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

Reviewer #1

The authors mainly addressed my comments. Some not as extensively as I would have thought to be useful, but there are.

A reamining point I find very unhelpful is that the authors refer to the BoxCox transformation as such in all tables and figures. However, this is based on a single choice of lambda value. This should be highlighted in tables/figures, e.g. include the chosen value in brackets).
It would also be appropriate for the authors to at least briefly describe in the text that different choices for lambda have been used in the hydrologic literature. I had given suggestions for references in my previous review.

We thank the reviewer for this suggestion. We added the value used for lambda in all captions of the tables and figures where the Box-Cox transformation was used in our work. In addition, we specified the value used for lambda in table 1, presenting a non-exhaustive list of references using streamflow transformations. We also added in table 1 the references the reviewer suggested in his/her previous review. We apologize for this omission.

Reviewer #2

I would like to thank the authors for considering my suggestions. After the first round of reviews, the manuscript is technically much stronger, and the collection of figures and tables supports most of the conclusions. However, I think that the writing needs to be improved to maintain the high standards of HESS. Additionally, one conclusion is unsupported, and the manuscript lacks any discussions on how the results connect with the literature cited in the introduction.

We thank the reviewer for their general assessment of the manuscript, and for their very detailed reading and suggestions, which will improve the quality of the manuscript.

Major comments

1. The number of awkward/redundant sentences is so large that interrupts the flow of the reading in nearly all sections. I think that the authors need to make an important effort in revising the text before this manuscript can be considered for publication. Given all the available tools and resources (e.g., Google translator, Grammarly), a final version written in good English should be easy to achieve.

We thank the reviewer for these suggestions. We had already asked a professional translator to copyedit the manuscript before the initial submission. The revision process may have introduced incorrect style or grammar formulations. We asked for a new correction before resubmission and hope that this will help to meet the standards of the journal.

Some examples:
- L2-3: “It is a widespread technique that has been…”. This sentence is completely redundant and

should be deleted.
- L3: “Indeed” -> “Further”.
- L4-5: “Besides, the actual goal of the model application… is undertaken”. This sentence is redundant and distracting. Delete.
- L8-L9: “Typically, a logarithmic transformation…no transformation”. This sentence is out of place here.
- L14: “intermediate range” -> “medium range”.
- L19: “impact” -> “impacts”.
- L23: “Consequently” and “such as those mentioned above” are completely redundant and should be deleted.
- L25: “on the use of a criterion (sometimes a combination of criteria)” -> “on the use of one or more criteria”.
- L28 :“and two different chosen criteria” -> “and two different criteria”.
- L32: “wide panel” -> “wide range”.
- L33: “…has been introduced in the literature (Bennett et al., 2013). These transformations consist…” -> “…has been introduced in the literature (Bennett et al., 2013), which consist…”
- L46-47: “in a review of suitability…. of low flows” -> redundant and distracting. I suggest deleting.
- L47: “justify” -> “justified”. You should use past tense when referring to what was done in the past.
- L57: “To the best of our knowledge, the use and choice of transformation have not been thoroughly assessed”. Because you quote some studies afterwards, it would be more appropriate to write ”Only a few studies have assessed...” (or something similar).
- L52: delete “calibration that provides”.
- L53: delete “leads to”.
- L75: “possibly identify” -> “explore possible links”.
- L80-81: I suggest writing “Daily meteorological and hydrological data from the period 1985-2005 were used…”.
- L83: “…in Table 2. It illustrates the high diversity” -> “…in Table 2, illustrating the large diversity…”.
- In the caption of Table 2, I suggest removing “statistics of the” from the beginning. Replace the last sentence by “The maps with sample statistics for these catchment features are included in Appendix C”.
- L129: I suggest re-writing as “The hydrological models are calibrated by applying different transformations to streamflow values in the calculation of the objective functions” (or something like that).
- L139-140: I suggest rewriting as “In order to evaluate the impact of the transformations on model calibration, we use a common analysis framework that aims at…”.
- Caption of Figure 3: no need for capital letter after “;” in “…series; b) Absolute…”.
- L166: rewrite as “Figure 4 illustrates an example application for a single catchment…”.
- L190: delete “quite logically”. Let the readers judge on this.
- L209: delete “some trends can be observed”. Or replace by “some general features can be observed” (the word “trend” is typically associated with “temporal trends”).
- L223: replace “This result is interesting since most of the time” by “Typically”.
- L237-238: “However, the purpose of using models is to apply them under conditions that are different from those they are calibrated on.”
- L286-287: awkward sentence. Please re-word.
- L313: “on the basis of” -> “using”.
- L346: “flood peak” -> “peak flows”.
- Appendix A and B: I suggest deleting “On the” from the title.
- L359: “the most error” -> “the largest error”.

