https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

Metrics that Matter: Objective Functions and Their Impact on Signature Representation in Conceptual Hydrological Models

Peter Wagener[1,2], Wouter Knoben[2], Niels Schütze[1], and Diana Spieler[1,2]

[1]Institute of Hydrology and Meteorology, TUD Dresden University of Technology, Dresden, Germany
[2]Schulich School of Engineering, Department of Civil Engineering, University of Calgary, Canada

Correspondence: Peter Wagener (peter.wagener@ucalgary.ca)

Abstract. Although objective functions (OFs) are widely discussed in the literature, many modelling studies still default to a few common metrics, without much consideration of their relative strengths and weaknesses. This paper systematically investigates the impact of OF choice on the representation of various streamflow characteristics across 47 conceptual models and 10 hydro-climatically diverse catchments selected from the CARAVAN dataset. We use eight different OFs for calibration, including the Kling–Gupta efficiency (KGE), Nash–Sutcliffe efficiency (NSE), and their respective logarithmic variants, as well as four more recently proposed metrics. We evaluate the representation of 15 hydrological signatures that capture a relevant selection of streamflow characteristics to determine generalizable strengths and weaknesses of individual OFs across different models and catchments. Results show that the choice of OF can significantly affect a model's capability to simulate different hydrological signatures such as runoff ratios, extreme flow percentiles, and certain baseflow characteristics. While certain signatures, particularly those related to flow variability, are relatively insensitive to OF choice, others exhibit large performance shifts across different OFs. Generally, no single OF simultaneously achieved high performance across all tested signatures, highlighting that a single-objective calibration is unlikely to lead to an all-purpose model. Our results reinforce calls to choose objective functions deliberately and in line with the objectives of a study. They also provide initial guidance on which metrics highlight particular facets of streamflow behaviour.

1 Introduction

Setting up and running a hydrological model requires modellers to make many decisions. These typically include the choice of a suitable model (structure), sensible parameter boundaries, an appropriate calibration period, and which method to use for forcing data interpolation. Selecting which performance metric(s) to employ as the objective function during model calibration is another critical decision. Different choices of performance metrics can lead to substantially different model outcomes. For example, Mai (2023) tested six alternative metrics for calibration and found marked differences in the resulting hydrographs. Similarly, Garcia et al. (2017) compared variants of the KGE metric for low-flow simulation, observing large differences in annual runoff estimates. Studies such as Pool et al. (2017) and Vis et al. (2015) further demonstrated that the choice of performance metric for calibration influences how well the model reproduces key hydrological signatures (i.e., streamflow characteristics). Similarly, Mendoza et al. (2016) and Seiller et al. (2017) show the impact calibration choices can have on the

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

portrayal of climate change impacts.

Among the various decisions involved in model calibration, evaluation, and diagnostic analysis, the choice of objective function is often critical. It directly determines how model parameters are optimized and, consequently, how well the resulting simulations reproduce both observed data and underlying hydrological processes. Despite the well-established influence of objective function selection, and an extensive body of research discussing the individual strengths and weaknesses of various performance metrics (e.g. Thirel et al., 2024; Althoff and Rodrigues, 2021; Cinkus et al., 2023; Knoben et al., 2019b; Mizukami et al., 2019; Bennett et al., 2013; Gupta et al., 2009; Krause et al., 2005), Jackson et al. (2019) conclude that: "in most hydrologic modelling studies error metrics are chosen based on familiarity, [and] without consideration of the relative strengths and weaknesses" in their review of more than 60 different performance metrics. And indeed, a quick and informal review of 60 modelling studies in the existing literature library of the authors, supports this claim by showing that most studies provide little to no reasoning for their choice of objective function (Figure 1; details on review in Supplement S1). The papers that gave reasoning mainly referred to general characteristics and perceptions of their metric of choice. Reasons we documented (see Supplement S1) can roughly be classified into various forms of "the KGE is a more balanced metric than the NSE", "NSE is a common metric for high flows", "logNSE/logKGE focus on low flows", "NSE/KGE are widely used metrics" or "we use this metric because it is comparable with previous studies".

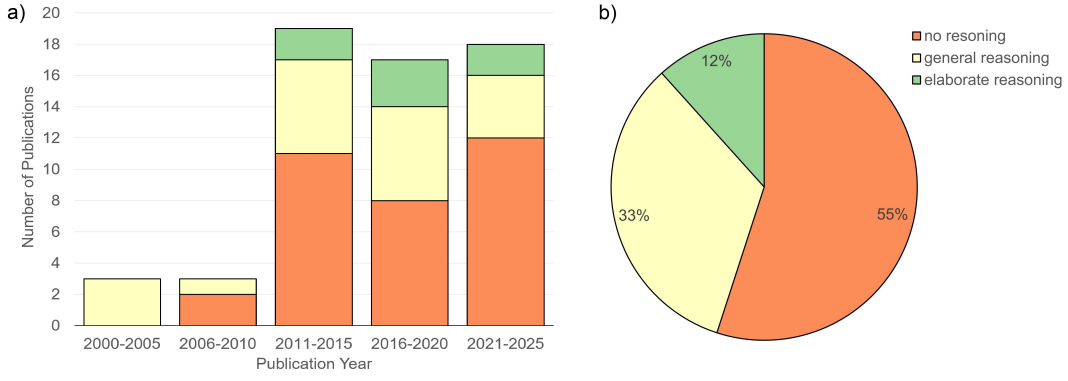

Figure 1. a) Number of analyzed studies published during different five-year periods b) Percentage of these studies that give no - general - or elaborate reasoning on their choice of objective function. This data stems from an informal/non-exhaustive literature review of 60 studies taken from the authors literature database. Details regarding the review are given in Supplement S1

We believe that part of the prevailing tendency to define objective functions in an often ad-hoc manner is the broad but scattered and often site- and/or model-specific research on the impacts of different performance metrics used as objective functions. Most good modelling practice guidelines (e.g. Waveren et al., 1999; Jakeman et al., 2006, 2018) simply advise to select an appropriate objective function that aligns with the purpose of the modelling study, but do not provide guidance on how

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

to do so. The question of how to link the choice for a suitable calibration metric to a specific modelling purpose thus remains unclear (Jacobs et al., 2024). While several authors have investigated connections between objective function choice and how the calibrated models represent certain streamflow characteristics, they are often limited in the number of models they use (1 model in e.g. Hallouin et al., 2020; Melsen et al., 2019; Pool et al., 2017; Vis et al., 2015) or signatures they consider (less than 5 in e.g. Araya et al., 2023; Mizukami et al., 2019; Seiller et al., 2017). As Table 1 highlights, studies with a similar focus to our study also mostly consider catchments of only one specific region or country and no study the authors are aware of tested more than 12 models while also considering several metrics and signatures. Consequently, few findings generalize beyond the familiar dictum that squared metrics emphasize the larger errors commonly associated with higher flows, and transformations help to emphasize low flows. The available site or model specific knowledge make a robust transfer to new studies uncertain and generalizable trade-offs of specific metric choices seem underdiscussed.

Publication	Main Focus	Catchments	Location	Models	OFs	Signatures
Althoff and Rodrigues (2021)	OF impact on streamflow simulations	179	Brazil	1	11*	7
Araya et al. (2023)	OF impact on streamflow forecast	22	Chile	3	12*	5
Hallouin et al. (2020)	OF impact on signatures	33	Ireland	1	6*	18
Hernandez-Suarez et al. (2018)	OF impact on signatures	1	ACF-Basin, USA	1	8*	167
Melsen et al. (2019)	calibration impact on flood/drought	9	Thur Basin, Switzerland	1	2	6
Mendoza et al. (2016)	calibration impact on climate projections	3	Colorado Basin, USA	4	4	6
Mizukami et al. (2019)	OF impact on high flows	492	USA	2	5	2
Pool et al. (2017)	OF impact on signatures	25	Tennessee, USA	1	29*	13
Seiller et al. (2017)	OF impact on climate projections	37	Quebec, Canada	12	3	4
Vis et al. (2015)	OF impact on signatures	27	Tennesse, USA	1	9*	12
This study	OF impact on signatures	10	worldwide	47	8	15

Table 1. Summary of similar studies that investigate the impact of objective function choice on the representation of hydrological signatures. We compare the number of investigated models, catchments, metrics and signatures to this study. Note that studies with an *asterix also tested combinations of different metrics (i.e. composite criteria or multi-metric objective functions) in addition to individual calibration metrics.

Building on this motivation, this study aims to develop a more generalized understanding of how the choice of performance metrics used for model calibration influences the representation of the hydrological regime. By systematically analyzing catchments spanning a broad range of climatic conditions and employing a large set of conceptual model structures, we seek to identify general benefits and limitations associated with different calibration metrics across diverse hydrological contexts. The selected set of catchments, though limited in number, was chosen based on differing climate indices such as aridity, seasonality, and fraction of snow, ensuring a representative spread of hydroclimatic conditions (cf. the 12 MOPEX basins selected in Duan et al. (2006); van Werkhoven et al. (2008); Carrillo et al. (2011)). This design enables us to examine how models calibrated with different performance metrics behave under varying hydrological regimes while also balancing depth with breadth of analysis (Gupta et al., 2014). In this study, we evaluate how 47 conceptual model structures, each calibrated using alternative

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

performance metrics, reproduce 15 selected hydrological signatures that describe distinct aspects of the flow regime, such as flow variability, baseflow contribution, and seasonal timing. This multi-dimensional setup — combining diverse catchments, model structures, and metrics — contributes to identifying generalizable strengths and weaknesses of individual performance metrics that extend beyond their known sensitivities to particular flow conditions, offering new insights into how calibration choices influence process realism in hydrological modelling. We will uncover potential drawbacks of specific metric choices and provide a more comprehensive guideline for modellers to decide which objective function may be most appropriate for their specific modelling goal. The following sections will introduce the catchments, models and signatures we used (Section 2), present and discuss the results (Section 3) and provide an overarching discussion of the implications metric choice can have on the representation of the hydrological regime (Section 4).

2 Methodology

Figure 2 gives an overview of the methodology used in this study. We calibrate 47 conceptual hydrological models from the MARRMoT Toolbox (Knoben et al., 2019a; Trotter et al., 2022) using eight different calibration metrics. The models and metrics used are introduced in Section 2.1.1 and 2.1.2 respectively. We calibrate all of these models (as described in Section 2.2) for a subset of 10 hydro-climatically differing catchments from the CARAVAN dataset (Kratzert et al., 2023) as introduced in Section 2.1.3. After selecting only the well-performing model structures according to a benchmark procedure (Section 2.3) we use the simulated discharge timeseries to calculate 15 selected signatures that represent varying aspects of the hydrological regime (as introduced in Section 2.4). Through a comparison of the simulated and observed signature values we are able to identify the strengths and weaknesses of the tested metrics in representing the hydrological regime over a broad range of hydro-climatic conditions and different model complexities.

2.1 Experiment Components

2.1.1 Models

We obtained the 47 models used in this study from the MARRMoT toolbox (Knoben et al., 2019a; Trotter and Knoben, 2022), partly because the toolbox allows easy and consistent use of the different models. These models are based on the scientific literature and mimic some widely used and well-known models such as HBV, GR4J, TOPMODEL, VIC and HYMOD. A full list of all models included in this analysis can be found in the Supplementary Material S2.1. Figure 2b shows how the model structures differ in their number of considered stores (i.e., state variables) and parameters and therefore gives an idea of the different models' complexity. The number of parameters ranges from one to 24 and the number of considered stores ranges from one to eight. The simplest model has one store and one parameter, whereas the most complex models have eight stores and 12 parameters, or three stores and 24 parameters. By using such a wide range of different complexities in model structures, we ensure that the strengths and weaknesses identified generalize across different model structures.

