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

Anonymous Reviewer #3:

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

This paper describes two versions of a national scale deep learning hydrological model for GB and compares them to 4 conceptual hydrological models from the FUSE framework. The effectiveness of LSTM has been well established in previous studies, and so the novelty of this paper lies in its application to GB catchments. As the code, data and outputs are all freely available, I consider this to be a useful study to hydrologists concerned with modelling GB catchments. I wonder if given the limited scientific insights of this paper may be better placed in the Journal of Hydrology: Regional studies, or Environmental Modelling and Software rather than HESS.

I would like to commend the authors on a very clearly written paper- it was very easy to follow and understand.

We thank Reviewer #3 for their comments and effective summary of the paper. We take on board the claims about scientific novelty and propose to update the paper to reduce the emphasis on outlining the performance improvement of the LSTM compared with the conceptual models. We will do this in three ways:

1) Reduce the length of our (re)introduction of LSTM/EA LSTM methods, pointing readers towards the original papers that introduced these methods to hydrology.
2) Reduce the number of figures that demonstrate performance improvement of the LSTM in comparison with the conceptual models.
3) Increase the emphasis on exploring what we can learn from the differences in performances. We propose to rename and restructure section 4.2. to better reflect the focus on intercomparison of simulations in different catchment attribute conditions.

My major criticism of the paper is that the authors never demonstrate the **model's applicability to a changing climate**. Even if the application of LSTM (and all models that rely entirely on calibration) is only for near term flood forecasting, it is likely that we will be modelling events outside of the training data of the model with increasing frequency. I think that an alternative calibration/validation strategy should be examined where extreme events are left out of the calibration of the model, to provide some confidence in its ability to model beyond its training dataset.

We agree that understanding model performances on out-of-sample events is an exciting area of study. However, we believe the calls from all reviewers for a more concise paper mean that a complete exploration of this question is beyond the scope of this study.

My other major criticism is that the authors never **discuss the insights gained from the LSTM model**. There is no discussion of the sensitivity of the model to the different inputs and how the model ends up being structured. **They never provide any evidence to answer their third research question**. I think this would add a lot more value to the paper and make it worthy of publication in HESS. In the conclusion the authors state that this will come in a subsequent paper, but I think it would be more valuable here (and some of the detail of the calibration/validation could be moved to the supplementary information).

We thank Reviewer #3 for their identification of research question 3 as the most scientifically valuable contribution of the paper. Given that we are shortening Section 2.3, removing a number of plots and reducing the emphasis on the performance comparison, we propose to expand the discussion of how we can learn from the performance differences. We propose to expand the discussion and results to address the question: 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.

Some more specific comments follow:

line 19: There are more modern PBSD models than SHE. Reference Parflow, SUMA, SHETRAN, Hydrogeosphere etc.
We will update the references to include links to more modern PBSD models. Thank you!

line 77: there are only 3 research questions
We will update this to read “three research questions”.

Figure 1: You can format text in python to include superscripts "$mm day^-1$". Reduce point size- they are overlapping and obscuring each other.
We will update this: thank you!

Table 1: Nice! Very useful table. Temperature should be referred to with a capital T. Should Xt actually be Xn if it is representing the concatenation of dynamic and static input data for a single catchment?
We will update this as proposed. X_t should probably be X_{t,n} to reflect that it contains information for the target time period and the target catchment. Great spot!

line 176: Include the link to the prediction and error metrics at the end of the article too.
We will update this link to include the metrics. Thank you!

Table 2: Why these attributes? Was LSTM sensitive to all of these?

