Hydrol. Earth Syst. Sci., 23, 4323–4331, 2019
https://doi.org/10.5194/hess-23-4323-2019

Hydrology and Earth System Sciences
Open Access
EGU

Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores

Wouter J. M. Knoben[1,a], **Jim E. Freer**[2,3], **and Ross A. Woods**[1,3]

[1]Department of Civil Engineering, University of Bristol, Bristol, BS8 1TR, UK
[2]School of Geographical Sciences, University of Bristol, Bristol, BS8 1BF, UK
[3]Cabot Institute, University of Bristol, Bristol, BS8 1UJ, UK
[a]now at: University of Saskatchewan Coldwater Laboratory, Canmore, Alberta, Canada

Correspondence: Wouter J. M. Knoben (wouter.knoben@usask.ca)

Received: 26 June 2019 – Discussion started: 1 July 2019
Revised: 6 September 2019 – Accepted: 22 September 2019 – Published: 25 October 2019

Abstract. A traditional metric used in hydrology to summarize model performance is the Nash–Sutcliffe efficiency (NSE). Increasingly an alternative metric, the Kling–Gupta efficiency (KGE), is used instead. When NSE is used, NSE $= 0$ corresponds to using the mean flow as a benchmark predictor. The same reasoning is applied in various studies that use KGE as a metric: negative KGE values are viewed as bad model performance, and only positive values are seen as good model performance. Here we show that using the mean flow as a predictor does not result in KGE $= 0$, but instead KGE $= 1 - \sqrt{2} \approx -0.41$. Thus, KGE values greater than -0.41 indicate that a model improves upon the mean flow benchmark – even if the model's KGE value is negative. NSE and KGE values cannot be directly compared, because their relationship is non-unique and depends in part on the coefficient of variation of the observed time series. Therefore, modellers who use the KGE metric should not let their understanding of NSE values guide them in interpreting KGE values and instead develop new understanding based on the constitutive parts of the KGE metric and the explicit use of benchmark values to compare KGE scores against. More generally, a strong case can be made for moving away from ad hoc use of aggregated efficiency metrics and towards a framework based on purpose-dependent evaluation metrics and benchmarks that allows for more robust model adequacy assessment.

1 Introduction

Model performance criteria are often used during calibration and evaluation of hydrological models, to express in a single number the similarity between observed and simulated discharge (Gupta et al., 2009). Traditionally, the Nash–Sutcliffe efficiency (NSE, Nash and Sutcliffe, 1970) is an often-used metric, in part because it normalizes model performance into an interpretable scale (Eq. 1):

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{t=T} \left(Q_{\mathrm{sim}}(t) - Q_{\mathrm{obs}}(t)\right)^2}{\sum_{t=1}^{t=T} \left(Q_{\mathrm{obs}}(t) - \overline{Q_{\mathrm{obs}}}\right)^2}, \quad (1)$$

where T is the total number of time steps, $Q_{\mathrm{sim}}(t)$ the simulated discharge at time t, $Q_{\mathrm{obs}}(t)$ the observed discharge at time t, and $\overline{Q_{\mathrm{obs}}}$ the mean observed discharge. NSE $= 1$ indicates perfect correspondence between simulations and observations; NSE $= 0$ indicates that the model simulations have the same explanatory power as the mean of the observations; and NSE < 0 indicates that the model is a worse predictor than the mean of the observations (e.g. Schaefli and Gupta, 2007). NSE $= 0$ is regularly used as a benchmark to distinguish "good" and "bad" models (e.g. Houska et al., 2014; Moriasi et al., 2007; Schaefli and Gupta, 2007). However, this threshold could be considered a low level of predictive skill (i.e. it requires little understanding of the ongoing hydrologic processes to produce this benchmark). It is not an equally representative benchmark for different flow regimes (for example, the mean is not representative of very seasonal regimes but it is a good approximation of regimes without a strong seasonal component; Schaefli and Gupta, 2007), and

it is also a relatively arbitrary choice (for example, Moriasi et al., 2007, define several different NSE thresholds for different qualitative levels of model performance) that can influence the resultant prediction uncertainty bounds (see e.g. Freer et al., 1996). However, using such a benchmark provides context for assessing model performance (Schaefli and Gupta, 2007).

