Dear editor, dear reviewers,

Thank you for the time taken to review our contribution. We have revised our manuscript in response to your suggestions. This document shows our responses to your comments on a point-by-point basis. For clarity, our responses are given in blue. Page and line numbers of these changes are given in our responses in the format “P[page number]L[line number]” and refer to the track-changes manuscript.

In addition, we have corrected a few typographical errors in the manuscript and changed the order of our discussion paragraphs, so that they more logically flow from the problem at hand (NSE and KGE being treated as approximately equivalent) to potential ways to deal with this problem.

Contents:

- Response to review by H. Gupta
- Response to anonymous reviewer 2
- Response to comment by John Ding

Kind regards,
On behalf of all co-authors,

Wouter Knoben

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

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

[1]Department of Civil Engineering, University of Bristol, Bristol, BS8 1TR, United Kingdom
[2]School of Geographical Science, University of Bristol, Bristol, BS8 1BF, United Kingdom
[3]Cabot Institute, University of Bristol, Bristol, United Kingdom, BS8 1UJ
*Now at University of Saskatchewan Coldwater Laboratory, Canmore, Alberta, Canada

Correspondence to: Wouter J. M. Knoben (~~w.j.m~~wouter.knoben@~~bristol.ac.uk~~usask.ca)

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 ~~often~~ viewed ~~in the literature~~ 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-√2 ≈ -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, ~~we argue that~~ modellers ~~that~~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 normalises model performance into an interpretable scale (Eq. (1)):

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

where T is the total number of time steps, $Q_{sim}(t)$ the simulated discharge at time t, $Q_{obs}(t)$ the observed discharge at time t, and $\overline{Q_{obs}}$ 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), albeit this threshold could be considered a low level of predictive skill (that is, 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:

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

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)):

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

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 $\overline{Q_{obs}}$). 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 > 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 $Q_{sim}(t) = \overline{Q_{obs}}$ for an arbitrary number of time steps, and where $\overline{Q_{obs}}$ is calculated from an arbitrary observed hydrograph. In this particular case, $\mu_{obs} = \mu_{sim}$, $\sigma_{obs} \neq 0$ but $\sigma_{sim} = 0$. Although the linear correlation between observations and simulations is formally undefined when $\sigma_{sim} = 0$, 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):

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

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

$$KGE = 1 - \sqrt{2}\,, \tag{6}$$

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

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 catchment(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 (a, 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 Figure S1 in Electronic Supplement 1 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 (that is, 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 Figures 1b, 1d and 1f, 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 Figures 1c, 1e and 1g, 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 Figure 1c, simulated flows are calculated as observed flows +0.45 mm/d (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 Figure 1e, simulated flows are calculated as observed flows +0.5 mm/d (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 mm/d (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 figures show that NSE values that are traditionally interpreted as high do not necessarily translate into high KGE values, ~~nor~~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.

