Dear editor, dear reviewers,

After thorough evaluation of the comments on our initial submission, we decided to make major changes to the analysis. These changes fall into two main areas, following the suggestion of the editor.

1. Parametric Uncertainty

The reviewers and editor ask for an assessment of parameter uncertainty in our analysis, and we acknowledge their reasons for doing so. Because originally only the final calibration outcomes were saved, we recalibrated all runs with five random seeds, calculating and storing signatures during run time. For the parameter uncertainty analysis, we now evaluate the top 25, 100, and 500 runs across all seeds. Although we found that parametric uncertainty was low in the case of our study, revisiting the analysis highlighted other forms of uncertainty, particularly model structural uncertainty, which we found can be remarkably large in certain basins. These additional findings are now presented in the revised manuscript.

2. Novelty

The choice of objective function is not a new topic in the hydrologic community, and some of the findings of this study will be unsurprising to some readers. Nevertheless, it remains a crucial topic, because by now most models applied for a hydrologic purpose are subject to a calibration procedure. We believe this study is a valuable addition because it (1) tests a large number of models in a systematic way, (2) examines the impact of model structure, (now also) parameters, objective function, and catchment in a comparative way, (3) applies a large sample of signatures, and (4) provides recommendations for combinations of objective function components and signatures based on the findings. In particular, the findings on model structure uncertainty in analyses such as these is a relevant addition to the current literature, where authors mostly focus on 1 or 2 models (for an overview of current studies, see Table 1 in our manuscript). This aspect of our study is now highlighted more clearly in the revised manuscript.

In addition, we have made the following changes:

- The decision to recalibrate all models enabled two further changes:
    - Following Clerc-Schwarzenbach et al. (2024), we replaced the PET data available in CARAVAN with new, more plausible estimates.
    - Following the recommendations in Thirel et al. 2024, we replaced log KGE with a power law transformation instead.

- To ease interpretation, we also updated our model benchmarking routine to always remove non-snow models in the three basins with a considerable snow fraction (C12, HYS, LAM at 0.49, 0.38 and 0.32 respectively). Previously, some models without a snow routine could still (barely) pass the benchmark through highly unrealistic parameter estimates. Such cases are now avoided.
- We extended the random forest analysis with an application to signature errors to investigate if it is possible to predict where and which signatures are more difficult to predict than others.
- We added detailed signature descriptions and additional supporting figures to the Supplementary Material. These are referred to throughout the manuscript where needed.

Overall, these revisions have had a considerable impact on the analysis, both in computational effort and in the conclusions drawn. We particularly believe that the revisions have increased the reliability of the findings and have clarified under which circumstances the choice of objective function is relevant, and which objective function components are most beneficial for which goal.

We thank the editor Markus Hrachowitz, the reviewers Guillaume Thirel and an anonymous reviewer, as well as Bettina Schaefli, Keith Beven, and John Ding for their community comments.

Below, we respond to the individual reviewer comments. In black, we show the reviewer comment, and blue shows the changes made to the manuscript, as well as any further thoughts deemed relevant.

Kind regards,

On behalf of all co-authors,

Peter Wagener

RC1: Guillaume Thirel

Wagener et al. present an analysis of the impact of objective functions when used to calibrate 47 conceptual hydrological models across 10 catchments from the CARAVAN dataset, on 15 hydrological signatures. The rationale of the authors is that the choice of the objective function is often poorly justified by authors. Wagener et al. identified several studies that have investigated the impact of objective function choice on the representation of hydrological signatures. However, in this work, they aim to develop a more generalized understanding of how the choice of performance metrics used for model calibration influences the representation of the hydrological regime. They conclude that no objective function provides satisfactory results for all signatures, that some signatures are insensitive to the objective function choice, and they provide some recommendations of objective function choice according to the objective of the model application.

Any knowledge about our understanding of which modelling (in the broad sense of model structure + parameters) we could use for which purpose is valuable. However, this is a tremendous work and the conclusions are rarely satisfying, as they are often unclear and they often depend on the methodological setup of the study. Nevertheless, the present study addresses a topic of interest to the HESS journal, and it is overall well written. If have several comments, some of them being major, and others being more minor.

Thank you for the thorough review of our submission. Please find our responses to your individual comments below.

Please note that I would be interested to see the authors' responses to the two community comments before the closure of the open discussion, as these comments were submitted quite early. In particular, the comment from Prof. Beven questions whether we really advance science with such studies that do not directly consider uncertainty in model preparation, and I guess that some (most?) of the studies I co-authored in the past could equally have received this comment. To some extent, I understand the comment. However, in the same time, most often hydrological models are deployed using an optimal parameter set, because it is simpler, because it is faster, because we always did like this, or because hydrological being a very applied science, it has to be simple enough to be used by stakeholders, forecasters (I do not say these are good reasons, but this is the way it is). Therefore, advancing knowledge regarding the identification of optimal parameter values can help such model uses. I would be interested to hear the authors' thoughts on this. From a similar philosophical viewpoint, I wonder what we can learn from such a large-sample study. As the authors acknowledge, it is difficult to draw strong conclusions when so many factors vary in the analysis and the differences in performance are not always significant.

