hess-2022-380: RC1

Having carried out hydrological modelling for the past 30 years it is interesting to see how the use of different performance criteria has developed. The Nash-Sutcliffe Efficiency (NSE) criteria has been the main criteria for flooding issues (and very often a general criteria on the overall performance of a model) for a long time. It has a number of well documented drawbacks but has the advantage of the values being widely understood. Kling-Gupta Efficiency (KGE) and its variants have become more popular recently but my feeling is that it is less well understood and some of the issues associated with its use have not been fully explored.

This paper is a useful addition to the subject of different performance criteria as it clearly shows that in the KGE there can be counterbalancing errors (i.e sometimes an over estimation and sometimes an under estimation of discharge) which produce a higher value without there being an improvement in the model. Whereas these counterbalancing error do not occur for the NSE. The authors summarize the issue and their contribution very well when they say "The aim of this paper is primarily to raise awareness among modellers. Performance criteria generally comprise several aspects of the characteristics of a model into a single value, which can lead to an inaccurate assessment of said aspects. Ultimately, all criteria have their flaws and should be carefully selected with regards to the aim of the model"

The paper is well written and presented. There is a good summary of the current state in the use of different performance criteria in hydrological models. The use of both a sythentic time series and a real case study gives more confidence in the issue of these counterbalancing errors. Overall, it is a good bit of work with a clear conclusion reached. I am happy to accept the paper with minor revisions

Thank you very much for you for your positive feedback, and also for your careful review of the manuscript. We are pleased that you find the paper interesting, well-constructed and meaningful.

Specific comments:

- **I wonder if it would be useful to mention some of the benchmarking studies in hydrological modelling (e.g. Seibert et al. 2018) which I feel are a useful addition with regards to the performance criteria of the models.**

Thank you for pointing out this interesting publication. We agree that it is important to mention the possible use of more relevant benchmarks in hydrological modelling. We modified the manuscript accordingly.

L30: *"Improvements can be made by working on (i) input data, [...], (iv) model calibration (Beven, 2019), and also (v) appropriate benchmarks for assessing model performance (Seibert et al., 2018)."*

L45: *"In relation to the assessment of model performance, Seibert et al. (2018) argued that the current benchmarks poorly reflect what could and should be expected of a model. They suggested to define lower and upper benchmarks based on the performance of a simple bucket-type model with few parameters, using the same data set."*

Added reference: *"Seibert, J., Vis, M. J. P., Lewis, E., and van Meerveld, H. j.: Upper and lower benchmarks in hydrological modelling, Hydrological Processes, 32, 1120–1125, https://doi.org/10.1002/hyp.11476, 2018."*

- **L98 (Equation 11). This is the Gupta(2009) equation. This is surely wrong as the last term should be minus not plus.**

Indeed, you are right, thanks for pointing this out. We corrected it.

- **L133-135. I do not understand this bit. I can see there are 361 transformation between -0.36 and 0.36 but I need not understand where the logarithmic scale comes in and how you get from here to the w values**

We wanted to study counterbalancing errors on a set of homothetic transformations ranging from roughly half to twice the discharges of the reference time series, which correspond to ω values of about 0.5 to 2. A linear sampling would have been uneven between underestimated and overestimated transformations (0.5–1 have a lower sample rate than 1–2). Sampling on the log-transformed interval allows to have an even distribution of the values below and above ω=1. As demonstrated in Eq. (1) and Eq. (2), the common logarithmic transformations (base 10) of 0.5 and 2 nearly equal -0.30103 and 0.30103, respectively; these values are equidistant to 0, which is the common logarithmic transformation of 1 (Eq. (3)).

$$\log_{10}(0.5) \approx -0.30103 \quad (1)$$
$$\log_{10}(2) \approx 0.30103 \quad (2)$$
$$\log_{10}(1) = 0 \quad (3)$$

We decided to take a slightly larger interval to ease the reading of the graphs, i.e. [-0.36, 0.36], sampled at a 0.002 step. The exponentiation in base 10 of the sampled values allows to get an even distribution of ω values around ω=1:

$$10^{-0.36} \approx 0.4365158 \quad (4)$$
$$10^{0.36} \approx 2.290868 \quad (5)$$
$$10^{0} = 1 \quad (6)$$

As $\frac{0.36}{0.002} = 180$, there are 361 transformations in total:

- 180 transformations below the ω=1 homothety in the [-0.36; -0.002] interval, with the minimum ω for -0.36 (Eq. (4))
- 180 transformations above the ω=1 homothety in the [0.002; 0.36] interval, with the maximum ω for 0.36 (Eq. (5))
- 1 transformation corresponding to the ω=1 homothety (Eq. (6))

We modified the manuscript with a better explanation of the logarithmic sampling procedure.

