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 #2

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.

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.

Dear reviewer, as mentioned in the manuscript, only 5000 MC simulations are run instead of the 95000 from which the behavioural models are identified. The emulators normally take few seconds to predict the response surfaces for the 95000 samples. And this justifies how much the computational cost has reduced as a result of using the MLMs to predict the response surface for the 95000 samples instead of using MC simulations.

As mentioned in the manuscript (Page 2, line 13; Page 4, line 22), GLUE pLoA is a time-relaxed variant of GLUE LoA which was introduced in our previous study (Teweldebrhan et al., 2018). Thus, the main goal of this study is to minimize the computational cost when using GLUE pLoA rather than proposing the methodology or comparing against other variants of the GLUE methodology. But we would like to reiterate that it was proposed as part of the endeavour to minimize the rejection of useful models when using the original GLUE LoA formulation rather than to dealing with the subjectivity. Useful models were effectively identified using GLUE pLoA, while all of the 100000 simulations were rejected as non-behavioural models when using the original GLUE LoA formulation (Teweldebrhan et al., 2018).

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.

As indicated in the manuscript (e.g. page 3, lines 33-35), the emulators were used to predict the response surfaces for new parameter sets. The identification of behavioural models is, however, a result from the **coupled effect** of the emulators in reproducing the response surfaces and the GLUE pLoA in identifying the behavioural parameter sets. Thus, first capability of the emulators to reproduce the response surface was evaluated through comparison of the predicted against MC simulation based values. Then, performance of the behavioural models was evaluated through comparison of their streamflow simulation result against observed values.

We appreciate for the alternative insight you provided us to dealing with the problem. However, in GLUE pLoA, the models are evaluated as ensemble, based on their capability to produce a CR value close to the predefined value, rather than as individual models. For this reason estimating the response surface using a regression method was found to be more relevant than generating binary values (behavioural/non-behavioural) using classification algorithms.

P3, L32: define Score.

This term was defined earlier in Page 3, line 20

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.

Since no stage-discharge relationship exists for estimating the streamflow uncertainty using the usual practice, i.e. by fitting different rating curves, an assumed value of 25% was adopted based on certain literature values and observational errors analysed for a neighbouring catchment. This value also takes into account incommensurability and uncertainty in the input dataset. The streamflow observational error bounds (limits) of each observation are estimated as ±25% of the corresponding observation, instead of the mean observation. Yet, as the reviewer mentioned since model errors are heteroscedastic mainly in response to the variability in input dataset errors, it would be too strict to expect a given model to satisfy the limits of acceptability criteria in 100% of the observations. And it is this phenomenon that has called the need to introduce the time relaxed variant of the original GLUE LoA formulation.

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.

As indicated in P4, L32, the acceptable pLoA is the one that yields a calculated CR value close to the predefined acceptable CR value.

As mentioned in line P4, L17, The CR value is expressed as the number of observations falling within their respective prediction bounds to the total number of observations (Eq. 1). In this study, the CR value obtained using the residual based GLUE methodology was used for the ease of comparing the result obtained from both methodologies. However, the modeller may also set the acceptable CR value based on previous experience, although this involves some degree of subjectivity.

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?

The iteration to get an acceptable pLoA value starts from 100% and decreases further, i.e. relaxed until the desired level of CR is achieved. The reason for relaxing this criterion is provided under the response to the P4, L14 comment. We would, however, like to iterate that relaxation in the GLUE LoA approach in order to overcome the rejection of useful models is not a new phenomenon. The difference with the previous approaches lies on use of the time relaxed approach than, for example, extending the limits (e.g. Choi and Beven, 2007).

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

As suggested, we will do that in the revised version of the manuscript.

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 $\mu_Q = 0, e \leq L_e$ might not be correct. It is better to illustrate Equation (3) with a figure.

Thank you, the notations l and u respectively correspond to $L_e$and L_u. Thus, we will replace them with the latter notations in order to be consistent with the notations in Equation 2. We will also provide an illustrative figure accompanying Equation 3 similar to the suggested one.

The notation e is not absolute and thus the expression $\mu_Q = 0, e \leq L_e$ is correct, since a model producing a negative error value of less than the lower observational error bound (which is also a negative value) has 0 degree of membership.

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

The 5000 samples were used for training and testing of the machine learning emulators. While the behavioural parameter sets that actually are less than 5000 were identified from the 95000 samples (Section 2.3). The reason for the low number of behavioural samples is partly attributed to the use of uniform parameter distribution and the simple Monte Carlo method for parameter sampling. However, analyses conducted using 50000 and 100000 samples in our previous study have yielded similar parameter and streamflow uncertainty results. We will include a text about this in the discussion session of the revised manuscript.

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

Here the validation dataset refers to S3 and the corresponding response surface values estimated using the MC simulations. We will clarify this in the revised manuscript.

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.

Thank you for the suggestion to the alternative cross-validation method. In this study we have preferred to test the model using the worst case scenario, i.e. if we have only one hydrological year for model calibration. Further, as presented in the discussion section, this method allows us to examine the performance of models identified in a given hydrological year when applied under a highly different hydrological condition. A similar approach was used in previous hydrological studies; and it was considered as a more rigorous validation method than the commonly used split-sample methods (e.g. Kirchner, 2009).

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

Thank you for the suggestion. We will replace this table with distribution plots displaying the distribution of each parameter under the different emulators.

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

As discussed in previous studies (e.g. Ratto et al., 2012), sensitivity analysis is often performed in tandem with uncertainty analysis in order to determine which of the input parameters are more important in influencing the uncertainty in the model output. Conducting sensitivity analysis using the inbuilt algorithms of the ML models also helps us to further evaluate their capability through comparison against the result obtained from other well established techniques.

P18, l17, row?
Thank you, we will correct this to “raw” in the revised version of the manuscript

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.