Hydrol. Earth Syst. Sci., 27, 2397–2411, 2023
https://doi.org/10.5194/hess-27-2397-2023

Hydrology and
Earth System
Sciences
Open Access

When best is the enemy of good – critical evaluation of performance criteria in hydrological models

Guillaume Cinkus[1], **Naomi Mazzilli**[2], **Hervé Jourde**[1], **Andreas Wunsch**[3], **Tanja Liesch**[3], **Nataša Ravbar**[4], **Zhao Chen**[5], **and Nico Goldscheider**[3]

[1]HydroSciences Montpellier (HSM), CNRS, IRD, Univ. Montpellier, 34090 Montpellier, France
[2]UMR 1114 EMMAH (AU-INRAE), Université d'Avignon, 84000 Avignon, France
[3]Institute of Applied Geosciences, Karlsruhe Institute of Technology (KIT), Kaiserstr. 12, 76131 Karlsruhe, Germany
[4]Karst Research Institute, ZRC SAZU, Titov trg 2, 6230 Postojna, Slovenia
[5]Institute of Groundwater Management, Technical University of Dresden, 01062 Dresden, Germany

Correspondence: Guillaume Cinkus (guillaume.cinkus@umontpellier.fr)

Received: 8 November 2022 – Discussion started: 15 November 2022
Revised: 29 April 2023 – Accepted: 18 May 2023 – Published: 3 July 2023

Abstract. Performance criteria play a key role in the calibration and evaluation of hydrological models and have been extensively developed and studied, but some of the most used criteria still have unknown pitfalls. This study set out to examine counterbalancing errors, which are inherent to the Kling–Gupta efficiency (KGE) and its variants. A total of nine performance criteria – including the KGE and its variants, as well as the Nash–Sutcliffe efficiency (NSE) and the modified index of agreement (d_1) – were analysed using synthetic time series and a real case study. Results showed that, when assessing a simulation, the score of the KGE and some of its variants can be increased by concurrent overestimation and underestimation of discharge. These counterbalancing errors may favour bias and variability parameters, therefore preserving an overall high score of the performance criteria. As bias and variability parameters generally account for two-thirds of the weight in the equation of performance criteria such as the KGE, this can lead to an overall higher criterion score without being associated with an increase in model relevance. We recommend using (i) performance criteria that are not or less prone to counterbalancing errors (d_1, modified KGE, non-parametric KGE, diagnostic efficiency) and/or (ii) scaling factors in the equation to reduce the influence of relative parameters.

1 Introduction

Hydrological models are fundamental to solving problems related to water resources. They help characterise hydrosystems (Hartmann et al., 2014), predict floods (Kauffeldt et al., 2016; Jain et al., 2018), and manage water resources (Muleta and Nicklow, 2005). A lot of research efforts are thus dedicated to improving the reliability, the robustness, and the relevance of such models. Improvements can be made by working on (i) input data, (ii) model parameters and structure, (iii) uncertainty quantification, (iv) model calibration (Beven, 2019), and also (v) appropriate benchmarks for assessing model performance (Seibert et al., 2018). In this study, we focus on the proper use of performance criteria for calibrating and evaluating hydrological models – an important part that can easily be overlooked (Jackson et al., 2019).

A performance criterion aims to evaluate the goodness of fit of a model to observed data. It is generally expressed as a score, for which the best value corresponds to a perfect fit between predictions and observations. In hydrology, the Nash–Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) is still one of the most commonly used criteria (Kling et al., 2012), although the past decade has seen a gain in popularity of alternatives (Clark et al., 2021), e.g. the Kling–Gupta efficiency (KGE) (Gupta et al., 2009). Many authors have pointed out the inherent limitations of using performance criteria, especially the fact that a single-score metric cannot reflect all relevant hydrological aspects of a model

(Gupta et al., 2009). The use of a multi-criteria framework is thus often emphasised to quantify different aspects of a model (Clark et al., 2021; Moriasi et al., 2015; Gupta et al., 1998; Jackson et al., 2019; van Werkhoven et al., 2009; Knoben et al., 2019; Althoff and Rodrigues, 2021; Ritter and Muñoz-Carpena, 2013; Krause et al., 2005; Legates and McCabe, 1999) alongside a scientific evaluation of the results (Biondi et al., 2012). Knoben et al. (2019), Althoff and Rodrigues (2021), and Clark et al. (2021) pointed out that modellers should carefully think about which aspects they consider to be the most important in their hydrological model and how to evaluate them. In relation to the assessment of model performance, Seibert et al. (2018) argued that the current benchmarks poorly reflect what could and should be expected of a model. They suggested that one should define lower and upper benchmarks based on the performance of a simple bucket-type model with few parameters using the same data set.

Performance criteria also have shortcomings at a distinctive level. A number of studies have identified several limitations of the NSE: (i) the contribution of the normalised bias depends on the discharge variability of the basin; (ii) discharge variability is inevitably underestimated because the NSE is maximised when the variability equals the correlation coefficient, which is always smaller than unity; and (iii) mean flow is not a meaningful benchmark for highly variable discharges (Gupta et al., 2009; Willmott et al., 2012). The KGE aims to address these limitations but also has its own issues (Gupta et al., 2009). Santos et al. (2018) identified pitfalls when using the KGE with a prior logarithmic transformation of the discharge. Knoben et al. (2019) warned against directly comparing NSE and KGE scores as the KGE has no inherent benchmark. Ritter and Muñoz-Carpena (2013) and Clark et al. (2021) showed that NSE and KGE scores can be strongly influenced by few data points, resulting in substantial uncertainties in the predictions.

What is not fully addressed yet is the trade-off between individual components (Wöhling et al., 2013) and especially the impact of counterbalancing errors induced by bias and variability parameters, which are integrated in many performance criteria. While accurate bias and variability are desired aspects of hydrological models, sometimes good evaluations may accidentally result from negative and positive values cancelling each other (Jackson et al., 2019; Massmann et al., 2018). This can be particularly detrimental to model calibration and evaluation as it generates an increase in the criterion score without necessarily being associated with a better model relevance. Some performance criteria naturally address this problem by using absolute or squared error values, but other criteria such as the KGE and its variants do not as they use relative errors. The aim of this study is to assess the extent to which criteria scores can be trusted for calibrating and evaluating hydrological models when predictions have concurrent over- and underestimated values. The influence of counterbalancing errors is evaluated on nine performance criteria including the NSE and KGE. This selection is far being from exhaustive but includes widely used and recently proposed KGE variants, as well as more traditional criteria such as the NSE or the modified index of agreement (d_1) for comparison purposes. We first use synthetic time series to highlight the counterbalancing-errors mechanism. Second, we show how counterbalancing errors can impair the interpretation of hydrological models in a real case study. Finally, we provide some recommendations about the use of scaling factors and the choice of appropriate performance criteria to nullify or reduce the influence of counterbalancing errors.

2 Performance criteria

2.1 Parameter description

All the performance criteria considered in this study are based on the same or similar statistical indicators, which are first described to avoid repetition.

We use $x_\mathrm{o}(t)$ and $x_\mathrm{s}(t)$ to refer to observed and simulated values of calibration variable x at a specific time step t; r and r_s correspond to the Pearson and the Spearman rank correlation coefficients (Freedman et al., 2007), respectively.

β is the ratio between the mean of simulated values μ_s and the mean of observed values μ_o:

$$\beta = \frac{\mu_\mathrm{s}}{\mu_\mathrm{o}}. \quad (1)$$

β_n corresponds to the bias (mean error) normalised by the standard deviation of observed values σ_o:

$$\beta_\mathrm{n} = \frac{\mu_\mathrm{s} - \mu_\mathrm{o}}{\sigma_\mathrm{o}}. \quad (2)$$

α is the ratio between the standard deviation of simulated values σ_s and the standard deviation of observed values σ_o:

$$\alpha = \frac{\sigma_\mathrm{s}}{\sigma_\mathrm{o}}. \quad (3)$$

γ is the ratio between the coefficient of the variation of simulated values ($\mathrm{CV_s} = \sigma_\mathrm{s}/\mu_\mathrm{s}$) and the coefficient of the variation of observed values ($\mathrm{CV_o} = \sigma_\mathrm{o}/\mu_\mathrm{o}$):

$$\gamma = \frac{\mathrm{CV_s}}{\mathrm{CV_o}}. \quad (4)$$

$\overline{B}_\mathrm{rel}$ and $|B_\mathrm{area}|$ (Schwemmle et al., 2021) are based on the flow duration curve (FDC). $B_\mathrm{rel}(i)$ is defined as the relative bias of the simulated and observed flow duration curves at the exceedance probability i:

$$B_\mathrm{rel} = \frac{x_\mathrm{s}(i) - x_\mathrm{o}(i)}{x_\mathrm{o}(i)}, \quad (5)$$

where $x_\mathrm{s}(i)$ and $x_\mathrm{o}(i)$ correspond to the simulated and observed values of the calibration variable at exceedance probability i. $\overline{B}_{\text{rel}}$ is the mean of $B_{\text{rel}}(i)$ when looking at n observations:

$$\overline{B}_{\text{rel}} = \frac{1}{n}\sum_{i=0}^{i=1} B_{\text{rel}}(i). \tag{6}$$

$|B_{\text{area}}|$ is calculated as follows:

$$|B_{\text{area}}| = \int_{0}^{1} |B_{\text{res}}(i)|\,\mathrm{d}i, \tag{7}$$

with B_{res} being the residual bias:

$$B_{\text{res}} = B_{\text{rel}}(i) - \overline{B}_{\text{rel}}. \tag{8}$$

α_{NP} (Pool et al., 2018) is also based on the FDC:

$$\alpha_{\text{NP}} = 1 - \frac{1}{2}\sum_{k=1}^{n}\left|\frac{x_{\text{s}}(I(k))}{n\mu_{\text{s}}} - \frac{x_{\text{o}}(J(k))}{n\mu_{\text{o}}}\right|, \tag{9}$$

where $I(k)$ and $J(k)$ stand for the time steps of the kth largest discharge for the simulated and observed time series, respectively.

