We thank Benedikt Heudorfer for the detailed and overall positive comments. These will certainly help us to improve the clarity of the manuscript. Below, we respond to the individual comments (our responses are in blue italics).

General comment:

This paper deals with very relevant questions w.r.t. proper comparison of model quality on large sample hydrological datasets. It uses the HBV conceptual model to provide upper and lower performance boundaries on dataset and catchment basis. This is a highly relevant contribution that can also guide the general machine learning community as it provides (among other things) reasonable lower boundaries of predictability of individual catchments from a conceptual hydrological model. Since neural networks are inherently random, the performance of the model can vary substantially between for each catchment between model runs. The study presented here can aid this shortcoming by telling us where good performance can be expected and where not.

Thank you for the kind words

Major comment:

However the main shortcoming of this paper is that it needs to provide better documentation of how the models are actually run, on which catchments they are run, on which time periods they are trained, what the exact performance is for each respective dataset and/or catchment, etc.. While this is (mostly) not sensible to report in the main body of the paper, these details should be included in the supplement files, in a data repository, published along with the code, anywhere (the link in the data availability section is dead). These details are crucial to allow what the study sets out to do: enabling comparison/benchmarking of models in future studies, because without these details, future studies will not be able to replicate the experimental setup properly. See also specific comments on this topic.

We fully agree that the details to replicate the modelling need to be given. We provide this information in the data repository. Please see the footnote: "This will be the link to the data on Zenodo once the paper is accepted. A temporary link is made available to the editor and reviewers."

The reason is that once we activate the link, we cannot make changes in the data (structure) on Zenodo without creating a new version. However, it seems the temporary link might not have reached the reviewer. We have sent this by email to the reviewer (and will send it to others upon request).

Specific comments:

Line 46 introduce NPE properly please.

We will add the full names for all the performance measures.

Line 98: dot missing

Thanks for spotting this. The full-stop dot will be added.

Section 2.3 is missing crucial information e.g. on test periods. To allow future studies to compare performance, the exact test periods should be named for each dataset. Background: The choice of data is really important how the value of the performance metric turns out (as the "elephant in the room" paper by Maier 2023 points out: https://doi.org/10.1016/j.envsoft.2023.105779). If exact periods and catchment lists are not reported for each dataset, valid comparison with your results are practically impossible, rendering the attempt to provide benchmark for future reference futile. Also, exact catchment list should be reported somewhere, on which you train. This is needed for people to match their model setup exactly in the future, in case they want to benchmark the results.

We fully agree. We will add a sentence and the (interesting) reference here, to better refer to the data repository, where this information can be found.

Line 117-119: "The first one to two years were used as a warming-up period to obtain reasonable initial conditions for the storage components of HBV. The remaining years were then used for model calibration (upper benchmark) and evaluation (lower benchmark) with streamflow." Does that mean no test period for calibrated models, and the reported upper benchmarks are from the calibration period metrics? If yes, that should be stated clearly. I don't know how it is treated in the conceptual modelling community, but in machine learning this would be considered data leakage and not a reliable metric for model performance, as it will mostly speak for the degree of overfitting. Could you please elaborate whether the phenomenon of overfitting is relevant in the conceptual modelling domain, or why the choice to report calibration period metric makes sense? Or if I misunderstand the section and this all is beside the point, please clarify.

We have indeed chosen not to include a validation (or test) period in this study. We want to show how well the streamflow (during the entire periods) can be reproduced by the model. There is, of course, always the risk of overfitting. However, with typical bucket-type (or conceptual) models and their limited number of free parameters, the risk of overfitting is much lower than with other approaches. The important issue in this study was to compute how well the observed could/should be reproduced when using the entire periods. This approach also followed a recommendation expressed by Shen et al. (2022): "Calibrating models to the full available data period and skipping temporal model

validation entirely is the most robust choice and eliminates additional subjective decisions." (p.20)

Line 131-134: very good strategy and reflects the authors knowledge about the dependency of metrics to calculation procedure. Speaks for the rigidity of the study.

Thanks for your kind words.

135-138: here as well, the exact catchment IDs associated with the lower benchmark should be made public somewhere, if not already present in a published codebase. That is, if the random selection of the 10 catchments was not repeated with replacement, in which case tracing it may be pointless. Background: The choice of data is really important how the value of the performance metric turns out.

Please see response above to your comment on section 2.3, the catchment IDs (which are crucial to use the values we provide) are given in the data repository.

Section 3.1: only evaluation based on NPE is reported, but 2.3.1 claims it was evaluated based on NSE, KGE and NPE; since NPE is more novel and NSE and KGE are more widely used, KGE and NSe should be reported as well; also, to allow proper use of this benchmark in future studies, tables with exact metric values (incl. any uncertainty of distribution indices like IQR etc.) should be reported, if only in the supplements.

These values are given in the data repository. A summary (10^{th} quantile, median, and 90^{th} quantile) of NPE, KGE, and NSE values is currently provided in Table 3. We feel that including more detailed values for all three metrics in the main text would overload the manuscript.

Section 3.2: Having this effect of the influence of the calculation procedure on the final metric value finally worked out in a paper is super exiting to me. We have briefly discussed this effect in our 2025 paper (https://doi.org/10.1029/2024GL113036) and internal tests show metric value deviations of up to 0.1 NSE depending on the calculation procedure, but I am not aware that this was specifically addressed anywhere yet. Very nice result, can be highlighted more prominently in my humble opinion.

We agree that this result is interesting. Thanks for pointing us to your study where you found a similar effect.

Line 208-209: I don't understand exactly how this parameter range width is applied; how can a negative parameter range be? Also range values should be included in the figure 3 xlabs as well.

Here we report the values for the lower benchmark when different parameter ranges for the random selection of parameter values were used. Depending on

ranges (wider or narrower) and catchments, this could result in negative NSE-values.

Table 3: ah, here is the table on exact upper and lower benchmark values ; but should be reported further split into the individual datasets, so people can compare their models against it in the future. These are surprisingly strong lower benchmark by the way. Reading this paper made me think how a lower benchmark might be computed for ML/DL models. If you have any ideas, I'd appreciate elaborating and/or putting them into the paper.

Yes, comparisons with other modelling approaches should be done on the basis of individual catchments. For this, we provide lists of benchmarks (model performance values) in the data repository.

Regarding using ML/DL for lower benchmarks: bucket-type (or conceptual) models have the advantage of being based on the water balance, which is the main reason for the (indeed) surprisingly good performances of the lower benchmarks (see Seibert et al., 2018, for more on this). The water balance, together with some (rough) idea of model parameter values, is something that might be missed for ML/DL approaches. It would indeed be interesting to explore more, how ML/DL could be used for lower benchmarks. One approach could be to use a number of models, that have been trained on other catchments, as 'general' model.

Reference

Shen, H., Tolson, B. A., & Mai, J. (2022). Time to update the split-sample approach in hydrological model calibration. Water Resources Research, 58(3), e2021WR031523. https://doi.org/10.1029/2021WR031523