Interactive comment on "Hydrologically Informed Machine Learning for Rainfall-Runoff Modelling: Towards Distributed Modelling"

By Herath Mudiyanselage Viraj Vidura Herath, Jayashree Chadalawada, Vladan Babovic

In this manuscript the authors use a machine learning method (Generic Programming) to identify optimal model structures from building blocks of two flexible rainfall-runoff modelling frameworks, i.e. FUSE and SUPERFLEX under semi-distributed catchment setting. This way the authors aim to eliminate the subjectivity in model structure selection. Further, the authors apply GLUE methodology to conduct parameter uncertainty analysis for the selected optimal model structures. The proposed approach was evaluated using data from the Red Creek Catchment, United States. This is very interesting research worthy to be encouraged as it lies within an area where extensive research is needed, i.e. in the area of machine learning applications in hydrology. However, I have some comments both on the content and structure of the manuscript that need to be addressed before the manuscript gets accepted for final publication.

General comment:

The authors have provided an extensive literature review on the subject matter. However, some of the topics are less relevant and may lead astray for the reader from the main subject matter. For example, it could suffice to present the literature on machine learning applications in water resources simply in one paragraph as part of the introduction section than providing own literature review section (section 3). Similarly, the sub-section focused on lumped and distributed models can be removed from the manuscript since this is too common topic in hydrology. On the other hand, less coverage was given to details of certain methodologies followed in this research. For example, it would have been more helpful to provide the reader with further details on set up and components of the Genetic Programming (GP), FUSE and SUPERFLEX by removing the literature review on less relevant topics including the sub-section focused on Artificial Neural Networks (ANN) (since ANN was not used in this research). Generally, with the exception of the sub-topics focused on GP, FUSE, and SUPERFLEX the remaining contents of Section 2 (Fundamental approaches in Hydrological modelling) and Section 3 (Machine Learning in Water Resources) can be either omitted, or placed under the introduction or discussions sections in a concise and relevant form.

The methodology and scientific background contents appear blended in many sections of this manuscript. Thus, I would recommend having a separate Methodology section with only those methodologies followed in your study placed under this section. Similarly, the 'Discussions' section is missing and some of the paragraphs in sections 2 to 6 appear more suited to the discussions section. Under this section, you may compare and contrast these previous research works in relation to yours with regards to the methodologies followed and results obtained in your research.

Although the authors have effectively applied machine learning methods for model structure identification under distributed setting, I am a beat skeptical on some of the conclusions arrived in relation to the methodologies followed. For example, the available dataset was divided into four categories, i.e. spin-up, calibration, validation, and test. From the manuscript it can be noticed that both model calibration and validation datasets were used in training the hydrological and machine

learning models (e.g. L448 and L464 in section 5.3). Thus, out of the total length of the dataset (i.e. 11 years), only one year was allocated for model testing (2013-2014) or for actual validation of the hydrological model. The question would then be if we can conclude that the proposed methodology achieved the intended goal. Application of the hydrological model for a single hydrologic year may provide the possibility to assess the dominant hydrologic processes in relation to the prevalent climatic and physiographic conditions in that particular year. But it would have been more helpful to use multiple single testing years, for example, using a cross-validation technique. This way the reader may get a better insight into the resulting model structures and the model test results under the different conditions and there by a more robust model evaluation. Similarly, the uncertainty analysis procedure lacks information on how the parameter bounds and threshold for behavioral models are set. A threshold NSE value of 0.6 was used in this study, which I think is very low for many practical applications of a hydrological model. Capability of the prediction bounds in bracketing the observed values is inversely related with the threshold NSE value. This may have yielded to the low modelling uncertainty (high percentage of bracketed observations) of the selected model structures reported in this manuscript.

It would also make easier for readers who are less familiar with GP to follow the manuscript if the GP terminologies can be re-written in hydrological context. For example, what do we mean by initial population here? is it a particular hydrological model component from FUSE/SUPERFLEX ? or is it a set of hydrological model parameters ?

Individual comments:

L6- ‘limited use in scientific fields’ seems too broad area to comment on.
Consider rephrasing it, e.g. ' in rainfall runoff modelling' (accompanied by a relevant reference)

L15- rephrase ‘decreasing meaningfulness of lumped models’.
Lumped models might be preferable under certain conditions, e.g. for very small catchments in data scarce areas where distributed or semi-distributed model settings might be less practical.

L20- ‘without any subjectivity in model selection’ seems less realistic since all model selection algorithms involve certain level of subjectivity, albeit at varying degrees. In your case, for example, setting the model parameter bounds (for the hydrological model) and many of the constants and assumptions related to set up of the machine learning model shown in Table 3 involve certain level of subjectivity.