As indicated in our initial response, we agree that it is still common practice to work with a single optimized parameter set. We do, however, also agree with the points made by Prof. Beven and the Editor, highlighting the importance of parameter uncertainty. We therefore

decided to substantially revise the analysis to address the uncertainty concern directly. We recalibrated every model-catchment-objective function combination across five random seeds, retaining the top 500 parameter sets across all seeds and computing signatures during the run time. In our study, we found the impact of parametric uncertainty (close to the optimum of the objective function value) to be small compared to the influence from model or objective function choice. In the revised submission, we now explicitly include model and parameter uncertainty in figures 3 through 6 by considering 500 parameter sets per model and catchment, adding the parameter uncertainty aspect to our analysis. Additional figures in the supplement help to assess the role of parameter uncertainty, showing e.g., how the results vary when a smaller or higher number of parameter sets is considered (S.2.10).

Major comments:
Although all the works presented in table 1 have limitations regarding the number of catchments, models, objective functions or signatures used, all of them explore one of these four items with a rather reasonable number of options. Therefore, it seems necessary for the introduction to present and, if needed, criticize the conclusions of all these studies. In the end, we can expect from the present study to highlight potential erroneous conclusions that may have been drawn in previous studies because the setup was too light on some aspect. We could also ask ourselves whether a meta-analysis of all these studies could lead to similar conclusions as the present work, and, eventually, we could question whether it is really necessary to go even bigger in terms of catchments, OFs, model types, or if a kind of IPCC-like review would lead to a consensus.

Thank you for this thought. We revisited the conclusions of the studies presented in Table 1 and did indeed find some similarities, but also differences that would make a potential meta-analysis difficult. We now explain why the previous studies cannot lead to the sort of generalized understanding of metric-signature connections we aim for in our study in a new paragraph added to the introduction.

"Some of the existing studies show recurring tendencies across regions and models. For instance, Pool et al. (2017) and Hallouin et al. (2020) found that traditional objective functions such as NSE or KGE can perform as well or better than tailored signature-based metrics for predicting flow characteristics not explicitly included in calibration. Mizukami et al. (2019) similarly noted that application-specific metrics may excel at their target but can degrade performance on other metrics. Generally, most studies can identify a preferred metric or metric combination. However, that selection is strongly conditioned to the modelling purpose, the flow conditions of interest, and the hydroclimatic setting of the individual study (e.g. seasonal forecasting in snow-influenced catchments for Araya et al. (2023) or tropical low-flow applications for Althoff and Rodrigues (2021)). A robust transfer of findings to new contexts is therefore difficult and further complicated by inconsistent experimental designs across studies (differing catchment sets, model selections, and signature choices), which prevent a straightforward synthesis on how metric choices influence signature representation."

In our discussion section, we further discuss the differences between our results and the recommendations found in these previous studies, e.g. in lines 499-508 or 525-528.

The authors apply their setup on 10 catchments, which is not a lot! In addition, all catchments are of a very similar area. I think that the classification used to limit the number of catchments should also consider the area, as the processes at stake at the catchment scale can differ a lot from a 100-300 km² catchment, as we have here, to a > 5000 km² catchment. The justification given in SM2.2 does not explain why larger catchments are excluded: "First, we limited the catchment size from 100 to 1000 km2 to increase the probability of an adequate signature representation through the calibrated models. Smaller catchments are less likely to dampen catchment responses and therefore better contain scale-dependent effects that also affect signature representation. Second, we ensured that the catchments had an data length of at least 20 years." While automatic procedures for (e.g.) catchment selection are useful, I think that hydrological expertise must intervene if this classification avoids a large category of catchments.

Excluding catchments larger than 1000km2 was a deliberate choice in this study to ensure signature values such as flashiness are not dampened by the 'smoothing out' of hydrologic response in larger basins. We have moved the justification for this step from the supplement to the catchment section of the methodology (line 174).

While I do agree that the impact of OFs is more direct on the calibration period, I am not convinced from SM Figure S3 that signatures are consistent between the two periods. Did you check to which extent the results are valid on the evaluation period? In my opinion, hydrological models are useless on calibration periods, therefore we want to know if the conclusions of such a work apply over the evaluation period. But of course, that makes the analysis even more complex.

In our initial response we briefly investigated the stability of conclusions when evaluation data was used. These findings suggested that results are broadly comparable, but with larger signature errors overall as well as some changes in the individual errors for specific signatures. However, the new calibration results, including the now 500 parameter sets being tested, means those initial results no longer hold. Below is an updated version of this analysis, which quantifies overall error per signature for each objective function (note: the analysis is limited to the signatures for which we found statistically significant impacts of OF choice on signature representation). We now see that the findings are more consistent between calibration and evaluation. The skill of signature representation decreases somewhat (as may be expected), but the relative strengths and weaknesses of each signature and even the ranking of objective functions stays similar. Please note this comparison was done using only the highest performing parameter set and not an ensemble of best performing parameter sets as in most other updated plots.