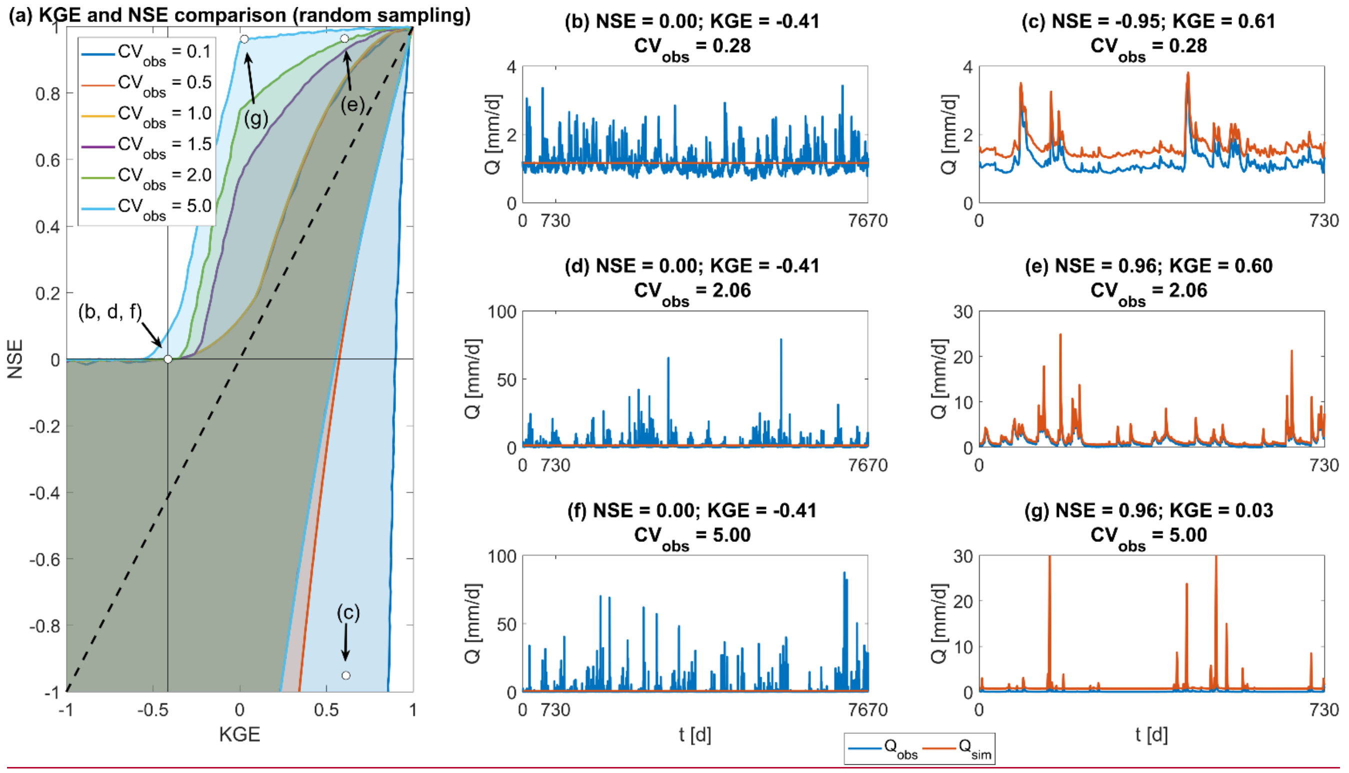


Figure 1: Overview of the relationship between NSE and KGE. (a) Comparison of KGE and NSE values based on random sampling of the r, a and b components used in KGE and NSE, using 6 different values for the Coefficient of Variation of observations (see appendix for method and separate plots of each plane). Internal axes are drawn at KGE = 1-√2 and NSE = 0. The dashed diagonal is the 1:1 line. Locations of figures 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 on every time step, increasing the bias of observations. (e) Q_{sim} is created as Q_{obs}+0.5 mm/d on every time step. (g) Q_{sim} is created as Q_{obs}+0.7 mm/d on every time step. The y-axis is capped at 30 mm/d to better visualise the difference between observations and synthetic simulations.

3.~~1~~ 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 $\overline{Q_{obs}}$, $KGE \approx -0.41$, 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.~~2~~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):

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

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 $KGE_{skill\ score} = 0$, 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 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 range [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. ~~A~~The scaled $KGE_{skill\ score}$ ~~score of~~= 0.95 cannot so readily be interpreted.

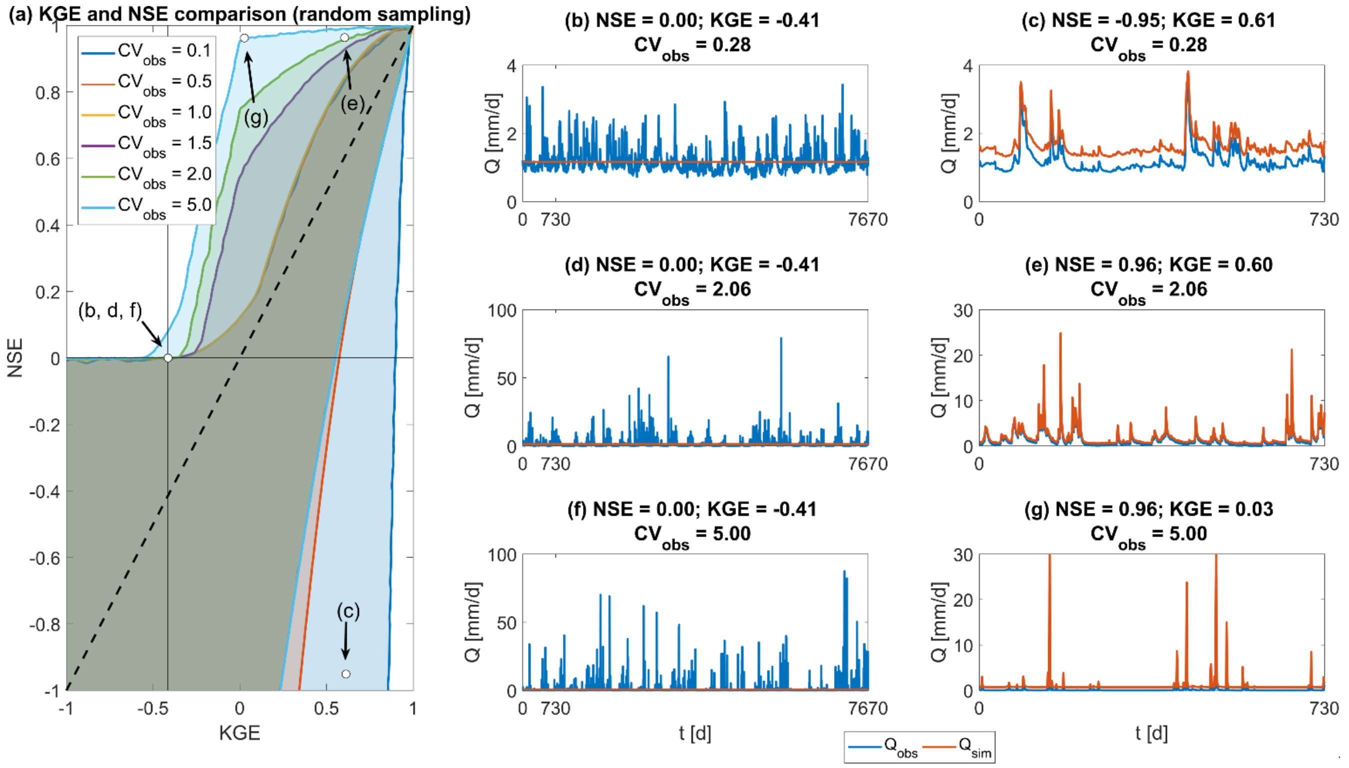
~~3.2~~3 ~~NSE and KGE values cannot be directly compared and should not be treated as approximately equivalent~~