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

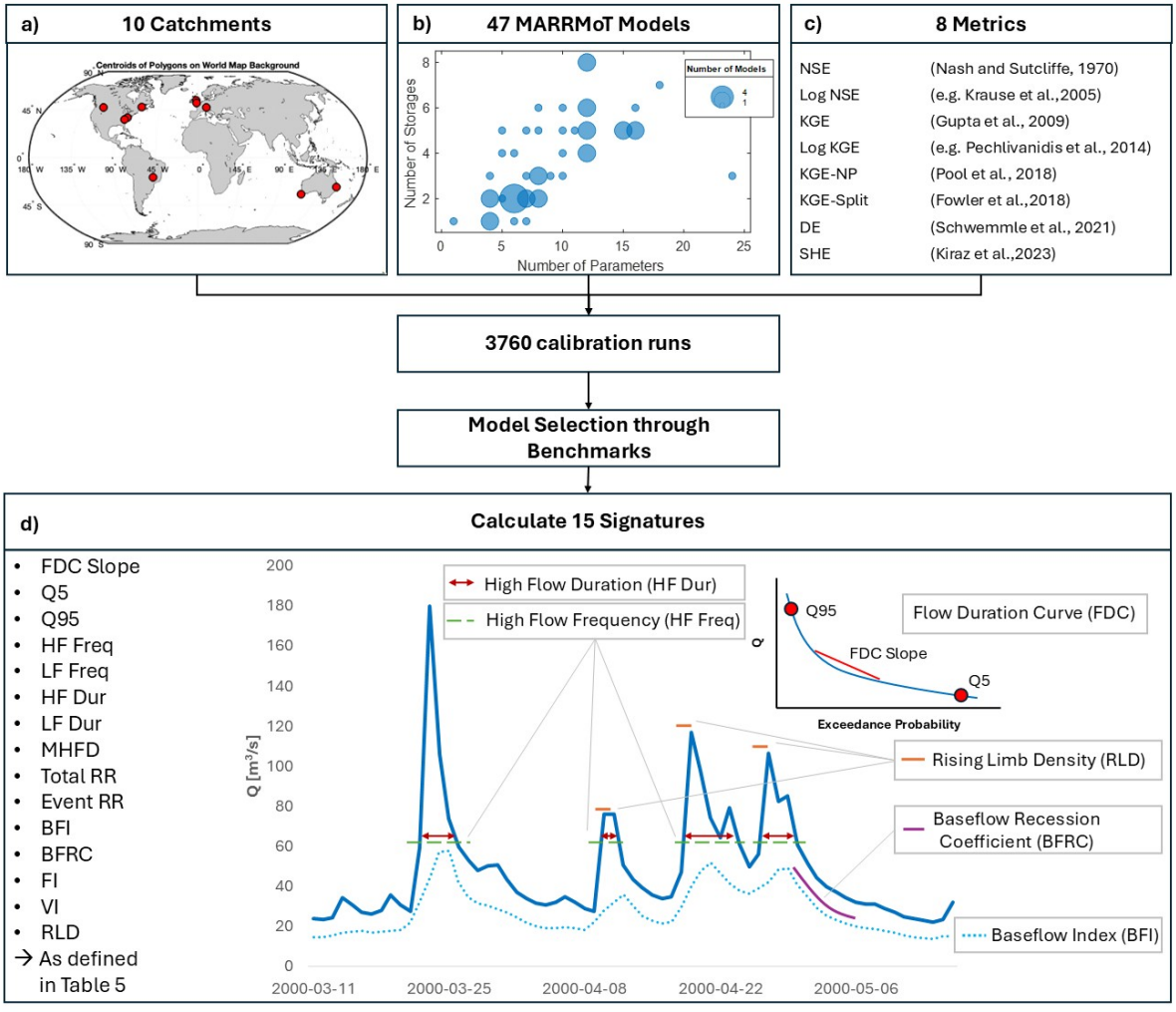

Figure 2. Overview of the experiment design of this study. a) Shows the geographical location of the study catchments. b) Visualizes the different complexities (number of storages vs number of parameters) of the investigated MARRMoT Models. c) Lists the eight objective functions used for calibration. d) Lists the 15 hydrological signatures investigated and shows a schematic visualization of some of them.

2.1.2 Objective Functions

We calibrate the 47 hydrological models with eight objective functions each using different performance metrics, and analyze how well the calibrated models simulate 15 streamflow signatures. We differentiate between model accuracy (objective function value) and model adequacy (signature representation) and aim to investigate their relationship. We selected four widely known and frequently used metrics and four metrics that have only recently been introduced in attempts to improve on known weaknesses of the more commonly used metrics. In general, "performance metrics" can be seen as any metric that quantifies

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

model performance, while "objective function" indicates a more specific use of any performance metric as a target during model calibration. In the following, we use these terms interchangeably. In this study, the eight different objective functions we use are defined through the eight individual performance metrics introduced below. To describe them, we introduce the following set of common symbols: Q and P denote streamflow and precipitation respectively. σ and r are standard deviation and correlation. Subscripts are used to indicate the specifics of all mentioned variables, e.g. Q_s and Q_o show simulated and observed streamflow respectively. Subscript t refers to individual time steps. For correlation, the subscripts r_{pe} and r_{sp} are used for Pearson and Spearman correlation. Overlines are used to indicate temporal means. In sums, the letter n is used to indicate the total length of the time series.

Table 2. Equations of the eight performance metrics used as objective functions in this study. Numbers (1–8) correspond to metric references in the text.

No.	Metric	Equation
(1)	Nash–Sutcliffe Efficiency (NSE)	$NSE = 1 - \frac{\sum_{t=1}^{n}(Q_{o,t} - Q_{s,t})^2}{\sum_{t=1}^{n}(Q_{o,t} - \overline{Q_o})^2}$
(2)	Kling–Gupta Efficiency (KGE)	$KGE = 1 - \sqrt{\left(\frac{\overline{Q_s}}{\overline{Q_o}} - 1\right)^2 + \left(\frac{\sigma_s}{\sigma_o} - 1\right)^2 + (r_{pe} - 1)^2}$
(3)	log NSE	$\log \text{NSE} = 1 - \frac{\sum_{t=1}^{n}(\log(Q_{o,t}) - \log(Q_{s,t}))^2}{\sum_{t=1}^{n}(\log(Q_{o,t}) - \overline{\log(Q_o)})^2}$
(4)	log KGE	$\log \text{KGE} = 1 - \sqrt{\left(\frac{\overline{\log(Q_s)}}{\overline{\log(Q_o)}} - 1\right)^2 + \left(\frac{\sigma_{\log,s}}{\sigma_{\log,o}} - 1\right)^2 + (r_{p,\log} - 1)^2}$
(5)	Non-parametric KGE (KGE–NP)	$KGE_{NP} = 1 - \sqrt{\left(\frac{\overline{Q_s}}{\overline{Q_o}} - 1\right)^2 + \left(1 - \frac{1}{2}\left\lvert\frac{Q_s(I(k))}{n\overline{Q}_s} - \frac{Q_o(J(k))}{n\overline{Q}_o}\right\rvert\right)^2 + (r_{sp} - 1)^2}$
(6)	KGE Split	$\text{KGE Split} = \frac{1}{Y}\sum_{y=1}^{Y}\left(1 - \sqrt{\left(\frac{\overline{Q_{s,y}}}{\overline{Q_{o,y}}} - 1\right)^2 + \left(\frac{\sigma_{s,y}}{\sigma_{o,y}} - 1\right)^2 + (r_{pe,y} - 1)^2}\right)$
(7)	Signature-based Hydrologic Efficiency (SHE)	$SHE_g = 1 - \sqrt{\left(\frac{\overline{Q_s}/\overline{P_o}}{\overline{Q_o}/\overline{P_o}} - 1\right)^2 + \left(\frac{\sigma_s/\sigma_p}{\sigma_o/\sigma_p} - 1\right)^2 + (r_{sp} - 1)^2}$
(8)	Diagnostic Efficiency (DE)	$DE = \sqrt{\left(\frac{1}{n}\sum_{i=0}^{n}\frac{Q_s(i) - Q_o(i)}{Q_o(i)}\right)^2 + \left(\frac{1}{n}\int_0^1\left\lvert\frac{Q_s(i) - Q_o(i)}{Q_o(i)} - \overline{\frac{Q_s(i) - Q_o(i)}{Q_o(i)}}\right\rvert di\right)^2 + (r_{pe} - 1)^2}$

The first two metrics are likely the most common objective functions used in hydrology. That is, (1) the Nash-Sutcliffe-Efficiency (NSE; Nash and Sutcliffe, 1970) and (2) the Kling-Gupta-Efficiency (KGE; Gupta et al., 2009) as defined in Table 2. The next two metrics are just as well known, and transformations of the first two metrics that aim to improve the representation of low flows, the logarithmic versions of NSE (3) and KGE (4). In addition, we include four recently developed metrics that intend to improve certain aspects of NSE or KGE or diagnose hydrological model behaviour.

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

Firstly, this is the non-parametric version of the KGE (KGE-NP; Pool et al., 2018). The major difference to the standard KGE is that the variability component is no longer based on the ratio of variances, but instead replaced by a flow-duration curve-based term. Additionally, the correlation term uses the Spearman rank correlation instead of the pearson correlation. Pool et al. (2018) argue that the resulting benefit of their proposed metric is an overall better agreement between observations and simulations, except for high flows. Their reasoning is that more information is contained within the metric because the FDC is more complex than the standard deviation and the Spearman rank correlation leads to improvements for low flows. In the equation of (5) $I(k)$ and $J(k)$ indicate time steps in which the k-th largest flow occurs in the simulated and observed time series respectively.

Secondly, another adaption of the KGE, the KGE Split (KGE Split; Fowler et al., 2018) is evaluated. It is calculated like the regular KGE, but instead of calculating one value for the entire time series, a value for each year is calculated, and the mean over these values becomes the actual objective function value. Fowler et al. (2018) developed it to put more emphasis on dry years, since the typically used "least squares" methods give more attention to high flows rather than low flows. The variables are identical to those used in the KGE, with y indicating the evaluated year and Y the total number of years as shown in equation (6).

Thirdly, the signature-based hydrologic efficiency (SHE; Kiraz et al., 2023) normalizes both the bias and the variance term by the mean and variance of the precipitation. Like the KGE-NP, it also uses the Spearman description for the correlation term to improve on the low flow representations. The SHE as used in Kiraz et al. (2023) is defined in equation (7).

Fourthly and lastly, we analyze the Diagnostic Efficiency (DE; Schwemmle et al., 2021) where the bias term is replaced with a mean relative error and the variance term is the integral over all residuals of the relative error based on the flow duration curve. As the name suggests, this metric is designed as a diagnostic tool but will be applied here as an objective function. The general structure, with three parts for bias, variance, and timing, is similar to the KGE, but particularly the variance term proposes an interesting alternative to the KGE. To convert it into a similar objective function as the other candidates we use: $DE_{OF} = 1 - DE$. In equation (8), i denotes the exceedance probability of observed and simulated streamflow.

All of the eight used metrics can be disaggregated into three components representing bias, variability and correlation. For an easier overview of the differences between the individual metrics we summarized how they represent each component in Table 3.

2.1.3 Study Catchments

This study uses catchments from the CARAVAN database (Kratzert et al., 2023) which combines several CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) datasets of different countries or regions in a standardized way. While the latest update to CARAVAN now includes more than 20000 different catchments (Färber et al., 2025) from over seven existing large-sample hydrology datasets, at the beginning of this study CARAVAN included 2901 catchments. At the time, CARAVAN included catchments in the United States (Addor et al., 2017), Great Britain (Coxon et al., 2020), Brazil (Chagas et al., 2020), Chile (Alvarez-Garreton et al., 2018) and Australia (Fowler et al., 2021) from their respective CAMELS datasets, as well as the LamaH-CE (Central Europe, Klingler et al., 2021) and HYSETS (North America, Arsenault et al., 2020) datasets.

