



Pros and cons of various efficiency criteria for hydrological model performance evaluation

Charles Onyutha

Department of Civil and Environmental Engineering, Kyambogo University,
P.O. Box 1, Kyambogo, Kampala, Uganda

Correspondence: Charles Onyutha (conyutha@kyu.ac.ug)

Received: 25 May 2022 – Revised: 28 January 2023 – Accepted: 30 March 2023 – Published: 18 April 2024

Abstract. Confidence in hydrological predictions is linked to the model’s performance in reproducing available observations. However, judgment of a model’s quality is challenged by the differences which exist among the available efficiency criteria or objective functions. In this study, model outputs based on several objective functions were compared and found to differ with respect to various circumstances of variability, number of outliers, and model bias. Computational difficulty or speed of a model during calibration was shown to depend on the choice of the efficiency criterion. One source of uncertainty in hydrological modelling is the selection of a particular calibration method. However, this study showed that the choice of an objective function is another sub-source of calibration-related uncertainty. Thus, tackling the issue of uncertainties on model results should comprise combination of modelled series obtained based on (i) various objective functions separately applied to calibrate a model, (ii) different calibration methods, and (iii) several hydrological models. The pros and cons of many new and old efficiency criteria which can be found explored in this study highlight the need for modellers to understand the impact of various calibration-related sub-sources of uncertainties on model outputs.

Keywords. UPH 20; Modelling; Efficiency criteria; Model evaluation; Model calibration uncertainty

1 Introduction

Model quality can be judged in terms of “goodness-of-fit” (GOF) or how well a model fits through observations or measured data points (Onyutha, 2022). Mathematical measures of model quality can be regarded as efficiency criteria (Beven, 2012). There are several efficiency criteria such as coefficient of determination (R^2) (also called R -squared), revised R -squared (RRS) (Onyutha, 2022), index of agreement (IOA) (Willmot, 1981), Nash Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970), Kling Gupta efficiency (KGE) (Gupta et al., 2009), hydrological model skill score or Onyutha Efficiency (OE) (Onyutha, 2022), Liu mean efficiency (LME) (Liu, 2020), Taylor skill score (TSS) (Taylor, 2001), Root mean squared error (RMSE), and Mean absolute error (MAE). Differences exist among these various efficiency criteria and/or objective functions. Furthermore,

each efficiency criterion has its advantages and disadvantages. Thus, outputs of a model calibrated using various objective functions can also differ.

Question number 20 of the unsolved problems in hydrology (UPH 20) (Blöschl et al., 2019) deals with the need to reduce model uncertainty. As highlighted in UPH20, uncertainty in model results can be due to model structure, parameter, or inputs. However, other sources of model uncertainty also exist as explored in this study.

It is worth noting that many old and new efficiency criteria can be found in literature. However, no any paper (which explored the pros and cons of the various old and new efficiency criteria) could be found in literature by the time of conducting this study. Therefore, this study was aimed at filling this knowledge gap to give modellers wide-ranging information which could influence decision regarding the choice of the best performing model in making hydrological prediction.

2 Materials and methods

2.1 Data and models

To evaluate efficiency criteria, several series were required. One way to obtain the required series for analysis was to make use of outputs from hydrological models. Here, conceptual models were preferred to physical models. This was because conceptual models (i) have few parameters, (ii) require few inputs, and (iii) are easy to calibrate compared with physical models. In this line, two lumped conceptual hydrological models including Hydrological Model focusing on Sub-flows' Variation (HMSV) (Onyutha, 2019) and Nedbør-Afstrømnings-Model (NAM) (Nielsen and Hansen, 1973) were applied to model river flow of the Jardine River catchment (with area 2500 km²) in Australia. Each of the selected models require catchment-wide averaged rainfall and potential evapotranspiration as inputs.

2.2 Application of the selected hydrological models

Each selected model was calibrated using the Generalized Likelihood Uncertainty Estimation (Beven and Binley, 1992) framework. To do so, one of the metrics (R^2 , RRS, IOA, NSE, KGE, OE, LME, TSS, RMSE, and MAE) was selected as the objective function for model optimization by running HMSV or NAM using 10 000 sets of parameters randomized within the stipulated parameter space. The optimal set of parameters was that which yielded the best value of the objective function or the most promising fit between the observed and modelled series. The procedure was repeated to ensure each of the efficiency criteria was used to generate a set of model outputs. Finally, model outputs based on the various efficiency criteria were compared.

