - O_i)^2\right)}}{\max(O) - \min(O)}, \quad (8)$$

$$\mathrm{MARE} = \frac{\sum_{i=1}^{n}\left|\frac{(S_i - O_i)}{O_i}\right|}{n}, \quad (9)$$

where, $\mathrm{K}_{1_S}$ and $\mathrm{K}_{1_O}$ are the first knowable moment of simulated and observed streamflow time series, and $\mathrm{K}_{2_S}$ and $\mathrm{K}_{2_O}$ are dispersion relying on the second knowable moments of simulated and observed streamflow time series. Notice that the square operator in $\mathrm{K}_2$ is not necessary in Eq. (7) but intentionally used to be in line with classical statistics and KGE formulation (see Eq. 5). $S$ and $O$ mean simulated and observed streamflow time series, respectively. $n$ is the length of the analysed period (at daily scale). RMSE, NRMSE and MARE have 0 at the perfect ideal value, whereas their values range from 0 to positive infinite. NSE and KGEkm have a range from minus infinite to 1, being 1 the ideal value.

Additionally, and with a particular focus on different runoff characteristics, 50 hydrological signatures were computed. Observed runoff, simulations with model calibrated with KGE, and simulations with model calibrated with RUMI were considered. Hydrological signatures were computed with the Toolbox for Streamflow Signatures in Hydrology (TOSSH, Gnann et al. (2021)). Table 1 shows the 50 computed signatures.

**Table 1: 50 hydrological signatures computed with the Toolbox for Streamflow Signatures in Hydrology (TOSSH). The computed hydrological signatures follow TOSSH nomenclature (e.g., TotalRR is the total runoff ratio). A description of the signatures is also included.**

| Nº | Hydrological signature (using TOSSH nomenclature) | Description |
|---|---|---|
| 1 | Q_mean | Mean streamflow |
| 2 | TotalRR | Total runoff ratio |
| 3 | QP_elasticity | Streamflow-precipitation elasticity |
| 4 | FDC_slope | Slope of the flow duration curve |
| 5 | BFI | Baseflow index |
| 6 | HFD_mean | Half flow date |
| 7 | Q5 | $5^{th}$ streamflow percentile |
| 8 | Q95 | $95^{th}$ streamflow percentile |
| 9 | high_Q_freq | High flow frequency |
| 10 | high_Q_dur | High flow duration |
| 11 | low_Q_freq | Low flow frequency |
| 12 | low_Q_dur | Low flow duration |
| 13 | AC1 | Lag-1 autocorrelation |
| 14 | AC1_low | Lag-1 autocorrelation for low flow period |
| 15 | RLD | Rising limb density |
| 16 | PeakDistribution | Slope of distribution of peaks |
| 17 | PeakDistribution_low | Slope of distribution of peaks for low flow period |
| 18 | IE_effect | Infiltration excess importance |
| 19 | SE_effect | Saturation excess importance |
| 20 | IE_thresh_signif | Infiltration excess threshold significance (in a plot of quickflow volume vs. maximum intensity) |
| 21 | SE_thresh_signif | Saturation excess threshold significance (in a plot of quickflow volume vs. total precipitation) |
| 22 | IE_thresh | Infiltration excess threshold location (in a plot of quickflow volume vs. maximum intensity) |
| 23 | SE_thresh | Saturation excess threshold location (in a plot of quickflow volume vs. total precipitation) |
| 24 | SE_slope | Saturation excess threshold above-threshold slope (in a plot of quickflow volume vs. total precipitation) |

| | | |
|---|---|---|
| 25 | Storage_thresh_signif | Storage/saturation excess threshold significance (in a plot of quickflow volume vs. antecedent precipitation index + total precipitation) |
| 26 | Storage_thresh | Storage/saturation excess threshold location (in a plot of quickflow volume vs. antecedent precipitation index + total precipitation) |
| 27 | min_Qf_perc | Minimum quickflow as a percentage of precipitation |
| 28 | EventRR | Event runoff ratio |
| 29 | RR_Seasonality | Runoff ratio seasonality |
| 30 | Recession_a_Seasonality | Seasonal variations in recession parameters |
| 31 | AverageStorage | Average storage from average baseflow and storage-discharge relationship |
| 32 | MRC_num_segments | Number of different segments in master recession curve (MRC) |
| 33 | BaseflowRecessionK | Exponential recession constant |
| 34 | First_Recession_Slope | Steep section of MRC = storage that is quickly depleted |
| 35 | Spearmans_rho | Non-uniqueness in the storage-discharge relationship |
| 36 | EventRR_TotalRR_ratio | Ratio between event and total runoff ratio |
| 37 | VariabilityIndex | Variability index of flow |
| 38 | BaseflowMagnitude | Difference between maximum and minimum of annual baseflow regime |
| 39 | FlashinessIndex | Richards-Baker flashiness idex |
| 40 | HFI_mean | Half flow interval |
| 41 | Q_CoV | Coefficient of variation |
| 42 | Q_mean_monthly | Mean monthly streamflow |
| 43 | Q_7_day_max | 7-day maximum streamflow |
| 44 | Q_7_day_min | 7-day minimum streamflow |
| 45 | Q_skew | Skewness of streamflow |
| 46 | Q_var | Variance of streamflow |
| 47 | RecessionK_part | Recession constant of early/late (exponential) recessions |
| 48 | ResponseTime | Catchment response time |
| 49 | SnowStorage | Snow storage derived from cumulative P-Q regime curve |

| | | |
|---|---|---|
| 50 | StorageFromBaseflow | Average storage from average baseflow and storage-discharge relationship |