All these modifications were made. The reviewer's suggestions that were not strictly followed are detailed below.

- L28: "will impact differently on the calibration process" -> "will impact differently the calibration process".

We rephrased as follows: "will impact the calibration process differently".

- L32: "criteria" -> do you mean objective functions? It would be good idea to be more specific.

Here we actually wanted to be rather general. This sentence is true for objective functions, but also when just assessing the quality of a model simulation, i.e., for a performance criterion.

- L39: "Since many metrics are squared metrics" reads weird. You could replace by "Since many metrics rely on squared transformations…".

We agree. However, we do not want to use the word « transformation » here, as the square operator is not applied to simulation and observation time series, but rather to the difference between these time series. We suggest "Since many metrics rely on squared errors…".

- L42-43: since you have a summary table, you could replace this sentence with something like "A non-exhaustive list of transformations is listed in Table 1, being XX and YY the most popular".

We rephrased as "A non-exhaustive list of transformations is listed in Table 1, with the square root, the logarithmic, the reciprocal of squared root, the inverse or other power–law transformations being the most popular".

- L58-66: I find the number of quotes from past papers excessive. I motivate the authors to describe previous findings that are relevant for their research using their own words.

We understand that the quotes may appear excessive here. However, we wanted to remain as close as possible to the authors' explanations, not to distort the points of view they express. Therefore, we would prefer to stick to the current formulation.

- L237-238: "However, the purpose of using models is to apply them under conditions that are different from those they are calibrated on."

We did not understand the reviewer's suggestion, since it corresponds to the original sentence. If the reviewer found it unclear, we propose rephrasing it as: "However, the purpose of using models is to apply them on periods different from those used for calibration" (see our answer to comment 16).

2. L268-269: "We can therefore conclude that the analysis is only slightly dependent on the model used" (also in L333-334). I think that the experimental setup does not enable to conclude this, unless *they repeat their calibration experiments using model structures with very different degrees of process and spatial complexity (sampling, for example, the model space described by Hrachowitz and Clark 2017 in Figure 1). Despite this is a major concern, it can be easily addressed by replacing that sentence by "For the models used here, the relative performance of transformations is very similar across the streamflow range", and re-wording similar statements in the conclusions section.*

We thank the reviewer for this remark. We agree that models with conceptual differences much more important than those between GR4J and GR6J could lead to additional understanding of the actual contribution of model structure on such a conclusion. While we think that models with different spatial complexities would open another research question, lumped models with different conceptualization exist. Although we believe that the end of the sentence pointed out by the reviewer conveys a similar message as the reviewer's remark (", although we must bear in mind that these models are partially similar"), we modified the sentence as suggested by the reviewer, and reworded the conclusions section.

3. I think that the authors should make an effort to connect their results with the existing literature. This can be done in a separate section named "Discussion", or within the results section, renaming it to "Results and Discussion".

Thank you for this comment. We agree that connecting our results with the existing literature could provide added value to the manuscript. Due to the limited literature about this specific research question addressed in this work, we opted to add a subsection in the results section, renaming it Results and Discussion as suggested by the reviewer.

We also modified the following sentence in the conclusions: "They show that no a priori assumption on streamflow transformations can be taken as warranted." as follows : "They show that, although some common beliefs about the impact of transformations are confirmed by this study, no a priori assumption on streamflow transformations can be taken as warranted."

Minor comments

4. L67-71: I think this is a good place to clearly describe the gap(s) that this study intents to fill or, even better, state the research question(s) justifying the existence of this paper.

The reviewer is right. We attempted to improve this paragraph accordingly.

5. L80: I suggest moving the link to the data availability statement.

Done.

