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:

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.

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?

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)?

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.

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

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

Yours sincerely,

Uwe Ehret