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

Table 3. Overview of the three components of the eight metrics

Metric	Bias	Variability	Correlation
KGE	$\frac{\overline{Q_s}}{\overline{Q_o}}$	$\frac{\sigma_s}{\sigma_o}$	Pearson
NSE	$\frac{\overline{Q_s}-\overline{Q_o}}{\sigma_o}$	$\frac{\sigma_s}{\sigma_o}$	$2 \cdot \text{Pearson} \cdot \frac{\sigma_s}{\sigma_o}$
log KGE	as KGE	as KGE	as KGE
log NSE	as NSE	as NSE	as NSE
KGE-NP	as KGE	FDC-based	Spearman
KGE-Split	as KGE	as KGE	as KGE
SHE	as KGE	as KGE	Spearman
DE	Bias FDC-based	Residuals FDC-based	Spearman

CARAVAN provides standardized meteorological forcing, streamflow data, and static catchment attributes (e.g., geophysical, sociological, climatological) with a median data length of 31 years. Our primary goal is to investigate the sensitivity of a large number of different model structures to objective function choice. To keep the analysis manageable (i.e., to balance depth with breadth (Gupta et al., 2014)), we defined a subset of ten hydro-climatically differing catchments for our analysis. To determine these catchments, we used the climate classification procedure outlined in (Knoben et al., 2018) to quantify the aridity, seasonality and fraction snow in each of the CARAVAN basins. We then used a k-means clustering algorithm to divide the catchments into 10 clusters, and selected the catchment closest to each cluster centroid for use in this work. The location of the selected catchments is shown in Figure 2a and a brief description of some catchment properties is given in Table 4. The full clustering procedure is described in more detail in the Supplementary Material S2.2.

Table 4. Overview of catchments properties for the selected 10 CARAVAN basins

Cluster	Abbrev.	Dataset	Area	Aridity	Seasonality	Snow Frac	Precip	Runoff	Mean Temp
–	–	–	(km^2)	(–)	(–)	(–)	(mm/yr)	(mm/yr)	(°C)
1	AUS1	CAMELS-AUS	125.74	-0.39	0.59	0.00	836.5	117.3	18.27
2	BR6	CAMELS-BR	194.00	-0.28	1.30	0.00	1451.2	490.0	22.45
3	AUS6	CAMELS-AUS	124.94	-0.04	1.67	0.00	680.4	165.8	15.43
4	C12	CAMELS	148.69	0.02	1.66	0.49	773.5	301.5	2.53
5	C02	CAMELS	285.39	0.03	1.07	0.07	1120.1	361.8	10.91
6	GB3	CAMELS-GB	136.53	0.16	1.46	0.00	792.6	185.7	9.74
7	C03	CAMELS	127.83	0.33	0.85	0.08	1411.5	656.6	10.59
8	HYS	HYSETS	328.44	0.34	1.28	0.38	1211.3	660.8	3.59
9	GB2	CAMELS-GB	283.60	0.42	1.25	0.00	1209.2	664.9	8.00
10	LAM	LamaH-CE	102.29	0.58	0.58	0.32	1777.0	1299.5	0.91

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

The 10 selected catchments span a wide range of climatic and hydrological settings. They differ substantially in aridity and mean temperature, ranging from humid, cool basins such as LAM and HYS (aridity ≈ 0.3–0.6, temperatures near 0–4 °C) to warm, dry catchments like AUS1 and BR6 (negative aridity indices, > 18 °C). Snow fraction varies markedly—from snow-free catchments in Australia and Brazil to snow-influenced basins such as C12 and LAM. Precipitation and runoff also contrast strongly, with wetter, high-runoff regions (e.g. LAM, GB2) versus more arid, low-runoff sites (AUS1, AUS6), illustrating the climatic and geographic diversity captured by the dataset.

2.2 Model Calibration Procedure

All models were calibrated with the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) algorithm by Hansen et al. (2003) as implemented in the MARRMoT Toolbox. We used a 10-year time period from 2004 to 2014 for calibration and a 10-year time period from 1993 to 2003 for evaluation with a one-year warm-up period each. The evaluation period of the catchment AUS6 had to be shortened to eight years, however, because of large data gaps. We conducted 3760 (47x8x10) calibration runs to calibrate each of the 47 models on all eight metrics in each of the ten catchments. For all conducted analyses, we decided to analyze the signature representation during the calibration period only, because the impact of the objective function is more direct. Because the signature representation is generally consistent between calibration and validation periods (see Figure S3 in Supplementary Material) we believe this to be a feasible approach.

2.3 Model Benchmarking

As we do not want to contaminate the analysis through poorly performing models, we establish a baseline performance. Based on earlier arguments and examples (Schaefli and Gupta, 2007; Knoben et al., 2020; Knoben, 2024), we use the daily mean flow as a benchmark, because it accounts for seasonality and provides a more informative baseline than the average flow. For each catchment, we compute this benchmark from observed discharge data in the calibration period and evaluate it with eight performance metrics. Models that fail to outperform the benchmark are excluded from further analysis, ensuring that only sufficiently reliable models are carried forward to the analysis of signature performance.

2.4 Signature Selection

We selected 15 streamflow signatures to investigate the extent to which objective function choice influences a model's ability to replicate streamflow signatures. Our goal with this selection is to cover different aspects of the flow regime, as well as a range of hydrological processes (McMillan, 2020). The selected signatures are shown in Table 5 and were all calculated using the TOSSH toolbox (Gnann et al., 2021). Most of the considered signatures describe streamflow properties such as the slope of the flow duration curve (FDC Slope), the 5th and 95th streamflow percentile (Q5 and Q95), the high and low flow duration (HF Dur & LF Dur) and frequency (HF Freq & LF Freq) as well as the mean half flow date (MHFD). The slope of the FDC describes the variability of the streamflow regime, indicating how quickly a river responds to rainfall events and how sustained the flows are during dry periods. Representing it correctly therefore indicates a good representation of in-catchment storage and

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

flashiness behaviour. The 5^{th} and 95^{th} streamflow percentile describe the magnitude of low and high flows. Representing these signatures well indicates a good representation of extreme flow behaviour, while representing the high and low flow duration and frequency well ensures that also the important characteristics of extreme flow can be met. The MHFD helps to determine if the general timing and seasonality of streamflow is represented well. Additionally, we evaluate the impact of the objective function on the total runoff ratio (Total RR), event runoff ratio (Event RR), baseflow index (BFI), baseflow recession coefficient (BFRC), flashiness index (FI), variability index (VI) and rising limb density (RLD). These are intended to deliver information on the different catchment processes such as the general water balance (Total RR), individual storm responses (Event RR), the baseflow processes (BFI, BFRC) or general water storage behaviour (flashiness index, variability index, rising limb density).

Table 5. Overview of selected signatures.

Signature	Abbreviation	Category	Units
Slope of Flow Duration Curve	FDC Slope	Streamflow	-
5^{th} Streamflow Percentile	Q5	Streamflow	mm/day
95^{th} Streamflow Percentile	Q95	Streamflow	mm/day
High Flow Frequency	HF Freq	Streamflow	-
Low Flow Frequency	LF Freq	Streamflow	-
High Flow Duration	HF Dur	Streamflow	days/event
Low Flow Duration	LF Dur	Streamflow	days/event
Mean Half Flow Date	MHFD	Streamflow	day of year
Total Runoff Ratio	Total RR	Water Balance	-
Event Runoff Ratio	Event RR	Partitioning/Connectivity	-
Baseflow Index	BFI	Baseflow	-
Baseflow Recession Coefficient	BFRC	Baseflow	1/day
Flashiness Index	FI	Water Storage	-
Variability Index	VI	Water Storage	-
Rising Limb Density	RLD	Channel Processes	rises/day

2.5 Analysis of Objective Function Influence

2.5.1 Signature Error Metric

As most signature values have individual ranges and can significantly differ between individual catchments, we analyze normalized signature errors for a better comparison. The metric is calculated as follows (additional details are provided in the Supplementary Material S2.4):

$$EM_{jk} = \frac{med_i\left(S_{ijk} - y_j\right)}{\left|max\left(med_i\left(S_{ijk} - y_j\right)\right)\right|} \tag{1}$$

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

where S_{ijk} is the simulated signature value depending on model i, catchment j, and objective function k. y_j refers to the observed signature value depending on the location, and $med_i(\cdot)$ denotes the median taken over the ensemble index i.

The range for this metric is limited to [-1,1], with positive values indicating overestimation and negative values indicating underestimation of the observed signature value. These values are calculated on a catchment-by-catchment basis so that if one OF is worst in all catchments, all error metric values would indicate ± 1. A value of 0 indicates that the observed value was matched exactly.

This initial error metric can be aggregated further by using the median value over all catchments. This gives an error metric specific to a signature and objective function, which can be compared to find the objective function that best represents each signature.

$$EM_{k,abs} = abs\left(med_j\left(EM_{jk}\right)\right) = abs\left(med_j\left(\frac{med_i\left(S_{ijk} - y_j\right)}{|max\left(med_i\left(S_{ijk} - y_j\right)\right)|}\right)\right) \tag{2}$$

2.5.2 Statistical Testing of Objective Function Influence

To better understand which signatures are most affected by the objective function choice, we conducted a paired sample t-test: a pairwise comparison for two objective functions at a time by testing their median value of signature values across catchments. The null hypothesis for the t-test used here is H_0 : The mean values of the distributions are equal to each other. With 8 objective functions this leads to $\sum_{k=1}^{n} k = \frac{n(n-1)}{2} = \frac{8*7}{2} = 28$ data points for each signature.

This analysis highlights which signatures are most strongly influenced by the choice of objective function. We report the resulting p-values for each pairwise comparison and consider differences statistically significant at $\alpha = 0.05$. Not all distributions are expected to differ significantly, as some objective functions are likely to produce similar results. Based on these outcomes, we restrict subsequent analyses to signatures that are commonly significantly impacted by the choice of objective function, which we define as 30% of the p-values being lower than 0.05.

2.5.3 Random Forest for Attribute Importance

For signatures that show significant sensitivity to the choice of objective function (based on the previous subsection), we apply a random forest (RF) regression to assess the relative importance of catchment, objective function (OF), and model choice. The RF was trained with the vectorized signature error values (taken from 2.5.1) as the response, and the corresponding factor indices for OF (k), model (l), and catchment (j) as predictors.

We used the `TreeBagger` implementation in MATLAB with 300 trees and regression mode, extracting out-of-bag permuted predictor importance scores. For each signature, importance values were normalized to sum to one, providing a direct measure of the relative influence of the three factors on the signature values. The resulting matrix reports the mean relative importance of catchment, OF, and model choice for the representation of the analyzed signatures.

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

3 Results

The following sections will cover the outcomes of the benchmarking procedure (Section 3.1). Based on this, the general distribution of signature representation (Section 3.2) for (i) the detailed example of Total Runoff Ratio and (ii) all signatures (aggregated over the models) are shown. Afterwards, we will use the error metric (Equation 1 from Section 2.5.1) to highlight the skills and shortcomings of each objective function in Section 3.3. Further analysis is then used to test the statistical significance of objective function influence (Section 3.4) and assess the relative predictive importance of objective function choice (Section 3.5). Lastly, we use the aggregated error metric (Equation 2 from Section 2.5.1) to find the best performing objective functions regarding signature representation.

3.1 Benchmarking the Models

Figure 3 shows the performance of all 47 calibrated models (violins) compared to the calculated benchmarks (red dash). The benchmark scores are calculated as any of the model scores, by comparing a timeseries of benchmark "simulations" against the observations in each basin. The chosen benchmark is the seasonal cycle, given by a repeating sequence of daily mean flows (see Section 2.3). The goal of the benchmarking procedure is to retain only plausible models for the remainder of the analysis.

The benchmark scores vary considerably between catchments and also between objective functions, as do the number of models that are able to beat the benchmark (shown as the number above each violin). In some basins, the interannual mean (i.e. the benchmark) is a good predictor of the daily flow, and in these basins the benchmark scores tend to be high. Basins C12 and LAM are catchments where snow plays a large role, and in these catchments this leads to a strongly seasonal flow regime that is relatively stable between years, and thus well captured by the benchmark. In these basins few models beat the benchmark, though this is not entirely unexpected because only eight of the MARRMoT models have a snow module capable of representing these local processes. Therefore there is a large spread in the number of models that can outperform the benchmark. The C12 and LAM basins, as well as HYS in most cases, contrast with the remaining catchments where typically a larger number of models outperform the benchmark. In a handful of cases all models can beat the benchmark (AUS1 and GB2 for the NSE objective function), and for catchment C12 (high snow tendency, high seasonality), there are two objective functions (log NSE, SHE) for which no model was able to outperform the interannual mean benchmark. Models that fail to beat the benchmark might be missing key controls such as snow accumulation and melt, glacier or deep-groundwater storage, so calibration cannot recover the correct timing or magnitude of seasonal peaks, allowing the simple daily-mean predictor to do better. For all following analysis, only those models that outperform the benchmark are used.