~~Through long use, hydrologic modellers have developed intuitive assessments about which NSE values can be considered acceptable for their preferred model(s) and/or catchment(s). However, this intuition of acceptable NSE values cannot easily be mapped onto corresponding KGE values. There is no unique relationship between NSE and KGE values (Figure 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 Figure S1 in Electronic Supplement 1 for a comparison of where NSE = 0 and~~ ~~$KGE = 1 - \sqrt{2}$~~ ~~fall in the space described bey KGE’s r, a and b components).~~

~~This has important implications when NSE or KGE thresholds are used to distinguish between behavioural and non-behavioural 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 Figures 1b, 1d and 1f, 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 Figures 1c, 1e and 1g, 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 Figure 1c, simulated flows are calculated as observed flows +0.45 mm/d (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 Figure 1e, simulated flows are calculated as observed flows +0.5 mm/d (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 mm/d (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 figures show that NSE values that are traditionally seen as high do not necessarily translate into high KGE values, nor that standards of acceptability developed through extensive use of the NSE metric are 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.~~



~~**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 6 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-√2 and NSE = 0. The dashed diagonal is the 1:1 line. Locations of figures 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 on every time step, increasing the bias of observations. (e) Q_{sim} is created as Q_{obs}+0.5 mm/d on every time step. (g) Q_{sim} is created as Q_{obs}+0.7 mm/d on every time step. The y-axis is capped at 30 mm/d to better visualise the difference between observations and synthetic simulations.**~~

3.~~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, ~~E~~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 mm/d 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 ~~'good'~~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 in where models fail (e.g. Beven et al., 2014).

Regardless whether KGE or some other metric is used, t~~T~~he final step in any modelling exercise would ~~then~~ be comparing the obtained efficiency score against a certain benchmark that dictates which kind of model performance might be expected ~~in this particular catchment~~ (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 ~~bad~~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.

~~The final step in any modelling exercise would then be comparing the obtained efficiency score against a certain benchmark that dictates which kind of model performance might be expected in this particular catchment (e.g. Seibert et al., 2018) and decide whether the model is truly skilful. How these purpose-dependent benchmarks should be set is an open question to the hydrologic community.~~.

4 Conclusions

There is a tendency in current literature to interpret Kling-Gupta Efficiency (KGE) values ~~analogous~~ in the same way as~~to~~ 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 ~~(an inherent feature of NSE,~~that ~~resulting~~ results in NSE = 0 and the likely origin of this 'bad/good' model distinction~~)~~, results in $KGE = 1 - \sqrt{2}$. 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 ~~thus~~ be specific about ~~which~~ the benchmark against which they compare their model performance~~ against~~. If the mean flow is used as a KGE benchmark, all model simulations with -0.41 < KGE ≤ 1 ~~could be considered as reasonable performance~~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.

Appendix 1

The relation between possible KGE and NSE values shown in Figure 1a have been determined through random sampling of 1000000 different combinations of the components r, a and b of KGE (Eq. 2), for 6 different coefficients of variation (CV; 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 Equation 2. The corresponding NSE value for each sampled combination of r, a and b is found through:

$$NSE = 2ar - a^2 - \frac{(b-1)^2}{CV_{obs}^2}, \qquad (7)$$

Figure 2 shows the correspondence between KGE and NSE values for the 6 different CVs. Axis limits have been capped at [-1,1] for clarity. Equation 7 can be found by starting from Equation 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 $CV_{obs} = \frac{\sigma_{obs}}{\mu_{obs}}$.