The Kling–Gupta efficiency (KGE; Eq. 2, Gupta et al., 2009) is based on a decomposition of NSE into its constitutive components (correlation, variability bias and mean bias), addresses several perceived shortcomings in NSE (although there are still opportunities to improve the KGE metric and to explore alternative ways to quantify model performance) and is increasingly used for model calibration and evaluation:

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

where r is the linear correlation between observations and simulations, α a measure of the flow variability error, and β a bias term (Eq. 3):

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{\mathrm{sim}}}{\sigma_{\mathrm{obs}}} - 1\right)^2 + \left(\frac{\mu_{\mathrm{sim}}}{\mu_{\mathrm{obs}}} - 1\right)^2}, \quad (3)$$

where σ_{obs} is the standard deviation in observations, σ_{sim} the standard deviation in simulations, μ_{sim} the simulation mean, and μ_{obs} the observation mean (i.e. equivalent to $\overline{Q_{\mathrm{obs}}}$). Like NSE, KGE = 1 indicates perfect agreement between simulations and observations. Analogous to NSE = 0, certain authors state that KGE < 0 indicates that the mean of observations provides better estimates than simulations (Castaneda-Gonzalez et al., 2018; Koskinen et al., 2017), although others state that this interpretation should not be attached to KGE = 0 (Gelati et al., 2018; Mosier et al., 2016). Various authors use positive KGE values as indicative of "good" model simulations, whereas negative KGE values are considered "bad", without explicitly indicating that they treat KGE = 0 as their threshold between "good" and "bad" performance. For example, Rogelis et al. (2016) consider model performance to be "poor" for $0.5 > \mathrm{KGE} > 0$, and negative KGE values are not mentioned. Schönfelder et al. (2017) consider negative KGE values "not satisfactory". Andersson et al. (2017) mention negative KGE values in the same sentence as negative NSE values, implying that both are considered similarly unwanted. Fowler et al. (2018) consider reducing the number of occurrences of negative KGE values as desirable. Knoben et al. (2018) cap figure legends at KGE = 0 and mask negative KGE values. Siqueira et al. (2018) consider ensemble behaviour undesirable as long as it produces negative KGE and NSE values. Sutanudjaja et al. (2018) only count catchments where their model achieves KGE > 0 as places where their model application was successful. Finally, Towner et al. (2019) use KGE = 0 as the threshold to switch from red to blue colour coding of model results, and only positive KGE values are considered "skilful". Naturally, authors prefer higher efficiency values over lower values, because this indicates their model is closer to perfectly reproducing observations (i.e. KGE = 1). Considering the traditional use of NSE and its inherent quality that the mean flow results in NSE = 0, placing the threshold for "good" model performance at KGE = 0 seems equally natural. We show in this paper that this reasoning is generally correct – positive KGE values do indicate improvements upon the mean flow benchmark – but not complete. In KGE terms, negative values do not necessarily indicate a model that performs worse than the mean flow benchmark. We first show this in mathematical terms and then present results from a synthetic experiment to highlight that NSE and KGE values are not directly comparable and that understanding of the NSE metric does not translate well into understanding of the KGE metric.

Note that a weighted KGE version exists that allows specification of the relative importance of the three KGE terms (Gupta et al., 2009), as do a modified KGE (Kling et al., 2012) and a non-parametric KGE (Pool et al., 2018). These are not explicitly discussed here, because the issue we address here (i.e. the lack of an inherent benchmark in the KGE equation) applies to all these variants of KGE.

2 KGE value of the mean flow benchmark

Consider the case where $Q_{\mathrm{sim}}(t) = \overline{Q_{\mathrm{obs}}}$ for an arbitrary number of time steps, and where $\overline{Q_{\mathrm{obs}}}$ is calculated from an arbitrary observed hydrograph. In this particular case, $\mu_{\mathrm{obs}} = \mu_{\mathrm{sim}}$, $\sigma_{\mathrm{obs}} \neq 0$ but $\sigma_{\mathrm{sim}} = 0$. Although the linear correlation between observations and simulations is formally undefined when $\sigma_{\mathrm{sim}} = 0$, it makes intuitive sense to assign $r = 0$ in this case, since there is no relationship between the fluctuations of the observed and simulated hydrographs. Equation (3) becomes (positive terms shown as symbols) the following:

$$\mathrm{KGE} = 1 - \sqrt{(0-1)^2 + \left(\frac{0}{\sigma_{\mathrm{obs}}} - 1\right)^2 + \left(\frac{\mu_{\mathrm{obs}}}{\mu_{\mathrm{obs}}} - 1\right)^2}, \quad (4)$$

$$\mathrm{KGE} = 1 - \sqrt{(0-1)^2 + (0-1)^2 + (1-1)^2}, \quad (5)$$

$$\mathrm{KGE} = 1 - \sqrt{2}. \quad (6)$$

Thus, the KGE score for a mean flow benchmark is $\mathrm{KGE}(\overline{Q_{\mathrm{obs}}}) \approx -0.41$.

