Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2019-161-RC1, 2019

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Calibration of hydrological models for ecologically-relevant streamflow predictions: a trade-off between performance and consistency" *by* Thibault Hallouin et al.

Anonymous Referee #1

Received and published: 3 June 2019

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

Hydrological models are an important tool for evaluating the effect of altered runoff regimes on the ecological integrity of a freshwater system. However, the accurate prediction of multiple specific hydrograph aspects (SFCs) from a single streamflow simulation is challenging. Hallouin et al. address this challenge by evaluating model simulations from calibrations with different objective functions. The selected objective

functions are the Kling-Gupta efficiency (and variants) and some objective functions consisting of ecologically relevant SFCs. Using 14 split-sample tests with a moving window, model performance, performance stability, and consistency in the selection of parameter sets are evaluated. Results from simulations in 33 Irish catchments indicate that model performance for SFCs tends to be higher when parameter sets are selected using SFC-based objective functions. However, SFCs based objective functions tend to have a lower consistency than traditional objective functions such as KGE.

This study addresses the current challenge of calibrating hydrological models to make accurate and robust predictions of multiple SFCs. The combination of performance, stability and consistency makes the results of this study especially valuable and highly relevant for the prediction of SFCs in places or times without data. I really like the concept of the three-fold model evaluation and I think it is presented in a clear well structured way. Generally, I think it would be important to do a more in-depth analysis of the results to improve the understanding of why there are differences in performance, stability and consistency. Also, I think that more references to related work would improve the scientific background of the research questions and could enhance the quality of the discussion.

I hope that the comments below will be helpful for the authors to improve their manuscript.

General comments

The differences in performance, stability, and consistency are evaluated in terms of KGEhi, KGEav, KGElo, Dinv, Dfsh, and Dall. These are all multi-objective criteria and their final value is an aggregated metric over multiple aspects of the hydrograph performance. Looking at the single components of KGE and also at single SFCs could potentially allow us to gain some insight into why differences in performance, stability,

and consistency can be observed (e.g. is the performance of a particular bespoke objective function lower than the one of KGEhi because of a few SFCs that are very poorly predicted?). I think it is especially the "why" that helps us to learn more about the effect of objective functions on simulations and that ultimately helps us to improve predictions of various hydrograph aspects.

Using the concept of consistency, this study looks at the ability of objective functions to consistently select the same set of behavioural parameters across multiple calibration time windows. Consistency is evaluated and compared across objective functions using a fraction-based score (i.e. fraction of parameter sets that is constantly behavioural). I think it could be interesting to also look at the parameter values themselves by e.g. plotting the values of the most sensitive parameters against model efficiency. Eventual patterns in parameter values might then be helpful to understand differences in model performance from different objective functions.

Sometimes differences in model performance between objective functions are rather small (as mentioned by the authors in section 4.2). Is it possible that these differences only seem to be small because model performance cannot get too bad in the study catchments? It might be worth to compare results against a benchmark, such as a random selection of 1

The study is based on three hypotheses addressing model performance, stability, and consistency. The introduction provides the reader mostly the background and motivation for hypothesis H1 (model performance). However, there is not much information about the current knowledge and experience on the stability of model performance and the consistency in parameter selection. I recommend to extend the introduction so that it covers the whole range of questions addressed in the study.

Specific comments

Abstract: More than half of the text in the abstract is introduction. I suggest to shorten

that part and provide more information on the methods and results of the three hypotheses of the study (performance, stability, and consistency).

Introduction: The studies of Kiesel et al. (2017) and Pool et al. (2017) are often cited as examples for using streamflow characteristics in model calibration. However, there are many more studies using hydrological signatures for model calibration. Although these studies don't aim at simulating ecological flow indices it might be worth to cite some of them.

P4 L 4-14: It makes clearly sense to me that you select the same SFCs as have been used in previous studies. However, I am not sure about the use of the term "ecologically relevant" in e.g. your title. To my knowledge, the relevance of SFCs depends on the species of interest. I wonder if the same species are important (or exist) in Ireland as in Germany and the Southeastern US? If not, can the selected SFCs still be considered as ecologically relevant?

P4 L20: Seven years are selected for calibration. Are seven years enough to get a robust estimate of the SFCs used in this study? I think it would be worth to discuss that and think about the consequences for the results on the stability of performance.

