We thank Tam Nguyen 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 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 will address. ML in the revised version¡

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.will.clarify.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¡‹

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¡?L8❺❸_966.for.data.quality.and.resolution·.L98❶_98❹for.model.complexity–~~st~~ructure).and.will.further.extend.this¡

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.will.clarify.before.that.the.evaluation.is.based.on.streamflow

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.0¡7)

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.will.add.a.sentence.to.clarify.the.meaning.of.lower.benchmarks¡. Note.that.we.also.refer.to.Seibert.et.al¡?(867❹).where.these.terms.are.discussed.in. more.detail¡

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.there.can.be.performance.differences.between.various.model. approaches¡.Important.here.is.the.second.part.of.the.sentence?i¡e¡?the.goal.of. using.a.model.to.get.an.indication.of.upper.and.lower.benchmarks?although.it.is. important.to.note.that.these.benchmarks.are.only.indications¡.Most.importantly?. the.upper.benchmark.is.not.an.absolute.optimum.and.could.be.exceeded.if. another.model.were.used¡.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.provides.a.useful.indication.of. which.performance.is.possible.to.achieve?even.if.it.is.not.the.theoretical.best. value.that.could.be.achieved.with.any.model¡.We.will.clarify.this.by.adding.some. text.before.the›less.crucial".statement¡

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.actually.address.this.later.in.the. manuscript?see.section.9¡8¡

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.will.clarify.this¡

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¡.The.necessary.details.are.provided.in.the.data.repository¡.We.will.consider.providing.there?in.addition.to.the.information.already.available?also.the.model.setup.for.an.example.catchment¡

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.(L77❶_786)¡.Thus?calibration–evaluation.period.and.its.length.differed.among.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.76?666.sets.from.a.pool.of.766?666.sets?meaning.that.76?666.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¡.While.we.do.not.expect.the.results.to.change?we.will.test.the.effect.of.increasing.the.number.of.repetitions¡

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.LL7_UL7?LL8_UL8?¡¡¡.and.LL❸_UL❸¡.We.will.clarify.this.by.stating.the.number.of.ranges.being.tested.(❸)¡.

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.will.add.an.explanation.of.the.abbreviation.NPE.at.its.first.occurrence.(above).and.refer.to.the.paper.by.Pool.et.al¡.(867❹).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.will.add.a.table.with.the.equations.for.the.three.objective.functions¡

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.7a.is.one.(representative).example.catchment¡.When.looking.at.all.78❹catchments?the.increase.of.performance.with.increasing.sample.sizes.is.much.more.moderate.and.can.be.attributed.to.the.increasing.robustness¡.We.will.add.a.sentence.to.comment.on.this¡

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¡ ?8674¡ We will clarify in the text that we focus on NPE in the manuscript? but provide the results for NSE and KGE in the data repository¡

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 will clarify (in the methods) the parameter range used for calibration¡

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.will.consider.further.seasonality.indices?but.do.not.expect.very.strong.correlations¡.We.agree.that.this.is.an.aspect.worth.exploring.further¡.However?we.also.decided.to.work.only.with.the.signatures.already.available.across.all.data.sets.to.simplify.the.reproduction.of.the.results¡.(see.also.Araki.et.al¡?868❷)

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¡?96?9❷0❸-9❷3❾?https¿-doi¡org-76¡❶7❺0-hess_96_9❷0❸_868❷