Manuscript hess-2022-380 – Responses to Reviewers

We thank the referees for their careful reading and helpful comments. Our reply is given below. The page and line numbers (in the "modification to manuscript" sections) correspond to the modifications done on the revised manuscript with changes marked.

Reviewer 1

1.1 Quoting: "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."

Response. 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.

Modification to manuscript.

- **Page 2. Line 31.** The sentence was changed into: "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)."
- **Page 2. Line 45.** A sentence was added: "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."
- The reference was added: "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."

1.2 Quoting: "L98 (Equation 11). This is the Gupta(2009) equation. This is surely wrong as the last term should be minus not plus."

Modification to manuscript. The equation was changed into: "$NSE = 2\alpha r - \alpha^2 - \beta_n{}^2$".

1.3 Quoting: "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"

Response. 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.

Modification to manuscript.

- **Page 6. Line 138.** The sentence was changed into: "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."
- **Page 6. Line 140.** The sentence was changed into: "The exponentiation in base 10 of the sampled values results in 361 ω values evenly distributed around the $\omega = 1$ homothety, which corresponds to the reference time series (i.e. absence of transformation)."
- **Page 6. Line 142.** The sentence was changed into: "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."

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

Response. 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.

Modification to manuscript. Figure 4. The point corresponding to the "Bad-Bad" model in Figure 2 was added to the figure.

1.5 Quoting: "Change "consisting in" to "consisting of""

Modification to manuscript. Page 11. Line 215. The sentence was changed into: "This is likely due to the nature of the equation consisting of 3 parameters [...]"

1.6 Quoting: "Maybe change "both ways" to "both sides" or "both directions""

Modification to manuscript. Page 12. Line 219. The sentence was changed into: "[...] show similar envelopes with a break point near the maximum transformation score in both directions around $\omega_1 = 1$."

1.7 Quoting: "Change "succeed to reproduce" to "succeed in reproducing" or "successfully reproduce""

Modification to manuscript. Page 15. Line 271. The sentence was changed into: "The models have overall good dynamics and successfully reproduce the observed discharges."

1.8 Quoting: "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."

Response. 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.

Modification to manuscript.

- **Page 15. Line 273.** The sentence was changed into: "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."
- **Page 15. Line 290.** The sentence was moved two lines earlier: "Some events are not well simulated by both models (e.g. the May 2017 flood), which may be due to uncertainties in the input data."
- **Page 15. Line 294.** A paragraph was added: "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 bucket-type model has an overall higher bias than the ANN 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 a hydrosystem, and more recently (iv) the study of the impact of climate change on water resources. Most studies thus put the emphasis on volumes, and also 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."

1.9 Quoting: "There is no reference to Figure 7, should it be here."

Modification to manuscript.

- **Page 17. Line 311.** The sentence was changed into: "The visual assessment is confirmed only by a few performance criteria: the NSE, d_1 and KGE_{NP} (Fig. 7a)."
- **Page 17. Line 315.** The sentence was changed into: "Looking at the values of the equations' parameters (Fig. 7b), we find that [...]"

1.10 Quoting: "L300-L303. I do not think this bit adds anything. I would remove these lines."

Modification to manuscript.

- The following sentences were removed from the text: **"**Figure 6b shows that there is a consistent greater or equal overestimation of the reservoir model compared to the ANN model, except for the May-June period where the difference is small and insignificant compared to the February or September events. The underestimated values are similar for both approaches, except when the reservoir model overestimates the flooding events."
- **Page 18. Line 337.** The sentence was changed into: "Interestingly, the cumulative sum of the absolute bias error between simulated and observed values (Fig. 6b) is smaller for [...]"

1.11 Quoting: "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"

Response. 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. α, β, r.

Modification to manuscript.

- **Figure 8.** The y-axis number were changed to the following order: variability, bias, timing (α-β-r for the KGE and γ-β-r for the KGE').
- **Figure 8.** The caption was changed into: "The y-axis numbers correspond to the scaling factors of the variability, bias and timing parameters, with the default being 1-1-1."
- **Page 21. Lines 378–381.** The text was modified accordingly.

1.12 Quoting: "Change "associated to" to "associated with""

Modification to manuscript. Lines 22, 62 and 401. "associated to" was changed to "associated with".

1.13 Quoting: "Change "include to" "include the""

Modification to manuscript. Page 23. Line 404. The sentence was changed into: "Recommendations also include the use of scaling factors to emphasise [...]"

Reviewer 2

2.1 Quoting: "I do understand why NSE is included as "recommended skill scores" given the context of this paper, but still not sure if it is good idea to state so because NSE has one separate issue (underestimating variability so, peak-flow is underestimated and low flow is overestimated). I would suggest stating NSE is less impacted by counter-balance error in the hydrograph, but has its own issue for the practical applications."

