Review of Manuscript

'Simulation-Based Inference for Parameter Estimation of Complex Watershed Simulators'

By R. Hull et al.

Dear Editor,

I have reviewed the manuscript. My conclusions and comments are as follows:

1. Scope

The article is within the scope of HESS.

2. Summary

In their manuscript, the authors address the question of efficient parameter estimation for distributed process-based hydrological models. They suggest simulation-based inference using a surrogate model (LSTM) for the original model (Parflow) for rapid generation of parameter – simulation output data sets to support training of a second neural network to learn the joint distribution of parameters and simulation output in an nonparametric way. With their approach, they address both the intractability problem of parameter estimation (distribution cannot be properly estimated due to theoretical or computational reasons) and the epistemic uncertainty problem, here more specifically the problem of uncertainty about the correct model structure. They explore the effects of the various parts of their workflow by various virtual reality studies with different levels of simplification (experiments 1-4). They conclude that i) SBI works well if the surrogate LSTM is accurate (experiment 1), that surrogate misspecification leads to errors in parameter estimation (experiment 2), that the problem of overconfident parameter inference of experiment 2 can be partially solved by ensemble (boosted) approaches (experiment 3), and by an ensemble approach with informal weighting of the members (experiment 4).

3. Evaluation

This is a thoroughly conducted study on a relevant topic, reported in a complete, concise and balanced manner. In short, it was a pleasure to read. So I have only very minor specific points:

We thank the reviewer for their thorough review of our work and for their positive evaluation. We have address all of the specific comments below.

Line 63-64: I do not agree with the general claim that "DL methods are not widely used in watershed prediction due to the inadequacy of available data in representing the complex spaces of hypotheses". There are in fact many examples of DL-only or DL-conecptual hydrological modeling applications in the literature. I'd agree if the authors meant that DL methods are not widely used for distributed prediction of a large number of hard-to-observe hydrological variables. Please explain.

We agree with the reviewer's comment that DL applications have become more common in recent years. Our intent was to emphasize they are not yet widely used to predict distributed variables in hydrology. We will revise the paragraph to emphasize that the novelty is in distributed prediction, in this case using an emulator of ParFlow, a distributed physically-based model.

Line 357: For demonstration purposes, only two parameters, Manning's roughness and hydraulic conductivity are investigated. Can you say a word about how you expect the method to scale to larger number of parameters?

We briefly address this question in discussion [lines 847-865]. We will lengthen this discussion in the revision to better cover the following points:

Expanding the model, both in terms of the number and distribution of parameters, is essential to finding adequate representations of real hydrological processes. SBI is well-suited for inference in high-dimensional space relative to some approaches, and has had many adaptations (Cranmer, 2021). As with any approach to inference, scaling to a greater number of parameters will bump into computational constraints. Those constraints come from the cost of simulation (i.e. ParFlow), and the cost of inference (i.e. neural density estimation). In our study, the cost of simulation from ParFlow is high, and this has a compounding effect on the cost of inference. Utilizing a surrogate can in some ways reduce the cost of inference, by reducing the need to resort to ParFlow, but as we show this comes at a tradeoff of accurate parameter estimates if the surrogate is not adequate. Focusing inference on the most informative parts of higher-dimensional parameter space is important if SBI is conducted directly with a costly simulator. Papamarkarios' early work with SBI developed sequential neural sampling techniques, which might be less wasteful than other approaches to sampling parameter space (i.e. Papamakarios et al., 2018; Lueckmann et al., 2017; Greenberg et al., 2019). Lastly is the option of compressing or reducing the dimensionality, which could be important for the case of estimating distributed parameters. The topic of compression and SBI is explored by Asling, 2019.

Line 359: 183 is not a very large ensemble. I assume this is due to the high computational effort of ParFlow? Also, can you say a word about the computational effort of the PB Model (ParFlow) vs. the NN model (LSTM)?

The reviewer is correct that this is primarily due to the cost of the process based model. As noted in the previous response, this is the largest part of our computational demand. The computational cost of training of the LSTM is many orders of magnitude smaller compared to the cost of ParFlow simulations. Once trained, simulation from the LSTM is essentially free. In the revised manuscript we will include specific details related to the cost of ParFlow vs. LSTM.

Line 374: "They [LSTMs] have had some use for predictions in hydrology" really is an understatement. They are in very widespread use these days. Please change.

At the time of drafting (2021), they were still somewhat nascent. We will update in the revised manuscript.

Eq [11]: Just a comment: This could also be done by Kullback-Leibler divergence without introducing a threshold chosen by trial and error.

Thanks for the comment. We could see how this approach will be advantageous in future evaluations of SBI.

Eq [12]: Why is here RMSE used, instead of KGE as in Sect. 3.8?

As referenced in the text from [490], we choose to use the Kling Gupta Efficiency (KGE; Gupta et al., 2009) as the likelihood metric for its utility and history rainfall-runoff model assessment. We appreciate the reviewer for noting alternatively that the RMSE was selected to evaluate the posterior prediction. However, using KGE instead for the posterior predictive check would not change the results and conclusions in the manuscript.

Yours sincerely,

Uwe Ehret