As β, β_{n}, and $\overline{B}_{\text{rel}}$ all represent the bias, they are therefore designed as bias parameters in this study.

2.2 Score calculation

A total of nine performance criteria are analysed in this study: the NSE, KGE, 2012 version of the KGE or modified KGE (KGE′), 2021 version of the KGE (KGE″), non-parametric KGE (KGE$_{\text{NP}}$), diagnostic efficiency (DE), Liu mean efficiency (LME), least-squares combined efficiency (LCE), and d_1. The value considered to be the best score is equal to 1 for all criteria, except for the DE, for which it is equal to zero.

The NSE (Nash and Sutcliffe, 1970) is a normalised variant of the mean squared error (MSE) and compares a prediction to the observed mean of the target variable:

$$\text{NSE} = 1 - \frac{\sum (x_{\text{s}}(t) - x_{\text{o}}(t))^2}{\sum (x_{\text{o}}(t) - \mu_{\text{o}})^2}. \tag{10}$$

Gupta et al. (2009) algebraically decomposed the NSE into correlation, variability, and bias components:

$$\text{NSE} = 2\alpha r - \alpha^2 - \beta_{\text{n}}^2. \tag{11}$$

The KGE was proposed by Gupta et al. (2009) as an alternative to the NSE. The optimal KGE corresponds to the closest point of the three-dimensional Pareto front – of α, β, and r – to the ideal value of [1; 1; 1]:

$$\text{KGE} = 1 - \sqrt{(\alpha - 1)^2 + (\beta - 1)^2 + (r - 1)^2}. \tag{12}$$

A modified Kling–Gupta efficiency was proposed by Kling et al. (2012). The coefficient of variation is used instead of the standard deviation to ensure that bias and variability are not cross-correlated:

$$\text{KGE}' = 1 - \sqrt{(\gamma - 1)^2 + (\beta - 1)^2 + (r - 1)^2}. \tag{13}$$

Tang et al. (2021) proposed another variant (KGE″) by using the normalised bias instead of β to ensure that the score is not overly sensitive to mean values – μ_{o} or μ_{s} – that are close to zero (Santos et al., 2018; Tang et al., 2021):

$$\text{KGE}'' = 1 - \sqrt{(\alpha - 1)^2 + \beta_{\text{n}}^2 + (r - 1)^2}. \tag{14}$$

Pool et al. (2018) cautioned against the implicit assumptions of the KGE – data linearity, data normality, and absence of outliers – and proposed a non-parametric alternative (KGE$_{\text{NP}}$) for limiting their impact. The non-parametric form of the variability is calculated using the FDC, and the Spearman rank correlation coefficient is used instead of the Pearson correlation coefficient:

$$\text{KGE}_{\text{NP}} = 1 - \sqrt{(\alpha_{\text{NP}} - 1)^2 + (\beta - 1)^2 + (r_{\text{S}} - 1)^2}. \tag{15}$$

In a similar way, Schwemmle et al. (2021) used FDC-based parameters to account for variability and bias in another KGE variant: the diagnostic efficiency. This criterion is based on constant, dynamic, and timing errors and aims to provide a stronger link to hydrological processes (Schwemmle et al., 2021):

$$\text{DE} = \sqrt{\overline{B}_{\text{rel}}^2 + |B_{\text{area}}|^2 + (r - 1)^2}. \tag{16}$$

In this study, we used a normalised diagnostic efficiency (DE′) so that the best error score is equal to 1 for facilitating the comparison with other performance criteria:

$$\text{DE}' = 1 - \sqrt{\overline{B}_{\text{rel}}^2 + |B_{\text{area}}|^2 + (r - 1)^2}. \tag{17}$$

Liu (2020) proposed another alternative, the Liu mean efficiency, to improve the simulation of extreme events. The LME thus aims to address the underestimation of variability of the KGE, which is still a concern despite not being as severe as with the NSE (Gupta et al., 2009; Mizukami et al., 2019):

$$\text{LME} = 1 - \sqrt{(r\alpha - 1)^2 + (\beta - 1)^2}. \tag{18}$$

Lee and Choi (2022) proposed the least-squares combined efficiency to address the shortcomings of the LME identified by Choi (2022), namely (i) an infinite number of solutions for the maximum score and (ii) an inclination to overestimate high flows and underestimate low flows. The LCE is based on the least-squares statistics combined from both-way regression lines $r\alpha$ and r/α:

$$\text{LCE} = 1 - \sqrt{(r\alpha - 1)^2 + (r/\alpha - 1)^2 + (\beta - 1)^2}. \tag{19}$$

Willmott et al. (1985) proposed a modified index of agreement, which aim to address the issues associated with r and the coefficient of determination, as well as the sensitivity of the original index of agreement to outliers (Legates and McCabe, 1999):

$$d_1 = 1 - \frac{\sum |x_\mathrm{s}(t) - x_\mathrm{o}(t)|}{\sum (|x_\mathrm{s}(t) - \mu_\mathrm{o}| + |x_\mathrm{o}(t) - \mu_\mathrm{o}|)}. \tag{20}$$

3 Synthetic time series

3.1 Generating synthetic time series with homothetic transformations

A simulation performance can be assessed in terms of bias, variability, and timing errors (Gupta et al., 2009). Bias and variability errors correspond to a difference in the volume and amplitude of discharges. Timing errors correspond to a shift in time. We created a synthetic hydrograph corresponding to one flood event as the reference (observed) time series. We also generated synthetic transformations – of the reference time series – with different errors in terms of bias and variability corresponding to the time series simulated by a model. We did not consider any timing errors as our aim is to assess counterbalancing errors induced by bias and variability parameters. Synthetic transformations were generated by multiplying the reference time series by the coefficient ω:

$$Q_\mathrm{s}(t) = Q_\mathrm{o}(t) \cdot \omega, \tag{21}$$

where $Q_\mathrm{s}(t)$ stands for the transformed discharge at the time t, $Q_\mathrm{o}(t)$ stands for the reference discharge at the time t, and ω stands for the coefficient. ω values were sampled uniformly on the log-transformed interval $[-0.36, 0.36]$ at a defined step of 0.002 to ensure a fair distribution between underestimated and overestimated transformations. The exponentiation in base 10 of the sampled values results in 361 ω values evenly distributed around the $\omega = 1$ homothety, which corresponds to the reference time series (i.e. absence of transformation). We defined ω bounds such that the transformed peak discharge roughly ranges from half ($\omega \approx 0.437 \approx 10^{-0.36}$) to twice ($\omega \approx 2.291 \approx 10^{0.36}$) that of the reference time series. Note that (i) the data linearity between simulated and observed values is verified and (ii) ω homotheties still induce small timing errors – which were considered to be negligible – because the correlation coefficients (r and r_s) also slightly account for the shape of the transformation.

To study counterbalancing errors induced by bias and variability parameters, we generated time series that consist of two successive flood events and considered all possible combinations of the 361 transformations for the simulated time series (Fig. 1). This results in a total of $361^2 = 130\,321$ transformations with two flood events, including (i) a perfect transformation with $\omega = 1$ for both flood events, (ii) bad–good (BG) or good–bad (GB) transformations when $\omega = 1$ for only one out of the two flood events, and (iii) bad–bad (BB) transformations when $\omega \neq 1$ for both flood events. The performance of the transformations with regards to the reference time series were evaluated using the nine performance criteria presented in Sect. 2.

3.2 Identifying counterbalancing errors on a straightforward example

Figure 2 presents the following two hydrographs extracted from the set of transformations: (i) a BB model with the combination $[\omega_1 = 0.75; \omega_2 = 1.2]$ and (ii) a BG model with the combination $[\omega_1 = 0.75; \omega_2 = 1]$. The BG model stands out as a better model because it perfectly reproduces the second flood event and is identical to the BB model during the first flood ($\omega_1 = 0.75$). Nevertheless, the KGE and its variants – KGE′, KGE″, $\mathrm{KGE_{NP}}$, DE′, LME, and LCE – all favour the BB model, whereas only the NSE and d_1 evaluate the BG model as better (Fig. 3a). Further results for common and recently developed performance criteria are presented in Fig. A1 in the Appendix.