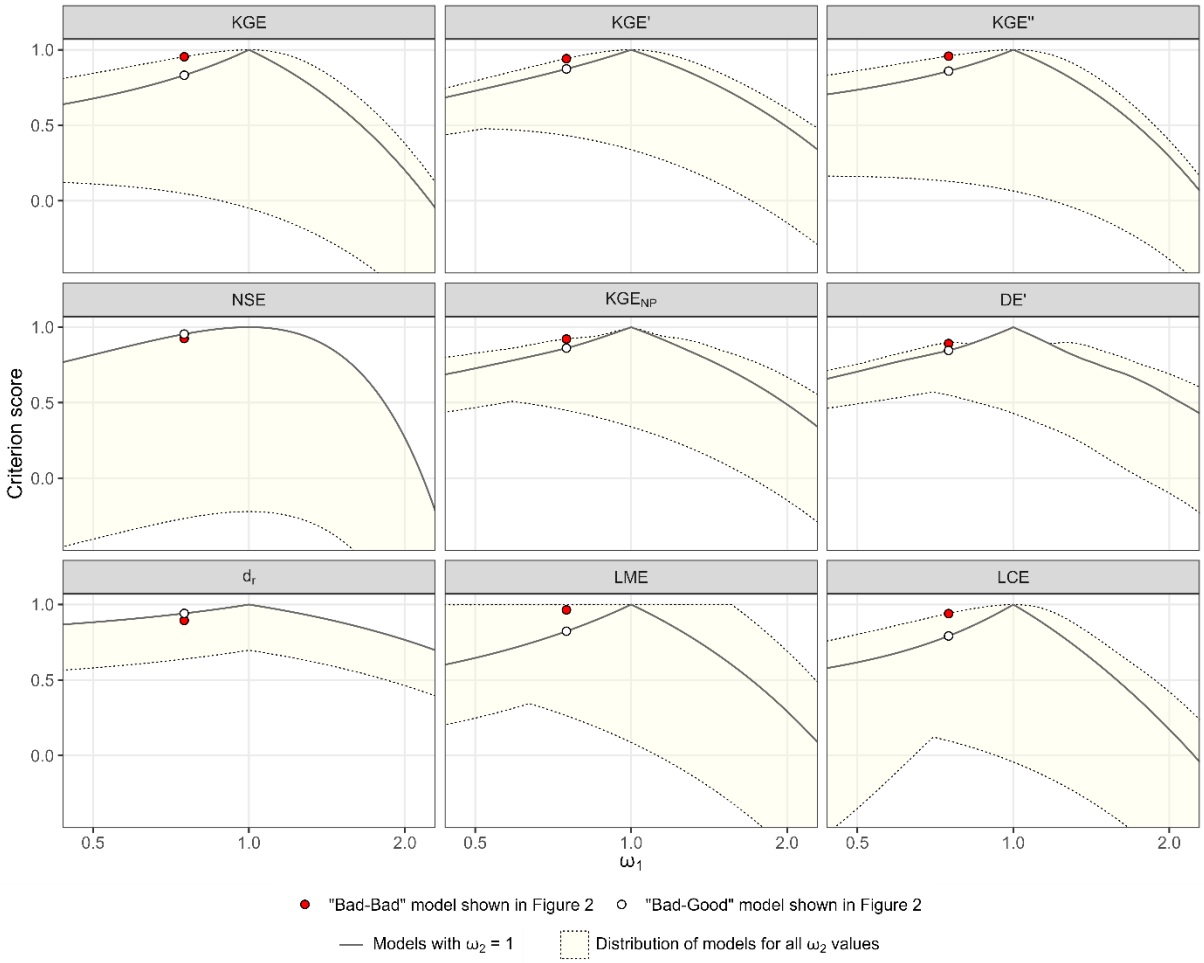
L133: *"ω values were sampled uniformly on the log-transformed interval [-0.36, 0.36] at a defined step of 0.002 to ensure a fair distribution between underestimated and overestimated transformations."*

L134: *"The exponentiation in base 10 of the sampled values results in 360 ω values evenly distributed around the* $\omega = 1$ *homothety, which corresponds to the reference time series (i.e. absence of transformation)."*

L136: *"We defined ω bounds such that the transformed peak discharge roughly ranges from half (*$\omega \approx 0.437 \approx 10^{-0.36}$*) to twice (*$\omega \approx 2.291 \approx 10^{0.36}$*) compared to the references time series."*

- **L195 (Figure 4). Would it be useful to also show the "Bad-Bad" model on this figure?**

Thank you, this is a great suggestion. It can be useful because it would show that, for some criteria, the "Bad-Bad" model is not even the most affected by counterbalancing errors.



