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 have made four major changes to the paper based on the reviewer's comments.

1) We have rewritten the *Section 2 Methods*, shortening *Section 2.3 An Overview of the LSTM and EALSTM* and moving a large part of the model description to *Appendix A: LSTM and EA LSTM Model Description*.
2) We have made the *Section 3 Results* more concise, reducing the number of figures in the main body of text. We have focussed more attention on the interesting patterns of LSTM performance, and away from the improved performance of the LSTM compared with the conceptual models. We have also focussed attention towards interpreting the model performances in the drier, groundwater dominated catchments of the South East.
3) We have expanded the discussion of these results, exploring more clearly our three research questions:
    a) *Section 4.1.1: How well do LSTM-based models simulate discharge in Great Britain?*
    b) *Section 4.1.2: How does the LSTM performance compare with the conceptual models used as benchmark?*
    c) *Section 4.1.3: Can we extract information from the spatial and temporal patterns in diagnostic measures?*
4) We have more critically engaged with our experimental structure and the intercomparison with the lumped conceptual models (*Section 2.4.1: Benchmark Models & Section 4.1.2: How does the LSTM performance compare with the conceptual models used as benchmark?*).

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.

We have changed the abstract to more accurately reflect the gaps that our study is seeking to address. **L2-6** "*Previous studies have demonstrated the applicability of LSTM based models for rainfall-runoff modelling, however, LSTMs have not been tested on catchments in Great Britain (GB). Moreover, opportunities exist to use spatial and seasonal patterns in model performances to improve our understanding of hydrological processes, and to examine the advantages and disadvantages of LSTM-based models for hydrological simulation.*".

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 have substantially changed the introduction to address the reviewer's comment. We explicitly outline the research gaps on **L57-68**. Our research questions then follow these gaps and are outlined on **L69-72**.

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

We have updated our literature review to more accurately represent the diversity of data-driven modelling approaches. This can be seen on **L24-31**.

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

Updated as proposed.

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.

We have moved the discussion of the LSTM and EA LSTM structure to **L507-557** *Appendix A: LSTM and EA LSTM Model Description*.

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

Updated as proposed.

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

This has been updated as proposed.

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 include this information in the complete description of the model architectures in Appendix A, **L556-557**.

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 include a more complete description of the lumped conceptual models from **L196-201**.

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 have kept the CDFs of catchment metrics (Figure 1), the maps showing seasonal NSE scores (Figure 2). Previous figures that are no longer included have been moved to Appendices, Appendix E: Spatial Performances of Error Metrics. We have reduced the focus of the results on the outperformance of the LSTM compared with the benchmark models.

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).

We have updated the text to reflect how connectivity information could be learned: “Connectivity information could be represented by the hidden state (ht), or cell state vectors (Ct)” **L460-461**.

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 have flagged this ablation study (removing inputs from the model training) as a topic warranting further discussion in our upcoming paper on LSTM interpretability and added a sentence to explain our intentions of pursuing this in future work.

We have updated the text to “*In order to test this hypothesis, one could perform an ablation study, removing input features and determining the impact on model performances.*

Alternatively, sensitivity analysis could be used to determine the relative contribution of the input features to the discharge prediction, thus revealing what input features are important for the model simulations. We intend to pursue this idea in upcoming papers. " (**L466-470**).

Anonymous Reviewer #2:

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

This manuscript investigates the following research questions:

- Are (regional) LSTM-based models able to simulate the rainfall-runoff process in GB and how do they compare against different, well established hydrological models?
- Can we extract any insights from this comparison, e.g. are there certain types of catchments that are consistently modeled better by one class of models (here LSTMs), which may hint to a missing representation for a dominant hydrological process in the other model class (here the benchmark models)?

Overall, I think this is a very good manuscript that only needs minor modifications. No additional experiments are required. The list of my suggestions might seem long in the first place, however most things should be very easy to fix. I tried to list everything I found, because I hope that will make the manuscript better. However, I am happily open to discuss any of my points.

Before I start with my comments, I want to highlight the points of the manuscript that I liked:
- The study is conducted on a large-sample, public dataset that was never used before for this kind of studies.
- All code is published and it seems trivial to reproduce the results of this manuscript.
- The benchmarking includes model outputs from different research groups, making it less likely that the model comparison is biased.
- The evaluation is performed on multiple metrics that account for different parts of the hydrograph.
- The discussion and analysis of the results w.r.t. the hydrological context/region was insightful.

We thank Reviewer #2 for their careful reading of the manuscript and thoughtful comments. They have identified the research aims of our paper and we are keen to incorporate their suggestions into our revision.

Next, a few general points:

Format

- The mathematical notation is inconsistent and not in line with the HESS guidelines. E.g.
- Vectors (e.g. all gates, the cell and hidden state, and the inputs at a particular timestep) should be boldface italics lower case.
- Matrices (e.g. weights) should be printed in upper boldface roman (upright) font.

- Abbreviations are no in line with the HESS guidelines. E.g. “Figure” and “Equation” (as the entire word) should only be used at the beginning of a sentence. Mid-sentence “Fig.” and “Eq.” should be used.
- Dates are not in line with the HESS guidelines. The format of the dates should be dd mm yyyy (e.g. 31 December 2008).

We have updated all notation and abbreviations to be in line with the HESS guidelines. Thank you very much for the relevant information included in this comment.

Paper length

- The manuscript is quite long but there is potential for shortening certain parts. I think a shorter, more concise paper will ultimately make this manuscript more read by people. I do have a few suggestions, where the manuscript could be shortened but feel free to ignore all of them if you would like these sections as is:

We have moved a number of figures to the appendices (from the old manuscript Fig 1, Fig 2, Fig 4, Fig 5) and removed others (from the old manuscript Fig 7). We have kept the CDFs of catchment metrics (*Figure 1*), the maps showing seasonal NSE scores (*Figure 2*), the correlation of catchment attributes and model performance scores (*Figure 4*) the Budyko curves (*Figure 5*) and histograms showing model performances in “leaky” catchments (*Figure 5*).