3.2 How accurate are the signature representations when different OF are used?

To begin the analysis of how objective functions affect hydrological adequacy, we compare the simulated signature values with the observed signature values for all eight objective functions. Figure 4 shows the distribution of absolute errors (y-axis) over all catchments (x-axis) for each metric (subplots) for one of the 15 signatures, the Total Runoff Ratio. The plots for all other signatures can be found in the Supplementary Material S2.7 (Figure S5 through S18).

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

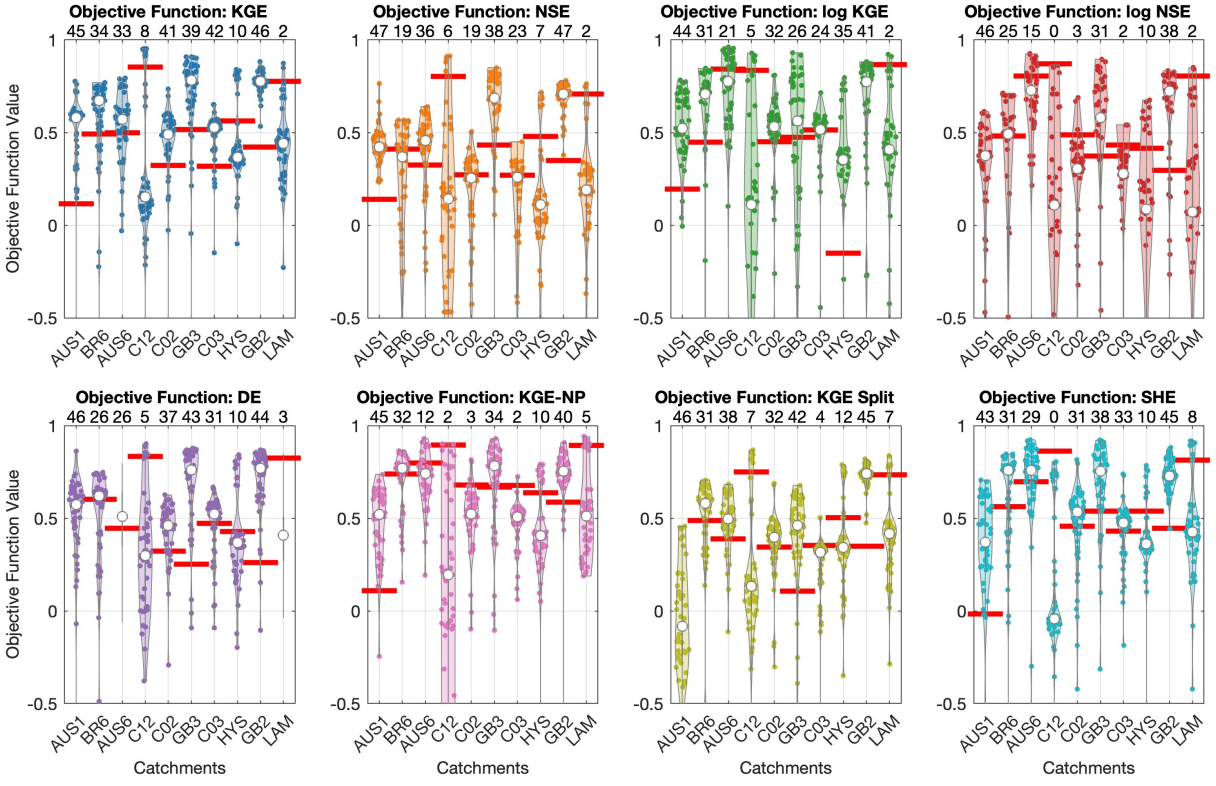

Figure 3. Benchmark Performance (driest catchment on left to wettest catchment on right). The dots represent the 47 individual models while the red line show the benchmark score of the objective function. The numbers above each violin indicate the number of models that outperform the benchmark.

In Figure 4, there are three things of note. First, the location of the violin plots shows whether the signature values are underestimated (negative values), overestimated (positive values) or relatively unbiased (centered around zero). Second, the size of the violins indicates whether the signature predictions are well-constrained (small violin) or not (large violin). Third, the catchments are ordered from driest (left) to wettest (right).

Interestingly, there is no clear pattern in Runoff Ratio errors related to the catchments' aridity values. Instead, there appear to be two categories of results: objective functions that return mostly unbiased models in all catchments (KGE and KGE-NP), and objective functions where the signature representation of the resulting models varies strongly per catchment (e.g. NSE: relatively unbiased in some basins, large variability in signature errors in others). The KGE-NP contains the variability within the models much better and is therefore assessed as the best available objective function for the general water balance as represented through the Total RR signature.

Figure 4 shows that the Runoff Ratio tends to be underestimated rather than overestimated by the given selection of objective functions. This means that generally the models tend to put too much water into storage and/or evaporation or other sink terms, which may become relevant when modelling climate change impacts.

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

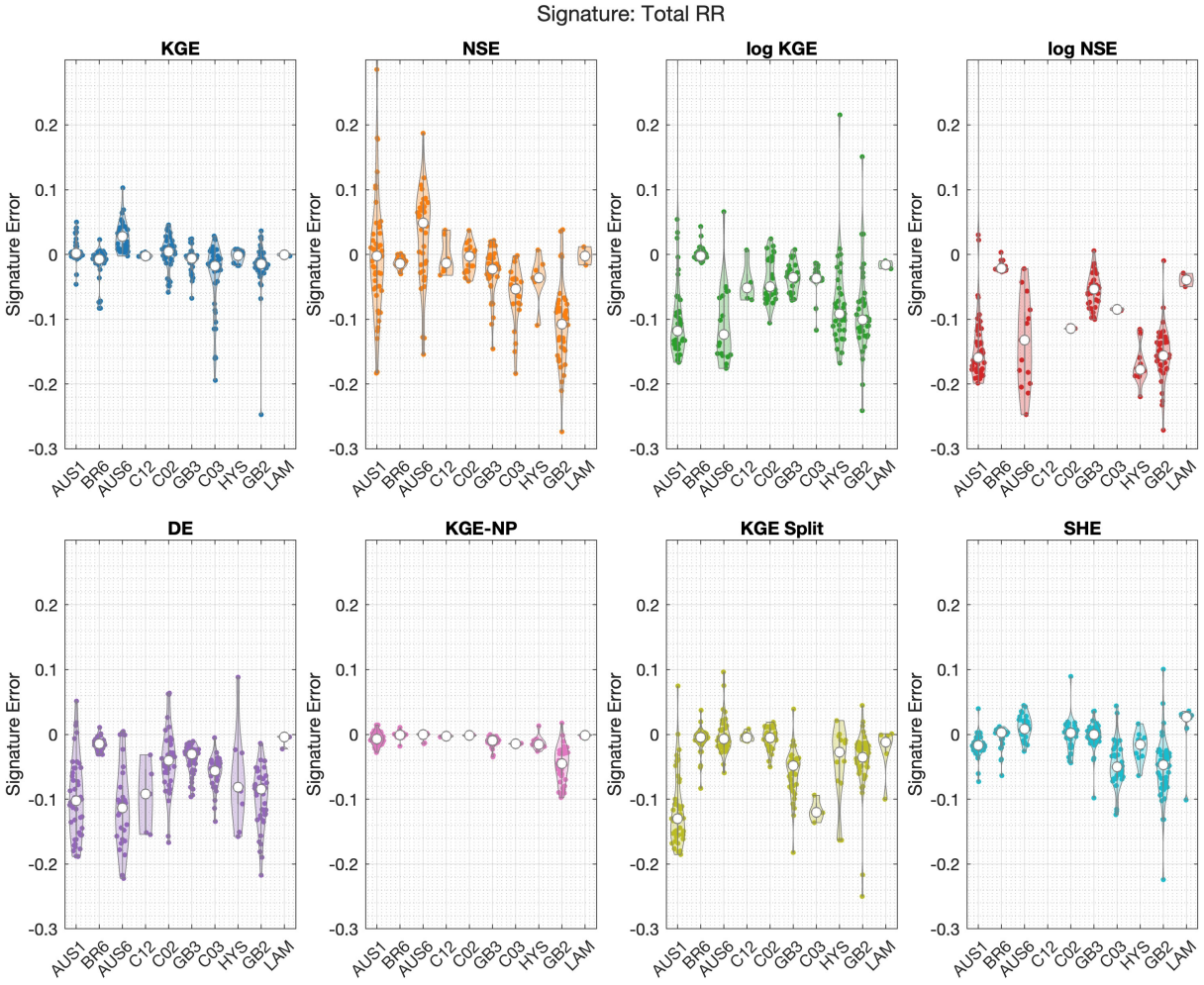

Figure 4. Absolute signature error on Total RR values for all objective functions, calculated as the difference between simulated and observed signature value. For the combination of catchment C12 and objective functions log NSE and SHE, no model outperformed the benchmark.

To show the results for more signatures in a comprehensive way, it is necessary to aggregate the results in some form. This is why we are using the median over all models (exceeding the benchmark in performance) in Figure 5. The median was chosen to achieve the most representable representation of the model spread, which is not susceptible to outliers. This gives an overall impression on the capability of the models to represent the hydrological signatures of interest. It also allows a first impression of the impact different objective functions have on signature representation. In Figure 5, the coloured dots indicate the median modeled signature value for each objective function, while the black line shows the signature values calculated from observations. The grey violins indicate the variability in the modelled signature value if all individual model results are considered (every model that beats the benchmark for any objective function). The plot therefore allows insights on how well a signature value can generally be represented through the different tested models and gives a first impression on the influence