- **Change "consisting in" to "consisting of"**
- **Maybe change "both ways" to "both sides" or "both directions"**
- **Change "succeed to reproduce" to "succeed in reproducing" or "successfully reproduce"**

Thank you for the corrections. We changed it in the revised manuscript.

- **L273-L274. "In general, the ANN model can be described as better because it is closer to the observed values in the high and low flow periods". As a hydrological modeller I agree the ANN model is better. But surely the whole point of performance criteria is to objectively decide which model is better. So how do you decide it is better when the performance criteria do not agree? There is no easy answer but I feel it is an important question that should be considered in more detail.**

You are questioning a valid point and we agree that it deserves a better and more detailed explanation. Thank you for considering the fact that the answer is not easy. We added a paragraph in the manuscript to better explain how the ANN model can be considered as better without using performance metrics (i.e. from a subjective assessment). We also improved the description of the models for the first and second flood events.

L262: *"The first flood event (February 2017) is slightly underestimated by the ANN model and highly overestimated by the bucket-type model. The second flood event (March 2017) is similarly underestimated by both models but the bucket-type model demonstrates a slightly better performance."*

L274: *"While this statement cannot be supported by performance metrics, we believe that an expert assessment based on intuition and experience is still valuable despite being intrinsically subjective. In this particular case, one can assess the main, distinctive flaws of each model: (i) the ANN model has continuous oscillations – especially on recession and low flow periods – and lacks of accuracy during recession periods; (ii) the bucket-type model highly overestimates several flood events and is inaccurate during a lot of recession and low flow periods. Figure 6b also shows that the ANN model has an overall lower bias than the bucket-type model. Hydrological models are generally used for (i) the prediction/forecast of water flood/inrush, (ii) the management of water resources, (iii) the characterisation of hydrosystems, and more recently (iv) the study of the impact of climate change on water resources. Most studies thus put the emphasis on extremes events (i.e. dry and flood periods), which in this case are more satisfactorily reproduced by the ANN model – in terms of volume estimate, timing and variability."*

- **There is no reference to Figure 7, should it be here.**

We added the reference to Figure 7 in the text.

L280: *"The visual assessment is confirmed only by a few performance criteria: the NSE, d_r and KGE_{NP} (Fig. 7a)"*

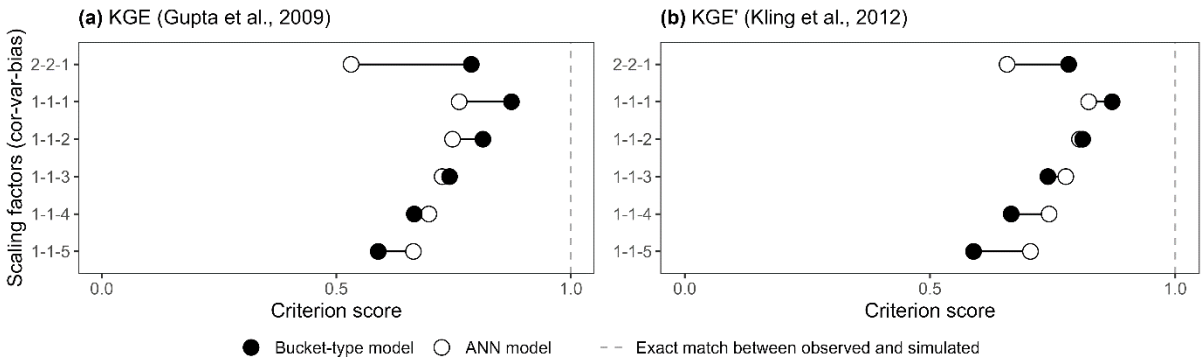
L283: *"Looking at the values of the equations' parameters (Fig. 7b), we find that [...]"*

- **L300-L303. I do not think this bit adds anything. I would remove these lines.**

You are right, this does not add a lot to the discussion, so we removed the sentences in the revised manuscript.

- **In Equation 22 the order of the parameters is alpha, beta, r. On Line 344 and subsequent lines it is r, alpha, beta. This is confusing. So when you look at (1-2-2) and you look at equation 22 everything needs to be swapped around as the 1 corresponds to r which is the last term in the equation**

Indeed, this is confusing. We changed the order of the y-axis number to be the same of the order of the equation, i.e. alpha, beta, r. The caption was edited: *"The y-axis numbers correspond to the scaling factors of the variability, bias and timing parameters, with the default being 1-1-1."*



- **Change "associated to" to "associated with"**
- **Change "include to" "include the"**

Thanks for the corrections. We changed it in the revised manuscript.