## 3 Results

Fig. 4 shows a graphical example of RUMI-based hydrological modelling of two of the catchments in calibration (Fig. 4a, catchment number: 8123001) and validation (Fig. 4b, catchment number: 9437002) over the years 1996 and 2016, respectively. Additionally, it shows observed and simulated stream flows, which were calibrated with KGE (red continuous line) and RUMI (blue continuous line is the mean of the stochastic simulation). 97.5 and 2.5 percentiles (computed with BLUECAT and RUMI) are shown with a violet band. Fig. 4a.2 and Fig. 4b.2 show observed and simulated stream flows over the complete period of analysis (performance of KGE-based simulations was 0.89 (0.80) and 0.95 (0.91) in calibration (validation) as well as the performance of RUMI-based simulations was 0.27 (0.20) and 0.46 (0.48) in calibration (validation), respectively). Worth to mention is that observed streamflow was between the 97.5 and 2.5 percentiles (i.e., the violet band) all the time except 4.93% and 0.19% of the time, presenting higher and lower observed streamflow, respectively (see, e.g., one event in June 1996 in Fig. 3a and one event in July 2016 in Fig. 3b).

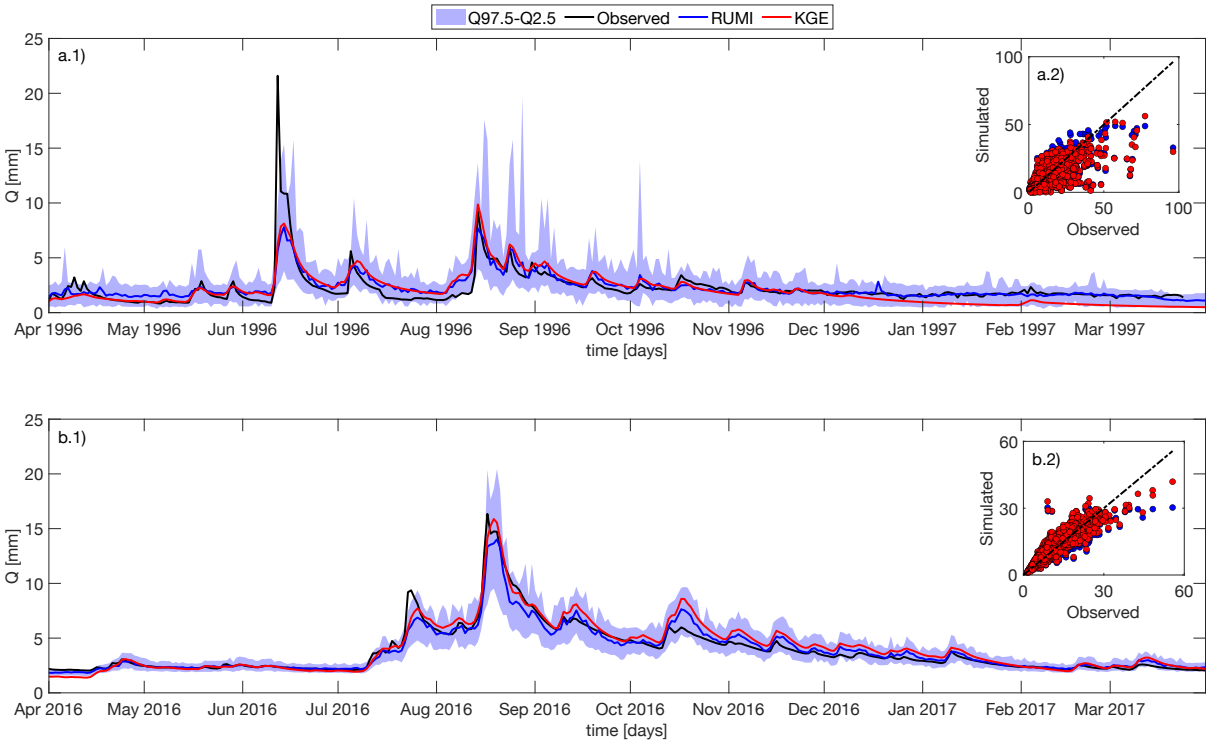


**Figure 4: Observed and simulated stream flows for the hydrological year 1996-1997 (a) and 2016-2017 (b). a.1) Catchment ID: 8123001 in calibration; b.1) Catchment ID: 9437002 in validation. Black: observed streamflow; Red: simulated by the deterministic model calibrated with KGE; Blue: simulated with the model calibrated with RUMI (mean stochastic simulation).**

**The filled violet band is the area between the 97.5 and 2.5 percentiles of simulation estimated by BLUECAT. The dashed line represents the perfect agreement between observed and simulated streamflow.**

In terms of other performance metrics, Fig. 5 shows NSE (a.1, b.1), KGEkm (a.2, b.2), NRMSE (a.3, b.3), and MARE (a.4, b.4) in calibration (a.1, a.2, a.3, a.4) and validation (b.1, b.2, b.3, b.4). Red markers are outliers, and grey dots represent the mean values (as a function of RUMI- and KGE-based simulations) which are linked with a grey line.

