hess-2022-380: CC1

The study investigated counterbalancing errors which are inherent in Nash-Sutcliffe Efficiency and its variants. Reliability of the performance criteria is important to boost confidence with which a particular model can be chosen. There are some important points which the authors could address to strengthen their paper.

We would like to thank you for taking the time to carefully review our manuscript and for providing valuable comments and suggestions.

1. **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.

Indeed, the explanation around the use of a multi-criteria framework is unclear and we will clarify this aspect. As you mentioned, the KGE already includes several components for evaluating the bias, variability and linear correlation of a model, which help to evaluate different aspects of a model. We are currently considering providing additional guidelines for evaluating "how much" a model is prone to counterbalancing errors, which can help to assess the relevance of the performance criteria used for the calibration and evaluation.

2. **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.

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. We will add some text (and maybe some scatter plots) on this aspect. 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.

Pool, S., Vis, M., and Seibert, J.: Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency, Hydrol. Sci. J., 63, 1941–1953, https://doi.org/10.1080/02626667.2018.1552002, 2018.

3. **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).**

Thank you for the suggestion of these two innovative performance criteria and the associated code. We will really consider adding them in the study as they aim to address the shortcomings of widely used performance criteria. It can be interesting to see if these metrics are prone to counterbalancing errors.

4. **ON EQUATION 20**

   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?

Thank you for pointing out this interesting discussion about the refined index of agreement. We will make use of the original for of the index of agreement in the revised version of the manuscript.