Sivarajah Mylevaganam
Alumnus, Spatial Sciences Laboratory, Texas A&M University, College Station, USA.

The history of hydrological models goes to many decades. The progress that has been made to improve hydrological models through scientific findings has been showing the endless road to guide the next generation of hydrologist and associated specialists to quest for the betterment and mark the dead end. In this manuscript, the authors, who are not from the same hydrological basin, have employed an ML model to predict streamflow by training the model with hydrologically diverse data that is spatially and temporally large in extent. Based on the research, the authors draw a concrete conclusion that the previous modeling works (that is found based on ad-sense free search) using ML models have failed to understand the underlying philosophy in training and testing ML models.

In my opinion, the current version of the manuscript has many flaws. Moreover, the way the manuscript has been presented gives an impression that the authors are far off from the field of hydrology. Therefore, the current version of the manuscript needs an expert in hydrology to go through in detail in an unbiased manner. Furthermore, the language that has been used in the manuscript needs to be edited by a language specialist as the way the manuscript has been presented and the words that have been coined throughout the manuscripts are beyond from what is expected in a scientific manuscript that is submitted to a journal office that advertises to have achieved a high impact factor based on rigorous reviews by a panel of experts in the fields of specializations listed in the scope of the journal.

The following comments are posted.

1) In my opinion, the abstract of the manuscript needs to be re-written. I would say that an ML model with limited and poor data has been employed in writing the abstract. Moreover, what has been highlighted ("there is one MISTAKE in particular that is common") is not well documented.

2) Line 10-Line 14:
Hydrology models based on machine learning (ML) are different – ML models work best when trained on data from **many watersheds** (Nearing et al., **2021**). This citation needs to be evaluated against the conclusion (i.e., ML models are best when trained with a large amount of hydrologically diverse data) that the authors draw from the manuscript that is submitted to this journal office.

3) Line 10-Line 14:
Because ML models are trained with data from **multiple watersheds**, they are able to learn hydrologically diverse rainfall–runoff responses (Kratzert et al., **2019b**). This citation goes to 2019. The comment 2 goes to **2021**. Are the cited manuscripts giving the same thoughts? If so, the rationales in citing these manuscripts are not well understood.

4) Line 17-Line 23:

The paragraph needs to be critically reviewed by a specialist. The sentences need to be evaluated. The paragraph gives an impression that the authors lack fundamental knowledge in the subject. The terminologies and words from an English dictionary are thrown without understanding the exact meaning.

5) The crux of the manuscript that is highlighted by the authors (i.e., LSTM stream models are best when trained with a large amount of hydrologically diverse data) is a well established fact in the scientific field. Basically, the authors are hitting the concept of SAMPLING SIZE of an experiment. Therefore, instead of going through a painful path of running models to determine the number of basins, it would be a wise man thought to go through some statistical methods to answer this question (i.e., sampling size). In fact, even the current version of this manuscript does not give the exact number that would be required. It is a random number (531) that the authors have ended up with based on what has been analyzed (see Fig.5).

6) Line 30-Line 36:
Does the order of the KEYWORDS have an influence in your search? Is the search from the engine not prioritized by the engine provider based on the business model employed? What was the reason to limit the search to 2021. Based on Fig.1, it is understood that there are more than 3500 publications. Even if we consider the authors' statement that review was initiated in September 2022, an iota of incompleteness surface.

7) Line 25-Line 29:
The authors claim that the use of LSTMs for rainfall–runoff modeling has increased exponentially in the last several years. A figure (Fig.1) to support the claim is found in the manuscript. As per the figure, a rough estimation considering the heights of the bars gives an indication that around 8500(=3500+2500+1500+500+400+100) manuscripts have been found in the topic that the authors have invested. Referring to the previous comment, the authors have considered 100 manuscripts based on the search from a search engine of their interest. In other words, this manuscript is based on 100/8500*100%=1% of the manuscript found in the literature. What can be inferred from the training dataset that is employed in reviewing the literature? Will it lead to conclude that similar to an ML model the limited publications reviewed lead to draw wrong conclusions?

8) The title of the manuscript needs to be assessed by a specialist. What is a basin? What is a watershed? What is a catchment? What is a region? What is the amount of data that a single basin possesses? What is the spatial and temporal extent of the basin that the authors are defining in the title of the manuscript? What is the heterogeneity level of the basin that the authors are defining in the title of the manuscript?

9) Refer to Part III

Acknowledgement and Disclaimer

The author is an alumnus of Texas A&M University, Texas, USA. The views expressed here are solely those of the author in his private capacity and do not in any way represent the views of Texas A&M University, Texas, USA.