Response. Thank you for the relevant suggestion. We modified the text to remove the ambiguity between the recommended skill scores and the skill scores that are less impacted by counterbalancing errors. As you mentioned, NSE is less impacted by counterbalancing errors but it does not seem appropriate to recommend it because of its known issue for practical applications.

Modification to manuscript.

- **Page 19. Line 352.** The sentence was changed into: "However, they have other drawbacks that are not associated with counterbalancing errors, especially the NSE with its limitation related to variability (Gupta et al., 2009)."
- **Page 19. Line 354.** The sentence was changed into: "We thus recommend using performance criteria that are not or less prone to counterbalancing errors (d_1, KGE', KGE_{NP}, DE)."
- **Abstract.** The sentence was changed into: "We recommend using (i) performance criteria that are not or less prone to counterbalancing errors (d_1, modified KGE, non-parametric KGE, Diagnostic Efficiency), and/or [...]"

2.2 Quoting: "Section 4. In real case study, the paper use "reservoir model" for actually some bucket type, conceptual hydrologic model. I suggest avoiding using "reservoir model" because some readers (including me) are confused with reservoir "operation" model (i.e., lake model)."

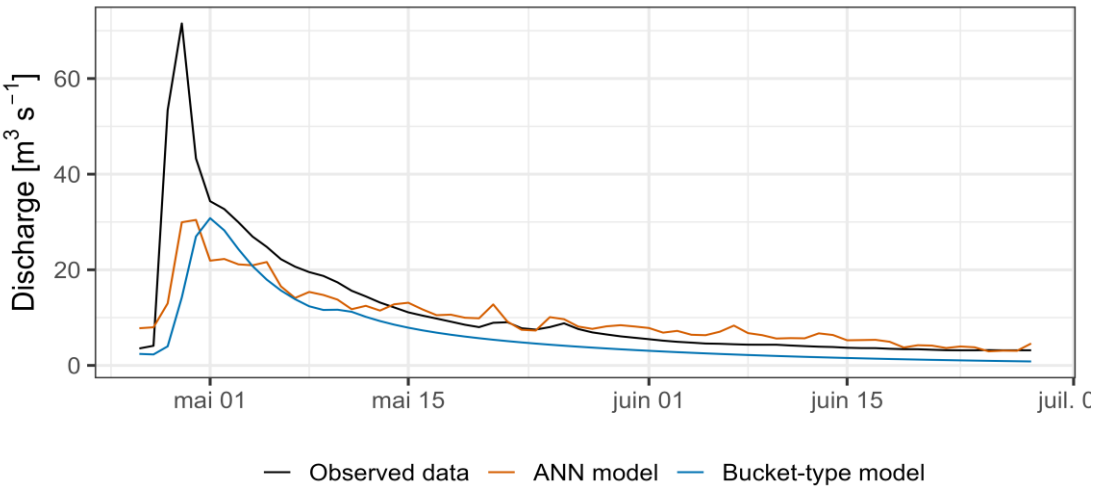
Response. Indeed, the term "reservoir model" can be confusing for some readers. Thank you for the nice suggestion.

Modification to manuscript.

- "reservoir model" was replaced with "bucket-type model" in both text and figures.
- "reservoir" was replaced with "bucket" in the text.

2.3 Quoting: "L264. The paper said "third flood event (May 2017) is better simulated by the ANN". I don't see this. Both ANN and reservoir models similarly underestimate the flows. Also, the statements after because are unclear to me."

Response. We agree that the statement of the ANN simulation being better is inappropriate for so little of a difference between the two models, which both have a poor performance on this flood event. However, we can state that the ANN simulation is *slightly* better on the timing of the flood peak ($\boldsymbol{r = 0.88}$) and the overall volumes ($\boldsymbol{\beta = 0.87}$), while the bucket-type model has a better recession coefficient and variability ($\boldsymbol{\gamma = 0.99}$) – see the figure and table below. We modified the manuscript accordingly.



Model	NSE	KGE'	γ	β	r	KGE_{NP}
ANN	0.59	0.51	0.55	0.87	0.88	0.79
Bucket-type	0.35	0.49	0.99	0.56	0.73	0.56

Modification to manuscript.

- **Page 15. Line 277.** The sentence was changed into: "The third flood event (May 2017) is poorly simulated by the models – both underestimate the flood peak – but the ANN model is more accurate in terms of timing and volume estimate, while the bucket-type model has a better recession coefficient and flow variability."

2.4 Quoting: "L269-273. I suggest using the dates to point which events are referred to."

Response. Thank you for the suggestion. We modified the paragraph so that the dates directly refer to the events.

