

Dear Editor, lieber Ralf,

We thank you and the two reviewers for the detailed and overall positive comments. These helped us improve the manuscript's clarity. Below, we respond to the individual comments (our responses are in blue italics). Where suitable, we provide line numbers for the implemented changes.

Best regards,

Jan Seibert, Sandra Pool and Marc Vis

Editor comments by Ralf Loritz:

Both referees and I assessed the manuscript positively and consider it a valuable contribution. I believe it can be published after a round of technical corrections and additions. I therefore trust the authors to address the remaining comments, particularly those concerning transparency and reproducibility. All information required to reproduce the experiments should be provided. Even if it is available on Zenodo, I recommend also including this information in the manuscript's supplementary material. Please also clarify that the upper benchmark is based on calibration-period performance and discuss its interpretation accordingly.

Overall, this is very nice work, and I look forward to seeing the final version published.

Thanks for your kind words.

We clarified that the upper benchmark depends on the calibration period and the model used.

Regarding the suggestion to provide the tables in the data repository also as a supplement in HESS, we would prefer to keep these data in one place only. The Zenodo option is our preference as this will allow us to update the tables to, for instance, add benchmarks for further CAMELS datasets in the future.

Comments by Benedikt Heudorfer:

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 in the data availability section: "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. The link provided in the text is now active.

Specific comments:

Line 46 introduce NPE properly please.

We added the full names for all the performance measures and added a table. (L48ff and Table 1)

Line 98: dot missing

Thanks for spotting this. The full stop has been 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 added a sentence and the (interesting) reference here, to better refer to the data repository, where the information on catchments and calibration periods can be found. (L137f)

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 key issue in this study was to assess how well the observed could/should be reproduced 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). We added the reasons for not using a validation period to the manuscript in section 2.3. (L133ff)

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 (10th quantile, median, and 90th quantile) of NPE, KGE, and NSE values is 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, which we now included as reference. (L224f)

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 NPE-values. We adapted the figure caption to better explain the meaning of the parameter ranges. (Fig 3)

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 based on individual catchments. For this, we provide benchmark lists (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>

Comments by Tam Nguyen:

General Assessment

From my understanding (please correct me if I have misunderstood any aspect of the methodology), the study derives lower and upper ranges for model performance metrics (NSE, KGE, and NPE). The lower benchmark is obtained from random or regional parameter sets, whereas the upper benchmark is derived from calibrated parameter sets. This framework is applied across multiple large-sample hydrological datasets (CAMELS and LamaH) using the HBV model. The overarching objective appears to be the establishment of performance ranges that can serve as references for future studies, allowing researchers to contextualize model performance in relation to catchment characteristics.

Yes, this was our objective, thanks.

1. Strengths of the Manuscript

The manuscript represents, to my knowledge, the first study to establish the HBV model at such a large scale, covering 12 CAMELS and LamaH datasets. This extensive analysis provides a valuable contribution to the hydrological modelling community, especially in terms of model evaluation and inter-comparison.

Thanks for your kind assessment.

Considering that nowadays there are many machine learning models and papers, this paper raises an important question to the ML community: “Do we need another ML model for streamflow prediction?” Considering the lack of trust and explainability in many ML models, this paper shows that a bucket-type model with “low data demand and ease of application” can be highly scalable, provide physically explainable results, and still achieve high performance. Therefore, ML models may only be needed if they can outperform bucket-type models such as this one.

While we did not explicitly address ML in our study, the reviewer is, of course, right that our results show that an ML approach must achieve a certain (and rather high) performance to be justified. Since both reviewers mentioned ML, we addressed ML shortly in the revised version. (L91ff)

2. Points for Discussion and Potential Improvements

From my understanding, the upper benchmark is derived from the calibration period. If this is correct, it should be stated explicitly in the manuscript. While overfitting may be less problematic for a relatively parsimonious model such as HBV, the interpretation of an upper benchmark derived from calibration data warrants discussion (e.g., if such model is evaluated for other period, the upper benchmark could be lower). For more complex models with larger parameter spaces (e.g., SWAT, mHM), calibration

performance may approach the theoretical optimum and therefore provide an unrealistically optimistic benchmark.

We agree that the calibrated performance values provide an indication of what performances are possible to be achieved and clarified this in the text by adding the following sentence: “The performance values thus provide an indication of what is possible to be achieved. These values must be expected to decrease over independent periods.” (L181f).

The upper benchmark is likely influenced by several factors, including: model structure, calibration period, objective function, data quality and completeness, record length, model complexity. These factors deserve further discussion and analysis in the manuscript.

Yes, both the lower and the upper benchmarks are influenced by a number of factors. While we analyse the impact of objective functions in detail, a further analysis on the impact of other factors on benchmark values would indeed be interesting. However, we feel this would be beyond the scope of this manuscript. Instead, we discuss some of these factors in the discussion (e.g., L326ff) for data quality and resolution, and for model complexity/structure) (and now also ML) (L356ff).

