Response to Reviewers for Manuscript: Benchmarking Data-Driven Rainfall-Runoff Models in Great Britain

Anonymous Reviewer #1:

Comments/Text of Anonymous Referee posted in **black**, our text in **blue**.

The manuscript outlines an application of LSTM-based runoff models, which were introduced in previous studies (Kratzert et al, 2018; 2019). In the present contribution, the focus of analysis is catchments in Great Britain. Similar to previous studies, the objective is to demonstrate the competitive ability of LSTM in rainfall-runoff simulations over traditional process-based models. The authors made considerable efforts to set up experiments and perform relevant analyses. Results are compared with four lumped conceptual models and show that the LSTM models outperform the traditional models as well when applied in Great Britain.

The manuscript is generally well written and organized, figures and tables support the results.

My main concern is the degree of innovation and scientific significance of this work compared to already published works. This is a critical aspect of the manuscript that should be improved.

A large section of the manuscript is dedicated to a discussion of the advantages reported in the previously developed LSTM model. This discussion focuses on predictive ability, without much methodological improvement and innovations in ideas, that in turn may impair the scientific importance of the research.

In recent years, LSTM models have been broadly assessed. Most of these studies indicate the generally better performance of LSTM models over lumped models. The results reported in this manuscript seem to confirm the previously reported conclusions. By comparison, the analysis in Sections 4.2 and 4.3 is limited, whereas IMHO this is the most insightful section of the paper which deserves additional in-depth discussion. I think the authors should dedicate more space to discuss the implication of their findings.

Below are more detailed comments, questions, and suggestions that hopefully initiate a fruitful discussion and help improve the paper.

We thank the reviewer for their sincere comments and suggestions. We intend to make two major revisions to the paper based on these comments.

1) **Shorten certain sections**: Make the description of the LSTM/EA-LSTM experiment and the presentation of the predictive ability results more concise, moving some material and plots into supplementary information.
2) **More explicit description and discussion of novelty**: We propose to expand the sections that discuss in what conditions the LSTM produces different simulations to the lumped conceptual models (Sect 4.2) and the different performance in catchments where water is balanced vs. imbalanced (Sect 4.3).

We recognise the need to discuss the implications of these findings more fully. We propose three ways to do this:

1) More critically engaging with the experimental structure and outlining what is and is not possible to conclude with the given experiment.
2) Focus attention on the interpretation of the performance differences and away from the performance improvement of the LSTMs.
3) Expand our analysis of specific events and catchments in order to provide case studies of where there are large discrepancies in model performance.

ABSTRACT:

I would suggest mentioning the challenges in present LSTM applications for hydrological modeling and what is to be addressed, otherwise, it is difficult to tell the significance and necessity of the work.

Some major challenges of LSTM applications include:

1) How to incorporate uncertainty in inputs and outputs? See Klotz et al 2020.
2) How well do LSTM based models perform under changing climate conditions? See Sungmin, Dutra and Orth 2019.
3) How we learn from the performance improvement of LSTM based models, e.g. to diagnose missing process representations by comparison with the benchmark models.
4) How does the number of parameters in the LSTM model interact with issues of equifinality;

We are addressing the third challenge in this paper. One of the promises of deep learning (DL) is that DL models can help identify any limitations of hydrological data and process representations, conditional on there being a pattern to these limitations. The LSTM is a demonstrably effective architecture for modelling systems with short (long-term) and fast (short-term) signals, such as hydrological systems. The LSTM can efficiently extract information from large sample datasets to link meteorological inputs to discharge, simulating the catchment system. The next step is then learning how to extract this information from the LSTM. We have proposed and explored a number of ways of doing this, and outline one of our methods in this paper, intercomparing diagnostic measures (the different goodness-of-fit metrics for different parts of the hydrograph).

We will make these challenges and the key contribution and novelty of the study clearer in the abstract.

INTRODUCTION:

I do not think research gaps are well defined in the introduction. The research objectives should be motivated by the research gaps. The latter two of the three questions raised in the manuscript are related to overcoming limitations and model diagnosis, without indication of the explicit research gaps to be addressed. Are there some additional studies that investigate the correlation between LSTM model performance and catchment attributes?

The background should be more concise and emphasizes more about what is still to be investigated regarding the usage of LSTM models.

We agree with the need to make the introduction more concise and focused on the underlying goals. The introduction and objectives will be rewritten to more clearly reflect the research gaps remaining for LSTM based rainfall-runoff models.

The existing literature on LSTMs has focussed primarily on the skill of these models, as applied to CONUS, but has not:

a) Performed a comparison with conceptual models in GB
b) Used LSTM performance improvements to provide insights into the catchment characteristics and time-periods where LSTM models provide better simulations, and using this to diagnose the conditions in which the LSTM has a significant advantage.