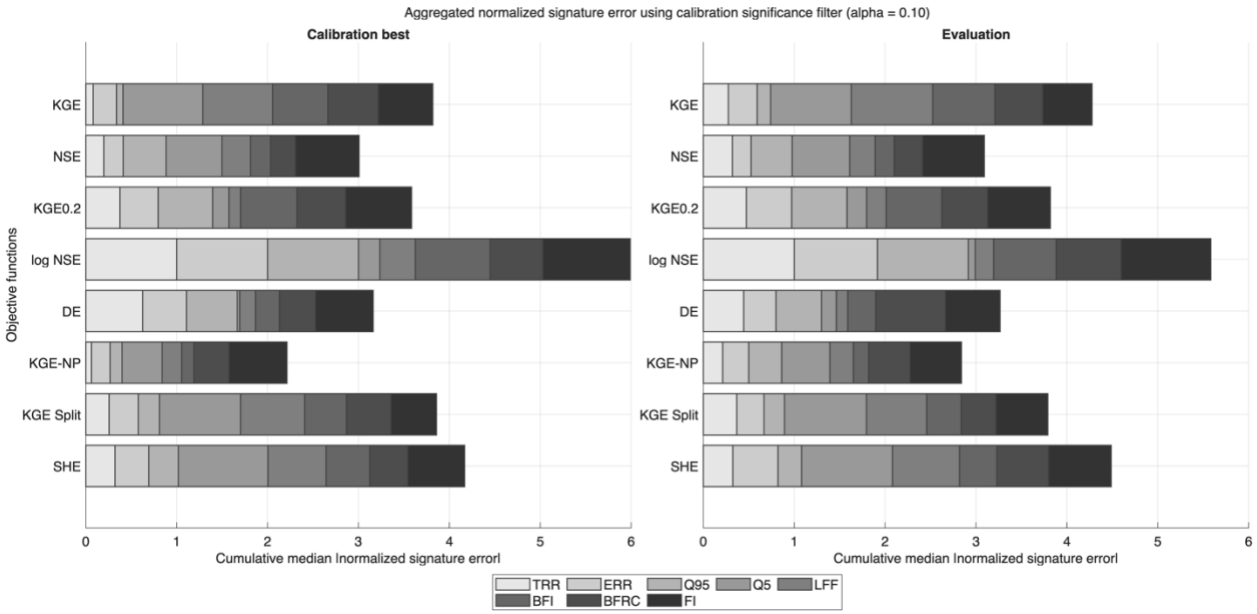

Figure 1: Comparison of Calibration and Evaluation Period for Error Metric

Model description and lines 249-251: The snow models used must be specified in the methods section. Do they use the same snow model? Do they use their companion snow model (e.g. CemaNeige for GR4J)? I also strongly suggest just discarding hydrological models that do not represent snow evolution in the analysis for the three catchments with the most snow, as it only adds noise in the analysis. Did you verify that those models were properly discarded from the further analysis because they perform less than the benchmark?

We have added text to the model description section clarifying that these models use their own snow model (Line 101-102), and implemented your suggestion of stricter benchmarking in snowy basins

The models selected in the three catchments with the most snow all possess a degree-day-based snow module. These models are Alpine 1 and 2 (IDs: 6, 12), Mopex 2-5 (IDs: 30, 31, 32, 35), Flexis (ID: 34), GSM-SOCONT (ID: 43), ECHO (ID: 44), HBV (ID: 37), NAM (ID: 41) and PRMS (ID: 45). This is a total of 12 out of 47 models. For internal consistency (i.e. numerical implementation, process equations, parameter ranges) we only use models from within the MARRMoT toolbox. Cema-Neige was not implemented as part of MARRMoT's mimicking of GR4J, and GR4J is thus not coupled with a snow module in our setup.

In the revised analysis, we added an additional filter to the benchmark routine that removes all models without a snow routine from the three basins that experience considerable snowfall.

Line 254-258: While I can definitely understand that for some models/OFs/catchments combinations the benchmark outperforms them, I am surprised that in some cases all models fail on the same catchment! What about the observed data quality in those cases?

The cases where no model could beat the benchmark are no longer present in the recalibrated results, likely due to improved PET estimations (see earlier response). Below is a plot showing a comparison of the forcing data currently available in CARAVAN, using basin C12 as an example. The plots for all other catchments can be found in the supplement (S.2.3)

