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.

Thank you for these thoughts. As indicated in our response to Prof. Beven, we understand his frustration that it is still not common to address multiple sources of uncertainty simultaneously. However, we agree with your own statement that the way we approach

parameter uncertainty (by simply optimizing model parameters through a calibration algorithm) is in line with many studies we still see published. In addition to your own explanations for why this is still commonly done, we also think it is worth mentioning the extra dimensionality that accounting for parameter uncertainty would add to our existing work. The analysis already covers models, basins, objective functions and signatures, adding parameter uncertainty would increase the scope of the study beyond what is feasible. We believe this is a tricky situation many scientists face: despite knowing that accounting for all sources of uncertainty simultaneously would lead to more robust conclusions, it remains difficult to do so in practice.

We see the solution for our current study in a more elaborate theoretical (and partly practical) consideration of uncertainty as outlined in the response to Prof. Beven as well as further below. But we also believe that it might be helpful for the community to develop guidelines on minimum standards for uncertainty quantification in hydrological modelling studies, as explicitly addressing uncertainty should be considered good scientific practice. While we will not be able to solve this issue in this study, we will expand the discussion on the influence of parameter and input uncertainty throughout the paper.

Regarding the question on what we can learn from such a large-sample study, we still believe there is value in exploring the relationship between objective functions and signatures in a setting that mimics common applications (i.e., calibrate once) but with a much broader model sample than is typical, because this gives insight into whether a model calibrated on a specific metric is generally "fit for purpose". We particularly appreciate Dr. Schaefli's note that this is in fact a helpful application of our work. We believe that identifying systematic tendencies is helpful for providing guidance to broader applications, raises awareness of potential influences that should be considered, and points to directions that might need further investigation.

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.

We agree that an additional paragraph to introduce the previous endeavors and their outcomes in more detail would be beneficial. We will add a more in-depth analysis on the

conclusions of the individual papers and relate this more directly to our findings. This might serve as an initial step in the meta-analysis the reviewer describes.

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.

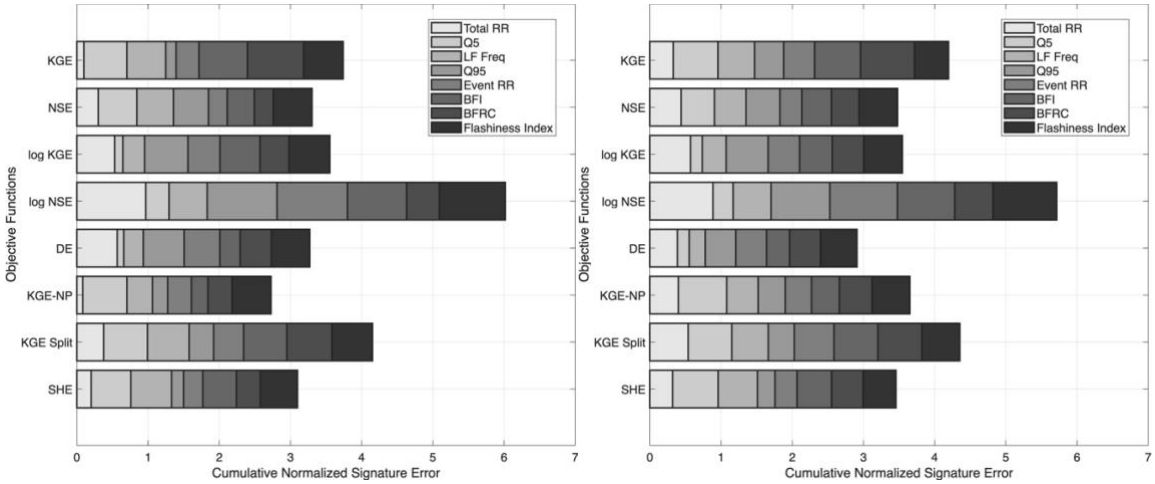
Excluding catchments larger than 1000km2 was a deliberate choice in this study to ensure signature values such as flashiness are not dampened by the 'smoothing out' of hydrologic response in larger basins, We will modify the explanation of this decision the reviewer cites to clarify this point, and emphasize that the selection procedure is informed by hydrologic reasoning, and not a purely random selection of the basins available in the Caravan dataset.