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

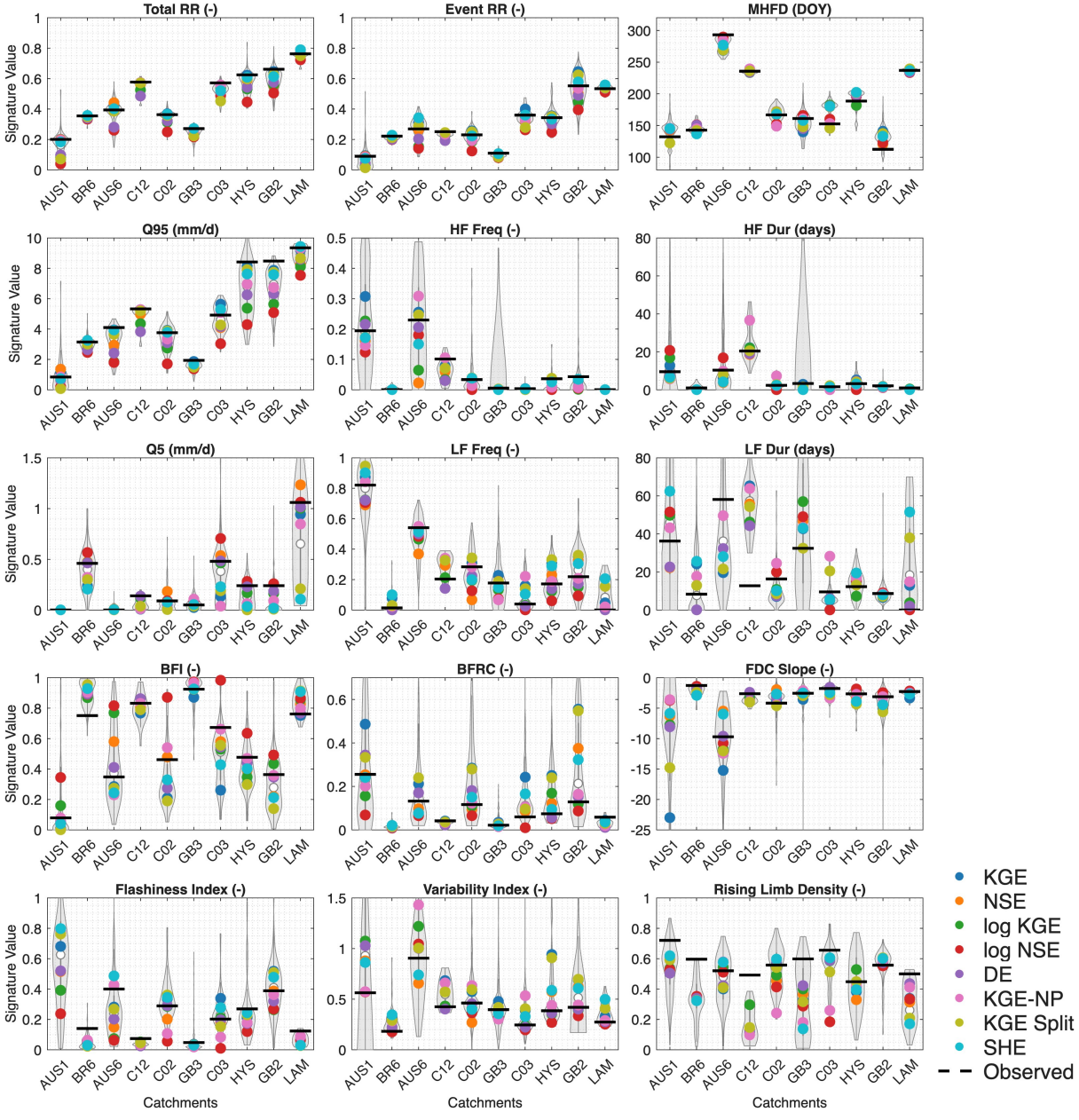

Figure 5. Comparison of signature representation capabilities across the different objective functions. The coloured points represent the median signature (across all models calibrated with a specific OF). The black dashes indicate the observed signature values. The grey violins represent the variability of the modelled signature values of the individual models. In catchment AUS1, the observed value for the FDC slope is NaN, because the observed flow is 0 mm/d for more than 66% of the time steps in calibration period.

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

the objective function has on these representations. The x-axis denotes the 10 selected catchments and the y-axis gives the signature value in its native unit (see Table 5).

Initially, we can use this plot to compare the violin ranges to the observed values of signatures. For the vast majority of cases, the observed value is always within the boundaries of the model spread, which indicates that the tested conceptual models are able to capture the hydrologic variability across diverse catchments. The spread of the colored dots around the observed values indicates different degrees of accuracy in the signature representation depending on the calibration metric used. For some signatures, the range of signature values is notably smaller (e.g. Total RR) than for others (e.g. Rising Limb Density). There are signatures (e.g. Q95, BFI) where the objective functions differ strongly from each other and signatures where different objective functions lead to similar modelled values (e.g. MHFD, HF Dur). To move beyond these very broad conclusions, the next section will use the previously introduced error metric (Equation 1 from Section 2.5.1) to assess the relative skill of each objective function in more detail.

3.3 Which OFs show strengths or weaknesses for a specific hydrological signature?

In this section we use Equation 1 to normalize the signature errors using the observed signature value.

Figure 6 shows the normalized error in signature representation for all objective functions. Generally, we can see that the performance varies largely depending both on the signature (each violin in a plot) and the catchment (each dot in a violin).

The KGE in Figure 6 has good performance for the Total Runoff Ratio and the high flow percentile Q95 (violin is clustered closely around the zero error line). This does not only apply to the median performance (white dot in the violin) but also has a high consistency among catchments (spread of violin). However, particularly for the low flow percentile Q5, FDC slope, BFI, and Variability Index the representation obtained from calibrating to KGE relative to the other objective functions is often worse.

The NSE performs mediocre on almost all of the evaluated signatures: the range within catchments (dots in each violin) is large and there is no clear pattern in signature estimation. This indicates that the suitability of the NSE metric can be very catchment-dependent. High Flow Frequency, Flashiness Index and Q95 are usually underestimated by models calibrated on the NSE. Compared to the KGE, the performance on Q5, FDC slope, and BFI is better in specific cases, but subject to large variability.

For log KGE and log NSE, we see similar patterns. Both (log KGE and log NSE) underestimate the runoff generation and high flows as well as the reactivity of the catchment (Flashiness Index). The log NSE is often the worst objective function among the eight investigated, indicated by values close to ± 1. Both metrics show improvements compared to regular KGE and NSE for FDC Slope and Q5, but not necessarily for low flow duration (LF Dur) and frequency (LF Freq). It is important to notice that the FDC Slope can also be met well if the runoff is systematically biased, which is the case with underestimation here. In many of the evaluated 15 signatures, the log KGE shows a better performance than the log NSE (violin closer to 0), but this cannot be generalized across locations based on this assessment.

The diagnostic efficiency generally performs similarly to the log KGE. It shows better performance for Q5 and BFI but has comparable issues for Total RR and Q95. Compared to the KGE, the errors are constrained better across catchments. In

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

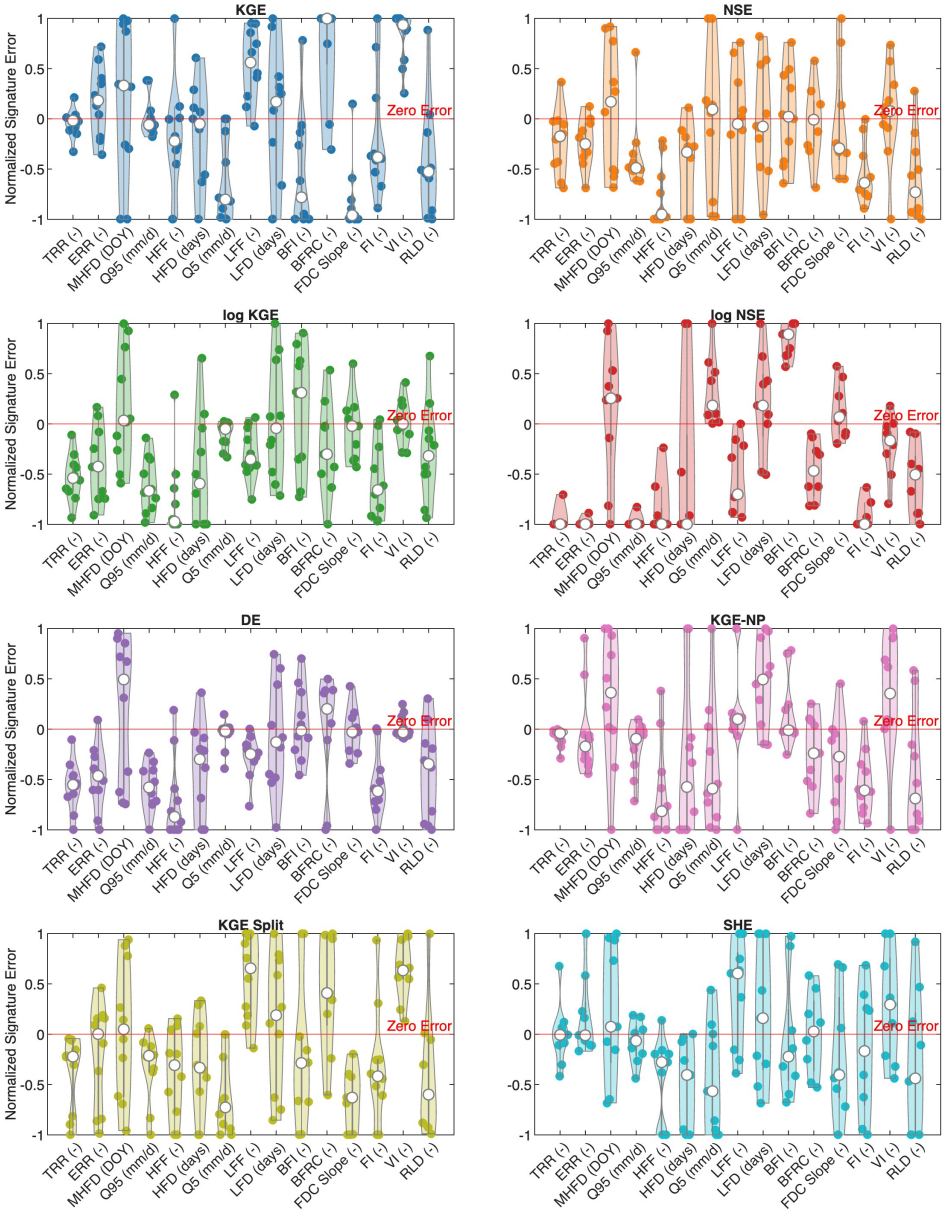

Figure 6. Error Metric as given in Equation 1 for all Objective Functions. The error metric is calculated for each objective function (subplot) per signature (x-axis) and catchment (dots in violin). A value of 0 indicates no error in signature representation and a value of ± 1 indicates the largest under- or overestimation error

conclusion, the diagnostic efficiency proves to be a veritable alternative for low-flow calibration, while offering benefits on additional signatures such as BFI and VI).

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

The non-parametric version of the KGE performs slightly worse than the KGE on high flows, but improves other signatures strongly, such as the BFI and FDC Slope. The KGE-NP is among the objective functions that perform best overall (smaller violins). The expectations (general improvement due to more information being included) for KGE-NP are thus largely met, with the main benefit being better incorporation of flow variability for a larger fraction of flow percentiles.

For KGE Split, we see similar improvements as for KGE-NP but they are not as strong as for the non-parametric KGE as the violin spread remains large. Like the NSE, it has no signature that is represented both accurately and consistently. Compared to the regular KGE, there are slight improvements on selected signatures (FDC Slope, BFI, Variability Index), but none are clear enough to recommend this OF for a specific aspect.

Lastly, the SHE is one of the objective functions with the largest range in signature representation. The general patterns are similar to the KGE (as expected due to their similar formulation), except for selected signatures like FDC slope. This indicates that a change from value-based correlation to rank-based can be influential for signature representation.

The violin plots also show that some signatures are affected very little by the choice of objective function such as the Mean Half Flow Date and the Rising Limb Density, this is indicated both by a large spread of the violin plots and similar distributions among objective functions. This suggests that either the models or the catchment mainly influence the results. As the signature value cannot be simulated well from any of the models, this shows the limits of objective function influence.

This Section showed the skills and shortcomings of the individual objective functions regarding signature representation. However, it is not easy to assess whether the observed influences that were described can be considered significant for the process representation. To investigate this, the next section uses statistical testing.

3.4 Does the OF choice significantly affect signature representation?

Figure 7 shows the distributions of p-values, when testing the significance of signature values calculated with different OFs against each other. The red and black lines mark the commonly used significance levels of 0.05 and 0.1. Every signature that is below these values can be considered to have a statistically significant difference between the paired samples of signature value distributions (in other words, the two objective functions in the pair lead to statistically different values for the signature across all included models). For these signatures, the choice of objective function thus plays a large role.

All signatures show a large range in p-values, which implies that there are both objective functions which have similar and different representation for each signature. Some OFs typically behave similar, e.g. the distributions of log KGE and log NSE are often almost identical for low flow signatures. Figure 7 shows that the choice of objective function is relevant only for the following tested signatures: the Total RR, Event RR, Q5, Q95, LF Freq, BFI, BRFC, and the Flashiness Index. For the remaining 7 signatures (Low and High Flow Duration, HF Freq, Mean Half Flow Date, FDC Slope, Variability Index, Rising Limb Density) a change of objective function does not lead to a clear shift in signature representation for most of the combinations of objective functions. Table S4 in the Supplementary Material S2.8 shows the frequency of objective functions pairs reaching p-values of below 0.10 or 0.05 respectively.