The investigation of the components of the criteria (Fig. 3b) reveals how a seemingly better model (i.e. the BG model) can have a lower score than expected. Bias parameters are systematically better for the BB model, with 0.98 over 0.88 for β, -0.02 over -0.08 for β_n, and -0.04 over -0.12 for $\overline{B}_\mathrm{rel}$. Timing parameters are systematically better for the BG model, with 0.99 over 0.96 for r and 0.99 over 0.98 for r_s. Variability parameters are mixed: (i) α favours the BB model with 1.01 over 0.89, (ii) γ favours the BG model with 1.01 over 1.04, (iii) α_NP slightly favours the BG model with 0.94 over 0.93, and (iv) $|B_\mathrm{area}|$ is equal for both models. The $r\alpha$ and r/α parameters are better for the BB model; $2\alpha r$ is better for the BG model.

The β, β_n, $\overline{B}_\mathrm{rel}$, α, $r\alpha$, and r/α parameters all provide a better evaluation of bias and variability for the BB model. Concurrent over- and underestimation of discharges over the time series result in a good water balance: close to 1 for β and $\overline{B}_\mathrm{rel}$ and 0 for β_n. Depending on the criterion, the variability parameter can also affect the score in a similar counter-intuitive manner. α is heavily impacted by the counterbalance, whereas it seems to be mitigated for γ, α_NP, and $|B_\mathrm{area}|$. The timing parameters (r and r_s) have an expected score that favours the BG model. However, the score difference in terms of timing errors between BB and BG models is very small (0.03 at best for r). The impact on the overall score is thus minimised compared to the one induced by bias and variability parameters, which can be cumulated (e.g. both β and α counterbalancing errors in the KGE) or have a larger difference – up to 0.12 for α. Counterbalancing errors can thus result in better values for bias and variability, which increase the overall score. In this case, the highest score may not be the most appropriate indicator of model relevance.

The largest differences in score appear for the LME and LCE criteria as all their parameters are affected by counter-

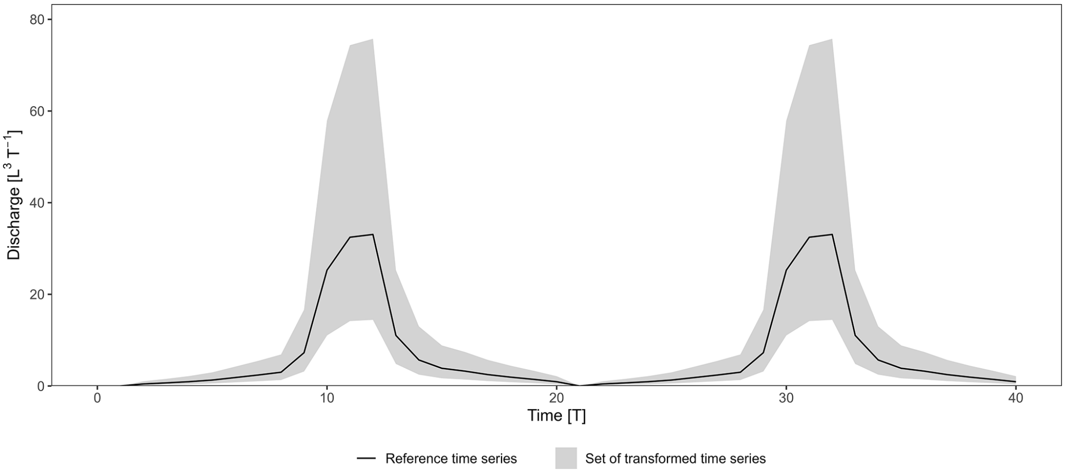

Figure 1. Synthetic hydrograph corresponding to two flood events.

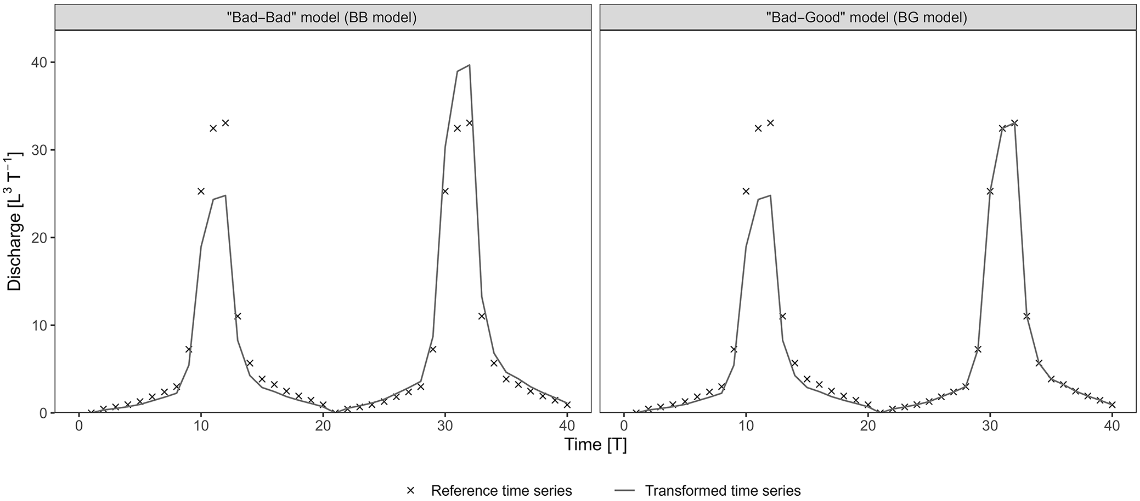

Figure 2. Synthetic examples extracted from the set of transformations. The first and second flood events of the bad–bad and bad–good transformations were shifted with [$\omega_1 = 0.75$; $\omega_2 = 1.2$] and [$\omega_1 = 0.75$; $\omega_2 = 1$] combinations, respectively.

balancing errors (β, $r\alpha$, and r/α). The KGE and KGE$''$ also show significant differences as they accumulate the counterbalancing errors of α and β. The KGE$'$ demonstrates a smaller difference than the KGE due to the use of γ. Both FDC-based criteria KGE$_{\text{NP}}$ and DE$'$ show the smallest differences due to α_{NP} and $|B_{\text{area}}|$, which have a nearly equal value for both BB and BG models. The NSE has a slightly better score on the BG model, while the difference is more pronounced for d_1.

This example demonstrates how relative error metrics can cancel each other out and affect the design and the evaluation of hydrological models. The counterbalancing errors especially affect bias parameters (β, β_{n}, and $\overline{B}_{\text{rel}}$) but also the variability parameter α.

3.3 Exploring counterbalancing errors with synthetic transformations

Figure 4 shows the score distribution of the synthetic set of hydrographs presented in Sect. 3.1. For each value of ω_1, the minimum and maximum criteria scores of the transformations resulting from all combinations with ω_2 provide the dashed envelope of the score distribution, with the maximum transformation score at the top (1 corresponding to a perfect model) and the worst at the bottom. The transformations corresponding to the BG models (with $\omega_2 = 1$) are represented by the black line. All transformations included in the dashed envelope can be identified as bad–bad models, except when $\omega_1 = 1$ or $\omega_2 = 1$ (black line).

It is obvious that the KGE and its variants – KGE$'$, KGE$''$, KGE$_{\text{NP}}$, DE$'$, LME, and LCE – always evaluate one or several BB models as being better than the BG model for the

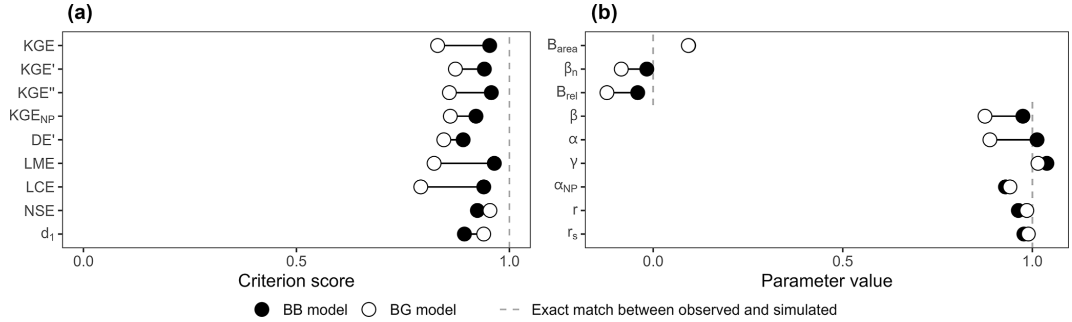

Figure 3. **(a)** Score of the BB and BG transformations according to the different performance criteria. **(b)** Values of the parameters used in the calculation of the performance criteria.

