https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

Improving low and high flow simulations at once: An enhanced metric for hydrological model calibration

Andrea Ficchì[1,⋆], Davide Bavera[2,⋆], Stefania Grimaldi[3], Francesca Moschini[3,4], Alberto Pistocchi[3], Carlo Russo[5], Peter Salamon[3], and Andrea Toreti[3]

[1]Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy
[2]Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC), European Institute on Economics and the Environment (EIEE), Milan, Italy
[3]European Commission, Joint Research Centre, Ispra, Italy
[4]Rey Juan Carlos University, Madrid, Spain
[5]Unisystems Luxembourg Sàrl, Bertrange, Luxembourg
⋆These authors contributed equally to this work.

Correspondence: Andrea Ficchì (andrea.ficchi@polimi.it) and Andrea Toreti (andrea.toreti@ec.europa.eu)

Abstract. The choice of an objective function for hydrological model calibration is a critical step that directly influences model performance and suitability for the intended use cases. While calibration functions should ideally be tailored to specific modeling objectives, such as flood forecasting or drought monitoring, general-purpose metrics are typically used in practice. The two most widely adopted objective functions are the Nash–Sutcliffe Efficiency (NSE) and the Kling–Gupta Efficiency (KGE). While the NSE is a simple normalization of the mean square error, the KGE overcomes some of the NSE limitations and is often preferred due to its decomposable structure, capturing bias, relative variability, and correlation. However, KGE still suffers from limitations, including sensitivity to outliers and assumptions of linearity and normality in the error distribution, which particularly limit performance under low-flow conditions. Although several alternatives to NSE and KGE have been proposed, none has clearly outperformed these standard metrics across the full flow duration curve (FDC), especially for improving low flows without degrading performance elsewhere. To address these limitations, we propose a new metric, the Joint Divergence Kling-Gupta Efficiency (JDKGE), that enhances the KGE by incorporating an additional component based on the Jensen–Shannon Divergence (JSD). We evaluate the JDKGE metric using two hydrological process-based models (GR6J and OS-LISFLOOD), applied to two large and diverse samples of catchments spanning a broad range of hydroclimatic conditions. Calibrated using a suite of objective functions, both models are then evaluated with multiple performance metrics, including KGE, JSD, quantile ratios, and FDC-based signatures. Results show that calibrations using JDKGE significantly improve low-flow simulations compared to KGE, NSE and other competitors, while maintaining comparable or improved performance in other regimes, including high flows. Multi-objective calibration experiments further reveal that substantial gains in distributional similarity (i.e., reductions in JSD) can be achieved with only marginal changes in overall performance (KGE). Moreover, the JDKGE objective function leads to a balanced compromise between KGE and JSD and a reduction in model equifinality. This study highlights the importance of carefully selecting the objective function for hydrological model calibration and proposes JDKGE as an effective solution for improving low-flow performance while retaining general-purpose applicability for floods and water management.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

1 Introduction

Effective and widely accepted numerical performance metrics are essential in environmental modeling, as they provide a quantitative basis to evaluate model accuracy and other performance aspects, guide model developments, and enable intercomparison and benchmarking across studies (e.g., Rykiel, 1996; Bennett et al., 2013; Hipsey et al., 2020; Gauch et al., 2023). In hydrology, system-scale performance metrics assess the goodness of fit of a modeled hydrograph (i.e., streamflow over time) with respect to the corresponding observed hydrograph (Beven, 2025; Clark et al., 2021). These metrics are generally used as objective functions in automatic optimization algorithms for model calibration, aiming to find the best set of model parameters. As many different metrics exist and several variants of the most popular ones can be found, model users and developers should ideally choose the most suitable one based on their specific goals and applications (e.g., Krause et al., 2005; Mizukami et al., 2019). In particular, when used for model calibration, tailored objective functions are of utmost importance as they aim to maximise the performance of the calibrated model for the relevant application. For example, the most suitable objective function for a hydrological model used solely for flood forecasting should differ from the one to be used with a model designed exclusively for low-flow analysis (Pushpalatha et al., 2012). However, in practice, most hydrological models are used for multiple applications and the same standard general-purpose function is adopted for their calibration, regardless of the use case.

Among many existing metrics, the Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970) and, more recently, the Kling-Gupta Efficiency (KGE; Gupta et al., 2009; Kling et al., 2012) have both risen as the "gold-standard", or at least the most popular, general-purpose metrics for model evaluation and calibration in hydrology (Klotz et al., 2024; Melsen et al., 2025). While the NSE is a normalized version of the common Mean Square Error (MSE), which has been used in many fields for decades, KGE is a related, more recent metric, rewritten by decomposing the MSE into three components: correlation, bias, and relative variability following the work of Murphy (1988). The components of the MSE that are used in the KGE can be seen as a measure of linear dependence, of the nonsystematic (i.e., conditional) bias, and systematic (unconditional) bias between model outputs and observations. Because of this decomposability and interpretability, in recent years KGE has become *de facto* the standard in hydrological modeling studies for both calibration and evaluation, probably surpassing the NSE in popularity (Melsen et al., 2025). Moreover, its use has begun to expand beyond hydrology; for example, it has been used in water quality and ecosystems modeling (e.g., Fu and Zhang, 2024) as well as in climate modeling and downscaling (e.g., Admasu et al., 2023).

The limitations of NSE and KGE have been discussed by several authors (e.g., Schaefli and Gupta, 2007; Clark et al., 2021; Fu and Zhang, 2024). From a mathematical and statistical point of view, these limitations include several interlinked issues stemming from their formulation: (i) unrealistic assumptions of linearity and normality (e.g., Fu and Zhang, 2024), (ii) extreme sensitivity to outliers (e.g., Clarke, 1973), (iii) substantial sampling uncertainty in the metrics estimates, due to the heavy-tailed nature of model residuals (e.g., Clark et al., 2021), (iv) the neglect of the well-known statistical properties of model residuals, including their auto-correlation and heteroscedasticity (e.g., Sorooshian and Dracup, 1980; Evin et al., 2014), (v) the proneness to counterbalancing errors (Cinkus et al., 2023), (vi) the large sensitivity to the sample size (e.g., Pushpalatha et al., 2012), and (vii) the so-called "divide and measure nonconformity", for which the overall score (NSE or KGE) of a dataset is not bounded

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

by the same scores (NSE or KGE) computed on all its partitions (Klotz et al., 2024). Some of these aspects and limitations have been known for decades, but are still overlooked in most of the hydrological literature and in the general practice. For example, Clarke (1973) discussed the problem of unrealistic assumptions of independent Gaussian residuals that underlies least-squares objective functions, like NSE, which are generally invalid for hydrological models. This critique, which came out only three years after the paper defining the NSE, has been mostly neglected, possibly due to the simplicity of the interpretation of the NSE, as reported and put into perspective by Beven (2025). Already more than a decade ago, some authors (e.g., Weijs et al., 2010; Nearing and Gupta, 2015) advocated for the use of information theory principles and metrics to replace these popular least-squares-based standards. However, these proposals have not been successful in establishing new standards or popular metrics that could be taken up by the hydrological modeling community and compete with the KGE and NSE benchmarks. According to the recent study by Melsen et al. (2025) the primacy of NSE and KGE owes more to a fortuitous historical momentum and social factors than to any proven technical superiority with respect to other metrics. The early adoption of NSE to calibrate widely used models, its widespread use in benchmark studies (which reached a critical mass), the convenience of comparability across models and the similarity of the KGE with NSE created a feedback loop where these two metrics became default choices, reinforcing their status through repetition.

One of the most concerning consequences of the overlooked limitations of these standard objective functions in hydrology is their poor capacity in characterising hydrological model performance on low-flows (e.g., Pool et al., 2018; Garcia et al., 2017, among others). It is well recognised that current general-purpose calibration metrics such as the KGE and NSE underperform in low-flow conditions, as they weigh more the model errors during high flow periods (both timing and magnitude errors). For example, Lin et al. (2017) demonstrate that NSE is not sensitive to poor model performance during droughts, as better NSE values can be obtained while degrading a model's ability to simulate minimum daily or 7-day flows. Despite the well-known poor representativity of NSE and KGE for droughts, the use of these scores for such applications remains widespread (Melsen et al., 2025). This "accepted" bias towards high flows is likely rooted in the historical major focus of hydrological modeling on flood prediction. With growing attention to drought risk assessment and forecasting (Baatz et al., 2025), driven by climate change, increasing pressures from water use and demand, there is an urgent need for new improved metrics and objective functions that would better capture low-flow behavior (e.g., Pushpalatha et al., 2012; Pool et al., 2018).

Different strategies have been proposed in the literature to improve low-flow simulations by acting on the model calibration, especially on the objective function. Following Garcia et al. (2017), these efforts can be categorised into three avenues:

1. Modified objective functions: some authors propose modified versions of existing objective functions to improve low-flow simulations. For example, non-parametric variants of the KGE (Pool et al., 2018) and FDC-based objective functions (e.g., Westerberg et al., 2011; Price et al., 2012) have been proposed. Other authors suggest calibrating models directly on user-relevant low-flow indices (e.g., Olsen et al., 2013) or use-case relevant streamflow characteristics (e.g., Hallouin et al., 2020) or combinations of these with traditional metrics (e.g., Pool et al., 2017), but selecting them remains very subjective and challenging. No such solution has provided improvements on low-flows with non negligible performance deterioration (or let alone improvements) on high flows or across the full range of the FDC yet.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

2. Transformation of flows within traditional general-purpose metrics: transforming river flow values before computing traditional metrics like KGE or NSE can emphasize low-flow values (Thirel et al., 2024). Examples include the use of log-transformed flows (Oudin et al., 2006; Seeger and Weiler, 2014) or inverse flow values (Pushpalatha et al., 2012; Knoben et al., 2020). Such transformations bring technical issues related to zero-flow values, and the need to introduce workarounds such as adding a small positive constant (Pushpalatha et al., 2012) or using Box–Cox transformations. Additionally, Santos et al. (2018) showed that KGE exhibits erratic behavior when small constants are introduced. Alternative methods, like subsetting model errors on chosen flow ranges (Deckers et al., 2010) or weighting the errors (Krause et al., 2005), can be considered, but are rarely applied.

3. Multi-objective calibration: this approach combines multiple objective functions to explore trade-offs among different objectives and find compromise solutions (e.g., Madsen, 2000). For example, by using the KGE and an entropy-based metric Pechlivanidis et al. (2014) found better results to balance performance on low and high flows than single-objective calibration (with traditional metrics using either raw or log-transformed values). Formal multi-objective optimization methods for Pareto-front analysis (e.g., Fenicia et al., 2007; Monteil et al., 2020) or simpler strategies aggregating multiple criteria (Oudin et al., 2006; Booij and Krol, 2010; Nicolle et al., 2014) can be used, but selecting optimal weights and combinations remains a challenge (Vis et al., 2015; Khu and Madsen, 2005).

The literature also offers examples of combined use of the approaches described above. For example, Garcia et al. (2017) test combinations of the first two approaches, finding that while results are not robust when using objective functions based on the simulation of low-flow indices (e.g., annual minimum monthly discharge) or a single transformation of flows within general-purpose scores (KGE), using the mean of the KGE applied to the discharge and to its inverse values offers a good compromise solution to improve the simulation of low-flow indices, while keeping good overall water balance (mean annual runoff). This combined objective function, also used in a previous study to calibrate the GR6J model (Nicolle et al., 2014), aims to balance the effects of its two components. However, the evaluation focused only on low-flows, regimes and seasonality, neglecting high-flows.

Despite several years of research on this topic, there is still a clear need in hydrology for metrics that improve model calibration for low flows or drought conditions. This improvement for low flows should be brought at no cost for high flows, preserving or even improving performance along the whole Flow Duration Curve (FDC) for general usability of models, which are often used for multiple purposes (e.g., water resource management, drought and flood monitoring and forecasting). For instance, the OS–LISFLOOD model (Knijff et al., 2010) is currently calibrated using KGE (on streamflow values), and yet deployed for operational services for both floods and droughts under the Copernicus Emergency Management Service (CEMS; https://emergency.copernicus.eu). Evidence from recent model intercomparison studies suggests that substantial room for improvement remains in the calibration of widely used hydrological models (like OS-LISFLOOD or GR models), particularly for specific flow regimes, like high- or low-flows (e.g., Cantoni et al., 2022; Chang et al., 2024; Orth et al., 2015; Nicolle et al., 2014). For example, Chang et al. (2024) compared the performance of GR6J and LISFLOOD, as implemented in the CEMS European Flood Awareness System (EFAS), for low-flow and drought assessment at 34 Alpine catchments. Their results high-

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

light the need for a better calibration of OS-LISFLOOD, especially as low flow conditions intensify, and makes well-calibrated models for drought conditions even more important.

While calibration is often guided by performance in streamflow simulations, it would be important to look at all the components of the water balance. Such an approach is expected to be more robust under evolving climate conditions and different land-use scenarios (e.g., Liu et al., 2024). This may bring benefits for monitoring and management of water resources as well as testing and optimizing adaptation solutions (e.g., Giuliani et al., 2016). However, only a few studies have tried to calibrate on other variables beyond streamflow, e.g., soil moisture (e.g., Li et al., 2018; Mei et al., 2023), groundwater levels (e.g., Pelletier and Andréassian, 2022), actual evapotranspiration (e.g., Mei et al., 2023), or total water storage (e.g., Pool et al., 2024; Döll et al., 2024). Despite following a traditional calibration approach based on streamflow, we here argue that an enhanced objective function can better constrain the hydrological model and contribute to improve the representation of all the relevant hydrological processes.

To overcome the aforementioned gaps, here we introduce a new metric to be used as calibration objective function, following the first category of approaches defined by Garcia et al. (2017) and summarized above, with the aim to propose an enhanced general-purpose metric. We achieve this goal by augmenting the traditional KGE components with an additional metric based on information theory, following a direction advocated in previous literature (e.g., Weijs et al., 2010). The additional component is based on the Jensen-Shannon Divergence (JSD; Lin, 1991) that can measure the similarity between two probability distributions via a bounded and symmetric score (e.g., Squicciarini et al., 2025). To prove the effectiveness of our solution and its robustness, we calibrate two different well-known hydrological models (OS-LISFLOOD and GR6J) and validate across two large samples of diverse catchments (240 in France and 45 across the world) using multiple metrics. We carry out an extensive comparison of the hydrological models calibrated with several alternative objective functions, including standard benchmarks and recent variants, evaluating their performance across the flow spectrum.

2 Methods

2.1 The new objective function: JDKGE

We propose an enhanced calibration objective function, the Joint Divergence Kling-Gupta Efficiency (JDKGE), obtained by integrating a divergence component, based on the Jensen-Shannon Divergence (JSD; Lin, 1991), into an augmented (modified) Kling-Gupta Efficiency (KGE′; Kling et al., 2012). The aim is to improve model calibration by targeting enhanced accuracy over low flows, while maintaining appropriate weight on the entire distribution including flood peaks. To strengthen model performance across the full streamflow spectrum, JSD was selected as it measures the similarity between two probability distributions; here, the empirical distributions of observed and simulated flows need to be compared and the JSD can serve this purpose. Defined by Lin (1991) based on Jensen's inequality and the Shannon entropy, JSD addresses the key limitations of the Kullback-Leibler Divergence (KLD; Kullback and Leibler, 1951). The JSD is particularly attractive due to its symmetry, boundedness, and smoothness (Briët and Harremoës, 2009; Squicciarini et al., 2025). Unlike KLD, JSD is always finite and

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

provides a more robust basis for evaluating differences between empirical distributions with zero-probability regions, which are common when comparing observed and simulated hydrological time series (as during calibration), which may have non-overlapping or partially disjoint supports (leading to zero-probability regions for at least one distribution). Applications of JSD beyond hydrology demonstrate its effectiveness in evaluating distributional (dis-)similarity. For example, it has been employed in climate science (e.g., Winderlich et al., 2024), in medical research (e.g., Yan et al., 2021), in bioinformatics / genetic research (e.g., Guo, 2020), in ecology (e.g., Netzel et al., 2024; Verbruggen et al., 2018), in digital soil mapping (Saurette et al., 2023), and in machine learning (ML; e.g., Squicciarini et al., 2025). In hydrology, only a few examples can be found, where JSD was used to compare hydrologic responses across experiments (Nicótina et al., 2008) or to compare the similarity of hydrological signatures across catchments (Loritz et al., 2019).

The original KGE (Gupta et al., 2009; Kling et al., 2012) is defined as:

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\alpha-1)^2} \tag{1}$$

where: r is the Pearson's correlation coefficient between simulated ($\mathbf{s}$) and observed ($\mathbf{o}$) streamflow time series; β is the bias ratio ($\beta = \mu_s/\mu_o$), i.e., the ratio of mean simulated flow (μ_s) to mean observed flow (μ_o); α is the variability ratio, i.e., the ratio of either the standard deviations ($\alpha = \sigma_s/\sigma_o$) or the coefficients of variation ($\alpha = \mathrm{cv_s}/\mathrm{cv_o}$) of simulated and observed streamflow. As notation, we will refer to the so-called *modified KGE* (Kling et al., 2012) as KGE', which uses the ratio of coefficients of variations for α. The ideal value of the KGE (and KGE') is clearly 1, as for its three components.

The proposed JDKGE function is defined as:

$$JDKGE = 1 - \left\{(r-1)^2 + (\beta-1)^2 + (\alpha-1)^2 + JSD^2\right\}^{\frac{1}{2}} \tag{2}$$

where JSD is an estimate of the Jensen-Shannon Divergence. As the ideal value of JSD is 0, the JDKGE optimum is still 1. For the definition of the α component in JDKGE we take the ratio of coefficients of variation, as in KGE′ (and drop the apostrophe to keep a simpler notation).

Different formulations for the JSD exist, due to its continuous nature and the need for an approximation. Our JSD discrete and smoothed formulation can be defined by following the choices, steps and equations outlined in the sub-sections below (see Sections 2.1.1, 2.1.2, and 2.1.3).

2.1.1 JSD estimation: adopted methods and choices

To compute the JSD, an inference step is required, as the continuous probability density functions (PDFs) of the observed and simulated streamflows are not known. Different strategies to estimate the JSD may be followed, for example by adopting methods inspired by either the Riemann or the Lebesgue integration. Here we follow the former one by relying on the partitioning (binning) of the data domain into intervals (bins) to compute empirical distributions (histograms). To determine an appropriate bin width, we adopt the Freedman-Diaconis (FD) rule (Freedman and Diaconis, 1981), as also recommended by previous studies (e.g., Saurette et al., 2022, 2023). Since both the bin width and the number of bins determined by the FD rule are sensitive

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

to the streamflow timeseries length and timestep, we adjust the FD equation to ensure invariance to the timestep. Moreover, we constrain the number of bins between 25 and 100, preventing either overly coarse discretizations for short timeseries (that would smooth out important flow variability), or excessively fine binning for longer timeseries. An excessive granularity in the JSD has to be avoided as it could introduce instability in the calibration process by emphasizing minor differences in flow values that are not hydrologically meaningful, rather than capturing relevant differences in distributional similarity. This thresholding and time-scale invariant approach also prevents spurious effects that would otherwise disproportionately amplify the JSD influence relative to other components of the JDKGE metric as the time series lengthens or the time step decreases.

Consistently with general guidelines from the literature on the choice of an appropriate time series length for calibration (e.g., Arsenault et al., 2018), which are expected to hold for the calibration problem with the JDKGE, a minimum length of 4 years of daily observed data is recommended. For daily time series of approximately 4 years, the lower threshold of 25 bins is triggered in most cases (about 75% of cases, as estimated on our sample of 240 catchments; see Section 2.2.1), depending on the interquartile range of observed flows, while the upper threshold of 100 bins is rarely activated. Conversely, for longer calibration periods, exceeding roughly 20 years, the upper cap of 100 bins is reached more frequently. Together, these constraints and the FD rule help ensure a stable and informative approximation of the flow distribution across different lengths of the calibration period.