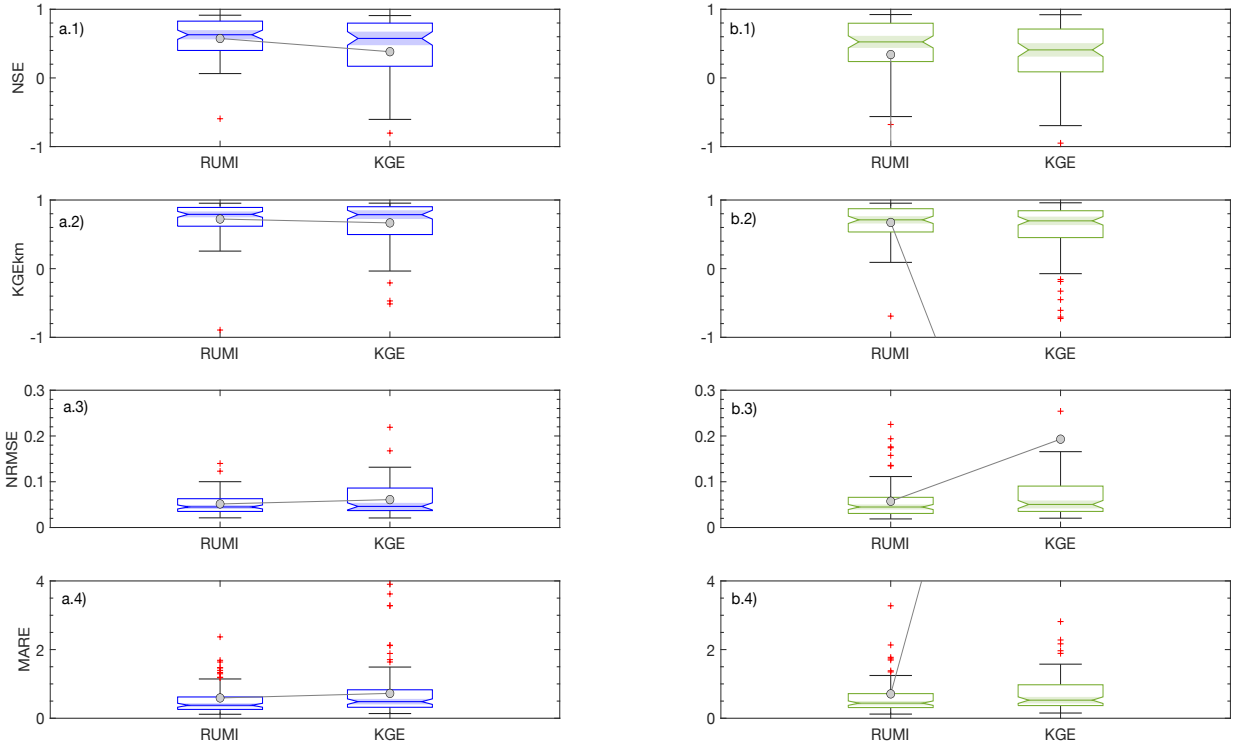


**Figure 5: Performance metrics in calibration (a.1, a.2, a.3, a.4) and validation (b.1, b.2, b.3, b.4). Red markers denote outliers. Grey dots represent the mean values computed with RUMI and KGE, which are linked to grey lines. Note that the y-axis limits are truncated for visualisation purposes.**

Remarkably, RUMI-based simulations outperform KGE-based ones in calibration and validation, and for the four performance metrics analysed. The latter in terms of variability (e.g., the interquartile range – IQR), median of boxplots, and number of outliers for both calibration and validation periods. Table 2 summarises the four considered performance metrics in terms of: a) calibration and validation; b) RUMI and KGE; and, c) minimum, maximum, median, IQR, and mean values.

**Table 2: Statistic's summary of boxplots results (see also Fig. 5).**

| | | Calibration | | | | Validation | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | NSE | KGEkm | NRMSE | MARE | NSE | KGEkm | NRMSE | MARE |
| Min | RUMI | -0.59 | -0.89 | 0.02 | 0.12 | -14.11 | -0.69 | 0.02 | 0.12 |
| | KGE | -1.80 | -0.51 | 0.02 | 0.14 | -299732 | -616 | 0.02 | 0.15 |
| Max | RUMI | 0.91 | 0.95 | 0.14 | 7.81 | 0.92 | 0.95 | 0.23 | 5.56 |
| | KGE | 0.91 | 0.95 | 0.22 | 3.90 | 0.92 | 0.96 | 12.58 | 1755 |
| Median | RUMI | 0.63 | 0.79 | 0.04 | 0.38 | 0.53 | 0.71 | 0.04 | 0.44 |
| | KGE | 0.58 | 0.79 | 0.05 | 0.48 | 0.41 | 0.70 | 0.05 | 0.53 |
| IQR | RUMI | 0.43 | 0.27 | 0.03 | 0.36 | 0.56 | 0.34 | 0.04 | 0.41 |
| | KGE | 0.63 | 0.41 | 0.05 | 0.51 | 0.62 | 0.39 | 0.06 | 0.61 |
| Mean | RUMI | 0.57 | 0.72 | 0.05 | 0.59 | 0.34 | 0.67 | 0.06 | 0.71 |
| | KGE | 0.38 | 0.67 | 0.06 | 0.72 | -3027 | -5.67 | 0.19 | 18.76 |

Based on Fig. 4 and Table 2, RUMI-based simulations showed more stable and consistent performance than KGE in calibration and validation phases. While KGE can achieve high accuracy (see, e.g., the maximum value of NSE for RUMI and KGE), it exhibits more variability and more extreme outliers (see, e.g., the minimum values of NSE: -14.11 vs -299732 for RUMI and KGE; the mean values of NSE: 0.34 vs -3027 for RUMI and KGE; the minimum values of KGEkm: -0.69 vs -616 for RUMI and KGE; the maximum values of NRMSE: 0.23 vs 12.58 for RUMI and KGE; and, the maximum values of MARE: 5.56 vs 1755 for RUMI and KGE). The latter, particularly during validation, indicates a lack of robustness. On the other hand, RUMI presented lower variability, more consistent results, and the opportunity to consider the confidence intervales in calibration.