- The model description of the LSTM and EA-LSTM is quite long and little information is added in comparison to the original manuscript that is cited. Personally, I don’t think that the equations and the entire formal explanation is needed and in my opinion it could be removed. I think it would suffice to have a short, one paragraph explanation of the main difference between those two models (maybe with Fig. 2) and then link the reader to the original citation. We have seen quite a few LSTM publications recently, and similar to papers with traditional hydrological models, in my opinion it is not needed again and again to write down the model equations.

We have rewritten *Section 2.2 An Overview of the LSTM and EALSTM* (**L101-124**) to summarise the differences between the two models, and moved a more complete analysis into *Appendix A: LSTM and EA LSTM Model Description*. The wiring diagram has been moved into the Appendix (*Figure A1*).

- Very similar in my opinion is the section about the evaluation protocol. Personally, I don’t think we need to see the equation of e.g. the Nash-Sutcliffe-Efficiency in every hydrology paper. Also the other equations could maybe be removed and you only keep a short explanation of the metrics with a reference to the original manuscripts. If you want to keep all equations, why not list them in e.g. a compact table like in Best et al. (2015).

We have updated *Section 2.4.2: Evaluation Metrics* to be much more concise. We have kept only a short explanation of the metrics and a reference to the original papersas proposed.

- The manuscript contains a lot of figures. Generally I like that. However, due to the length of the manuscript, there are some figures that I think could be shortened or removed (or moved to the supplement). E.g. Figure 4 spans 3 pages, however most of these maps show the same pattern.

We moved old Figure 4 and Figure 5, to the appendices (*Figure E1, E2*). We have also removed the old Figure 7. This has substantially shortened the paper.

- Lastly, I was a bit confused about the section starting in L 452ff. At this point, we reached the discussion of the results. You evaluated and compared the LSTM-based models in hundreds of basins with established hydrological models. Why do you add another comparison in the discussion with two additional models in (only) 13 basins? I see no real motivation for this additional comparison and there are no new insights gained from this comparison. Personally, I think this section can be removed entirely, or I would like to see a better explanation why this additional comparison is wanted/needed and what we get out of it that we did not know from the first, very large-sample, comparison.

We have removed this section as proposed.

Minor line-by-line comments:

- L 38: “account for temporal dependence using a series of recurrent layers”: RNNs account for temporal dependencies by processing the input time series timestep by timestep. This can (as in your LSTM model) also be done by a single layer. Using a “series of recurrent layers” would mean to stack RNN layers, each with it’s own set of weights.
We have removed the line in question.

- L 42: “Long Short Term Memory” -> “Long Short-Term Memory”
Updated as proposed.

- L 50: The cited reference for the EA-LSTM did not investigate prediction in ungauged basins. It was done in a different publication, by the same authors though, in which they did not use the EA-LSTM however.
We have rewritten this subsection and removed the sentence in question.

- L 77: “Our study poses the following four research questions:” Your enumeration only contains three research questions.
We have updated as proposed (**L57**).

- L 109: Table 2 mentioned before Tab 1.
We have updated this as proposed (**L97**).

- L 114: “The static attributes we use to train the LSTM models are listed in Table 1.” Table 1 does not list the static attributes but the notation of the mathematical symbols.
We have updated this section, the reference to the static variables is now as above (**L97**).

(Next points are only important if you decide to keep Section 2.3)
We have updated Section 2.3, moving the majority of offending sentences to *Appendix A*.

- In Eq. 1-4: To simplify, you can write [[X_t, A], h_t-1] as [X_t, A, h_t-1], since all three vectors are stacked. Although note, as stated above, that it should be boldface italics lowercase for all three vectors
We have updated everything in line with HESS guidelines (**L527-532**; **L536-540**).

- L 159 “hs” is explained in L. 219 but not here. Maybe better to explain “hs” at the first occurrence and remove the explanation in L 219.
This is now explained in *Table 2* and introduced in the text in **L164**.

- Figure 2: caption “we have 365 cells” maybe confusing with cells also regularly used to describe the number of memory cells. Maybe simply remove the last part of the Sentence.
We have updated this as proposed.

- L 185 y_hat is not explained.
We have updated *Table 2* to include y_hat.

- L 185, Eq. 12 M_theta (which is a function) should be typeset in roman (upright) font, see HESS guidelines for mathematical symbols and functions.
We have updated this as proposed (Table 2; **L139**).

- L 200 The following two sentences could be rephrased or maybe one could be removed: “We used 21 static inputs (A). Each catchment was characterised using 21 individual features describing the topographic, soil, land-cover, and climatic properties.”. Maybe just write “Each catchment was characterised using 21 individual features (A) describing the topographic, soil, land-cover, and climatic properties.”.
We have updated on **L153**: “*We selected 21 individual features describing each catchment's topographic, soil, land-cover, and climatic properties as static inputs (A)*”.

- Table 2, The median for low_prec_freq is missing.
We have updated *Table 1*.

- L 207-208 Slightly repetitive to the preceding paragraphs. Can maybe be deleted!?
Deleted as proposed.

- L 212 Just to be sure, are you using the average discharge of the ensemble or are you later reporting the average metric value, calculated as the mean/median over the ensemble members? I think it is the former, but it would maybe help to be more explicit here.

We have updated text on **L238-239** which reads: "*For the LSTM-based models the evaluation metrics are calculated given the average discharge of the ensemble*". We have also updated the caption of *Table 3* to read: "*We have shown the median catchment score for the metric given the mean simulated discharge of our ensemble*".
(End of Section 2.3 comments)

- L 247: "...chose parameters for the 4 lumped models from a grid of 10,000 parameters..." I think the formulation is slightly wrong. It is not a "grid of 10.000" parameters", but they sampel 10.000 parameter sets from a grid defined by the user defined parameter boundaries. Maybe "...chose parameters for the 4 lumped models by sampling 10,000 random parameter sets from a grid with predefined parameter boundaries..."?