Modification to manuscript. Page 15. Line 284. The sentence was changed into: "The small flood events are better simulated by the ANN model than the bucket-type model: (i) the ANN model simulates them satisfactorily, except for the second one (mid-April), where the simulated discharges are overestimated; (ii) the bucket-type model does not simulate the first two events at all (mid-January and mid-April) and largely overestimates the last two (early and late June), in addition to timing errors."

Community comment 1 (CC1)

3.1 Quoting: "In the last sentence of the abstract, the authors mention the use of multi-criteria framework in their recommendation. On the need to consider a particular "goodness-of-fit" metric within the multi-criteria framework, the authors could clarify on other specific requirements apart from the general condition that the performance criteria should be less or not prone to counterbalancing. Furthermore, the use of several criteria for a particular calibration can complicate the applications of automation of famous search strategies or algorithms (Onyutha 2022). It is upon this basis that a number of performance criteria which are not mathematically and statistically related tend to be formed into single metric. For instance, Kling-Gupta Efficience combines three components including measures of bias, variability and linear correlation between observed (X) and modelled (Y) series. Thus, the authors should provide more considered justification for their recommendation of the use of multi-criteria framework for calibration of hydrological models."

Response. Thank you for this relevant examination. Indeed, the explanation around the use of a multi-criteria framework is unclear. All things considered, it seems that the recommendation of using a multi-criteria framework is not appropriate in this article, as its scope is to identify and raise awareness about the problem of counterbalancing errors. The mention of the multi-criteria framework, its benefit and also relevant references are already mentioned is the introduction, so we changed the recommendations to be consistent with the findings of this study.

Modification to manuscript.

- **Abstract.** The sentence was changed into: "We recommend using (i) performance criteria that are not or less prone to counterbalancing errors (d_1, modified KGE, non-parametric KGE, Diagnostic Effiency), and/or (ii) [...]"
- **Page 19. Line 354.** The sentence was changed into: "We thus recommend using performance criteria that are not or less prone to counterbalancing errors (d_1, KGE', KGE_{NP}, DE)."
- **Page 21. Line 384.** The sentence was changed into: "This is why a multi-criteria framework can strengthen the evaluation of models and reduce the uncertainty associated with the interpretation of individual performance criteria scores."

3.2 Quoting: "Most of (if not all) the metrics used in this study rely on the assumption that X and Y are linearly related. Note that X and Y can be so highly dependent yet it may be nearly impossible to detect the dependence using classical dependence metric (Székely et al. 2007). In other words, the authors should clarify on whether the model performance results of this study may not have been affected by the said assumption.

Response. Thank you for pointing out this implicit assumption of the KGE and its variants. Although this study focuses on counterbalancing errors in widely used performance criteria and not (so much) on the correlation between X and Y, it is important to clarify whether the data linearity between X and Y is skewed – which is often the case in hydrological modelling – to better appreciate the model performance results. Note that we also included the non-parametric KGE (Pool et al., 2018), which is based on the Spearman correlation coefficient and the flow duration curve, and has no assumption of data linearity.

Modification to manuscript.

- **Page 6. Line 144.** The sentence was changed into: "Note that (i) the data linearity between simulated and observed values is verified, and (ii) ω homotheties still induces [...]"
- **Page 15. Line 292.** A sentence was added: "Also, the data linearity between simulated and observed values is slightly skewed for both models, which can affect the relevance of *r* (Barber et al., 2020)."

3.3 Quoting: "Most of the performance criteria (especially Nash Sutcliffe Efficiency NSE (Nash and Sutcliffe, 1970) and its variants) comprise some forms of the well-known coefficient of determination (R-squared) (see Onyutha, 2022). R-squared is known to have various short comings. To address these short comings, new metrics including the revised R-squared (RRS) and hydrological model skill score E (Onyutha 2022) were developed. Thus, instead of focussing on NSE and its variants, the authors should compare results of many other performance criteria such as RRS and E. Accordingly, Figure 7 and Table 1 in this manuscript can be updated. The MATLAB codes to compute RRS and E can be downloaded via https://doi.org/10.5281/zenodo.6570905 and the codes can also be found as supplementary material to the paper by Onyutha (2022)"

Response. Thank you for the suggestion of these two innovative performance criteria and the associated code. The results for the synthetic models and the real case study have been added as a figure in appendix alongside other commonly used performance criteria (RMSE, R^2, d, d_r).

Modification to manuscript.