3 Consequences

3.1 NSE and KGE values cannot be directly compared and should not be treated as approximately equivalent

Through long use, hydrologic modellers have developed intuitive assessments about which NSE values can be considered acceptable for their preferred model(s) and/or catch-

ment(s); however, this interpretation of acceptable NSE values cannot easily be mapped onto corresponding KGE values. There is no unique relationship between NSE and KGE values (Fig. 1a, note the scatter along both axes; see also Appendix 1), and where NSE values fall in the KGE component space depends in part on the coefficient of variation (CV) of the observations (see animated Fig. S1 in the Supplement for a comparison of where NSE = 0 and KGE = $1-\sqrt{2}$ fall in the space described by KGE's r, a and b components for different CVs, highlighting that many different combinations of r, a and b can result in the same overall NSE or KGE value).

This has important implications when NSE or KGE thresholds are used to distinguish between behavioural and non-behavioural models (i.e. when a threshold is used to decide between accepting or rejecting models). Figure 1b–g are used to illustrate a synthetic experiment, where simulated flows are generated from observations and a threshold for behavioural models is set midway between the value for the mean flow benchmark (NSE = 0 and KGE = −0.41) and the value for a perfect simulation (NSE = KGE = 1): simulations are considered behavioural if NSE > 0.5 or KGE > 0.3. Each row shows flows from a different catchment, with increasing coefficients of variations (i.e. 0.28, 2.06 and 5.00 respectively). In Fig. 1b, d and f, the simulated flow is calculated as the mean of observations. NSE values are constant at NSE = 0 for all three catchments, and KGE values are constant at KGE = −0.41. In Fig. 1c, e and g, the simulated flow is the observed flow plus an offset, to demonstrate the variety of impacts that bias has on NSE and KGE (similar examples could be generated for other types of error relating to correlation or variability, but these examples are sufficient to make the point that NSE and KGE behave quite differently). In Fig. 1c, simulated flows are calculated as observed flows $+0.45\,\mathrm{mm\,d^{-1}}$ (bias +39 %). With the specified thresholds, this simulation would be considered behavioural when using KGE (0.61 > 0.3), but not with NSE (−0.95 < 0.5). In Fig. 1e, simulated flows are calculated as observed flows $+0.5\,\mathrm{mm\,d^{-1}}$ (bias +40 %). In this case, however, these simulations are considered behavioural with both metrics (NSE: 0.96 > 0.5; KGE: 0.60 > 0.3). Figure 1g shows an example where simulated flows are calculated as observations $+0.7\,\mathrm{mm\,d^{-1}}$ (bias +97 %), which is considered behavioural when NSE is used (0.96 > 0.5), but not when KGE is used (0.03 < 0.3).

These examples show that NSE values that are traditionally interpreted as high do not necessarily translate into high KGE values and that standards of acceptability developed through extensive use of the NSE metric are not directly applicable to KGE values. Instead, hydrologists who choose to use the KGE metric need to develop new understanding of how this metric should be interpreted and not let themselves be guided by their understanding of NSE.

3.2 Explicit statements about benchmark performance are needed in modelling studies

The Nash–Sutcliffe efficiency has an inherent benchmark in the form of the mean flow, giving NSE = 0. This benchmark is not inherent in the definition of the Kling–Gupta efficiency, which is instead an expression of distance away from the point of ideal model performance in the space described by its three components. When Q_{sim} is $\overline{Q_{obs}}$, KGE ≈ −0.41, but there is no direct reason to choose this benchmark over other options (see e.g. Ding, 2019; Schaefli and Gupta, 2007; Seibert, 2001; Seibert et al., 2018). Because KGE itself has no inherent benchmark value to enable a distinction between "good" and "bad" models, modellers using KGE must be explicit about the benchmark model or value they use to compare the performance of their model against. As succinctly stated in Schaefli and Gupta (2007),

> Every modelling study should explain and justify the choice of benchmark [that] should fulfil the basic requirement that every hydrologist can immediately understand its explanatory power for the given case study and, therefore, appreciate how much better the actual hydrologic model is.

If the mean flow is chosen as a benchmark, model performance in the range −0.41 < KGE ≤ 1 could be considered "reasonable" in the sense that the model outperforms this benchmark. By artificially and consistently imposing a threshold at KGE = 0 to distinguish between "good" and "bad" models, modellers limit themselves in the models and/or parameter sets they consider in a given study, without rational justification of this choice and without taking into account whether more catchment-appropriate or study-appropriate thresholds could be defined.

3.3 On communicating model performance through skill scores

If the benchmark is explicitly chosen, then a so-called skill score can be defined, which is the performance of any model compared to the pre-defined benchmark (e.g. Hirpa et al., 2018; Towner et al., 2019):

$$\mathrm{KGE_{skill\,score}} = \frac{\mathrm{KGE_{model}} - \mathrm{KGE_{benchmark}}}{1 - \mathrm{KGE_{benchmark}}}.$$

The skill score is scaled such that positive values indicate a model that is better than the benchmark model and negative values indicate a model that is worse than the benchmark model. This has a clear benefit in communicating whether a model improves on a given benchmark or not with an intuitive threshold at $\mathrm{KGE_{skill\,score}} = 0$, where negative values clearly indicate a model worse than the benchmark and positive values a model that outperforms the benchmark.