We use log-transformed flows in the JSD component to place greater emphasis on low flows when computing the distributional similarity. Using log-transformed flows for the other components of the JDKGE should be avoided due to critical issues of stability and interpretability, which were well documented for the KGE (Santos et al., 2018). These concerns do not apply to our JSD component, as the use of log-transformed flows is conceptually and mathematically different in this case (see Sections 2.1.2 and 2.1.3).

2.1.2 Main JSD computational steps

Let $\mathbf{o} = (o_t)_{t=1}^{n_o}$ and $\mathbf{s} = (s_t)_{t=1}^{n_o}$ be observed and simulated streamflow timeseries, respectively; both timeseries are sampled over the same period of n_o timesteps, at a regular sampling frequency (timestep), Δt (seconds).

The five key steps to compute the JSD are summarized below (and further detailed in Section 2.1.3):

1. **Pre-processing:** Remove any timestep with missing, negative or non-finite values in either $\mathbf{o}$ or $\mathbf{s}$. Set an ϵ lower than the minimum of any non-zero value in either series ($\epsilon \leq \min\{\min_{o_t>0} o_t, \min_{s_t>0} s_t\}$), and replace zeros by ϵ.

2. **Log transformation:** Log-transform flows to $\mathbf{x}^{(o)} = \log \mathbf{o}$ and $\mathbf{x}^{(s)} = \log \mathbf{s}$; set the common support of log-transformed flows, $[x_{\min}, x_{\max}]$, from $\mathbf{x}^{(c)} = \mathbf{x}^{(o)} \cup \mathbf{x}^{(s)}$.

3. **Bin width calculation with FD rule and time-scale adjustment:** Compute FD-based bin width h_{FD} from $\mathbf{x}^{(o)}$, applying fallback rules (for edge cases). Set a time-scale invariance factor, $ts_{\text{f}} = \Delta t / 86400$, and compute the number of bins so that is invariant with respect to changing time steps, and thresholded between $n_{\min}$ and $n_{\max}$.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

4. **Histograms construction with α-smoothing and normalization:** Create n_{bins} equal-width bins in log space on $[x_{\min}, x_{\max}]$; compute the histogram densities, i.e. bin heights $(d_i^{(o)}, d_i^{(s)})$, and convert them to rescaled piecewise-uniform distributions p_i and q_i (i.e., uniform within each bin i).

5. **JSD estimation:** Define the mixture distribution, $m_i = \frac{1}{2}(p_i + q_i)$, and estimate the JSD as sum of logarithms of ratios of the discrete probability distributions of observations ($\mathbf{p}$) over mixture and simulations ($\mathbf{q}$) over mixture (with logarithm in base $b = 2$).

2.1.3 JSD formulation with log-transformed flows and Freedman-Diaconis-rule based binning

Our JSD formulation includes a pre-processing step to ensure numerical stability and allow for a log transformation of the observed ($\mathbf{o}$) and simulated ($\mathbf{s}$) streamflow time series. For this, only finite and strictly positive values are kept. Any exact zero in either $\mathbf{o}$ or $\mathbf{s}$ is replaced by a small $\epsilon > 0$ chosen not to exceed the smallest positive value seen in either series:

$$\epsilon \le \min\Big\{ \min_{t:o_t>0} o_t,\ \min_{t:s_t>0} s_t \Big\}, \qquad o_t = \begin{cases} \epsilon, & o_t = 0 \\ o_t, & o_t > 0 \end{cases}, \quad s_t = \begin{cases} \epsilon, & s_t = 0 \\ s_t, & s_t > 0 \end{cases}.$$

Having defined the combined set $\mathbf{c}$ as $\mathbf{c} = \mathbf{o} \cup \mathbf{s}$, after removing any zeros, we set $\epsilon = \min\left(10^{-6}; 0.1 \cdot \min(\mathbf{c})\right)$. This definition is needed to ensure ϵ is lower than the smallest non-zero flow in either series, for numerical stability and consistency with the data scale.

After the pre-processing step, the logarithm of the two series can be computed:

$$x_t^{(o)} = \log o_t, \qquad x_t^{(s)} = \log s_t.$$

Then, we define $x_{\min} = \min(\mathbf{x}^{(o)} \cup \mathbf{x}^{(s)})$ and $x_{\max} = \max(\mathbf{x}^{(o)} \cup \mathbf{x}^{(s)})$.
At this point, we can derive the Interquartile Range (IQR) of the observations in the log space, i.e., $\text{IQR} = \text{IQR}(\mathbf{x}^{(o)})$, and define the Freedman–Diaconis (FD; Freedman and Diaconis, 1981) bin width h_{FD} as:

$$h_{\text{FD}} = \begin{cases} 2\,\text{IQR} \cdot n_o^{-\frac{1}{3}}, & \text{IQR} > 0, \\ \dfrac{x_{\max} - x_{\min}}{n_{\min}}, & \text{IQR} = 0 \quad \text{(fallback rule 1)}. \end{cases}$$

If $\text{IQR} = 0$, the fallback is the bin width corresponding to the minimum number of bins ($n_{\min} = 25$).

Then, a minimum width is enforced $h_{\min}$ (fallback rule 2):

$$h \;=\; \max\Big(h_{\text{FD}},\, h_{\min}\Big), \qquad h_{\min} = \min\big(10^2\,\epsilon,\ 10^{-1}\big).$$

Subsequently, a time-scale invariance factor (adimensional) is introduced and derived as the ratio of the timestep length over a daily time step (both in seconds):

$$ts_{\text{f}} = \frac{\Delta t}{86400},$$

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

and the number of bins is set as:

$$n_{\text{raw}} = \left\lceil ts_{\text{f}}^{1/3} \cdot (x_{\max} - x_{\min}) \cdot h^{-1} \right\rceil, \qquad n_{\text{bins}} = \max(\min(n_{\text{raw}}, n_{\max}), n_{\min}),$$

where the raw number of bins derived from the FD bin width is thresholded between the minimum and maximum number of bins, i.e., respectively $n_{\min} = 25$ and $n_{\max} = 100$.

At this point, we construct equally spaced bin edges in the log space:

$$\{b_j\}_{j=0}^{n_{\text{bins}}} \text{ with } b_0 = x_{\min},\ b_{n_{\text{bins}}} = x_{\max},\ b_j = x_{\min} + \frac{j}{n_{\text{bins}}}(x_{\max} - x_{\min}).$$

For this common set of bin edges $\{b_j\}_{j=0}^{n_{\text{bins}}}$, let

$$w_i = b_i - b_{i-1}, \qquad i = 1, \ldots, n_{\text{bins}},$$

denote the width of bin i.

Accordingly, for each series the empirical bin counts are defined as:

$$c_i^{(o)} = \begin{cases} \#\{\, x_t^{(o)} : b_0 \le x_t^{(o)} \le b_1 \,\}, & i = 1, \\ \#\{\, x_t^{(o)} : b_{i-1} < x_t^{(o)} \le b_i \,\}, & i = 2, \ldots, n_{\text{bins}}, \end{cases}$$

$$c_i^{(s)} = \begin{cases} \#\{\, x_t^{(s)} : b_0 \le x_t^{(s)} \le b_1 \,\}, & i = 1, \\ \#\{\, x_t^{(s)} : b_{i-1} < x_t^{(s)} \le b_i \,\}, & i = 2, \ldots, n_{\text{bins}}. \end{cases}$$

Then, the bin counts are converted to the histogram density function for observations ($d_i^{(o)}$) and simulations ($d_i^{(s)}$), as follows:

$$d_i^{(o)} = \frac{c_i^{(o)}}{n_o\, w_i}, \qquad d_i^{(s)} = \frac{c_i^{(s)}}{n_s\, w_i}, \qquad i = 1, \ldots, n_{\text{bins}},$$

so that each density (for observations and simulations) integrates to one:

$$\sum_{i=1}^{n_{\text{bins}}} d_i^{(o)}\, w_i = 1, \qquad \sum_{i=1}^{n_{\text{bins}}} d_i^{(s)}\, w_i = 1.$$

When all bins have equal width ($w_i = w$ for all i), as in our case, the density reduces to the relative frequency:

$$d_i^{(o)} = \frac{c_i^{(o)}}{n_o\, w} \propto \frac{c_i^{(o)}}{n_o}, \qquad d_i^{(s)} = \frac{c_i^{(s)}}{n_o\, w} \propto \frac{c_i^{(s)}}{n_o}.$$

Subsequently, we apply additive α-smoothing (also called "Laplace smoothing") to prevent zero-probability bins and extend by continuity due to the use of a common support of the log-transformed flows $[x_{\min}, x_{\max}]$. Here, we set $\alpha = \epsilon$, where ϵ is

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

the small positive constant defined in the pre-processing step. After smoothing, the resulting bin densities are renormalized to obtain valid discrete probability distributions:

$$\tilde{p}_i = d_i^{(o)} + \alpha, \qquad \tilde{q}_i = d_i^{(s)} + \alpha, \qquad p_i = \frac{\tilde{p}_i}{\sum_{k=1}^{n_{\text{bins}}} \tilde{p}_k}, \quad q_i = \frac{\tilde{q}_i}{\sum_{k=1}^{n_{\text{bins}}} \tilde{q}_k}.$$

Then, we define the mixture distribution (m) as:

$$m_i = \tfrac{1}{2}(p_i + q_i), \qquad i = 1, \ldots, n_{\text{bins}}.$$

Finally, the discrete estimate of the JSD is computed as follows:

$$\text{JSD}_b(\mathbf{p} \,\|\, \mathbf{q}) \;=\; \frac{1}{2}\sum_{i=1}^{n_{\text{bins}}} p_i \log_b\!\left(\frac{p_i}{m_i}\right) \;+\; \frac{1}{2}\sum_{i=1}^{n_{\text{bins}}} q_i \log_b\!\left(\frac{q_i}{m_i}\right),$$

where the logarithm base b is set to $b = 2$.

In theory, the additive α-smoothing ensures that all bin probabilities (p_i, q_i, and m_i) remain strictly positive. However, due to numerical issues in computational implementations (such as floating-point underflow or numerical rounding) exact zeros may still arise in practice in p_i and q_i. For this reason, following standard continuity arguments in information theory (e.g., Cover and Thomas, 2005), we adopt the following limiting-value convention, whereby any terms with $p_i = 0$ or $q_i = 0$ contribute 0 to $\text{JSD}_b(\mathbf{p} \,\|\, \mathbf{q})$. This corresponds to the limiting value of x:

$$\lim_{x \to 0^+} x \log_b\!\left(\frac{x}{m_i}\right) = 0 \quad (m_i > 0),$$

which ensures that the JSD remains well-defined and finite, even when numerical approximations may produce zero-valued probabilities.

2.2 Experimental design

2.2.1 Hydro-meteorological data and catchment samples

The model calibration and validation experiments were performed over two datasets to cover a wide range of characteristics in terms of catchment, climate and hydrological characteristics, as well as data quality and quantity. The two samples include:

1. A dataset of 240 catchments across France (Figure 1), set up by Ficchì et al. (2016) at daily and sub-daily time steps and used in other hydrological modeling studies (e.g., Ficchì et al., 2019; Santos et al., 2018): this is an established quality-controlled dataset with a large number of natural basins, with limited regulation and human influence, across different climate regimes over France. These catchments were selected according to the availability of continuous streamflow and precipitation measurements (with limited gaps), the rain gauges density, and the limited presence of human regulation and snow influence. Meteorological data were obtained from the SAFRAN mesoscale atmospheric reanalysis developed by Météo-France (Vidal et al., 2010), at 8-km resolution and hourly timestep, and from automatic rain gauges for rainfall at sub-daily resolutions (up to 6 minutes), while hydrological data were derived from the HydroPortail (previously

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

known as BanqueHydro) database (https://www.hydro.eaufrance.fr/). Potential evapotranspiration was estimated from temperature data (from the SAFRAN reanalysis) using the formula proposed by Oudin et al. (2005). Here we report and analyse extensively only the daily model results, but hourly datasets and models were also used to verify the consistency of the findings across temporal resolutions (results not shown).

2. A dataset of 45 global catchments (Figure 2) from the computational domains of the CEMS Global Flood Awareness System (GloFAS) and European Flood Awareness System (EFAS) (Matthews et al., 2025a, b). The meteorological data for this catchment sample is obtained from ERA5 for the subset from CEMS-GloFAS domain (40 catchments), with daily resolution, and from EMO-1 (Gomes et al., 2025; Thiemig et al., 2022) for the subset from CEMS-EFAS (5 catchments), with a 6-hourly resolution. Potential evapotranspiration is computed using the Penman-Monteith equation. The inclusion of catchments with sub-daily forcing data further allows us to verify the robustness of the proposed objective function (JDKGE) across different temporal resolutions. For the 5 catchments with 6-hourly data, the model is run at 6-hourly time step, and then model outputs are aggregated at daily time step for a common evaluation across all 45 catchments. This dataset covers catchments distributed across even more varied regimes from diverse global regions, and with more challenging characteristics for modeling, including intermittent or ephemeral rivers, heavily regulated catchments (with reservoirs), river systems with lakes, and river gauge stations with less strict data quality requirements (e.g., larger data gaps, less homogeneous quality controls due to operational needs, and the retention of potentially *suspicious* streamflow observations). The inclusion of such challenging conditions is intended to perform a stress test for the new calibration function. Indeed, errors in observed streamflow time series are a well-known source of model instability and reduced robustness (Thébault et al., 2023). In addition, human impacts on the terrestrial hydrological cycle (including reservoir operations and water uses) are notoriously difficult to represent accurately in hydrological models due to their complex and non-stationary nature, as well as to the lack of information on human actions and decision-making processes (e.g., Galelli et al., 2025; Wada et al., 2017). Within this dataset, we estimate that 9 catchments are heavily regulated (i.e., they include at least one reservoir represented in LISFLOOD), 7 catchments are influenced by lakes, and at least 5 catchments exhibit potentially suspicious or lower-quality streamflow records (e.g., noisy or non-natural behaviour, abrupt drops in rivers not expected to be regulated or intermittent, and larger data gaps). As these factors may affect calibration functions differently, this dataset offers the opportunity to compare model performance under particularly demanding conditions, and to assess the relative robustness of alternative calibration functions.

The diversity in catchment size, morphological and hydro-climatic conditions across the two selected study samples (Figure 1 and 2) ensures that the model experiments are confronted with a comprehensive representation of the large variability that can be found in rainfall-runoff processes across climates and catchment characteristics. In particular, the two samples cover a wide range of aridity conditions, as captured by the aridity index, and baseflow contributions, as represented by the base flow index (BFI), calculated as by Gustard et al. (1992) from daily streamflow data. Both samples range from water-limited catchments with high aridity and low runoff efficiency to groundwater-dominated systems with high baseflow indices (BFI > 0.7). This ensures strong contrasts in the low-flow regimes which are a key focus of this study. Between the two catchment samples, the

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

global one presents a more skewed distribution of the analysed characteristics with more extreme cases, and a predominance of arid catchments and with large groundwater contributions.

2.2.2 Models

We selected two hydrological models that are used operationally for monitoring and forecasting in different countries across the world: GR6J and OS-LISFLOOD.

1. GR6J (Génie Rural à 6 paramètres Journalier; Pushpalatha et al., 2011) is a conceptual rainfall-runoff model developed by INRAE (see https://webgr.inrae.fr/eng/tools/hydrological-models), based on a soil-moisture accounting reservoir, non-linear routing stores and unit hydrographs. GR6J, part of the suite of Génie rural (GR) hydrological models (Coron et al., 2017), is a variant of the more popular GR4J model (Perrin et al., 2003). With respect to GR4J, it has an increased model structure complexity, specifically designed to improve the simulations of low flows and groundwater exchanges, thanks to an additional exponential store, to better simulate groundwater dynamics, and a more refined routing component. GR6J has 6 parameters to calibrate (Table A1) which include four water-balance parameters (the production store's capacity, the exponential store's capacity and two inter-catchment groundwater exchange coefficients) and two routing parameters (routing store capacity and base time of the unit hydrographs). The structure has been refined over decades of research to maximise performance over large catchment samples, with a range of climate characteristics and distributed globally, while ensuring model parsimony. GR6J is used operationally for low-flow forecasting and water resources applications in different Countries, including France, within the PREMHYCE operational tool (Nicolle et al., 2020) and the UK (Harrigan et al., 2018; Hannaford et al., 2023) where it has replaced GR4J in 2023 due to its superior performance (Hannaford et al., 2023; UKCEH, 2025). Both GR4J and GR6J are open-source and freely available from different sources. We use the version implemented within the suite of models provided in the airGR R package (Coron et al., 2017).

2. OS–LISFLOOD, *aka* LISFLOOD (De Roo et al., 2000; Knijff et al., 2010), is an open-source, distributed physics-based model developed by the European Commission's Joint Research Centre (JRC) and used for several applications and operational services of the Copernicus Emergency Management Service (CEMS), ranging from flood forecasting (see https://global-flood.emergency.copernicus.eu) at European (Thielen et al., 2009; Matthews et al., 2025b) and global scale (Alfieri et al., 2013; Matthews et al., 2025a), to drought monitoring (see https://drought.emergency.copernicus.eu; Cammalleri et al., 2017, 2020) and water-resources management, including under climate change (Bisselink et al., 2020). Furthermore, LISFLOOD has also recently been used to quantify terrestrial water storage (TWS) changes on Earth for geodetic applications (Jensen et al., 2025) or to study projected climate change impacts on floods and droughts (Ekolu et al., 2025; Diop et al., 2025). The LISFLOOD model has up to 14 model parameters to calibrate (Table A2, Appendix A), where the exact number depends on the setup. In this work, we use two LISFLOOD setups: (i) for the CEMS-GloFAS catchments (40 out 45), we use the GloFAS v5.0 setup (current development version of LISFLOOD), where 12

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

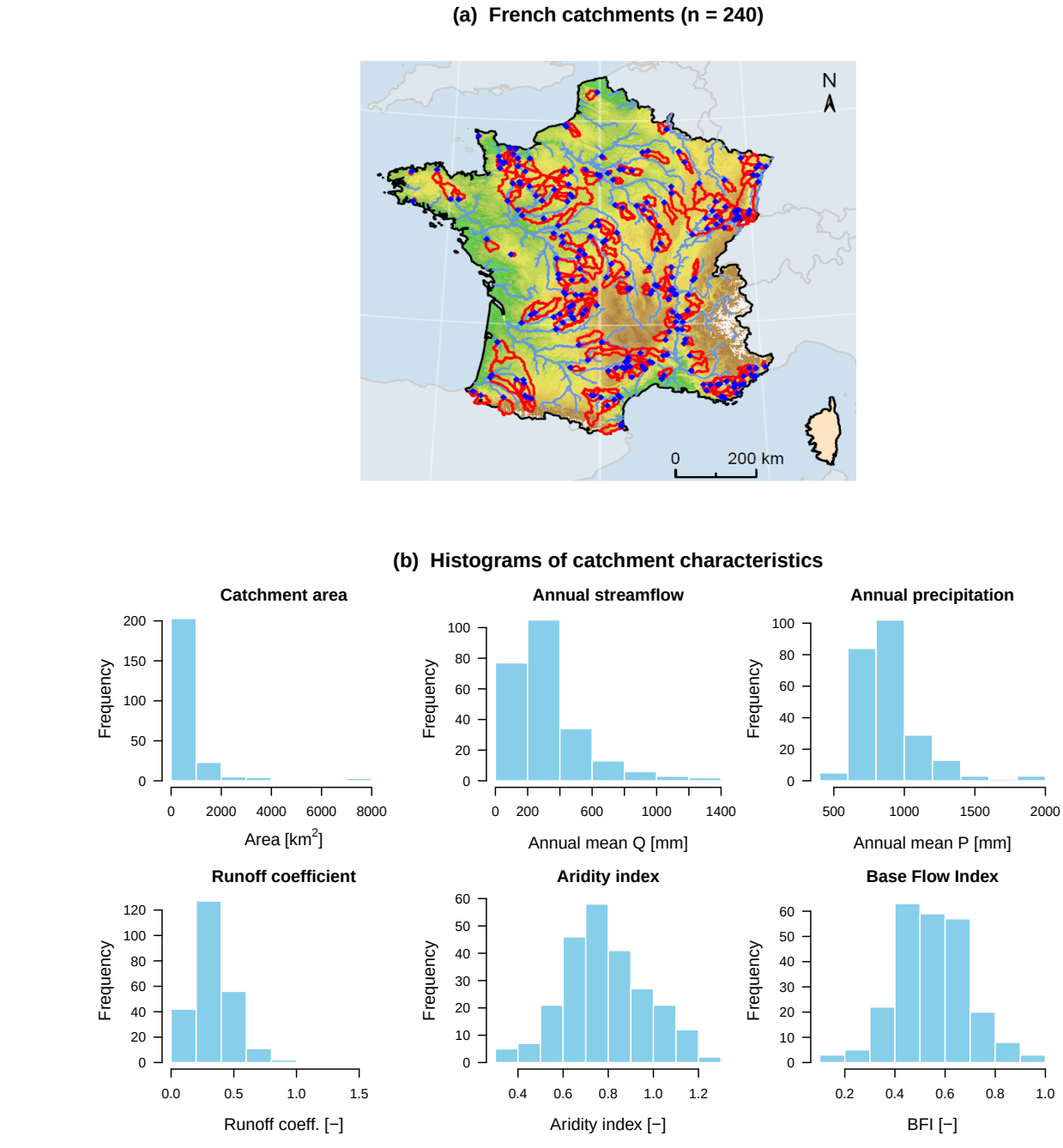

Figure 1. French catchment dataset (Ficchì et al., 2016) used for the first set of modeling experiments with GR6J: (a) spatial distribution of the 240 catchments (red lines: catchment boundaries; blue points: river gauges at catchment outlet), and (b) histograms of key catchment characteristics. The reported characteristics are: catchment area, annual mean streamflow (Q), annual mean rainfall (P), runoff coefficient (mean annual Q/P), aridity index (mean annual potential evapotranspiration divided by P), and Base Flow Index (BFI).