- **Appendix A.** An appendix was added: "Appendix A: Common and recently developed performance criteria applied to the synthetic time series and the real case study"
- **Figure A1.** A figure was added in Appendix A with the following caption: "Figure A1: Score of the BB and BG transformations according to other common and recently developed performance criteria: the Root Mean Square Error (RMSE), the coefficient of determination R^2, the index of agreement d (Willmott, 1981), the refined index of agreement d_r (Willmott et al., 2012), the Onyutha efficiency E and the revised R-squared RRS (Onyutha, 2022).
- **Figure A2.** A figure was added in Appendix A with the following caption: "Figure A2: Score of the ANN and bucket-type models according to other common and recently developed performance criteria: the Root Mean Square Error (RMSE), the coefficient of determination R^2, the index of agreement d (Willmott, 1981), the refined index of agreement d_r (Willmott et al., 2012), the Onyutha efficiency E and the revised R-squared RRS (Onyutha, 2022).
- **Page 7. Line 162.** A sentence was added: "Further results for common and recently developed performance criteria are presented in Fig. A1."
- **Page 17. Line 314.** A sentence was added: "Further results for common and recently developed performance criteria are presented in Fig. A2."
- **Page 17. Line 315.** The sentence was changed into: "It is interesting to note how similar these results are to those of the synthetic example (Fig. 3a, Fig. A1)."

3.4 Quoting: "According to Legates & McCabe (2013), the refinement of Index of Agreement (IOA) (Willmott, 1981) made by Willmott et al. (2012) especially regarding the extension of the IOA bound from 1 to 0 was unnecessary. Check Legates & McCabe (2013) for other limitations of the refined IOA. Therefore, could the authors make use of the original form of IOA for their model performance evaluation and analyses?"

Response. Thank you for pointing this interesting discussion about the refined index of agreement. Following your suggestion, we changed the refined index of agreement to the modified index of agreement (Willmott et al., 1985). We chose to use the modified index of agreement because it is less sensitive to outlier, which is relevant in our case study.

Modification to manuscript.

- **Page 3. Line 67.** The sentence was changed into: "[...] as well as more traditional criteria such as the NSE or the modified index of agreement (d_1) for comparison purpose."
- **Page 5. Line 124.** The sentence was changed into: "Willmott et al. (1985) proposed a modified index of agreement, which aim to address the issues associated with *r* and the coefficient of determination, as well as the sensitivity of the original index of agreement to outliers (Legates and McCabe Jr., 1999):"
- Equation 20 was changed to the one of the modified index of agreement.
- **Abstract.** The sentence was changed into: "[...] as well as the Nash-Sutcliffe Efficiency (NSE) and the modified index of agreement (d_1) [...]"
- **Abstract.** The sentence was changed into: "We recommend using (i) performance criteria that are not or less prone to counterbalancing errors (d_1, modified KGE, non-parametric KGE, Diagnostic Efficiency) [...]"
- "d_r" was replaced with "d_1" throughout the manuscript.
- Figure 3, 4, 5 and 7 were updated accordingly.
- **Table 1.** The advantages of the modified index of agreement were changed into: "Address the shortcomings of *r* and the coefficient of determination (Willmott et al., 1981)" and "The score is less sensitive to errors concentrated in outliers in comparison to the original index of agreement (Willmott et al., 1985)"

Additional modification to the manuscript

4.1 Grammar and error check

Modification to manuscript.

- **Page 5. Line 109.** As the FDC abbreviation is already defined Line 84, the sentence was changed into: "The non-parametric form of the variability is calculated using the FDC [...]"
- Remove repetition of *the*, e.g. "The NSE and ~~the~~ d_1".
  - Page 3. Line 66.
  - Page 7. Line 162.
  - Page 9. Line 188.
  - Page 9. Line 189.
  - Page 11. Line 209.
  - Page 12. Line 219.
  - Figure 5. Caption.
  - Page 13. Line 230.
- The reference of a Python library was corrected: "[...] Bayesian Optimization (Nogueira, 2014) [...]"
- **Page 17. Line 312.** Add missing "of": "However, the KGE and most of its variants [...]"

References

Barber, C., Lamontagne, J. R., and Vogel, R. M.: Improved estimators of correlation and R^2 for skewed hydrologic data, Hydrological Sciences Journal, 65, 87–101, https://doi.org/10.1080/02626667.2019.1686639, 2020.

Onyutha, C.: A hydrological model skill score and revised R-squared, Hydrology Research, 53, 51–64, https://doi.org/10.2166/nh.2021.071, 2022.

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.

Willmott, C. J.: On the validations of models, Phys. Geogr., 2, 184–194, https://doi.org/10.1080/02723646.1981.10642213, 1981.

Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. J., Klink, K. M., Legates, D. R., O'Donnell, J., and Rowe, C. M.: Statistics for the evaluation and comparison of models, J. Geophys. Res., 90, 8995, https://doi.org/10.1029/JC090iC05p08995, 1985.

Willmott, C. J., Robeson, S. M., and Matsuura, K.: A refined index of model performance, Intern. J. Climatol., 32, 2088–2094, https://doi.org/10.1002/joc.2419, 2012.