We have updated the text to read: "*The benchmark study provides an assessment of conceptual model simulation performances across a large sample of GB catchments, and also quantifies uncertainty in hydrological simulations due to parameter uncertainty and model structural uncertainty (Lane et al., 2019). Parameter values for each conceptual model were selected from 10,000 simulations of multi-dimensional parameter space. The best-estimate model parameter values were selected from these 10,000 samples using the Nash-Sutcliffe Efficiency score.*" (**L203-206**).

- L 245-254 What you describe here, especially the difference of how the model parameters are selected, is indeed a huge difference. The section however, is quite long and you could maybe try to rewrite this section in a more concise/structured way. If I understand you correctly, there are two main points you want to say:

- - Lane et al. "calibrate" their models by sampling 10.000 different parameter sets and evaluating the models on the entire data record, then picking the parameter set with the highest NSE. In contrast, you train your model on one data split, and you use a different data split (of unseen data) to evaluate the models and to calculate the metrics.
- - Lane et al. find individual parameter sets per basin. In contrast, you train one model with a single set of parameters for all basins at once.
- Maybe you find a way to shorten this section and boil it down to the main points.

We have expanded our discussion to more critically engage with the experimental differences between our experiment and the approach taken for the conceptual model experiments. This discussion can be found on **L202-234**.

L 256f "An important difference between the LSTMs and the traditional hydrological models, is that traditional models perform best when calibrated for individual basins. The parameters that they use to produce simulations are unique to each basin. This often represents the state-of-the-art for traditional hydrological models." I think the last sentence can be removed, as you already said that in the first sentence. Although I agree, it might be good to have a reference for such a statement.
Updated as proposed, "*Finally, the LSTM-based models are trained on all basins, with a single set of weights for the whole of GB. Therefore, these LSTM models are regional models that are*

able to reproduce behaviours across Great Britain. In contrast, most hydrological models perform best when calibrated on individual basins (Beven, 2006)." **L419-420**.

- L 256 I feel like such a sentence needs either a reference or an experiment.
We have updated as proposed "*In contrast, most hydrological models perform best when calibrated on individual basins (Beven, 2006).*" **L419-420**.

- L 261ff "The conceptual models were calibrated and evaluated to produce simulated streamflows by Lane et al. (2019). We did not run these benchmarks ourselves. This is important because we have not biased the calibration of these models to favour the deep learning models. We have used the published time-series of model outputs to calculate performance scores for the conceptual Models." You can maybe remove this sentences, since you already stated the same at the beginning of Sect. 2.5.1. The only thing added here is the sentence of being unbiased. You could maybe add this to the first sentences of this paragraph as well. E.g. Maybe (L. 229) "We compare the performance of the LSTM based models against a range of lumped, conceptual models. To be unbiased on the model calibration, we used predicted discharge time series from Lane et al. (2019) who utilised the FUSE framework to train and evaluate four lumped conceptual models across Great Britain (Clark et al., 2008)."
We updated this sentence to: "*To be unbiased on the model calibration,we used simulated discharge time series from Lane et al. (2019) who calibrated and evaluated these four conceptual models on 1000 catchments across Great Britain*" (**L189-191**).

- L 263 Why does using simulations from someone else help you to better understand the seasonal and geographical patterns?
We removed the sentence as it was unclear.

- Table 3: Are the LSTM metrics the mean/median over the 10 repetition, or the metric value given the ensemble mean discharge? If the former, you could/should report the std/interquartile range as an error metric.
The LSTM metrics are the metric value given the ensemble mean discharge. We have updated text on **L238-239** which reads: "For the LSTM-based models the evaluation metrics are calculated given the average discharge of the ensemble". We have also updated the caption of *Table 3* to read: "We have shown the median catchment score for the metric given the mean simulated discharge of our ensemble".

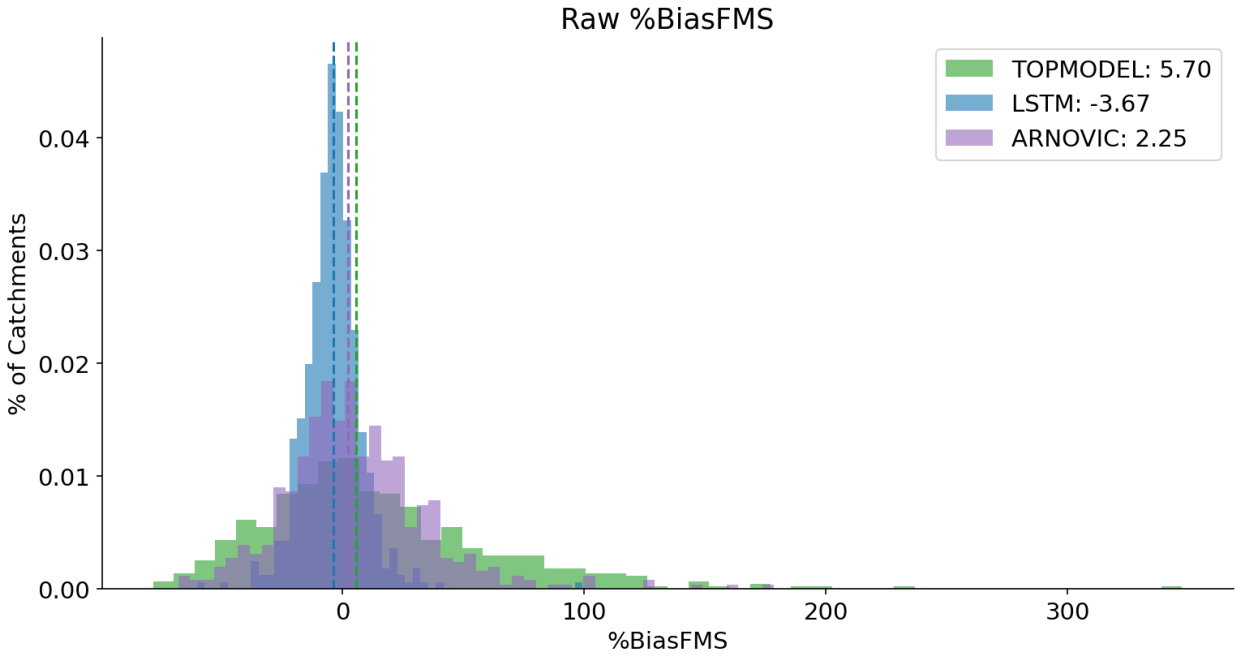
- Table 3 and L 313: Since all models model the same basins, you should use the "paired" Wilcoxon test. Furthermore, I am a bit surprised by the results of %BiasFMS: Did you test for significance using the absolute metric values of the signed metric values? Because the LSTM is actually closer to zero as TOPMODEL, making it actually the better model. Since, in my opinion, neither over- nor unterestimating is better, you should test for significance using the absolute metric values per basin for the metrics that go from -inf to +inf (with zero being the best).
We used the "Paired Wilcoxon Test" (Scipy Function with the "alternative" parameter set to "two-sided"). We clarified this in the revised manuscript: **L261**.