To compute the various efficiency criteria, consider \bar{X} and \bar{Y} as the mean of the observed (X) and modelled (Y) series, respectively. Take r as the coefficient of correlation between X and Y . Let r_a denote the maximum attainable r value (and it was taken to be 0.9975 in this study). Other terms to note include sample size (n), standard deviation of X (s_x), variance of X (s_x^2), standard deviation of Y (s_y), variance of Y (s_y^2), normalized s_y (\hat{s}_y), distance covariance of X (d_{xx}), distance covariance of Y (d_{yy}), and distance covariance of X and Y (d_{xy}). We can compute the selected efficiency criteria using

$$R^2 = \sum (X - \bar{X})(Y - \bar{Y}) / \sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2} \quad (1)$$

$$\text{RRS} = |r| \times \frac{\min(s_x, s_y)}{\max(s_x, s_y)} \times \frac{\min(s_x^2, s_y^2)}{\max(s_x^2, s_y^2)} \quad (2)$$

$$\text{where } s_{yc}^2 = (n - 1)^{-1} \times \sum (Y - \bar{Y})^2$$

$$\text{IOA} = 1 - \frac{\sum (X - Y)^2}{\sum (|(X - \bar{X})| + |(Y - \bar{Y})|)^2} \quad (3)$$

$$\text{NSE} = 1 - \frac{\sum (X - Y)^2}{\sum (X - \bar{X})^2} \quad (4)$$

$$\text{KGE} = 1 - \sqrt{(r - 1)^2 + \left(\frac{s_y}{s_x} - 1\right)^2 + \left(\frac{\bar{Y}}{\bar{X}} - 1\right)^2} \quad (5)$$

$$\text{OE} = r_d \times \frac{\min(d_{xx}, d_{yy})}{\max(d_{xx}, d_{yy})} \times \frac{\min(s_x^2, s_{yc}^2)}{\max(s_x^2, s_{yc}^2)} \quad (6)$$

$$\text{such that } r_d = d_{xy} \times (\sqrt{d_{xx} \times d_{yy}})^{-1}$$

$$\text{LME} = 1 - \sqrt{\left(\frac{s_y}{s_x} \times r - 1\right)^2 + \left(\frac{\bar{Y}}{\bar{X}} - 1\right)^2} \quad (7)$$

$$\text{TSS} = \frac{4(1 + Q_d)}{\left(\hat{s}_x + \frac{1}{\hat{s}_y}\right)^2 (1 + r_a)} \quad (8)$$

$$\text{while } Q_d = \sum (X - \bar{X})(Y - \bar{Y}) / (n \times s_x s_y)$$

$$\text{RMSE} = \sqrt{n^{-1} \times \sum (X - Y)^2} \quad (9)$$

$$\text{MAE} = n^{-1} \times \sum (|X - Y|) \quad (10)$$

Note that in Eqs. (2) and (6), when $s_x = s_{yc} = 0$, $\text{RRS} = \text{OE} = 0$. Results of the models were analysed while exploring the advantages and disadvantages of each of the considered efficiency criteria.

2.3 Simulation experiments

To demonstrate the pros and cons of the various efficiency criteria or objective functions, several experiments were conducted using synthetic series. This was to investigate the influence of various factors on model outputs obtained using each objective function. These factors included variability, bias, and presence of outliers. Various efficiency criteria were compared with respect to the computation time under varying sample sizes. In the experiment involving outliers, one series (A) of sample size n was generated. Here, the standard deviation and mean of A were purposely made to be approximately 0.25 and 0.5, respectively. Another series (B) was then obtained as a copy of A . One data point of A was made as an outlier and the resultant series was termed A_1 . An outlier was obtained by making a selected data point to be at least three times the 75th percentile in the series. The various efficiency criteria were applied to A_1 and B . The outliers' extent was computed as the ratio of the number of outliers to n in percentage. While keeping B unchanged, the procedure was repeated with the number of outliers in A increased to two, three, four, ..., G , to obtain $A_2, A_3, A_4, \dots, A_G$, respectively. Here, G is an integer which makes the outliers' extent to be approximately 5% of n .