6. Table 2: I don't think you need decimals for altitude. I suggest replacing "mean annual streamflow" by "mean annual runoff", since your units are not volume per time units.

Done.

7. L98-99: Did you check whether excluding CemaNeige affects your results for these catchments?

Not using CemaNeige strongly affects the hydrological simulations when the proportion of snow becomes high, as the hydrological regime cannot be well reproduced. We believe that the missing process (snow accumulation and melt) could be wrongly compensated by transformations and the calibration algorithm. Consequently, comparing transformations on snowfed catchments without using CemaNeige was not checked.

8. I suggest renaming section 2.3 to "Calibration metrics" or "Objective functions", since the optimization criteria also involves the choice of optimization algorithm (which was already described).

Done.

9. L148: it would be good to clarify that in Figure 3 you only have 1 or 2 because you have only two transformations (right?).

The reviewer is right. Done.

10. Figure 3: the numbers in the y-axis are too small.

We increased the size of the fonts.

11. Figure 4: Please increase the size of characters, especially for the transformations on the right.

We increased the size of the fonts.

12. L194-195: "To circumvent this issue…". This sentence should be in the methods section.

We respectfully disagree with this recommendation. The fact that we worked one a single catchment, and then on 325 catchments to generalize the results, is already explained at the end of the methods section. Here, we want to recall this fact, and to further justify it with the results obtained on a single catchment (which were by definition not available in the methods section). We slightly modified the sentence as follows: "To circumvent this issue, and to generalize the results, we perform a similar analysis over the 325-catchment set presented in section 2."

13. L197-198: do you mean the most number 1 ranks among the 325 catchments?

Yes.

14. L271: I do not see an arc shaped curve for QlogQ. Please revise and correct if needed.

The QlogQ transformation does not appear in this Figure (Figure 10), as the use of log transformations is not advised for KGE (see cited reference in the text).

15. L219-220: I do not see what the authors write for the boxcox transformation. The minimum is reached close to the high flow category.

The reviewer is right. We modified as follows: « Finally, transformation *boxcox* shows the best rank for medium to high flows, but not for the highest flows. »

16. L237: The authors state that "the purpose of using models is to apply them under conditions that are different from those they are calibrated on". I think that having this sentence here is misleading, since the hydroclimatic conditions defined in this study for the calibration and evaluation periods are very similar (L86-87). I suggest deleting or re-wording.

We partially agree, as we used an independent period that is by definition different from the calibration period. However, it is true that we did not seek a climatologically different period. We suggest the following reformulation: « However, the purpose of using models is to apply them on periods different from those used for calibration. »

17. L255-259: this text should be in the methods section.
18. L277-282: this text should be in the methods section.

The tests linked to this text were already mentioned in the methods section. However, recalling the additional tests in the results helps the reader to understand the rationale of performing them based on previous findings. We therefore prefer to keep them. We also added the experiment plan at the end of the methods section.

19. L283-284: Does this mean that catchments with low BFI and calibrated with QlogQ will have low ranks (i.e., poorer performance) compared to other transformations? I recommend the authors explaining here the practical implications of their results.

Thank you for this comment. We could indeed help the readers to understand the implications of this analysis. For this example, the reviewer is almost right: it means that catchments with low BFI and calibrated with QlogQ will have low rank values, and thus a higher performance (rank 1 being the lowest error). We added this sentence: "It means that the lower the BFI, the better the use of transformations giving intermediate weight between high and low flows. Conversely, the higher the BFI, the better the use of transformations that give a large weight to low flows."

20. Figure 10: I recommend to adjust the limits in the y-axis (e.g., by setting maximum values to 7) for comparison purposes.

We decided to keep for Figure 10 the same limits for the y-axis as for Figure 9 and previous figures, for the sake of homogeneity with previous figures.

21. L355-356: Please remove the last sentence since, in my opinion, speculations should be made in a Discussion section, and not in an Appendix.

Done

References

Hrachowitz, M., and M. P. Clark, 2017: HESS Opinions : The complementary merits of competing modelling philosophies in hydrology. Hydrol. Earth Syst. Sci., 21, 3953–3973, doi:10.5194/hess-21-3953-2017.