However, scaling the KGE metric might introduce a different communication issue. In absolute terms, it seems clear

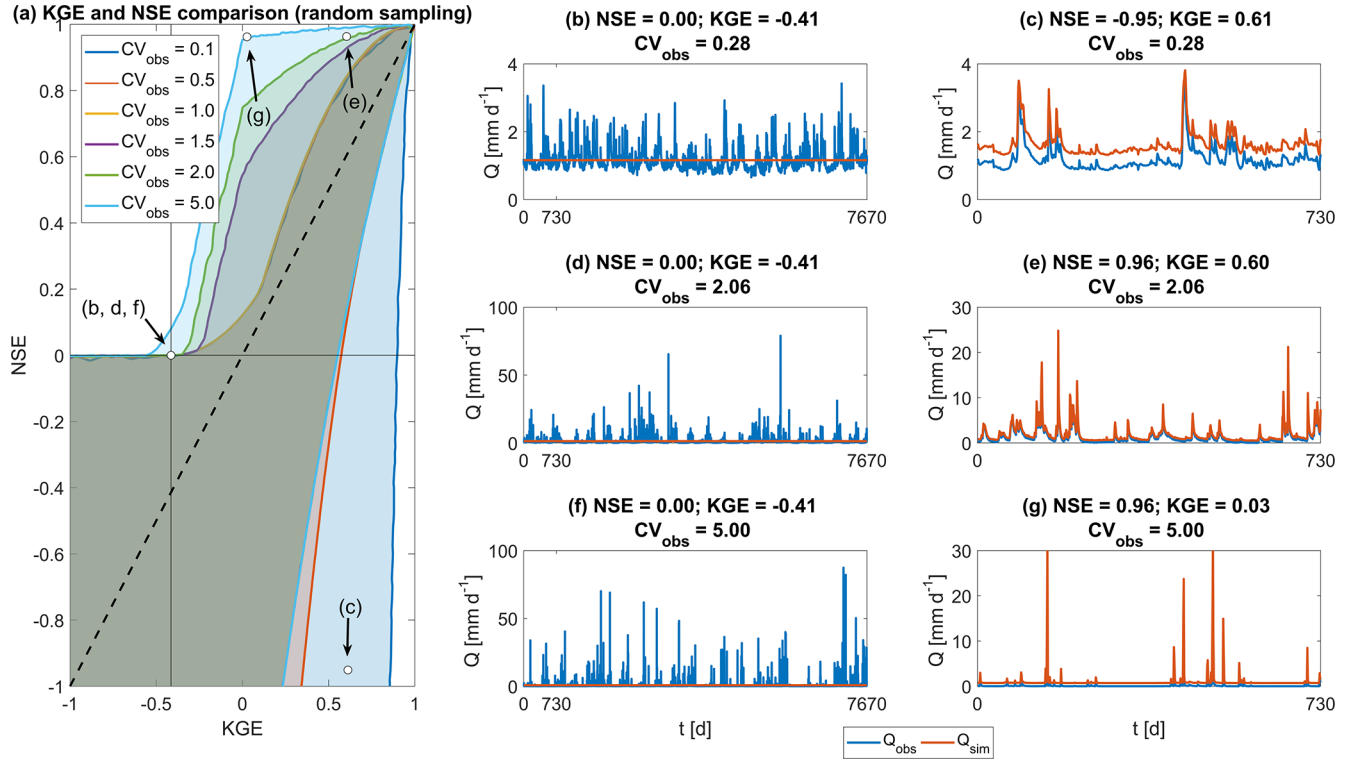

Figure 1. Overview of the relationship between NSE and KGE. **(a)** Comparison of KGE and NSE values based on random sampling of the r, a and b components used in KGE and NSE, using six different values for the coefficient of variation of observations (see Appendix for method and separate plots of each plane). Internal axes are drawn at KGE $= 1 - \sqrt{2}$ and NSE $= 0$. The dashed diagonal is the 1 : 1 line. Locations of panels **(b)**–**(g)** indicated in brackets. **(b, d, f)** Simulated flow Q_{sim} is created from the mean of Q_{obs}. **(c)** Q_{sim} is created as $Q_{obs} + 0.45\,mm\,d^{-1}$ on every time step, increasing the bias of observations. **(e)** Q_{sim} is created as $Q_{obs} + 0.5\,mm\,d^{-1}$ on every time step. **(g)** Q_{sim} is created as $Q_{obs} + 0.7\,mm\,d^{-1}$ on every time step. The y axis is capped at $30\,mm\,d^{-1}$ to better visualize the difference between observations and synthetic simulations. **(b)**–**(g)** Flow observations are taken from the CAMELS data set (Addor et al., 2017a), using catchments 04124000, 01613050 and 05507600 for the top, middle and lower plots respectively.

that improving on $KGE_{benchmark} = 0.99$ by using a model might be difficult: the "potential for model improvement over benchmark" is only $1 - 0.99 = 0.01$. With a scaled metric, the "potential for model improvement over benchmark" always has a range of [0,1], but information about how large this potential was in the first place is lost and must be reported separately for proper context. If the benchmark is already very close to perfect simulation, a $KGE_{skill\ score}$ of 0.5 might indicate no real improvement in practical terms. In cases where the benchmark constitutes a poor simulation, a $KGE_{skill\ score}$ of 0.5 might indicate a large improvement through using the model. This issue applies to any metric that is converted to a skill score.