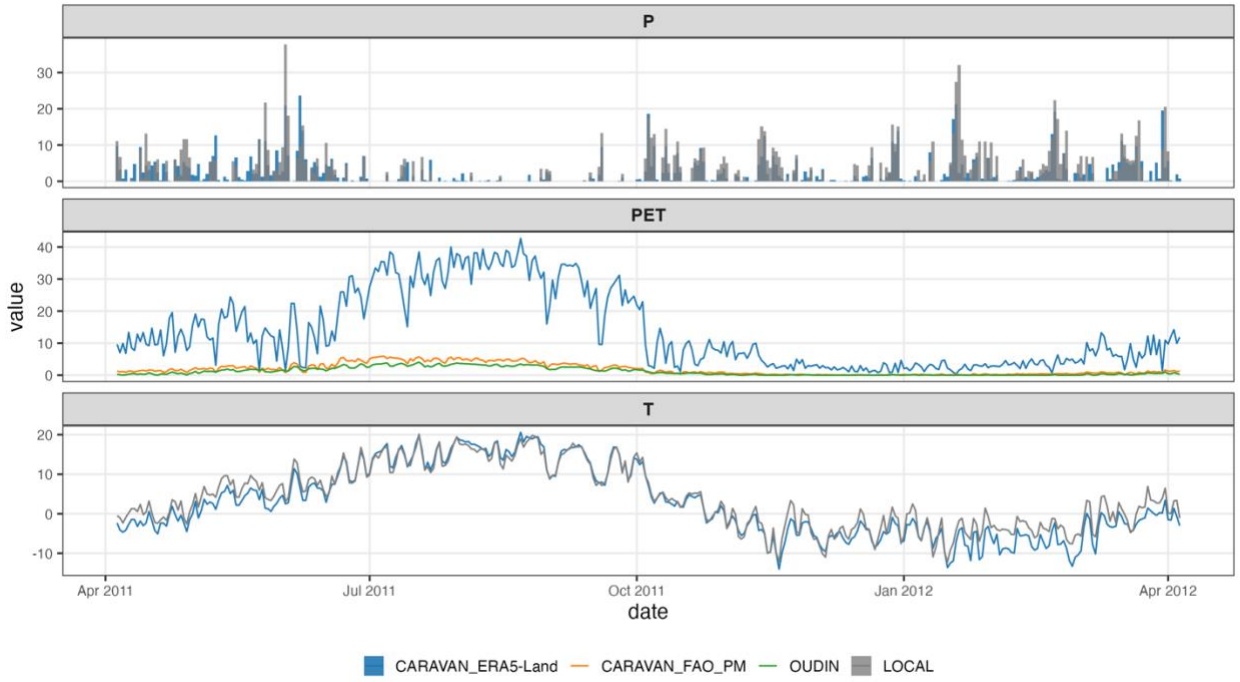

Figure 2: Comparison of precipitation, potential evapotranspiration and temperature for catchment HYS

Generally, we find that the largest deviations are seen for PET, which appear to be systematically overestimated in ERA5-Land. Hence, we decided to use the Penman-Monteith implementation now instead. For temperature, local and CARAVAN data tend to align closely. For precipitation, there were some deviations with the local data, but they were not systematic and general agreement was good, and we therefore kept the precipitation data from CARAVAN, which also simplified the modelling setup.

Did you check these 10 catchments datasets? You mention possible processes that are missing from the models and may cause that, but did you actually check that these processes are indeed at stake here?

As indicated in the response to the previous question and initial commentary, we have adjusted the PET forcing based on a testing of all catchments. Based on our inspection of the timeseries for all basins, we did not find concrete reasons to replace the precipitation or temperature data.

Figure S20 and uncertainty: The results presented in this figure are problematic and illustrate the danger of studies with huge numbers of calibrations/simulations. Indeed, in such cases, authors

cannot verify all calibrations/parameter values/simulations. In addition, authors might not be experts of all catchments or models they used. In this specific case, as a very regular user and developer of the GR4J model, I see that parameter values for all but the X2 parameter (and to some extent X3) indeed present very stable patterns for a large number of OFs. However, to me, the reason why it is so is not that uncertainty is not a large concern, it is simply because the optimisation algorithm reaches the boundaries of parameter values that the authors allowed or that are reachable. We see that X1 is often equal to the minimal value allowed (close to 0 mm), X3 reaches in two cases the maximal value (300 mm) and X4 is almost always equal to 0.5 d, the minimal value. This, according to my experience of GR4J, illustrates a few things. First, it illustrates that the catchments are small, therefore the catchment time concentration is rather short. With larger catchments, we would not necessarily have such a concentration of X4 values close to 0.5 d. Second, the X1 value is rather unusual (almost 0 in most cases) and highlights an issue. To make the simulations correspond to the observations, the optimisation algorithm proposes low X1 values, therefore minimising evaporation. In the same time, the X2 parameter adds a rather large amount of streamflow (values often around 4 mm/d, usually X2 values are negative or very lightly positive). This indicates a water balance issue. Not knowing this catchment, I do not know if this is a data issue or a specificity of the catchment processes, which could be missing in GR4J. We could accuse the OFs, especially when log transformations are used, because those lead to simulated time series that focus on low flows only, and therefore neglect completely certain processes in the model. However, some OFs with no log transformations are also affected. To conclude, seeing such a result i) does not help justifying that parameter uncertainty is not a large concern, and ii) raises interrogations about how the general results of this study might be impacted by such undetected dubious behaviours of parameter optimisation.

We appreciate your insight into the specific behaviour of this model. For context, MARRMoT mimics but does not 1:1 replicate the source models it includes, and the behaviour of MARRMoT's m07 is not identical to that of the airGR4J implementation (see Fig 4b here: https://gmd.copernicus.org/articles/12/2463/2019/), and this is why the models are typically referred to by their MARRMoT ID (i.e., m07 in this case) rather than their original name.

We investigated this issue after changes to the MARRMoT source code (we fixed a bug relating to X4) and the change in PET forcing. Specifically related to m07/GR4J, it seems that the issue for the basin (C03) shown has reduced significantly, predominantly through the improved PET estimation. Other catchments still possess extreme parameter values (see figure below, which only shows runs exceeding benchmark performance (up to the top-500 parameter sets retained per model). The layer cake appearance of the optimized parameter values happens because the figure combines results from all different objective functions. Depending on the objective function, e.g. parameters for the X2 parameter can vary largely (AUS1, BR6) or be consistently low (GB3). Hence, we conclude that the current setup is an improvement over the previous one, but some limitations persist and we therefore investigated this for all other models as well.