Our process was as follows:

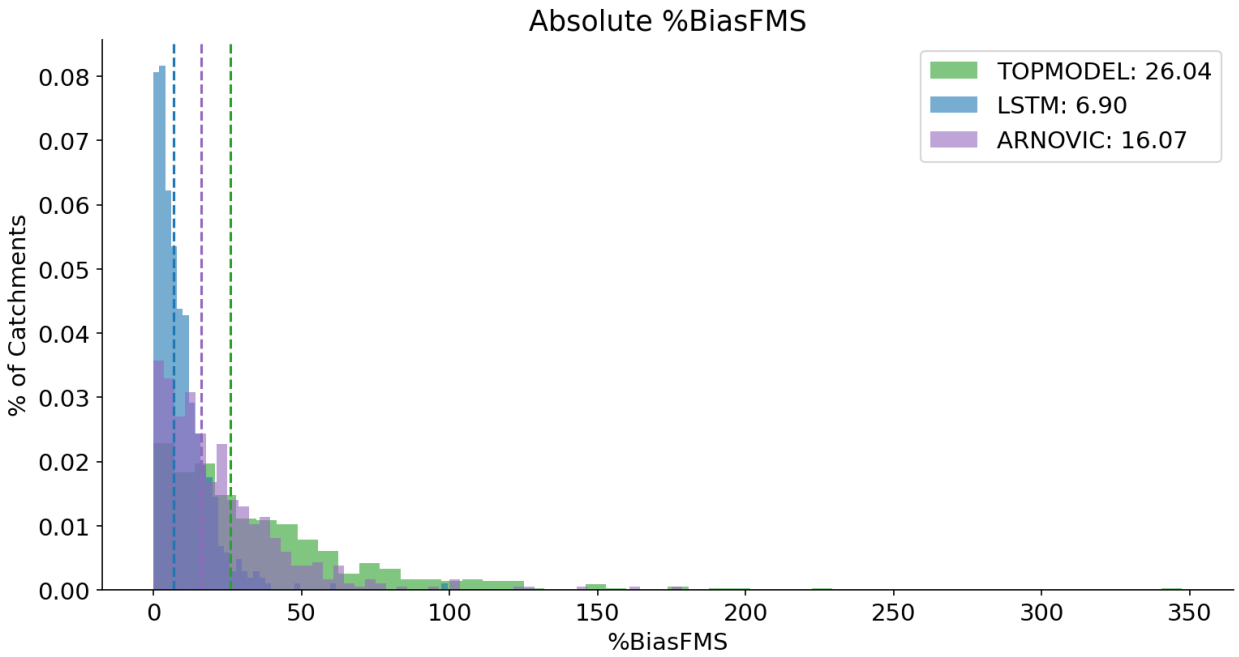
1) Calculate the paired wilcoxon test for each model intercomparison and for each statistic:
    a) LSTM vs. TOPMODEL, SACRAMENTO, PRMS, VIC, EALSTM
    b) EALSTM vs. TOPMODEL, SACRAMENTO, PRMS, VIC
    c) TOPMODEL vs. , SACRAMENTO, PRMS, VIC
    d) SACRAMENTO vs. PRMS, VIC
    e) PRMS vs. VIC
2) We presented only the results showing significant difference between the *best* model, (which was VIC for %BiasFMS, with the median score closest to zero). The difference was significant for the comparison with the LSTM but insignificant for TOPMODEL.

I think the confusion comes from the fact that **the median score obscures the similarity in the distributions** of catchment %BiasFMS scores (in this case between TOPMODEL, ARNOVIC and LSTM). Also the median of the raw BiasFMS is a little confusing, because the absolute BiasFMS scores clearly show that the LSTM outperforms the other models, but in the spirit of fairness I wanted to be consistent in the application of the metric across all metrics and all models.

The raw catchment fms: (medians shown as dashed lines and score in the keys)



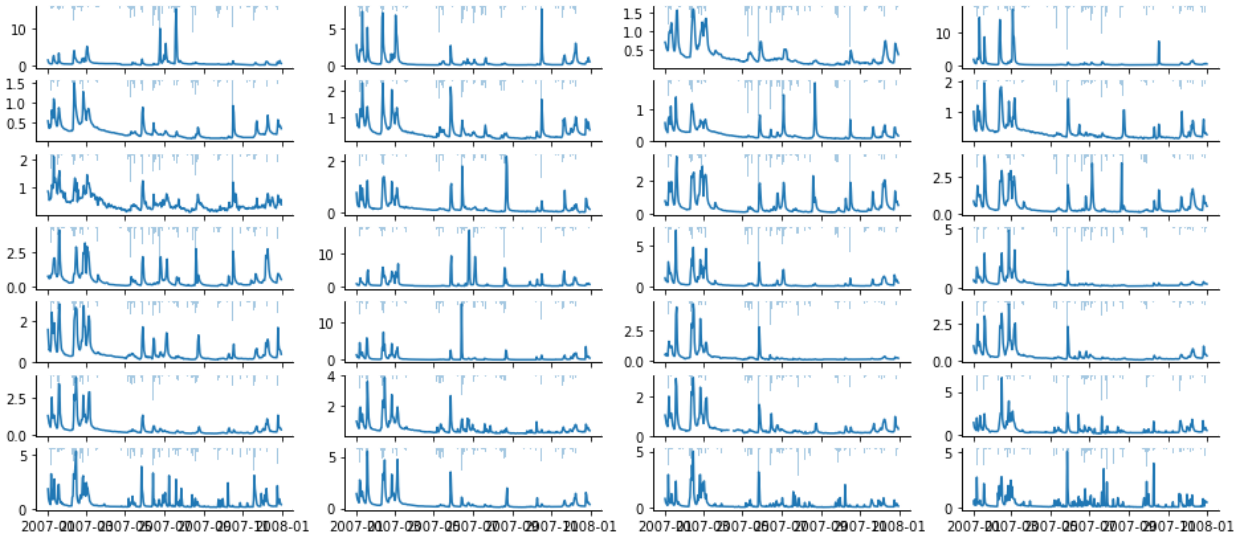
The absolute BiasFMS distributions:



L 332 The difference in low flow metrics is indeed interesting. I am not too familiar with the CAMELS GB data but I could imagine that one of the reasons might also be the difference between the two datasets (CAMELS US and GB). In CAMELS US, there are a number of basins that fall completely dry during long periods of the year, which are generally hard to model. What is “dry” in CAMELS GB, might not be in the spectrum that was reported as difficult for CAMELS US. So maybe the LSTM is not good at zero flow predictions but good for “lower flows” (with water)?!

We updated the text to reflect this comment, arguing that the results show a confirmation of the worse performance in drier catchments (**L439-440**). We removed the statement about the performance improvement relative to the results in the US.

These are the 28 catchments with the highest aridity, and we can see there are a small number of ephemeral streams in the GB dataset.



- Figure 4: As stated above, these are a lot of plots/pages. Maybe not all are necessary so the paper becomes shorter? E.g. for most metrics, the patterns are pretty much identical. There is only a visible difference in the pattern for %BiasFLV, where TOPMODEL is different to the other benchmark models. Maybe just include figures for one (or two?) metric(s) in the main paper and put all others into the supplementary?

We moved all of these spatial plots to *Appendix E: Spatial Performances of Error Metrics*, *Figure E1*.

- L 360f “An initial hypothesis is that hydrological conditions in the drier catchments with groundwater transfers remain difficult to model, requiring time-varying parameters and more detailed representation of hydrogeological properties.” Or maybe different/better inputs? Something like groundwater transfer might be hard to learn from the limited inputs that are used in this study.

We have significantly expanded our discussion about the difficulties of learning groundwater transfers, or subsurface dynamics in catchments with significant subsurface flow pathways. Key sentences relating to this comment:

- *“This suggests that the underlying data does not contain sufficient information to model the full range of processes that influence the hydrograph in these catchments, including groundwater and abstractions. The catchment averaged information on soil texture (sand-silt-clay)provides a coarse proxy for catchment porosity. Furthermore, further data, such as groundwater time-series, might be necessary to obtain more accurate discharge predictions.”* **L447-456**.
- And in the conclusions: “*Finally, the data may not contain sufficient information to capture the percolation and connectivity dynamics that drive hydrological behaviour in catchments with significant groundwater processes*” (L**507-509**)

- L 362 “...but further research should address how the LSTM might be further improved in these low-flow regimes.” Here, you are saying that LSTM performance suffers in low-flow regimes, which would be inline with the studies you referenced above (see L332f).

We have updated the text as proposed: "*The LSTM shows a performance decline in drier conditions (Fig. 4). This confirms the findings of other DL studies in the US, where the LSTM also struggled to reproduce hydrographs in drier conditions (Kratzert et al., 2019, 2018)*" (**L439-440**).

- L 376f "This means that the LSTM is overpredicting low flows, with a larger bias in the South East." Slightly repetitive to L 375, consider rephrasing.
We have removed this sentence.

- L 386 "The largest difference from GB average is 0.03 NSE". Isn't the difference 0.05?
The difference was compared to the GB average (|0.88 - 0.91| for SWESW and |0.88 - 0.85| for ANG). We have however, removed this section, and the figure has been moved to the Appendix *Figure E2*.

- L 389 "...the conceptual models show are clearly more capable in..." -> "the conceptual models are clearly more capable in..."
We have removed this section and replaced it with a discussion of where the conceptual models perform well, "*The catchments where the comparative performance difference is small, i.e. where the conceptual models perform almost as well as the LSTM, reflect areas where the conceptual models capture the majority of the information from the data, and the conceptual model well represents the hydrological process. This is the case in West Scotland, North West England & NorthWales and North East England (see Appendix Fig. E2).*" (**L422-425**).

- L 395 Delete "clearly".
Updated as proposed: "*The East-West gradient in model performances can be seen for all models, particularly in JJA*" **L291-292**.

- L 394ff I'm not sure if I agree with your summary. For me, it is almost easier to see the East-West gradient in the LSTM/EA-LSTM figures, because they switch from darker colors (all but JJA) to lighter colors (JJA), whereas the others have East-West gradient almost always (all but JJA), where as in JJA almost the entire map has lighter colors.
We updated this to read: "*The East-West gradient in model performances can be seen for all models, particularly in JJA. However, the range of errors is smaller for the LSTM based models when compared with the conceptual models. *" (**L291-293**).

- L 397ff. Figure 7 and Fig. 8 are not linked in the text and the order of the figures should probably be switched to account for the occurrence in the text (map before cdf). Also: I'm not sure if Fig. 7 and Fig. 8 are needed or if their results can be described with 2-3 sentences. Since the pattern and the results are basically always the same, i.e. LSTM is generally better everywhere and at any time. So this could be a good opportunity to shorten the paper.
We removed the old Figure 7 from the manuscript. Figure 8 results were kept in the new *Figure 3*, since the Delta NSE metrics are important for analysing the differences in model performances explicitly. **L305-329**.