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

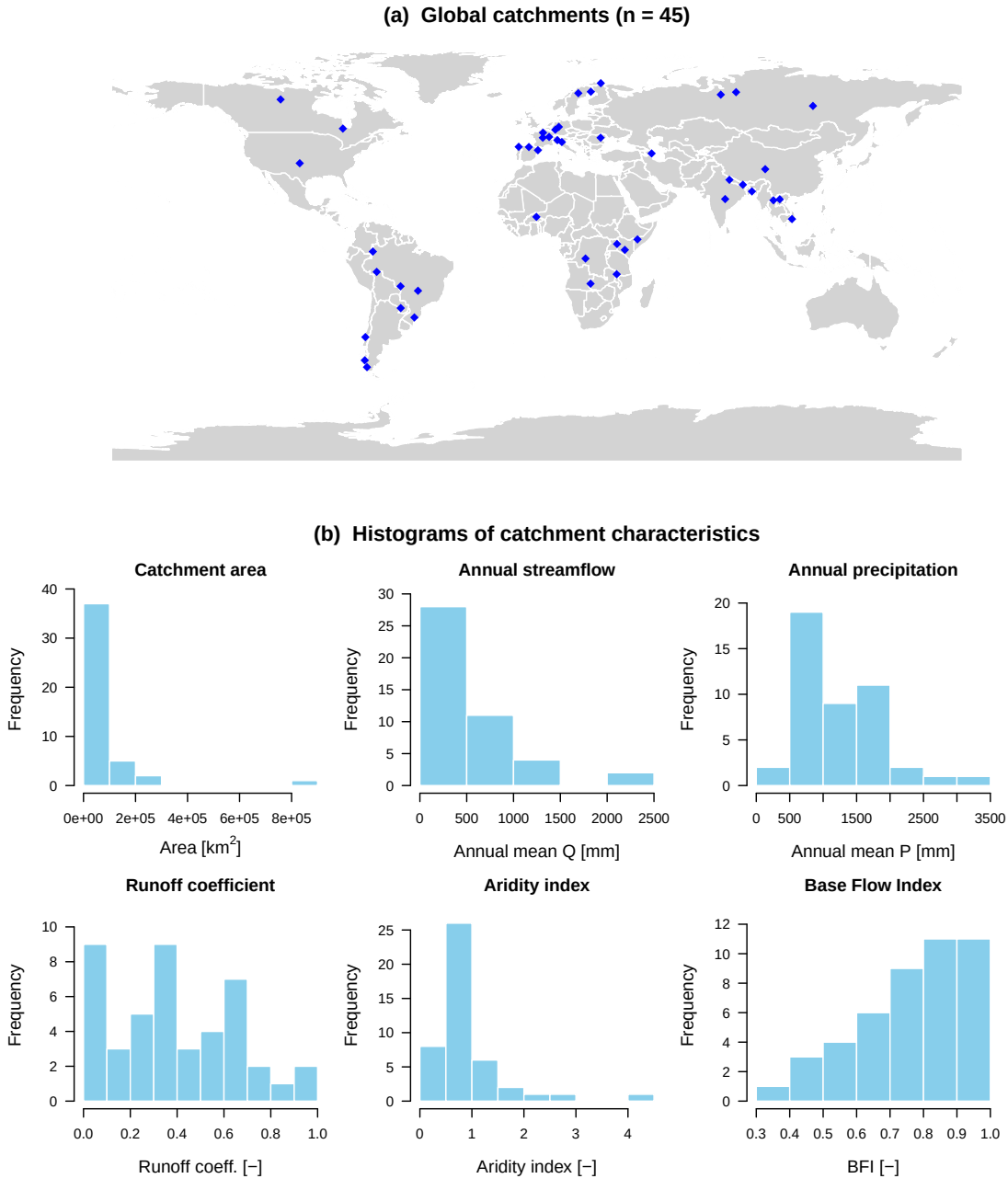

Figure 2. Global catchment dataset drawn from the CEMS-GloFAS and CEMS-EFAS computational domains: (a) spatial distribution of the 45 catchments (blue points: river gauges at catchment outlet), and (b) histograms of key catchment characteristics (same set of characteristics as in Fig. 1). This dataset is used for the second set of modeling experiments with both OS-LISFLOOD and GR6J.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

parameters are jointly calibrated (using the algorithm described in Section 2.2.4), while the 2 parameters on reservoirs are tuned via a separate procedure (independent of the objective function); (ii) for the CEMS-EFAS catchments (5 out 45), we use the EFAS v5.0 setup, where all 14 parameters (including the 2 reservoir-related parameters) are jointly calibrated (ECMWF, 2022). The 14 calibrated parameters include 8 water-balance parameters (characterizing the snowmelt rate, infiltration capacity, maximum percolation rate, residence time of the upper and lower saturated soil zones, groundwater-river connectivity, percolation to the deep aquifer, and transmission losses), 3 routing parameters (related to the channel Manning's coefficients and discharge level triggering floodplain flow), 2 parameters on reservoirs (calibrated only in our EFAS setup) and 1 on lakes (where reservoirs or lakes are present).

For both models, the free parameters (Tables A1 and A2) are tuned via automated calibration, by optimizing a goodness-of-fit metric as a function of simulated and observed flows, for which we test as alternatives our proposed metric (JDKGE) and several competitor functions (see Section 2.2.3).

2.2.3 Benchmark and competitor calibration functions

Our main benchmark calibration function is the modified Kling-Gupta Efficiency (KGE′; Kling et al., 2012), which is used for the calibration experiments of both models (GR6J and LISFLOOD) to provide the key reference to be compared with our new function (JDKGE). In addition, we consider a wider set of competitor (alternative benchmark) functions from the literature, using them in the GR6J tests only, given the lower computational costs of GR6J with respect to LISFLOOD which allows for more extensive tests. The objective functions tested in the main single-objective calibration experiments with GR6J are:

1. The original Kling-Gupta Efficiency (KGE; Gupta et al., 2009), defined as in Eq. (1) with α being the ratio of the standard deviations, i.e., $\alpha = \sigma_s/\sigma_o$.

2. The modified Kling-Gupta Efficiency (KGE'; Kling et al., 2012), defined as in Eq. (1) with α being the ratio of the coefficients of variation, i.e., $\alpha' = \mathrm{cv}_s/\mathrm{cv}_o$.

3. The Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970), defined as:

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T}(o_t - s_t)^2}{\sum_{t=1}^{T}(o_t - \bar{O})^2} \tag{3}$$

where o_t and s_t are the observed and simulated flow at time step t (with t in 1:T, with T being the time series length), and $\bar{O}$ is the mean observed flow.

4. The NSE on log-transformed flow data (NSE_log):

$$\mathrm{NSE\_log} = 1 - \frac{\sum_{t=1}^{T}(log(o_t) - log(s_t))^2}{\sum_{t=1}^{T}(log(o_t) - \bar{O_{log}})^2} \tag{4}$$

where $log(o_t)$ ($log(s_t)$) is the natural logarithm of observed (simulated) flow at time step t, and $\bar{O_{log}}$ is the mean of log-transformed observed flows. NSE_log is a traditional solution in the literature to focus on low-flows (e.g., Orth et al.,

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

2015) and has been shown to better reflect human judgment on drought performance with respect to other metrics (e.g., Gauch et al., 2023).

5. The non-parametric KGE variant (KGE_NP), proposed by Pool et al. (2018) to improve the calibration on low flows:

$$\mathrm{KGE_{NP}} = 1 - \sqrt{(r_s - 1)^2 + (\alpha_{\mathrm{NP}} - 1)^2 + (\beta - 1)^2}, \tag{5}$$

where r_s is the Spearman rank correlation coefficient between observed and simulated flows:

$$r_s = \frac{\sum_{t=1}^{T}\left(R_t^{(o)} - \overline{R^{(o)}}\right)\left(R_t^{(s)} - \overline{R^{(s)}}\right)}{\sqrt{\left[\sum_{t=1}^{T}\left(R_t^{(o)} - \overline{R^{(o)}}\right)^2\right]\left[\sum_{t=1}^{T}\left(R_t^{(s)} - \overline{R^{(s)}}\right)^2\right]}}, \tag{6}$$

with $R_t^{(o)}$ and $R_t^{(s)}$ denoting the ranks of o_t and s_t, respectively; the non-parametric variability component α_{NP} is defined from the (normalized) flow-duration curve as:

$$\alpha_{\mathrm{NP}} = 1 - \frac{1}{2}\sum_{k=1}^{T}\left|\frac{s_{I(k)}}{T\,\mu_s} - \frac{o_{J(k)}}{T\,\mu_o}\right|, \tag{7}$$

where $I(k)$ and $J(k)$ return the indices of the k-th largest simulated and observed flows, respectively; β is the usual bias ratio ($\beta = \mu_s/\mu_o$).

6. Our proposed JDKGE function, defined as in Eq. (2), by integrating the JSD into the modified KGE (see Section 2.1 for its full definition).

Additionally, we performed multi-objective calibration experiments with KGE and JSD as separate objectives to explore trade-offs between overall performance and accuracy on low flows (JSD) (see Section 2.2.4). These experiments were conducted using GR6J only, owing to its comparatively low computational cost; extending the same multi-objective analysis to LISFLOOD would be computationally prohibitive.

2.2.4 Calibration algorithms and adopted settings

In the single-objective calibration experiments, optimization algorithms are used to explore the parameter space and identify the set leading to the optimal value of the selected objective function. We use two automatic optimization algorithms that are the standard option in the open-source (R and Python) implementations of the two models. In particular:

1. GR6J: the calibration is performed using a two-step approach combining global and local search algorithms, derived from Michel (1991). First, a coarser screening of the parameters space is performed using either a rough predefined grid (over default ranges) or a user-defined list of initial parameter sets. Then, a local refinement is carried out using a steepest descent local search algorithm, starting from the best-performing set of parameters identified in the initial step. This algorithm, developed by INRAE and implemented in the `airGR` R package (Coron et al., 2017), has demonstrated superior performance with respect to other global optimization algorithms for the calibration of the GR models (Thirel et al., 2024).

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

2. OS-LISFLOOD: the Distributed Evolutionary Algorithm for Python (DEAP; Fortin et al., 2012; De Rainville et al., 2012) is used, as implemented in the open-source LISFLOOD calibration tool (https://github.com/ec-jrc/lisflood-calibration). This algorithm is based on the iterative evolution of a population of parameter sets, with a loop for offspring generation via crossover and Gaussian-based mutation (Fortin et al., 2012; De Rainville et al., 2014). At each iteration, the population of parameters is evaluated by model simulations to determine the best performing individuals until a stopping criterion is met.

Using two different optimization algorithms further supports the robustness and generality of our findings.

For the multi-objective calibration experiments with KGE′ and JSD as separate objectives (using GR6J), we used a genetic algorithm, i.e., the *caRamel* algorithm, as implemented in the R package by Monteil et al. (2020). The caRamel algorithm is a hybrid evolutionary multi-objective optimization method, capable of identifying a family of Pareto-optimal parameter sets for multi-objective problems. It combines a stochastic genetic algorithm, i.e., the non-dominated sorting genetic algorithm II (NSGA-II), with the multi-objective evolutionary annealing simplex (MEAS) and gradient-like deterministic search rules. As key settings, we used a population size of 5000 for the genetic algorithm, with 25 new parameter sets per generation and for each rule, and a maximum number of 100k simulations. These settings were refined through preliminary sensitivity tests, to ensure adequate exploration of the parameter space, convergence of the Pareto fronts, and stable identification of optimal solutions and trade-offs with respect to the two objective functions (KGE′ and JSD).

In the main experiments here described, model calibration was performed over the whole period of data availability (8 years for the French catchment set, and at least 4 years for the global dataset), where only a first part of the dataset was used as warm-up period (i.e., 2 years for the French catchment database and 3 years for the global dataset). This is in line with the recommendations from a recent extensive hydrological modeling study (Shen et al., 2022), demonstrating the value of calibrating using the whole period of data availability.

2.2.5 Evaluation metrics and their normalization

To evaluate how different calibration functions affect the performance along the whole FDC, we use specific quantile-based scores and FDC-based hydrological signatures, in addition to the competing calibration metrics (KGE, its three traditional components, and JSD). All metrics are computed over the whole time series of daily simulated and observed flows. For modeling experiments at sub-daily time steps, outputs are first aggregated at daily time scale before computing the evaluation metrics.

Quantile ratios

To analyse the calibrated models in terms of specific biases along the FDC, we compute quantile ratios, defined as the ratios between simulated and observed streamflow quantiles, as follows:

$$Qr_x = \frac{s_x + \epsilon}{o_x + \epsilon},$$

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

where s_x and o_x are the x-quantile of the paired simulated and observed time series, and ϵ is a small positive constant chosen to be negligible relative to non-zero flows. The small positive constant ϵ is added to both the numerator and denominator to avoid division by zero when the observed quantile is equal to zero (e.g., in intermittent or ephemeral rivers). This adjustment ensures that quantile ratios remain well defined and interpretable everywhere. Here we define ϵ as a negligible fraction of the smallest non-zero observed flow:

$$\epsilon = 10^{-6} \min\{\, o_t : o_t > 0 \,\},$$

ensuring that ϵ does not influence ratios at non-zero quantiles while providing numerical stability at the lowest flow levels.

Hydrological signatures based on the FDC

To assess the performance of specific parts of the FDC, we use three diagnostic hydrological signatures (Yilmaz et al., 2008):
(i) the percent bias in the FDC mid-segment slope (pBiasFMS, [%]),
(ii) the percent bias in the FDC low-segment volume (pBiasFLV, [%]), and
(iii) the percent bias in the FDC high-segment volume (pBiasFHV, [%]).

The mid-segment slope bias ($pBiasFMS$) quantifies how well the model reproduces the slope of the FDC mid-segment, between 0.2–0.7 flow exceedance probabilities:

$$\text{pBiasFMS} = \frac{[\log(Q_{\text{sim},m_1}) - \log(Q_{\text{sim},m_2})] - [\log(Q_{\text{obs},m_1}) - \log(Q_{\text{obs},m_2})]}{[\log(Q_{\text{obs},m_1}) - \log(Q_{\text{obs},m_2})]} \times 100 \tag{8}$$

where m_1 and m_2 refer to the 0.2 and 0.7 flow exceedance probabilities. Positive values of $pBiasFMS$ indicate a steeper simulated FDC mid-segment than observed, which corresponds to flashier simulated runoff response than observed; on the other hand, a negative $pBiasFMS$ indicates an underestimation of the FDC mid-segment slope (slower response than observed). Ideally, $pBiasFMS$ should be close to zero, indicating good agreement between observed and simulated mid-segment slopes.

High- and low-flow volume biases ($pBiasFHV$ and $pBiasFLV$) are computed respectively as:

$$\text{pBiasFHV} = \frac{\sum_{h=1}^{H} (Q_{\text{sim},h} - Q_{\text{obs},h})}{\sum_{h=1}^{H} Q_{\text{obs},h}} \times 100 \tag{9}$$

$$\text{pBiasFLV} = \frac{\sum_{l=1}^{L} (Q_{\text{sim},l} - Q_{\text{obs},l})}{\sum_{l=1}^{L} Q_{\text{obs},l}} \times 100 \tag{10}$$

where $h = 1, 2, \ldots, H$ are the flow indices for values within the respective exceedance probabilities intervals, i.e., for high flows [0–0.02], and for low flows [0.7–1.0].

Positive values of $pBiasFHV$ ($pBiasFLV$) indicate a model overestimation of high (low) flows, while negative values represent an underestimation of high (low) flows. The ideal values is zero when the model is able to accurately reproduce high/low flows.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

Following previous studies (e.g., Cislaghi et al., 2020; Pfannerstill et al., 2014), these percent biases are considered as acceptable or *satisfactory* when their values are within a $\pm 30\%$ *satisfactory range*, i.e., the bias of simulated FDC (mid, low, or high) segment slope or volumes is lower than $\pm 30\%$ with respect to observations.

Metric normalization

To enable effective joint comparisons and visualization of heterogeneous metrics, which differ in performance (ranking) direction, scale and domain, we normalize each metric m to the unit interval $[0,1]$ with the ideal value at 1 (see Appendix B). The resulting normalized scores are denoted as z_m (or z_m). This normalization is applied only when relevant, for example to perform non-parametric statistical tests or to visualize effectively multiple metrics in a radar (spider) chart.

2.2.6 Statistical testing

To assess the statistical difference between model results from two different calibrations, we use a two-sample statistical test based on the *2-Wasserstein distance* (wd, hereafter referred to as Wasserstein distance) and asymptotic theory, recently proposed by Schefzik et al. (2021) and implemented in the `waddR` package. The test was originally developed in the context of bioinformatics, but as suggested by the authors it can be used in a broad range of disciplines. In our application, the two samples represent two sets of model performance scores obtained for the same evaluation metric over the same catchment sample by a model calibrated with two different objective functions. Let F_A and F_B denote the empirical cumulative distribution functions (CDFs) of the two samples. Using the 2-Wasserstein distance between the two samples, the test evaluates the null hypothesis $H_0 : F_A = F_B$ (i.e., F_A and F_B are drawn from the same underlying distribution) against the alternative hypothesis $H_1 : F_A \neq F_B$ (i.e., the two distributions are different). The test is based on the asymptotic distribution of the test statistic under the null hypothesis, which is valid when the samples can be assumed to come from continuous distributions, as in our case (and in most hydrological modeling evaluation cases).