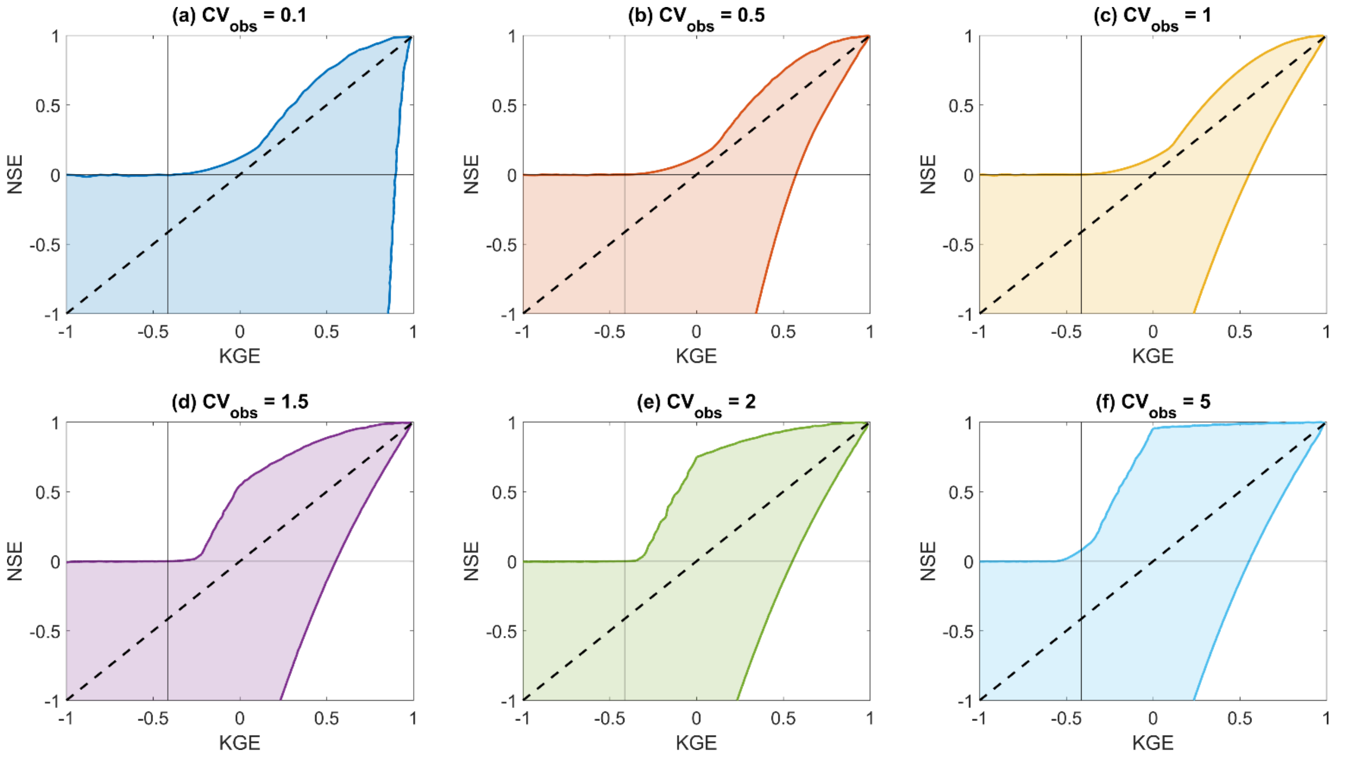


Figure 2: Correspondence between synthetic KGE and NSE values based on 1E6 random samples of components r, a and b, for different coefficients of variation (CV). Colour coding corresponds to the colours used in Figure 1a.

Acknowledgements

This work was funded by the Engineering and Physical Sciences Research Council WISE CDT, grant reference number EP/L016214/1. 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.

References

Abramowitz, G.: Towards a public, standardized, diagnostic benchmarking system for land surface models, Geosci. Model Dev., 5(3), 819–827, doi:10.5194/gmd-5-819-2012, 2012.

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(25), 4540–4554, doi: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), pp. 13–22, American Society of Civil Engineers, Liverpool, UK., 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, vol. 3, pp. 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., doi: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(12), doi: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(7), 2161–2173, doi: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(4), 2091–2115, doi: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(4), 751–763, doi: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(1–2), 80–91, doi:10.1016/j.jhydrol.2009.08.003, 2009.

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

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(March), 595–606, doi: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(7), 2069–2082, doi: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, doi: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(7), 5088–5109, doi: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(February), 858–869, doi: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(6), 2601–2614, doi: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, Trans. ASABE, 50(3), 885–900, doi: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, Cryosph., 10(5), 2147–2171, doi: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(3), 282–290, doi: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, Hydrol. Sci. J., 63(13–14), 1941–1953, doi: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., (March), 1–36, doi:10.5194/hess-2016-30, 2016.

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

Schönfelder, L. H., Bakken, T. H., Alfredsen, K. and Adera, A. G.: Application of HYPE in Norway., 2017.

Seibert, J.: On the need for benchmarks in hydrological modelling, Hydrol. Process., 15(6), 1063–1064, doi: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(8), 1120–1125, doi: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(9), 4815–4842, doi: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(6), 2429–2453, doi:10.5194/gmd-11-2429-2018, 2018.

Towner, J., Cloke, H. L., Zsoter, E., Flamig, Z. and Hoch, J. M.: Assessing the performance of global hydrological models for capturing peak river flows in the Amazon Basin, Hydrol. Earth Syst. Sci. Discuss., (February), 1–37, doi:10.5194/hess-2019-44, 2019.

Review 1 – Goshin Gupta

Review of HESS Technical note: "*Inherent benchmark or not? Comparing Nash- Sutcliffe and Kling-Gupta efficiency scores*", by Wouter J, M Knoben, JE Freer and RA Woods Review Provided by Hoshin Gupta (23rd July 2019)

Summary of the Paper:

The paper makes perhaps three main points:

Main Point Number (1): On Use of the "Mean Flow Benchmark" to interpret NSE and KGE

- The NSE normalizes model performance to an interpretable scale such that 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.
- NSE = 0 is regularly used as a benchmark to distinguish '*good*' and '*bad*' models, although this threshold could be considered a low level of predictive skill and is also a relatively arbitrary choice.
- KGE addresses several shortcomings in NSE and is increasingly used for model calibration and evaluation. Like NSE, KGE = 1 indicates perfect agreement between simulations and observations.
- Some users have tried to assign a similar scale/threshold as with NSE to be used in interpretation of KGE scores. Many authors use positive KGE values as indicative of '*good*' model performance, and negative KGE values as indicative of '*bad*' performance.
- However, this paper shows that placing the threshold for '*good*' model performance at KGE = 0 is generally correct (i.e., positive KGE values do indicate improvements upon the mean flow benchmark) but not complete. In fact, *negative KGE values do not necessarily indicate a model that performs worse than the mean flow benchmark*. The authors 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.
- Mathematically, if the model simulations of the system responses are in fact constant over time and equal to the mean of the observed flows (the mean flow benchmark), we actually have KGE ≈ -0.41.