By comparing Figure 7 to Figure 5 we can conclude that there can be a lot of spread in the representation of the signatures that do not show a significant difference. This is because the spread for signatures like LF Duration, Variability Index, or

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

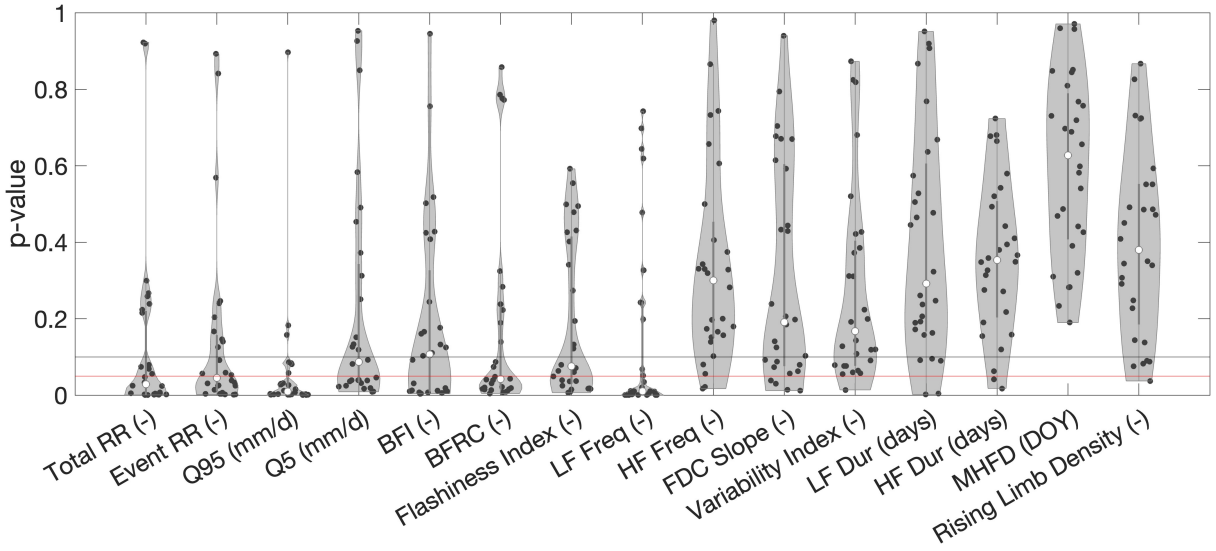

Figure 7. Significance Test between signature representations of the different objective functions. If the median value of a violin plot (white dot) lies below the significance levels of 0.05 or 0.1, the choice of objective function plays a significant role in the representation of this signature.

Rising Limb Density is often not consistent, meaning that when comparing the distribution of signature values of two objective functions with each other, there is variation as to which one has higher or lower values. Therefore, the significance often shows no significant impact because the signal is not clear. The significance test indicates that the variation is not driven by the objective function and we speculate that in these cases other influences such as catchment characteristics or models drive the spread seen in the signature representation.

3.5 What controls simulated signature values?

By running a random forest analysis regarding the importance of the three varying components of this modelling experiment (catchment, model, objective function), we can further evaluate the importance of the objective function choice by comparing it to the other two varying factors. Figure 8 shows the results of a random forest feature importance analysis.

We can see that all signatures are dominated by the influence of the catchment, implying that the strongest driver of signature representation is location change. This is reasonable as climatic and landscape attributes are typically viewed as the most important drivers of hydrologic behaviour. The choice of objective function, however, has relevant predictive importance for a large number of the signatures. Variation in objective function is typically even more impactful than a change of the hydrological model. Particularly for the baseflow signatures (BFI and BFRC) and Q5 the importance of the objective function combined with the model choice even challenges the catchment as the most important predictor. Model choice seems to play a more important role in representing Q5, BFRC and the Flashiness Index.

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

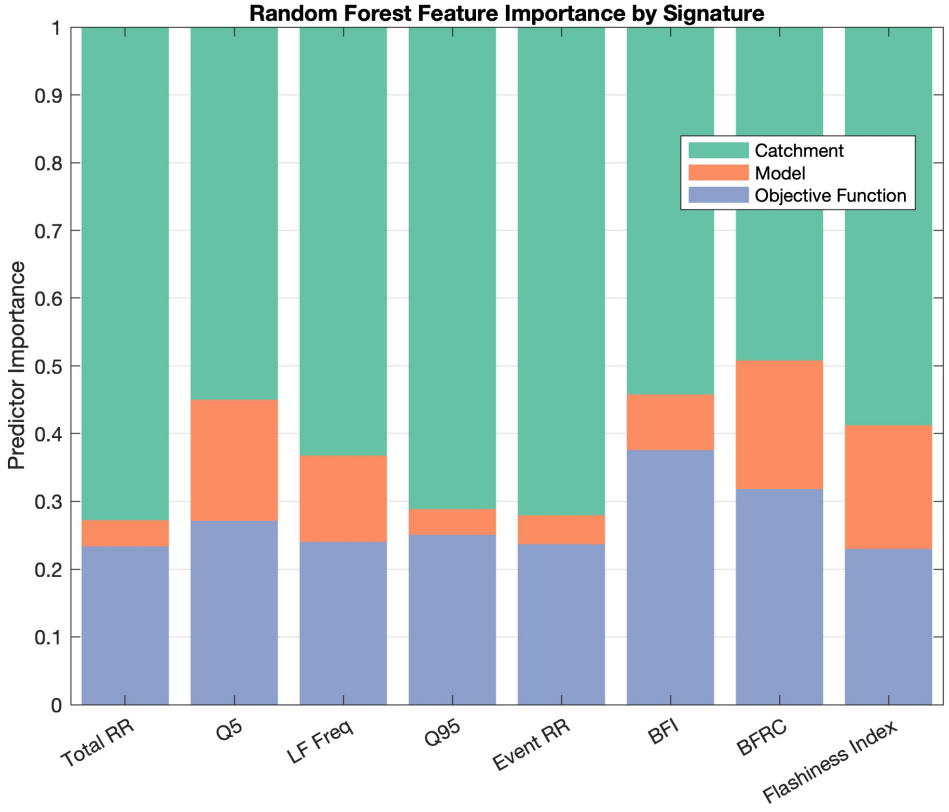

Figure 8. Random forest feature importance analysis of all varying components in this modelling experiment (catchments, models, objective functions)

3.6 Overall Skill of Signature Representation

Our results suggest that all objective functions can represent some signatures well while having shortcomings in others. This implies that, for our selection of objective functions, there is not a single one that will ensure realistic values across the spectrum of streamflow signatures we investigated. Yet, we are able to identify objective functions that generally have good process representation on most signatures. We will therefore investigate the accumulated errors of all significantly influenced signatures as identified from Figure 7. For this we use equation 2 to get one median error value per signature and objective function. Figure 9 shows this accumulated error for the eight tested objective functions.

This plot indicates a few interesting results. The modeled overall error in signature representation varies strongly depending on the objective function chosen. Considering the entire set of signatures with significant metric influence, the KGE-NP, SHE, and DE show the best overall performance. Conversely, this study identifies the log NSE as the worst objective functions overall. This, however, does not immediately imply that certain OFs should not be used as each has its strengths and weaknesses. As shown before, the KGE is the best metric for high flow conditions (small errors for Q95) and works well for Total RR, while it

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

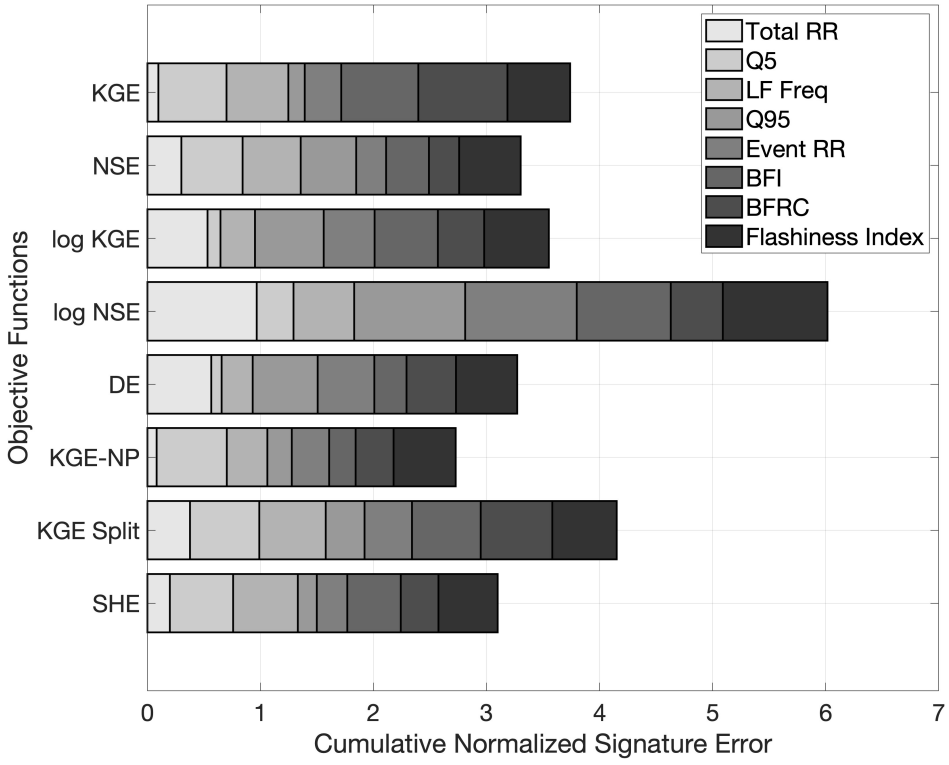

Figure 9. Cumulative error in significantly influenced signatures per tested objective function.

has considerable weaknesses for low flow representation (e.g. BFI, BFRC and Q5). And while the log NSE performs well on low flow conditions (Q5) it is considerably worse for all other significant signatures. Similarly, despite the KGE-NP showing the best overall performance, it still comes with weaknesses in Q5 and FI representation.

4 Discussion

We structured our discussion into three parts. First, we synthesize the main takeaways for signature representation across models, catchments, and objective functions. We condense our findings into some practical recommendations. Second, we discuss the correlation between different signatures and how they relate to the identified impact of objective function choice. Third, we outline the limitations of our study and potential future work.

4.1 Main Takeaways and Practical Recommendations

Across 47 model structures and 10 hydro-climatically diverse basins, OF choice exerts a systematic but selective influence on simulated streamflow signatures. A paired significance test indicated that 8 of 15 signatures vary significantly with OF choice (Fig. 7), whereas timing descriptors such as MHFD and RLD are largely OF-insensitive. This may in part be due to

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

the selection of OFs, as objective functions that specifically target time offset (e.g Mean Absolute Peak Time Error; MAPTE) are less frequently used (Ehret and Zehe, 2011) than the OFs we did select and were thus not tested in our study. Generally, no single OF excels at all tested signatures. Considering the cumulative error across OF-sensitive signatures (Fig. 9), KGE-NP yields the lowest overall error, with SHE, DE, and NSE performing competitively in some basins but less consistently across signatures. And while log-transformed OFs reproduce low flow metrics better (Q5, LF Freq), they come with considerable tradeoffs in the representation of other signatures such as Total RR, baseflow related signatures, or HF signatures.

Our analyses hence contributes to the available information of specific strengths and weaknesses of individual OFs. KGE, for example, most reliably reproduces Q95, and KGE-NP is one of the best OFs for total runoff ratio and baseflow characteristics (Figure 5 and 6). Log NSE in comparison showed few strengths for representing hydrological signatures in our experiments, but log KGE did well on Q5 and the variability index (VI). However, the DE shows those same strengths, while having a smaller overall error. The KGE-Split, NSE, and SHE show a balanced error field with individual differences in cumulative errors (Fig. 9). While information like this helps to guide the choice of OF for purpose-based model calibration, it remains important to consider weaknesses of the individual metrics that have not been part of our experiments. Log KGE, for example, has been shown to lead to numerical issues and should be avoided according to Santos et al. (2018), and the KGE has been criticized for leading to counterbalancing errors as described in (Cinkus et al., 2023). Thus, it remains important to consider all strengths and weaknesses of a specific OF and how it may influence the acquired modelling results.