Following Schefzik et al. (2021), we highlight here the main advantages of this test compared to statistical tests more commonly used in hydrological modeling studies (e.g., Friedman or Wilcoxon signed-rank tests):

1. it is a nonparametric test based on asymptotic theory, making it flexible, computationally efficient (more than permutation tests) and accurate (when the two distributions are continuous);

2. it assesses differences between entire distributions rather than focusing primarily (or only) on location or rank-based differences, allowing it to detect more complex distributional changes; moreover, the values of wd (in addition to p-values) can be used to summarize the distributional differences;

3. by exploiting the decomposition of wd into location, scale (size), and shape components (similarly to the decomposition behind KGE), it enables a quantitative attribution of the observed (significant) differences to these distinct distributional aspects (in %).

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

3 Results

This section first presents the results obtained by applying the novel JDKGE objective function against several benchmarks and alternative calibration functions (see Section 2.2.3) to the GR6J model across 240 French catchments (see Figure 1). Next, the calibration results of the LISFLOOD and GR6J models over 45 global catchments (Figure 2) is analyzed, comparing the main benchmark objective function (modified KGE') with our proposed function (JDKGE) for both models. Finally, the results of GR6J multi-objective calibration experiments using KGE' and JSD as separate objectives are summarized to highlight the tradeoff between the benchmark function (KGE') and the novel JSD component (defined as in JDKGE; see Section 2.1.3).

3.1 Calibration results of JDKGE vs. benchmarks and competitors using GR6J over 240 French catchments

The simulation results using GR6J over the sample of 240 French catchments are first evaluated based on 8 metrics (see Section 2.2.5) and 8 years of data. The distributions of the scores show that, depending on the evaluation metrics considered, our new function (JDKGE) either outperforms or matches the general-purpose benchmarks (KGE, KGE' and NSE) and the other competitors (NSE_log, KGE_NP) as calibration objective functions (Figure 3). In particular, according to the main (more generalist) performance evaluation metric considered (KGE'), i.e. modified KGE by Kling et al. (2012), the distribution of performances of the model calibrated with our new function matches the distribution obtained by calibrating the model using the same (benchmark) function, modified KGE' by Kling et al. (2012), or the original KGE (Gupta et al., 2009), as objective function. This is obviously a non-trivial (rather quite unexpected) and positive achievement, as one would expect a higher level of performance when evaluating the model with the same function as used in calibration, rather than when using a different function for calibration and evaluation. Moreover, what is particularly striking in the overall performance of different alternatives (Figure 3; see also Appendix Figure A1) is the large difference between our JDKGE function with respect to other benchmarks (NSE) and competitor alternatives proposed in the past to improve low flows, i.e., NSE_log and KGE_NP: while these other functions deteriorate the overall performance (KGE') more significantly to get an improvement on low-flows, our JDKGE function does also bring improvements on low flows (even larger), but at less significant cost in terms of overall performance with respect to the benchmark functions (KGE and KGE'), as found by using the 2-Wasserstein distance test (see Appendix Figure A2). Looking at the decomposition of the overall performance into correlation, variability and bias ratios (KGE components), the other competitor functions lead to a much larger degradation of specific performance aspects: this is particularly visible from the distributions (Figure 3 and Figure A1) and from the pairwise statistical differences (Appendix Figure A2), especially in terms of relative variability for NSE, NSE_log and KGE_NP, correlation for KGE_NP, and mean bias ratios for NSE and NSE_log. The distributions of correlation, relative variability and ratio of means are only slightly altered by JDKGE. In particular, no significant difference is found in the distribution of the correlation with respect to the calibration with KGE. A significant difference is found with the other two components, but corresponding to a small 2-Wasserstein distance (small distributional change) and a small performance loss on these simple ratios of means and variability measures with respect to KGE'. The medians of α and β change by 0.003 and 0.001, respectively; the mean of α moves slightly

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

from 0.993 (obtained calibrating with KGE') to 0.988 (with JDKGE), while the mean of β moves from 0.994 to 0.991. The value and decomposition of the Wasserstein distance indicate that there is no important change in the full distribution either. Moreover, our JDKGE function improves significantly the representation of the FDC and relative frequencies across the full flow spectrum, as captured well by the JSD metric and in part by the three FDC-based scores (percent bias in the FDC mid-segment slope, percent bias in the FDC high and low volume). In terms of low-flows (pbias_flv), our new function provides the best performance, shifting the distribution towards the satisfactory range. JDKGE outperforms the benchmarks but also the two popular variants of KGE and NSE proposed in the literature to improve the simulation of low-flows, i.e., NSE_log and KGE_NP. The improvement of JDKGE on low-flows is followed closely by NSE_log (not significantly different) and to a lesser extent by KGE_NP (more distant and significantly different). However, these competitor functions (especially NSE_log) lead to negative marked performance loss on high flows (see degradation in pbias_fhv) which are strongly significant (Appendix Figure A2). Finally, our new function performs best in terms of flows frequencies (JSD), as quite expected, followed closely by NSE_log. Only in terms of the mid-segment of the FDC (pbias_fms), the non-parametric KGE (KGE_NP) and to a lesser extent NSE_log slightly outperform the JDKGE function; this indicates that the non-parametric KGE is still an interesting option, especially when model users would be interested in the mid-segment of the FDC or regime flows more than low- and high- flows (as highlighted also by a slightly better bias ratio component, beta). Following Yilmaz et al. (2008), this suggests that the KGE_NP leads to a good representation of the flashiness of runoff or vertical redistribution of water in the model (as pbias_fms is related to these aspects), followed closely by NSE_log and JDKGE, which also outperform KGE and KGE' as calibration functions in optimizing pbias_fms. However, JDKGE leads to better performance over both high and low flows than these competitors (KGE_NP and NSE_log), striking a better trade-off when model users are interested in both extremes, i.e., floods and droughts.

To analyse further the performance along the whole FDC, we look at all quantile ratios (Qr) for all percentiles from 1 to 100 (with step 1%). As introduced in Section 2.2.5, these scores highlight the deviations between simulated and observed flows, with a perfect model that would result in ratios of 1 across all percentiles. We observe that the new calibration function (JDKGE) significantly improves the performance along the whole FDC, as the quantile ratios of the model calibrated with JDKGE are closer to 1 for almost all percentiles, with respect to the benchmark (KGE'). The relative improvements in terms of quantile ratios are larger for low flow (Q1–Q25) and mid-low flow (Q25-Q50). The tails of the distribution of quantile ratios across the 240 catchments improve consistently over all segments of the FDC, except for minor parts of the mid-high flow section (between 60th and 70th percentile flows), and slightly improving also over most percentile flows in the high-flow segment (Fig. 4). For the most extreme flow values (98th-100th percentile flows) the differences are minimal, with a slight degradation on the lower tail of the distribution of scores, but a more marked improvement in the location and variance of the distribution of Qr_100 around the ideal value brought by JDKGE.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

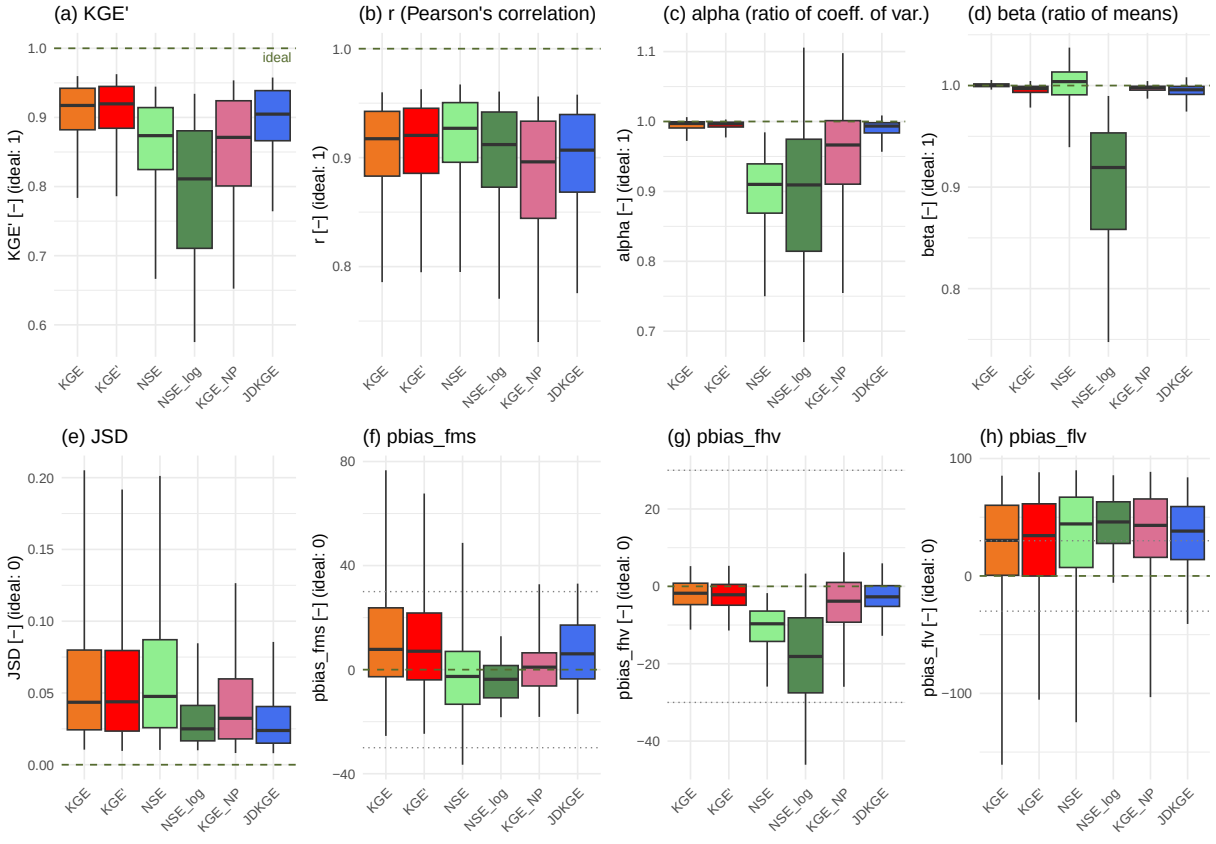

Figure 3. Distribution of the performance evaluation metrics for the GR6J model over 240 French catchments using 6 calibration function alternatives (reported on the x-axis), i.e., KGE, KGE', NSE, NSE_log, KGE_NP, JDKGE. The evaluation metrics are: (a) KGE, (b) Pearson's correlation, (c) variability ratio (alpha), (d) ratio of means (beta), (e) JSD, (f) the percent bias in the FDC midsegment slope (pBias_FMS [%]), (g) the percent bias in the FDC high-segment volume (pBias_FHV [%]), (h) the percent bias in the FDC low-segment volume (pBias_FLV [%]). The boxplot shows the median value and interquartile range, with the whiskers representing the 5-th and 95-th percentiles. The dashed horizontal line indicates the ideal value for each score. For the FDC-based percent biases (pBias_FMS, pBias_FHV, pBias_FLV), thin dotted lines mark the satisfactory range (±30%).

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

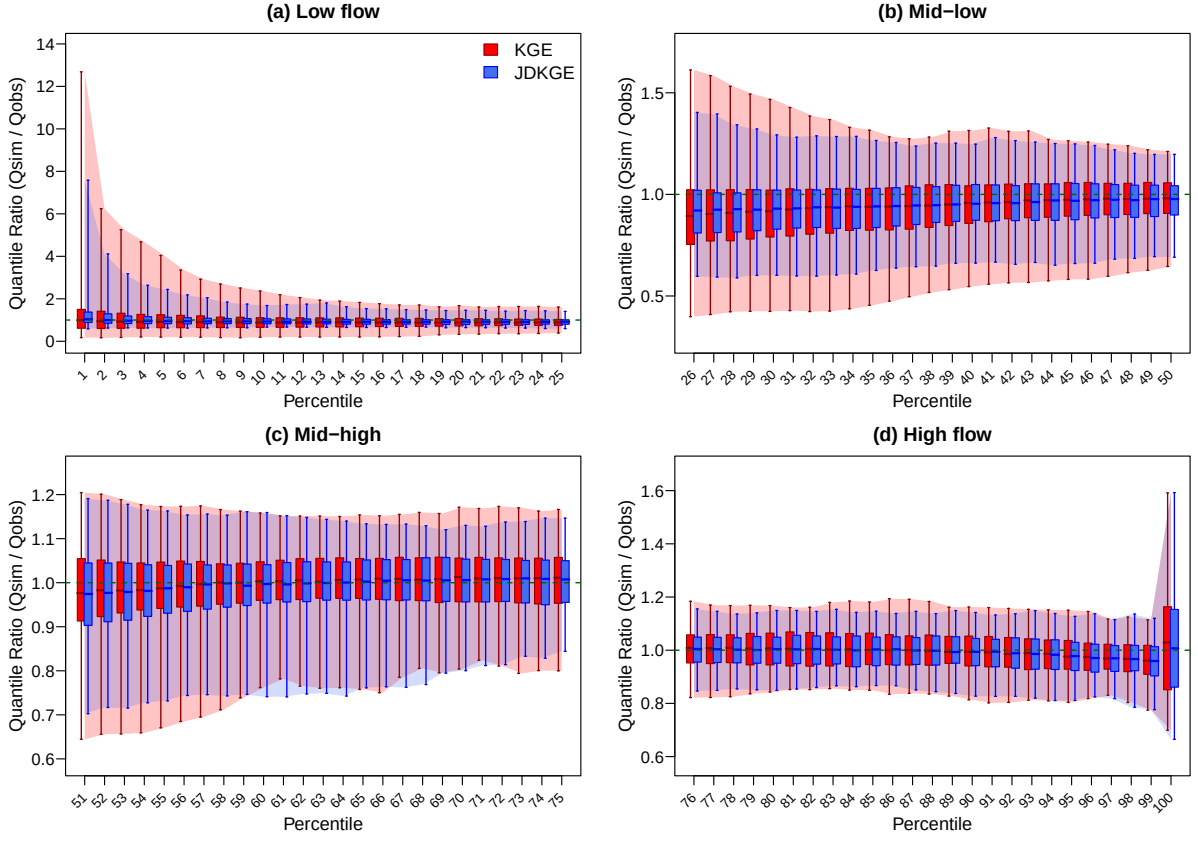

Figure 4. Quantile ratios (Qr_1–Qr_100), defined as the ratio of simulated to observed percentile flows, for the GR6J model calibrated using KGE' and JDKGE objective functions. Each boxplot represents the distribution of quantile ratios across all 240 basins: the box shows the interquartile range with the median, the whiskers span the 5th to 95th percentiles, and the shaded envelope indicates the same percentile range to better visualize the changing behavior across percentiles. The dashed horizontal line indicates the ideal value (Qr=1).

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

3.2 Calibration results of JDKGE vs. benchmark using LISFLOOD and GR6J over 45 global catchments

To test the level of generalizability of the performance improvements across different models and hydro-climate conditions, we performed calibration tests of both OS-LISFLOOD and GR6J across the same sample of 45 diverse global catchments with a even wider range of different hydroclimatic and morphologic characteristics. For both models, we selected the main benchmark (modified KGE') as alternative to our novel objective function (JDKGE), as among the various options tested they optimize the two more generalist metrics (KGE' and JSD) and perform best in terms of low-flows (JDKGE) and high-flows (both KGE' and JDKGE). Our results (Figure 5) show that the improvements brought by using JDKGE as objective function instead of KGE' are marked across the full flow spectrum and of a similar magnitude for both models over the global catchment sample (see absolute JSD deltas in Figure 5), although they are larger with LISFLOOD than with GR6J. Moreover, these improvements are significant for both models, according to the the 2-Wasserstein distance-based test (with p<0.01 for LISFLOOD and p<0.05 for GR6J). This improvement occurs at no significant expense for the KGE (based on the statistical test; see Figure 5) and reflects a very marked improvement on low flows (significant) and to a lesser extent on high flows (not significant), as highlighted by the distributions of two selected extreme quantile ratios (Figure 5). The minor differences in behaviour between the results with LISFLOOD and GR6J suggest a good level of generalizability of the performance improvements across different models. The improvements of simulations on low flows (quantile ratios) and frequencies across the full flow spectrum (JSD) are more significant for LISFLOOD than for GR6J. This might be due to the structural differences of the two models, with the lack of modeling routines for reservoirs and lakes in GR6J affecting its capacity of improving the accuracy on low flows (over a sample with human influences and lakes).

A more comprehensive multi-metric assessment (Fig. 6) shows that calibrating with JDKGE yields systematic, marked gains on average over KGE′ for around half of the 14 metrics considered for LISFLOOD (7 out of 14) and for a few metrics (3 out 14) for GR6J, while not changing or not degrading significantly the other metrics. By ranking improvements by the magnitude of change in the central tendency (mean or 25th/75th percentiles) for LISFLOOD, the metrics that see a marked improvement in mean scores (by more than 10^{-2} metric points) are: (i) JSD (as logically expected, given the explicit optimisation of this component in JDKGE), (ii) Qr_1, (iii) Qr_5, (iv) Qr_10, (v) pbias_fms, (vi) pbias_flv, and (vii) Qr_90. The same ranking applies to GR6J for JSD, Qr_1 and Qr_5, but a stable average performance is found for the other metrics. For both models, the remaining metrics, i.e., KGE′, r, α, β, Qr_95, Qr_99, and pbias_fhv, exhibit only more modest changes (lower than 10^{-2}), with central lines close to parity and quantile bundles largely overlapping between the two calibrations. Notably, the quantile envelopes (q10–q90) for the first group of improved metrics shift outward towards the ideal ring, indicating that improvements are not confined to a few basins but extend across the distribution. For both models, the most pronounced gains occur for metrics sensitive to low flows (i.e., Qr_1, Qr_5, Qr_10) and for LISFLOOD also for FDC-based (percent-bias) metrics, suggesting better reproduction of low- and moderate-flow regimes under JDKGE, especially for LISFLOOD. In particular, the lower quantiles of these scores over the catchment sample (i.e., 10th/25th percentiles) show a very marked relative improvement (Fig. 6), suggesting that the new metric is particularly valuable for catchments where the performance is lower, especially in terms of low flows and flow distribution (JSD and percent biases on FDC mid- and low-segments). Overall, the radar chart

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

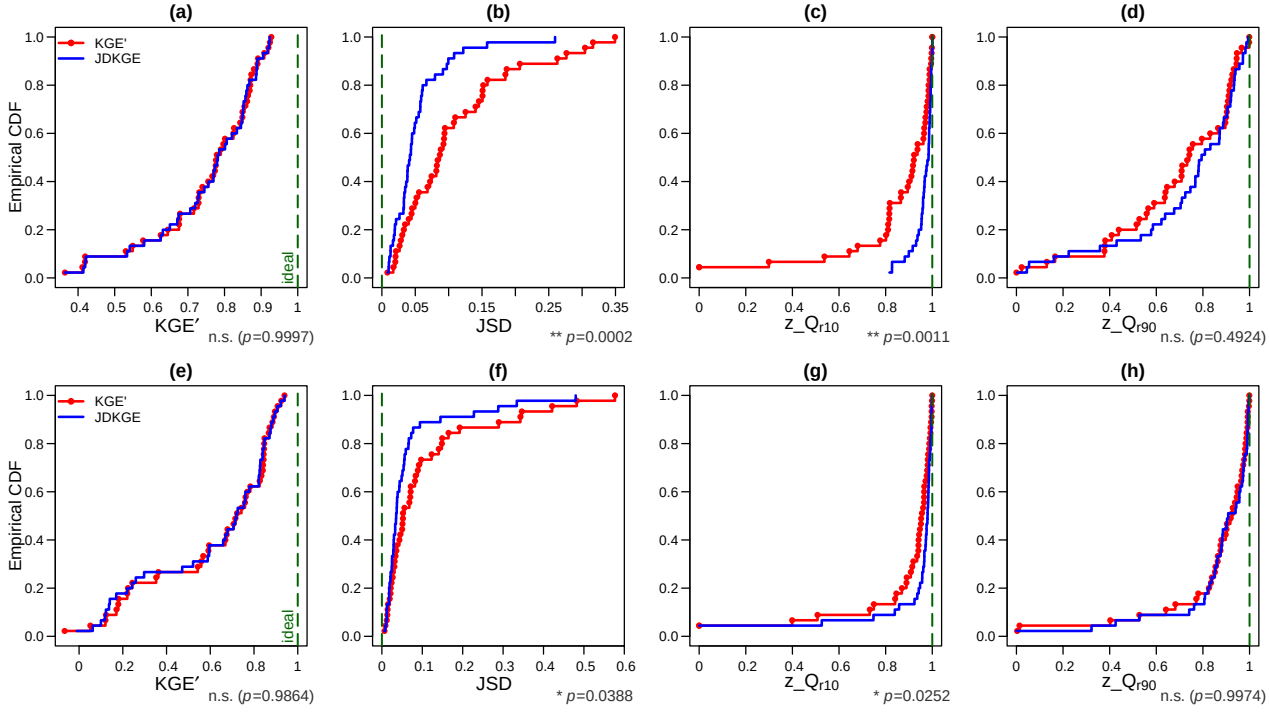

Figure 5. Empirical cumulative distribution functions (ECDFs) of four key performance metrics (KGE′, JSD, $z_{Q_{r,10}}$, and $z_{Q_{r,90}}$) across the global catchment set for OS–LISFLOOD (panels a–d, top row) and GR6J (e–h, bottom row). For each metric, the two curves compare the calibration with KGE′ (red line with points) and JDKGE (blue solid line). The ideal value is marked by a vertical (dashed) line as reference: KGE′ = 1, JSD= 0, and $z_{Q_{r,x}} = 1$ (quantile ratios are normalized as described in Appendix B). Statistical differences between the two distributions are assessed using the 2-Wasserstein distance-based test with asymptotic p-values; p-values are reported along the bottom margin of each panel, with a "*" when differences are significant ($p < \alpha$, $\alpha = 0.05$; "**" denotes $p < 0.01$); non-significant cases are labeled "n.s.".

therefore points to broader, distribution-wide benefits in these metrics, while changes in correlation, bias and relative variability remain comparatively small, with overlapping distributions of scores obtained by the two calibration approaches. For GR6J, there are only a few cases of degradation of the percent bias of the high-flow FDC segment and bias component of KGE (ratio of means) which drive the two marked degradations visible in the lower quantiles (q10) of these scores, with no marked change in average performance.