Main Point Number (2): On the Need to Explicitly Consider Benchmark Performance

- NSE and KGE values cannot be directly compared and should not be treated as approximately equivalent. There is no unique relationship between NSE and KGE values and where NSE values fall in the KGE component space depends in part on the coefficient of variation (CV) of the observations.
- NSE values that are traditionally seen as high do not necessarily translate into high KGE values. 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.
- Whereas NSE has an inherent benchmark in the form of the mean flow, this benchmark is not inherent in the definition of KGE, which is instead an expression of distance away from the point of ideal model performance in the space described by its three components.
- There is no direct reason to use the "*mean flow*" as a benchmark over other options.

- Because KGE has no inherent benchmark value to enable a distinction between '*good'* and '*bad'* models, *modelers using KGE must be explicit about the benchmark model or value they use to compare the performance of their model against*.
- By choosing the mean flow as a benchmark to distinguish between '*good'* and '*bad'* models, practitioners limit themselves in the models and/or parameter sets they consider in a given study, without rational justification.

Main Point Number (3): On the Need to Recognize that Metrics and Benchmarks are Purpose- Dependent

- There is no single perfect model performance metric that is suitable for every study purpose. Indeed, global metrics that lump complex model behaviour and residual errors into a single value are not useful for exploring model deficiencies and diagnostics regarding how models fail or lack certain processes.
- In the choice of metrics, modellers should make conscious and well-founded choices 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.
- When using KGE, 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.
- This purpose-dependent score should then be compared against a purpose-dependent benchmark to determine whether the model can be considered '*good'*.
- How these purpose-dependent benchmarks should be set is an open question to the hydrologic community.

My Review Remarks:

[1] I thoroughly enjoyed reading this Technical Note contribution by *Wouter*, *Knoben*, *Freer* and *Woods*, and I thank them for (re)raising some very important issues, and for their new/original contribution regarding the value that the KGE criterion takes on when using the mean flow as a benchmark. As such, I have no critique per se to offer regarding this paper, and compliment the authors on an excellent contribution to the literature.

Thank you for these kind words.

[2] Instead I would like to focus on some interesting points raised by this work. This review opportunity allows me to take the liberty of reminding the readers of some interesting points that were previously raised in *Schaefli and Gupta (2007)* and *Gupta et al (2009)*, that the authors allude to, *but which perhaps could be strengthened by the authors of the current work in their presentation*. Text between quotes is reproduced from the original papers.

We have made changes to the manuscript in order to strengthen our message like you suggest. Changes are detailed in response to your remaining comments.

[3] Beginning first with *Schaefli and Gupta (2007)*, that paper was about benchmarking. In it, we discussed the fact that the process by which anyone assesses and communicates model performance evaluation is of primary importance, and that "*the basic 'rule' is that every modelling result should be put into context, for example, by indicating the model performance using appropriate indicators, and by highlighting potential sources of uncertainty*".

We have added a sentence to the introduction to emphasize that benchmarks provide context for model performance (P3L1):

"However, using such a benchmark provides context for assessing model performance (*Schaefli and Gupta (2007)*)."

[4] We pointed out therein (as have others before and after us) that: a) the *"NSE value, while a convenient and normalized measure of model performance does not provide a reliable basis for comparing the results of different case studies*" b) the "*use of the mean observed value as a reference can be a very poor predictor (e.g. for strongly seasonal time series), or a relatively good predictor (e.g. for time series that are essentially fluctuations around a relatively constant mean value)*". For example, "*In the case of strongly seasonal time series, a model that only explains the seasonality but fails to reproduce any smaller time scale fluctuations will report a good NSE value; for predictions at the daily time step, this (high) value will be misleading. In contrast, if the model is intended to simulate the fluctuations around a relatively constant mean value, it can only achieve high NSE values if it explains the small time-scale fluctuations*".

We have added a sentence to the introduction to highlight the weakness of NSE of being not comparable between different flow regimes (P2L34, addition in bold):

"albeit this threshold could be considered a low level of predictive skill **(that is, 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 …"

[5] Therefore, the definition of an appropriate benchmark model is particularly important … to properly communicate how good a model really is, it is necessary to establish an appropriate reference (or benchmark model) for a given case study and a given modelling time step. In that paper we mention some examples, including: a) the interannual mean value for every calendar day proposed by *Garrick et al. (1978)* for systems having strong but relatively constant seasonality b) a simple adjusted precipitation benchmark (APB) where the rainfall is scaled to match the mean discharge and shifted in time by some optimum lag that reflects the time of concentration of the basin, and c) a smoothed version of the APB where a simple dispersion process (moving average) is added to adjust the smoothness of the scaled-down and translated precipitation to match the smoothness of the observed discharge, for example by maximizing the correlation between the adjusted precipitation and the observed flow *(Morin et al., 2002)*. Of course, many other possible benchmarks can be conceived, such as "*persistence*" (the next time steps' simulated flow is the same as the current time step's observed flow), some kind of linear or non-linear extrapolation into the future, and some kind of data-based time-series analytical model projection such as can be constructed by ARMAX or ANN methods.

