

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. I have several comments, some of them being major, and others being more minor.

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 users. 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.

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.

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.

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.

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?

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? 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?

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.

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-and-exploration/no_pie-charts)

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)

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

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 "General reasoning", using the authors' classification used in Figure 1. I am also quite surprised that no multi-criteria 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.

Line 171: prefer the word evaluation.

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.

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

Section 2.4: Please provide the equations of all signatures in SM

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)"

Line 258: Does that mean that for some catchment / OF combinations, there are no models remaining?

Figure 3: What are the large circles? What happened for model LAM and OF DE? Where are the dots and the violin?

Line 342: Please consider putting that in the caption

Line 352: Please make sure that all SM elements are provided in the order that they are cited in the manuscript

Figure 7: Please make the lines wider

Line 362: How does that analysis deal with the fact that you have different numbers of items for catchments, 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?

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.

Line 422 and 521: I cannot agree with that, KGElog should definitely be discarded from model calibration possibilities, and I would not recommend it.

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?

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.