Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., author comment AC2
https://doi.org/10.5194/hess-2021-325-AC2, 2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.

Reply on RC2

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-AC2, 2021

Anonymous Referee #2
The study analyses the performance of three frequently used automatic optimization algorithms including Monte Carlo simulations, simulated annealing (SA) and a genetic algorithm (GA) for a multivariable calibration in a small glaciated catchment considering streamflow, snow cover area and glacier mass balance. The results are evaluated based on the objective function values achieved by the best 100 and best 10 parameter sets as well as by uncertainty widths regarding prediction and parameter uncertainties. The authors conclude that the genetic algorithm outperformed the two other methods as it achieved better solutions and narrower confidence intervals than the other two methods. The paper is generally well structured. The problem of model calibration is within the scope of HESS and the question of which search technique to select is a practical and relevant question that modelers have to decide upon.
AR: we thank the reviewer for this positive feedback.

However, I have several major concerns with this paper.
The authors see the novelty of their work in "confronting for the first time these three metaheuristics most frequently applied in hydrology within a multi-output calibration framework to derive practical recommendations for further applications". This seems overstated. What about other comparison of optimization techniques? (See e.g. studies cited in Efstratiadis and Koutsoyiannis (2010)). Are Monte Carlo, GA and SA really the three most frequently applied optimization methods in hydrologic model calibration? What has been found by other studies that compare different optimization techniques and what are the research gaps that are addressed in this study?

AR: We thank the reviewer for this meaningful comment. We agree with the reviewer that metaheuristics have been investigated numerous times in several studies. We also apologize for not having included all previous works. In a revised version we will include a thorough literature review and include all relevant works not cited so far, especially works cited in Efstratiadis and Koutsoyiannis. The novelty of our comparison study lies however in the fact that we use a non-weighted, multi-dataset calibration, where all datasets are equally weighted, and apply three algorithms of similar computational efforts, i.e., Monte Carlo (MC), Genetic Algorithm (GA) and Simulated Annealing (SA). We agree that these three metaheuristics may not be the most commonly applied but they are computationally equal and simple to implement, which is a great benefit of using them. We will highlight this point more in the revised manuscript. To our knowledge, there is no other study like that in hydrology that has used muti-dataset calibration with these three metaheuristics in

order to investigate the Pareto front. We will emphasize the novelty and the distinction of our study to previously published works in a revised version of this manuscript.

A fundamental problem of the current study is that the authors emphasize the multiobjective nature of the problem and aim at analyzing which of the three optimization methods provides the most balanced pareto front. However, as far as I understand, the authors did not apply optimization techniques that are designed for this task (e.g. multiobjective variants of GA). Instead it seems that the multi-objective problem was summarized to a single-objective problem (using a weighted sum approach with fixed weightings) and SA and GA were applied in their single-objective forms (which is fine in principle but not if the aim is to study the pareto front).

AR: We agree with the argumentation of the reviewer but believe that there is a misunderstanding. We complemented the standard multi-objective calibrations (GA, SA, and MC) with a ranking of all runs according to each criterion individually and subsequent averaging of the ranks to obtain the Pareto front (see below for more details). We do not use a single-objective form, but a fully independent multi-dataset calibration with three equally weighted objective criteria. These three criteria are used during the calibration but, as correctly noticed by the reviewer, we do not distinguish weights for them but apply the same weighting. The reason for not varying weights is that we simply want all of the three variables (i.e. discharge, snowmelt, and glacier mass balance) to be simulated similarly well. Introducing different weights would possibly have improved the simulation of one of the variables but at cost of lower performance for the other two variables. Further details of the way how best runs are chosen are described in Finger et al. (2011), but in principle, we rank all model runs according to the performance of each dataset, then average the ranks and finally select the 10 or 100 best runs that received the highest average rank. We use this ensemble of "good" runs to illustrate our Parento frontier. This allows us also to obtain a Pareto frontier for each method (MC, SA, and GA). We will make an extra effort to clarify the method in a revised version of the manuscript.

For analyzing which method performs best in representing the Pareto front, the study focuses on objective function values and the number of non-dominated solutions, concluding that GA performs best. In a multi-objective setting, one should additionally consider the diversity of the solutions, i.e. how well they are spread along the pareto front.

AR: The reviewer is correct! We will look at the Pareto frontier as a complementary tool to support the comparison of our three calibration algorithms. We will add additional text on the Pareto spread in the revised manuscript as criteria for method comparison.

Some of the conclusions cannot be drawn from the results of this study. The study concludes that the results demonstrated the value of multi-dataset calibration for realistically simulating different runoff components. However, while this might have been a finding from a previous study I cannot see how this can be concluded based on results from the current study. The study also states that "it appears to be essential to give equal weights to all modelled runoff components". However, only one weight configuration has been tested so that this statement cannot be derived from the presented results.

AR: Thank you for this comment which showed us that we must clarify our statements and conclusions. We believe that this comment is based on a misunderstanding and would like to clarify our point. It is true that similar results regarding the value of multi-objective calibration were found in the study of Finger (2011), however, that study was using only MC calibration. Our results indeed reconfirm previous findings from Finger et al (2011, 2012, 2015, 2018), Etter et al. (2018), J de Niet et al (2020) and complement them by demonstrating the value of SA and GC and their impacts on the Paret front. This is the major novelty of our study, which we will better describe in the revised manuscript.

It is also true that we use the same weights for all three calibration criteria, i.e., the Nash-Sutcliffe coefficient for Discharge (Q), RMSE for Glacier mass balances (MB) between measured and simulate MB and the ratio of correctly predicted snow cover area for snow cover (SC). We did not test different set-ups of the weights as we aim at having all variables simulated equally well. We will modify these conclusions in our revised manuscript.