REVIEW of the paper

Coupled machine learning and the limits of acceptability approach applied in parameter identification for a distributed hydrological model

Authors: Aynom T. Teweldebrhan, John F. Burkhart, Thomas V. Schuler, Morten Hjorth-Jensen
Manuscript Number: hess-2019-464

Submitted: **HESSD**

This paper presents machine learning methods (MLMs) to emulate MC simulations to identifying behaviour parameter sets of hydrological model. Three MLMs were trained on limited number of MC samples to predict some sort of error or loss function of the MC simulations. Trained models were then used to predict loss function for a large number of samples from which the behavioural parameter sets were identified. While the results look reasonable, there are two main fundamental issues in this manuscript. Authors claimed that the proposed method overcomes computational burden of MC simulations and subjectivity in choosing the likelihood and the threshold value in GLUE. Manuscript fails to provide sufficient evidence to support both claims (see comments below).

I am struggling to find main motivation of this work. It is mentioned that emulators are used to minimize the computational burden of the MC simulation. But this is not completely true. Emulators are used only to predict some sort of likelihood values of the simulation to know whether it should be rejected or not in GLUE framework. Then hydrological models are run with behavioural parameter sets to quantify predictive uncertainty. In other words, MC simulations are still used. Indeed, the proposed method does not save computational time when it is required e.g., in real time forecast. For example flood emergency managers want to know the probability of exceeding major flood level at tomorrow noon. There are other ways to emulate MC simulations which are saving computational time in real time application (e.g., Shrestha et al., 2009; Shrestha et al., 2014).

Another issue in this manuscript is that proposed GLUE pLoA is not convincing. Authors mentioned that the original GLUE has issue in subjectively choosing a likelihood and threshold value for identification of behavioural and non-behavioural parameter sets. They proposed GLUE pLoA to overcome these limitations, however it introduces two additional settings to choose: error bounds and percentage of the model predictions that fall within the error bounds to identify whether given simulation is behavioural and non-behavioural. So proposed method is also subjective, indeed more complex than the original GLUE and requires iterations to choose percentage of the model predictions that fall within the error bounds that satisfy the acceptable CR value.

Verification scores used in this manuscript do not directly test accuracy of emulators to identify behavioural or non-behavioural parameters sets. In this manuscript, RMSE and related measures were used as performance measures of the emulators. However, the problem should be formulated as classification rather than regression if the objective of emulators is to identify whether given simulation is behavioural or non-behavioural. Classification problem is very straightforward:

- Classify each MC simulation to behavioral or non-behavioral model using GLUE pLoA
- Train and test emulators to classify whether given MC simulation is behavioral or non-behavioral model
- Verify the emulators to test accuracy of the classification using a 2 by 2 contingency table similar to used in weather forecast. In this table “hit” represents number of the cases when the MLM correctly identifies or classifies the behavioral parameter sets (i.e. classification from MLM is behavioral for behavioral parameter sets). From this table it is possible to compute various scores including hit rates (Hits/(Hits+Misses) etc.

MLM Emulators	Parameter Sets	
	Behavioural	Non-Behavioural
Behavioural	Hits	False alarms
Non-Behavioural	Missed	Correct negatives

Minor comments

P3, L32: define Score.

P4, L14: What is the basis for 25% as mean observational uncertainty? It is not clear how streamflow limits are computed using this observation uncertainty. Since hydrological model errors are heteroscedastic, applying same value of 25% of the mean observation as error bounds for all time steps would be problematic.

P4, L27: Define acceptable pLoA. Is it CR from the original GLUE? I wonder what GLUE CR value is. I think this is another subjectivity in this method. Importantly the proposed method relies on original GLUE method to identify acceptable CR. In other words, the proposed GLUE pLoA is not completely independent method, it relies on residual GLUE method to compute its hyper parameters such as acceptable CR.

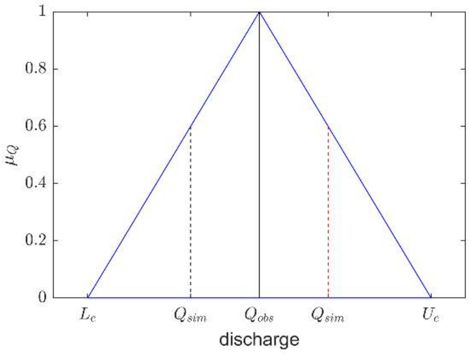
P4, Step 3: “… specified percentage of the total observations.” Here is one of subjectivity to identify whether the model simulation is behavioral or non-behavioral. What value is used?

P5, L1: Equation 2 should be defined before steps.

P5, L9: Since all terms of Equation 3 are not defined (e.g. l, u) and assuming L_e in this equation is same as L_e defined in equation 2, I am not sure if the equation is correct. It is not clear whether e is absolute. In either case, for example first expression $\mu_Q = 0, e \le L_e$

might not be correct. It is better to illustrate Equation (3) with a figure similar to the following

$$\mu_Q = \begin{cases} 0, Q_{sim} \le L_e \\ \dfrac{Q_{sim} - L_e}{Q_{obs} - L_e}, L_e < Q_{sim} \le Q_{obs} \\ \dfrac{U_e - Q_{sim}}{U_e - Q_{obs}}, Q_{obs} < Q_{sim} \le U_e \\ 0, Q_{sim} \ge U_e \end{cases}$$



P6, Line 31: 5000 samples may not truly represent the parameter uncertainty. I suggest to use convergence analysis to know the number of samples.

P11, L5, what is the validation data set? Is it S3?

P13, Table 4: Another widely used cross-validation method is leave out cross-validation. For example, for leave-one-year-out cross-validation, generate simulations in 2011 using model calibrated (e.g., behavioral parameter sets identified) in all data except year 2011, generate simulations in 2012 using model calibrated in all data except year 2012 and so on. Then all simulation data from year 2011, 2012, 2013, and 2014 can be collated to verify the results. This cross validation procedure is expected to produce results that are comparable to those obtainable under operational conditions as the number of data used to fit the model will be similar to that available for operational applications.

P15, Table 5: I strongly suggest replacing Table 5 with distribution plots which is more readable.

P15,L3: Section 4.3 is not relevant to this study, so can be deleted.

P18, l17, row?

References

Shrestha, D.L., Kayastha, N., Solomatine, D., 2009. A novel approach to parameter uncertainty analysis of hydrological models using neural networks. Hydrol. Earth Syst. Sci., 13: 1235-1248.

Shrestha, D.L., Kayastha, N., Solomatine, D., Price, R., 2014. Encapsulation of parametric uncertainty statistics by various predictive machine learning models: MLUE method. Journal of Hydroinformatics, 16(1): 95-113.