Similarly, a skill score reduces the ease of communication about model deficiencies. It is generally difficult to interpret any score above the benchmark score but below the perfect simulation (in case of the KGE metric, KGE $= 1$) beyond "higher is better", but an absolute KGE score can at least be interpreted in terms of deviation-from-perfect on its a, b and r components. A score of KGE $= 0.95$ with $r = 1$, $a = 1$ and $b = 1.05$ indicates simulations with 5 % bias. The scaled $KGE_{skill\ score} = 0.95$ cannot be so readily interpreted.

3.4 The way forward: new understanding based on purpose-dependent metrics and benchmarks

The modelling community currently does not have a single perfect model performance metric that is suitable for every study purpose. Indeed, global metrics that attempt to lump complex model behaviour and residual errors into a single value may not be useful for exploring model deficiencies and diagnostics into how models fail or lack certain processes. If such metrics are used however, a modeller should make a conscious and well-founded choice about which aspects of the simulation they consider most important (if any), and in which aspects of the simulation they are willing to accept larger errors. The model's performance score should then be compared against an appropriate benchmark, which can inform to what extent the model is fit for purpose.

If the KGE metric is used, emphasizing certain aspects of a simulation is straightforward by attaching weights to the

individual KGE components to reduce or increase the impact of certain errors on the overall KGE score, treating the calibration as a multi-objective problem (e.g. Gupta et al., 1998) with varying weights assigned to the three objectives. An example of the necessity of such an approach can be found in Fig. 1g. For a study focussing on flood peaks, an error of only $0.7\,\mathrm{mm\,d^{-1}}$ for each peak might be considered skilful, although the bias of these simulations is very large (+97 %). Due to the small errors and the high coefficient of variation in this catchment, the NSE score of these simulations reaches a value that would traditionally be considered as very high (NSE = 0.96). The standard formulation of KGE however is heavily impacted by the large bias, and the simulations in Fig. 1g result in a relatively low KGE score (KGE = 0.03). If one relies on this aggregated KGE value only, the low KGE score might lead a modeller to disqualify these simulations from further analysis, even if the simulations are performing very well for the purpose of peak flow simulation. Investigation of the individual components of KGE would show that this low value is only due to bias errors and not due to an inability to simulate peak flows. The possibility to attach different weights to specific components of the KGE metric can allow a modeller to shift the metric's focus: by reducing the importance of bias in determining the overall KGE score or emphasizing the importance of the flow variability error, the metric's focus can be moved towards peak flow accuracy (see Mizukami et al., 2019, for a discussion of purpose-dependent KGE weights and a comparison between (weighted) KGE and NSE for high-flow simulation). For example, using weightings $[1, 5, 1]$ for $[r, a, b]$ to emphasize peak flow simulation (following Mizukami et al., 2019), the KGE score in Fig. 1g would increase to KGE = 0.81 This purpose-dependent score should then be compared against a purpose-dependent benchmark to determine whether the model can be considered fit for purpose.

However, aggregated performance metrics with a statistical nature, such as KGE, are not necessarily informative about model deficiencies from a hydrologic point of view (Gupta et al., 2008). While KGE improves upon the NSE metric in certain ways, Gupta et al. (2009) explicitly state that their intent with KGE was "not to design an improved measure of model performance" but only to use the metric to illustrate that there are inherent problems with mean-squared-error-based optimization approaches. They highlight an obvious weakness of the KGE metric, namely that many hydrologically relevant aspects of model performance (such as the shape of rising limbs and recessions, as well as timing of peak flows) are all lumped into the single correlation component. Future work could investigate alternative metrics that separate the correlation component of KGE into multiple, hydrologically meaningful, aspects. There is no reason to limit such a metric to only three components either, and alternative metrics (or sets of metric components) can be used to expand the multi-objective optimization from three components to as many dimensions as are considered necessary or hydrologically informative. Similarly, there is no reason to use aggregated metrics only, and investigating model behaviour on the individual time-step level can provide increased insight into where models fail (e.g. Beven et al., 2014).