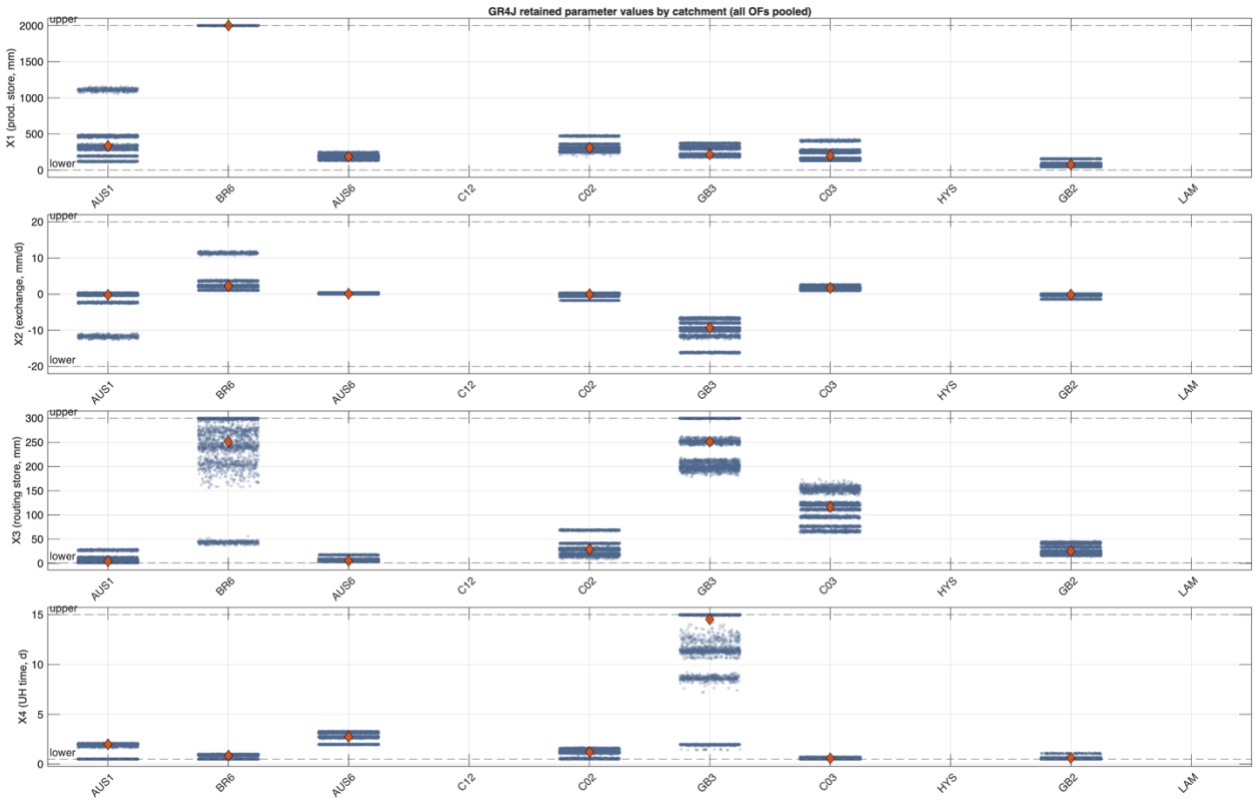

Figure 3: Overview of Parameter Sets for GR4J Model across Basins. The red dots show the median across all runs, blue dot indicate the single runs.

On average, around 10% of all parameters hit a boundary across all the 18,800 conducted runs. We found no indication of systematic issues with selected basins, models, or objective functions. Given (1) the broad range of applied models and the hydroclimatic diversity in the selected catchments, (2) the fact that parameters can be insensitive in a given catchment (e.g. degree day parameters in a snow-free location), and (3) that MARRMoT uses literature-based, standardized parameter ranges to limit differences between individual models, we believe that these results are acceptable, or at least illustrative of the difficulties inherent to evaluating conceptual hydrologic models in large sample and large model studies. Unfortunately, there is little literature to contextualize our findings in the large-sample context. We recognize the importance of this comment, however, and agree that this is an issue that more attention should be paid towards. We have added a brief discussion of the above to the "Limitations and Future work" section (lines 569-574).

Minor comments:
Figure 1: plot b) does not seem necessary. In addition, pie plots are misleading, barplots must be preferred (see e.g. https://scc.ms.unimelb.edu.au/resources/data-visualisation-andexploration/no_pie-charts)
We adapted the figure accordingly and removed the pie plot.

Section 2.1.1: I think I missed the information regarding the time step of the hydrological models used in this study (although I guess this is a daily time step)
We mention that these are daily models in line 191 (first sentence in model calibration procedure)

The numbering of sections in the Supplementary Material 2 is wrong, it should be 2.1, 2.2… instead of 1.1, 1.2…
We fixed this.

Section 2.1.2: In this section the objective functions selected for this study are presented. To be fair, we could qualify the justification of the choice of these eight objective functions as "Generalreasoning", using the authors' classification used in Figure 1. I am also quite surprised that no multicriteria objective functions were selected, although one of the objectives of this work was to potentially identify an objective function that could potentially be relevant for a reasonable range of signatures. Streamflow transformations are also rather neglected in this work. Two log transformations are used, that's all. That might deserve discussion, as transformations are a useful mean to impact the hydrological signatures simulated by models.
We deliberately decided to only use single-objective calibration in this study. We now provide a clearer justification for this in the introduction (line 68-70). Our goal is to learn about the specific weaknesses and strengths of individual metrics, which can then provide information on how to potentially build complementary multi-objective objective functions that can compensate for individual metric weaknesses.