We already provided references to some of these possible other benchmarks in our discussion (P5L14:"… 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)."). Comparison of possible benchmarking options is not the focus of our paper and we have therefore chosen not to specifically mention what these alternatives are.

[6] In the conclusions to *Schaefli and Gupta (2007)*, we argued that the definition of an appropriate baseline for model performance, and in particular, for measures such as NSE (and by extension, KGE or any other model performance measure), should become part of the 'best practices' in hydrologic modelling, that "*Every modelling study should explain and justify the choice of benchmark*", and that "*the benchmark should fulfill 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*".

This is a good and succinct description of why benchmarks are needed and we have quoted it verbatim in our revision (P5L17):

"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.'*""

[7] Moving next to *Gupta et al (2009)*, we discussed the fact that the NSE, which is a dimensionless mathematical normalization of the mean squared error (MSE) criterion can be viewed as a classic skill score (*Murphy, 1988*), where 'skill' is interpreted as the comparative ability of a model with regards to a baseline 'model'. Further, as shown by *Murphy (1988)* and *Weglarczyk (1998)*, it is possible to decompose the NSE criterion into components (correlation, conditional bias, and unconditional bias) that facilitates a better understanding of what is causing a particular model performance to be 'good' or 'bad', while providing insight into possible trade-offs between the different components.

We have modified our introduction of the KGE metric to be clearer about its origin and also to include two concerns you mention further in this review, namely that KGE is not a perfect metric by itself and that other options could (should?) be explored (P3L3, additions in bold):

"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:"

[8] Our own particular diagnostic decomposition of NSE (and hence MSE) was developed in the context of our interest in hydrological modelling where, as we showed, interactions among these components (correlation, mean bias, and variance bias) can cause problems during model calibration – possibly leading to parameter estimates that are associated with large volume balance errors and/or underestimation of the variability in the flows. Further, we pointed out that many different combinations of the three components can result in the same overall value for NSE, leading to considerable ambiguity in the comparative evaluation of alternative model hypotheses.

We have added this notion about ambiguity of the overall NSE/KGE value to the text (P4L15, addition in bold):

"… see animated Figure S1 in Electronic Supplement 1 for a comparison of where NSE = 0 and KGE=1-√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**"

[9] Importantly, we also pointed out that, rather than trying to come up with a 'corrected' version of the NSE criterion, *the whole calibration problem can instead be viewed from the multi-objective perspective* (see e.g., *Gupta et al., 1998*), by focusing on the correlation, variability error and bias error as separate criteria to be optimized. *When we do so, if a compromise solution is desired, we can use the solution provided by the KGE or one of its alternatively weighted variants*.

We have added the mention of calibration as a multi-objective problem to the text (P7L20, additions in bold):

"… 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**."

[10] We presented some comparative experimental results that show that when optimizing on KGE, there is a strong correlation between the values obtained for the KGE and NSE criteria, but when optimizing on NSE, the correlation between the values obtained for NSE and KGE is lower due to the fact that optimization on KGE strongly controls the values that the mean and variance ratio components can achieve, whereas optimization on NSE constrains these components only weakly. Overall, the use of KGE instead of NSE for model calibration tends to improve the bias and variability measures considerably while only slightly decreasing the correlation.

Our current Figure 1 shows that there is indeed some correlation between NSE and KGE values, but that there is large scatter in both directions. We have not changed the text in response to this comment.

[11] Finally, we pointed out that the NSE/MSE or KGE performance metric decomposition relates to the idea of diagnostic model evaluation, as proposed by *Gupta et al. (2008)*, *which is to move beyond aggregate measures of model performance that are primarily statistical in meaning, towards the use of (multiple) measures and signature plots that are selected for their ability to provide hydrological interpretation*. While the theoretical development behind the KGE provides one simple, statistically founded approach to the development of a strategy for diagnostic evaluation and calibration of a model, *we also pointed out that all other statistical properties beyond the mean and standard deviation (which are two long-term statistics of the data), such as timing of the peaks, and shapes of the rising limbs and the recessions of the hydrograph (i.e. autocorrelation structures), are lumped into the (linear) correlation coefficient as an aggregate measure*.

See below.

[12] We therefore suggested that a logical next step would be to consider other relevant diagnostic properties (such as for example, different aspects of flow timing and shape), but left those considerations are left for future work. For example, although not mentioned explicitly in *Gupta et al (2009)*, there is no reason that other (statistical or otherwise) aspects of model performance, such as "*skewness*", or "*particular quantiles*" etc., should not be integrated into the basis for model performance evaluation and, if desired, built into a "*KGE-like*" metric.

See below.

[13] However, the *explicitly stated purpose* of the *Gupta et al (2009)* study *was NOT to design an improved measure of model performance*, but instead: a) to show clearly that there are systematic problems inherent with any optimization that is based on mean squared errors (such as NSE), b) that "*the alternative criterion KGE was simply used for illustration purposes*" (many different alternative criteria would also be sensible), and c) that "*Ultimately the decision to accept or reject a model must be made by an expert hydrologist, where such a decision is best based in a multiple-criteria framework*", where tracking the mean bias, variance bias and correlation (and other possible) components can help.