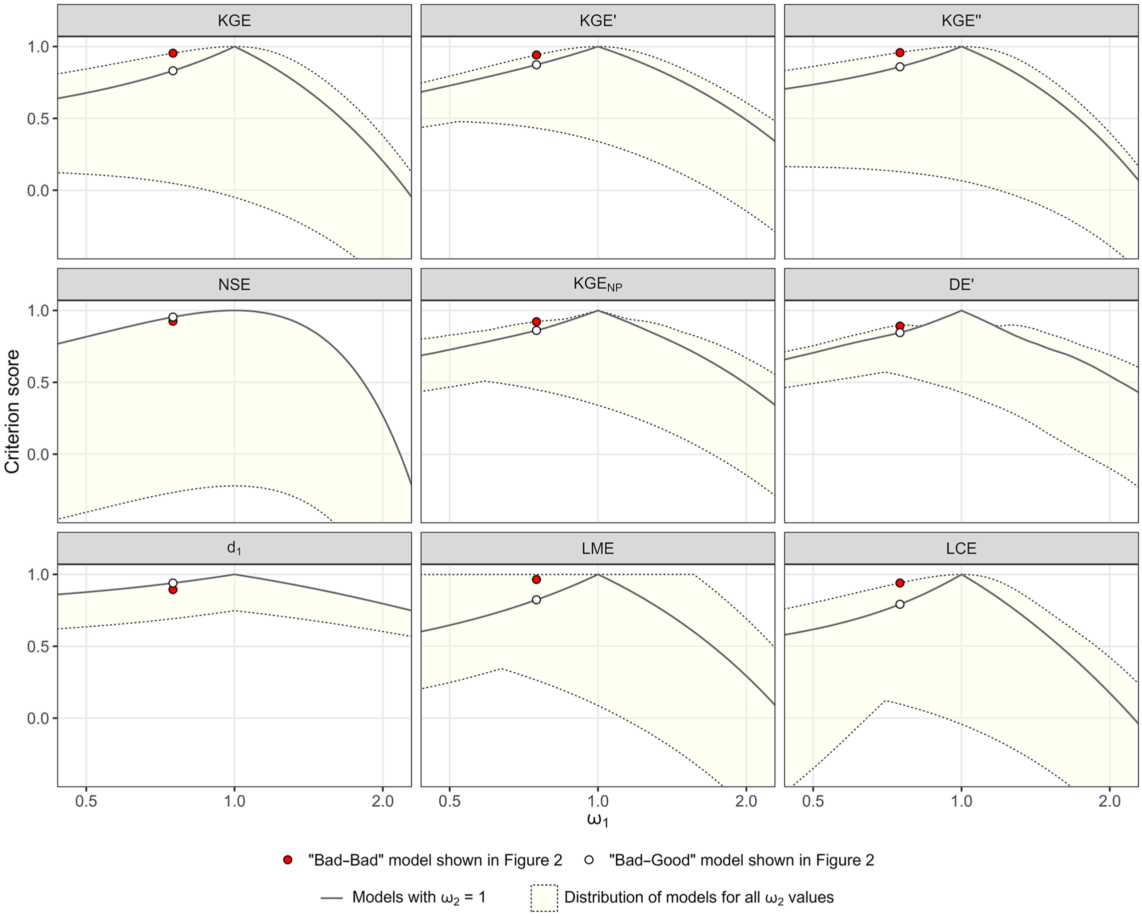

Figure 4. Score of each transformation for all $[\omega_1; \omega_2]$ combinations by performance criteria.

same ω_1 value, except for $\omega_1 = 1$. On the other hand, the NSE and d_1 correctly identify the BG model as the best transformation for all combinations of $[\omega_1; \omega_2]$; i.e. the black line is always above the dashed envelope. The envelopes of the KGE, KGE$'$, and KGE$''$ criteria are similar, but they do not display the same difference between the best scores and the scores of the BG models. These differences are smaller for the latter two because the KGE$'$ is based on γ instead of α, and the KGE$''$ is based on β_n instead of β, for which it is demonstrated in Sect. 3.2 that they both soften counterbalancing errors. The envelope of the LCE criterion looks like that of the KGE. However, the difference between the best scores and the scores of the BG models is much higher. This is likely due to the nature of the equation consisting of three parameters affected by counterbalancing errors (β, $r\alpha$, and r/α). The LME criterion has a very distinctive envelope, for which the maximum score of 1 is reached for a lot of BB models, even when both ω_1 and ω_2 are different from 1. This can be explained by the interaction between r and α that leads to an infinite number of solutions (Choi, 2022). The KGE$_{\mathrm{NP}}$ and DE$'$ (FDC-based criteria) both show similar envelopes with a break point near the maximum transformation score in both directions around $\omega_1 = 1$. This is especially pronounced for the DE$'$, for which the BG model is nearly the best model between $\omega_1 = 0.83$ and $\omega_1 = 1.17$. These results show that counterbalancing errors can happen on a large range of parameters, and when using the KGE or its variants, there is a possibility for the more meaningful model (i.e. BG model) to have a lower score than a compensated or bad–bad model.

Figure 5 shows the value of ω_2 corresponding to the best evaluation for a given ω_1 by performance criteria. As identified above, the NSE and d_1 both evaluate the BG models as the best transformations (NSE and d_1 black lines coincide at $\omega_2 = 1$, Fig. 5). Counterbalancing errors are apparent for the KGE and its variants. For $\omega_1 \neq 1$, the best transformations are always BB models and follow two conditions: (i) if $\omega_1 < 1$ then $\omega_2 > 1$, and (ii) if $\omega_1 > 1$ then $\omega_2 < 1$. This means that, in this case, such performance criteria will always be flawed towards concurrent under- and overestimation of discharges in a transformation.

4 Real case study

To highlight how counterbalancing errors can affect the assessment of hydrological models in a real case study, we used two different modelling approaches: artificial neural networks (ANNs) and bucket-type models. The simulations of the karst spring discharges of both models were evaluated for the same 1-year validation period. To clearly highlight the problem, we deliberately chose a bucket-type simulation that is noticeably affected by counterbalancing errors yet is still realistic. Further information on the modelling approaches, the input data, the calibration strategy, and the simulation procedure can be found in Cinkus et al. (2023).

4.1 Study site

The Unica springs are the outlet of a complex karstic system influenced by a network of poljes. The recharge area is about 820 km^2 and is located in a moderate continental climate with a strong snow influence. Recharge comes from both (i) allogenic infiltration from two sub-basins drained by sinking rivers and (ii) autogenic infiltration through a highly karstified limestone plateau (Gabrovšek et al., 2010; Kovačič, 2010; Petric, 2010). The network of connected poljes constitutes a common hydrological entity that induces a high hydrological variability in the system and long and delayed high discharges at the Unica springs (Mayaud et al., 2019). The limestone massif can reach a height of 1800 m a.s.l. (above sea level) and has significant groundwater resources (Ravbar et al., 2012). A polje downstream of the springs can flood when the Unica discharge exceeds 60 m^3 s^{-1} for several days. If the flow reaches 80 m^3 s^{-1}, the flooding can reach the gauging station and influence its measurement. The flow data are from the gauging station in Unica-Hasberg (ARSO, 2021a). Precipitation, height of snow cover, and height of new snow data are from the meteorological stations in Postojna and Cerknica (ARSO, 2021b). Temperature and relative humidity data are from the Postojna station. Potential evapotranspiration is calculated from the Postojna station data with the Penman–Monteith formula (Allen et al., 1998).

4.2 Modelling approaches

The first modelling approach is based on convolutional neural networks (CNNs) (LeCun et al., 2015), which is a specific type of ANN that is powerful in processing image-like data but also very useful for processing sequential data. The model consists of a single 1D convolutional layer with a fixed kernel size of 3 and an optimised number of filters. This layer was complemented by a MAX-POOLING LAYER, a Monte Carlo dropout layer with 10 % dropout rate and two dense layers. The first dense layer had an optimised number of neurons, and the second layer had a single output neuron. We programmed our models in Python 3.8 (van Rossum, 1995) using the following frameworks and libraries: Bayesian optimisation (Nogueira, 2014), Matplotlib (Hunter, 2007), NumPy (van der Walt et al., 2011), Pandas (Reback et al., 2021; McKinney, 2010), scikit-learn (Pedregosa et al., 2018), TensorFlow 2.7 (Abadi et al., 2016), and Keras API (Chollet, 2015).

The second modelling approach is a bucket-type model, which is a conceptual representation of a hydrosystem consisting of several buckets that are supposed to be representative of the main processes involved. We used the adjustable modelling platform KarstMod (Mazzilli et al., 2019). The model structure consists of one upper bucket for simulating soil and epikarst processes (including a soil available water capacity) and two lower buckets corresponding to matrix and conduits compartments. A very reactive transfer function from the upper bucket to the spring is used to reproduce very fast flows occurring in the system.

