hess-2022-380: RC2

The paper examines several goodness-of-fit, skill scores used for hydrologic model evaluations based on the synthetic flow data and real simulation data. In the paper, the types of the skill scores are classified into two—1) multi-variant-based skill scores such as KGE (the skill score computed based on multiple metrics like bias, variability error, correlation etc. using distance measures) and its variants and 2) NSE. The paper focused on the impacts on the skill scores (also its components - bias and variability error if applicable), originating from the situation where under- and over-estimation on the peak flow can exist in one hydrograph. Also, the paper discusses compensation between the components of the skill scores, namely bias and variability. The paper concludes KGE type scores can be inflated for the hydrograph include both under- and over-estimation of the event (because of lower bias and variability error over the time series), which does not necessarily represent "accurate" simulations. The paper suggests that weighting KGE components mitigates this misleading score values.

I think the hydrologic modeling community intuitively realizes this counter-balancing issue in the KGE type scores. The paper explicitly illustrated the issues clearly and would be nice reference for the hydrologic modelers. I think the paper is also in fairly good in shape in terms of the presentations and writing, and I don't find any major comments, and only several minor comments.

Thank you very much for your thoughtful comment and for taking the time to carefully reading the manuscript. We appreciate that you find this work relevant to the hydrological modelling community.

Another thought: the overall results are mostly due to the fact that the skill scores use bias, instead of the error in the magnitude (e.g., root-mean square of error, absolute error). I wonder if it is worth trying modifying KGE components into two components - absolute error and correlation. I am not requesting the authors do (I don't even know this is a good idea), but reading the paper makes me think about it.

Thank you for the suggestion, this is an interesting idea that could be developed in a further study. Here are some preliminary results coming from a small, quick experimentation using a kind of absolute error for the volume estimate instead of β.

$$KGE'_{abs} = 1 - \sqrt{(\gamma - 1)^2 + {\beta_{abs}}^2 + (r - 1)^2}$$

With KGE'_{abs} the modified KGE using β_{abs}, γ the ratio between the coefficient of variation of simulated values and the coefficient of variation of observed values, and r the Pearson correlation coefficient. β_{abs} corresponds to the ratio of absolute errors to the sum of all observations, and is calculated as follows:

$$\beta_{abs} = \frac{\sum |x_s(t) - x_o(t)|}{\sum x_0(t)}$$

With $x_s(t)$ and $x_o(t)$ the simulated and observed values of a variable x at a specific time step t. For β_{abs}, 0 would thus correspond to a perfect fit.

For synthetic models (Sect. 3 of the article) and the real case study (Sect. 4 of the article), it gives interesting results which correspond to what is expected from the visual assessment.

For the synthetic model – better score is in bold:

Model	Bad-Bad	Bad-Good
β	**0.98**	0.88
β_{abs}	0.22	**0.12**
KGE'	**0.94**	0.87
KGE'$_{abs}$	0.77	**0.87**
NSE	0.92	**0.95**

And also, for the real case study – better score is in bold:

Model	ANN	Bucket-type
β	0.92	**1.00**
β_{abs}	**0.27**	0.31
KGE'	0.82	**0.87**
KGE'$_{abs}$	**0.69**	0.66
NSE	**0.87**	0.82

In the future, we will consider to do an extensive study on whether this approach could be relevant for the calibration and evaluation of the performance of hydrological models.

Minor comments

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.

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.

L318: *"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)."*

L319: *"We thus recommend using performance criteria that are not or less prone to counterbalancing errors (d_r, KGE', KGE_{NP}, DE), preferably [...]"*

L22: *"We recommend using (i) performance criteria that are not or less prone to counterbalancing errors (d_r, modified KGE, non-parametric KGE, Diagnostic Efficiency) in a [...]"*

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).

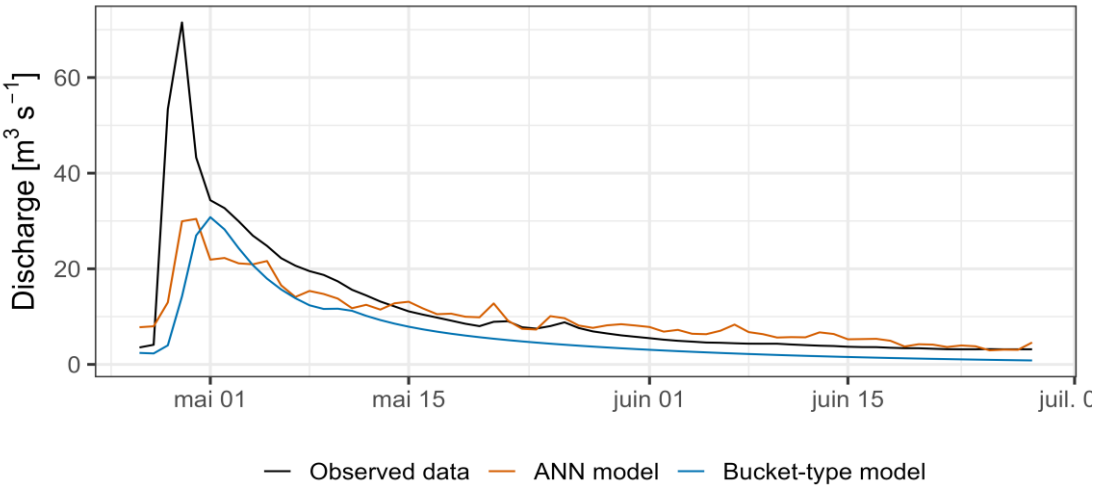
Indeed, the term "reservoir model" can be confusing for some readers. Thank you for the nice suggestion. "reservoir model" was replaced with "bucket-type model" in both text and figures.

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.

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 ($r = 0.88$) and

the overall volumes ($\beta = 0.87$), while the bucket-type model has a better recession coefficient and variability ($\gamma = 0.99$) – see the figure and table below for the detailed analysis of the flood event. We modified the manuscript accordingly.

L263: *"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."*



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

L269-273. I suggest using the dates to point which events are referred to.

Thank you for the suggestion. We modified the paragraph so that the dates directly refer to the events, L268: *"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."*