Review of the manuscript „Calibration of hydrological models for ecologically-relevant streamflow predictions: a trade-off between performance and consistency?" by Hallouin et al.

Dear editor and authors,

I thank the authors for their well-structured answers and for rigorously answering and addressing many of the reviewers comments. I especially appreciate that some major concerns have been addressed, such as a more detailed analysis of the three KGE components and single SFCs, the introduction of a lower benchmark, restructuring the methods chapter, and the increase of study-specific information in the introduction. I very much enjoyed reading this version of the manuscript and I personally think this is a study of good quality with a sound and novel modelling approach and nice figures supporting the conclusion. I only have a few minor comments left that could be addressed in the review.

Comments

P1 L2: I am not sure if "In catchments with little human-induces hydro-morphological changes" is needed, because there are hydrological model that use e.g. land use as input and should be able to deal with changes.

P4 L15-23: To me this paragraph is like a short summary of the methods. I preferred the previous version with the hypotheses. Having hypotheses or questions helps to guide the readers focus on specific aspects in the results.

P5 L20: Could you add a sentence explaining the principle behind the quality check?

P5 L 30 (Rainfall runoff model): Is the model used in a lumped or semi-distributed way? I was wondering if you have any snowfall in the 33 catchments and if the model has the option to simulate snow accumulation and snowmelt?

P6 L5: I don't think the statement "…needed for an EPA funded project Pathways…" is relevant for the reader.

P6 L 24-27: I don't fully understand what the difference between the studies of de Lavenne et al. (2016) and Coron et al. (2012) is. Could you reformate the sentences that to make it more clear?

P11 L1: I would suggest to extend to section title to "Consistency in the selection of model parameter values" to make clear that consistency is about the parameters whereas stability and robustness are about performance.

P13 L29-30: The statement that tailored objective functions suffer more from overfitting than traditional objective function is based on Figure 6. However, to make a fair comparison of the robustness, the robustness of the traditional objective functions should be calculated by calculating the difference between a traditional objective function (and not the ESFCs) in calibration and validation when calibrated on the traditional objective function.

P15 L27: I couldn't find the evaluation of model simulations on 156 SFCs in the methods section. I recommend to add it somewhere to have a complete description of the evaluation in the methods section.