Influence of human on the upper and lower benchmarks, e.g., catchments where streamflow is substantially modified by reservoirs, weirs, diversions, or water transfer schemes may exhibit much lower lower and upper benchmarks. Additional analysis of how anthropogenic influences may affect both lower and upper benchmarks would strengthen the manuscript.

Yes, anthropogenic influences will also affect the benchmark values, even if catchments with significant human impacts are excluded in most of the data sets. Similar as for other factors, see above, a detailed analysis would be interesting but quite challenging for such a large number of catchments, and clearly beyond the scope of this study.

3. Specific Comments

L48–49

“A perfect fit (value of 1) is usually impossible to achieve due to model and data errors and uncertainties. In the case of NSE, a value of zero can be considered a benchmark representing the prediction of a constant discharge equal to the annual mean discharge.”

This is the first explicit mention of discharge in the manuscript. Since the study focuses specifically on streamflow prediction, it may be helpful to introduce this earlier in the introduction so that readers immediately understand the scope of the performance evaluation.

Good point, we clarified that the evaluation is based on streamflow earlier in the text. (L46f, L63ff)

L54–58

“...using any fixed value for a performance measure to judge model performance ... might not be appropriate...”

While I agree with the authors’ argument regarding the limitations of “absolute” performance thresholds, it may be worth acknowledging that absolute performance measures still provide valuable information. For example, NSE = 1 represents a theoretically optimal model and remains an important reference point for model development. Relative benchmarks are highly useful for contextual comparisons, but they may not fully replace absolute performance measures, particularly in regions where all available models perform poorly.

We agree and discuss this issue later in the text (see section 4.1)

L63

“both upper and lower benchmarks”

The terms "upper" and "lower" benchmark should be defined more explicitly when first introduced. Initially, I interpreted the lower benchmark as the minimum achievable model performance, whereas it actually appears to represent performance obtained under minimal calibration effort (i.e., random parameter sets).

We agree and added a sentence to clarify the meaning of lower benchmarks. Note that we also refer to Seibert et al., (2018) where these terms are discussed in more detail. (L63ff)

L66–67

“the choice of model is less crucial...”

I am not fully convinced by this statement. For example, Kratzert et al. (2024, <https://doi.org/10.5194/hess-28-4187-2024>) demonstrated substantial differences in predictive performance among VIC, mHM, and LSTM models. Additional justification or supporting evidence would strengthen this conclusion.

We agree that performance can differ across model approaches. The second part of the sentence is important: the goal is to use a model to provide upper and lower benchmarks, although these benchmarks are only indications. Most importantly, the upper benchmark is not an absolute optimum and could be exceeded with another model. Still, we argue that the differences between the upper benchmarks for different catchments are larger than those between different models. In other words, the upper benchmark indicates what performance is achievable, even if it is not the theoretical best value any model could achieve. We clarified this by adding some text before the ‘less crucial’ statement. (L74f)

L71–72 and L76–77

The manuscript discusses using either the median ensemble performance or the performance of the ensemble mean simulation as a lower benchmark.

From a parameter-estimation perspective, the ensemble mean simulation may not correspond to any physically realizable parameter set. Therefore, the median performance across ensemble members appears conceptually more consistent for model comparison purposes. It would be helpful if the authors could provide a stronger rationale for preferring one aggregation method over the other.

 *We agree that this is an interesting question and address it later in the manuscript; see Section 4.2.*

L73–74

“The purpose of this study was two-fold. Firstly, we evaluated different approaches to derive lower benchmark values, and secondly, we provide upper and lower benchmark values for existing large-sample datasets.” I think the term model performance for “streamflow” should be introduced before or here.

We agree and clarified this. (L83)

Section 2.1

Providing one detailed example of the HBV model setup and input data would greatly enhance reproducibility and facilitate adoption of the framework by other researchers.

We fully agree on the importance of reproducibility. We provide the necessary details in the data repository.

In addition, the simulation, calibration, warm-up, and evaluation periods are not clearly described.

Table 1: Consider adding: simulation period, warm-up period, calibration period, evaluation period.. This information would improve transparency and reproducibility.

The simulation period was divided into a warm-up period and a calibration/evaluation period, whereby the first one to two years served for warm-up and all the remaining data was used for calibration/evaluation (L129ff). Thus, calibration/evaluation period and its length differed among datasets and/or catchments. The information needed to reproduce the simulations, most importantly the simulation periods, are given in the data repository.

L117

“The remaining years...”

Please specify the exact calibration and evaluation periods rather than referring to them generically.

The exact periods differed between the different datasets. While we fully agree on the importance of this information, to avoid going into too much technical detail in the main text, we have included the exact periods in the data repository.