3.3 Multi-objective calibration results with KGE and JSD as competing objectives using GR6J

We performed multi-objective calibration experiments using GR6J on the 240 French catchments to explore the potential trade-offs between overall performance, as represented by the KGE′, and distributional similarity, particularly on low flows, as represented by the JSD. To illustrate the trade-offs between KGE′ and JSD, we present exemplary results for three catch-

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

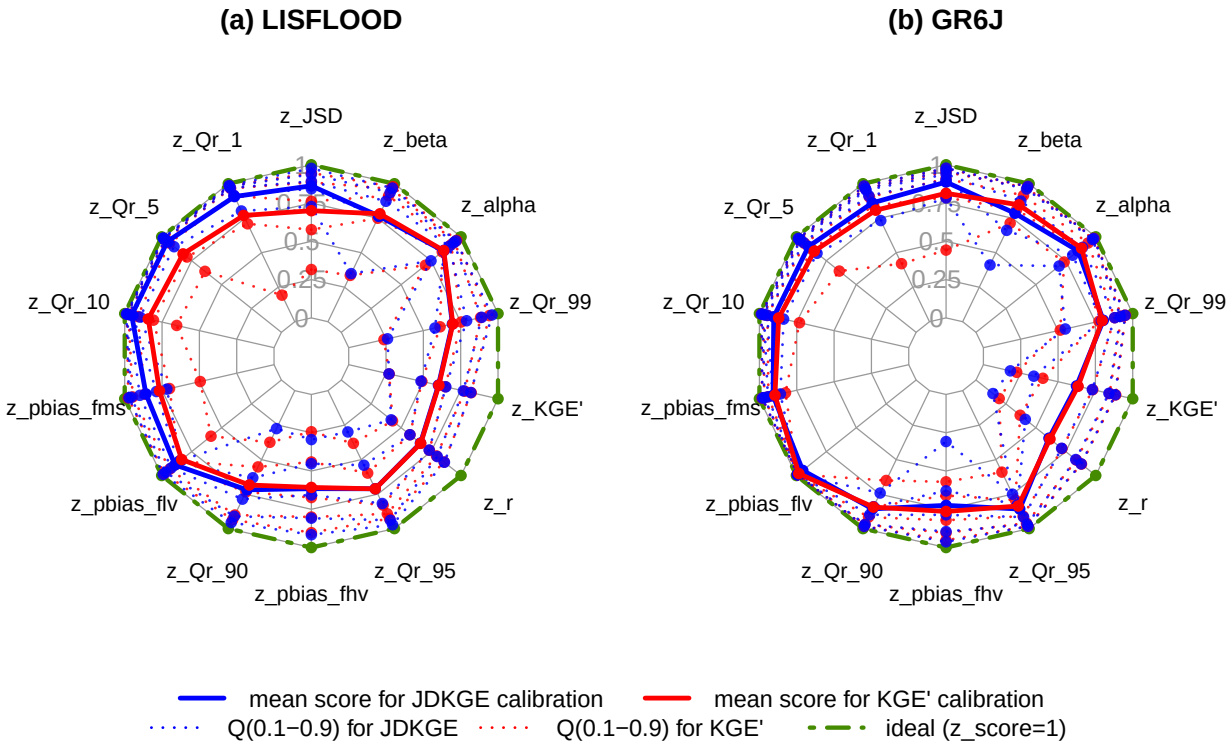

Figure 6. Radar chart of hydrological model performance scores (quantiles and averages), after normalization to a common [0-1] scale with an ideal value of 1 (see Appendix B), for: (a) OS-LISFLOOD and (b) GR6J, calibrated using either KGE' or JDKGE. Quantiles and averages are computed after normalization across the 45 global catchments for the multiple evaluation metrics. The radius of the reported quantiles (colored, dotted lines with points) correspond to the 10th, 25th, 50th, 75th and 90th percentiles of the normalized scores. For reference, the ideal radius (z=1) is drawn (green dot-dashed ring line). For LISFLOOD (panel (a)) metrics are ordered in terms of differences in means (scores obtained from calibration with JDKGE - scores from calibration with KGE'), from the largest difference (top-center) in decreasing order (anticlockwise); the same order is then used for GR6J.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

ments from the French database, selected to reflect typical patterns observed across the full sample and to highlight different behaviours in low-flow and overall performance with changing KGE′ and JSD scores.

Across the 240 catchments, approximately two thirds (65.4%) exhibit a *negligible difference in KGE′* (i.e., absolute KGE′ difference < 0.01) when moving from a KGE′-based calibration to JDKGE calibration. In contrast, 32.9% of catchments show a *non-negligible* KGE′ decrease (drop > 0.01), while for 1.7% of cases KGE′ even improves when moving to the JDKGE calibration, which is particularly striking and could be linked to equifinality and optimization issues (in finding the global optimum of KGE′). Thus, based on this distribution, we select one catchment with a KGE′ decrease larger than 0.01 (corresponding to the 80-th percentile of KGE′ drop) and a large change in JSD (about 95-th percentile), i.e., the Anglin River at Mérigny. Moreover, we select two catchments with negligible differences in KGE′, i.e., a near-zero drop ($<= 0.01$), but different levels of improvements in JSD (close to the median and 90-th percentile of all changes), i.e., the Mortagne River at Gerbéviller and Arroux River at Dracy-Saint-Loup. This choice is motivated as cases with lower changes in JSD and KGE scores would result in less illustrative examples, with more similar simulations from different calibrations.

The trade-offs between overall performance (KGE) and accuracy on low flows (JSD) are clearly visible in the Pareto fronts obtained by multi-objective calibration using KGE′ and JSD as separate objectives (Figure 7, panels a, c, e)). The Pareto-dominant solutions, i.e., optimal parameter sets identified by the multi-objective calibration with KGE' and JSD, generate simulations with markedly different low-flow performance, as illustrated by the corresponding ensembles of FDCs (Figure 7, panels (b, d, f)). These examples illustrate that KGE′ values alone are very poor indicators of low-flow performance. In fact, along the Pareto front, small gains in KGE′ are often associated with increasing divergence between simulated and observed FDCs, particularly in the low-flow segment. Simulations achieving the highest KGE′ scores tend to exhibit the largest deviations from observed low flows. Conversely, solutions with lower (better) JSD values more closely reproduce the observed FDC; moreover, the sensitivity of JDKGE to both low-flow and overall performance is also good (Figure 7.

Importantly, the streamflow simulations obtained from single-objective calibration using JDKGE consistently achieve a favourable compromise between overall performance (KGE') and distributional similarity or low-flow performance (JSD), typically located near the elbow of the Pareto front or half-way between the KGE′ optimal solution and the elbow of the curve (Figure 7, panels (a, c, e)). In other words, thanks to the integration of the JSD with the KGE' components, JDKGE approaches the best compromise solution between JSD and KGE' (Figure 7). This behaviour highlights a key benefit of the JDKGE objective function: substantial improvements in distributional similarity and low-flows accuracy (large reductions in JSD) can be achieved with only marginal changes in KGE′ (most often lower than 0.01 points of KGE'). The presence of solutions (parameter sets) with nearly identical KGE′ values but strongly contrasting JSD scores reflects the well-known problem of equifinality in hydrological modeling and its clear dependence on the objective function. By incorporating a divergence-based constraint, JDKGE effectively reduces the equifinality of model parameter sets by favouring solutions that better reproduce the full flow distribution, particularly low-flow behaviour, an aspect that is not captured by the KGE'. The streamflow simulations obtained by single-objective model calibration with the JDKGE objective function strike a good balance between low- and high-flow performance, approaching the 'best' compromise solution (i.e., the elbow of the Pareto front). Indeed, this can be

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

explained by the fact that JDKGE keeps a dominant weight on the KGE′ components rather than the JSD, due to the presence and range of the three KGE′ components which are also present in the JDKGE function.

This good balance between low- and high-flow performance is also visible in the hydrographs of regime monthly streamflow for the three exemplary catchments (Figure 8). Indeed, as expected, substantial changes in the key chosen metrics (KGE′ and JSD) have a pronounced, visible impact on hydrographs. In relative terms, looking at the spread with respect to streamflow values, the Pareto solutions differ especially in terms of low flows (Figure 8) and this pattern is systematic across most of the 240 catchments, where large gains in low-flow accuracy are observed as the JSD and JDKGE improve. During boreal summer, when low flows in Europe (as in these French examples) are critical for environmental objectives (e.g., aquatic ecosystems health) and water resource management (to satisfy water demands), giving more weight to the JSD produces more accurate simulations (Figure 8), effectively avoiding the frequent under-estimation (or the less frequent over-estimation) of baseflow that characterizes the GR6J model calibrated using KGE′.

Large absolute changes are also visible on high flows and wet-to-dry transitions for some catchments, as illustrated in Figure 8, panels (a,b)) for the three French representative catchments, but this pattern is less systematic. Regarding high flows, for most catchments no such a systematic change as for low flows is visible in hydrograph regimes when moving from KGE' to JDKGE calibration. Approximately 79% of the 240 catchments do not see large changes in extreme quantile ratios (Figure 4), as also confirmed by hydrographs inspection. About 13% of the catchments show a marked performance improvement also on high flow regimes when moving from a KGE to JDKGE calibration, while only 8% show a smaller degradation, which can be explained by a more marked trade-off between model performance on low- and high-flows, that can depend on local model suitability and potential data limitations.

Our multi-objective calibration analysis shows that while JDKGE achieves a good compromise solution, giving too much weight to the JSD component (beyond what JDKGE does) may marginally enhance low-flows at the expense of the performance on high flows. In general, in catchments where calibrating the model using JDKGE does not lead to any substantial difference in KGE′ (e.g., about 65% of the catchments with a change lower than 0.01 KGE′ points) no such trade-off is to be expected, as also seen in the Pareto fronts and hydrographs of the three exemplary cases: the model calibrated using JDKGE significantly improves the regime hydrographs on low flows compared to the calibration with the traditional KGE', while high-flow regimes are more similar. Where larger differences are found also on high flows, improvements occur in most cases and degradation in less cases. As the analysis of quantile ratios showed, for peak flows, the JDKGE-calibrated model aligns closely with the KGE benchmark, demonstrating that the proposed metric does not compromise accuracy on crucial extreme events for the sake of improving performance on lower flows.

4 Discussion

Our enhanced objective function, JDKGE, delivers improved low-flow simulations with respect to KGE' and other benchmarks, without degrading the performance at regime and high-flows on most catchments, over which it even improves in part,

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

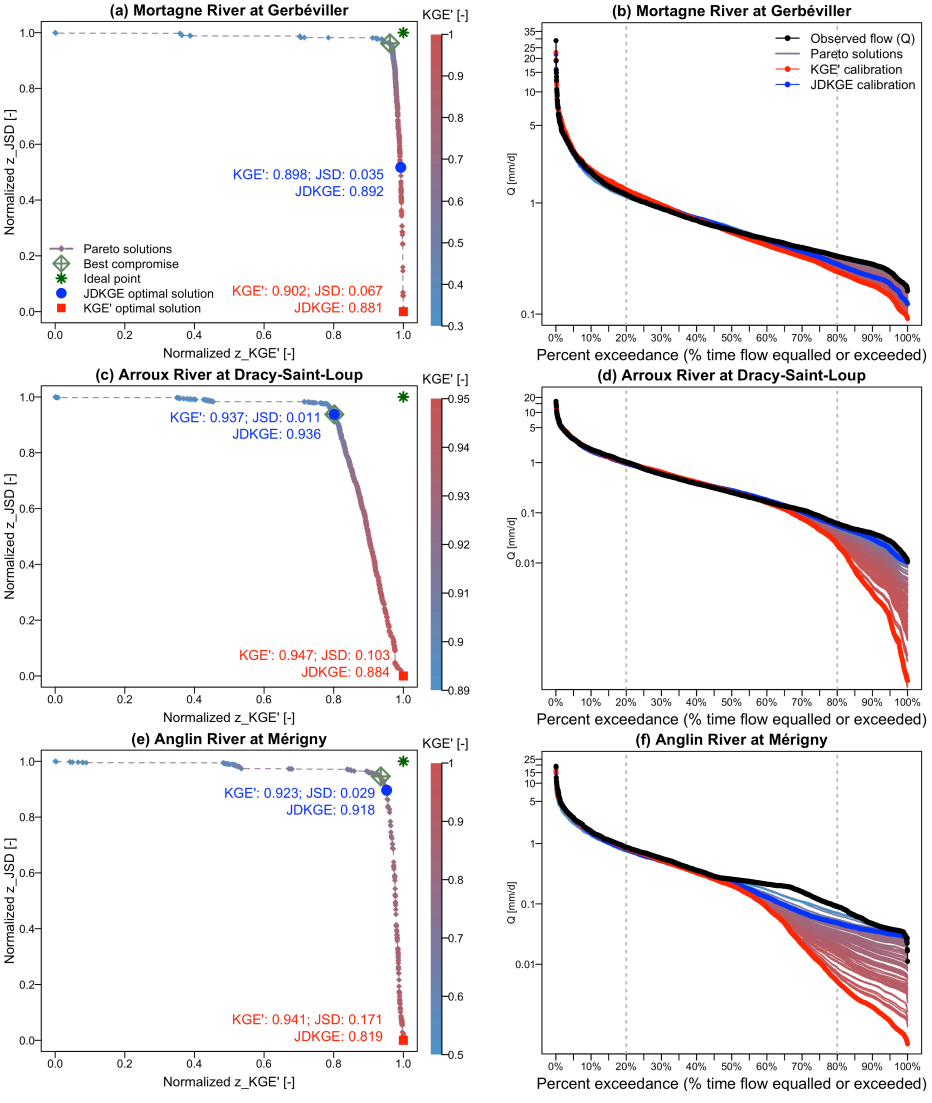

Figure 7. Multi-objective calibration results with GR6J on three exemplary French catchments, ordered by increasing changes in KGE′ and JSD between KGE′ and JDKGE optimal solutions (from top to bottom): the Mortagne River at Gerbéviller (catchment area: 493 km^2), Arroux River at Dracy-Saint-Loup (776 km^2), and Anglin River at Mérigny (1637 km^2). Left panels (a, c, e) show the Pareto fronts between z_KGE′ and z_JSD, with normalized scores in $[0, 1]$ (ideal value: 1); each coloured point (small rhombus) represents a Pareto-optimal solution and is coloured according to its raw KGE′ value (see colour bar); the solutions selected by single-objective calibration using KGE′ and JDKGE are highlighted by red squares and blue circles, respectively; the *best compromise* solution (green rotated square) identifies the closest solution to the ideal point ([1,1], asterisk). Right panels (b, d, f) show the flow duration curves (FDCs) of observed streamflow alongside the ensemble of simulated FDCs corresponding to the Pareto-optimal solutions; simulated FDCs are coloured by KGE′ value, from lower (blue) to higher KGE′ (red); the thicker red and blue curves correspond to the simulations obtained by single-objective calibration with KGE′ and JDKGE, respectively; the vertical dashed lines highlight the high-flow and low-flow FDC segments (top and bottom 20%).