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.

We understand the reviewer's point that the signature values are at times different during evaluation compared to calibration. We performed a preliminary analysis using validation data instead, and found that conclusions are broadly robust but different in some specifics (see figure). When comparing one of the plots in the discussion between calibration (left) and evaluation (right) the attributed skill is typically less strong in evaluation compared to calibration (as was to be expected), and the strengths and weaknesses get more diluted. Additionally, the best overall metric shifts from KGE-NP to DE, indicating a more consistent performance for DE.

The main messages remain, however, showing that objective functions have different skills, and that the choice of OF should be guided by the purpose of the modelling endeavor. Furthermore, the recommended best objective functions (of Table 6) remain similar, e.g. KGE for high flows or KGE-NP/DE for Baseflow Index.

We will repeat the full analysis currently in the paper with evaluation data and ensure the reader has access to these. We will specifically address the robustness of findings in the discussion.



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 will add text to the model description section clarifying that these models use their own snow model. Practically, most of these rely on the degree-day snow model implementations. The specific models selected in the three catchments with the most snow all possess a snow module, mostly based on a degree-day approach. These models are 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 10 out of 47 models, so GR4J is not coupled with a snow module in our setup.

There is one exception: for the basin selected from the HYSETS dataset, 35 models exceed the benchmark for log KGE, many of which lack a snow module. This happens because the bias term is rather close to 0 for the logarithmic values, and the bias term of the log KGE becomes large. This significantly decreases the performance compared to the default KGE in agreement with Santos et al., 2018. Therefore, the large number of models exceeding the benchmark is an artifact of the logarithmic transformation. Regarding one of the minor comments, this is probably a case study for why the log KGE should not be used as an objective function, and e.g., KGE(1/Q) should be preferred to emphasize low flows.

To address this pragmatically, we agree that it is appropriate to proceed with the analysis only using models with a snow module in basins with snow. Therefore, we will include an

additional methodological step where we exclude models without a snow component from basins with snow fraction >0.1.

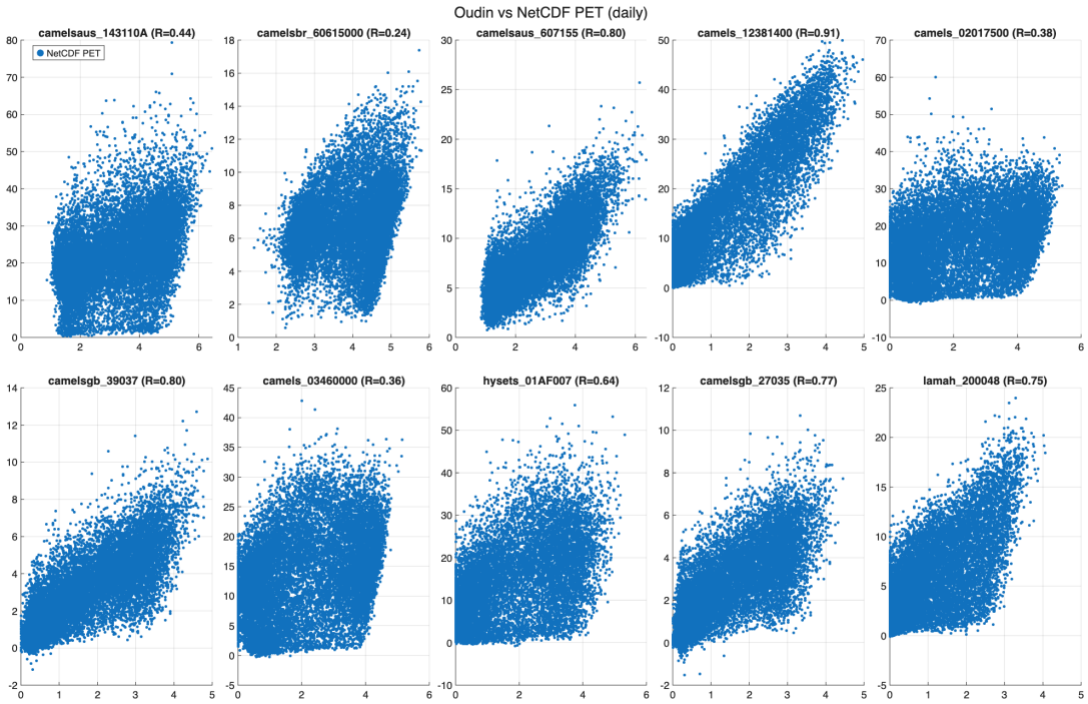
Line 254-258: While I can definitely understand that for some models/OFs/catchments combinations the benchmark outperforms them, I am surprised that in some cases all models fail on the same catchment! What about the observed data quality in those cases?