P4 L24-30: Can you say how many catchments are nested? How does that affect the generalization of the results? You mention that the catchments represent the diversity in soil types and geology - therefore, I would provide more explicit information about soil type and geology. Is there any snowfall in the study catchments?

P5 L26: Model set-up: How was potential evapotranspiration calculated? How was the warming-up period selected in case of the moving split-sample tests?

P6 L20: It is a common practice to not only transform flows to put more emphasize on low flows but also put more emphasize on mean flows (by e.g. calculating the sqrt of flows). What was the reason for calculating the mean of KGEhigh and KGElow instead of making a transformation?

P7 L12: Given that the focus of the paper is on the comparison of traditional and bespoke objective functions, I recommend to explicitly write down the equation for each of the three bespoke objective functions so that the reader doesn't have to check Table 2 to know which SFCs and how many are in each bespoke objective function.

P7 L12: The traditional objective functions (KGE variants) have their optimal value at 1 whereas the bespoke objective functions have an optimal value of 0. This can be very confusing when interpreting the results in Figs. 6, 7 and 8. I strongly recommend to define the bespoke objective functions in the same way as the KGE, i.e. 1 - Euclidean distance.

P7 L16: Model evaluation: A description of the concepts of stability and consistency is missing in the section on model evaluation. On the other hand, many sections of the results (P8 L 20-25; P8 L27-P9 L2; P9 L 25-L31; P10 L9-15) start with an extended paragraph on methods. I would move these paragraphs to the section on model evaluation to make a clear separation between methods and results.

P8 L 10-14. How were the 14 combinations of the 7-year periods chosen?

P11 L 4: There are more studies looking at the effect of objective functions on the prediction of streamflow characteristics than those of Vis et al. (2015), Kiesel et al. (2017), and Pool et al. (2017) (two examples are Hernandez-Suarez et al., 2018; Zhang et al., 2016)

Discussion: About half of the discussion is about the implication of the results for climate change studies and studies on the prediction in ungauged basins. However, climate change and prediction in ungauged basins are not directly addressed in this study. I agree that the results are interesting and relevant for these two topics and it is important that you do discuss the implication of your results for climate change and prediction in ungauged basins. But generally, I recommend to rearrange the discussion to focus much more on the findings and limitations of this study: which hydrological processes are represented by different objective functions? What causes

differences in performance and consistency? Why is there not much difference in the stability of model performance?

Detailed comments

Title: Given that the results don't show very strong differences between performance and consistency between traditional and bespoke objective functions I would think about using a question mark at the end of the title.

P3 L29-32: H1: It is hypothesized that bespoke objective functions lead to a better model performance than traditional objective functions. This is a rather vague formulation and I would explicitly state that it is about a better model performance. H2: What is a "small" number of SFCs? Again, I would be more explicit in the formulation of the hypothesis.

Fig. 2: I agree that the information in Fig. 2 is interesting for someone working with the same data set. But its information is maybe not so relevant for the general reader. You could think about placing the figure in the supplemental material.

Fig. 4: i) For me step f and g go hand in hand and I would merge them as you do with the calculation of the objective function in step c. Is there a reason why you use once the term "mean" and once the term "average"?

P6 L30: I would suggest to shortly explain how exactly flows were normalized.

Fig. 5: I am not sure if this figure is needed. The concept of the split sample test with a moving window is already well explained in the text.

Fig. 7: Are the axis labels in 7a and 7b switched, i.e. should 7a be Dinv and 7b be Dfsh?

Fig. 7: You often use the term "Euclidean distance" when talking about the SFCs based objective functions. These objective functions are defined in section 3.2 and I don't think it is necessary to repeat that they are based on the Euclidean distance.

P9 L29: "... more holistic definition that traditional objective functions represent...". I think this is a delicate statement and needs some explanation, especially since KGE consists of three components and the bespoke objective functions consist of many more components.

P10 L27: I think it is rather the model performance that is consistent than the catchments themselves.

References

Hernandez-Suarez, J. S., Nejadhashemi, A. P., Kropp, I. M., Abouali, M., Zhang, Z., Deb, K. (2018). Evaluation of the impacts of hydrologic model calibration methods on predictability of ecologically-relevant hydrologic indices. Journal of hydrology, 564, 758-772.

Zhang, Y., Shao, Q., Zhang, S., Zhai, X., She, D. (2016). Multi-metric calibration of hydrological model to capture overall flow regimes. Journal of Hydrology, 539, 525-538.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-161, 2019.