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

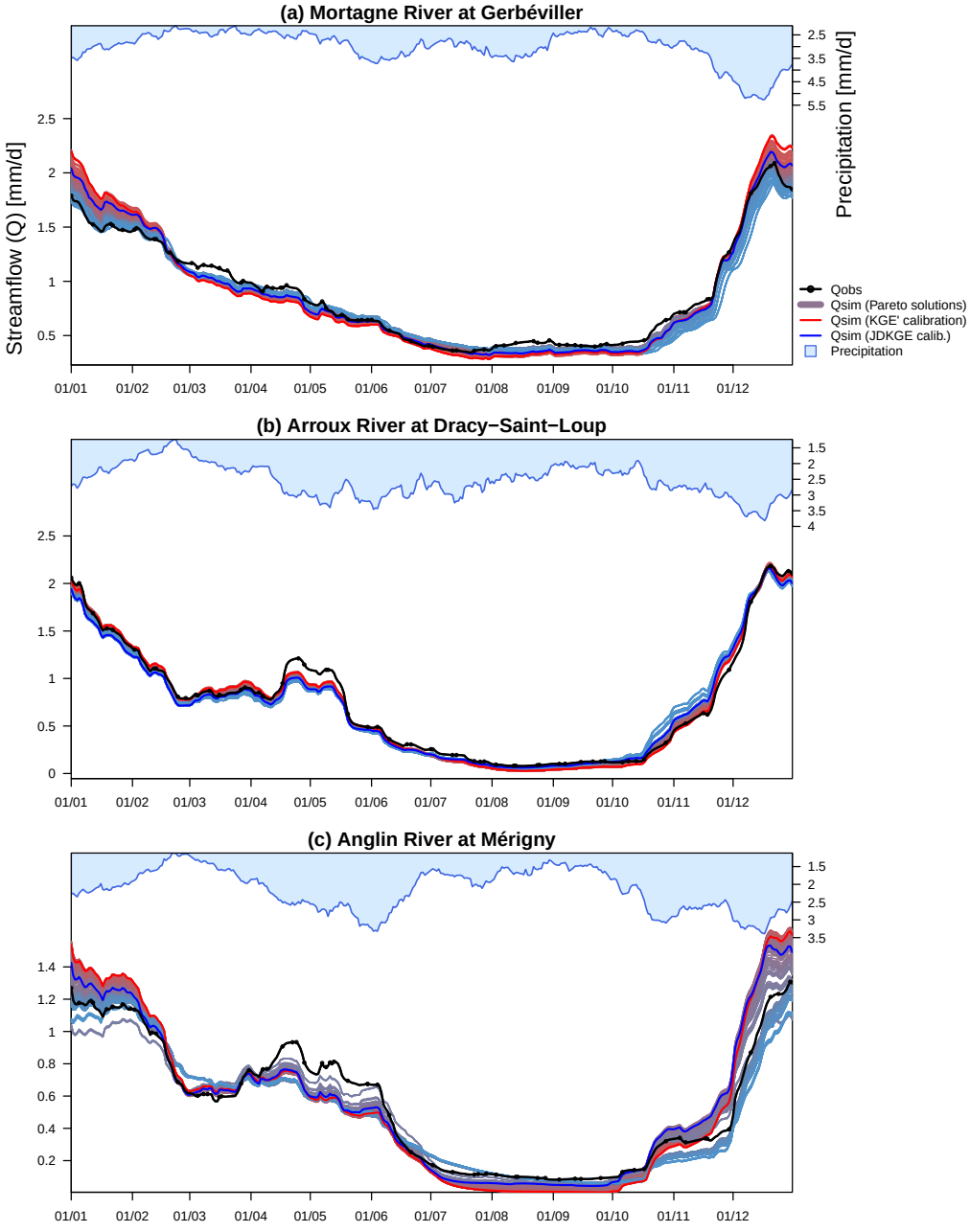

Figure 8. Hydrographs of observed and simulated monthly regime flows (31-day moving average of daily streamflow over 8 years) for three representative French catchments: the Mortagne, Arroux and Anglin Rivers (same as in Fig. 7). The ensemble of simulations (lines ranging from red to blue) shows the spread of the Pareto-optimal solutions from the multi-objective calibration of GR6J using KGE′ and JSD as separate objective functions. The single-objective optimal solutions obtained with KGE′ and JDKGE are shown as thicker red and blue lines, respectively. Observed streamflow is shown as black solid lines with point markers.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

depending on the catchment and on the section of the FDC of interest (see Figures 3, 4, 5 and 6). This represents a clear advantage over previously proposed solutions in the literature, such as Pool et al. (2018). The practical advantages of our function over existing metrics are multifaceted. Firstly, the simulations from our enhanced calibration retain strong correlation and very similar bias and variability ratios with respect to the original and modified KGE, with no significant change (Gupta et al., 2009; Kling et al., 2012). Secondly, simulations of low flows are further improved with respect to previous solutions from the literature such as the NSE_log or the KGE_NP (Pool et al., 2018), which already improve low flows over the original and modified KGE. Thirdly, the introduction of a distribution divergence component makes the calibration more robust and accurate over the whole FDC, with significant improvements in terms of meaningful hydrological signatures (Yilmaz et al., 2008) and good performance on high-flows.

The implications of our enhanced objective function are particularly valuable for hydrological models used for both drought and flood applications, such as OS-LISFLOOD, which is used in CEMS Drought observatories (EDO and GDO), as well as in the European and Global Flood Awareness Systems (EFAS and GloFAS). The next release of these systems in 2026 (CEMS EFAS v.6.0 and GloFAS v.5.0) will include the calibration of LISFLOOD based on the JDKGE objective function presented in this study. By providing a better model performance in the frequency domain, outperforming KGE' and other benchmarks on low- and high-flows, while keeping a good water balance and variability, this enhanced objective function can provide added value for multi-purpose applications, ranging from flood and drought forecasting to water management activities, with the largest benefits expected in drought-prone regions.

We expect that our enhanced objective function may also offer other benefits (beyond model performance in terms of streamflow outputs), including better identification of model parameters, with calibrated values being more constrained, and internal water fluxes being more "physically consistent" and realistic. In other words, our additional constraint, i.e., the use of a divergence component in the augmented objective function, which was added to an otherwise poorly constrained model (using only KGE as objective function), should contribute to a better representation of all the non-discharge values thanks to the equifinality reduction. These expected benefits could be further enhanced and assessed in conjunction to multivariate calibration strategies, as advocated by other authors (e.g., Pool et al., 2024). However, further work should aim to prove these expectations, as this goes beyond the scope of this study. There are some other limitations and potential areas for future work. For instance, the increased complexity of the objective function with respect to traditional metrics (NSE and KGE) may lead to difficulties in interpreting results for users unfamiliar with the divergence (JSD) component.

Related to this, in contrast to NSE which has a inherent benchmarking with respect to the mean observed flow, the benchmark value for the JDKGE function with respect to a mean observed flow reference must be evaluated explicitly. For the original KGE, Knoben et al. (2019) derived the reference value, i.e., $1 - \sqrt{2} \approx -0.41$, representing the expected score of a mean-flow benchmark. When incorporating the JSD component, this reference point shifts, and the corresponding benchmark value becomes approximately -0.65, with slight variations depending on the observed flow distribution (on the order of 10^{-2}).

Regarding a possible suggested reference value for our JSD metric, we found that in general a value of $JSD < 0.1$ corresponds to low flows that are represented in a satisfactory or acceptable way. For some particular settings of catchments and data, where lower-quality data issues or human influences on river flows are significant, and where model limitations do not

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

allow a good representation of the full flow spectrum (e.g., lack of relevant processes in the model), we noticed a greater difficulty in optimising the new JSD component (that can be detected for example by optimized JSD values remaining above 0.1). In these cases, the increased complexity of the objective function (trying to achieve a good low-flow representation) combined with a poor-quality data target or unfit-for-purpose model means (lack of structural components) may result in a degradation of performance in the more basic components of the KGE (e.g., bias ratio and relative variability).

To address these limitations, a *fallback strategy* can be implemented, where the calibration is reverted to the original (or modified) KGE if results with the more complex JDKGE are not satisfactory. Our tests suggest that when the JSD component does not fall below 0.1 at the end of the optimization, this can suggest that the model struggles to improve the low flows and any improvements in JSD would come at a larger expense of the basic KGE components (correlation, mean and variability). Thus, a simple fallback strategy that we recommend is to revert to using KGE as objective function when the optimized JSD component of JDKGE does not drop below 0.1. Other criteria may be added, by looking at whether the correlation or original (modified) KGE are satisfactory (i.e., if they are not worse for 0.1 points with respect to the best value seen during the optimization). This fallback strategy has been already implemented in the OS LISFLOOD calibration tool that will be applied for the calibration of GloFAS and EFAS, using the criteria on $JSD >$ 0.1, as well as on KGE' and r (i.e., if worse by more than 0.1 than the best value seen during the evolutionary optimization).

Furthermore, the potential sensitivity of the JDKGE metric to the discretization solution adopted for the computation of the JSD-based component (Freedman-Diaconis rule) could be investigated further, and alternative formulations of the JSD could be explored, such as estimating the densities of observations and simulations via discrete CDF derivatives (e.g., Perez-Cruz, 2008). Other potential areas for future work include analyzing the implications of our function on model parameters identifiability and on the physical consistency of modeled water balance components (e.g., Ficchì et al., 2019), assessing the benefits of our improved calibration function, possibly alongside multi-variate calibration. Future work could also further assess the potential of the new objective function for the calibration of hydrological models in snow-dominated or tropical basins, where existing models and calibration functions are particularly challenged. Some issues of KGE and NSE, such as the assumption of linearity of residuals, and the "divide and measure nonconformity" issue (Klotz et al., 2024) have not been dealt with in this study and could be addressed in future works. Finally, the new JDKGE metric could be used as loss function for machine learning (ML)-based or hybrid hydrological models, but further work is needed to ensure the differentiability of the function and test its performance in ML frameworks.

5 Conclusions

In this study, we propose and extensively validate a new enhanced metric, the Joint Divergence Kling–Gupta Efficiency (JDKGE), to be used as objective function for hydrological model calibration. The JDKGE is designed with the goal of improving performance on low flows, while maintaining balanced results along the whole flow duration curve (FDC). The JDKGE is obtained by augmenting the popular (modified) Kling-Gupta Efficiency (KGE) metric (Gupta et al., 2009; Kling et al., 2012) with a new component derived from information-theory, i.e., an estimate of the Jensen-Shannon Divergence (JSD; Lin,

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

1991). This divergence component is chosen as it enables a robust assessment of the distributional similarity between observed and simulated streamflows, via a bounded and symmetric score, which remains well defined even for empirical distributions with zero-probability regions.

Our comprehensive evaluation on 285 catchments based on 14 metrics and 100 quantile ratios (for all percentile flows) suggests that when moving from a calibration with KGE′ to JDKGE the performance on low flows improves significantly, and high-flow or regime performance remains unchanged or improves for most catchments. In particular, both high (Q90, i.e., 90-th percentile flows) and extreme peak daily flows (Q100) slightly improve on average when moving from calibrating with KGE to JDKGE. Our multi-objective calibration experiments highlight the successful mechanism behind the JDKGE: as the objectives of KGE′ and JSD vary, a substantial spread of the Pareto-optimal solutions is found, offering calibration ensembles with markedly enhanced performance on low-flows when giving more weight to the JSD optimisation, as the JDKGE does. The streamflow simulations obtained by single-objective model calibration with JDKGE strikes a good balance between accuracy on low flows and overall performance, approaching the best compromise solution (the elbow of the Pareto front) between JSD and KGE′ optimization. This means that in most cases accuracy on low flows (JSD) improves substantially with only marginal changes in KGE′.

Moreover, our new JDKGE objective function outperforms standard, popular benchmarks, as the NSE (Nash and Sutcliffe, 1970), the NSE on log-transformed flows, the KGE (in its original and modified versions), or the non-parametric KGE_NP (Pool et al., 2018) across several evaluation aspects. In particular, the JDKGE improves simulations of low flows, without compromising model performance on other regimes and particularly on high flows as other existing metrics do. While alternative metrics such as the non-parametric KGE (KGE_NP) or the NSE on log-transformed flows may still offer advantages with respect to KGE for specific aspects, like for accuracy on the mid-segment of the FDC, the JDKGE provides a more balanced and robust calibration across flow regimes, particularly for both tails of the distribution. As such, the new calibration function represents an effective solution for multi-purpose hydrological modeling, especially when models are used for both types of hydrological extremes, floods and droughts. For this reason, JDKGE is being used for the new operational calibration of OS-LISFLOOD within the flood and drought monitoring systems of Copernicus Emergency Management Service (CEMS), i.e., EFAS/GloFAS and EDO/GDO, which support both flood forecasting and drought monitoring at the European and global scales. The next versions of these systems (EFAS v.6.0 and GloFAS v5.0, to be released in 2026) will rely on the calibration of OS-LISFLOOD using JDKGE.

While in most catchments the JDKGE and KGE benchmarks yield comparable performance for high-flow events, the new objective function maintains a greater consistency with observed data across the full hydrograph range. However, as our multi-objective calibration analysis suggests, the effects of changing the objective function are not the same on all catchments. For a minor portion of catchments a non-negligible drop in KGE′ (> 0.1) can be observed when moving from a KGE- to a JDKGE-based calibration, corresponding to cases where the trade-offs between low-flow and high-flow accuracy should be more carefully analysed. In general, the biggest advantage of JDKGE lies in its ability to better resolve low-flow dynamics, which are frequently neglected in calibrations using standard objective functions, as KGE or NSE, while preserving overall

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

and high-flow performance.

Our results underscore the value of incorporating information-theoretic measures into objective functions for model calibration, as the Jensen-Shannon Divergence (JSD) into the JDKGE, as these measures provide a more holistic representation of hydrological behavior across all flow regimes. Future work could aim to further enhance the JDKGE function or propose alternative metrics targeting further improvements on both tails of the flow distribution, to improve even more high- and low-flows, or strike a different balance between performance aspects (e.g., average regime and extremes), tailored to specific use cases and modeling goals. For this, we encourage further exploration of enhanced, hybrid single-objective functions, leveraging on the power of divergence measures. While multi-objective calibration frameworks could be further used to explore trade-offs between different performance aspects (as shown in our experiments), single-objective functions, like KGE and our enhanced JDKGE, remain a practical and effective solution for multi-purpose hydrological modeling. In this context, the JDKGE objective function represents a step forward toward calibration strategies that better align with the growing operational needs for robust hydrological models supporting both flood and drought applications.

Code and data availability. The code of the hydrological models and the calibration tools are openly available: for OS-LISFLOOD on GitHub (https://github.com/ec-jrc/lisflood-code), in Python, and for GR6J (and the GR models family) in the `airGR` R package (https://cran.r-project.org/web/packages/airGR/). The code for the new metric (JDKGE) will be made available in both Python and R in a GitHub repository (https://github.com/AndreaFicchi/JDKGE) from mid-February 2026. The scripts for the analysis of the results of this article will also be made available. The multi-objective optimization experiments have been performed using the `caRamel` R package (https://cran.r-project.org/web/packages/caRamel/). The statistical testing methods are implemented in the R/Bioconductor package `waddR`, which is freely available on GitHub (https://github.com/goncalves-lab/waddR). For the French catchment database, the observed streamflow data were originally obtained from the HydroPortail platform (https://www.hydro.eaufrance.fr/, formerly Banque Hydro), while meteorological data were obtained from Météo-France. For the global catchment database, the CEMS-GloFAS and EFAS forcing data were derived from the original ERA5 and EMO-1 datasets, openly accessible respectively via the Copernicus Climate Change Service (C3S) Climate Data Store (CDS) and the JRC Earth Observation Data and Processing Platform (JEODPP). Parts of the processed forcing and streamflow data that support the findings of this study are available upon request.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

Appendix A: Calibrated parameters of GR6J and LISFLOOD

Table A1. GR6J model parameters optimized during the calibration procedure.

Parameter name	Description
X1	Maximum capacity of the soil moisture (production) store [mm]
X2	Inter-catchment groundwater exchange coefficient [mm/day]
X3	Maximum capacity of the routing store [mm]
X4	Time constant of the unit hydrograph (UH) [day]
X5	Groundwater exchange threshold [-]
X6	Exponential store depletion coefficient [mm]

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

Table A2. OS-LISFLOOD model parameters optimized during the calibration procedure (ECMWF, 2022, 2025). A flag (Setup flag) is reported representing the setups used in this study (CEMS-GLOFAS v5.0, CEMS-EFAS v5.0 or both) in which the parameters are jointly calibrated using the Distributed Evolutionary Algorithm for Python (DEAP) and the objective functions tested in this study. In GloFAS v5.0, the reservoirs parameters were tuned using a Random Forest-based approach outside of DEAP.

Parameter name	**Description**	**Setup flag (EFAS v5.0 / GloFAS v5.0 / both)**
bInfilt	Exponent in Xinanjiang equation for infiltration capacity of the soil [-]	Both
PowerPrefFlow	Exponent in the empirical function describing the preferential flow (i.e. flow that bypasses the soil matrix and drains directly to the groundwater) [-]	Both
SnowMeltCoef	Snow melt rate in degree day model equation [mm/(C day)]	Both
UpperZoneTimeConstant	Time constant for upper groundwater zone [days]	Both
GwPercValue	Maximum percolation rate from upper to lower groundwater zone [mm/day]	Both
LowerZoneTimeConstant	Time constant for lower groundwater zone [days]	Both
LZThreshold	Threshold to stop outflow from lower groundwater zone to the channel [mm]	Both
GwLoss	Maximum loss rate out of lower groundwater zone expressed as a fraction of lower zone outflow [-]	Both
TransSub	Transmission loss function parameter [-]	GloFAS v5.0
QSplitMult	Multiplier to adjust discharge triggering floodplains flow [-]	EFAS v5.0
CalChanMan1	Multiplier for channel Manning's coefficient n for riverbed (for mid-high slope rivers) [-]	Both (in GloFAS v.5.0 used only for mid-high slope rivers)
CalChanMan2	Multiplier for channel Manning's coefficient n for floodplains [-]	EFAS v.5.0
CalChanMan3	Multiplier for channel Manning's coefficient n for riverbed for mild-slope rivers [-]	GloFAS v.5.0
adjustNormalFlood	Multiplier to adjust reservoir normal filling (balance between lower and upper limit of reservoir filling) [-]	EFAS v5.0
ReservoirRnormqMult	Multiplier to adjust normal reservoir outflow [-]	EFAS v5.0
LakeMultiplier	Multiplier to adjust lake outflow [-]	Both

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

Appendix B: Normalization of evaluation metrics

To enable joint comparison and visualization of heterogeneous evaluation metrics (with different ranges, directions and optima), we map each metric to a unit interval $z \in [0,1]$, with $z = 1$ representing the ideal value. Let x denote the original value of a metric m for a given basin/model configuration. For metrics requiring min–max scaling, let $\min_m$ and $\max_m$ be the minimum and maximum values of m pooled across all basins and all model configurations being compared, and let $\mathrm{rng}_m = \max_m - \min_m$ (with rng_m set to 1 if degenerate). The normalized values z are then computed using metric-specific, direction-aware transformations and finally clipped to $[0,1]$.

Lower-is-better metrics with ideal value at 0

For metrics where smaller values indicate better performance (e.g., JSD), we use an inverted min–max transformation:

$$z = \frac{\max_m - x}{\mathrm{rng}_m}. \tag{B1}$$

Higher-is-better metrics with ideal value at 1

For metrics bounded above by 1 with an optimum at 1 (e.g., KGE′ and r), we normalize based on the (non-negative) deficit to the optimum and cap it using the pooled minimum:

$$d = \max(1 - x, 0), \qquad d_{\mathrm{cap}} = \max\big(1 - \min_m, \varepsilon\big), \qquad z = 1 - \frac{\min(d, d_{\mathrm{cap}})}{d_{\mathrm{cap}}}, \tag{B2}$$

where $\min_m$ denotes the pooled minimum of metric m across all basins and model configurations, and ε is a small constant (here $\varepsilon = 10^{-6}$) used to avoid division by zero.