Table 3 shows the Pearson's correlation coefficient for the 50 computed hydrological signatures considering observed and simulated streamflow data ("Obs vs KGE" means the Pearson's correlation coefficient using observed and simulated-with-KGE stream flows to compute any hydrological signature. "Obs vs RUMI" means the Pearson's correlation coefficient using observed and simulated-with-RUMI stream flows to compute any hydrological signature). On average, RUMI outperforms KGE-based simulations (average values: 0.72 vs 0.48) and minimum and maximum values (-0.07 vs -0.10 and 1.00 vs 0.96, respectively). RUMI-based simulations outperform KGE-based ones by 82% of the considered hydrologic signatures. Fig. 6 shows four examples of this comparison in terms of the runoff ratio (TotalRR, Fig. 6a), streamflow-precipitation elasticity (QP_elasticity, Fig. 6b); 5-th flow percentile of streamflow (Q5, Fig. 6c), and 95-th flow percentile of streamflow (Q95, Fig. 6d). Colours of the dots are related to the five different defined macroclimatic zones depicted in Fig. 1.

**Table 3: 50 used hydrological signatures. Performance was assessed using Pearson's correlation coefficient. Hydrological signatures were computed with TOSSH. Obs vs KGE means the Pearson's correlation coefficient using observed and simulated-with-KGE stream flows to compute any hydrological signature. Obs vs RUMI means the Pearson's correlation coefficient using observed and simulated-with-RUMI stream flows to compute any hydrological signature. The average, minimum, and maximum values were computed and added at the end of the list.**

| Hydrological signature | Obs versus KGE | Obs versus RUMI |
|---|---|---|
| Q_mean | 0.90 | 1.00 |
| TotalRR | -0.06 | 1.00 |
| QP_elasticity | 0.30 | 0.63 |
| FDC_slope | 0.30 | 0.86 |
| BFI | 0.74 | 0.83 |
| HFD_mean | 0.75 | 0.94 |
| Q5 | 0.96 | 0.99 |
| Q95 | 0.41 | 0.99 |
| high_Q_freq | 0.52 | 0.91 |
| high_Q_dur | 0.27 | 0.28 |
| low_Q_freq | 0.56 | 0.95 |
| low_Q_dur | -0.09 | 0.61 |
| AC1 | 0.67 | 0.69 |
| AC1_low | 0.61 | 0.59 |
| RLD | 0.16 | 0.15 |
| PeakDistribution | 0.28 | 0.76 |
| PeakDistribution_low | 0.07 | 0.57 |
| IE_effect | 0.53 | 0.51 |
| SE_effect | 0.68 | 0.67 |
| IE_thresh_signif | 0.63 | 0.50 |
| SE_thresh_signif | 0.51 | 0.41 |
| IE_thresh | -0.04 | 0.53 |
| SE_thresh | -0.06 | 0.65 |
| SE_slope | 0.71 | 0.72 |
| Storage_thresh_signif | 0.49 | 0.53 |
| Storage_thresh | -0.04 | 0.70 |
| min_Qf_perc | -0.02 | 0.63 |

| | | |
|---|---|---|
| EventRR | 0.96 | 0.98 |
| RR_Seasonality | 0.83 | 0.86 |
| Recession_a_Seasonality | 0.20 | 0.37 |
| AverageStorage | 0.72 | 0.87 |
| MRC_num_segments | -0.10 | -0.07 |
| BaseflowRecessionK | 0.33 | 0.65 |
| First_Recession_Slope | 0.34 | 0.40 |
| Spearmans_rho | 0.48 | 0.65 |
| EventRR_TotalRR_ratio | 0.85 | 0.97 |
| VariabilityIndex | 0.06 | 0.91 |
| BaseflowMagnitude | 0.95 | 0.97 |
| FlashinessIndex | 0.86 | 0.91 |
| HFI_mean | 0.63 | 0.86 |
| Q_CoV | 0.89 | 0.82 |
| Q_mean_monthly | 0.74 | 0.99 |
| Q_7_day_max | 0.77 | 0.94 |
| Q_7_day_min | -0.04 | 0.95 |
| Q_skew | 0.45 | 0.58 |
| Q_var | 0.10 | 0.98 |
| RecessionK_early | 0.82 | 0.67 |
| ResponseTime | 0.42 | 0.25 |
| SnowStorage | 0.95 | 0.98 |
| StorageFromBaseflow | 0.79 | 0.84 |
| **Average** | 0.48 | 0.72 |
| **Min** | -0.10 | -0.07 |
| **Max** | 0.96 | 1.00 |

