Review of

Continuous streamflow prediction in ungauged basins: Long Short-Term Memory Neural Networks clearly outperform hydrological models

by Richard Arsenault, Jean-Luc Martel, Frédéric Brunet,
François Brissette, and Juliane Mai
[doi.org/10.5194/hess-2022-295]

General evaluation

This paper deals with a highly current subject matter. It presents a methodological framework for streamflow prediction in ungauged basins using a leave one out cross-validation approach (LOOCV) for three hydrological models and an LSTM network. The authors compare the performance of these four models for 148 basins in Northeast North America. The evaluation of the models at such a scale and working with these number of watersheds is quite impressive. The paper is well written, illustrated and organized. However, some points need to be taken into consideration to improve the manuscript, namely:

(i) A more comprehensive literature review should be provided; that is, the authors should acknowledge the works of several other authors who have dealt with streamflow prediction using LSTMs, not just focusing on the works of a specific research group. Please homogenize the diversity of the literature review and cover the works of others that have provided significant achievements in the field of hydrological modelling using Deep Learning (DL) models.

(ii) In the proposed LOOCV for LSTM modeling, the LSTM model was trained using a large dataset (N-1 basins), while keeping one basin as a pseudo-ungauged basin for validation. This approach departs from the basic philosophy of training DL models. Indeed, to avoid introducing a bias during training of DL models such as LSTM, overfitting should be avoided by considering a considerable proportion of the whole dataset as a testing dataset. What was the rationale behind this methodological approach?

(iii) The authors propose an LSTM modeling approach for ungauged basins that will, without a doubt, spur the interest of the readers. However, the literature has provided several good performances of LSTM models for similar regions in Northeastern North America. Perhaps the authors could provide some insights for future work in dry regions where the presence of extreme flows may not be as prevalent and whether they expect that there approach would need to be modified or not accordingly.

At this point, I am looking forward to reading the authors' point of view as I believe they have earned an opportunity to provide sound rebuttal comments as I feel the paper has the potential to be a valuable contribution to *Hydrology and Earth System Sciences*. Thus, for the time being, I would say that major revisions are necessary and required.

Please find additional suggestions/recommendations and editorial comments below that will need to be addressed thoroughly before the paper can be recommended for publication.

Comments/suggestions/recommendations

P4 The following sentence, « In the Kratzert *et al*. (2018) study, the regional LSTM models performed on average just as well as the local LSTM with the median NSE difference of 0. » Local LSTM should be clarified compared to regional LSTM.

P5 As illustrated in Figure 1 and Table 1, very large basins are included in the dataset, while including these basins during LSTM modeling has been quite a challenge since the input data are at the basin scale. How do the authors evaluate their results by assigning just one point to a basin with an average area of almost 31,900 km^2?

P10 Why did the authors choose the leaky ReLU activation function? The authors should provide a table presenting the tested functions and values of the specificities of the LSTM model and the optimal ones; that would provide more insights to the readers.

P11 Correct me if I am wrong, but according to the following sentence: « The twelve static descriptors presented in Table 1 allow the model to distinguish between each catchment ». Which one of them did the authors exactly use? Please provide another table introducing the list of twelve basin descriptors used for LSTM modeling.

P11 According to the following sentence: « Static descriptors were normalized between 0 and 1 using a min-max scaler, while the dynamic variables were standardized by the mean and the standard deviation, which is a standard practice ». Did the authors include streamflow (target) during this normalization process? If not, how do they analyze their results after denormalization? Later, on the same page, it is mentioned, « The specific streamflow was used as the target variable by diving streamflow records by the drainage area, then converted from m^3s^{-1} to $mm.d^{-1}$.». Please further clarify.

P14 According to the following sentence: « This is important, considering that a strong performing hydrological model with the

best regionalization method is still outperformed on average by a relatively simple LSTM model. », the authors claim to use a simple LSTM model while using 2 LSTM layers each with 512 units, based on my experience, this is not considered a simple LSTM model. Please modify the text accordingly.

P14 Please be specific. According to the following sentence: « It is also important to note that the training (80%) and validation (20%) basins are categorized as such randomly, so the training step is performed on different catchments for each of the 5 runs #4a-#4e. », the authors should provide more details on how they couple this splitting approach with LOOCV, this needs to be clarified.

P14 Figure 8 shows the sensitivity of the hyperparameter selection and the assessment of the LSTM model structure. The authors claimed that the performance generally increases with a more complex model structure, meanwhile Figure 8 shows that increments are very minor between the simple structure models and the complicated models. In real-world practices, training and calibration of complex models face major challenges, how do the authors explain the choice of the selected complex model?

P15 According to the following sentence: « First, the nature of the LSTM model makes it extremely difficult or practically impossible to determine the logical flow of data between the observations and the predicted streamflow », readers may find it misleading since understanding the relationships between inputs and output of data-driven models can be achieved using sensitivity analysis. It is the authors' responsibility to provide such analysis as it would provide a way of following the logical flow of data. Thus, this sentence should be clarified accordingly.

P16 Based on the following sentence: « However, in this study, regularization failed to improve results ». Did the authors test all the possible values of dropout rates to reach such a conclusion? For instance, the value of 0.5 for the dropout rate has shown to be promising in improving the accuracy of streamflow modeling in other studies. Did the authors test this value?

Figures and Tables

None, all the tables and figures are well organized.

Editorial comments

None, this is a well-written paper.