The specific case the reviewer refers to is basin C12. Somewhat related to the previous question, C12 is one of the basins with strong snow influence, therefore the benchmark is typically high (>0.85 for both log NSE and SHE) and the pool of plausible models is already small. The reason is a strong seasonal streamflow pattern, which can be predicted well by the interannual mean. This is in line with findings in Knoben et al. (2020; Section 4.1), who find that in 11 snow-dominated basins in the US, no MARRMoT model is able to outperform the simple daily-mean benchmark.

However, based on the reviewer's next comment, we also investigated data quality for all basins. Specifically, we compare the PET estimates from CARAVAN to a simple estimate from Oudin et al., 2005. For the C12 catchment (first row, 2nd plot from the right), the CARAVAN PET values were much higher than those estimated by Oudin. The underlying temporal pattern is good (corr=0.91), but the values are larger by a factor of 4.6 on average relative to the Oudin values.

As PET can be estimated in many ways (McMahon et al., 2013), it is not immediately clear if one is preferable over the other. It will also be important to check whether the performance and streamflow characteristics stay similar with different forcings and if these model parameters compensate for differences in PET, which we have begun to investigate for the GR4J model (see later in this document).

We have started to investigate this for all catchments in the analysis (please see the figure below).



Generally, the PET estimates from CARAVAN are higher than those from Oudin. Both the temporal fit and the absolute values can be rather close (GB27) or very different (AUS1, BR6). For neither is it immediately clear which estimation is "better" as potential evapotranspiration can be understood in different ways, and the estimation of PET in ERA5-Land is more complex than Oudin's formula. Hence, we suggest that the more plausible parameter sets (and potentially higher performance) will indicate the more appropriate choice of PET estimation, which will have to be tested on all catchments and multiple models to find a general conclusion.

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?

This is a very broad question. Generally, we assumed the data provided in a community dataset, such as CARAVAN, were adequate. But we are also now aware of issues with the PET calculation from ERA5-Land used in CARAVAN (Clerc-Schwarzenbach et al., 2024).

Regarding further relationships, we will check the forcing data that was used from the CARAVAN dataset, specifically potential evapotranspiration, precipitation and temperature. So far, we can clearly see a strong relationship between PET and precipitation across all catchments, presumably stemming from cloud coverage estimates that lead to lower PET values.

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.

The issue pointed out in this paragraph is very important, and we appreciate your insight. We started investigating the impact of the forcing data on the parameters and model performance for our model setup with two initial steps:

Firstly, there was a bug in the MARRMot implementation of GR4J, which could affect the parameter set (*https://github.com/wknoben/MARRMoT/issues/54*). We found that the new MARRMoT GR4J version has a minor influence on the parameter and objective function values for 3 of the 4 catchments tested so far.

Secondly, based on the previous analysis, it seemed appropriate to check the influence of the PET forcing on the parameter sets because of the known issues with Caravan. We will test that for all catchments, below are the initial results for 4 of the 10 catchments.