4.3 Impact of counterbalancing errors on model evaluation

Figure 6a shows the results of the two hydrological models for Unica springs. The models have overall good dynamics and successfully reproduce the observed discharges. Regarding high-flow periods, both models show a small timing error, inducing a delay in the simulated peak flood. The first flood event (February 2017) is slightly underestimated by the ANN model and highly overestimated by the bucket-type model. The second flood event (March 2017) is similarly underes-

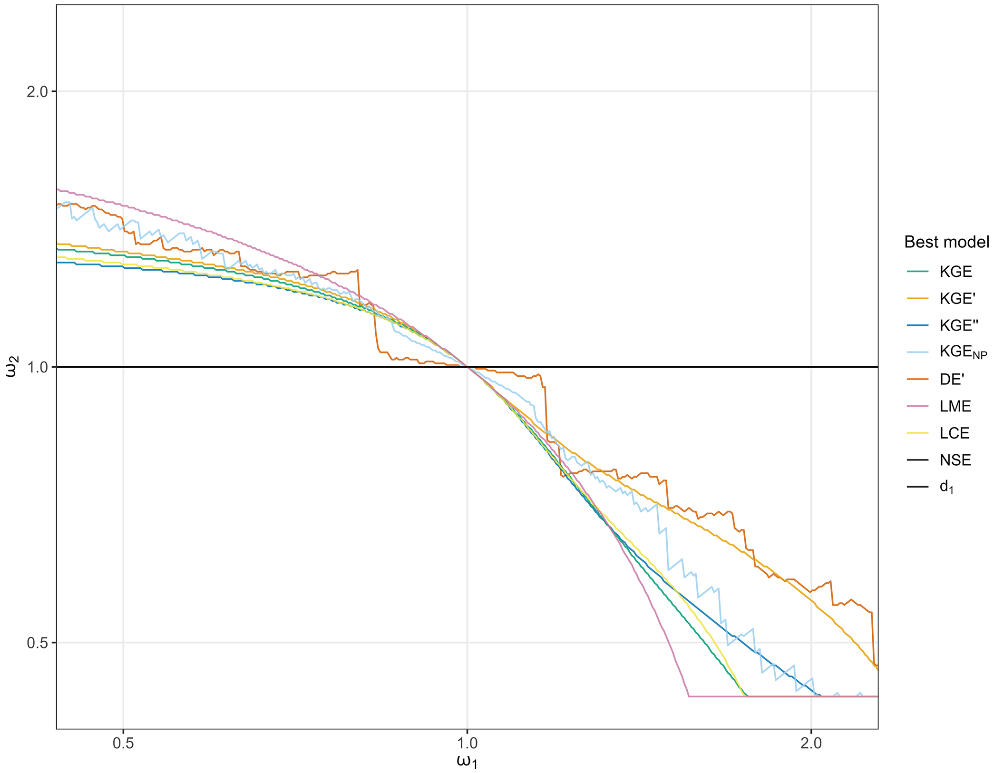

Figure 5. Graph of each [ω_1; ω_2] combination identified as the best transformation by each performance criteria. The NSE and d_1 black lines coincide at $\omega_2 = 1$.

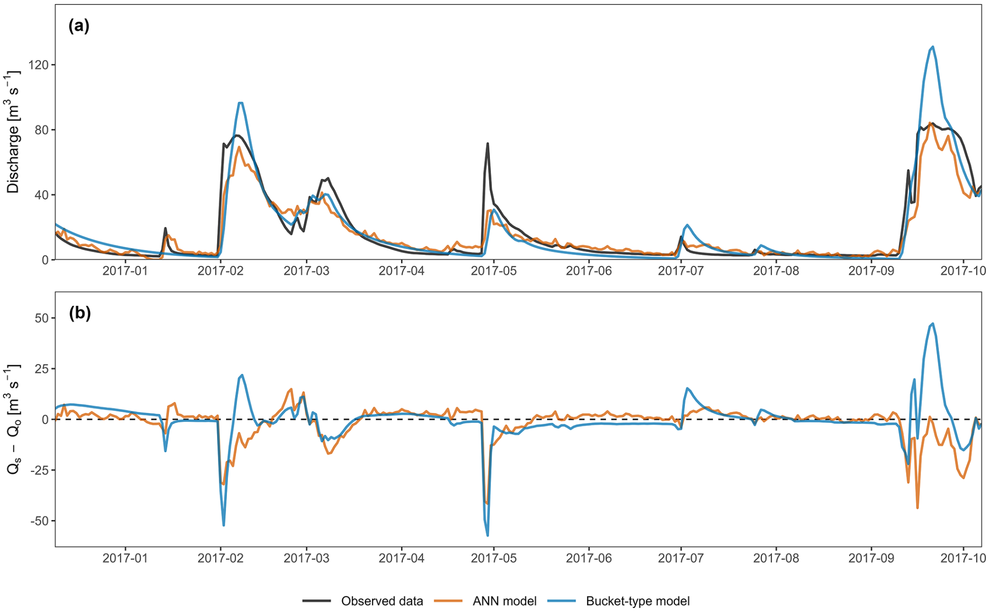

Figure 6. **(a)** Observed and simulated spring discharge time series for the validation period. **(b)** Relative difference between simulated and observed discharge for the validation period.

timated by both models, but the bucket-type model demonstrates a slightly better performance. The third flood event (May 2017) is poorly simulated by the models, with both underestimating the flood peak, but the ANN model is more accurate in terms of timing and volume estimate, while the bucket-type model has a better recession coefficient and flow variability. The last flood event (September 2017) comprise a small peak followed by a very high and long-lasting flood. Both models fail to account for the small peak. The following important flood event is highly overestimated by the bucket-type model while being nicely simulated by the ANN model despite the small underestimation and timing error. The small flood events are better simulated by the ANN model than the bucket-type model: (i) the ANN model simulates them satisfactorily, except for the second one (mid-April), where the simulated discharges are overestimated; (ii) the bucket-type model does not simulate the first two events at all (mid-January and mid-April) and largely overestimates the last two (early and late June) in addition to timing errors. Both models can be improved during recession and low-flow periods. The ANN model is rather close to the observed discharges but seems to be too sensitive to precipitation (continuous oscillations). On the other hand, the bucket-type model shows no oscillations but either overestimates or underestimates the observed discharges. Some events are not well simulated by both models (e.g. the May 2017 flood), which may be due to uncertainties in the input data. Also, the data linearity between simulated and observed values is slightly skewed for both models, which can affect the relevance of r (Barber et al., 2020).

In general, the ANN model can be described as better because it is closer to the observed values in the high- and low-flow periods. While this statement cannot be supported by performance metrics, we believe that an expert assessment based on intuition and experience is still valuable despite being intrinsically subjective. In this particular case, one can assess the main, distinctive flaws of each model: (i) the ANN model has continuous oscillations – especially during recession and low-flow periods – and lacks accuracy during recession periods; (ii) the bucket-type model highly overestimates several flood events and is inaccurate during a lot of recession and low-flow periods. Figure 6b also shows that the bucket-type model has an overall higher bias than the ANN model. Hydrological models are generally used for (i) the prediction or forecast of water flooding or inrush, (ii) the management of water resources, (iii) the characterisation of hydrosystems, and more recently (iv) the study of the impact of climate change on water resources. Most studies thus put the emphasis on volumes and also on extremes events (i.e. dry and flood periods), which in this case are more satisfactorily reproduced by the ANN model in terms of volume estimate, timing, and variability.

This visual assessment is confirmed by only a few performance criteria: the NSE, d_1, and KGE$_{\mathrm{NP}}$ (Fig. 7a). These criteria evaluate the ANN model as better, although the performances of both models are quite close for the d_1. However, the KGE and most of its variants (except the KGE$_{\mathrm{NP}}$) all favour the bucket-type model over the ANN model – sometimes by a large margin. Further results for common and recently developed performance criteria are presented in Fig. A2 in the Appendix. It is interesting to note how similar these results are to those of the synthetic example (Figs. 3a and A1 in the Appendix). Looking at the values of the equations' parameters (Fig. 7b), we find that bias parameters are systematically better for the bucket-type model, with 1 over 0.92 for β, 0 over -0.06 for β_{n}, and -0.07 over 0.18 for $\overline{B}_{\mathrm{rel}}$. Timing errors are systematically better for the ANN model, with 0.95 over 0.92 for r and 0.94 over 0.83 for r_{s}. Variability parameters favour the bucket-type model with 1.1 over 0.78 for α, 1.1 over 0.85 for γ, 0.22 over 0.3 for $|B_{\mathrm{area}}|$, and a very close value that is better by 0.005 for the α_{NP} parameter. In summary, all bias and variability parameters have better values for the bucket-type model, while timing and shape parameters are better for the ANN model.

As the KGE and its variants are generally composed of equally weighted bias, variability, and timing, their overall score is heavily affected by compensation effects – except in the case of a large error for one parameter. In our case, all parameters have similar errors, which results in a better KGE for the bucket-type model compared to the ANN model. This applies to all the KGE variants except the KGE$_{\mathrm{NP}}$, where the error for r_{s} is significant, resulting in a better score for the ANN model. The LME score is extremely high (0.99) for the bucket-type model, which is probably due to the compensation of r and α identified by Choi (2022). Also, using γ instead of α for assessing the variability seems to lower counterbalancing errors.

