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

Aynom T. Teweldebrhan, John F. Burkhart, Thomas V. Schuler, Morten Hjorth-Jensen

Response to Reviewer #1

Dear Reviewer, we are grateful for your thoughtful comments and suggestions. Following is our reply to the points raised in your feedback; and it is structured as comment from reviewer (light blue text) followed by our response to the comment.

Reply to the general impression of the reviewer

Dear reviewer, as you have pointed out under the specific comments (1), the identification of behavioural models through coupling of emulators is affected by multiple factors. It depends on nature of the likelihood measure and its predictability as independent variable (for example in this study, between pLoA and Score). It also depends on the type of fitting model (emulator) used to estimate value of the likelihood measure (in this case the machine learning models).

Although residual-based likelihood measures were used in previous similar studies, as of our best knowledge none of **the emulator based studies** have used pLoA or Score as a response surface, and the limits of acceptability approach in general. And it is for this reason that the first objective of this study was focused on assessing the possibility of using pLoA for the identification of behavioural models using the **coupled** MLMs and the limits of acceptability approach. Further, since the three machine learning models are applied to predict the same response variables followed by the identification of behavioural models using the limits of acceptability approach, the relative performance of RF and KNN (that were not applied in previous studies) can be easily evaluated against the standard ML model, i.e. NNET. And this forms the basis for the second objective of this study, for which the authors believe gives a new insight into the possibility of using RF and KNN as emulators of the MC simulation for application in parameter identification.

To what does one ascribe this conclusion - ploA or emulation?: "ML emulators and the limits of acceptability approach have performed very well in reproducing the median streamflow prediction both during the calibration and validation periods."

The median streamflow prediction is the result from the **coupled** effect of both the likelihood measure (pLoA) and the specific emulator used to predict the likelihood values.

1. A good emulator (in this case a mapping between $\mathbb{R}^n \rightarrow \mathbb{R}$?) may not help to improve the streamflow predictions if the identification metric or the hydrologic models are bad. So the performance of emulation is a somewhat independent question from that of the performance of an identification metric.

This comment is consistent with the response provided above for "the general impression of the reviewer".

From the manuscript, the conclusions suggest that both emulation and pLoA together happen to work well. But even that is doubtful as the paper does not comment on many aspects of emulation.

(a) How do these techniques perform when the models are run fewer number of times, say only 400 times instead of 4000?

Thank you, we will accommodate this comment in the revised version. A preliminary analysis using 400 samples shows that some of the coupled emulators fail to produce any behavioural model in certain years.

(b) How do these techniques perform with a parameter space of higher dimensionality (n) such that $\mathbb{R}^n \rightarrow \mathbb{R}$?)?

Sensitivity of the emulation-based parameter identification to parameter space dimension was not conducted since running the hydrological model used in this study under a distributed setting requires a long time. The model is structured in such a way that, at each time step, the main processes of the model run on each of the grid-cells. This challenge becomes more pronounced when we consider the need for high number of model runs in order to overcome the non-identifiability problem for high parameter space dimensions. Thus, the assessment for effect of parameter space on emulation-based parameter identification might be the subject of our future studies.

(c) Also, what is the added utility of the 95000 simulations in comparison to the already 4000 runs? Any recommendations/comments on the number of samples required for convergence?

Like most studies based on the GLUE methodology, the main focus of this study was also to get as much behavioural models as possible so as to encapsulate future uncertain conditions. However, only little to no improvement was obtained in most cases when assessed using the available evaluation dataset and the streamflow evaluation metrics used in this study.

(d) How does the emulator perform in extrapolation phase (the 80% calibration, 20% validation separation will not be adequate to show how the emulator may diverge when one uses parameter values away from the training data set. This implication will be more severe when the emulators are used in Bayesian inference and the prior distribution of parameters is not hard-bounded).

As presented in the manuscript (Validation columns in Table 3), capability of the emulators to reproduce the response surface generated directly from the Monte Carlo simulations was further assessed using the 95, 000 samples (S3) in addition to the 20% (test) samples.

(e) And perhaps analysing or commenting on the time efficiency of emulators.

We will accommodate this comment in the revised version of the manuscript. The emulators normally take few seconds to generate the response surfaces for the 95000 samples. And when it comes to the Monte Carlo simulation, it was assumed that each of the iterations requires same amount of time. Accordingly, the amount of time required would be proportional to the number of iterations.

2. What new insights do we get from the application of emulation tools to this pLoA metric, apart from the fact that it is a possibility to emulate?

Since the predictability of independent variables varies from one to another, application of emulation methods to predict pLoA gives us a further insight on the potential and scope of the emulators to predict different response surfaces in addition to the residual-based likelihood measures that were applied in previously studies.

“the three MLMs were able to adequately mimic the response surfaces directly estimated from MC simulations”. This needs to be made clear (preferably using numbers) in the abstract, discussion and conclusions.

The detailed result supporting this conclusion is presented in Table 3 and explained in section 4.1 of the manuscript. As suggested, we will also provide some metric values in the abstract, discussion, and conclusions in the revised version of the manuscript.

3. What is the interpretation of the output generated from behavioral parameters? Do we expect the observations to lie within these bands with a certain frequency? (please refer to Stedinger et al. 2008, for more insights on this debate) If yes, then the reader would like to see reliability (q-q) plots to gauge the performance.

Thank you for your suggestion to the reading material. It provides further insight on uncertainty analysis in hydrological modelling. This theme has been the subject of debate in many hydrology literatures. In order to avoid any confusion with the confidence level expected from the formal Bayesian approach, we will include the following text in the revised version:

When using the GLUE methodology, the observations are not expected to lie within the prediction bands at a percentage that equals the given certainty level. However, the modeller can adopt the certainty level specified for producing the prediction limits as a kind of standard for assessing the efficiency of the prediction limits in enveloping the observations (Beven, 2006).

4. How much of the statements made about the efficiency of the emulator are dependent on the choice of the specifications of those machine learning techniques? A paragraph on the meta parameters of this study will be appreciated.

As suggested, we will include a paragraph on hyper-parameters of the machine learning models in the revised version of the manuscript.

5. Some hydrographs will be a useful addition to the existing plots.

As suggested, hydrograph plots will be included in the revised version of the manuscript.

6. Please explain why an assumption of 25% for observation error and what will be the effect of choosing a different value on the performance of either GLUE pLoA and the emulation.

In the GLUE LoA methodology, the limits are set with due consideration to the observation and input errors. Since observational error values were not available for the study area, this value was set based on literature value and observations from a neighbouring catchment plus assumed allowance for input errors. In our previous study, a preliminary assessment on effect of relaxing the limits further, i.e. over 25% while keeping the threshold pLoA at 100% have yielded to the inclusion of non-behavioural models, leading to very low performance during the validation period.

References:

Beven, K.: A manifesto for the equifinality thesis. Journal of Hydrology, 320, 2006.

Choi, H. T. and Beven, K.: Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework, Journal of Hydrology, 332, 2007.

Kirchner, J. W.: Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward, Water resources research, 45, 2009.

Ratto, M., Castelletti, A., and Pagano, A.: Emulation techniques for the reduction and sensitivity analysis of complex environmental models, Environmental Modelling & Software, 34, 2012.

Stedinger, J. R., Vogel, R. M., Lee, S. U., and Batchelder, R.: Appraisal of the generalized likelihood uncertainty estimation (GLUE) method, Water resources research, 44, 2008.