For experiments involving biases, the starting point was series A. Using small increments in terms of $\Delta = 0.05$, several other series (B 's) were obtained using $B_{i,j} = (\Delta \times j + A_j)$ while considering ($1 \leq j \leq F$) and ($1 \leq i \leq n$) where F is a whole number such that the percentage bias is equal to or approximately 10%. Efficiency criteria were applied to A and each of the B 's. Like for biases, the experiments for variability were based on series A as the starting point. Several other series (B 's) were generated in terms of $\gamma = 0.001$ using $B_{i,j} = (\gamma \times j \times A_i)$ where ($1 \leq j \leq W$) and ($1 \leq i \leq n$). Here, W is a whole number which makes the standard deviation of B_j equal to 1. Relative coefficient of variation (RCV) was computed as the ratio of the standard deviation of B_j to the mean of A . Values of efficiency criteria applied to A and each of the B 's were compared. When $\gamma \times j = 1$, it meant, $B_j = A$ and here, $RCV \approx 0.5$ while all the efficiency criteria were expected to be at their values which indicate an ideal model performance.

3 Results and discussion

3.1 Comparison of various GOF metrics

Figure 1 shows result for comparison of the GOF values based on various criteria. The computation time (in seconds) for OE was slightly larger than those for other metrics (Fig. 1a). This is because OE comprises distance correlation, a term which has substantial run time of old algorithms. The fastest algorithm (known by the time of writing this paper) for computing distance correlation is the one provided by Chaudhuri and Hu (2019). Computation times for the other efficiency criteria are comparable. For series of n up to 1×10^6 , it took less than one second to obtain a value of each efficiency criterion.

The values of R^2 remained constant regardless of the variation in the RCV (Fig. 1a). As the RCV increased from zero to 0.5, values of the other efficiency criteria (NSE, KGE, OE, RRS, TSS, IOA) increased. On the other hand, as the RCV increased from 0.5 to one, the values of NSE, KGE, OE, RRS, TSS, and IOA decreased. MAE and RMSE decreased linearly to zero as RCV increased from zero to 0.5. However, both MAE and RMSE increased linearly as RCV increased from 0.5 to one. The best value of each GOF metric was obtained when the RCV was approximately 0.5 (or when the two series being compared were identical).

Generally, increase in the number of outliers leads to poor performance of the model (Fig. 1c–d). Except for TSS and R^2 , GOF metrics which occur over the range 0–1 reduced as the extent of outliers increased (Fig. 1c). However, an increase in the number of outliers makes the values of NSE, KGE, and LME less negative while MAE and RMSE increase in magnitudes (Fig. 1d). As bias increases, the magnitudes of all the selected GOF metrics (except for TSS and R^2) generally decrease (Fig. 1e) or increase (Fig. 1f) thereby