Two-sided metrics with ideal value at 1 (ratio-type metrics)

For ratio-type (strictly positive) metrics with an ideal value of 1 (e.g., quantile ratios and KGE components such as α and β), we adopt a multiplicative, log-symmetric deviation measure, so that deviations of x and $1/x$ from the ideal value are penalized equally. Let $x^+ = \max(x, \varepsilon)$ and define the deviation as

$$d = \left|\log\big(x^+\big)\right|. \tag{B3}$$

To limit the influence of extreme outliers (below the 1st or above the 99th percentile) and to ensure stable scaling across datasets, the deviation is capped using robust quantiles of the pooled distribution of x^+:

$$d_{\mathrm{cap}} = \max\big(\,|\log(q_\ell)|,\, |\log(q_h)|,\, d_{\min}\big), \qquad z = 1 - \frac{\min(d, d_{\mathrm{cap}})}{d_{\mathrm{cap}}}, \tag{B4}$$

where q_ℓ and q_h are the ℓ and h empirical quantiles of the pooled sample (here $\ell = 0.01$ and $h = 0.99$), and $d_{\min} > 0$ is a small floor ($d_{\min} = \log(1 + 0.01)$) that prevents very narrow spreads from causing excessive penalization.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

Two-sided metrics with ideal value at 0 (bias-type metrics)

For metrics with an ideal value of 0 and a two-sided distribution around this ideal value (e.g., FDC-based percent biases), we normalize the absolute deviation from the optimum:

$$d = |x|, \qquad d_{\mathrm{cap}} = \max\big(q_h(|x|), d_{\min}\big), \qquad z = 1 - \frac{\min(d, d_{\mathrm{cap}})}{d_{\mathrm{cap}}}, \tag{B5}$$

where $q_h(|x|)$ is a robust upper quantile of the pooled absolute deviations (here $h = 0.99$), and $d_{\min} > 0$ is a small floor value (here $d_{\min} = 0.01$). Using a high quantile to define d_{cap} limits the influence of extreme outliers that could otherwise dominate the normalization and compress the range of typical values. The additional floor $d_{\min}$ ensures numerical stability when the pooled spread of deviations is very small (e.g., in the unlikely case when all biases are uniformly close to zero), preventing disproportionate sensitivity to minor numerical differences. The chosen floor corresponds to a small but practically meaningful deviation (1%), preserving resolution among near-optimal scores while avoiding instability in the normalized metric.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

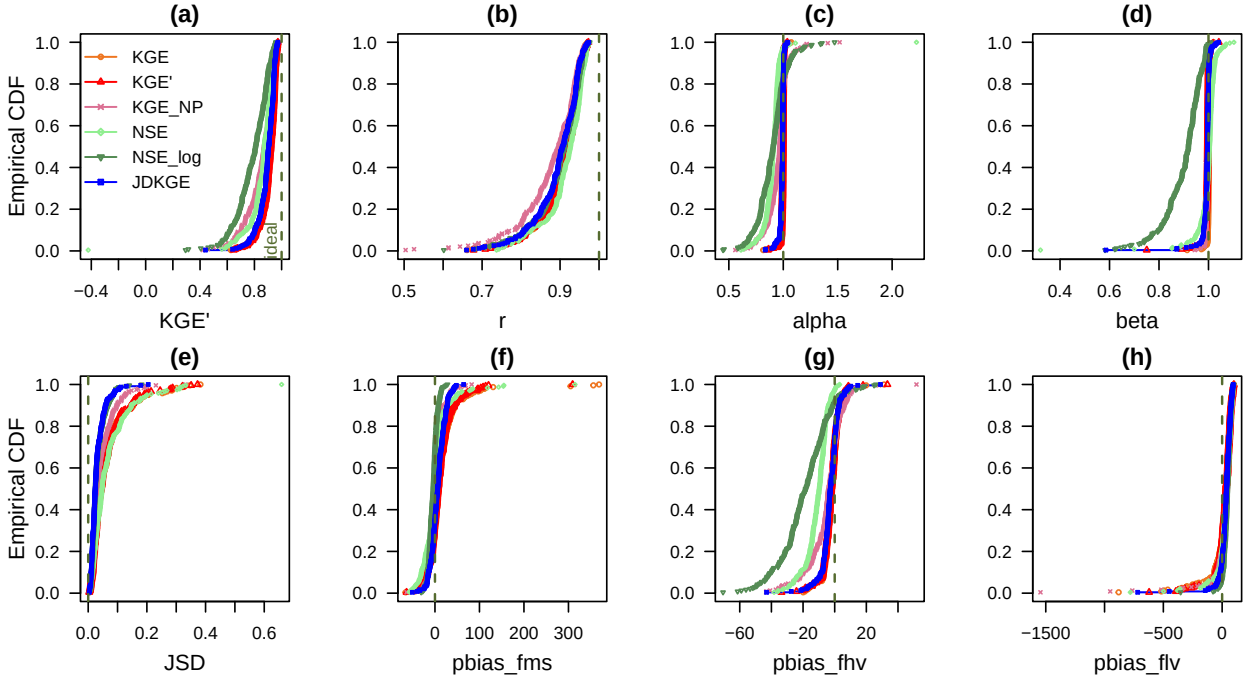

Figure A1. Empirical cumulative distribution functions (ECDFs) of eight performance evaluation metrics across the 240 French catchments for 6 calibration objective functions (KGE, KGE′, NSE, NSE_log, KGE_NP, and JDKGE) using the GR6J model. The evaluation metrics are: (a) KGE′, (b) Pearson correlation (r), (c) variability ratio (alpha), (d) bias ratio (beta), (e) JSD, and (f-h) FDC-based percent biases. The ideal value is marked by a vertical (dashed) line.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

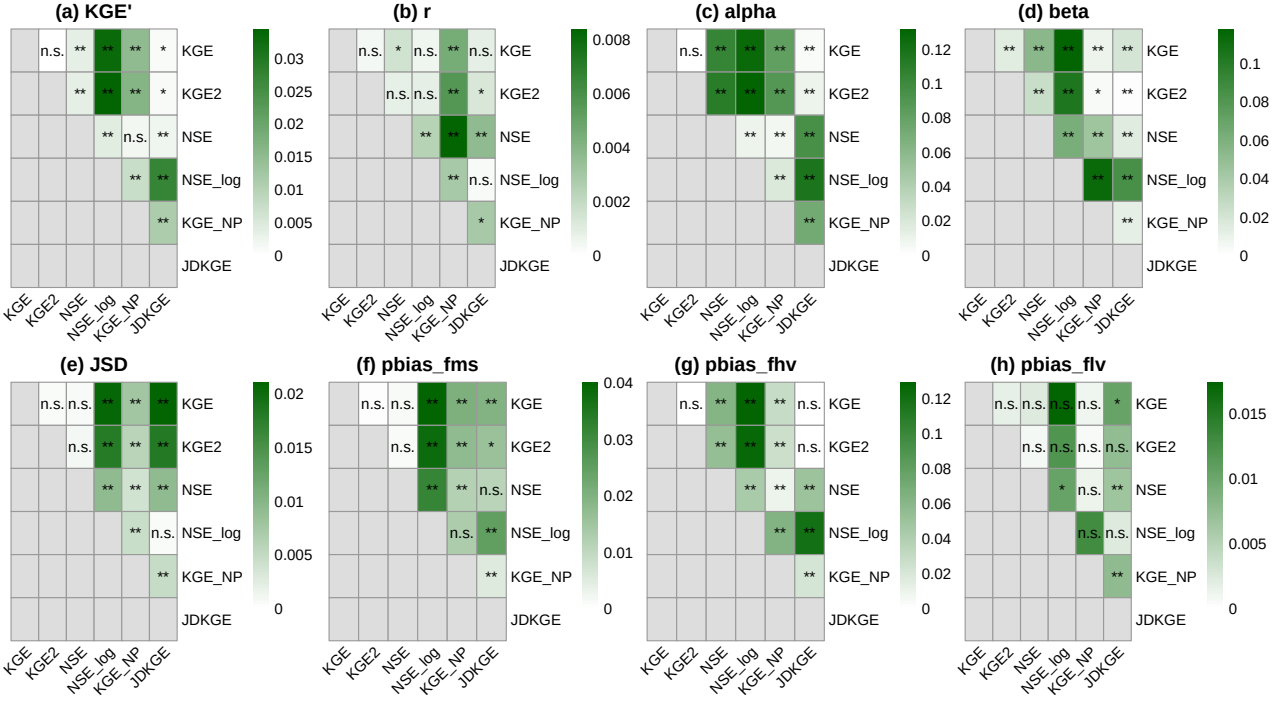

Figure A2. Heatmaps showing pairwise statistical comparisons between empirical CDFs of eight performance evaluation metrics across the 240 French catchments for six different calibration functions (KGE, KGE′, NSE, NSE_log, KGE_NP, and JDKGE) using the GR6J model. Statistical differences for each pair of calibration objective functions are analysed using the 2-Wasserstein distance-based test with asymptotic p-values: in each cell, "**" denotes strong evidence ($p < 0.01$), "*" moderate evidence (less significant differences), and "n.s." indicates non-significant differences. Each cell's color represents the squared 2-Wasserstein distance (wd) between the ECDFs of each pair of calibration functions. The evaluation metrics are: (a) KGE′, (b) Pearson correlation (r), (c) variability ratio (alpha), (d) bias ratio (beta), (e) JSD, and (f-h) FDC-based percent biases.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

Author contributions. AF, DB and AT led the design of the novel calibration objective function (JDKGE). AF and DB co-led the interpretation of results and the writing of the manuscript, with AF preparing the first draft. DB and AT formulated a preliminary framework for potential enhancements of the KGE metric. AF, DB, SG, FM contributed significantly to the conceptualization of the study's methodology and its implementation and testing phases. AF designed and carried out the GR6J (and other GR models) calibration experiments and simulations, and developed the code for the analyses of results for both GR6J and LISFLOOD models. SG, FM, CR designed the LISFLOOD model configurations and led the LISFLOOD calibration experiments, developing code for the operational implementation of the calibration workflow. AT supervised the project, alongside PS and AP. All authors discussed the results and contributed to the final manuscript.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Acknowledgements. We thank Christel Prudhomme, Cinzia Mazzetti, and colleagues at ECMWF and JRC for their work on the Copernicus Emergency Management Service (CEMS) EFAS v6.0 and GloFAS v5.0 calibration prototypes using JDKGE and KGE′ as objective functions, which supported the validation of JDKGE as chosen objective function for the OS-LISFLOOD model calibration and its operational implementation in CEMS-EFAS and GloFAS.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

References

Admasu, L. M., Grant, L., and Thiery, W.: Exploring Global Climate Model Downscaling Based on Tile-Level Output, Journal of Applied Meteorology and Climatology, 62, 171 – 190, https://doi.org/10.1175/JAMC-D-21-0265.1, place: Boston MA, USA Publisher: American Meteorological Society, 2023.

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., and Pappenberger, F.: GloFAS – global ensemble streamflow forecasting and flood early warning, Hydrology and Earth System Sciences, 17, 1161–1175, https://doi.org/10.5194/hess-17-1161-2013, 2013.

Arsenault, R., Brissette, F., and Martel, J.-L.: The hazards of split-sample validation in hydrological model calibration, Journal of Hydrology, 566, 346–362, https://doi.org/10.1016/j.jhydrol.2018.09.027, 2018.

Baatz, R., Ghazaryan, G., Hagenlocher, M., Nendel, C., Toreti, A., and Rezaei, E. E.: Drought research priorities, trends, and geographic patterns, Hydrology and Earth System Sciences, 29, 1379–1393, https://doi.org/10.5194/hess-29-1379-2025, 2025.

Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., and Andreassian, V.: Characterising performance of environmental models, Environmental Modelling & Software, 40, 1–20, https://doi.org/10.1016/j.envsoft.2012.09.011, 2013.

Beven, K. J.: A short history of philosophies of hydrological model evaluation and hypothesis testing, WIREs Water, 12, e1761, https://doi.org/10.1002/wat2.1761, publisher: John Wiley & Sons, Ltd, 2025.

Bisselink, B., Bernhard, J., Gelati, E., Adamovic, M., Guenther, S., Mentaschi, L., Feyen, L., and de Roo, A.: Climate change and Europe's water resources, Tech. Rep. EUR 29951 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-10398-1, https://doi.org/10.2760/15553, jRC118586, 2020.

Booij, M. J. and Krol, M. S.: Balance between calibration objectives in a conceptual hydrological model, Hydrological Sciences Journal, 55, 1017–1032, https://doi.org/10.1080/02626667.2010.505892, publisher: Taylor & Francis, 2010.

Briët, J. and Harremoës, P.: Properties of classical and quantum Jensen-Shannon divergence, Physical Review A, 79, 052 311, https://doi.org/10.1103/PhysRevA.79.052311, publisher: American Physical Society, 2009.

Cammalleri, C., Vogt, J., and Salamon, P.: Development of an operational low-flow index for hydrological drought monitoring over Europe, Hydrological Sciences Journal, 62, 346–358, https://doi.org/10.1080/02626667.2016.1240869, publisher: Taylor & Francis, 2017.

Cammalleri, C., Barbosa, P., and Vogt, J. V.: Evaluating simulated daily discharge for operational hydrological drought monitoring in the Global Drought Observatory (GDO), Hydrological Sciences Journal, 65, 1316–1325, https://doi.org/10.1080/02626667.2020.1747623, publisher: Taylor & Francis, 2020.

Cantoni, E., Tramblay, Y., Grimaldi, S., Salamon, P., Dakhlaoui, H., Dezetter, A., and Thiemig, V.: Hydrological performance of the ERA5 reanalysis for flood modeling in Tunisia with the LISFLOOD and GR4J models, Journal of Hydrology: Regional Studies, 42, 101 169, https://doi.org/10.1016/j.ejrh.2022.101169, 2022.

Chang, A. Y.-Y., Ramos, M.-H., Harrigan, S., Prudhomme, C., Tilmant, F., Domeisen, D. I., and Zappa, M.: Exploring hydrological system performance for alpine low flows in local and continental prediction systems, Journal of Hydrology: Regional Studies, 56, 102 056, https://doi.org/10.1016/j.ejrh.2024.102056, 2024.

Cinkus, G., Mazzilli, N., Jourde, H., Wunsch, A., Liesch, T., Ravbar, N., Chen, Z., and Goldscheider, N.: When best is the enemy of good – critical evaluation of performance criteria in hydrological models, Hydrology and Earth System Sciences, 27, 2397–2411, https://doi.org/10.5194/hess-27-2397-2023, 2023.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

Cislaghi, A., Masseroni, D., Massari, C., Camici, S., and Brocca, L.: Combining a rainfall–runoff model and a regionalization approach for flood and water resource assessment in the western Po Valley, Italy, Hydrological Sciences Journal, 65, 348–370, https://doi.org/10.1080/02626667.2019.1690656, 2020.

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., Shook, K. R., and Papalexiou, S. M.: The Abuse of Popular Performance Metrics in Hydrologic Modeling, Water Resources Research, 57, e2020WR029 001, https://doi.org/10.1029/2020WR029001, publisher: John Wiley & Sons, Ltd, 2021.

Clarke, R.: A review of some mathematical models used in hydrology, with observations on their calibration and use, Journal of Hydrology, 19, 1–20, https://doi.org/10.1016/0022-1694(73)90089-9, 1973.

Coron, L., Thirel, G., Delaigue, O., Perrin, C., and Andréassian, V.: The suite of lumped GR hydrological models in an R package, Environmental Modelling & Software, 94, 166–171, https://doi.org/10.1016/j.envsoft.2017.05.002, 2017.

Cover, T. M. and Thomas, J. A.: Entropy, Relative Entropy, and Mutual Information, in: Elements of Information Theory, pp. 13–55, Publisher: John Wiley & Sons, Ltd, ISBN 978-0-471-74882-3, https://doi.org/10.1002/047174882X.ch2, 2005.

De Rainville, F.-M., Fortin, F.-A., Gardner, M.-A., Parizeau, M., and Gagné, C.: DEAP: a python framework for evolutionary algorithms, in: Proceedings of the 14th Annual Conference Companion on Genetic and Evolutionary Computation, GECCO '12, pp. 85–92, Association for Computing Machinery, New York, NY, USA, ISBN 978-1-4503-1178-6, https://doi.org/10.1145/2330784.2330799, event-place: Philadelphia, Pennsylvania, USA, 2012.

De Rainville, F.-M., Fortin, F.-A., Gardner, M.-A., Parizeau, M., and Gagné, C.: DEAP: enabling nimbler evolutions, SIGEVOlution, 6, 17–26, https://doi.org/10.1145/2597453.2597455, place: New York, NY, USA Publisher: Association for Computing Machinery, 2014.

De Roo, A. P. J., Wesseling, C. G., and Van Deursen, W. P. A.: Physically based river basin modelling within a GIS: the LISFLOOD model, Hydrological Processes, 14, 1981–1992, https://doi.org/10.1002/1099-1085(20000815/30)14:11/12<1981::AID-HYP49>3.0.CO;2-F, publisher: John Wiley & Sons, Ltd, 2000.

Deckers, D. L. E. H., Booij, M. J., Rientjes, T. H. M., and Krol, M. S.: Catchment Variability and Parameter Estimation in Multi-Objective Regionalisation of a Rainfall–Runoff Model, Water Resources Management, 24, 3961–3985, https://doi.org/10.1007/s11269-010-9642-8, 2010.

Diop, S. B., Ekolu, J., Tramblay, Y., Dieppois, B., Grimaldi, S., Bodian, A., Blanchet, J., Rameshwaran, P., Salamon, P., and Sultan, B.: Climate change impacts on floods in West Africa: new insight from two large-scale hydrological models, Natural Hazards and Earth System Sciences, 25, 3161–3184, https://doi.org/10.5194/nhess-25-3161-2025, 2025.

Döll, P., Hasan, H. M. M., Schulze, K., Gerdener, H., Börger, L., Shadkam, S., Ackermann, S., Hosseini-Moghari, S.-M., Müller Schmied, H., Güntner, A., and Kusche, J.: Leveraging multi-variable observations to reduce and quantify the output uncertainty of a global hydrological model: evaluation of three ensemble-based approaches for the Mississippi River basin, Hydrology and Earth System Sciences, 28, 2259–2295, https://doi.org/10.5194/hess-28-2259-2024, 2024.

ECMWF: EFAS v5.0 - Calibration Methodology and Data, https://confluence.ecmwf.int/display/CEMS/EFAS+v5.0+-+Calibration+Methodology+and+Data, accessed May 25, 2025, 2022.

ECMWF: GloFAS versioning system, https://confluence.ecmwf.int/display/CEMS/GloFAS+versioning+system, accessed September 25, 2025, 2025.

Ekolu, J., Dieppois, B., Diop, S. B., Bodian, A., Grimaldi, S., Salamon, P., Villarini, G., Eden, J. M., Monerie, P.-A., van de Wiel, M., and Tramblay, Y.: How could climate change affect the magnitude, duration and frequency of hydrological droughts and floods in West Africa during the 21st century? A storyline approach, Journal of Hydrology, 660, 133 482, https://doi.org/10.1016/j.jhydrol.2025.133482, 2025.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

Evin, G., Thyer, M., Kavetski, D., McInerney, D., and Kuczera, G.: Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, Water Resources Research, 50, 2350–2375, https://doi.org/10.1002/2013WR014185, publisher: John Wiley & Sons, Ltd, 2014.

Fenicia, F., Solomatine, D. P., Savenije, H. H. G., and Matgen, P.: Soft combination of local models in a multi-objective framework, Hydrology and Earth System Sciences, 11, 1797–1809, https://doi.org/10.5194/hess-11-1797-2007, 2007.

Ficchì, A., Perrin, C., and Andréassian, V.: Impact of temporal resolution of inputs on hydrological model performance: An analysis based on 2400 flood events, Journal of Hydrology, 538, 454–470, https://doi.org/https://doi.org/10.1016/j.jhydrol.2016.04.016, 2016.

Ficchì, A., Perrin, C., and Andréassian, V.: Hydrological modelling at multiple sub-daily time steps: Model improvement via flux-matching, Journal of Hydrology, 575, 1308–1327, https://doi.org/https://doi.org/10.1016/j.jhydrol.2019.05.084, 2019.

Fortin, F.-A., Rainville, F.-M. D., Gardner, M.-A., Parizeau, M., and Gagné, C.: DEAP: Evolutionary Algorithms Made Easy, Journal of Machine Learning Research, 13, 2171–2175, http://jmlr.org/papers/v13/fortin12a.html, 2012.

Freedman, D. and Diaconis, P.: On the histogram as a density estimator: L2 theory, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 57, 453–476, https://doi.org/10.1007/BF01025868, 1981.

Fu, T. and Zhang, C.: Towards a generic model evaluation metric for non-normally distributed measurements in water quality and ecosystem models, Ecological Informatics, 80, 102 470, https://doi.org/10.1016/j.ecoinf.2024.102470, 2024.

Galelli, S., Turner, S. W. D., Pokhrel, Y., Yi Ng., J., Castelletti, A., Bierkens, M. F. P., Pianosi, F., and Biemans, H.: Advancing the Representation of Human Actions in Large-Scale Hydrological Models: Challenges and Future Research Directions, Water Resources Research, 61, e2024WR039 486, https://doi.org/10.1029/2024WR039486, publisher: John Wiley & Sons, Ltd, 2025.

Garcia, F., Folton, N., 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, publisher: Taylor & Francis, 2017.

Gauch, M., Kratzert, F., Gilon, O., Gupta, H., Mai, J., Nearing, G., Tolson, B., Hochreiter, S., and Klotz, D.: In Defense of Metrics: Metrics Sufficiently Encode Typical Human Preferences Regarding Hydrological Model Performance, Water Resources Research, 59, e2022WR033 918, https://doi.org/10.1029/2022WR033918, publisher: John Wiley & Sons, Ltd, 2023.