See below.

Concluding Remarks:

[14] With this context, it would actually be useful for the community to strategically move beyond the use of single metrics for model performance assessment (and/or selection), whether NSE or KGE or any other that might be conceived, and to follow the spirit of *Gupta et al (2008)* by designing some reasonable and rational basis for selecting "*sets*" of metrics that provide meaningful diagnostic evaluation of a model.

We have combined the previous four remarks into a single new discussion paragraph (P8L8):

"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). In fact, 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 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 in where models fail (e.g. Beven et al., 2014)."

[15] As pointed out by the current authors, to be meaningful, any such metrics should be accompanied by meaningful benchmarks. To be meaningful, these benchmarks should *not* be specified in an ad-hoc manner (such as NSE > 0.5 etc.) but should have some meaningful theoretical basis that conveys useful information to the decision maker.

See below.

[16] Indeed, I have often been contacted by researchers asking for some "*threshold*" values to use with KGE in their studies, and have always responded by discouraging such a practice and instead encouraging the use of the individual diagnostic components of KGE (and others that might be imagined) and setting associated thresholds using some meaningful basis.

See below.

[17] I do understand that, when performing studies involving large samples of data and/or many models, there is a tendency to want to use simple "*aggregate*" metrics in order to select or focus on a sub-set of "*good*" or "*poor*" models. However, there is arguably little to be gained by doing so by following the (arguably lazy) approach of using an aggregate metric that is not meaningfully interpretable.

See below.

[18] I sincerely hope that this current authored contribution will help to move the bulk of the community of hydrologic practitioners in the direction of using a more informative, and powerful, *diagnostic* (and necessarily multi-criteria) basis for model evaluation *that points to the nature of model deficiencies* and therefore to the modeling issues that need fixing.

See below.

[19] It might be helpful therefore, for the current authors to make some stronger arguments/comments in this direction, to encourage movement beyond the use of NSE and/or KGE, and thereby to a more powerful and robust approach to model assessment, as has been (slowly) pursued the case in some closely related communities (*Abramowitz 2012*).

We have combined the previous five remarks into a single rewrite of our concluding paragraph of the discussion section (P8L20):

"Regardless whether KGE or some other metric is used, the final step in any modelling exercise would then be comparing the obtained efficiency score against a certain benchmark that dictates which kind of model performance might be expected in this particular catchment (e.g. Seibert et al., 2018) and decide whether the model is truly skillful. These benchmarks should not be specified in an ad-hoc manner (e.g. our earlier example where the thresholds are 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."

We have added a final sentence to the conclusions that reflects the changes made above:

"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 efficiency metrics and benchmarks that allows for more robust model adequacy assessment."

We have added a final sentence to the abstract stating the same (P2L15, changes in bold):

"Therefore, we argue that 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 efficiency metrics and benchmarks that allows for more robust model adequacy assessment.**"

Reviewer 2

Summary:

The technical note provides interesting discussions on an interpretation of two metrics widely used in hydrologic community: NSE and KGE. First, the author reminds the readers that NSE is the metrics that quantify the performance compare to observed mean flow benchmark (NSE=0 indicates model performance is equivalent to this benchmark). The authors then state that there are many past studies that used KGE=0 as a threshold between bad and good model performance, same as NSE threshold. The authors point out KGE=0 does not hold the same meaning as NSE=0, and analytically show that KGE > -0.41 indicates that the model performs better than observed mean flow (if a modeler compares the model to mean flow using KGE). The authors made a direct comparison between NSE and KGE by random sampling of each KGE component and corresponding NSE, showing there is no unique relationship between two metrics, but their range of NSE value given a KGE partly depends on Coefficient of variation of the observed flow, indicating NSE and KGE cannot be directly compared. Finally, the authors that single, aggregated metrics like NSE and KGE might be misleading if the modeler looks for a specific model application (i.e., flood forecast need accuracy of high flow), and the modelers need to look more targeted metrics related to the application.

Comment:

I agree on all the major statements made in this technical note. I think one Figure presented in the note is unique contribution. It is similar to Fig 6d Gupta et al., 2009, but is expanded version and generated in the different context. I think this is very informative article, and great particularly for hydrologic practitioners who tend to quickly and intuitively evaluate the model with either NSE or KGE. I did not find any corrections/suggestions I can offer and therefor I recommend publish as is.

Thank you for your kind words. We appreciate you taking the time to read this manuscript and providing us with this review.

Interactive comment by John Ding

Equating the NSE and KGE scores

The authors raise an interesting question of whether or not the mean observed flow is an inherent benchmark of the NSE and KGE criteria. The mean flow is a base value intended by Nash and Sutcliffe (1970) to scale their NSE score to between 0 and 1. Corresponding KGE scores are -0.41 and 1 (Page 3, Line 10). Rescaling the KGE criterion to (KGE+0.41)/1.41 would produce a 0 to 1 scale.