We also compared the influence of OF choice on signature representation with other influencing factors such as catchment characteristics or model structure choice. Our random-forest analysis showed that catchment characteristics are the dominant control of signature representation, with OF choice typically more influential than model choice and therefore placing second (Figure 8). Model choice does, however, gain relevance for signatures connected to low flow or baseflow representation, while OF choice seemed especially relevant for baseflow related signatures. Together, our results support the argument for a purpose-based model calibration, that considers multiple aspects of the flow regime, and multi-objective calibration setups, rather than defaulting to a familiar single metric (Mai, 2023; Jackson et al., 2019).

For the signatures that were found to be significantly impacted by OF choice, we used our results to develop guidelines for which OF will perform well in reproducing a given signature as shown in Table 6. We can conclude that every OF leads to the best representation for at least one signature, highlighting that no OF is best across all signatures. However, if the goal is a balanced representation of the water-balance and variability across hydrologic regimes, KGE-NP is a strong option. If low flows are central (ecological flows, drought assessment), preference should be given to DE or log-KGE. If peak flows are the priority (flood risk, infrastructure design), KGE generally performs best for Q95 and related high-flow signatures. Reflecting back on our review of metric choice justification (Fig. 1), we strongly recommend to document the trade-offs one is willing to accept explicitly (e.g., low-flow degradation under KGE) and, where possible, to verify unaffected signatures (e.g., MHFD, RLD) since OF changes have limited influence there.

When multiple aspects of the flow regime matter simultaneously, multi-objective calibration has been shown to provide considerable benefits (Hernandez-Suarez et al., 2018; Nemri and Kinnard, 2020). Multi-objective formulations can expose the trade-offs explicitly via a Pareto front and allow practitioners to choose solutions that best satisfy competing goals (Yapo et al.,

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

Table 6. Recommendations for objective functions that can represent the 15 considered signatures well according to our study results. Where a signature was not significantly affected by OF choice, we report "not significant".

Signature	**Best OF**
Total Runoff Ratio	KGE-NP
Event Runoff Ratio	SHE, KGE-NP
Mean Half Flow Date	Not significant
FDC Slope	Not significant
5th Flow Percentile (Q5)	log KGE, DE
Low-Flow Duration	Not significant
Low-Flow Frequency	DE, log KGE
Baseflow Index (BFI)	KGE-NP, DE
95th Flow Percentile (Q95)	KGE, SHE
High-Flow Duration	Not significant
High-Flow Frequency	Not significant
Baseflow Recession Coefficient (BFRC)	SHE, NSE
Rising Limb Density (RLD)	Not significant
Flashiness Index (FI)	Split-KGE, SHE
Variability Index (VI)	DE (where relevant)

1998; Efstratiadis and Koutsoyiannis, 2010; Mai, 2023). If a single composite metric must be used, a transparent weighting that reflects decision priorities can be helpful. Vis et al. (2015), for instance, argue for directly incorporating multiple ecological-flow characteristics in the OF to improve the relevance of the model for management applications. Similarly, signature values have directly been used in calibration, either as additional objectives or as constraints, to target process realism (Gupta et al., 2008; Pool et al., 2017). This can potentially improve baseflow behaviour, low-flow frequency, or water-balance partitioning when those aspects are central. However, incorporating signatures does not guarantee a better hindcast/forecast skill in all settings, and can even degrade time-series fit if misapplied (Araya et al., 2023).

4.2 Impact of Signature Correlations

Signatures are not independent from each other. When investigating correlations among observed signatures (Fig. 10) we noticed three clusters with similar patterns. First, a *runoff/high-flow* cluster (Total RR, Event RR, and Q95) shows strong positive association, linking long-term water partitioning to the reaction to events. This aligns with aridity being a primary control on mean runoff (Berghuijs et al., 2017). Second, a *baseflow/storage* cluster, including BFI, BFRC and FI, exhibits tight coupling (higher BFI ↔ lower BFRC and FI) and a strong relationship to low-flow frequency, reflecting how sustained baseflow suppresses low-flow events. Similar to the first cluster, this group was also sensitive to objective-function (OF) choice. Third, a *variability/threshold* cluster (FDC slope, HF/LF frequency, low-flow duration, VI) captures distributional shape and

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

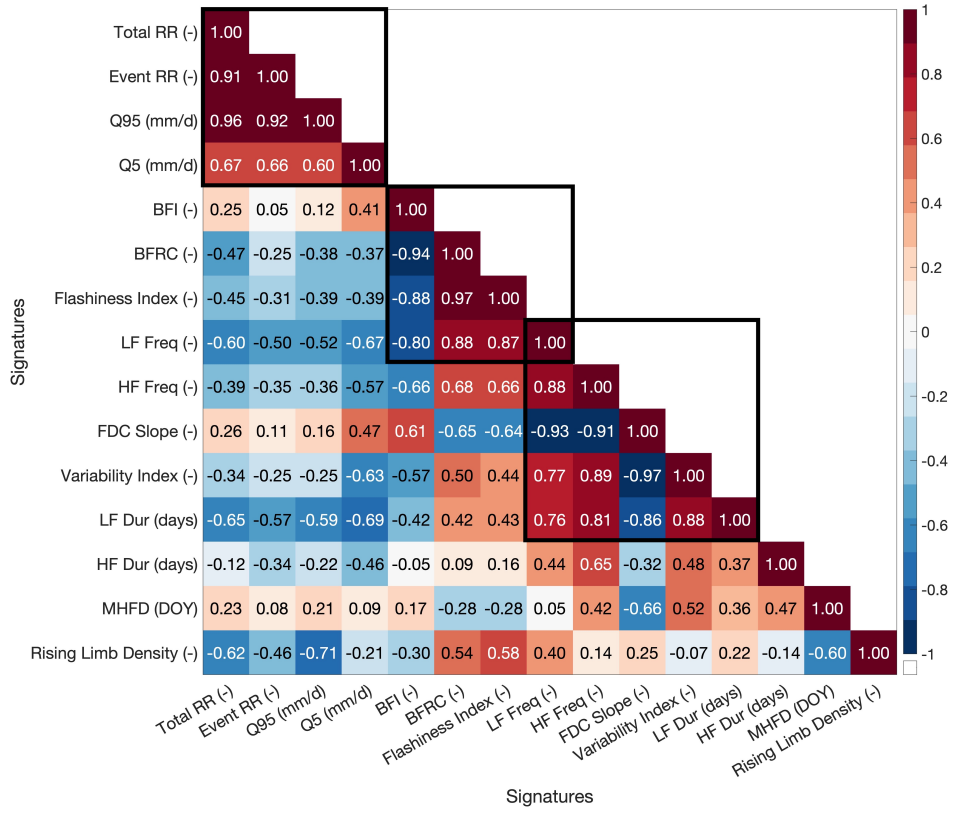

Figure 10. Correlation test between signatures based on Observations. Please note, that the signatures have been reorganized in this plot to highlight the emerging patterns. Supplement S2.9 Fig. S19 provides a similar plot for the correlations between the modelled signatures.

threshold exceedance. These signatures, derived from the hydrograph, partly encode duplicate information and were mostly OF-insensitive. Several other signatures showed weak or inconsistent ties to these clusters, such as High-Flow Duration, Mean Half-Flow Date, Rising-Limb Density.

These clusters help explain why improving one signature typically moves its cluster-mates, yet they do not imply that different OFs are needed for each cluster. For example, KGE-NP performed best for many signatures in the first two clusters (e.g., Total/Event RR, BFI).

We identified some overlap between signatures that are strongly/weakly affected by OF choice and those that were identified as more/less predictable from hydrologic drivers by Addor et al. (2018): some signatures (e.g., FDC slope, durations) tend to be OF-insensitive with low spatial prediability, whereas water-balance and high-flow signatures were more OF-sensitive and spatially predicable. A notable exception is Mean Half-Flow Date, which was well predicted in Addor et al. (2018) but showed little OF sensitivity in our study suggesting a stronger climatic control.

If we consider the structural differences among OFs some of these patterns become clearer. For *bias/mean* terms, KGE's mean component directly constrains volumes, explaining its systematic advantage for Total RR and Q95. OFs that deviate from

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

this approach (NSE, DE, log-KGE/NSE; Table 3) typically perform worse for these metrics (Gupta et al., 2009; Mizukami et al., 2019). Across objective functions, Total RR tends to be underestimated (Fig. 4), reflecting systematic water-balance biases related to excessive storage or evapotranspiration. This pattern underscores the importance of bias terms in objective functions, which directly constrain mean flow and promote a more realistic overall water balance (Mizukami et al., 2019). This underestimation is strongest for low-flow-focused objectives (log-metrics), as also noted by Vis et al. (2015). The claim that KGE resolves NSE's low-flow insensitivity or performs well simultaneously for high and low flows (Althoff and Rodrigues, 2021), but note that Althoff and Rodrigues (2021) investigates more basins, while we test a larger number of models.

For *variability* terms, replacing variance ratios with FDC-based components (as done for the KGE-NP) or integrating relative errors along the FDC (as done for the DE) re-targets calibration toward distributional shape and mid/low flows, improving BFI, Q5, and Event RR in our results (Fig. 6). While log-metrics and DE enhance low-flow behaviour, log transformations carry numerical pitfalls (Santos et al., 2018); $1/Q$ variants can retain sensitivity with fewer issues (Pushpalatha et al., 2012).

For *correlation* terms, moving from Pearson to Spearman (KGE-NP, SHE, DE) reduces sensitivity to extremes and stabilizes correlation. A direct comparison of KGE and SHE, where the dependence component is the only structural change, suggests gains for FDC slope, VI, and baseflow metrics, with similar or slightly worse performance for high-flow signatures (Kiraz et al., 2023). Annual reweighting via Split-KGE yielded mixed, context-dependent benefits and no systematic improvement of low-flow characteristics over baseline KGE in our experiments (Fig. 9; Fowler et al., 2018).

Overall, each OF emphasizes the hydrograph component it encodes (bias, variability, dependence), none dominates across all signatures. This aligns with evidence that flood characteristics are more sensitive to metric choice than droughts (Melsen et al., 2019) and with comparative studies of KGE vs. NSE (Mizukami et al., 2019).

4.3 Limitations and Further Research

Our study aimed to fill a specific research gap in current understanding of the impact of objective function choice on signature representation; using a larger number of models for this type of analysis (see Table 1). To enable a thorough investigation of the results we necessarily had to impose certain limits in other parts of the experimental design (i.e., having to "balance depth with breadth", Gupta et al. (2014)).

First, compared to most studies in Table 1 we use a limited number of catchments. This keeps computational cost manageable and allows more detailed analysis into individual modelling results than would otherwise be possible. We selected these basins to ensure a hydro-climatic spread, as is common practice in these scenarios (see for example the 12 MOPEX basins that have been used for a large number of studies Duan et al. (e.g. 2006); van Werkhoven et al. (e.g. 2008); Carrillo et al. (e.g. 2011)). Compared to existing studies, our selected number of objective functions and signatures is fairly typical, while our number of models is clearly higher.

Second, our experiments necessarily exclude parameter uncertainty: every combination of model, catchment and objective function is calibrated only once. To briefly investigate the impact of parameter uncertainty, we re-calibrated one model (GR4J) with multiple random seeds (Fig. S20 in the Supplementary Material S2.10, catchment C03). Optima were stable across seeds for KGE, NSE, and DE, whereas log-KGE, KGE-NP, and Split-KGE returned different parameter sets with similar OF scores.

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

Importantly however, the best-performing parameter sets had very similar parameters and signature representation was nearly identical across these sets, showing that in this specific case parameter uncertainty is not a large concern. However, more work on this is needed.