Regardless of whether KGE or some other metric is used, the final step in any modelling exercise would be comparing the obtained efficiency score against a certain benchmark that dictates which kind of model performance might be expected (e.g. Seibert et al., 2018) and decide whether the model is truly skilful. These benchmarks should not be specified in an ad hoc manner (e.g. our earlier example where the thresholds are arbitrarily set at NSE = 0.5 and KGE = 0.3 is decidedly poor practice) but should be based on hydrologically meaningful considerations. The explanatory power of the model should be obvious from the comparison of benchmark and model performance values (Schaefli and Gupta, 2007), such that the modeller can make an informed choice on whether to accept or reject the model and make an assessment of the model's strengths and where current model deficiencies are present. Defining such benchmarks is not straightforward because it relies on the interplay between our current hydrologic understanding, the availability and quality of observations, the choice of model structure and parameter values, and modelling objectives. However, explicitly defining such well-informed benchmarks will allow more robust assessments of model performance (see for example Abramowitz, 2012, for a discussion of this process in the land-surface community). How to define a similar framework within hydrology is an open question to the hydrologic community.

4 Conclusions

There is a tendency in current literature to interpret Kling–Gupta efficiency (KGE) values in the same way as Nash–Sutcliffe efficiency (NSE) values: negative values indicate "bad" model performance, whereas positive values indicate "good" model performance. We show that the traditional mean flow benchmark that results in NSE = 0 and the likely origin of this "bad/good" model distinction, results in $\mathrm{KGE} = 1 - \sqrt{2}$. Unlike NSE, KGE does not have an inherent benchmark against which flows are compared and there is no specific meaning attached to KGE = 0. Modellers using KGE must be specific about the benchmark against which they compare their model performance. If the mean flow is used as a KGE benchmark, all model simulations with $-0.41 < \mathrm{KGE} \leq 1$ exceeds this benchmark. Furthermore, modellers must take care to not let their interpretation of KGE values be consciously or subconsciously guided by their understanding of NSE values, because these two metrics cannot be compared in a straightforward manner. Instead of relying on the overall KGE value, in-depth analysis of the KGE components can allow a modeller to both better understand what the overall value means in terms of model errors and to modify the metric through weighting of the

components to better align with the study's purpose. More generally, a strong case can be made for moving away from ad hoc use of aggregated efficiency metrics and towards a framework based on purpose-dependent evaluation metrics and benchmarks that allows for more robust model adequacy assessment.

Data availability. The CAMELS catchment data can be accessed as open-source data through the provided reference (Addor et al., 2017b).

Appendix A

The relation between possible KGE and NSE values shown in Fig. 1a has been determined through random sampling of 1 000 000 different combinations of the components r, a and b of KGE (Eq. 2), for six different coefficients of variation (CVs; 0.1, 0.5, 1.0, 1.5, 2.0, 5.0 respectively). Values were sampled in the following ranges: $r = [-1, 1]$; $a = [0, 2]$; $b = [0, 2]$. The KGE value of each sample is found through Eq. (2). The corresponding NSE value for each sampled combination of r, a and b is found through

$$\mathrm{NSE} = 2ar - a^2 - \frac{(b-1)^2}{\mathrm{CV}_{\mathrm{obs}}^2}. \tag{A1}$$

Figure A1 shows the correspondence between KGE and NSE values for the six different CVs. Axis limits have been capped at $[-1, 1]$ for clarity. Equation (A1) can be found by starting from Eq. (4) in Gupta et al. (2009) and expressing $\beta_n = \frac{\mu_s - \mu_o}{\sigma_o}$ in terms of $b = \frac{\mu_s}{\mu_o}$, using $\mathrm{CV}_{\mathrm{obs}} = \frac{\sigma_{\mathrm{obs}}}{\mu_{\mathrm{obs}}}$.