Giuliani, M., Li, Y., Castelletti, A., and Gandolfi, C.: A coupled human-natural systems analysis of irrigated agriculture under changing climate, Water Resources Research, https://doi.org/10.1002/2016WR019363, 2016.

Gomes, G., Salamon, P., Lemke, C.-D., Sperzel, T., Radke-Fretz, M., Schweim, C., Ziese, M., Russo, C., and Grimaldi, S.: EMO: A high-resolution multi-variable gridded meteorological data set for Europe, https://doi.org/10.2905/0BD84BE4-CEC8-4180-97A6-8B3ADAAC4D26, 2025.

Guo, X.: JS-MA: A Jensen-Shannon Divergence Based Method for Mapping Genome-Wide Associations on Multiple Diseases, Frontiers in Genetics, Volume 11 - 2020, https://doi.org/10.3389/fgene.2020.507038, 2020.

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/https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Gustard, A., Bullock, A., and Dixon, J.: Low Flow Estimation in the United Kingdom, Tech. Rep. Report 108, UK Institute of Hydrology, 1992.

Hallouin, T., Bruen, M., and O'Loughlin, F. E.: Calibration of hydrological models for ecologically relevant streamflow predictions: a trade-off between fitting well to data and estimating consistent parameter sets?, Hydrology and Earth System Sciences, 24, 1031–1054, https://doi.org/10.5194/hess-24-1031-2020, 2020.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

Hannaford, J., Mackay, J. D., Ascott, M., Bell, V. A., Chitson, T., Cole, S., Counsell, C., Durant, M., Jackson, C. R., Kay, A. L., Lane, R. A., Mansour, M., Moore, R., Parry, S., Rudd, A. C., Simpson, M., Facer-Childs, K., Turner, S., Wallbank, J. R., Wells, S., and Wilcox, A.: The enhanced future Flows and Groundwater dataset: development and evaluation of nationally consistent hydrological projections based on UKCP18, Earth System Science Data, 15, 2391–2415, https://doi.org/10.5194/essd-15-2391-2023, 2023.

Harrigan, S., Prudhomme, C., Parry, S., Smith, K., and Tanguy, M.: Benchmarking ensemble streamflow prediction skill in the UK, Hydrology and Earth System Sciences, 22, 2023–2039, https://doi.org/10.5194/hess-22-2023-2018, 2018.

Hipsey, M. R., Gal, G., Arhonditsis, G. B., Carey, C. C., Elliott, J. A., Frassl, M. A., Janse, J. H., de Mora, L., and Robson, B. J.: A system of metrics for the assessment and improvement of aquatic ecosystem models, Environmental Modelling & Software, 128, 104 697, https://doi.org/10.1016/j.envsoft.2020.104697, 2020.

Jensen, L., Dill, R., Balidakis, K., Grimaldi, S., Salamon, P., and Dobslaw, H.: Global 0.05° water storage simulations with the OS LISFLOOD hydrological model for geodetic applications, Geophysical Journal International, 241, 1840–1852, https://doi.org/10.1093/gji/ggaf129, 2025.

Khu, S. T. and Madsen, H.: Multiobjective calibration with Pareto preference ordering: An application to rainfall-runoff model calibration, Water Resources Research, 41, 2005.

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.

Klotz, D., Gauch, M., Kratzert, F., Nearing, G., and Zscheischler, J.: Technical Note: The divide and measure nonconformity – how metrics can mislead when we evaluate on different data partitions, Hydrology and Earth System Sciences, 28, 3665–3673, https://doi.org/10.5194/hess-28-3665-2024, 2024.

Knijff, J. M. V. D., Younis, J., and Roo, A. P. J. D.: LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation, International Journal of Geographical Information Science, 24, 189–212, https://doi.org/10.1080/13658810802549154, publisher: Taylor & Francis, 2010.

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrology and Earth System Sciences, 23, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019, 2019.

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., and Woods, R. A.: A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments, Water Resources Research, 56, e2019WR025 975, https://doi.org/10.1029/2019WR025975, publisher: John Wiley & Sons, Ltd, 2020.

Krause, P., Boyle, D. P., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Advances in Geosciences, 5, 89–97, https://doi.org/10.5194/adgeo-5-89-2005, 2005.

Kullback, S. and Leibler, R. A.: On Information and Sufficiency, The Annals of Mathematical Statistics, 22, 79–86, https://doi.org/10.1214/aoms/1177729694, 1951.

Li, Y., Grimaldi, S., Pauwels, V. R., and Walker, J. P.: Hydrologic model calibration using remotely sensed soil moisture and discharge measurements: The impact on predictions at gauged and ungauged locations, Journal of Hydrology, 557, 897–909, https://doi.org/10.1016/j.jhydrol.2018.01.013, 2018.

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, 05017 023, https://doi.org/10.1061/(ASCE)HE.1943-5584.0001580, _eprint: https://ascelibrary.org/doi/pdf/10.1061/%28ASCE%29HE.1943-5584.0001580, 2017.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

Lin, J.: Divergence measures based on the Shannon entropy, IEEE Transactions on Information Theory, 37, 145–151, https://doi.org/10.1109/18.61115, 1991.

Liu, C., Xu, C., Zhang, Z., Xiong, S., Zhang, W., Zhang, B., Chen, H., Xu, Y., and Wang, S.: Modeling hydrological consequences of 21st-Century climate and land use/land cover changes in a mid-high latitude watershed, Geoscience Frontiers, 15, 101 819, https://doi.org/10.1016/j.gsf.2024.101819, 2024.

Loritz, R., Kleidon, A., Jackisch, C., Westhoff, M., Ehret, U., Gupta, H., and Zehe, E.: A topographic index explaining hydrological similarity by accounting for the joint controls of runoff formation, Hydrology and Earth System Sciences, 23, 3807–3821, https://doi.org/10.5194/hess-23-3807-2019, 2019.

Madsen, H.: Automatic calibration of a conceptual rainfall–runoff model using multiple objectives, Journal of Hydrology, 235, 276–288, https://doi.org/10.1016/S0022-1694(00)00279-1, 2000.

Matthews, G., Baugh, C., Barnard, C., Carton De Wiart, C., Colonese, J., Grimaldi, S., Ham, D., Hansford, E., Harrigan, S., Heiselberg, S., Hooker, H., Hossain, S., Mazzetti, C., Milano, L., Moschini, F., O'Regan, K., Pappenberger, F., Pfister, D., Rajbhandari, R. M., Salamon, P., Ramos, A., Shelton, K., Stephens, E., Tasev, D., Turner, M., van den Homberg, M., Wittig, J., Zsótér, E., and Prudhomme, C.: Chapter 15 - On the operational implementation of the Global Flood Awareness System (GloFAS), in: Flood Forecasting (Second Edition), edited by Adams, T. E., Gangodagamage, C., and Pagano, T. C., pp. 299–350, Academic Press, ISBN 978-0-443-14009-9, https://doi.org/10.1016/B978-0-443-14009-9.00014-6, 2025a.

Matthews, G., Baugh, C., Barnard, C., De Wiart, C. C., Colonese, J., Decremer, D., Grimaldi, S., Hansford, E., Mazzetti, C., O'Regan, K., Pappenberger, F., Ramos, A., Salamon, P., Tasev, D., and Prudhomme, C.: Chapter 14 - On the operational implementation of the European Flood Awareness System (EFAS), in: Flood Forecasting (Second Edition), edited by Adams, T. E., Gangodagamage, C., and Pagano, T. C., pp. 251–298, Academic Press, ISBN 978-0-443-14009-9, https://doi.org/10.1016/B978-0-443-14009-9.00005-5, 2025b.

Mei, Y., Mai, J., Do, H. X., Gronewold, A., Reeves, H., Eberts, S., Niswonger, R., Regan, R. S., and Hunt, R. J.: Can Hydrological Models Benefit From Using Global Soil Moisture, Evapotranspiration, and Runoff Products as Calibration Targets?, Water Resources Research, 59, e2022WR032 064, https://doi.org/10.1029/2022WR032064, publisher: John Wiley & Sons, Ltd, 2023.

Melsen, L. A., Puy, A., Torfs, P. J. J. F., and and, A. S.: The rise of the Nash-Sutcliffe Efficiency in hydrology, Hydrological Sciences Journal, 0, https://doi.org/10.1080/02626667.2025.2475105, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02626667.2025.2475105, 2025.

Michel, C.: Hydrologie appliquée aux petits bassins ruraux, Hydrology handbook, CEMAGREF, Antony, France, 1991.

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.

Monteil, C., Zaoui, F., Le Moine, N., and Hendrickx, F.: Multi-objective calibration by combination of stochastic and gradient-like parameter generation rules – the caRamel algorithm, Hydrology and Earth System Sciences, 24, 3189–3209, https://doi.org/10.5194/hess-24-3189-2020, 2020.

Murphy, A. H.: Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient, Monthly Weather Review, 116, 2417–2424, https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2, place: Boston MA, USA Publisher: American Meteorological Society, 1988.

Nash, J. 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.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

Nearing, G. S. and Gupta, H. V.: The quantity and quality of information in hydrologic models, Water Resources Research, 51, 524–538, https://doi.org/10.1002/2014WR015895, publisher: John Wiley & Sons, Ltd, 2015.

Netzel, P., Tyminska, L., Dana Feleha, D., and Socha, J.: New approach to assess forest fragmentation based on multiscale similarity index, Ecological Indicators, 158, 111 530, https://doi.org/10.1016/j.ecolind.2023.111530, 2024.

Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., Le Lay, M., Besson, F., Soubeyroux, J.-M., Viel, C., Regimbeau, F., Andréassian, V., Maugis, P., Augeard, B., and Morice, E.: Benchmarking hydrological models for low-flow simulation and forecasting on French catchments, Hydrology and Earth System Sciences, 18, 2829–2857, https://doi.org/10.5194/hess-18-2829-2014, 2014.

Nicolle, P., Besson, F., Delaigue, O., Etchevers, P., François, D., Le Lay, M., Perrin, C., Rousset, F., Thiéry, D., Tilmant, F., Magand, C., Leurent, T., and Jacob, E.: PREMHYCE: An operational tool for low-flow forecasting, Proceedings of the International Association of Hydrological Sciences, 383, 381–389, https://doi.org/10.5194/piahs-383-381-2020, 2020.

Nicótina, L., Alessi Celegon, E., Rinaldo, A., and Marani, M.: On the impact of rainfall patterns on the hydrologic response, Water Resources Research, 44, https://doi.org/https://doi.org/10.1029/2007WR006654, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2007WR006654, 2008.

Olsen, M., Troldborg, L., Henriksen, H. J., Conallin, J., Refsgaard, J. C., and Boegh, E.: Evaluation of a typical hydrological model in relation to environmental flows, Journal of Hydrology, 507, 52–62, https://doi.org/10.1016/j.jhydrol.2013.10.022, 2013.

Orth, R., Staudinger, M., Seneviratne, S. I., Seibert, J., and Zappa, M.: Does model performance improve with complexity? A case study with three hydrological models, Journal of Hydrology, 523, 147–159, https://doi.org/10.1016/j.jhydrol.2015.01.044, 2015.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall–runoff model? Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling, Journal of Hydrology, 303, 290–306, https://doi.org/10.1016/j.jhydrol.2004.08.026, 2005.

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., and Michel, C.: Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations, Water Resources Research, 42, https://doi.org/10.1029/2005WR004636, publisher: John Wiley & Sons, Ltd, 2006.

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, publisher: John Wiley & Sons, Ltd, 2014.

Pelletier, A. and Andréassian, V.: On constraining a lumped hydrological model with both piezometry and streamflow: results of a large sample evaluation, Hydrology and Earth System Sciences, 26, 2733–2758, https://doi.org/10.5194/hess-26-2733-2022, 2022.

Perez-Cruz, F.: Kullback-Leibler divergence estimation of continuous distributions, in: 2008 IEEE International Symposium on Information Theory, pp. 1666–1670, ISBN 2157-8117, https://doi.org/10.1109/ISIT.2008.4595271, journal Abbreviation: 2008 IEEE International Symposium on Information Theory, 2008.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275–289, https://doi.org/10.1016/S0022-1694(03)00225-7, 2003.

Pfannerstill, M., Guse, B., and Fohrer, N.: Smart low flow signature metrics for an improved overall performance evaluation of hydrological models, Journal of Hydrology, 510, 447–458, https://doi.org/10.1016/j.jhydrol.2013.12.044, 2014.

Pool, S., Vis, M. J. P., Knight, R. R., and Seibert, J.: Streamflow characteristics from modeled runoff time series – importance of calibration criteria selection, Hydrology and Earth System Sciences, 21, 5443–5457, https://doi.org/10.5194/hess-21-5443-2017, 2017.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

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, publisher: Taylor & Francis, 2018.

Pool, S., Fowler, K., and Peel, M.: Benefit of Multivariate Model Calibration for Different Climatic Regions, Water Resources Research, 60, e2023WR036 364, https://doi.org/https://doi.org/10.1029/2023WR036364, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023WR036364, 2024.

Price, K., Purucker, S. T., Kraemer, S. R., and Babendreier, J. E.: Tradeoffs among watershed model calibration targets for parameter estimation, Water Resources Research, 48, https://doi.org/10.1029/2012WR012005, publisher: John Wiley & Sons, Ltd, 2012.

Pushpalatha, R., Perrin, C., Le Moine, N., Mathevet, T., and Andréassian, V.: A downward structural sensitivity analysis of hydrological models to improve low-flow simulation, Journal of Hydrology, 411, 66–76, https://doi.org/10.1016/j.jhydrol.2011.09.034, 2011.

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.

Rykiel, E. J.: Testing ecological models: the meaning of validation, Ecological Modelling, 90, 229–244, https://doi.org/10.1016/0304-3800(95)00152-2, 1996.

Santos, L., Thirel, G., and Perrin, C.: Technical note: Pitfalls in using log-transformed flows within the KGE criterion, Hydrology and Earth System Sciences, 22, 4583–4591, https://doi.org/10.5194/hess-22-4583-2018, 2018.

Saurette, D. D., Berg, A. A., Laamrani, A., Heck, R. J., Gillespie, A. W., Voroney, P., and Biswas, A.: Effects of sample size and covariate resolution on field-scale predictive digital mapping of soil carbon, Geoderma, 425, 116 054, https://doi.org/10.1016/j.geoderma.2022.116054, 2022.

Saurette, D. D., Heck, R. J., Gillespie, A. W., Berg, A. A., and Biswas, A.: Divergence metrics for determining optimal training sample size in digital soil mapping, Geoderma, 436, 116 553, https://doi.org/10.1016/j.geoderma.2023.116553, 2023.

Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrological Processes, 21, 2075–2080, https://doi.org/10.1002/hyp.6825, publisher: John Wiley & Sons, Ltd, 2007.

Schefzik, R., Flesch, J., and Goncalves, A.: Fast identification of differential distributions in single-cell RNA-sequencing data with waddR, Bioinformatics, 37, 3204–3211, https://doi.org/10.1093/bioinformatics/btab226, _eprint: https://academic.oup.com/bioinformatics/article-pdf/37/19/3204/50338109/btab226.pdf, 2021.

Seeger, S. and Weiler, M.: Reevaluation of transit time distributions, mean transit times and their relation to catchment topography, Hydrology and Earth System Sciences, 18, 4751–4771, https://doi.org/10.5194/hess-18-4751-2014, 2014.

Shen, H., Tolson, B. A., and Mai, J.: Time to Update the Split-Sample Approach in Hydrological Model Calibration, Water Resources Research, 58, e2021WR031 523, https://doi.org/10.1029/2021WR031523, publisher: John Wiley & Sons, Ltd, 2022.

Sorooshian, S. and Dracup, J. A.: Stochastic parameter estimation procedures for hydrologie rainfall-runoff models: Correlated and heteroscedastic error cases, Water Resources Research, 16, 430–442, https://doi.org/10.1029/WR016i002p00430, publisher: John Wiley & Sons, Ltd, 1980.

Squicciarini, A., Trigano, T., and Luengo, D.: Jensen–Tsallis divergence for supervised classification under data imbalance, Machine Learning, 114, 162, https://doi.org/10.1007/s10994-025-06791-4, 2025.

Thielen, J., Bartholmes, J., Ramos, M.-H., and de Roo, A.: The European Flood Alert System – Part 1: Concept and development, Hydrology and Earth System Sciences, 13, 125–140, https://doi.org/10.5194/hess-13-125-2009, 2009.

https://doi.org/10.5194/egusphere-2026-43
Preprint. Discussion started: 13 February 2026

Thiemig, V., Gomes, G. N., Skøien, J. O., Ziese, M., Rauthe-Schöch, A., Rustemeier, E., Rehfeldt, K., Walawender, J. P., Kolbe, C., Pichon, D., Schweim, C., and Salamon, P.: EMO-5: a high-resolution multi-variable gridded meteorological dataset for Europe, Earth System Science Data, 14, 3249–3272, https://doi.org/10.5194/essd-14-3249-2022, 2022.

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.

Thébault, C., Perrin, C., Andréassian, V., Thirel, G., Legrand, S., and Delaigue, O.: Impact of suspicious streamflow data on the efficiency and parameter estimates of rainfall–runoff models, Hydrological Sciences Journal, 68, 1627–1647, https://doi.org/10.1080/02626667.2023.2234893, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02626667.2023.2234893, 2023.

UKCEH: UK Hydrological Outlook - River Flows, https://hydoutuk.net/about/methods/river-flows, accessed November 28, 2025, 2025.

Verbruggen, E., Sheldrake, M., Bainard, L. D., Chen, B., Ceulemans, T., De Gruyter, J., and Van Geel, M.: Mycorrhizal fungi show regular community compositions in natural ecosystems, The ISME Journal, 12, 380–385, https://doi.org/10.1038/ismej.2017.169, 2018.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., and Soubeyroux, J.-M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, International Journal of Climatology, 30, 1627–1644, https://doi.org/10.1002/joc.2003, 2010.

Vis, M., Knight, R., Pool, S., Wolfe, W., and Seibert, J.: Model Calibration Criteria for Estimating Ecological Flow Characteristics, Water, 7, 2358–2381, https://doi.org/10.3390/w7052358, 2015.

Wada, Y., Bierkens, M. F. P., de Roo, A., Dirmeyer, P. A., Famiglietti, J. S., Hanasaki, N., Konar, M., Liu, J., Müller Schmied, H., Oki, T., Pokhrel, Y., Sivapalan, M., Troy, T. J., van Dijk, A. I. J. M., van Emmerik, T., Van Huijgevoort, M. H. J., Van Lanen, H. A. J., Vörösmarty, C. J., Wanders, N., and Wheater, H.: Human–water interface in hydrological modelling: current status and future directions, Hydrology and Earth System Sciences, 21, 4169–4193, https://doi.org/10.5194/hess-21-4169-2017, 2017.

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, 2010.

Westerberg, I. K., Guerrero, J.-L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E., and Xu, C.-Y.: Calibration of hydrological models using flow-duration curves, Hydrology and Earth System Sciences, 15, 2205–2227, https://doi.org/10.5194/hess-15-2205-2011, 2011.

Winderlich, K., Dalelane, C., and Walter, A.: Classification of synoptic circulation patterns with a two-stage clustering algorithm using the modified structural similarity index metric (SSIM), Earth System Dynamics, 15, 607–633, https://doi.org/10.5194/esd-15-607-2024, 2024.

Yan, J., Li, P., Gao, R., Li, Y., and Chen, L.: Identifying Critical States of Complex Diseases by Single-Sample Jensen-Shannon Divergence, Frontiers in Oncology, Volume 11 - 2021, https://doi.org/10.3389/fonc.2021.684781, 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, publisher: John Wiley & Sons, Ltd, 2008.