We agree that the selection of the eight tested metrics is based on very general reasoning and added a sentence that many more metrics could be tested as well (lines 111-113). In the updated submission, we additionally replaced one of the eight objective functions with a power law streamflow transformation to include an additional streamflow transformation approach and remove the problematic log KGE metric. Furthermore, we briefly discuss the role of transformations in the discussion (lines 529-534)

Line 171: prefer the word evaluation.
This was adjusted.

Line 175-176: I am not sure to understand: do you mean the interannual regime of daily mean flow vs the annual average flow? Line 243 does not make it much clearer unfortunately.
Yes, we meant the daily mean flow. We have updated the section on benchmarking and hope the new text is now clearer.

Line 177: "we compute this benchmark": actually, you compute the performance of the benchmark, as you define the benchmark as the daily mean flow.
Line 178: "benchmark" -> performance of benchmark
We have updated the section on benchmarking and hope the new text is now clear.

Section 2.4: Please provide the equations of all signatures in SM
This was added to the supplement under section S2.5

Line 204: please consider replacing "the location" with "catchment j"
Line 226: "section" 2.5.1
Line 227: please replace "model (l)" with "model (i)"
We have changed the text accordingly. Thanks.

Line 258: Does that mean that for some catchment / OF combinations, there are no models remaining?
This was indeed the case. Due to the new PET data this is not the case anymore in the updated version.

Figure 3: What are the large circles? What happened for model LAM and OF DE? Where are the dots and the violin?
The large circle depicts the median of the distribution. We have now added this to the caption of the figure. The updated plot does not have any empty spots anymore and should be more visually appealing.

Line 342: Please consider putting that in the caption
The caption was updated accordingly.

Line 352: Please make sure that all SM elements are provided in the order that they are cited in the manuscript
We have checked this and it should be in the right order now.

Figure 7: Please make the lines wider
This was adjusted.

Line 362: How does that analysis deal with the fact that you have different numbers of items forcatchments, models and OFs? Does that influence the importance of your components? Isn't this result also affected by the actual variety in models, catchments and OFs? Wouldn't this result be impacted by different selections of models, catchments and OFs? If so, to which extent?
The random forest analysis was performed using median signature values and errors from each combination of catchment, model and objective function. The benchmark filtering influences the resulting feature importance as poorer performing combinations are removed. Comparison with an alternative analysis including all combinations (see figure 4)

indicated that this filtering tends to increase relative importance attributed to the catchment while reducing importance of model structure and objective function (mainly because the large impact of very poorly calibrated models is taken out). Overall patterns remained similar, particularly for the RF training on signature errors (right) over signature values (left).

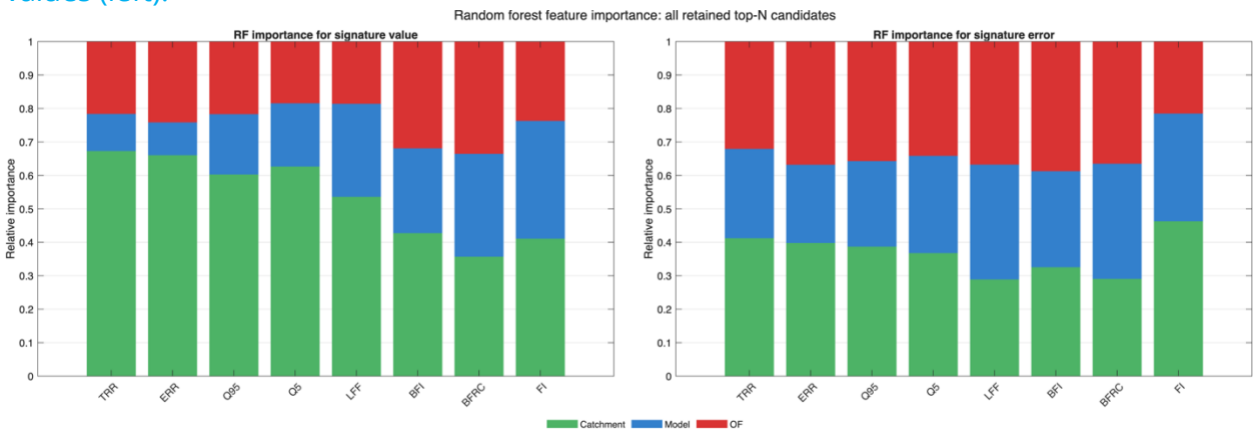

Figure 4: Alternative Random Forest without Benchmarking routine

Run-level variability (i.e. different parameter sets) cannot be included directly as an explanatory factor. Individual retained runs do not correspond to a systematic categorical or continuous predictor analogous to catchment, model, or objective function, so including them as independent observations would mix parameter-level variability with the factors of interest and obscure attribution. Run-level variability was therefore analyzed separately throughout the study (particularly shown in the results) using the top-N ensembles.

The reviewer is of course right that the specific selection of catchments, OFs and models will also be influential on the results, but that is to some extent unavoidable. We have aimed for a reasonable spread in hydroclimatic conditions, model structures and OFs to ensure conclusions are not overly skewed in a particular direction. Assessing the sensitivity of the RF analysis to these choices seems beyond the scope of the paper.