We incorporate these attributes since they cover three families of catchment characteristics that are important for hydrological modelling: soil structure, landcover types and climatic conditions. We used the same attributes as the previous LSTM/EA LSTM paper which was completed on data from CAMELS-US (Kratzert et al 2019).

line 220: What is an epoch? how does this relate to number of catchments/years of data?
An epoch is a single pass through all of the data. We will clarify this point in the text. So the LSTM is trained using an iterative gradient based optimisation method, stochastic gradient descent. An epoch means that every single sample (catchment-time) in the training period is used to update the weights. It does not affect the number of catchments or years of data, but it reflects the iterative nature of the training process, i.e. that at each iteration the model will see all of the data and make steps in multi-dimensional parameter space towards a more effective representation of the hydrological system (through better predictions, defined by our loss function, NSE).

Table 3: How is statistical significance calculated here? Double check that it is the appropriate method.
We used the “Paired Wilcoxon Test” (Scipy Function with the “alternative” parameter set to “two-sided”). We did not use the absolute values and will change that as proposed. We propose to clarify this in the text.

Figure 3: Nice figure
Thankyou! We also liked this composite figure.

line 366: I don't think that the **catchments with significant snowfall** should be included in the comparison if the snow modules of the conceptual models have not been turned on- this does not seem like a fair comparison. Recalculate the statistics leaving these catchments out.
This is a very interesting point and something that we propose to expand our discussion about. One of the key benefits of using the LSTMs, and data-driven approaches, is that we do not need to pre-specify the modules/structures that need to be included. Instead we can learn this from the data. We propose to recalculate the statistics for the intercomparison as suggested and to include this information in the supplementary information, however, we believe that by providing a GB-wide benchmark it is important to show the performance across all of the catchments that have been modelled, especially since these results are being published as a comparison for future work. We propose to expand our discussion and critical evaluation of the experimental setup.

line 367-371: this is a repetition of the previous paragraph.
Thank you for drawing our attention to this. We will remove the repeated sentence.

Figure 5: cut. This is a long paper with a lot of figures. I don't think this figure adds much to the maps.
We agree and propose to keep table 3 (the overall median goodness-of-fit metrics) Figure 3 (CDFs) and perhaps Figure 6 (seasonal NSE spatially) and remove the other figures or move

into supplementary information. 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.

Figure 6. Label missing on the colorbar
We will update the colorbar as proposed.

Discussion: Cut all references to the physically based models. The comparisons are not rigorous and so should not be presented.
We agree and will remove this section from the manuscript.

figure 9: significant correlations are not clear. consider showing this in an alternative way.
We propose using larger font for the “*” signifying significant correlations. We agree they are currently too small to be useful.

line 537: I think this is the most interesting point in the whole paper- I would love to read a lot more about this in the discussion.
We agree that this warrants further discussion and propose to expand our discussion of this point, drawing on references from outside hydrology, and particularly from atmospheric science.

Uncertainty: I would like to see some discussion of training models to uncertain flows and uncertain inputs.

We agree that addressing uncertainty in inputs and outputs is of vital importance for hydrological modelling. While we feel that full treatment of uncertainty in inputs and outputs is beyond the scope of this manuscript, we want to propose two methods for addressing this important point.

1) Expand discussion of recent advances in LSTM based modelling with uncertainty inputs. In particular, see Kratzert et al (2021) on the performance boost of using multiple rainfall datasets, highlighting that the LSTM can flexibly incorporate new information from highly co-linear input datasets. Secondly, work by Klotz et al (2021) demonstrating three different methods for uncertainty quantification.
2) We have trained an ensemble of 8 models. We propose to make better use of this ensemble of models to represent uncertainty. Firstly, we propose to demonstrate the hydrological conditions in which discharge estimates are most uncertain. We propose to include a variance-based metric to reflect the uncertainty of the underlying ensemble, and to present the hydrographs in Appendix 2 with uncertainty bands reflecting the interquartile range of predictions.

References:
Klotz, Daniel, et al. "Uncertainty Estimation with Deep Learning for Rainfall–Runoff Modelling." Hydrology and Earth System Sciences Discussions (2021): 1-32.

Kratzert, Frederik, et al. "A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling." Hydrology and Earth System Sciences 25.5 (2021): 2685-2703.