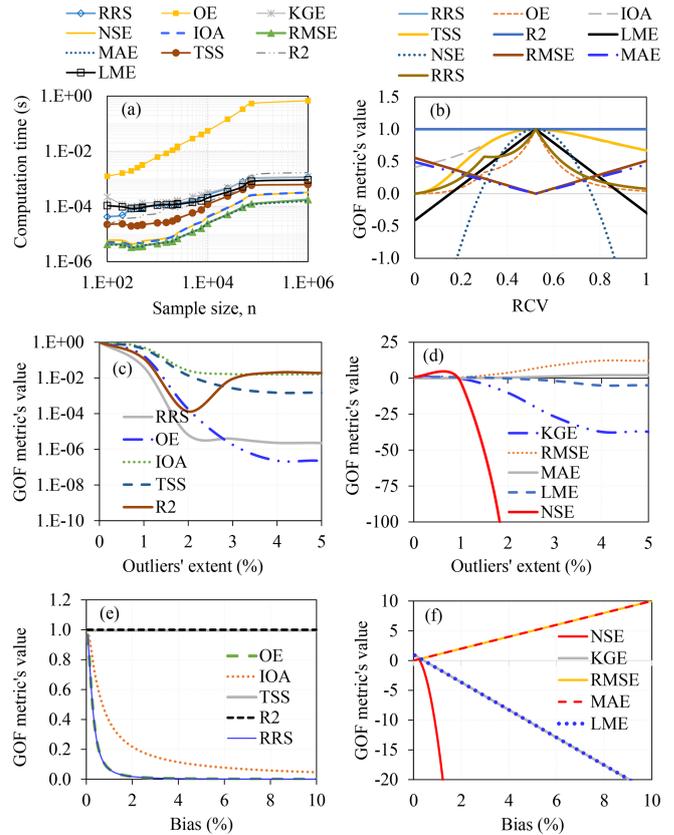


Figure 1. Comparing GOF metrics based on (a) computational load, (b) RCV, (c–d) outliers' extent, and (e–f) model bias.

showing reduction in model quality. TSS and R^2 do not quantify bias (Fig. 1e).

3.2 Application of HMSV and NAM

Figure 2 shows impact of the choice of an efficiency criterion on the modelled output. Modelled series based on various efficiency criteria resonated with observations to various extents (Fig. 2a–b). Considering HMSV (Fig. 2a), observed flow was more over-estimated by the modelled results based on R^2 than those obtained using other objective functions. However, modelled outputs based on RMSE exhibited the largest under-estimation compared with results from other objective functions. While considering NAM (Fig. 2b), results based on calibration using almost all selected objective functions resulted into under-estimation of the maximum observed flow. The largest and smallest under-estimation was from results based on R^2 and TSS, respectively.

The water balance closure was noted to depend on the selected efficiency criterion. The dissimilarity in the extent of spread of the right tails of the cumulative flow generated by HMSV and that of NAM was due to the difference in the structures of the models (Fig. 2c–d). One key area in which

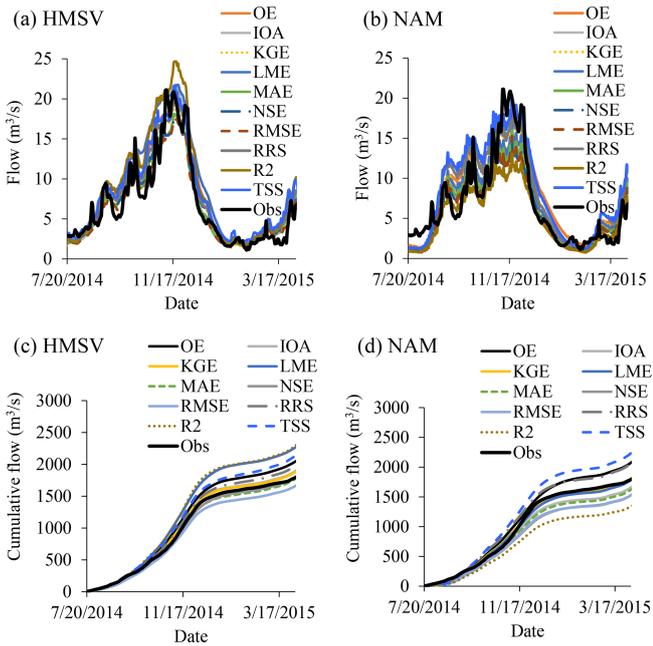


Figure 2. Observed (Obs) flow versus outputs of (a, c) HMSV and (b, d) NAM.

these hydrological models tended to differ was in reproducing observed peak flows.

3.3 Discussion

Known sources of uncertainty in hydrological models include model structure, input data errors, parameters, and calibration methods. However, this study, based on results from Figs. 1 and 2, showed that the choice of an objective function is another sub-source of calibration-related uncertainty. In other words, the choice of an efficiency criterion influences judgment of a model's quality (Onyutha, 2016). Expressly, the choice of an objective function influences the outputs of a model. Thus, it is important for a modeller to know their pros and cons of the various GOF metrics or objective functions.