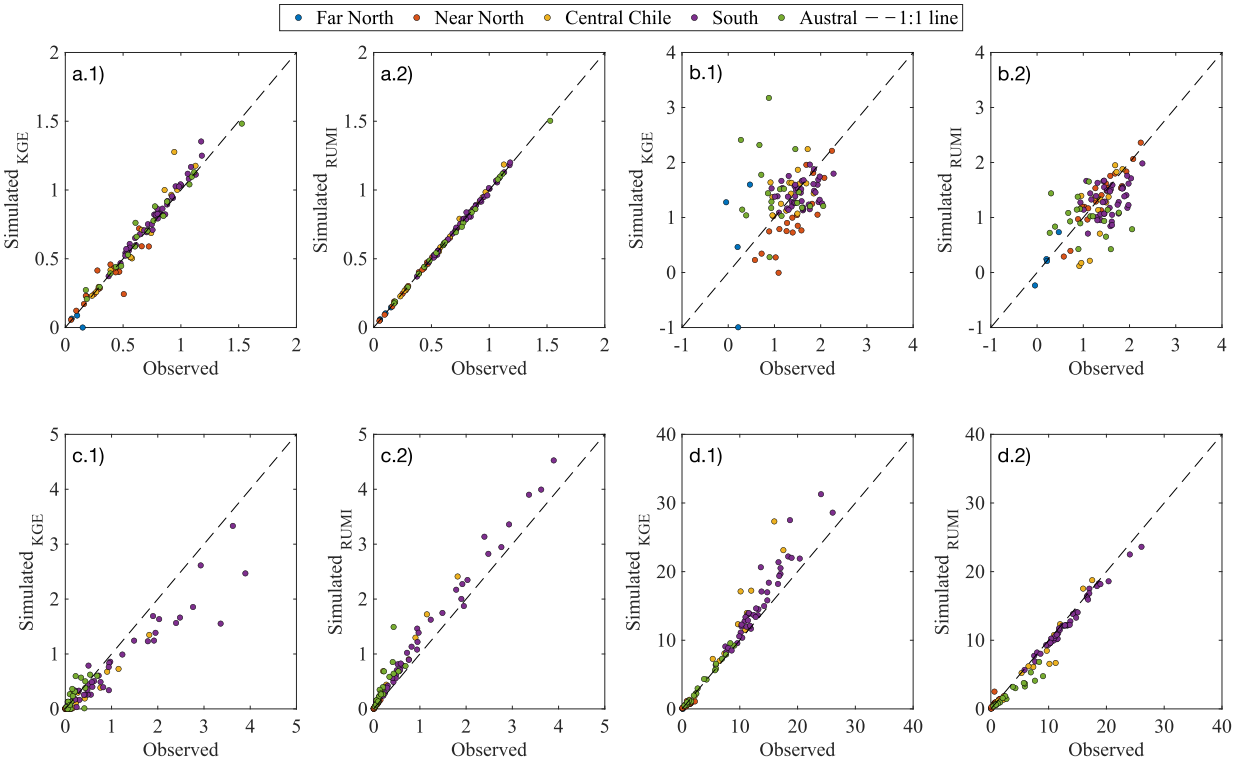


**Figure 6: Observed and simulated hydrological signatures for each case (a.1, b.1, c.1, d1: simulated with KGE; and, a.2, b.2, c.2, d.2: simulated with RUMI). a: runoff ratio (TotalRR); b: streamflow-precipitation elasticity (QP_elasticity); c: 5-th flow percentile of streamflow (Q5); d: 95-th flow percentile of streamflow (Q95). Colours of dots are related to the five considered macroclimatic zones. The dashed line represents the perfect agreement between observed and simulated hydrological signature. Note that the y-axis limits for the a.1 plot are truncated for visualisation purposes (original y-axis range: [0, 30]).**

## 4 Strengths and limitations

One of the main strengths of this study was the proposal of a new dimensionless metric to be used as objective function for rainfall-runoff model calibration. The proposed approach provides a comprehensive measure of the shared information between observed and simulated stream flows, normalises this measure for comparability, and integrates uncertainty quantification in the calibration process. The rescaling of the performance metric ensures intuitive interpretation (RUMI ranges between 0 and 1, being the latter the optimal value), aligning with standard efficiency metrics and making it easy to understand. This study presented a large-sample rainfall-runoff modelling experiment, analysing 99 catchments in a pseudo-natural hydrologic regime that covers five different macroclimatic zones and, therefore, giving robustness to the analysis. The latter ensures a diverse representation of hydrological characteristics and a broad evaluation of the RUMI-based modelling approach. The simplicity of the approach, its capacity to quantify confidence intervals and, therefore, also

uncertainty quantification are significant strengths. As demonstrated by the IQR, the median of results, and outliers (see Table 2), simulations during validation are also seen to improve (alongside with calibration results). Also, using the 50 hydrological signatures, the RUMI-based approach was compared considering different runoff dynamics characteristics showing improvements for most (82% of the analysed signatures showed a better correlation with observed data compared to KGE). RUMI-based performances rely on the combination of available information (in terms of observed quantities) and physically based consistency of modelled hydrological processes (BLUECAT alongside entropy-based computations and deterministic rainfall-runoff model). RUMI-based modelling implementation is also facilitated by the codes provided in this manuscript (see Code and Results availability statement), which enhances the reproducibility of the methodology.

In terms of limitations – and considering that RUMI considers uncertainty quantification in its computing process – we emphasise the fact that other methodologies for such purposes should be testing (such as multi-model ensemble methods or time-varying model parameters. See Gupta and Govindaraju 2023 for a recent review in this regard). The latter with the intention to quantify the sensibility of RUMI as a function of those additional methodologies. Additionally, RUMI calculations can be computationally intensive. The method's accuracy depends on high-quality input data and length of the time series (BLUECAT assumes that the calibration dataset is extended enough to upgrade from the deterministic to the stochastic model). It also assumes that observed and simulated stream flows can be effectively described by these measures, which may not capture all dependencies and non-linearities. Finally, entropy and mutual information might be sensitive to outliers.

## 5 Conclusions

The RUMI-based hydrological modelling approach outperforms KGE-based modelling in both calibration and validation phases across various performance metrics. This method demonstrates lower variability and a consistent performance improvement. RUMI's capability to quantify uncertainty and incorporate it into the calibration process ensures more reliable predictions. The analysis of hydrological signatures further confirms the superiority of RUMI, with 82% of the signatures showing a better correlation with observed data compared to KGE. RUMI offers a valuable tool for hydrological modelling, enhancing the understanding and prediction of streamflow under different hydrological conditions.