Going forward, we see four priorities to build on the individual strengths of existing work (Table 1). First, broaden external validity with a larger, stratified basin sample (including ephemeral and groundwater-dominated systems) and potentially a wider model palette spanning physically-based and ML models, using hierarchical/mixed-effects analyses to partition variance among catchment, OF choice, and model structure. Second, test cluster-aware multi-objective calibration that deliberately pairs complementary OFs—e.g., a water-balance/high-flow metric (KGE or KGE-NP) with a low-flow/storage metric (DE or log-KGE), optionally adding an FDC-shape term—and evaluate gains via Pareto hypervolume and the change in worst-signature error within each cluster. Third, make component-level experiments less ambiguous by swapping KGE-type components one-by-one in a controlled scaffold (mexwan: runoff-ratio vs. P-normalized mean; variability: stdev ratio vs. FDC-shape; dependence: Pearson vs. Spearman vs. short-lag correlation), and test weighted KGE-type variants; an accompanying analytical test case should trace how components propagate into variability metrics (e.g., VI). Finally, propagate forcing/observation and signature-method uncertainty (e.g., baseflow separation) and explore information-rich objectives that go beyond the bias-variability-correlation trio, such as distributional divergences along the flow-duration curve, mutual-information-style dependence measures, or terms tied to process diagnostics (recession-slope distributions, intermittency, precipitation–discharge hysteresis), to link calibration more directly to hydrologic processes (Kirchner, 2006). Designs like KGE-NP already go in this direction and offer a principled path to retain low-flow sensitivity without relying solely on log transforms.

5 Conclusions

We calibrated 47 conceptual model structures across 10 hydro-climatically diverse basins using 8 objective functions (OFs) and evaluated 15 streamflow signatures. Three results stand out. First, OF choice exerts a *selective* influence: 8 of 15 tested signatures are OF-sensitive, while others (e.g. mean half-flow date and rising-limb density, and often event durations) change only little with OF choice. Second, no single OF performs best across all signatures; aggregated across OF-sensitive signatures, KGE-NP yields the smallest overall error across signatures. Third, the direction of change is interpretable from OF design: metrics emphasizing mean volumes favour water balance and high flows; FDC-based and rank-based terms improve storage/low-flow behaviour.

These patterns translate into simple steps for guiding OF choice. For accurate water balance and high-flow magnitude, KGE (or KGE-NP if distributional shape matters) is effective. When low flows and storage-related signatures are central, DE or log-KGE (noting log-transform caveats) can be preferred. When multiple signature clusters matter simultaneously, multi-objective calibration (e.g. pair KGE-NP with DE or a high-flow metric) can help to manage trade-offs explicitly.

Catchment differences dominate overall variability, with OF choice typically more influential than model structure in our setup. Our findings are bounded by the selected basins, conceptual structures, signature methods, and deterministic

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

calibration. Priorities for future work include cluster-aware multi-objective formulations, broader regime and structure coverage (snow/glacier and groundwater-explicit models), and low-flow objective functions that retain sensitivity without log pitfalls.

In short, knowing the trade-offs in metric selection offers a practical path to models that reproduce the signatures that matter for the question at hand.

Code and data availability

. All code used in this analysis will be provided in a Github repository. The used data can be downloaded from the Caravan repository on Zenodo.

Author contributions

. PW: methodology, data curation, formal analysis, investigation, visualization, writing - original draft, writing - review and editing; WJMK: supervision, investigation, writing - review and editing; NS: writing - review and editing, supervision; DS: conceptualization, methodology, investigation, writing - original draft, writing - review and editing, supervision

Competing interests

. No competing interests are being declared.

Acknowledgements

. The authors are grateful to the Center for Information Services and High Performance Computing [Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH)] at TUD Dresden University of Technology, for providing its facilities for high throughput calculations. PW and WK were supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) with funding under award NA22NWS4320003 from the NOAA Cooperative Institute Program. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the opinions of NOAA. During the preparation of this work, some of the authors used ChatGPT in order to improve language and readability of the original draft, with caution. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrology and Earth System Sciences, 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Vine, N. L., and Clark, M. P.: A Ranking of Hydrological Signatures Based on Their Predictability in Space, Water Resources Research, 54, 8792–8812, https://doi.org/10.1029/2018WR022606, 2018.

Althoff, D. and Rodrigues, L. N.: Goodness-of-fit criteria for hydrological models: Model calibration and performance assessment, Journal of Hydrology, 600, 126 674, https://doi.org/10.1016/j.jhydrol.2021.126674, 2021.

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., and Ayala, A.: The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset, Hydrology and Earth System Sciences, 22, 5817–5846, https://doi.org/10.5194/hess-22-5817-2018, 2018.

Araya, D., Mendoza, P. A., Muñoz-Castro, E., and McPhee, J.: Towards robust seasonal streamflow forecasts in mountainous catchments: impact of calibration metric selection in hydrological modeling, Hydrology and Earth System Sciences, 27, 4385–4408, https://doi.org/10.5194/hess-27-4385-2023, 2023.

Arsenault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez, M. C., Ameli, A., and Poulin, A.: A comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds, Scientific Data, 7, 243, https://doi.org/10.1038/s41597-020-00583-2, 2020.

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.

Berghuijs, W. R., Larsen, J. R., Van Emmerik, T. H. M., and Woods, R. A.: A Global Assessment of Runoff Sensitivity to Changes in Precipitation, Potential Evaporation, and Other Factors, Water Resources Research, 53, 8475–8486, https://doi.org/10.1002/2017WR021593, 2017.

Carrillo, G., Troch, P. A., Sivapalan, M., Wagener, T., Harman, C., and Sawicz, K.: Catchment classification: hydrological analysis of catchment behavior through process-based modeling along a climate gradient, Hydrology and Earth System Sciences, 15, 3411–3430, https://doi.org/10.5194/hess-15-3411-2011, 2011.

Chagas, V. B. P., Chaffe, P. L. B., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C. D., and Siqueira, V. A.: CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil, Earth System Science Data, 12, 2075–2096, https://doi.org/10.5194/essd-12-2075-2020, 2020.

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.

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J. K., Lane, R., Lewis, M., Robinson, E. L., Wagener, T., and Woods, R.: CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain, Earth System Science Data, 12, 2459–2483, https://doi.org/10.5194/essd-12-2459-2020, 2020.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H., Gusev, Y., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., and Wood, E.: Model Parameter Estimation

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, Journal of Hydrology, 320, 3–17, https://doi.org/10.1016/j.jhydrol.2005.07.031, 2006.

Efstratiadis, A. and Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: a review, Hydrological Sciences Journal, 55, 58–78, https://doi.org/10.1080/02626660903526292, 2010.

Ehret, U. and Zehe, E.: Series distance – an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events, Hydrology and Earth System Sciences, 15, 877–896, https://doi.org/10.5194/hess-15-877-2011, 2011.

Fowler, K., Peel, M., Western, A., and Zhang, L.: Improved Rainfall-Runoff Calibration for Drying Climate: Choice of Objective Function, Water Resources Research, 54, 3392–3408, https://doi.org/10.1029/2017WR022466, 2018.

Fowler, K. J. A., Acharya, S. C., Addor, N., Chou, C., and Peel, M. C.: CAMELS-AUS: hydrometeorological time series and landscape attributes for 222 catchments in Australia, Earth System Science Data, 13, 3847–3867, https://doi.org/10.5194/essd-13-3847-2021, 2021.

Färber, C., Plessow, H., Mischel, S. A., Kratzert, F., Addor, N., Shalev, G., and Looser, U.: GRDC-Caravan: extending Caravan with data from the Global Runoff Data Centre, Earth System Science Data, 17, 4613–4625, https://doi.org/10.5194/essd-17-4613-2025, 2025.

Gnann, S. J., Coxon, G., Woods, R. A., Howden, N. J., and McMillan, H. K.: TOSSH: A Toolbox for Streamflow Signatures in Hydrology, Environmental Modelling and Software, 138, 104 983, https://doi.org/10.1016/j.envsoft.2021.104983, 2021.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrological Processes, 22, 3802–3813, https://doi.org/10.1002/hyp.6989, 2008.

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

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrology and Earth System Sciences, 18, 463–477, https://doi.org/10.5194/hess-18-463-2014, 2014.

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.

Hansen, N., Müller, S. D., and Koumoutsakos, P.: Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES), Evolutionary Computation, 11, 1–18, https://doi.org/10.1162/106365603321828970, 2003.

Hernandez-Suarez, J. S., Nejadhashemi, A. P., Kropp, I. M., Abouali, M., Zhang, Z., and Deb, K.: Evaluation of the impacts of hydrologic model calibration methods on predictability of ecologically-relevant hydrologic indices, Journal of Hydrology, 564, 758–772, https://doi.org/10.1016/j.jhydrol.2018.07.056, 2018.

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

Jacobs, B., Tobi, H., and Hengeveld, G. M.: Linking error measures to model questions, Ecological Modelling, 487, 110 562, https://doi.org/10.1016/j.ecolmodel.2023.110562, 2024.

Jakeman, A., Letcher, R., and Norton, J.: Ten iterative steps in development and evaluation of environmental models, Environmental Modelling & Software, 21, 602–614, https://doi.org/10.1016/j.envsoft.2006.01.004, 2006.

Jakeman, A. J., Sawah, S. E., Cuddy, S., Robson, B., McIntyre, N., and Cook, F.: Good Modelling Practice Principles, Tech. rep., 2018.

https://doi.org/10.5194/egusphere-2025-5413
Preprint. Discussion started: 27 November 2025

Kiraz, M., Coxon, G., and Wagener, T.: A Signature-Based Hydrologic Efficiency Metric for Model Calibration and Evaluation in Gauged and Ungauged Catchments, Water Resources Research, 59, https://doi.org/10.1029/2023WR035321, 2023.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resources Research, 42, https://doi.org/10.1029/2005WR004362, 2006.

Klingler, C., Schulz, K., and Herrnegger, M.: LamaH-CE: LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe, Earth System Science Data, 13, 4529–4565, https://doi.org/10.5194/essd-13-4529-2021, 2021.

Knoben, W. J. M.: Setting expectations for hydrologic model performance with an ensemble of simple benchmarks, Hydrological Processes, 38, https://doi.org/10.1002/hyp.15288, 2024.

Knoben, W. J. M., Woods, R. A., and Freer, J. E.: A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data, Water Resources Research, 54, 5088–5109, https://doi.org/10.1029/2018WR022913, 2018.

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

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, 2019b.

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, https://doi.org/10.1029/2019wr025975, 2020.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community dataset for large-sample hydrology, Scientific Data, 10, https://doi.org/10.1038/s41597-023-01975-w, 2023.

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.

Mai, J.: Ten strategies towards successful calibration of environmental models, Journal of Hydrology, 620, 129 414, https://doi.org/10.1016/j.jhydrol.2023.129414, 2023.

McMillan, H.: Linking hydrologic signatures to hydrologic processes: A review, Hydrological Processes, 34, 1393–1409, https://doi.org/10.1002/hyp.13632, 2020.

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

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

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

Nash, J. and Sutcliffe, J.: 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-2025-5413
Preprint. Discussion started: 27 November 2025

Nemri, S. and Kinnard, C.: Comparing calibration strategies of a conceptual snow hydrology model and their impact on model performance and parameter identifiability, Journal of Hydrology, 582, 124 474, https://doi.org/10.1016/j.jhydrol.2019.124474, 2020.

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.

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

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

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.

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

Schwemmle, R., Demand, D., and Weiler, M.: Technical note: Diagnostic efficiency – specific evaluation of model performance, Hydrology and Earth System Sciences, 25, 2187–2198, https://doi.org/10.5194/hess-25-2187-2021, 2021.

Seiller, G., Roy, R., and Anctil, F.: Influence of three common calibration metrics on the diagnosis of climate change impacts on water resources, Journal of Hydrology, 547, 280–295, https://doi.org/10.1016/j.jhydrol.2017.02.004, 2017.

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.

Trotter, L. and Knoben, W. J. M.: MARRMoT v2.1, https://doi.org/10.5281/zenodo.6484372, 2022.

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

van Werkhoven, K., Wagener, T., Reed, P., and Tang, Y.: Characterization of watershed model behavior across a hydroclimatic gradient, Water Resources Research, 44, https://doi.org/https://doi.org/10.1029/2007WR006271, 2008.

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.

Waveren, R. v., Groot, S., Scholten, H., Geer, F. v., Wösten, J., Koeze, R., and Noort, J.: Good Modelling Practice Handbook, Tech. rep., 1999.

Yapo, P. O., Gupta, H. V., and Sorooshian, S.: Multi-objective global optimization for hydrologic models, Journal of Hydrology, 204, 83–97, https://doi.org/10.1016/S0022-1694(97)00107-8, 1998.