3.3.1 Coefficient of determination (R^2)

The metric R^2 (Eq. 1) varies from 0 to 1. Advantages of R^2 are that: it is popular, can be computed fast, and R^2 does not have the interpretability limitations. Interpretation of R^2 is straightforward in terms of the percentage of the variability explained by the model. Furthermore, R^2 is suitable for variables which are linearly related when there is no bias. There are several disadvantages of R^2 . It does not quantify bias (Onyutha, 2022) and can be sensitive to outliers in the sample. It provides invalid results when the data has measurement errors (Cheng et al., 2014). There are several common misunderstandings on the use of R^2 ; thus, the use and interpretation of R^2 is confusing (Alexander et al., 2015).

C. Onyutha: Pros and cons of various efficiency criteria

The metric R^2 can be low and high for an accurate and imperfect model, respectively (Onyutha, 2022). The most commonly used R^2 formula or the version based on correlation is unsuitable for analysis of relationships which are not linear. It yields the same value when we regress X on Y and *vice versa* (Onyutha, 2022) thereby invalidating the terming of R^2 as coefficient of determination. Finally, R^2 lacks justification for its use as a descriptive statistic (Cameron, 1993).

3.3.2 Revised R -squared (RRS)

The metric RRS (Eq. 2) (Onyutha, 2022) varies from 0 to 1 and it has a number of advantages. Its computation is fast. We get different values in the two cases when regressing X on Y and *vice versa* (unlike R^2). Furthermore, RRS does not have interpretability limitations (Onyutha, 2022). The metric RRS indicates the amount of the total variance in observations explained by the model. A modeller can use RRS to evaluate model performance in terms of bias, correlation and variability (Onyutha, 2022). The formula for RRS does not comprise direct squaring of the error term, an aspect responsible for making other metrics (such as NSE) sensitive to large model errors. The main disadvantage of RRS is that it assumes linear relationship between X on Y .

3.3.3 Index of agreement (IOA)

The metric IOA (Eq. 3) (Willmot, 1981) ranges from 0 to 1 and the advantages of IOA are that it is widely applied and can be computed fast. The limits of the IOA do not have the interpretability limitations. Furthermore, IOA can detect additive differences in the means and variances of X and Y (Moriasi et al., 2015). The disadvantages of IOA are that it can yield high values even for poorly fit model (Krause et al., 2005; Onyutha, 2022). It is sensitive to extreme values in the sample. Furthermore, it lacks physical meaning since it does not have unit. In other words, the IOA values between zero and one are difficult to interpret.

3.3.4 Nash Sutcliffe efficiency (NSE)

The metric NSE (Eq. 4) (Nash and Sutcliffe, 1970) yields values which range from $-\infty$ to 1 and the advantages of NSE are that it is simple and popular. Its computation can be fast. Values of NSE in the two cases when we regress X on Y and *vice versa* are different (unlike R^2). The metric NSE has several disadvantages. It has interpretability limitation compared to other metrics which vary from zero to one (Onyutha, 2022). Furthermore, NSE lacks physical meaning since it does not have a unit. Values between the limits $-\infty$ and 1 are difficult to interpret. The metric NSE can be sensitive to outliers. The sampling uncertainty in NSE estimator is substantial (Clark et al., 2021). The reliance of NSE on mean of observations leads to exaggerated model efficiency when analysing highly seasonal river flow (Gupta et al., 2009). The

metric NSE also does not indicate bias (Jackson et al., 2019) and it is not suitable for single-event simulations since it can be inadequate in quantifying differences in the time and magnitude of peak flows (Jackson et al., 2019).

3.3.5 Kling Gupta efficiency (KGE)

The metric KGE (Eq. 5) (Gupta et al., 2009) varies from $-\infty$ to 1 and it has a number of advantages. It is becoming popular. It considers correlation, variability, and bias measures. It takes little time to be computed. The value of KGE in the two cases when we regress X on Y and *vice versa* are different (unlike R^2). On the other hand, there are many disadvantages of KGE. It varies from $-\infty$ to 1 and this engenders interpretability limitations compared to other metrics which vary from zero to one (Onyutha, 2022). A model's improvement starts from KGE equal to -0.41 even if the KGE values are still negative (Knoben et al., 2019). It can be sensitive to outliers. The sampling uncertainty of KGE is substantial (Clark et al., 2021). The metric KGE assumes linear relationships among variables and it can be inadequate in evaluating model performance when this assumption is violated. Furthermore, KGE assumes that the data is normally distributed and has no outliers; thus, non-normal distribution and presence of outliers all affect the metric. While optimizing KGE, the means of the simulation or values which under-estimate \bar{X} , especially in the high flows will tend to be favourably selected (Liu, 2020). The metric KGE lacks physical meaning since it does not have unit. Values between the limits $-\infty$ to 1 are difficult to interpret (not as it is clear for the values of R^2).