Interestingly, the cumulative sum of the absolute bias error between simulated and observed values (Fig. 6b) is smaller for the ANN model (1394 m^3) than the bucket-type model (1611 m^3), but still the relative bias and variability parameters are better for the bucket-type model. This observation highlights how counterbalancing errors can impair the evaluation of hydrological models: seemingly better parameters values (bias and variability) that increase criteria scores are not necessarily associated with an increase in model relevance.

5 Recommendations

The aim of this paper is primarily to raise awareness among modellers. Performance criteria generally aggregate several aspects of the characteristics of a model into a single value, which can lead to an inaccurate assessment of said aspects. Ultimately, all criteria have their flaws and should be carefully selected with regards to the aim of the model.

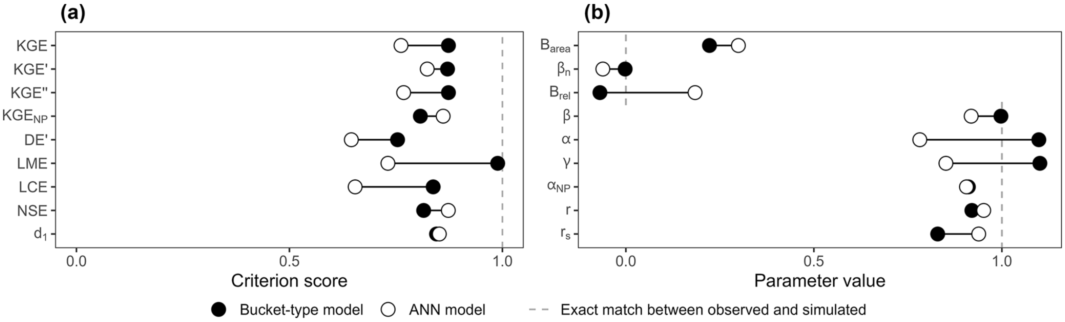

Figure 7. (a) Score of the ANN and bucket-type models according to the different performance criteria. **(b)** Values of the parameters used in the calculation of the performance criteria.

5.1 Use of relevant performance criteria

Table 1 summarises the presence and impact of counterbalancing errors, as well as the advantages and drawbacks (as reported in other studies) of the different performance criteria. The recommendations for counterbalancing errors are based on the results of this research – i.e. synthetic and real case studies. The KGE and all its variants are affected by counterbalancing errors with varying degrees of intensity: (i) mildly impacted (+) for the KGE′, KGE$_{\mathrm{NP}}$, and DE; (ii) moderately impacted (++) for the KGE, KGE″, and LCE; and (iii) strongly impacted (+ + +) for the LME. In this study, the NSE and d_1 stand out as clearly better since they have no counterbalancing errors. However, they have other drawbacks that are not associated with counterbalancing errors, especially the NSE, with its limitations related to variability (Gupta et al., 2009). We thus recommend using performance criteria that are not or less prone to counterbalancing errors (d_1, KGE′, KGE$_{\mathrm{NP}}$, DE).

5.2 Use of scaling factors

The assessment of the hydrological models in the real case study shows how concurrent over- and underestimation can generate counterbalancing errors in bias and variability parameters. For the case study considered in this paper, the ANN model, although offering a better simulation, is evaluated as – sometimes considerably – worse than the bucket-type model because it slightly underestimates the total volume. This has a great impact on the overall score, as the KGE and its variant are calculated with both bias and variability parameters accounting for two-thirds of the overall criterion score.

While the overall balance (bias) may be a desired feature in a model, we showed that a good value may be accidental and result from counterbalancing errors. The common use of the KGE neglects one of the original proposals, which is to weight the parameters β, α, and r in the equation. Gupta et al. (2009) proposed an alternative equation for adjusting the emphasis on the different aspects of a model:

$$\mathrm{KGE_s} = 1 - \sqrt{[s_\alpha(\alpha-1)]^2 + \left[s_\beta(\beta-1)\right]^2 + [s_\mathrm{r}(r-1)]^2}, \quad (22)$$

with s_r, s_β, and s_α being the scaling factors of r, β, and α, respectively. By default, these factors are equal to 1, which induces a weight of one-third for the parameter in absolute value (r) and two-thirds for the parameters in relative values (β, α). To the best of our knowledge, only Mizukami et al. (2019) ever considered changing the scaling factors when using the KGE. We suggest that one carefully consider such scaling factors for the calibration and the evaluation of hydrological models using the KGE and its variants. Depending on the purpose of the model, they can help to emphasise particular aspects of a model or reduce the influence of relative parameters and counterbalancing errors.

Figure 8 shows how emphasising absolute parameters with scaling factors helps to reduce the influence of counterbalancing errors for the KGE (Fig. 8a) and its most used variant KGE′ (Fig. 8b). The default value (1–1–1) – corresponding to scaling factors of 1 for α (KGE) or γ (KGE′), 1 for β, and 1 for r, respectively – is compared to other factor combinations with different ratios between absolute and relative parameters. The 2 : 1 ratio (2–2–1) increases counterbalancing errors as the emphasis is on the relative parameters, while the 1 : 2, 1 : 3, 1 : 4, and 1 : 5 ratios decrease counterbalancing errors. The ANN model is evaluated as better with the 1 : 4 ratio for the KGE and the 1 : 3 ratio for the KGE′, highlighting that the KGE′ is less sensitive to counterbalancing errors. This also shows how the score of a performance criterion and, by extension, its interpretation can be radically different depending on the parameters used in the equation. This is why a multi-criteria framework can strengthen the evaluation of models and reduce the uncertainty associated with the interpretation of individual performance criteria scores.

6 Conclusion

This study sets out to explore the influence of counterbalancing errors and to raise awareness among modellers about the

Table 1. Presence and impact of counterbalancing errors (CEs) on the assessment of model performance according to different performance criteria. The impact of CEs is denoted as null (–), mild (+), moderate (++), or strong (+ + +).

Criterion	Year	Affected by CE	Impact of CE	Advantages	Drawbacks*
KGE	2009	Yes	++	Variability is not underestimated (Gupta et al., 2009)	Still a slight underestimation of high discharges (Gupta et al., 2009) Bias and variability are cross-correlated (Kling et al., 2012) Implicit assumptions of data linearity, data normality, and absence of outliers (Pool et al., 2018) No inherent benchmark (Knoben et al., 2019) Not suited to logarithmic transformation of discharge (Santos et al., 2018)
KGE′	2012	Yes	+	Bias and variability are not cross-correlated (Kling et al., 2012)	
KGE″	2021	Yes	++	The score is not overly sensitive to mean values close to zero (Santos et al., 2018; Tang et al., 2021)	
KGE$_{\text{NP}}$	2018	Yes	+	Reduces the impact of implicit assumptions of data linearity, data normality, and absence of outliers by using non-parametric parameters (Pool et al., 2018)	
DE	2021	Yes	+	Aims to provide a stronger link to hydrological processes (Schwemmle et al., 2021)	
LME	2020	Yes	+ + +	Improves the simulation of extreme events (Liu, 2020)	Infinite number of solutions for the maximum score (Lee and Choi, 2022) Inclination to overestimate high flows and underestimate low flows (Lee and Choi, 2022)
LCE	2022	Yes	++	Improve the simulation of extreme events (Lee and Choi, 2022)	
NSE	1970	No	–		The contribution of β_{n} depends on the variability (Gupta et al., 2009) Variability is underestimated (Gupta et al., 2009) The benchmark is inappropriate for highly variable discharges (Gupta et al., 2009)
d_1	2012	No	–	Addresses the shortcomings of r and the coefficient of determination (Willmott, 1981) The score is less sensitive to errors concentrated in outliers in comparison to the original index of agreement (Willmott et al., 1985)	

* KGE drawbacks may likely apply to KGE variants, but this has not been studied extensively.

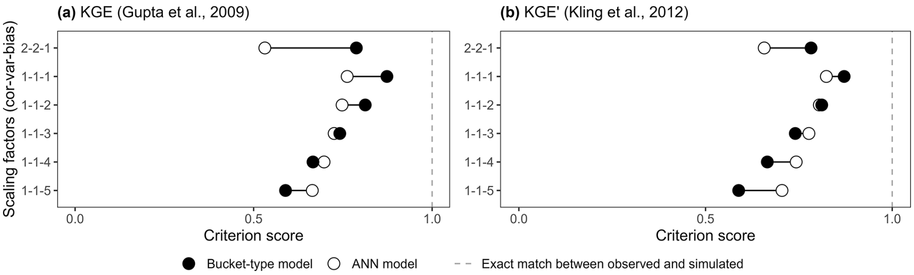

Figure 8. **(a)** KGE and **(b)** KGE′ scores of the ANN and bucket-type models (Fig. 6a) according to different scaling factors. The y-axis numbers correspond to the scaling factors of the variability, bias, and timing parameters, with the default being 1–1–1.

use of performance criteria for calibrating and evaluating hydrological models. A total of nine performance criteria (NSE, KGE, KGE′, KGE″, KGE$_{\mathrm{NP}}$, DE, LME, LCE, and d_1) are analysed. The investigation of synthetic time series and real hydrological models shows that concurrent over- and underestimation of multiple parts of a discharge time series may favour bias and variability parameters. This especially concerns the bias parameters (β, β_{n}, and B_{rel}) as their values are all influenced by counterbalancing errors in both synthetic time series and the real case study. On the other hand, the impact of counterbalancing errors on the variability parameters seems to depend on the time series: only the value of α is influenced in the synthetic time series, while the values of all variability parameters (α, γ, $|B_{\mathrm{area}}|$, and α_{NP}) are influenced in the real hydrological models. As bias and variability parameters generally account for two-thirds of the weight in the equation of certain performance criteria, this can lead to an overall higher criterion score without being associated with an increase in model relevance. This is especially concerning for the KGE and its variants as they generally use relative parameters for evaluating bias and variability in hydrological models. These findings highlight the importance of carefully choosing a performance criterion adapted to the purpose of the model. Recommendations also include the use of scaling factors to emphasise different aspects of a hydrological model and to reduce the influence of relative parameters on the overall score of the performance criterion. Further research could explore the appropriate values of scaling factors to be used, depending on the modelling approach and the purpose of the study.

