https://doi.org/10.5194/hess-2019-327
Preprint. Discussion started: 1 July 2019

Hydrology and Earth System Sciences
Discussions

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

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

[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

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

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 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, we argue that modellers 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.

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)):

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

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 and is also a relatively arbitrary choice (for example, Moriasi et al., 2007, define several different NSE thresholds for different qualitative levels of model performance).

The Kling-Gupta Efficiency (KGE, Eq. (2), Gupta et al., 2009) addresses several shortcomings in NSE and is increasingly used for model calibration and evaluation:

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

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)):

https://doi.org/10.5194/hess-2019-327
Preprint. Discussion started: 1 July 2019

Hydrology and Earth System Sciences
Discussions

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

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 $\overline{Q_{obs}}$). 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):

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

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

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

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

https://doi.org/10.5194/hess-2019-327
Preprint. Discussion started: 1 July 2019

Hydrology and Earth System Sciences
Discussions

3 Consequences

3.1 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. 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.

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 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 be 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.

https://doi.org/10.5194/hess-2019-327
Preprint. Discussion started: 1 July 2019

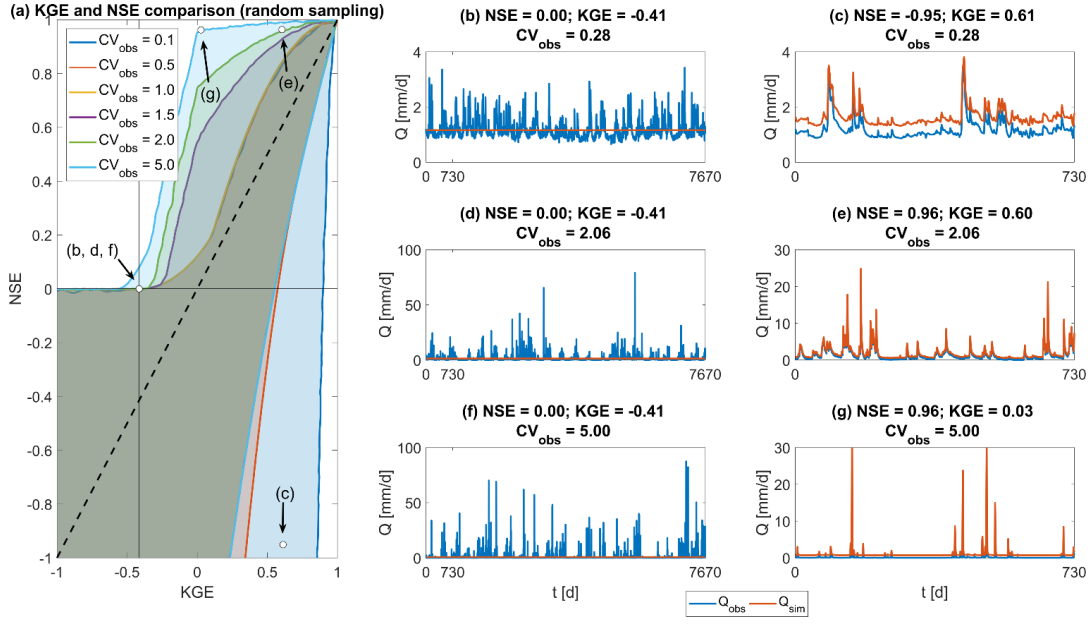

Figure 1: Overview of the relationship between NSE and KGE. (a) Comparison of KGE and NSE values based on random sampling of the r, a and b components used in KGE and NSE, using 6 different values for the Coefficient of Variation of observations (see appendix for method and separate plots of each plane). 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 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. 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'.

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).

https://doi.org/10.5194/hess-2019-327
Preprint. Discussion started: 1 July 2019

Hydrology and Earth System Sciences
Discussions

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. 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 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, resulting 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 must thus be specific about which benchmark they compare their model performance against. If the mean flow is used as a KGE benchmark, all model simulations with -0.41 < KGE $\leq$ 1 could be considered as reasonable performance. Furthermore, modellers must take care to not let their interpretation of KGE values be (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.

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:

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

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 $CV_{obs} = \frac{\sigma_{obs}}{\mu_{obs}}$.

https://doi.org/10.5194/hess-2019-327
Preprint. Discussion started: 1 July 2019

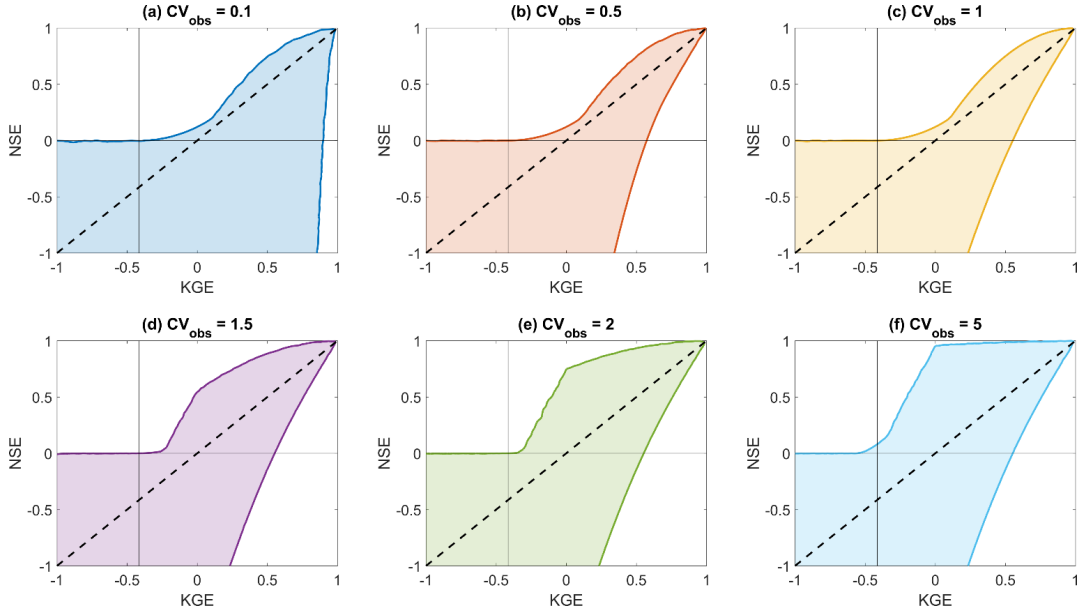

Figure 2: Correspondence between synthetic KGE and NSE values based on 1E6 random samples of components r, a and b, for different coefficients of variation (CV). Colour coding corresponds to the colours used in Figure 1a.

References

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.

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.

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.

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., 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.

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.

https://doi.org/10.5194/hess-2019-327
Preprint. Discussion started: 1 July 2019

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.