Thank you very much for your review. We provide below some answers to the reviewer's remarks.

[1] In lines 36-46, the authors cite a lot of literature where transformations have been used. I find this paragraph very difficult to read. Would it not be useful to place all these papers in a table and simply report percentages of time a particular transformation has been used? It is quite difficult to find the non-reference text in this paragraph.

A1: We thank the reviewer for this suggestion, which will indeed increase the readability of this paragraph. We will fill in and comment a Table such as the one below:

Article	$Q^{1/2}$	$Q^{-1/2}$	Box-Cox	Other power-law transformations	Inverse
Smith et al. (2023)	✓	×	×	✓	×
Doe et al. (2023)	✓	✓	✓	×	×
…					

[2] More explanation would be helpful in places to be clearer about what previous authors found and what the state of knowledge is. The authors cite studies, but it is not clear what relevance the conclusions of these papers have. A couple of examples:

"Peña-Arancibia et al. (2015) showed that a squared root transformation with the Nash–Sutcliffe efficiency leads to a better calibration and a reduced parameter uncertainty than no transformation or a logarithmic transformation." – In how far did it lead to better calibration? What does better calibration mean in this context? A better NSE value?

"Sadegh et al. (2018) investigated the role of several transformations in three catchments and two models and deduced that data transformations might be more helpful for evaluation and analysis of model behaviour than model inference." – Why did they conclude that? Why the difference in result for evaluation and inference? Is this conclusion not in conflict with the conclusion of *Peña-Arancibia et al.*? What does 'analysis' mean in this context.

A2: We agree, we will provide more details about these studies.

[3] Why do the authors select these objective functions shown in section 2.3. The authors state that they analyze the following: '*in order to estimate how transformations impact the simulated time series*' . But this is not really what the authors do. They assess performance difference with respect to a couple of popular metrics, they do not analyze how the actual time series changes beyond assessing model performance.

A3: These objective functions were selected for their popularity in hydrological model calibration. In order to assess some sensitivity to the objective function choice, we selected two of them. The piece of sentence '*in order to estimate how transformations impact the simulated time series*' relates to the following: '*the 1995–2005 independent evaluation period is also used*', not to the choice of the objective function. This means that we do not aim to assess only what happens on the calibration period, but also what happens over an independent period after such a calibration, because being applied on periods different from the calibration period is how hydrological models are the most useful.

We respectfully disagree regarding the second part of the reviewer's comment: we do not 'assess the performance difference with respect to a couple of popular metrics'. These metrics are only used as objective functions (combined with transformations). Then, as described in section 3, the performance assessment is purely based on closeness of simulations to observations, i.e. how simulated time series are impacted by the transformations used for calibration.

[4] I am a bit confused by the transformations introduced in section 2.4. Aren't some of the transformations included in others? E.g. the log transformation is a specific case of the Box-Cox transformation. Why not use the minimum number of transformations and then test the influence of the scaling parameter used in the transformation. Using just the Box-Cox transformation and a Q^x transformation with lambda and x varying would capture most and would allow for a more general analysis. You could use the two flexible transformations and plot the result against the lambda and x values used and against the streamflow percentiles to get a better fundamental overview about what is happening!?

A4: We agree on the fact that 'the log transformation is a specific case of the Box-Cox transformation'. However, we decided to stick on the denomination of the actual transformations the most found in the literature. This is a deliberate choice, as we wanted to provide some feedback on transformations that are commonly used. The suggestion of the reviewer would have rather answered to another question, i.e. a try to assess (any?) possible transformations in a more systematic way, in order to finally try to identify a (set of) best transformation(s). Consequently, we prefer to keep the transformations that we selected.

[5] What lambda value has been used for the Box-Cox transformation? The result should be dependent on that choice given that the transformation is flexible. Previous studies suggested a lambda value of 0.3 to suitable for streamflow data to gain a more balanced calibration results (e.g. Vrugt et al. (2006), Journal of Hydrology, doi: 10.1016/j.hydrol.2005.10.041). How much does the result depend on that choice?

A5: We chose a value of 0.25, as suggested by Vazquez et al. (2008) and further used in Santos et al. (2018). We will specify this in the manuscript. The results should indeed depend on that choice but the difference between a lambda value of 0.25 and 0.3 may remain small.

[6] In line 245 you state: *"In addition, the transformations that show the best average rank are not widely used in the literature (0.2, log and boxcox)."*– Are you sure about this? Log and BoxCox (lambda of 0.3) transformations are such a standard to reduce the focus on high flows. They might not have been a focus in very recent years, but certainly from the late 90s to some years ago, they were widely used.

Some (random) examples:

Lerat et al. (2020). Journal of Hydrology, doi.org/10.1016/j.jhydrol.2020.125129

van Werkhoven et al. (2008). Water Resour. Res., doi:10.1029/2007WR006271

Huang et al. (2023). Journal of Hydrology, doi.org/10.1016/j.jhydrol.2023.129347

A6: Thank you for this comment and for the additional references. We agree on the fact that these transformations are used in the literature, we just meant that they are not the most common transformations. We will rephrase as follows "*In addition, the transformations that show the best average rank are not* ***the most*** *widely used in the literature (0.2, log and boxcox).*"

[7] For section 4.3, could the authors not organize the catchments into those dominated by slow and fast behavior, e.g. using the (central) slope of the flow duration curve or some other signature metric? There might be different reasons why a catchment varies in this regard (snow, pervious geology, …), which might not be easily captured by the characteristics available.

All in all, an interesting study, though I think the authors could (should?) provide some more fundamental insight still. For example by varying the parameter of the Box-Cox or other flexible transformations.

A7: Thank you for these suggestions. We will add the calculation of the central slope of the flow duration curve, as defined in McMillan et al. (2017). We will also add other indicators such as the ones detailed in answer A25 to the reviewer 2.

Regarding the reviewer's last remark, we agree that assessing in a more systematic way many transformations (e.g. by varying the Box-Cox parameter or by testing many values of power transformations and also potentially by combining criteria) could be interesting. However, we believe that such a study is beyond the scope of the present study, as here we wish to assess transformations selected among those commonly used. We will however add this perspective in the conclusions.

References:

McMillan, H., Westerberg, I., Branger, F.: Five guidelines for selecting hydrological signatures. Hydrol. Process., 31, 4757– 4761, https://doi.org/10.1002/hyp.11300, 2017.

Santos, L., Thirel, G., and Perrin, C.: Technical note: Pitfalls in using log-transformed flows within the KGE criterion, Hydrol. Earth Syst. Sci., 22, 4583–4591, https://doi.org/10.5194/hess-22-4583-2018, 2018.

Vázquez, R. F., Willems, P., and Feyen, J.: Improving the predictions of a MIKE SHE catchment-scale application by using a multi-criteria approach, Hydrol. Process., 22, 2159–2179, https://doi.org/10.1002/hyp.6815, 2008.