- L 426f: “The catchment attributes alone are not sufficient to determine what information needs to be passed into the cell memory (Equation 8 compared with Equation 2). In other words, the LSTM learns more about the catchments’ hydrological response to rainfall from the hydrographs themselves than from the static catchment attributes.” I agree with the finding that EA-LSTM seems to be worse than LSTM, however I am not sure if I agree with the statement of these sentences. The EA-LSTM has the same discharge available to “learn from” and the LSTM has the same static attributes. You mention in other places that probably the main reason for the difference is that the EA-LSTM “freezes” one of the gates, making the EA-LSTM less flexible, which I think is the main reason for the difference.

We have updated this as proposed, the text has been updated to read: “*The EA LSTM is constrained to treat information that does not vary over time (catchment attributes) separately from information that varies over time (hydro-meteorological forcings). However, the constraint penalizes performance, which was also found by (Kratzert et al. 2019). The EA LSTM, in contrast, is forced to keep the input gate static through time. The input gate receives only information about catchment attributes. This means that no time-varying information is passed through the EA LSTM input gate. In contrast, the LSTM gates receive information from both time-varying meteorological inputs and static catchment attributes. The under performance of the EA LSTM relative to the LSTM suggests that this regularisation hurts performance in out-of-sample conditions.*” (**L394-399**)

- L 428 “For example, we can imagine a snowy catchment where we also need the temperature information to decide whether to store snow water in the network memory. The LSTM has this temperature information fed through the input gate (Equation 2), whereas the EA LSTM does not (Equation 8).” The EA-LSTM is still able to model snow, as you described in this example. The only difference is that for the EA-LSTM, the cell-update gate has to model the entire process (i.e. “that there is snow” and “how much” snow is added to the cell). In the standard LSTM, the two things can be modeled by two gates (input gate + cell update gate), making it more flexible to learn this process.
We removed the discussion of the snow process as above and focused the discussion on the key points, as shown above (**L394-399**).

- L 436 I think you mean the correct thing but just to clarify. As far as I understand, both models run on GPUs, the difference is that the standard LSTM makes use of a CUDA optimized implementation in the background, while the EA-LSTM is custom code.
We have updated this to read: “*It is worth noting that the LSTM and EA-LSTM also differ in terms of practical computational requirements. The LSTM trains much faster than the EA-LSTM. The LSTM will train 30 epochs in 1 hour, compared with 30 epochs in 10 hours for the EA-LSTM. This is due to the LSTM being an in-built Pytorch (v.1.7.1) function that makes use of CUDA optimised code (for running the models on a GPU). In contrast, the EA-LSTM relies on custom code without the CUDA enabled optimisations.*” (**L401-404**)

- L 438 At first, I was confused if the deltas are the differences of the means or medians, which is explained then in the next sentence. Maybe you could move this explanation to the beginning?
We have updated this as proposed (**L307-315**).

- L445ff Coming back to an earlier point of my review: I feel like it is worth repeating that you compare against the calibration period of the benchmark models. Most likely, a fair comparison, where you compare to out-of-sample periods of the benchmark models, would further increase the performance difference.

We have updated the text as proposed and included this point in the methods and reiterated in the discussion:

- **L223-225**: "*Therefore, the LSTM is evaluated on out-of-sample (in time) data, whereas, the conceptual model parameters were calibrated on data included in the evaluation period (in-sample evaluation).*"
- **L415-416**: "*Another difference is that the LSTM diagnostic scores are calculated on out-of-sample predictions, compared with the in-sample predictions for the benchmark conceptual models.* "

- L 478ff: I'm not familiar with all 4 conceptual models but if I'm not wrong, at least not all of them contain a snow-module in the setting used for this study. Maybe this does also explain some of the differences in North East Scotland?

We have updated the text to make this a clear conclusion:

- **L286-288**: "*We suggest that these differences in performance are due to the low rainfall and chalk aquifer in the South East of England, and to the lack of snow modules incorporated into the conceptual models for North East Scotland.*"
- **L324-326**: "*The conceptual models lack a snow module, and are therefore unable to capture snow melt or frozen ground processes, which are especially important in winter (DJF) and spring (MAM)*"
- **L429-432**: "*The performance differences in North East Scotland are very likely a result of the ability of the LSTM to learn a representation of snow processes from the input data, whereas, the conceptual models were simulating these catchments without a snow module*."

- L 490ff: Isn't another possible option based on the way how those models are trained? Imagine a basin that has constant low flow (or zero flow) for an extended period each year. All those timesteps yield little information that can be used to update the weights, since for all different meteorological inputs, the output would always have to be the same. So the underlying physical processes can only be inferred from those timesteps with varying discharge.
We have included this as a hypothesis for the results: **L440-442**: "*Basins that have long periods of low flow contain little information, since changing meteorological inputs co-occurs with very little change in the target discharge. Therefore, the physical process relating meteorological inputs to river discharge can only be inferred from those catchments with varying discharge.* "

- L516ff I am not an expert with these models, but how strict is mass conservation really? We can't see more water than what has fallen as precipitation (upper bound) but there is no lower bound, or? Since the models are not calibrated on evapotranspiration, it can vary this model output at will, to e.g. remove less water from the system than it would evaporate in reality. Additionally, some conceptual hydrology models (e.g. SACRAMENTO) have an additional option to remove water from the system that then does not reach the channel, which is the baseloss flow. This is another degree of freedom, which can be fitted at will, since the models are only calibrated on discharge. What I want to say is: Conceptual models can't "invent" water (e.g. by water transfer from a different catchment) but water can be removed at will. So personally, I don't think that a "leaking catchment" (L. 518) has to be a problem, or?