Possible additional research is mentioned as follows: (a) Testing the RUMI-based approach with other rainfall-runoff models (lumped, semi-distributed, and distributed hydrological models); (b) Testing the RUMI-based approach under other hydroclimatological catchment characteristics and in a higher number of catchments; (c) Testing alternative uncertainty quantification methods; (d) Exploring the impact of varying data quality on RUMI performance to establish guidelines for data requirements; (e) Testing with higher resolution data to reduce discretisation issues; and, (f) Exploring the applicability of RUMI in other disciplines such as meteorology, environmental science, and ecology where modelling and uncertainty quantification are critical.

## Code and Data availability

RUMI codification – in Matlab and R – is available in Pizarro et al. (2024): https://www.doi.org/10.17605/OSF.IO/93N4R. Data used in this study are available in the CAMELS-CL dataset (Alvarez-Garreton et al., 2018): https://doi.pangaea.de/10.1594/PANGAEA.894885

## Author contribution

AM and DK developed the BLUECAT code in R. AP developed RUMI codes and performed simulations. AP prepared the manuscript with contributions from all co-authors.

## Competing Interest

The authors declare that they have no conflict of interest.

## Financial support

AP was supported by The National Research and Development Agency of the Chilean Ministry of Science, Technology, Knowledge and Innovation (ANID), grant no. FONDECYT Iniciación 11240171. AM was partially supported by (1) the RETURN Extended Partnership which received funding from the European Union Next-GenerationEU (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$_2$Water)", grant number J53C23003860001. DK was not supported at all.

## References

Acuña, P. and Pizarro, A.: Can continuous simulation be used as an alternative for flood regionalisation? A large sample example from Chile, Journal of Hydrology, 626, 130118, https://doi.org/10.1016/j.jhydrol.2023.130118, 2023.

Alexander, A. A., Kumar, D. N., Knoben, W. J. M., and Clark, M. P.: Evaluating the parameter sensitivity and impact of hydrologic modeling decisions on flood simulations, Advances in Water Resources, 181, 104560, https://doi.org/10.1016/j.advwatres.2023.104560, 2023.

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., and Garreaud, R.: The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies–Chile dataset, Hydrology and Earth System Sciences, 22, 5817–5846, 2018.

Amorocho, J. and Espildora, B.: Entropy in the assessment of uncertainty in hydrologic systems and models, Water Resources Research, 9, 1511–1522, 1973.

Auer, A., Gauch, M., Kratzert, F., Nearing, G., Hochreiter, S., and Klotz, D.: A data-centric perspective on the information needed for hydrological uncertainty predictions, Hydrology and Earth System Sciences, 28, 4099–4126, https://doi.org/10.5194/hess-28-4099-2024, 2024.

Bai, Z., Wu, Y., Ma, D., and Xu, Y.-P.: A new fractal-theory-based criterion for hydrological model calibration, Hydrology and Earth System Sciences, 25, 3675–3690, https://doi.org/10.5194/hess-25-3675-2021, 2021.

Barber, C., Lamontagne, J. R., and Vogel, R. M.: Improved estimators of correlation and R2 for skewed hydrologic data, Hydrological Sciences Journal, 65, 87–101, https://doi.org/10.1080/02626667.2019.1686639, 2020.

Beven, K.: Prophecy, reality and uncertainty in distributed hydrological modelling, Advances in water resources, 16, 41–51, 1993.

Beven, K.: A manifesto for the equifinality thesis, Journal of Hydrology, 320, 18–36, https://doi.org/10.1016/j.jhydrol.2005.07.007, 2006.

Beven, K.: Environmental modelling: an uncertain future?, CRC press, 2018.

Beven, K.: A brief history of information and disinformation in hydrological data and the impact on the evaluation of hydrological models, Hydrological Sciences Journal, 69, 519–527, 2024.

Beven, K. and Binley, A.: The future of distributed models: model calibration and uncertainty prediction, Hydrological processes, 6, 279–298, 1992.

Beven, K. and Lane, S.: On (in) validating environmental models. 1. Principles for formulating a Turing-like Test for determining when a model is fit-for purpose, Hydrological Processes, 36, e14704, 2022.

Beven, K., Page, T., Smith, P., Kretzschmar, A., Hankin, B., and Chappell, N.: UPH Problem 20–reducing uncertainty in model prediction: a model invalidation approach based on a Turing-like test, Proceedings of IAHS, 385, 129–134, 2024.

Blasone, R.-S., Vrugt, J. A., Madsen, H., Rosbjerg, D., Robinson, B. A., and Zyvoloski, G. A.: Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling, Advances in Water Resources, 31, 630–648, https://doi.org/10.1016/j.advwatres.2007.12.003, 2008.

Blazkova, S. and Beven, K.: Flood frequency estimation by continuous simulation for a catchment treated as ungauged (with uncertainty), Water Resources Research, 38, 14-1-14–14, https://doi.org/10.1029/2001WR000500, 2002.

Blazkova, S. and Beven, K.: Flood frequency estimation by continuous simulation of subcatchment rainfalls and discharges with the aim of improving dam safety assessment in a large basin in the Czech Republic, Journal of Hydrology, 292, 153–172, 2004.

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J. W., McDonnell, J. J., Savenije, H. H., and Sivapalan, M.: Twenty-three unsolved problems in hydrology (UPH)–a community perspective, Hydrological sciences journal, 64, 1141–1158, 2019.

Chapman, T. G.: Entropy as a measure of hydrologic data uncertainty and model performance, Journal of Hydrology, 85, 111–126, 1986.

