Reviewers' comments are in italics. Our responses are in blue.

Dear Editor,

We send you a revised version of our manuscript and a reply to the review comments we received.

We would like to stress that the first reviewer who had already reviewed the manuscript is satisfied with the revised version, and that the major modifications asked by the second reviewer are linked to the fact that he/she did not reviewed the revised version but the initial manuscript submission. Most of the modifications requested by this reviewer were already addressed in the past two review rounds. Therefore the new revised version includes only minor modifications, corresponding to the remaining suggestions not already addressed (see our reply for more detail).

We hope that this confusion will not lead you to ask for the article to be reviewed again by external reviewers, given the efforts we made to comply with the comments received in the three review rounds, and that you will be able to make a decision on the basis of our responses and the changes made in the article.

We thank you and the reviewers for your continued efforts to help us improving this manuscript.

Sincerely,

Guillaume Thirel

Reviewer #1

The authors addressed my concerns. There are some small technical things that would be good to correct. There is a bracket missing in Table 1, it would be good to add to the caption of Figure 2 that this is a map of France and to explain the small land area in the bottom right. A final thorough look through the paper would be good to catch such things.

We thank the reviewer for its assessment of the manuscript.

We added the bracket that was indeed missing. We also added in caption of Figure 2 that this is a map of France.

The small land area in the bottom right is the Corsica Island, which is part of France. We believe this is not necessary to mention it.

Reviewer #2

The review of the manuscript "On the use of streamflow transformations for hydrological model calibrations" submitted to the Hydrology and Earth System Sciences.

The manuscript focuses on the analysis of the choice of the objective function and streamflow data transformation for the calibration and validation of the hydrological models. This study analysed three objective functions and 11 transformations using data from 1985-1995 and 1995-2005 from 325 catchments in France. The outcomes are presented for the specific single catchment, the Fecht at Wintzenhein, and averaged over 325 catchments.

The topic is very important and well-suited to the journal. I have some doubts regarding the applied methodology, the presentation of the results, and the quality of the discussion that should be improved.

We thank the reviewer for his/her assessment of the manuscript. We must however say that we were at first rather surprised by some comments, as they did not seem to correspond to the content of the revised version of the manuscript we had sent to the editor after the previous review round. We then understood that the review comments were made on the initial version of the manuscript, which explains why they are not necessarily relevant anymore. We however tried to address them below, or to refer to our previous answers to reviews from Rounds #1 and #2 of reviews.

General comments:

(1) *The results are presented in the form of ranks; therefore, it is difficult to describe relative differences in the outcomes between tested transformations. The differences between methods may be not statistically significant. So, the ranking of the transformation method will be different when the other criteria are applied.*

We thank the reviewer for this legit comment. We do agree that the ranking of the transformations can and will probably be different when the methodology, including the criteria used to compare them, differs. Consequently, the conclusions we draw are impacted by our methodological choices, as are any results of any study. Regarding the use of ranks rather than another way of comparing transformations, we already addressed this issue (see the Answer to Comment #1 of Reviewer #2 of the Round #1 of review; https://egusphere.copernicus.org/preprints/2023/egusphere-2023-775/egusphere-2023-775-AC2-supplement.pdf). In the 2-page answer to this specific comment, we justified that the "idea behind choosing to work with ranks, instead of direct (normalised) errors was to i) be less impacted by different orders of errors magnitudes between catchments or ranges of streamflows, and to ii) answer the question of what are the best transformations, rather than how good transformations are." To answer the comment properly, we assessed the impact of modifying the methodology, i.e. using absolute differences rather than ranks. Based on results over the 325 catchments, we found "similar results as when working with ranks: the general conclusions are not changed. The same groups of transformations still are the best over the same ranges of streamflows.

As a consequence, we prefer not to change the whole methodology used in the manuscript and we will stick to the one proposed so far." We however modified the manuscript (see the first revised version) to mention this alternative and to discuss it.

Code de champ modifié

(2) *The analysis covers two periods, 1985-1995, for the model's calibration and 1995-2005 and validation period. Are these periods overlapping? There is no description of hydroclimatic conditions during these periods. Was it dry, average or wet years? It is important as later data were classified into 200 intervals, and the ranks were evaluated for these classes.*

There is no overlapping between the periods. We indeed used hydrological years, i.e. from August to July. We added this information in the revised version of the manuscript.

We did provide information about the hydroclimatic conditions of both periods, through the calculation of mean air temperature, mean annual potential evapotranspiration, mean annual runoff, baseflow index, aridity index and aridity seasonality (see Table 2). As mentioned already in the

manuscript, this "shows that the climatic and hydrological conditions are similar between the two periods, with the evaluation period being only slightly warmer and wetter than the calibration period and other indicators showing only slight variations."