We tested a simple Oudin equation in comparison to PET from Caravan, here we saw that:

- For C03, Oudin PET leads to more sensible parameters (X1 around 120mm instead of close to minimum, X2 smaller, but still positive) than CARAVAN (ERA5-Land)

- For GB39, differences are much smaller, as PET estimates from Oudin and ERA5-Land are much more similar. X2 switches from positive to negative.
- For C12, we saw extreme values for X1 and X3 with CARAVAN PET, but X3 also converged to the maximum value when using Oudin's formula.
- For GB27, parameter ranges are comparable between Oudin and CARAVAN.

Hence, we conclude that the change in PET can have a large impact on parameter sets in GR4J. We will proceed to check whether this has to be addressed through the usage of a new PET estimation (which would require recalibration of all models) or can be addressed through less computationally demanding adjustments, such as replacing singular basins or a more thorough benchmarking procedure (e.g. checking if parameter bounds are reached).

We will also investigate additional models to determine whether this effect is specific to GR4J. In any case, we will extend the discussion of the impact of input data on our findings.

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)
Thank you, we will make the suggested replacement.

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 will make sure to add this information more prominently.

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 will fix 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 "General reasoning", 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 objective functions. This was done to learn about the specific weaknesses and strengths of individual metrics which can then provide information on how to potentially build 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 will highlight that many more metrics could be tested. We will also mention that we have not investigated streamflow transformations despite their usefulness as nicely investigated in Thirel et al., 2024.

Line 171: prefer the word evaluation.
We will adjust the formulation.

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.
We will clarify the wording here.

Line 177: "we compute this benchmark": actually, you compute the performance of the benchmark, as you define the benchmark as the daily mean flow.
We will clarify the wording here.

Line 178: "benchmark" -> performance of benchmark
We will adjust the wording.

Section 2.4: Please provide the equations of all signatures in SM
We will add the equations.

Line 204: please consider replacing "the location" with "catchment j"
We will adjust the wording here.

Line 226: "section" 2.5.1
This will be changed.

Line 227: please replace "model (l)" with "model (i)"
This will be changed.

Line 258: Does that mean that for some catchment / OF combinations, there are no models remaining?
Yes – there are combinations of catchments and objective functions for which no model exceeds the interannual mean as a benchmark. This is similar to findings by Knoben et al., 2020 as mentioned before.

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. For the LAM and DE combination, the objective function value is very low for one model, which affects the violin fit. We will investigate the cause of the low DE value and adjust the plot accordingly.

Line 342: Please consider putting that in the caption
We will adjust this.

Line 352: Please make sure that all SM elements are provided in the order that they are cited in the manuscript
We will adjust this.

Figure 7: Please make the lines wider
We will adjust this.

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?
Thank you for bringing this up. The only variable that changes throughout the analysis is the number of models per combination of catchments and objective functions based on

the performance compared to the benchmark performance. This was the main reason for conducting most of the analysis with a metric that incorporates the median values of signature distributions. This allows us to gather representative values while incorporating the model performance and should not influence the results methodologically.
Further analysis was either accounting for the spread in the variables (e.g. significance analysis) or was using the full set of signatures (e.g. random forest feature importance).

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.

Line 422 and 521: I cannot agree with that, KGElog should definitely be discarded from model calibration possibilities, and I would not recommend it.
Based on the preliminary analysis conducted above and the findings from Santos et al., 2018, we agree with you on this point. There are clear issues related to the use of log KGE, and it should therefore not be recommended. We will also clarify this in the discussion, particularly related to Table 6. Alternatively, DE should be preferred.

Figure S20: This figure is quite difficult to read, please increase fonts. Please also provide a selfexplaining caption. Shall we understand that 6 different calibrations were done for each OF?
Based on the wider discussion, we concluded that the figure does not serve its intended purpose and will therefore remove it. For completeness, the conducted calibrations differed only in the random seed chosen at the beginning. If the performance is close to the highest objective function values that were found, the model setups typically converge to the same parameter set.

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 is correct, and we agree that our initial phrasing could be misinterpreted. We will change this.

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