Appendix A: Common and recently developed performance criteria applied to the synthetic time series and the real case study

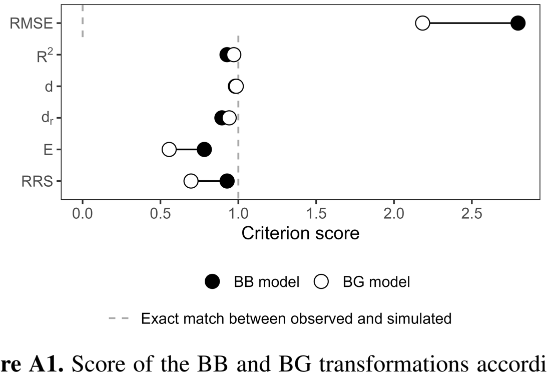

Figure A1. Score of the BB and BG transformations according to other common and recently developed performance criteria: the root mean square error (RMSE), the coefficient of determination R^2, the index of agreement d (Willmott, 1981), the refined index of agreement d_{r} (Willmott et al., 2012), the Onyutha efficiency E, and the revised R^2 (RRS) (Onyutha, 2022).

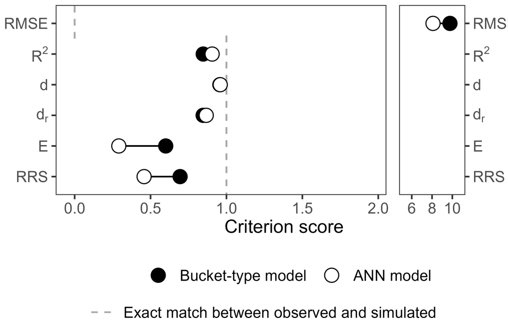

Figure A2. Score of the ANN and bucket-type models according to other common and recently developed performance criteria: the root mean square error (RMSE), the coefficient of determination R^2, the index of agreement d_{r} (Willmott, 1981), the refined index of agreement d_{r} (Willmott et al., 2012), the Onyutha efficiency E, and the RRS (Onyutha, 2022).

Code and data availability. We provide complete scripts for reproducing the results for the synthetic time series (Sect. 3), as well as the ANN model code and KarstMod *.properties* file (bucket-type model), on Zenodo (https://doi.org/10.5281/zenodo.7274031, Cinkus and Wunsch, 2022). Unica spring discharge time series and meteorological data are available from the Slovenian Environment Agency (http://vode.arso.gov.si/hidarhiv/, ARSO, 2021a; http://www.meteo.si/, ARSO, 2021b).

Author contributions. GC, NM, and HJ conceptualised the study and designed the methodology. GC and AW developed the software code. GC performed the experiments and investigated and visualised the results. AW provided the ANN results for the case study. GC wrote the original paper draft with contributions from AW and NR. All the authors contributed to the interpretation of the results and the review and editing of the paper draft. NM and HJ supervised the work.

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

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. For the data provided, we thank the Slovenian Environment Agency (ARSO, 2021a, b). The analyses were performed using R (R Core Team, 2021) and the following packages: readxl, readr, dplyr, tidyr, ggplot2, lubridate (Wickham et al., 2019), cowplot (Wilke, 2020), diageff (Schwemmle et al., 2021), flextable (Gohel, 2021), hydroGOF (Zambrano-Bigiarini, 2020), HydroErr (Roberts et al., 2018), and padr (Thoen, 2021). The paper was written with the Rmarkdown framework (Allaire et al., 2021; Xie et al., 2018, 2020).

Financial support. This research has been supported by the French Ministry of Higher Education and Research for the thesis scholarship of Guillaume Cinkus and the European Commission for its support through the Partnership for Research and Innovation in the Mediterranean Area (PRIMA) programme under the Horizon 2020 (KARMA project, grant agreement no. 01DH19022A). We were supported by the Slovenian Research Agency within the project "Infiltration processes in forested karst aquifers under changing environment" (grant no. J2-1743).

Review statement. This paper was edited by Yi He and reviewed by two anonymous referees.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, arXiv [preprint], https://doi.org/10.48550/arXiv.1603.04467, 2016.

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., and Iannone, R.: Rmarkdown: Dynamic documents for r, https://cran.r-project.org/package=rmarkdown (last access: 27 June 2023), 2021.

Allen, R. G., Pereira, L. S., Raes, D., Smith, M., and FAO (Eds.): Crop evapotranspiration: Guidelines for computing crop water requirements, Food and Agriculture Organization of the United Nations, Rome, https://appgeodb.nancy.inra.fr/biljou/pdf/Allen_FAO1998.pdf (last access: 27 June 2023), 1998.

Althoff, D. and Rodrigues, L. N.: Goodness-of-fit criteria for hydrological models: Model calibration and performance assessment, J. Hydrol., 600, 126674, https://doi.org/10.1016/j.jhydrol.2021.126674, 2021.

ARSO: Ministry of the Environment and Spatial Planning, Slovenian Environment Agency, Archive of hydrological data ARSO, http://vode.arso.gov.si/hidarhiv/ (last access: 27 June 2023), 2021a.

ARSO: Ministry of the Environment and Spatial Planning, Slovenian Environment Agency, Archive of hydrological data ARSO, http://www.meteo.si/ (last access: 27 June 2023), 2021b.

Barber, C., Lamontagne, J. R., and Vogel, R. M.: Improved estimators of correlation and R^2 for skewed hydrologic data, Hydrolog. Sci. J., 65, 87–101, https://doi.org/10.1080/02626667.2019.1686639, 2020.

Beven, K.: How to make advances in hydrological modelling, Hydrol. Res., 50, 1481–1494, https://doi.org/10.2166/nh.2019.134, 2019.

Biondi, D., Freni, G., Iacobellis, V., Mascaro, G., and Montanari, A.: Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice, Phys. Chem. Earth Pt. A/B/C, 42–44, 70–76, https://doi.org/10.1016/j.pce.2011.07.037, 2012.

Choi, H. I.: Comment on Liu (2020): A rational performance criterion for hydrological model, J. Hydrol., 606, 126927, https://doi.org/10.1016/j.jhydrol.2021.126927, 2022.

Chollet, F.: Keras, GitHub [code], https://github.com/keras-team/keras (last access: 12 October 2022), 2015.

Cinkus, G. and Wunsch, A.: Busemorose/KGE_critical_evaluation: Model code release, Zenodo [code], https://doi.org/10.5281/zenodo.7274031, 2022.

Cinkus, G., Wunsch, A., Mazzilli, N., Liesch, T., Chen, Z., Ravbar, N., Doummar, J., Fernández-Ortega, J., Barberá, J. A., Andreo, B., Goldscheider, N., and Jourde, H.: Comparison of artificial neural networks and reservoir models for simulating karst spring discharge on five test sites in the Alpine and Mediterranean regions, Hydrol. Earth Syst. Sci., 27, 1961–1985, https://doi.org/10.5194/hess-27-1961-2023, 2023.

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 Resour. Res., 57, e2020WR029001, https://doi.org/10.1029/2020WR029001, 2021.

Freedman, D., Pisani, R., and Purves, R.: Statistics: Fourth International Student Edition, W. W. Norton & Company, New York, ISBN 978-0-393-92972-0, 2007.

Gabrovšek, F., Kogovšek, J., Kovačič, G., Petrič, M., Ravbar, N., and Turk, J.: Recent Results of Tracer Tests in the Catchment of the Unica River (SW Slovenia), Acta Carsolog., 39, 27–37, https://doi.org/10.3986/ac.v39i1.110, 2010.

Gohel, D.: Flextable: Functions for tabular reporting, Manual, https://cran.r-project.org/web/packages/flextable/index.html (last access: 27 June 2023), 2021.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resour. Res., 34, 751–763, https://doi.org/10.1029/97WR03495, 1998.