L129–130

The analysis of subsets containing 1, 2, 5, and 10 parameter sets is acceptable, although it is not immediately clear how informative these very small sample sizes are in practice.

Additionally, since 10,000 represents the full parameter ensemble rather than a subset, the wording could be revised for clarity.

We agree that some of the sample sizes are very small. However, they are needed to properly analyze the impact of sample size and consequently justify which sample size might be large enough for robust results. We would like to emphasise that we selected 10,000 sets from a pool of 100,000 sets, meaning that 10,000 is a subset.

L131

“The selection of subsets was repeated ten times”. I think the authors can repeated more times (my rule of thumb is 30 times or more, please see this reference; https://web.stanford.edu/class/archive/cs/cs109/cs109.1212/lectureNotes/LN18_clt.pdf?utm_source=chatgpt.com) to have a robust result. As the simulation results are already there, doing this does not require much effort

Thanks for this comment. We increased the number of repetitions from 10 to 30 and included a new figure in the manuscript. However, as expected, results did not change considerably. (Figure 1)

Table 2: The parameter-range combinations are difficult to follow. It would help if the authors explicitly stated the number of combinations considered and perhaps provided examples such as [LL1, UL1], [LL1, UL2], etc.

We realised that there was some confusion. We actually used (only) seven ranges, namely LL1-UL1, LL2-UL2, ... and LL7-UL7. We clarified this by stating the number of ranges being tested (7). (L157f, Table 3, Figure 3)

L157 (NPE): Many readers may be less familiar with NPE than NSE or KGE. A brief description of NPE and its interpretation relative to the other metrics would improve accessibility.

We added an explanation of the abbreviation NPE at its first occurrence (above) and refer to Pool et al. (2018) twice, which describes NPE in more detail. We feel that repeating more of this information here would distract from the main objectives of the current study. However, we added a table with the equations for the three performance measures. (Table 1).

L162–163

“...generated maps of model performance for each country and predicted model performance using random forest regression trees.”

I only observed maps for the United States. If maps were generated for additional regions, it would be useful to present them or provide them as supplementary material.

Yes, these maps are included in the supplementary material.

Figure 1: I was surprised to observe that the median NPE appears to increase with increasing numbers of random parameter sets (Figure 1a). Additional discussion of this behaviour would be helpful (e.g., is this model-specific behavior?)

Honestly, we did not find this finding very surprising. Please note that Fig 1a is one (representative) example catchment. When looking at all 128 catchments, the increase of performance with increasing sample sizes is much more moderate and can be attributed to the increasing robustness.

Furthermore, the rationale for presenting NPE rather than NSE or KGE in this figure is not entirely clear. Including the mathematical definition of the reported statistics would also improve interpretability.

Maps for all objective functions are shown in the supplementary material. We selected NPE here as we found NPE advantageous compared to the other two alternatives (see Pool et al., 2018). We clarified in the text that we focus on NPE in the manuscript, but provide the results for NSE and KGE in the data repository. (L209f)

Section 3.2 It is not clear to me what the purpose of this section is

This section addresses the question of using the median performance of individual simulation time series or the performance of the simulated ensemble mean, this question also raised by the referee, see above.

Section 3.4. It is not clear to me which range was ultimately selected for deriving the lower benchmark

For the regional parameter sets, calibrated parameter sets were used. We clarified (in the methods) the parameter range used for calibration. (L195)

Figure 3

This figure clarified that seven benchmark ranges were considered rather than the 49 combinations I initially inferred from Table 2. It may be helpful to explain this more clearly earlier in the manuscript.

We realise that we described this in a confusing way, see comment above.

Figure 5

It would be highly valuable to provide a supplementary CSV file containing: catchment identifiers, dataset name, benchmark values, calibration period, evaluation period. This would substantially increase the utility and reproducibility of the dataset for future studies.

We fully agree. Such a file is included in the data repository on Zenodo.

L263

From my experience, seasonality in streamflow, precipitation, and other climatic variables is often a major determinant of hydrological predictability. Catchments with strong seasonal signals are frequently easier to model and may achieve higher performance metrics.

It would therefore be interesting to investigate whether seasonality metrics could explain some of the observed variation in benchmark performance and whether they should be included among the predictor variables considered in the analysis.

Good point. In general, the relationships between catchment characteristics and model performance are more tendencies than strong correlations. The seasonality indeed might also be related to model performances and we already included some catchment attributes related to seasonality (such as precipitation seasonality or snow fraction). We agree that this is an aspect worth exploring

further. However, we decided to use only the signatures already available across all datasets to simplify reproducing the results. (see also Araki et al., 2026)

Reference

Araki, R., Holt, A., Hammond, J. C., Husic, A., Coxon, G., and McMillan, H. K.: Continental-scale prediction of hydrologic signatures and processes, Hydrol. Earth Syst. Sci., 30, 3647–3673, <https://doi.org/10.5194/hess-30-3647-2026>