Summary

In this manuscript, the authors employ a specific set of machine-learning based emulators (random forests, kNN and ANN) for parameter and uncertainty estimation while using two different parameter identification metrics: (a) the absolute bias score of model predictions and (b) GLUE with time relaxed limits of acceptability (pLoA). The results from the work are mentioned in the abstract: "The three MLMs (machine-learning models) were able to adequately mimic the response surfaces directly estimated from MC simulations; and the models identified using the coupled 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."

General impression of the reviewer

While both (a) unbiased/improved parameter identification and (b) fast emulation of expensive hydrologic simulators are important research topics, I have major reservations with regard to the analysis and conclusions of this manuscript. The authors have already published a paper on the performance of ploA (Teweldebrhan et al., 2018) and there are several detailed papers/studies out there on the performance of various emulation techniques as cited by the authors (page 2, last paragraph, page 3 first paragraph). Combining these two separate research questions makes it difficult to judge whether the presented analysis is targeting the performance of ploA as a parameter/uncertainty estimation metric or of these three machine-learning techniques as efficient emulators. Such a joint treatment of these two separate questions doesn't necessitate any adequately new insight into either the emulator efficiency or parameter estimation. 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."

Additionally, while this manuscript can be a useful addendum of information for those interested in parameter estimation using the GLUE pLoA approach, it doesn't seem to qualify as a stand-alone novel application. The principal criteria for the acceptance of a paper in HESS are: **Scientific significance, Scientific quality, and Presentation quality**: I am afraid that as far as the first two criteria are concerned, the manuscript does not do better than a "fair" classification.

Therefore, I am inclined to reject this paper, and suggest a resubmission in a more suitable journal. However, in case other reviewers differ in their judgement, I request the authors to make these substantial revisions, in both the analysis and the writing structure, based on the following comments before it can be considered for publication in HESS.

Specific comments

1. A good emulator (in this case a mapping between $\mathbb{R}^8 \rightarrow \mathbb{R}$) 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. 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? (b) How do these techniques perform with a parameter space of higher dimensionality (n) such that $\mathbb{R}^n \rightarrow \mathbb{R}$? (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? (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 implications will be more severe when the emulators are used in Bayesian inference and the prior distribution of parameters is not hard-bounded). (e) And perhaps analysing or commenting on the time efficiency of emulators.

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? "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.

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.

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.

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

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.