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, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hartmann, A., Goldscheider, N., Wagener, T., Lange, J., and Weiler, M.: Karst water resources in a changing world: Review of hydrological modeling approaches, Rev. Geophys., 52, 218–242, https://doi.org/10.1002/2013RG000443, 2014.

Hunter, J. D.: Matplotlib: A 2D Graphics Environment, Comput. Sci. Eng., 9, 90–95, https://doi.org/10.1109/MCSE.2007.55, 2007.

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, Environ. Model. Softw., 119, 32–48, https://doi.org/10.1016/j.envsoft.2019.05.001, 2019.

Jain, S. K., Mani, P., Jain, S. K., Prakash, P., Singh, V. P., Tullos, D., Kumar, S., Agarwal, S. P., and Dimri, A. P.: A Brief review of flood forecasting techniques and their applications, Int. J. River Basin Manage., 16, 329–344, https://doi.org/10.1080/15715124.2017.1411920, 2018.

Kauffeldt, A., Wetterhall, F., Pappenberger, F., Salamon, P., and Thielen, J.: Technical review of large-scale hydrological models for implementation in operational flood forecasting schemes on continental level, Environ. Model. Softw., 75, 68–76, https://doi.org/10.1016/j.envsoft.2015.09.009, 2016.

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

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

Kovačič, G.: Hydrogeological study of the Malenščica karst spring (SW Slovenia) by means of a time series analysis, Acta Carsolog., 39, 201–215, https://doi.org/10.3986/ac.v39i2.93, 2010.

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

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521, 436–444, https://doi.org/10.1038/nature14539, 2015.

Lee, J. S. and Choi, H. I.: A rebalanced performance criterion for hydrological model calibration, J. Hydrol., 606, 127372, https://doi.org/10.1016/j.jhydrol.2021.127372, 2022.

Legates, D. R. and McCabe Jr., G. J.: Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–241, https://doi.org/10.1029/1998WR900018, 1999.

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

Massmann, C., Woods, R., and Wagener, T.: Reducing equifinality by carrying out a multi-objective evaluation based on the bias, correlation and standard deviation errors, in: EGU2018, Vienna, Austria, 4–13 April, 2018EGUGA..2011457M, 11457, 2018.

Mayaud, C., Gabrovšek, F., Blatnik, M., Kogovšek, B., Petrič, M., and Ravbar, N.: Understanding flooding in poljes: A modelling perspective, J. Hydrol., 575, 874–889, https://doi.org/10.1016/j.jhydrol.2019.04.092, 2019.

Mazzilli, N., Guinot, V., Jourde, H., Lecoq, N., Labat, D., Arfib, B., Baudement, C., Danquigny, C., Soglio, L. D., and Bertin, D.: KarstMod: A modelling platform for rainfall – discharge analysis and modelling dedicated to karst systems, Environ. Model. Softw., 122, 103927, https://doi.org/10.1016/j.envsoft.2017.03.015, 2019.

McKinney, W.: Data Structures for Statistical Computing in Python, in: Proceedings of the 9th Python in Science Conference, Austin, Texas, 56–61, https://doi.org/10.25080/Majora-92bf1922-00a, 2010.

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, Hydrol. Earth Syst. Sci., 23, 2601–2614, https://doi.org/10.5194/hess-23-2601-2019, 2019.

Moriasi, D. N., Gitau, M. W., Pai, N., and Daggupati, P.: Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria, T. ASABE, 58, 1763–1785, https://doi.org/10.13031/trans.58.10715, 2015.

Muleta, M. K. and Nicklow, J. W.: Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model, J. Hydrol., 306, 127–145, https://doi.org/10.1016/j.jhydrol.2004.09.005, 2005.

Nash, J. E. and Sutcliffe, J.: River flow forecasting through conceptual models: Part 1. A discussion of principles, J. Hydrol., 10, 282–290, 1970.

Nogueira, F.: Bayesian Optimization: Open source constrained global optimization tool for Python, GitHub [code], https://github.com/bayesian-optimization/BayesianOptimization (last access: 27 June 2023), 2014.

Onyutha, C.: A hydrological model skill score and revised R-squared, Hydrol. Res., 53, 51–64, https://doi.org/10.2166/nh.2021.071, 2022.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É.: Scikit-learn: Machine Learning in Python, arXiv [preprint], https://doi.org/10.48550/arXiv.1201.0490, 2018.

Petric, M.: Chapter 10.3 – Case Study: Characterization, exploitation, and protection of the Malenščica karst spring, Slovenia, in: Groundwater Hydrology of Springs, edited by: Kresic, N. and Stevanovic, Z., Butterworth-Heinemann, Boston, 428–441, https://doi.org/10.1016/B978-1-85617-502-9.00021-9, 2010.

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

Ravbar, N., Barberá, J. A., Petrič, M., Kogovšek, J., and Andreo, B.: The study of hydrodynamic behaviour of a complex karst system under low-flow conditions using natural and artificial tracers (the catchment of the Unica River, SW Slovenia), Environ. Earth Sci., 65, 2259–2272, https://doi.org/10.1007/s12665-012-1523-4, 2012.

R Core Team: R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/ (last access: 27 June 2023), 2021.

Reback, J., jbrockmendel, McKinney, W., Bossche, J. V. den, Augspurger, T., Cloud, P., Hawkins, S., Roeschke, M., gfyoung, Sinhrks, Klein, A., Petersen, T., Hoefler, P., Tratner, J., She, C., Ayd, W., Naveh, S., Garcia, M., Darbyshire, J. H. M., Schendel, J., Hayden, A., Shadrach, R., Saxton, D., Gorelli, M. E., Li, F., Zeitlin, M., Jancauskas, V., McMaster, A., Battiston, P., and Seabold, S.: Pandas-dev/pandas: Pandas 1.3.5, Zenodo [code], https://doi.org/10.5281/zenodo.5774815, 2021.

Ritter, A. and Muñoz-Carpena, R.: Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments, J. Hydrol., 480, 33–45, https://doi.org/10.1016/j.jhydrol.2012.12.004, 2013.

Roberts, W., Williams, G. P., Jackson, E., Nelson, E. J., and Ames, D. P.: Hydrostats: A Python Package for Characterizing Errors between Observed and Predicted Time Series, Hydrology, 5, 66, https://doi.org/10.3390/hydrology5040066, 2018.

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

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

Seibert, J., Vis, M. J. P., Lewis, E., and van Meerveld, H. J.: Upper and lower benchmarks in hydrological modelling, Hydrol. Process., 32, 1120–1125, https://doi.org/10.1002/hyp.11476, 2018.

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

Thoen, E.: Padr: Quickly get datetime data ready for analysis, CRAN, https://CRAN.R-project.org/package=padr (last access: 27 June 2023), 2021.

van der Walt, S., Colbert, S. C., and Varoquaux, G.: The NumPy Array: A Structure for Efficient Numerical Computation, Comput. Sci. Eng., 13, 22–30, https://doi.org/10.1109/MCSE.2011.37, 2011.

van Rossum, G.: Python Tutorial, https://ir.cwi.nl/pub/5008/05008D.pdf (last access: 27 June 2023), 1995.

van Werkhoven, K., Wagener, T., Reed, P., and Tang, Y.: Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models, Adv. Water Resour., 32, 1154–1169, https://doi.org/10.1016/j.advwatres.2009.03.002, 2009.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H.: Welcome to the tidyverse, J. Open Source Softw., 4, 1686, https://doi.org/10.21105/joss.01686, 2019.

Wilke, C. O.: Cowplot: Streamlined plot theme and plot annotations for “Ggplot2”, Manual, https://cran.r-project.org/web/packages/cowplot/index.html (last access: 27 June 2023), 2020.

Willmott, C. J.: On the validations of models, Phys. Geogr., 2, 184–194, https://doi.org/10.1080/02723646.1981.10642213, 1981.

Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. J., Klink, K. M., Legates, D. R., O’Donnell, J., and Rowe, C. M.: Statistics for the evaluation and comparison of models, J. Geophys. Res., 90, 8995, https://doi.org/10.1029/JC090iC05p08995, 1985.

Willmott, C. J., Robeson, S. M., and Matsuura, K.: A refined index of model performance, Int. J. Climatol., 32, 2088–2094, https://doi.org/10.1002/joc.2419, 2012.

Wöhling, T., Samaniego, L., and Kumar, R.: Evaluating multiple performance criteria to calibrate the distributed hydrological model of the upper Neckar catchment, Environ. Earth Sci., 69, 453–468, https://doi.org/10.1007/s12665-013-2306-2, 2013.

Xie, Y., Allaire, J. J., and Grolemund, G.: R markdown: The definitive guide, Chapman and Hall/CRC, Boca Raton, Florida, ISBN 978-1-138-35933-8, 2018.

Xie, Y., Dervieux, C., and Riederer, E.: R markdown cookbook, Chapman and Hall/CRC, Boca Raton, Florida, ISBN 978-0-367-56383-7, 2020.

Zambrano-Bigiarini, M.: hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series, Zenodo [code], https://doi.org/10.5281/zenodo.839854, 2020.