This is a very interesting point and something that we have discussed. The particular models that we benchmark against here were constrained to not remove any more water than the maximum defined by the input potential evapotranspiration. You are correct that conceptual models often have a baseloss flow, however, in the models used for comparison here, baseloss flow parameters were set to zero (i.e. excluded) and there is no baseloss flow. We have updated the text to read: "*One of the key hydrological conditions that hydrological models struggle with is the lack of closure of the catchment water balance. The conceptual models we test here explicitly maintain mass balance. They define the topographic surface water catchment as the surface over which water is conserved, i.e. the surface water catchment is not expected to leak, nor should any water enter the catchment other than through measured precipitation.* " (**L345-248**)

- L 533ff "Alternatively, the fact that both LSTMs and conceptual models struggle in catchments where data does not meet the water balance constraints might suggest that human impacts on the hydrograph are ultimately unpredictable, such as abstraction and effluent returns." Or maybe just unpredictable from the given model inputs? Even if anthropogenic influences are included in the catchment attributes, water extraction is most likely a dynamic process and would require additional dynamic inputs. But conceptually, I don't see why a data-driven model should not be able to learn this process? The process is either driven by physical processes or by a human factor, which is most likely driven by a management plan. Both things could in theory be learned, given enough (informative) inputs.

We have reformulated the argument to reflect the reviewers comments. "*This suggests that the underlying data does not contain sufficient information to model the full range of processes that influence the hydrograph in these catchments, including groundwater and abstractions. The catchment averaged information on soil texture (sand-silt-clay) provides a coarse proxy for catchment porosity. Furthermore, further data, such as groundwater time-series, might be necessary to obtain more accurate discharge predictions. We suggest that different input data sets should be tested to try and improve LSTM performances enabling the LSTM to more properly account for the complex percolation and infiltration dynamics in these catchments.* " (**L450-456**).

- L 539 Do you mean to link Figure 10? The link to Figure 11 is not clear to me.
We have updated this to link to the Budyko-curve figure (*Figure 5*) "*We tested whether the LSTM was better able to simulate discharge in catchments with "excess" water (i.e. the points below the curved lines in Fig. 5, which are then represented by the orange kernel density estimate in Fig. 6).* " (**L365-366**)

- Figure 11: x-axis label (NSE) should be in capital letters
We have updated this as proposed.

- L 551 Again, since this is the conclusion, worth that your comparison is biased (towards the lumped hydrology models), since you compare your hold-out period to their calibration period.
We included various statements critically engaging with the intercomparison of the two experiments. We updated the manuscript to reflect the reviewers comment:

- **L223-225**: "*Therefore, the LSTM is evaluated on out-of-sample (in time) data, whereas, the conceptual model parameters were calibrated on data included in the evaluation period (in-sample evaluation).*"
- **L415-416**: "*Another difference is that the LSTM diagnostic scores are calculated on out-of-sample predictions, compared with the in-sample predictions for the benchmark conceptual models.* "

References:
Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., ... & Vuichard, N. (2015). The plumbing of land surface models: benchmarking model performance. Journal of Hydrometeorology, 16(3), 1425-1442.

Beven, Keith. "A manifesto for the equifinality thesis." Journal of hydrology 320.1-2 (2006): 18-36.

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 have updated the paper to reduce the emphasis on outlining the performance improvement of the LSTM compared with the conceptual models. We have made four major changes to the paper based on the reviewers comments.

1) We have rewritten the *Section 2 Methods*, shortening *Section 2.3 An Overview of the LSTM and EALSTM* and moving a large part of the model description to *Appendix A: LSTM and EA LSTM Model Description*.
2) We have made the *Section 3 Results* more concise, reducing the number of figures in the main body of text. We have focussed more attention on the interesting patterns of LSTM performance, and away from the improved performance of the LSTM compared with the conceptual models. We have also focussed attention towards interpreting the model performances in the drier, groundwater dominated catchments of the South East.
3) We have expanded the discussion of these results, exploring more clearly our three research questions:
    a) *Section 4.1.1: How well do LSTM-based models simulate discharge in Great Britain?*
    b) *Section 4.1.2: How does the LSTM performance compare with the conceptual models used as benchmark?*
    c) *Section 4.1.3: Can we extract information from the spatial and temporal patterns in diagnostic measures?*
4) We have more critically engaged with our experimental structure and the intercomparison with the lumped conceptual models (*Section 2.4.1: Benchmark Models* & *Section 4.1.2: How does the LSTM performance compare with the conceptual models used as benchmark?*).

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. We have updated the results to more accurately address our third research question: "*Can we extract information from the spatial and temporal patterns in diagnostic measures? e.g. What is the relationship between LSTM performance and catchment attributes?*".

Section *3.3: In what hydrological conditions do model performances differ?* outlines the spatial (**L317-338**), and temporal (**L325-330**) patterns in model performances and explores the catchment attributes that correlate with model performances (**L335-344**). We explore in greater depth the impact of water balance closure on LSTM performances, highlighting that the LSTM performances are worse in catchments where the water balance does not close than in those catchments that are not "leaky" (Section 3.3.1, **L345-379**). We have also substantially increased the discussion of these results, outlining the reasons for the differences in model performances and learning from the differences in model performances. *Section 4.14* explicitly addresses the question that the Reviewer has highlighted. We first examine the performance differences in NE Scotland (**L430-433**) before exploring in more depth the conditions and explanation for differences in LSTM performances in the SE of England compared to elsewhere in GB (**L439-457**). We then explore two other aspects which warrant further discussion, firstly, improved performance in summer months (**L458-469**) and in catchments with a strong agricultural signal (**L470-475**).

Some more specific comments follow:

line 19: There are more modern PBSD models than SHE. Reference Parflow, SUMA, SHETRAN, Hydrogeosphere etc.
We have updated these references on **L16-18**.

line 77: there are only 3 research questions
We have updated this as proposed.

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 have removed this figure in order to make the paper more concise.

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 have updated *Table 2* as proposed. X_t has become 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 have updated this as proposed (**L520**).

Table 2: Why these attributes? Was LSTM sensitive to all of these?
We have updated the text to read: "*These attributes were chosen to reflect hydrological information that the model can use to distinguish between catchment rainfall-runoff behaviours \citep{kratzert2019_ealstm}.* " (**L154-155**).

line 220: What is an epoch? how does this relate to number of catchments/years of data?
We have updated the text to define an epoch: **L176-178:** "*An epoch reflects a single pass of the training dataset through the model, such that every sample in the training dataset has been used to update the model weights. This reflects the fact that during the training of DL models, the data are often split into batches to allow large datasets to be read into memory.* "

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 clarified this in the manuscript: **L262**.