From our initial online response:

We agree with your comment that KGE can be rescaled so that the KGE score of the mean flow equals 0. Both Feyera et al (2018) and Towner et al (2019) use a generalized scaled KGE as a skill score metric *[author's note: our thanks to Shaun Harrigan for pointing this out]*:

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

This could potentially be of use for clearer communication of whether any model's KGE score exceeds the benchmark (i.e. all positive scores of $KGE_{\text{skill score}}$) or not (i.e. all negative scores on $KGE_{\text{skill score}}$).

However, scaling the KGE metric might introduce a different communication issue. In absolute terms, it seems clear that improving on $KGE_{\text{benchmark}}$ = 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 range [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 $KGE_{\text{skill score}}$ of 0.5 might indicate no real improvement in practical terms. In cases where the benchmark constitutes a poor simulation, a $KGE_{\text{skill score}}$ of 0.5 might indicate a large improvement through using the model.

Similarly, scaling the metric might also reduce the ease of communication about model deficiencies. It is generally difficult to interpret any score above the benchmark score but below the perfect simulation (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 (assuming they are also reported). A score of KGE = 0.95 with r = 1, a = 1 and b = 1.05 indicates simulations with 5% bias. A scaled KGE score of 0.95 cannot so readily be interpreted.

Therefore, we think that a scaled metric could be of use in some cases (the clear meaning of positive and negative values is useful) but also has some drawbacks: a scaled metric is not necessarily a more efficient way of communicating model performance (because still two values must be reported for proper context) and scaling also reduces the ease with which individual KGE components can be interpreted in terms of simulation deficiencies. We will consider adding these thoughts to the discussion section in our manuscript.

We have added these considerations in a condensed way to the manuscript in a new section in the discussion (P5L25):

"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):

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

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 $KGE_{\text{skill score}} = 0$, 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 that improving on $KGE_{\text{benchmark}} = 0.99$ 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 range [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 $KGE_{\text{skill score}}$ of 0.5 might indicate no real improvement in practical terms. In cases where the benchmark constitutes a poor simulation, a $KGE_{\text{skill score}}$ 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 $KGE_{\text{skill score}} = 0.95$ cannot so readily be interpreted."

While worth searching for "a single perfect (hydrologic) model performance metric" (Page 4, Line 10), equally important if not more, in my opinion, is finding an alternate "starter" model to the mean flow one, the "no model" one in NSE. This will be a new benchmark or baseline against which the performances of other hydrologic models are to be measured. One of the "least skill(ful)" ones is a one–step linear extrapolation model of the observed flows. The predicted or forecast flow by extrapolation is: Qfore(t) =Qobs(t−1) + [Qobs(t−1)−Qobs(t−2)]. This is a simplest autoregressive model of order 2. It has been used on its own, i.e., outside the NSE, as a river forecast model. The NSE criterion may be modified by substituting the mean observed flow term, Qobs,in Equation (1), by the forecast flow. See Mizukami et al. (2019) cited by the authors for my previous comment on this (SC1 therein), the deficiency of the extrapolation model included.

Similar to our response to reviewer 1, we do not consider an overview of possible benchmark metrics within scope of this paper. We have therefore chosen not to explicitly mention autoregressive models in the text but have included a reference to this comment so that this may become part of future work on the topic (P5L14, addition in bold):

"… 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)."

References Cited:

Abramowitz G, 2012, Towards a public, standardized, diagnostic benchmarking system for land surface models, Geoscientific Model Development, vol. 5, pp. 819 - 827, http://dx.doi.org/10.5194/gmd-5-819-2012

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., doi:https://doi.org/10.5194/hess-2019-327-SC1, 2019

Garrick M, Cumane C, Nash JE. 1978. A criterion of efficiency for rainfall-runoff models. Journal of Hydrology 36: 375–381.

Gupta HV, S Sorooshian and PO Yapo (1998), Towards Improved Calibration of Hydrologic Models: Multiple and Non-Commensurable Measures of Information, Water Resources Research, Vol. 34, No. 4, pp. 751-763

Gupta HV, T Wagener and YQ Liu (2008), Reconciling Theory with Observations: Towards a Diagnostic Approach to Model Evaluation, Hydrological Processes, Vol. 22 (18), pp. 3802-3813, DOI: 10.1002/hyp.6989.

Gupta HV, H Kling, KK Yilmaz and GF Martinez-Baquero (2009), Decomposition of the Mean Squared Error & NSE Performance Criteria: Implications for Improving Hydrological Modelling, Journal of Hydrology, Vol. 377, pp. 80-91, doi: 10.1016/j.jhydrol.2009.08.003.

Morin E, Georgakakos KP, Shamir U, Garti R, Enzel Y. 2002. Objective, observations-based, automatic estimation of the catchment response timescale. Water Resources Research 38: 1212, DOI: 10・1029/2001WR000808.

Murphy A (1988), Skill scores based on the mean square error and their relationships to the correlation coefficient. Monthly Weather Review 116: 2417-2424

Schaefli B and HV Gupta (2007), Do Nash values have value? Hydrological Processes, 21(15), 2075-2080, simultaneously published online as Invited Commentary in Hydrologic Processes (HP Today), Wiley InterScience, doi: 10.1002/hyp.6825

Weglarczyk S (1998), The interdependence and applicability of some statistical quality measures for hydrological models. Journal of Hydrology 206: 98-103