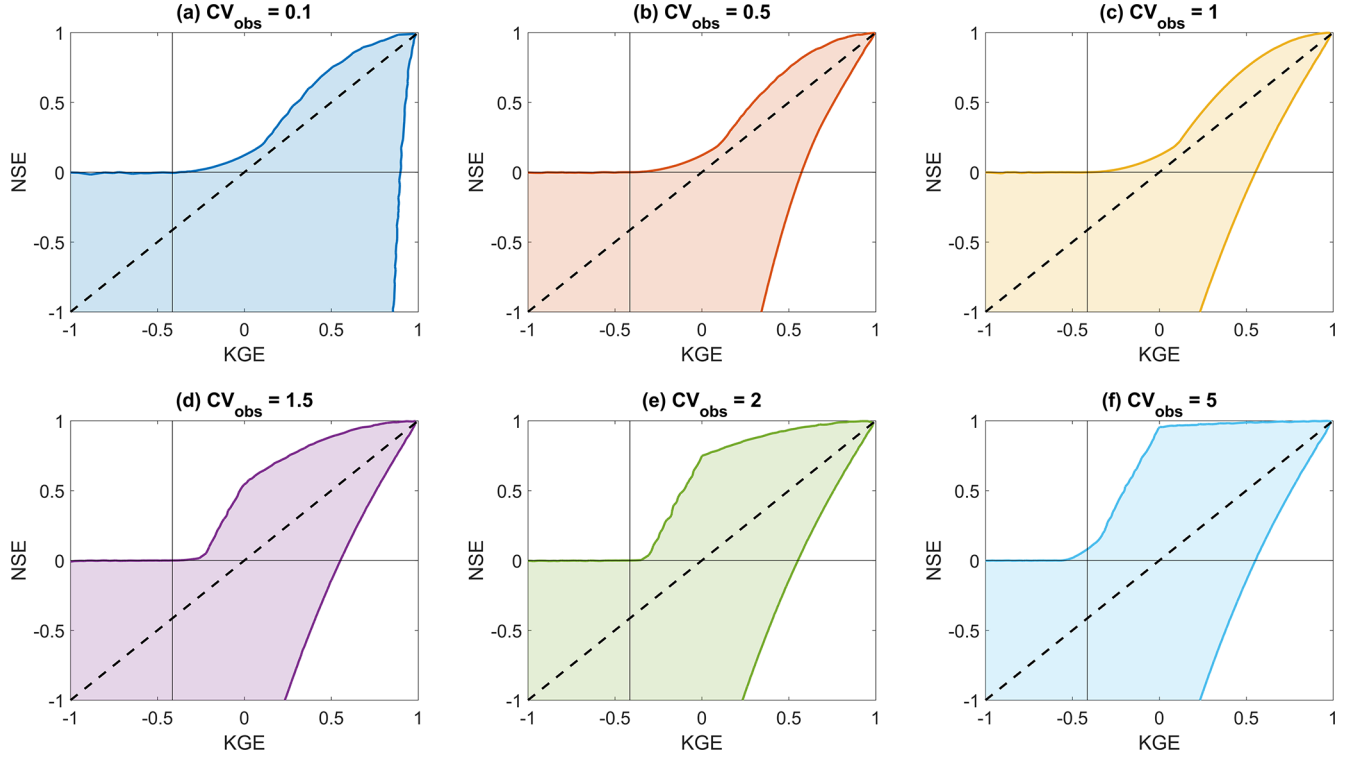

Figure A1. Correspondence between synthetic KGE and NSE values based on 1×10^6 random samples of components r, a and b, for different coefficients of variation (CVs). Colour coding corresponds to the colours used in Fig. 1a.

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/hess-23-4323-2019-supplement.

Author contributions. WJMK performed the initial analyses and draft text that outlined this paper. The idea was further developed in meetings between all authors. RAW provided the derivation of NSE in terms of a, b and r. The manuscript was written and revised by WJMK with contributions from JEF and RAW.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We are grateful to Hoshin Gupta, John Ding, Paul Whitfield and one anonymous reviewer, for their time and comments which helped us strengthen the message presented in this work.

Financial support. This research has been supported by the EPSRC WISE CDT (grant no. EP/L016214/1).

Review statement. This paper was edited by Nunzio Romano and reviewed by Hoshin Gupta and one anonymous referee.

References

Abramowitz, G.: Towards a public, standardized, diagnostic benchmarking system for land surface models, Geosci. Model Dev., 5, 819–827, https://doi.org/10.5194/gmd-5-819-2012, 2012.

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

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies. version 2.0., UCAR/NCAR, Boulder, CO, USA, https://doi.org/10.5065/D6G73C3Q, 2017b.

Andersson, J. C. M., Arheimer, B., Traoré, F., Gustafsson, D., and Ali, A.: Process refinements improve a hydrological model concept applied to the Niger River basin, Hydrol. Process., 31, 4540–4554, https://doi.org/10.1002/hyp.11376, 2017.

Beven, K. J., Younger, P. M., and Freer, J.: Struggling with Epistemic Uncertainties in Environmental Modelling of Natural Hazards, in: Second International Conference on Vulnerability and Risk Analysis and Management (ICVRAM) and the Sixth International Symposium on Uncertainty, Modeling, and Analysis (ISUMA), 13–16 July 2014, Liverpool, UK, American Society of Civil Engineers, 13–22, 2014.

Castaneda-Gonzalez, M., Poulin, A., Romero-Lopez, R., Arsenault, R., Chaumont, D., Paquin, D., and Brissette, F.: Impacts of Regional Climate Model Spatial Resolution on Summer Flood Simulation, in: HIC 2018, 13th International Conference on Hydroinformatics, 1–6 July 2018, Palermo, Italy, 3, 372–362, 2018.

Ding, J.: Interactive comment on "Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores" by Wouter J. M. Knoben et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-327-SC1, 2019.