Figure 3: Nice figure
Thank you!

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 have expanded 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 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 have recalculated the statistics excluding catchments with significant snow-processes and the results do not significantly change. We propose to leave the comparison as is for the reasons outlined above.

```
Original data all catchments (N=518)
\begin{tabular}{lrrrrrrr}
\toprule
{} &  nse &  bias\_error &  std\_error &  correlation &   fms &     flv &    fhv \\
\midrule
TOPMODEL    & 0.76 &       -0.04 &      -0.10 &         0.88 &  5.70 &   42.22 & -13.04 \\
ARNOVIC     & 0.78 &        0.06 &      -0.10 &         0.90 &  2.25 &  -60.34 & -14.66 \\
PRMS        & 0.77 &        0.03 &      -0.03 &         0.89 & 35.24 & -315.25 & -15.11 \\
SACRAMENTO & 0.80 &       -0.01 &      -0.07 &         0.90 & 27.91 & -195.92 & -16.19 \\
EALSTM      & 0.86 &       -0.02 &      -0.10 &         0.94 & -6.29 &   23.61 & -10.81 \\
LSTM        & 0.88 &       -0.02 &      -0.09 &         0.94 & -3.67 &   26.34 &  -9.09 \\
\bottomrule
\end{tabular}

Stations with frac_snow <= 0.03 (N=450)
\begin{tabular}{lrrrrrrr}
\toprule
{} &  nse &  bias\_error &  std\_error &  correlation &   fms &     flv &    fhv \\
\midrule
TOPMODEL    & 0.77 &       -0.03 &      -0.09 &         0.89 &  6.35 &   42.27 & -11.97 \\
ARNOVIC     & 0.79 &        0.07 &      -0.09 &         0.90 &  3.37 &  -56.83 & -14.11 \\
PRMS        & 0.78 &        0.04 &      -0.02 &         0.90 & 36.45 & -307.09 & -14.27 \\
SACRAMENTO & 0.81 &       -0.01 &      -0.06 &         0.91 & 29.20 & -194.96 & -15.61 \\
EALSTM      & 0.86 &       -0.02 &      -0.10 &         0.94 & -6.12 &   25.11 &  -9.82 \\
LSTM        & 0.88 &       -0.02 &      -0.08 &         0.95 & -3.67 &   27.95 &  -8.48 \\
\bottomrule
\end{tabular}

Stations with frac_snow > 0.03 N=(68)
\begin{tabular}{lrrrrrrr}
\toprule
{} &  nse &  bias\_error &  std\_error &  correlation &   fms &     flv &    fhv \\
\midrule
TOPMODEL    & 0.71 &       -0.06 &      -0.14 &         0.85 & -0.96 &   41.42 & -15.59 \\
ARNOVIC     & 0.75 &        0.04 &      -0.14 &         0.87 & -7.35 &  -83.36 & -18.01 \\
PRMS        & 0.73 &       -0.00 &      -0.09 &         0.86 & 25.33 & -361.99 & -19.45 \\
SACRAMENTO & 0.76 &       -0.03 &      -0.12 &         0.87 & 19.61 & -202.01 & -19.60 \\
EALSTM      & 0.84 &       -0.03 &      -0.16 &         0.93 & -8.17 &   15.66 & -16.88 \\
LSTM        & 0.87 &       -0.03 &      -0.11 &         0.93 & -4.11 &   21.38 & -12.35 \\
\bottomrule
\end{tabular}
```

line 367-371: this is a repetition of the previous paragraph.
We have removed 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 have moved this figure to the Appendix (*Fig. E2*) as proposed. We have kept the CDFs of catchment metrics (Figure 1), the maps showing seasonal NSE scores (Figure 2). Previous figures that are no longer included have been moved to Appendices, Appendix E: Spatial Performances of Error Metrics. We have reduced the focus of the results on the outperformance of the LSTM compared with the benchmark models.

Figure 6. Label missing on the colorbar
We have updated the colorbar as proposed on *Figure 2*.

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

Figure 9: significant correlations are not clear. consider showing this in an alternative way.
We have increased the size of the marks (*) as proposed on *Figure 4*.

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 have significantly expanded the discussion about the information included in the data. Section *4.1.3 Can we extract information from the spatial and temporal patterns in diagnostic measures?* explicitly deals with the information available in the underlying dataset, outlining what we can and cannot learn from the CAMELS-GB dataset. We have proposed two hypotheses about information that the LSTM captures that could explain performance improvements in summer months (**L457-469**), and in catchments with a strong agricultural signal (**L470-475**). We have also explored what are the limits to the information available in the underlying data, exploring the difficulty in modelling groundwater dominated catchments only with meteorological datasets and coarse geological information (**L447-456**). We offer an expansive concluding paragraph for this discussion section, outlining the conditions we should focus our model improvement efforts on given the information available in the underlying dataset (**L475-481**).

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 have addressed this important point in two ways.

1) Expanded discussion of recent advances in LSTM based modelling with uncertainty inputs.
    a) 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 (**L593-594**: "*Uncertainties in observations can be estimated and accounted for by using multiple forcing products \citep{kratzert2021synergy} or by resampling the input data.* ").
    b) Secondly, work by Klotz et al (2021) demonstrating three different methods for uncertainty quantification (**L603-605**: "*A more principled treatment of uncertainty, which benchmarks various methods for using DL models to directly simulate a distribution can be found in \citet{klotz2020}.*").

2) In *Appendix D: Model Uncertainty* **L574-594** we have explored the uncertainty represented by the variability in the ensemble of 8 LSTM models.
    a) We present the spatial distribution of ensemble variability as a % of discharge ("Coefficient of Variability" - *Figure D2*)
    b) We present the standard deviation of ensemble simulations for different flow exceedances (*Figure D1*)
    c) We presented hydrographs in *Appendix 2* with uncertainty bands reflecting one standard deviation of ensemble member simulations.