L35- ‘Therefore, the final goal of any successful hydrological model must be based on a physically meaningful model architecture along with a good predictive performance’
But the measure of success for a hydrological model may vary from one model to another depending on the specific purpose for which they are developed. For example, physically based models might be tailored to enhance our understanding of the underlying physical system. While conceptual models might be expected to have only a partial understanding of the processes with the main purpose being to yield predictions within the required acceptable accuracy for the intended purpose. Further, black-box models, though with little or no understanding of the underlying physical system, still have their own merits when the main goal of the modeler is just to get acceptable outputs from the set of inputs as you’ve mentioned in L212.

L36- ‘Data science models’. Do you mean data-driven-models? Provide reference for this sentence.

Section 2.3 – remove this section or take selected points from this section and concisely discuss in relation to your methodology, results or conclusions (under the Discussions section).

L167-174 – provide references

L212- 'Certainly, if we are only interested in better forecasting results then, the machine learning models might be the preferred choice over the conceptual or process-based models due to their better predictive capability'.
But can we give this generalization in light of the multiple factors affecting the relative performance of machine learning models, including length of the training dataset and nature of the training algorithm? Provide reference.

L215- 'actionable models'. Rephrase in hydrological context

L222- 'Further, data science models…' Do you mean: **However**, data…
This seems to contradict with your previous statements in L218: '… offer two reasons for the limited success of data driven models'

L286-288- sentence not clear. Re-write, for example, as: Individuals with better performance (based on the objective function values) are assigned higher probability of selection and thereby given the chance to create offspring through genetic operators (crossover, mutation, and elitism).

L306 – mentioned?

L306-310 – long sentence, re-write with shorter sentences

L329- regularized? regular?

L355- 'GP has been selected as the machine learning technique here due to its ability to optimize both model configuration and model parameters together'.
Was GP used to simultaneously optimize both the hydrological model structure and parameters in this study? If so, how was GP used for parameter optimization of the hydrological model? (The illustrated procedure was focused on model structure). If not, how were the hydrological model parameters optimized before conducting the uncertainty analysis (UA)?

L377- how was the shape parameter value (2.5) of the Gamma-distribution based routing function determined?

L425- Elaborate on how the number of independent runs and other algorithm settings of your framework (Table 3) were determined. For example, why not 10 or 30 independent number of runs instead of 20? Similarly, why a generation number of 50 or population size of 2000 etc.

L471 - The selection of Cross-sample entropy parameters (e.g. r) are quite critical for the evaluation result. How were these values determined in your study?

L487- what are these model parameters? elaborate and provide reference.

L498- '…is changed uniformly…'. Do you mean: …are generated … ?

L500- how many and which model parameters were allowed to vary and how were these parameters selected out of the total number of model parameters?

L500-‘ …while keeping the remaining model parameters at their calibrated values.’.
How were these model parameters calibrated before conducting the UA. Was GP applied for that purpose? It would be helpful for the reader if you can clarify this in relation to the comment mentioned under L355.

L501- why was this threshold NSE value of 0.6 chosen?

L502- The term ‘behavioral models’ in this case refers to the parameter sets rather than to their NSE or discharge values.

L512- ‘…to measure the uncertainty estimation capability of the selected optimal model’.
Do you mean: … to measure the level of modelling uncertainty of the selected optimal model structure? In your study, the GLUE methodology was used to estimate the level of uncertainty, while the selected optimal model structure was itself subjected to uncertainty analysis rather than being used as an uncertainty estimation tool.

L513- ‘If the uncertainty estimation capabilities are satisfactory, the model performance of the optimal model is tested for an independent time frame (2013/01/01 to 2014/12/31) which is not used in model selection or identification stages’.
Re-write this sentence as well in accordance to the previous comment. What was your criteria for a satisfactory level of modelling uncertainty (or as in your text- a satisfactory uncertain estimation capability)? It seems that both the model selection and validation periods were actually used for model identification (selection) and not for a hydrological model validation or testing. And a single year of model testing looks quite short period to arrive at a conclusion. Thus, if you have data limitation from additional periods or if some of the hydrological model identification (selection) years cannot be moved to the hydrological model testing (validation), you may consider using alternative model evaluation techniques such as the leave-one-out or other cross-validation techniques. This way you may get more validation (test) results that can help you arrive at a relatively robust conclusion.

L626-629 – ‘Out of the 33 model parameters only 5 parameters can be identified as sensitive parameters. … This demonstrates a lesser dependency on model parameters compared to the total model performance in semi-distributed modelling owing to the large number of model parameters.’.
Among other factors, sensitivity analysis results depend on the minimum and maximum values of a parameter dimension. How, were these values fixed in your study?

L629-‘FUSE_TOPO_M1 results in high value (94%) for the percentage of measured streamflow data within the confidence interval bands and hence shows a significant capability of estimating associated uncertainty.’
Re-write this sentence in accordance to the comment provided in L512.
The percentage of observation bracketed by the uncertainty bounds is highly dependent on the threshold value used during behavioral model identification. The threshold NSE used in this study (0.6) seems very low as compared to the reported calibration and validation results of the optimal model. Given this low threshold NSE, it is expected to get a high percentage of the observations falling within the uncertainty bounds. Thus, try to justify why you adopted this threshold NSE value (under the methodology section).