Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., author comment AC1
https://doi.org/10.5194/hess-2021-325-AC1, 2021

Reply on RC1

Silja Stefnisdóttir et al.

Author comment on "Improving the Pareto Frontier in multi-dataset calibration of hydrological models using metaheuristics" by Silja Stefnisdóttir et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-325-AC1, 2021

Dear editor, Prof. Harrie-Jan Hendricks-Franssen,
Dear reviewers,
Thank you for your critical review comments on our manuscript. Below we respond to your comments and outline how we will address your comments in a revised version of our manuscript (our responses are provided in italics):
We propose to change the manuscript title, a tentative title could be:
"The value of metaheuristic calibration techniques in muti dataset calibrations to improve hydrological modeling results and constrain parameter uncertainty",
Which better describes our study.

Anonymous Referee #1
This well-written manuscript compares three different non-likelihood based model calibration methods for a single case study with the aim of showing the relative value of each of the methods, given the same amount of model runs.

Authors' reply (AR): thank you for this positive comment.

The paper is based on a case study presented in the work of Finger et al. 2011. It takes from this earlier study the case study, the (multi-objective) data, the hydrological model, the metrics and part of the model calibration methods. What is added is the use of the simulated annealing method and the Genetic Algorithm.

While the idea of comparing the relative value of different model calibration methods given a fixed amount of simulations is somewhat interesting, I cannot recommend the publication of this paper because all results are conditional on the single case study and, more importantly, on the chosen algorithmic parameters of the compared search algorithms (simulated annealing, Genetic algorithm). These algorithmic parameters are not discussed, they are simply fixed. Furthermore, GA is treated as a single method whereas there is a multitude of implementations with different performances.

AR: Regarding the first comment, we agree with the reviewer that a single case study can only reveal case-specific results. However, the methodology applied and the framework developed are valid for any case study. Thus, the framework built for comparing different metaheuristic methods remains valid for other case studies and can be directly transferred to other catchments.
The choice of the case study for illustrating work is a general issue in any hydrological

study as every study has to be based on a selected catchment or a set of catchments. In our case, we chose the same case study as in Finger et al. 2011, i.e., the Rhoneglaciser catchment, which has been well investigated in numerous previous studies. Thus, our reasons for choosing this catchment are as follows: 1) the data availability is excellent, 2) we can refer current results to previous works done on this case study, and 3) its high glaciation fits the purpose of our study perfectly. We do envisage an expansion to other case studies in the future, but we believe this would be the subject of another study.

Regarding the algorithm parameters and settings, we decided to use the version of the GA algorithm that is implemented as a part of the HBV model and suggested by Seibert et al. (2000). There is a multitude of GA algorithms and parameters, but we feel that the selection process and the sensitivity analysis required to find the best version of GA and the parameter settings to optimize the performances are out of scope for this paper. We, therefore, selected three algorithms and their parameter settings from the literature, i.e. MC from Finger et al. 2011, SA from Stefnisdóttir et al. (2020), and GA from Seibert (2000) and the HBV model as representatives for the three different algorithm paradigms. We will clarify this issue in the revised manuscript.

Thus: The conclusion that GA outperforms the other two within a fixed amount of simulations is not really interesting: it certainly outperforms random search (MC); whether it outperforms or not simulated annealing depends on the problem at hand and on the algorithmic parameters (how the search algorithm is tuned).

AR: We agree that GA will certainly find faster an optimal solution than MC. However, the novelty of our approach relies on the complementary processing of the results from MC, SA, and GA to perform a non-weighted multiple dataset calibration, which allows us to assess the impact of the three algorithms on three independent evaluation criteria and on the parameter uncertainty. Furthermore, our results reveal that GA is able to find faster and better solutions than SA, based on the example study. As the number of interactions for the model rounds is one of the critical settings that have to be given for any optimization algorithm, in our opinion, this finding should be of interest to the hydrological community. We agree however that how much GA is able to outperform MC or SA depends on the problem at hand. We will emphasize this issue much stronger in our revised manuscript.

Moreover, we were able to demonstrate that the spread of the Pareto Frontier is much better in MC than in the other two methods as shown in Figure 9. One could argue that GA or SA will find a local optimum, without exploring the full spread of the Pareto Frontier. However, our results demonstrate the opposite of that, i.e., the Pareto Frontier of our multiple dataset calibration was improved compared to other methods, and parameter uncertainty could be significantly reduced. We do believe that this is a fundamental result that is of value for the entire hydrological modeling community.

All conclusions on the value of multi-data calibration for this case study re-iterate, reinforce earlier conclusions.

AR: We thank the reviewer for this supportive statement. We would like to point out that our results additionally provide proof that the Pareto Frontier can be identified with our method. We also highlight the importance of analyzing the Pareto Frontier in the post-processing analysis to support the evaluation of the methods. We will put more focus on this issue in our revised manuscript.

The paper does not present new methods on how to compare the algorithm outputs nor on how to analyze the optimization outputs (metrics taken from Finger et al., 2011). Accordingly, the paper does not present new methods nor new transferable insights into existing methods (except into the exact algorithms used in this paper) nor new insights

into hydrological processes.

AR: It is true that MC, SA and GA have been developed in previous works, which demonstrates that we incorporated previous works into our study. However, the comparison of non-weighted multiple dataset calibration and its effect on the Pareto Frontier has never been quantified in hydrology in the context provided in this manuscript. To the best of our knowledge, there is no other study that looked at multiple-data calibration using snow cover, discharge, and glacier mass balance data comparing these three optimization algorithms. In addition, we judge the three methods using strict computational requirements, represented here by the number of model runs, since the question is also how well the algorithms use the available computational resources.

As far as I see, this paper does thus not fit HESS.

AR: Given our argumentation above, we must disagree with the reviewer on this point. We believe that HESS is a great journal for our manuscript as it deals with the problem of multiple data calibration using different optimization algorithms. This should be of interest to HESS readers, as also noticed by the 2nd reviewer. Based on this reviewer's comments, we believe however that we need to better highlight the novelty of our work and clarify methodology.

Additional comment:
I do not understand how the paper can mix Pareto-optimality and multi-objective optimization via objective function weighing: if we optimize a weighted sum of objective functions, you cannot get the Pareto-frontier (or only if you explore different weighings). Either an algorithm looks specifically for solutions on the frontier or it does not. Judging a posteriori how many we found by chance seems a rather unfair criteria to compare different algorithms. But perhaps I misunderstood something her.

AR: We agree that the weighing function cannot simply define a Pareto frontier. However, we believe that there is a misunderstanding and we do have the firm intention to clarify this misunderstanding in a revised version of the manuscript. As stated in the reply to reviewer 2 (see below) we do not use function weighing. We complemented the standard calibration algorithms (MC, SA, and GA) with a ranking algorithm to quantify the trade-off between the three calibration criterion and identify the Pareto front (details are described in Finger et al. 2011). Each calibration function has the same weight. In short: we rank all runs according to each calibration criteria, average the ranks and select the 10 or the 100 best runs according to the average rank. This ensemble of "good runs" does not have "a best" run, it is simply the ensemble that describes the trade-off between the three criteria and accordingly defines the Pareto front. This method allows us to avoid function weighing and identifying a Pareto front (illustrated in Figure 9).
Thus, we believe that there is a misunderstanding and we do have the firm intention to clarify this misunderstanding in a revised version of the manuscript.