RC2: Anonymous

The value and relevance of different objective functions to quantify the performance of hydrologic models is a never-ending story. The authors present a study in which they connect several versions of hydrologic efficiency metrics with hydrologic signatures to study their interactions across catchments and models. The study is technically fine, and I see no obvious problem in what the authors have done. What I really would like the authors to reflect on more extensively, is what they have learned.

The authors state in their abstract that "Results show that the choice of OF can significantly affect a model's capability to simulate different hydrological signatures …". And that "Generally, no single OF simultaneously achieved high performance across all tested signatures, highlighting that a single-objective calibration is unlikely to lead to an all-purpose model. Our results reinforce calls to choose objective functions deliberately and in line with the objectives of a study."

OK, I believe you but – given the studies listed in Table 1 and many others before then – did we not know this already? No metric is giving us a perfect fit and hence we should assume that differences in metric formulation are reflected in differences in signature values. Let me cite an older paper (which was the first one I tried to find a suitable quote): "The selection of the best efficiency measures should reflect the intended use of the model and should concern model quantities which are deemed relevant for the study at hand (Janssen and Heuberger 1995)." This quote is from Krause et al. (2005, Advances in Geosciences), referencing Janssen and Heuberger (1995, Ecological Modelling). I am not suggesting that the current study is not providing new insights. Just that the rather general statements in the abstract are not it. Could you be more specific in your conclusions and hence in what this specific study adds to our knowledge base?

Thank you for highlighting this. We agree that we can improve our communication about what we have learned in our study compared to previous statements on "calibration metrics should be picked based on the modelling purpose". We will improve the abstract to highlight that we identify general strengths and weaknesses of individual calibration metrics regarding signature representation to help modellers make a more informed decision on purpose-aware metric selection.

Compared to existing studies, we believe that ours is able to broaden the existing literature mainly by using a larger sample of models for the selected catchments, signatures and objective functions. Specifically, we are able to quantify the effects consistently and thereby generalize existing findings. We will clarify the abstract in this regard.

Regarding the question of new findings, we believe that the study was able to both verify/falsify findings in the existing literature and highlight new relations. For example, we were able to confirm the ability of the KGE to represent high flows well (Mizukami et al.,

2019, who use 2 models, whereas we show that these findings hold for 47 models, strongly suggesting that these findings are broadly applicable and not specific to Mizukami's model selection), but could not clearly attribute improvements compared to the NSE regarding the insensitivity to low-flows (Althoff and Rodrigues, 2021; who use 1 model compared to our 47).

And finally, if you want to get hydrologic signatures right (which should reflect hydrologic processes more directly), then why optimize an efficiency metric and not directly a combination of those signatures you think best reflect your catchment and purpose?

We agree that the calibration of signatures is an appealing idea when specific hydrologic aspects are of relevance, and will clarify that this is an active area of research in the introduction. There are substantial efforts to indicate that this approach may be beneficial (Yilmaz et al., 2008; Euser et al., 2013; Wagener and Montanari, 2011). In our study, we deliberately tested individual common calibration approaches because (1) this remains the dominant approach in large-sample and operational hydrology, and (2) by providing insights on the strengths and weaknesses of individual metrics we provide information that can help modellers to fill the "gap" in their model performance by choosing additional metrics or building multi-metric or multi-objective calibration routines that highlight the different aspects important to their purpose. We are aware that there is merit to signature-based approaches, but rather wanted to evaluate how common objective functions affect hydrologic behaviour. We will make this clearer in our text.

In this sense, what functions of the catchment are represented by the chosen signatures? Can you discuss the signatures you selected in terms of the hydrologic behavior they (should) reflect?

The selection of hydrologic signatures is a critical choice. In our study, we aimed to both represent a large range of streamflow characteristics and linkages to catchment internal processes (e.g. Olden and Poff, 2003; McMillan, 2020). The intent was not exhaustive process coverage, but to include a set of interpretable, streamflow-based indicators that are expected to respond differently to calibration choices.

Related to this, we identified three relevant clusters in the discussion: (1) runoff tendency, (2) baseflow/storage and (3) variability. For example, runoff ratio and Q95 are strongly correlated, indicating that the tendency towards runoff generation and high flows are clearly linked. From a process perspective, this implies an overlap in the characteristics of the signatures, which might indicate similar underlying processes, at least for the sample we investigated. We will add more specific information on the relationship between signatures and hydrologic behavior in Table 5.

What do your signature clusters tell you about the underlying processes that a model should reflect? Can we learn something of diagnostic value from using these integrated metrics?

From a diagnostic perspective, these clusters help identify which aspects of hydrologic behaviour are systematically emphasized or neglected by different objective functions as the groups typically align well with best/worst performing OFs. This helps interpret calibration results in terms of behavioural trade-offs rather than overall performance and can indicate whether deficiencies are related to calibration choices or model structure.

At this point, we are unable to conclusively assess the diagnostic values of the integrated metrics, but we will return to that thought in the discussion and either present our findings or state present limitations.

The authors further conclude that: “Together, our results support the argument for a purpose-based model calibration, that considers multiple aspects of the flow regime, and multi-objective calibration setups, rather than defaulting to a familiar single metric (Mai, 2023; Jackson et al., 2019).” This is rather close to: “This paper suggests that the emergence of a new and more powerful model calibration paradigm must include recognition of the inherent multiobjective nature of the problem and must explicitly recognize the role of model error.” from Gupta et al. (1998, WRR). The multi-objective nature of the problem is clear, so why are we still looking for a single metric solution?

We agree that the multi-objective nature of model calibration has been well recognised for a long time. As mentioned before, single-objective calibration remains common practice, due to its simplicity and potentially the lack of clear guidance on how to define effective multi-objective criteria. In this sense, our results can help inform the construction of purpose-based, multi-objective calibration strategies by clarifying which metrics contribute to representing specific aspects of the flow regime. We will ensure that this intent will become clearer throughout the paper.

Minor comment:

The authors use the slope of the flow duration curve (FDC). I assume that this signature quantifies the slope of the central part of the FDC, though the authors never specify this. It would be good if they would.

This is fair and the assumption is correct. We will add the equations for all signatures to the supplement of the paper.

References:

Althoff, D., & Rodrigues, L. N. (2021). Goodness-of-fit criteria for hydrological models: Model calibration and performance assessment. Journal of Hydrology, 600, 126674. https://doi.org/10.1016/j.jhydrol.2021.126674

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., & Savenije, H. H. G. (2013). A framework to assess the realism of model structures using hydrological signatures. Hydrology and Earth System Sciences, 17(5), 1893–1912. https://doi.org/10.5194/hess-17-1893-2013

McMillan, H. (2020). Linking hydrologic signatures to hydrologic processes: A review. Hydrological Processes, 34(6), 1393–1409. https://doi.org/10.1002/hyp.13632

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

Olden, J. D., & Poff, N. L. (2003). Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. River Research and Applications, 19(2), 101–121. https://doi.org/10.1002/rra.700

Wagener, T., & Montanari, A. (2011). Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. Water Resources Research, 47(6), 2010WR009469. https://doi.org/10.1029/2010WR009469

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. Water Resources Research, 44(9), 2007WR006716. https://doi.org/10.1029/2007WR006716