3.3.6 Hydrological model skill score

The metric OE (Eq. 6) (Onyutha, 2022) varies from 0 to 1 and its advantages are that it does not assume linear relationship among variables. The values of OE in the two cases when we regress X on Y and *vice versa* are different (unlike R^2). The metric OE allows model performance evaluation in terms of bias, correlation and variability (Onyutha, 2022). It does not involve direct squaring of the error term, an aspect responsible for making other metrics such as NSE sensitive to large model residuals. Like other metrics, OE has a few disadvantages. It is slightly computationally slower than other metrics such as NSE. As mentioned before, the component of OE which slightly increases the computational time is the distance correlation.

3.3.7 Liu mean efficiency (LME)

The metric LME (Eq. 7) (Liu, 2020) varies from $-\infty$ to 1. Its advantages are that it can be computed fast. It considers correlation, variability, and bias measures. The metric LME has a number of disadvantages. It assumes linear relationship. Furthermore, it is characterized by underdetermined so-

lutions mainly approaching the excessive flow variation (Lee and Choi, 2022). Thus, the maximum potential LME can be characterized by an infinite number solutions (Lee and Choi, 2022). The variation of LME from $-\infty$ to 1 brings about the interpretability limitations compared to other metrics which vary from 0 to 1. Like a few other metrics, LME lacks physical meaning since it does not have a unit. Values other than one are difficult to interpret. The sampling uncertainty of LME could be substantial just like that of NSE or KGE (following the methodology of Clark et al., 2021).

3.3.8 Taylor skill score (TSS)

The metric TSS (Eq. 8) (Taylor, 2001) varies from 0 to 1. The advantages of TSS are that it is popular and can be computed fast. Limits of TSS are straightforward to interpret. However, TSS has a number of disadvantages. It does not quantify bias. It can yield high values even for poorly fit model (Onyutha, 2022). It can also be affected by the presence of outliers in the data. The formula for TSS has a term which requires case-specific calibration to determine the maximum possible correlation attainable. Furthermore, the metric TSS assumes linear relationship. TSS also lacks physical meaning since it does not have a unit. Values between zero and one are difficult to interpret.

3.3.9 Root mean squared error (RMSE)

The metric RMSE (Eq. 9) varies from 0 to $+\infty$ and it has several advantages. It has the same unit as the variable being modelled; thus, easy to interpret given the physical meaning. It is popular, simple and can be computed fast. The main disadvantages of RMSE are that its values remain the same in the two cases when we regress X and Y and *vice versa* (like that of R^2). The metric RMSE can be affected by the presence of outliers in the data.

3.3.10 Mean absolute error (MAE)

The metric MAE (Eq. 10) varies from 0 to $+\infty$. Its advantages are that it is simple and can be computed fast. It has the same unit as the variable being modelled; thus, easy to interpret given its physical meaning. There are a few disadvantages of MAE. The two cases of regressing X on Y and *vice versa* lead to the same value of the metric MAE. Furthermore, MAE can be dominated by even one outlier or large error.

4 Conclusions

The pros and cons of the various efficiency criteria or objective functions presented in this study highlight the need to understand the impact of each efficiency criterion on model performance assessment (or the effect of the choice of an objective function on model calibration results) before making

hydrological predictions. Several sources of uncertainty in hydrological models are known including model structure, input data errors, parameters, and calibration methods. However, this study showed that the choice of an objective function for calibrating a model is also another sub-source of uncertainty related to model calibration. Therefore, a modeller should make use of several efficiency criteria to judge the quality of a model. Finally, the modeller should obtain an ensemble of modelled series using results based on various (i) objective functions used individually to calibrate a model (ii) calibration methods, and (iii) hydrological models.

Code availability. MATLAB codes to compute OE and RRS can be found in a freely downloadable supplementary material from Onyutha (2022). To obtain HMSV, please contact the author (conyutha@kyu.ac.ug).