Line 416-418: I agree with that! It's a pity that no multi-criteria OFs were included in this analysis, though. They could help better considering the multiple aspects of the flow regime.
Agreed, but as argued before we wanted to provide insights on individual metrics to help others build more informed multi-criteria OF. Hopefully the changes to the text we discussed before clarify this sufficiently.

Line 422 and 521: I cannot agree with that, KGElog should definitely be discarded from model calibration possibilities, and I would not recommend it.
Thank you for bringing this to our attention. Based on this comment and the findings from Santos et al., 2018, as well as Thirel et al., 2024, we have decided to remove the log KGE from the study, replaced it with a power-law transformation, and re-calibrated all models for all basins.

Figure S20: This figure is quite difficult to read, please increase fonts. Please also provide a self explaining caption. Shall we understand that 6 different calibrations were done for each OF?
The plot was removed as we now consider parameter uncertainty across the entire study.

Line 521-522: "When multiple…": Authors should make clear that this was actually not shown in the study, and that this assertion is just an extrapolation based on the results of the present work.
This was adapted.

References:

Clerc-Schwarzenbach, F., Selleri, G., Neri, M., Toth, E., Van Meerveld, I., & Seibert, J. (2024). Large-sample hydrology – a few camels or a whole caravan? Hydrology and Earth System Sciences, 28(17), 4219–4237. https://doi.org/10.5194/hess-28-4219-2024

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., & Woods, R. A. (2020). A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments. Water Resources Research, 56(9), e2019WR025975. https://doi.org/10.1029/2019WR025975

McMahon, T. A., Peel, M. C., Lowe, L., Srikanthan, R., & McVicar, T. R. (2013). Estimating actual, potential, reference crop and pan evaporation using standard meteorological data: a pragmatic synthesis. Hydrology and Earth System Sciences, 17(4), 1331–1363. https://doi.org/10.5194/hess-17-1331-2013

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., & Loumagne, C. (2005). Which potential evapotranspiration input for a lumped rainfall–runoff model? Journal of Hydrology, 303(1–4), 290–306. https://doi.org/10.1016/j.jhydrol.2004.08.026

Santos, L., Thirel, G., & Perrin, C. (2018). Technical note: Pitfalls in using log-transformed flows within the KGE criterion. Hydrology and Earth System Sciences, 22(8), 4583–4591. https://doi.org/10.5194/hess-22-4583-2018

Thirel, G., Santos, L., Delaigue, O., & Perrin, C. (2024). On the use of streamflow transformations for hydrological model calibration. Hydrology and Earth System Sciences, 28(21), 4837–4860. https://doi.org/10.5194/hess-28-4837-2024

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?

We agree that our conclusions can be more specific. We believe that the main addition our study provides compared to previous studies is the guidance across models and catchments on the question of which metric has which effect on hydrological signature representation. In similar studies that have looked into this question (Table 1), the number of model structures and also the type of hydrological regimes considered tends to be somewhat limited. Results are therefore difficult to generalize across studies using different models in different hydrological regimes. Additionally, the different experimental setups across these studies make a generalization from previous results difficult. We believe that the consistent experimental setup of our study provides a considerable benefit for generalizing our insights across models and catchments compared with previous studies.

While a number of studies conclude that we need to pick metrics based on modelling purpose, we are not aware of any other study that provides similarly direct and generalizable insights on the strengths and weaknesses of different objective functions

regarding signature representation. We were also able to verify or falsify findings of the more specific experimental setups used in the studies of Table 1. For example, we were able to confirm the ability of the KGE to represent high flows better than NSE (Mizukami et al., 2019), who use 2 models across the USA, whereas we show that these findings hold for another 47 conceptual models, strongly suggesting that these findings are broadly applicable and not specific to Mizukami's model and catchment selection. In contrast to this, we could not clearly attribute the insensitivities to low flow that Althoff and Rodrigues (2021) related to the NSE. While this may be true for their single model tested in Brazilian savanna catchments, this observation does not generally seem to hold across our 47 models tested in different hydrological regimes.

We have clarified these advances in the abstract, conclusions, and throughout the text and hope that our work can help modellers make a more informed decision on purpose-aware metric selection than previously possible.

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 already an existing area of research (Yilmaz et al., 2008; Euser et al., 2013; Wagener and Montanari, 2011) in the discussion (lines 555-561). Our work provides generalizable insights that may help determine a complementary metric-metric or metric-signature combination. However, such approaches are both somewhat rare compared to metric-based optimization and also somewhat ad-hoc (in the sense that there is some general dissatisfaction about the signature simulations produced by metric-based calibration, but limited literature to base improvements on). We therefore deliberately decided to test 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 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.

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. Table 5 was updated accordingly, also following a suggestion from the community comment by Bettina Schaefli.

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.

It might also be possible to relate signatures to specific processes (as done, for example, in McMillan, 2020; McMillan, Gnann and Araki, 2022), but such work is still in its early stages and remains a bit speculative. The fact that conceptual models typically only simulate a subset of all processes explicitly might be an interesting angle to consider here (e.g., if metric A influences signature B which is linked to process C which is not included in model D, while model D also performs poorly on metric A, this might suggest that model D should include process C), but this seems like a separate investigation rather than an integral component of our current study.