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., and Shook, K. R.: The abuse of popular performance metrics in hydrologic modeling, Water Resources Research, 57, e2020WR029001, 2021.

Freer, J. E., McMillan, H., McDonnell, J., and Beven, K.: Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures, Journal of Hydrology, 291, 254–277, 2004.

Garcia, F., Folton ,Nathalie, and and Oudin, L.: Which objective function to calibrate rainfall–runoff models for low-flow index simulations?, Hydrological Sciences Journal, 62, 1149–1166, https://doi.org/10.1080/02626667.2017.1308511, 2017.

Garreaud, R.: The Andes climate and weather, Advances in Geosciences, 22, 3–11, 2009.

Gnann, S. J., Coxon, G., Woods, R. A., Howden, N. J. K., and McMillan, H. K.: TOSSH: A Toolbox for Streamflow Signatures in Hydrology, Environmental Modelling & Software, 138, 104983, https://doi.org/10.1016/j.envsoft.2021.104983, 2021.

Gong, W., Gupta, H. V., Yang, D., Sricharan, K., and Hero III, A. O.: Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach, Water Resources Research, 49, 2253–2273, https://doi.org/10.1002/wrcr.20161, 2013.

Gong, W., Yang, D., Gupta, H. V., and Nearing, G.: Estimating information entropy for hydrological data: One-dimensional case, Water Resources Research, 50, 5003–5018, https://doi.org/10.1002/2014WR015874, 2014.

Gupta, A. and Govindaraju, R. S.: Uncertainty quantification in watershed hydrology: Which method to use?, Journal of Hydrology, 616, 128749, https://doi.org/10.1016/j.jhydrol.2022.128749, 2023.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hansen, N. and Ostermeier, A.: Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation, Proceedings of IEEE international conference on evolutionary computation, 312–317, 1996.

Hansen, N., Müller, S. D., and Koumoutsakos, P.: Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES), Evolutionary computation, 11, 1–18, 2003.

Hemelrijk, J.: Underlining random variables, Statistica Neerlandica, 20, 1–7, 1966.

Hundecha, Y. and Bárdossy, A.: Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model, Journal of Hydrology, 292, 281–295, https://doi.org/10.1016/j.jhydrol.2004.01.002, 2004.

Jackson, E. K., Roberts, W., Nelsen, B., Williams, G. P., Nelson, E. J., and Ames, D. P.: Introductory overview: Error metrics for hydrologic modelling – A review of common practices and an open source library to facilitate use and adoption, Environmental Modelling & Software, 119, 32–48, https://doi.org/10.1016/j.envsoft.2019.05.001, 2019.

Jorquera, J. and Pizarro, A.: Unlocking the potential of stochastic simulation through Bluecat: Enhancing runoff predictions in arid and high-altitude regions, Hydrological Processes, 37, e15046, https://doi.org/10.1002/hyp.15046, 2023.

Kennedy, M. C. and O'Hagan, A.: Bayesian calibration of computer models, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63, 425–464, 2001.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424–425, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

Knoben, W. J. M., Freer, J. E., Fowler, K. J. A., Peel, M. C., and Woods, R. A.: Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations, Geoscientific Model Development, 12, 2463–2480, https://doi.org/10.5194/gmd-12-2463-2019, 2019.

Koutsoyiannis, D.: When Are Models Useful? Revisiting the Quantification of Reality Checks, Water, 17, 264, https://doi.org/10.3390/w17020264, 2025.

Koutsoyiannis, D. and Montanari, A.: Bluecat: A local uncertainty estimator for deterministic simulations and predictions, Water Resources Research, 58, e2021WR031215, 2022a.

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

Krzysztofowicz, R.: Bayesian system for probabilistic river stage forecasting, Journal of hydrology, 268, 16–40, 2002.

Kuczera, G., Kavetski, D., Franks, S., and Thyer, M.: Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, Journal of hydrology, 331, 161–177, 2006.

Lamontagne, J. R., Barber, C. A., and Vogel, R. M.: Improved Estimators of Model Performance Efficiency for Skewed Hydrologic Data, Water Resources Research, 56, e2020WR027101, https://doi.org/10.1029/2020WR027101, 2020.

Lin, F., Chen, X., and Yao, H.: Evaluating the Use of Nash-Sutcliffe Efficiency Coefficient in Goodness-of-Fit Measures for Daily Runoff Simulation with SWAT, Journal of Hydrologic Engineering, 22, 05017023, https://doi.org/10.1061/(ASCE)HE.1943-5584.0001580, 2017.

Liu, D.: A rational performance criterion for hydrological model, Journal of Hydrology, 590, 125488, https://doi.org/10.1016/j.jhydrol.2020.125488, 2020.

Melsen, L. A., Teuling, A. J., Torfs, P. J. J. F., Zappa, M., Mizukami, N., Mendoza, P. A., Clark, M. P., and Uijlenhoet, R.: Subjective modeling decisions can significantly impact the simulation of flood and drought events, Journal of Hydrology, 568, 1093–1104, https://doi.org/10.1016/j.jhydrol.2018.11.046, 2019.

Melsen, L. A., Puy ,Arnald, Torfs ,Paul J. J. F., and and Saltelli, A.: The rise of the Nash-Sutcliffe efficiency in hydrology, Hydrological Sciences Journal, 0, 1–12, https://doi.org/10.1080/02626667.2025.2475105, n.d.

Mendoza, P. A., Clark, M. P., Mizukami, N., Gutmann, E. D., Arnold, J. R., Brekke, L. D., and Rajagopalan, B.: How do hydrologic modeling decisions affect the portrayal of climate change impacts?, Hydrological Processes, 30, 1071–1095, https://doi.org/10.1002/hyp.10684, 2016.

Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V., and Kumar, R.: On the choice of calibration metrics for "high-flow" estimation using hydrologic models, Hydrology and Earth System Sciences, 23, 2601–2614, https://doi.org/10.5194/hess-23-2601-2019, 2019.

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

Montanari, A. and Koutsoyiannis, D.: A blueprint for process-based modeling of uncertain hydrological systems, Water Resources Research, 48, 2012.

Montanari, A. and Koutsoyiannis, D.: Uncertainty estimation for environmental multimodel predictions: The BLUECAT approach and software, Environmental Modelling & Software, 106419, 2025.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Onyutha, C.: A hydrological model skill score and revised R-squared, Hydrology Research, 53, 51–64, 2022.

Page, T., Smith, P., Beven, K., Pianosi, F., Sarrazin, F., Almeida, S., Holcombe, L., Freer, J., Chappell, N., and Wagener, T.: The CREDIBLE Uncertainty Estimation (CURE) toolbox: facilitating the communication of epistemic uncertainty, Hydrology and Earth System Sciences, 27, 2523–2534, 2023.

Pechlivanidis, I. G., Jackson, B., McMillan, H., and Gupta, H.: Use of an entropy-based metric in multiobjective calibration to improve model performance, Water Resources Research, 50, 8066–8083, https://doi.org/10.1002/2013WR014537, 2014.

Pechlivanidis, I. G., Jackson, B., Mcmillan, H., and Gupta, H. V.: Robust informational entropy-based descriptors of flow in catchment hydrology, Hydrological Sciences Journal, 61, 1–18, https://doi.org/10.1080/02626667.2014.983516, 2016.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of hydrology, 279, 275–289, 2003.

Pizarro, A. and Jorquera, J.: Advancing objective functions in hydrological modelling: Integrating knowable moments for improved simulation accuracy, Journal of Hydrology, 634, 131071, https://doi.org/10.1016/j.jhydrol.2024.131071, 2024.

Pool, S., Vis, M., and Seibert, J.: Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency, Hydrological Sciences Journal, 63, 1941–1953, https://doi.org/10.1080/02626667.2018.1552002, 2018.

Pushpalatha, R., Perrin, C., Moine, N. L., and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, Journal of Hydrology, 420–421, 171–182, https://doi.org/10.1016/j.jhydrol.2011.11.055, 2012.

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

Ruddell, B. L., Drewry, D. T., and Nearing, G. S.: Information Theory for Model Diagnostics: Structural Error is Indicated by Trade-Off Between Functional and Predictive Performance, Water Resources Research, 55, 6534–6554, https://doi.org/10.1029/2018WR023692, 2019.

Sarricolea, P., Herrera-Ossandon, M., and Meseguer-Ruiz, Ó.: Climatic regionalisation of continental Chile, Journal of Maps, 13, 66–73, https://doi.org/10.1080/17445647.2016.1259592, 2017.

Shannon, C. E.: A mathematical theory of communication, The Bell system technical journal, 27, 379–423, 1948.

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

Tang, G., Clark, M. P., and Papalexiou, S. M.: SC-Earth: A Station-Based Serially Complete Earth Dataset from 1950 to 2019, Journal of Climate, 34, 6493–6511, https://doi.org/10.1175/JCLI-D-21-0067.1, 2021.

Thirel, G., Santos, L., Delaigue, O., and Perrin, C.: On the use of streamflow transformations for hydrological model calibration, Hydrology and Earth System Sciences, 28, 4837–4860, https://doi.org/10.5194/hess-28-4837-2024, 2024.

Thomas, M. and Joy, A. T.: Elements of information theory, Wiley-Interscience, 2006.

Trotter, L., Knoben, W. J. M., Fowler, K. J. A., Saft, M., and Peel, M. C.: Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v2.1: an object-oriented implementation of 47 established hydrological models for improved speed and readability, Geoscientific Model Development, 15, 6359–6369, https://doi.org/10.5194/gmd-15-6359-2022, 2022.

Vrugt, J. A. and Beven, K. J.: Embracing equifinality with efficiency: Limits of Acceptability sampling using the DREAM (LOA) algorithm, Journal of Hydrology, 559, 954–971, 2018.

Vrugt, J. A. and de Oliveira, D. Y.: Confidence intervals of the Kling-Gupta efficiency, Journal of Hydrology, 612, 127968, https://doi.org/10.1016/j.jhydrol.2022.127968, 2022.

Vrugt, J. A., Ter Braak, C. J., Clark, M. P., Hyman, J. M., and Robinson, B. A.: Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, Water Resources Research, 44, 2008.

Vrugt, J. A., de Oliveira, D. Y., Schoups, G., and Diks, C. G.: On the use of distribution-adaptive likelihood functions: Generalized and universal likelihood functions, scoring rules and multi-criteria ranking, Journal of Hydrology, 615, 128542, 2022.

Weijs, S. V., Van Nooijen, R., and Van De Giesen, N.: Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition, Monthly Weather Review, 138, 3387–3399, 2010a.

Weijs, S. V., Schoups, G., and van de Giesen, N.: Why hydrological predictions should be evaluated using information theory, Hydrology and Earth System Sciences, 14, 2545–2558, https://doi.org/10.5194/hess-14-2545-2010, 2010b.

Ye, L., Gu, X., Wang, D., and Vogel, R. M.: An unbiased estimator of coefficient of variation of streamflow, Journal of Hydrology, 594, 125954, https://doi.org/10.1016/j.jhydrol.2021.125954, 2021.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resources Research, 44, https://doi.org/10.1029/2007WR006716, 2008.