To address these gaps the paper asks the following questions:

1) How does the LSTM **perform in GB**, and how do the results compare against commonly used conceptual models? We use benchmark model performances to give context to the LSTM model performances (i.e. to demonstrate if the LSTM is competitive against other models). We use results from a previously published model benchmarking paper.
2) How do we **extract information** from the spatial and temporal patterns in diagnostic measures? The aim is to relate these patterns back to the hydrological conditions in those catchments/time periods with the largest differences in model performance.

Furthermore, LSTM is but one of several machine learning frameworks used in rainfall-runoff modelling. Recent advances in evolutionary computation report theory guided and "hydrological informed" approaches that result in not only highly accurate but also readily interpretable models. See for example:

J Chadalawada, et al, 2020, Hydrologically Informed Machine Learning for Rainfall‐Runoff Modeling: A Genetic Programming‐Based Toolkit for Automatic Model Induction, Water Resources Research 56 (4), e2019WR026933

HMVV Herath, 2020, Hydrologically Informed Machine Learning for Rainfall-Runoff Modelling: Towards Distributed Modelling, Hydrology and Earth System Sciences Discussions, 1-42

I have read these papers and will include a more complete set of references with regards to machine learning based approaches to hydrological modelling. Although the major focus for our review here is on neural network / deep learning methods.

Line 77: It seems only THREE research questions are being proposed.

We will update this to reflect the three research questions outlined.

METHODS:

Section 2.3: It is more suitable to use the term “layer” (e.g., LSTM layer and EA LSTM layer) when describing the specific layer structure.

Yes that makes sense, and this allows us to incorporate the comment below about the “final layer”, the fully connected layer used to map the hidden state vector to a single discharge prediction.

Line 158: Please keep consistent notation using curly quotes or straight quotes throughout the manuscript.

We will check and update notation for consistency.

Figure 2: In EA LSTM cell, is the input gate “i_t” or “i”? (see Equation 8)

This should read “**i**”, since the input gate does not receive time varying inputs but only the catchment attributes (**A**). Hence the output of the input gate is a unique vector for each catchment (but *static over time*).

Lines 203-206: The fully connected layer should be a part of the model architecture. It seems strange to introduce them in this subsection (model training).

We agree, and will rewrite this.

Section 2.5.1: A brief description of the process-based models is required, especially what hydrological processes are included in the respective models because the discussion section involves the consideration of processes.

We completely agree that this is a necessary part of the methods section. We will include a more detailed description of the benchmark models, but aim not to repeat what has already been written by the original benchmarking paper (Lane et al 2019).

Furthermore, looking at all reviews, there is a consistent call for the paper to be made more concise. We will **rewrite the methods and experimental design** to be a shorter summary of the main components of LSTM based models that make them suitable for rainfall-runoff modelling. We will aim to draw parallels with the traditional hydrological models, outlining the key similarities and differences between these models. We agree with all reviewers that there is no need to repeat the equations for the LSTM based models, in line with similar papers for other hydrological models.

RESULTS AND DISCUSSION

Table 3, Figure 3, Figure 4, Figure 6, and Figure 7: All the results seem to merely be used to show the outperformance of LSTM models than other models in various cases. I think this part should be more concise if the result is not out of expectations, and more other implications should be discussed from the results.

We agree and will move most of these plots into the supplementary information. We believe that keeping table 3 (the overall median goodness-of-fit metrics) Figure 3 (CDFs) and perhaps Figure 6 (seasonal NSE spatially). This should give an overview of the overall pattern, the spatial pattern and the seasonal pattern which form the three key goodness-of-fit intercomparisons.

Lines 496-503: The speculation of "connectivity" is interesting, while how the connectivity can be "learned" by LSTM models should be clarified, say whether the connectivity can be represented by hidden information within data or the model architecture (such as the memory of LSTM).

Great point. We believe that the information captured within the model architecture may be used to tell us something interesting about connectivity. The idea is that the vectors that represent the fast and short information processed by the LSTM (h_t and C_t) have learned something useful about summer (semi-arid) hydrology that we can extract and interpret.

Lines 505-507: A simple strategy to examine the speculation is to train an LSTM model with/without crop_perc included for checking its role in improving the representation of hydrology in those catchments with a strong agricultural signal.

This is a very useful suggestion. Ultimately, in order to properly account for the contribution of many different factors we would require a more comprehensive analysis, rather than performing a somewhat ad-hoc analysis on a single variable. All reviewers agree that we need to make the paper more concise, however, we intend on expanding the discussion of the hydrological conditions in which the LSTM outperforms the benchmarking models. Therefore, we intend to flag this topic as warranting further discussion in our upcoming paper on LSTM interpretability and propose to add a sentence to explain our intentions of pursuing this in future work.