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, and we do not mean to advocate for the usage of a single-metric calibration. As mentioned before, single-objective calibration however remains common practice, due to its simplicity and possibly also due to the lack of clear guidance on how to define effective multi-objective criteria. We believe that our premise to evaluate commonly used metrics in

isolation has value in raising awareness of these metrics weaknesses and can inform the selection of informative multi-criterion objectives. This is reflected in the text through an extended discussion of limitations (lines 549-555).

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.

You are right that this information was missing. We now include it in Table 5 as well as in the supplementary material, where now all signatures have a detailed descriptions including their equations.

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

McMillan, H. K., Gnann, S. J., & Araki, R. (2022). Large scale evaluation of relationships between hydrologic signatures and processes. Water Resources Research, 58, e2021WR031751. https://doi.org/10.1029/2021WR031751

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

Community comments

We are grateful to Keith Beven (CC1, CC3), John Ding (CC2), and Bettina Schaefli (CC4, CC5) for their engagement during the open discussion. We responded in detail online (AC1, AC2, AC3, AC6) and summarise the resulting manuscript changes here.

Prof. Beven (CC1/CC3) raised the broader treatment of data, parameter, and epistemic uncertainty. While a full uncertainty assessment remains beyond the scope of this study, the revision engages the concern directly on several fronts. Most substantively, we recalibrated every model–catchment–objective-function combination across five random seeds and now retain the top 500 parameter sets rather than a single optimum, so that the near-optimal parameter spread Prof. Beven highlights is carried explicitly through the analysis (Figures 3–6). We agree that defining fitness-for-purpose under data and model uncertainty is the harder and more valuable question, and have flagged it as a priority for future work (line 581-584) while being clear that current common practice, which this study deliberately mirrors, remains somewhat divorced from that ideal.

Prof. Schaefli (CC4) noted that some signatures are inherently input-data-dominated while those reflecting partitioning or storage-release functions are more likely to be OF-sensitive. We now address this in the discussion (lines 472-475) and have refined the overly generic "streamflow" category in Table 5. We have also corrected the terminology ("hydrological response" rather than "hydrological regime") and repaired the incomplete sentence (line 510-513).

John Ding (CC2) suggested a missing reference, which we have added (Melsen et al., 2025). He also prompted an enjoyable excursion into the history of the Nash in Canada, including a search for who invited Nash to lecture in Ontario in 1968. We were unfortunately not able to get very far during the public discussion phase. What we found is given below.

Although not directly related to the work conducted in this study, community comment 2 posed a question that was difficult to resist putting some effort into.

Q: Whose idea it was to invite Nash of Wallingford, UK, and/or Galway, Ireland, to visit Guelph? Their identity remains buried in archives. The search continues.

We propose that Donald McMullen is a likely candidate.

Nash' lecture notes indicate he was "Ontario Visiting Professor of Hydrology under the auspices of The Ontario Committee for the International Hydrological Decade":

"I would like to express my deep appreciation of the many kindnesses shown to me by my colleagues at the Universities in Ontario [Guelph, Ottawa, Queen's, as per the manuscript's

cover page] *where I had the privilege of giving these lectures and to the Ontario Committee for I.H.D., who were responsible for inviting me to Canada. My particular thanks is due, however, to all those who so patiently listened to me lecture, at times extremely unclearly."*

- J. E. Nash (1968-1969). A course of Lectures on PARAMETRIC OR ANALYTICAL HYDROLOGY.

It is rather difficult to track down the members of the Ontario Committee for the IHD. Membership of the Canadian Committee for the IHD can however be found from the Government of Canada: https://publications.gc.ca/site/eng/9.900086/publication.html. Of note, this lists Prof. H. D. Ayers as a committee member employed at the University of Guelph. Neither of the other universities Nash visited (Queen's and Ottawa) have a representative on the committee, at the time of this report (1964).

A later report (1974; https://www.govinfo.gov/content/pkg/CZIC-gb1627-g8-l82-1974/html/CZIC-gb1627-g8-l82-1974.htm) provides another list of members of the Canadian Committee for the IHD. Prof. Ayers is still present, and so are several others, suggesting that these sources are reasonably accurate. Of note is the appearance of D. N. McMullen, listed as "Secretary, Ontario Committee for THD". Assuming "THD" is a typo for "IHD", we have found our first confirmed member of the Ontario Committee. The foreword of Volume 8, Number 1 of the Atmosphere journal (Canadian Meteorological Society, 1970) confirms he was indeed involved (https://cmosarchives.ca/Atmosphere/a0801.pdf).

This D. N. McMullen is Donald McMullen (https://cmosarchives.ca/BIO/dmcmullen.html) who, among his many achievements, formed the "Toronto Hydrology Group" that had the aim to bring together *"public and university sectors for interactive lectures on hydrology and general water management"*. McMullen was president of the group from 1965 to 1977. Combined with his role as secretary of the Ontario Committee for IHD, this makes him a likely candidate for inviting Nash to visit and present those lectures, possibly with Prof Ayers acting as host whilst at the University of Guelph.

This is, of course, entirely speculative but it was fun digging through this part of hydrologic history a bit and perhaps there is something here to help others answer this long-standing conundrum.