(3) *The quality of the calibration results is not addressed. What constitutes 'satisfactory performance'? Were optimal parameter values obtained for all catchments? How was calibration uncertainty considered? Were repeated calibrations conducted, and did they yield similar results (objective function values and ranks)?*

It is true that we did not provide much detail about the calibration results. The reason behind this is that the modelling platform and algorithms (i.e. the GR models, the airGR package and the optimization algorithm) we use have been used for years in many studies in many countries, including France of course, as they are developed there. The mentioned term "satisfactory results" was referring to that. We consequently added some references, two of them already cited in the manuscript, such as Perrin et al. (2003) who developed and assessed GR4J, Mathevet (2005) who compared this algorithm to SCE-UA among others, and Coron et al. (2017) who developed and assessed the airGR package.

As an additional indication allowing to state that calibration results were satisfactory, we calculated the average objective function among all models / objective functions / transformations / catchments, and we obtained a median value equal to 0.88 and an average value equal to 0.84. Please note that such scores only give as a qualitative assessment, since values coming from different objective functions and different transformations are hardly comparable. However, comparing these values to those of Table 3 of Crochemore et al. (2015), we can see that they can reasonably be considered as good or very good.

The question whether optimal parameter values were obtained for all catchments is difficult to answer. It is indeed impossible to answer that question without testing all combinations of parameter values, which is impossible to do due to the infinite number of possible combinations. Due to its deterministic nature, the optimization algorithm does not consider uncertainty and repeated calibrations necessarily lead to identical objective function values and parameters values and ranks. Assessing uncertainty in calibration certainly is an interesting topic, however we believe that it is beyond the scope of the current study.

(4) *Why was the Fecht at Wintzenhein chosen as an example catchment? Does it represent average conditions? Where is it located? Which hydrological model version was used?*

The Fecht River was chosen for illustrative purposes only. As a consequence, we only chose it because it could be used for showcasing the methodology. As we did not aim to use it to draw any conclusions (those are drawn from the results on the 325 catchments), we did not expect this catchment to represent any average conditions. The location of the Fecht catchment was added on Figure 2 in the previous rounds of review, as well as the name of the hydrological model (namely the GR4J model) in the caption of Figures 4 and 5.

(5) *The second part of the results presentation deals with the outcomes averaged over 325 catchments. How similar were the results across the 325 catchments? Were there distinct groups with different behaviours?*

Section 3.3 is already dedicated to an attempt to establish links between catchment characteristics and transformations. Unfortunately, as explained in this section, we failed to identify strong links between catchment characteristics and transformations, and consequently to identify groups with different behaviours.

(6) *The quality of the figures should be improved. In the case of Fig. 5, I recommend using a colour scale. Lines from Fig. 6, 9, 10, 11, and 12 have similar colours, so it is difficult to recognize what they are representing. The text regarding Figure 10 states that the transformation log and 0.2 show the best average ranks for the lowest flows. For me, there are four transformation methods that are characterized by almost the same values. Are these differences in the results significant?*

We thank the reviewer for the suggestion to improve the figures, which is consistent with comments we got in earlier reviews. The figure 5 from the initial submission was removed following reviewers' comments. Regarding initial figures 6, 9, 10, 11 and 12, following a reviewer's comment, we added symbols for the different transformations (see answer A20 in https://editor.copernicus.org/index.php?_mdl=msover_md&_jrl=778&_lcm=oc108lcm109w&_acm=get_comm_sup_file&_ms=110954&c=250088&salt=6749699781714668426). We believe these figures are much clearer now.

Code de champ modifié

Regarding initial Figure 10 (now Figure 8), transformations log and 0.2 show a lower average rank over a large portion of flows (more than half of the range). Although some other transformations are not that far sometimes and although we did not compute statistical tests to assess the significance of the difference, we believe that the fact that this difference is visible over more than half of the flows range makes it worth mentioning.

(7) *I miss the discussion of the results, taking into account other studies. It is necessary to supplement the manuscript with a discussion of the results!*

We added a discussion of the results in the previous round of revision, in section 3.5 (see answer to major comment #2 of Reviewer #2 of the Round #2 of reviews).

(8) *What does it mean "Anti-correlations"?*

This expression, which was only present in the initial submission and was removed since, was used for meaning "negative correlation".

Reference:

Crochemore, L., Perrin, C., Andréassian, V., Ehret, U., Seibert, S. P., Grimaldi, S., … Paturel, J. E. (2015). Comparing expert judgement and numerical criteria for hydrograph evaluation. Hydrological Sciences Journal, 60(3), 402–423. https://doi.org/10.1080/02626667.2014.903331

Mathevet, T.: Quels modèles pluie-débit globaux au pas de temps horaire ? Développements empiriques et comparaison de modèles sur un large échantillon de bassins versants, Phd thesis, Doctorat spécialité Sciences de l'eau, ENGREF Paris, https://hal.inrae.fr/tel-02587642, 2005.