Fowler, K., Coxon, G., Freer, J., Peel, M., Wagener, T., Western, A., Woods, R., and Zhang, L.: Simulating Runoff Under Changing Climatic Conditions: A Framework for Model Improvement, Water Resour. Res., 54, 9812–9832, https://doi.org/10.1029/2018WR023989, 2018.

Freer, J. E., Beven, K., and Ambroise, B.: Bayesian Estimation of Uncertainty in Runoff Prediction and the Value of Data: An Application of the GLUE Approach, Water Resour. Res., 32, 2161–2173, https://doi.org/10.1029/95WR03723, 1996.

Gelati, E., Decharme, B., Calvet, J.-C., Minvielle, M., Polcher, J., Fairbairn, D., and Weedon, G. P.: Hydrological assessment of atmospheric forcing uncertainty in the Euro-Mediterranean area using a land surface model, Hydrol. Earth Syst. Sci., 22, 2091–2115, https://doi.org/10.5194/hess-22-2091-2018, 2018.

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., Wagener, T., and Liu, Y.: Reconciling theory with observations?: elements of a diagnostic approach to model evaluation, Hydrol. Process., 3813, 3802–3813, https://doi.org/10.1002/hyp.6989, 2008.

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

Hirpa, F. A., Salamon, P., Beck, H. E., Lorini, V., Alfieri, L., Zsoter, E., and Dadson, S. J.: Calibration of the Global Flood Awareness System (GloFAS) using daily streamflow data, J. Hydrol., 566, 595–606, https://doi.org/10.1016/j.jhydrol.2018.09.052, 2018.

Houska, T., Multsch, S., Kraft, P., Frede, H.-G., and Breuer, L.: Monte Carlo-based calibration and uncertainty analysis of a coupled plant growth and hydrological model, Biogeosciences, 11, 2069–2082, https://doi.org/10.5194/bg-11-2069-2014, 2014.

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., Woods, R. A., and Freer, J. E.: A Quantitative Hydrological Climate Classification Evaluated with Independent Streamflow Data, Water Resour. Res., 54, 5088–5109, https://doi.org/10.1029/2018WR022913, 2018.

Koskinen, M., Tahvanainen, T., Sarkkola, S., Menberu, M. W., Laurén, A., Sallantaus, T., Marttila, H., Ronkanen, A. K., Parviainen, M., Tolvanen, A., Koivusalo, H., and Nieminen, M.: Restoration of nutrient-rich forestry-drained peatlands poses a risk for high exports of dissolved organic carbon, nitrogen, and phosphorus, Sci. Total Environ., 586, 858–869, https://doi.org/10.1016/j.scitotenv.2017.02.065, 2017.

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., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L.: Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations, T. ASABE, 50, 885–900, https://doi.org/10.13031/2013.23153, 2007.

Mosier, T. M., Hill, D. F., and Sharp, K. V.: How much cryosphere model complexity is just right? Exploration using the conceptual cryosphere hydrology framework, The Cryosphere, 10, 2147–2171, https://doi.org/10.5194/tc-10-2147-2016, 2016.

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

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.

Rogelis, M. C., Werner, M., Obregón, N., and Wright, N.: Hydrological model assessment for flood early warning in a tropical high mountain basin, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2016-30, 2016.

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

Schönfelder, L. H., Bakken, T. H., Alfredsen, K., and Adera, A. G.: Application of HYPE in Norway, Assessment of the hydrological model HYPE as a tool to support the implementation of EU Water Framework Directive in Norway, SINTEF Energy Research, report no. 2017:00737, available at: https://sintef.brage.unit.no/sintef-xmlui/handle/11250/2499427 (last access: 22 February 2019), 2017.

Seibert, J.: On the need for benchmarks in hydrological modelling, Hydrol. Process., 15, 1063–1064, https://doi.org/10.1002/hyp.446, 2001.

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.

Siqueira, V. A., Paiva, R. C. D., Fleischmann, A. S., Fan, F. M., Ruhoff, A. L., Pontes, P. R. M., Paris, A., Calmant, S., and Collischonn, W.: Toward continental hydrologic–hydrodynamic modeling in South America, Hydrol. Earth Syst. Sci., 22, 4815–4842, https://doi.org/10.5194/hess-22-4815-2018, 2018.

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., and Bierkens, M. F. P.: PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model, Geosci. Model Dev., 11, 2429–2453, https://doi.org/10.5194/gmd-11-2429-2018, 2018.

Towner, J., Cloke, H. L., Zsoter, E., Flamig, Z., Hoch, J. M., Bazo, J., Coughlan de Perez, E., and Stephens, E. M.: Assessing the performance of global hydrological models for capturing peak river flows in the Amazon basin, Hydrol. Earth Syst. Sci., 23, 3057–3080, https://doi.org/10.5194/hess-23-3057-2019, 2019.