Dear Dr John Ding,

Thank you for your comment on our manuscript and the reference to your earlier comments on this topic during the discussion of Mizukami et al (2019).

Communication through scaled metrics

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 $KGE_{skill\ score}$) or not (i.e. all negative scores on $KGE_{skill\ score}$).

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.

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.

Which benchmarks should be used?

Communicating model performance in comparison to benchmark values is a separate issue from *which* benchmark should be used, which is the focus of the second part of your comment. We agree that the traditional mean flow benchmark is not a particularly taxing baseline in many (although not all) cases (as mentioned on page 1, lines 28-29) and that it is worthwhile to carefully consider alternative options (page 1, lines 28-30; page 3, lines 5-7). We already provide references to several other possible options for benchmarking (Schaefli and Gupta, 2007; Seibert, 2001; Seibert et al.,

2018) and will add your suggestion of a linear extrapolation model to this list. We intend to change the text to emphasize that (an) appropriate benchmark(s) should be chosen as part of the experimental design, such that model scores that outperform the benchmark are a clear reflection of the model being closer to the modelling aim than the benchmark was.

On behalf of all authors,

Kind regards,

Wouter Knoben

References

Feyera A. Hirpa, Peter Salamon, Hylke E. Beck, Valerio Lorini, Lorenzo Alfieri, Ervin Zsoter, Simon J. Dadson (2018). Calibration of the Global Flood Awareness System (GloFAS) using daily streamflow data. Journal of Hydrology, Volume 566, 595-606, https://doi.org/10.1016/j.jhydrol.2018.09.052

Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V., & Kumar, R. (2019). On the choice of calibration metrics for "high-flow" estimation using hydrologic models. Hydrology and Earth System Sciences, 23(6), 2601–2614. https://doi.org/10.5194/hess-23-2601-2019

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

Seibert, J. (2001). On the need for benchmarks in hydrological modelling. Hydrol. Process., 15(6), 1063–1064, doi:10.1002/hyp.446

Seibert, J., Vis, M. J. P., Lewis, E. and van Meerveld, H. J. (2018). Upper and lower benchmarks in hydrological modelling. Hydrol. Process., 32(8), 1120–1125, doi:10.1002/hyp.11476

Towner, J., Cloke, H. L., Zsoter, E., Flamig, Z., Hoch, J. M., Bazo, J., Coughlan de Perez, E., and Stephens, E. M. (in press). Assessing the performance of global hydrological models for capturing peak river flows in the Amazon Basin. Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-44

Dear Dr John Ding,

Thank you for your comment on our manuscript and the reference to your earlier comments on this topic during the discussion of Mizukami et al (2019).

\textbf{Communication through scaled metrics}

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 \emph{[author's note: our thanks to Shaun Harrigan for pointing this out]}:

\begin{equation}

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

\end{equation}

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 KGEskill score) or not (i.e. all negative scores on KGEskill score).

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.

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 is not necessarily a more efficient way of communicating model performance (because still two values must be reported for proper context). We will consider adding these thoughts to the discussion section in our manuscript.

\textbf{Which benchmarks should be used?}

Communicating model performance in comparison to benchmark values is a separate issue from which benchmark should be used, which is the focus of the second part of your comment. We agree that the traditional mean flow benchmark is not a particularly taxing baseline in many (although not all) cases and (as mentioned on page 1, lines 28-29) and that it is worthwhile to carefully consider alternative options (page 1, lines 28-30; page 3, lines 5-7). We already provide references to several other possible options for benchmarking (Schaefli and Gupta, 2007; Seibert, 2001; Seibert et al., 2018) and will add your suggestion of a linear extrapolation model to this list.

On behalf of all authors,

Kind regards,

Wouter Knoben

\textbf{References}

Feyera A. Hirpa, Peter Salamon, Hylke E. Beck, Valerio Lorini, Lorenzo Alfieri, Ervin Zsoter, Simon J. Dadson (2018). Calibration of the Global Flood Awareness System (GloFAS) using daily streamflow data. Journal of Hydrology, Volume 566, 595-606, https://doi.org/10.1016/j.jhydrol.2018.09.052

Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V., & Kumar, R. (2019). On the choice of calibration metrics for “high-flow” estimation using hydrologic models. Hydrology and Earth System Sciences, 23(6), 2601–2614. https://doi.org/10.5194/hess-23-2601-2019

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

Seibert, J. (2001). On the need for benchmarks in hydrological modelling. Hydrol. Process., 15(6), 1063–1064, doi:10.1002/hyp.446

Seibert, J., Vis, M. J. P., Lewis, E. and van Meerveld, H. J. (2018). Upper and lower benchmarks in hydrological modelling. Hydrol. Process., 32(8), 1120–1125, doi:10.1002/hyp.11476

Towner, J., Cloke, H. L., Zsoter, E., Flamig, Z., Hoch, J. M., Bazo, J., Coughlan de Perez, E., and Stephens, E. M. (in press). Assessing the performance of global hydrological models for capturing peak river flows in the Amazon Basin. Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-44