Data availability. For more information about the used data, please contact the author (conyutha@kyu.ac.ug).

Competing interests. The author has declared that there are no competing interests.

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

Special issue statement. This article is part of the special issue "IAHS2022 – Hydrological sciences in the Anthropocene: Variability and change across space, time, extremes, and interfaces". It is a result of the XIth Scientific Assembly of the International Association of Hydrological Sciences (IAHS 2022), Montpellier, France, 29 May–3 June 2022.

Acknowledgements. The author acknowledges that modelling datasets from Onyutha (2022) were used in this study.

Financial support. This research has been supported by the International Association of Hydrological Sciences (IAHS) through an IAHS SYSTA award granted to the author to participate in the IAHS 2022 Scientific Assembly in Montpellier, France, where this paper was presented.

Review statement. This paper was edited by Christophe Cudenneq and reviewed by two anonymous referees.

References

- Alexander, D. L. J., Tropsha, A., and Winkler, D. A.: Beware of R²: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models, *J. Chem. Info. Mod.*, 55, 1316–1322, 2015.
- Beven, J. K.: *Rainfall-Runoff Modelling – The Primer*, 2nd Edition, Wiley-Blackwell, 488 pp., ISBN 978-0-470-71459-1, 2012.
- Beven, K. J. and Binley, A. M.: The future role of distributed models: model calibration and predictive uncertainty, *Hydrol. Process.*, 6, 279–298, 1992.
- Blöschl, G., Bierkens, F. P., Chambel, A., et al.: Twenty-three unsolved problems in hydrology (UPH) – a community perspective, *Hydrolog. Sci. J.*, 64, 1141–1158, <https://doi.org/10.1080/02626667.2019.1620507>, 2019.
- Cameron, S.: Why is the R squared adjusted reported?, *J. Quant. Econ.*, 9, 183–186, 1993.
- Chaudhuri, A. and Hu, W.: A fast algorithm for computing distance correlation, *Comput. Stat. Data Anal.*, 135, 15–24, 2019.
- Cheng, C.-L., Shalabh, and Garg, G.: Coefficient of determination for multiple measurement error models, *J. Multivar. Anal.*, 126, 137–152, 2014.
- 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 Papalexioiu S. M.: The abuse of popular performance metrics in hydrologic modelling, *Water Resour. Res.*, 57, e2020WR029001, <https://doi.org/10.1029/2020WR029001>, 2021.
- 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 modeling, *J. Hydrol.*, 377, 80–91, <https://doi.org/10.1016/j.jhydrol.2009.08.003>, 2009.
- 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. Modell. Softw.*, 119, 32–48, 2019.
- 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, *Hydrol. Earth Syst. Sci.*, 23, 4323–4331, <https://doi.org/10.5194/hess-23-4323-2019>, 2019.
- 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.
- 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.
- Liu, D.: A rational performance criterion for hydrological model, *J. Hydrol.*, 590, 125488, <https://doi.org/10.1016/j.jhydrol.2020.125488>, 2020.
- 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, 2015.
- Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – a discussion of principles, *J. Hydrol.*, 10, 282–290, 1970.

- Nielsen, S. A. and Hansen, E.: Numerical simulation of the rainfall-runoff process on a daily basis, *Nordic Hydrol.*, 4, 171–190, 1973.
- Onyutha, C.: Influence of hydrological model selection on simulation of moderate and extreme flow events: a case study of the Blue Nile basin, *Adv. Meteorol.*, 2016, 1–28, <https://doi.org/10.1155/2016/7148326>, 2016.
- Onyutha, C.: Hydrological model supported by a step-wise calibration against sub-flows and validation of extreme flow events, *Water*, 11, 244, <https://doi.org/10.3390/w11020244>, 2019.
- Onyutha, C.: A hydrological model skill score and revised R-squared, *Hydrol. Res.*, 53, 51–64, 2022.
- Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res.*, 106, 7183–7192, 2001.
- Willmott, C. J.: On the validation of models, *Phys. Geogr.*, 2, 184–194, <https://doi.org/10.1080/02723646.1981.10642213>, 1981.