Review of “Benchmarking Data-Driven Rainfall-Runoff Models in Great Britain: A comparison of LSTM-based models with four lumped conceptual models” by Lees et al.

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.

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).

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:
    - 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.
    - 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).
    - 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.
    - 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.

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.
- L 42: "*Long Short Term Memory*" -> "Long Short-Term Memory"
- 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.
- L 77: "*Our study poses the following four research questions:*" Your enumeration only contains three research questions.
- L 109: Table 2 mentioned before Tab 1.
- 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.

(Next points are only important if you decide to keep Section 2.3)

- 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
- 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.
- 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.
- L 185 y_hat is not explained.
- 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.
- 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.*".
- Table 2, The median for low_prec_freq is missing.
- L 207-208 Slightly repetitive to the preceding paragraphs. Can maybe be deleted!?
- 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.

(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..."?
- 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.
- 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.
- L 256 I feel like such a sentence needs either a reference or an experiment.
- 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)."
- L 263 Why does using simulations from someone else help you to better understand the seasonal and geographical patterns?
- 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.
- 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 he 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).

- 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 there is 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)?!
- 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?
- 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.
- 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).
- 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.
- L 386 “*The largest difference from GB average is 0.03 NSE*”. Isn’t the difference 0.05?
- L 389 “...*the conceptual models show are clearly more capable in...*” -> “the conceptual models are clearly more capable in...”
- L 395 Delete “clearly”.
- 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.
- 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.
- 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.

- 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.
- 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.
- 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?
- 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.
- 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?
- 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.
- 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?

- 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.
- L 539 Do you mean to link Figure 10? The link to Figure 11 is not clear to me.
- Figure 11: x-axis label (NSE) should be in capital letters
- 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.

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.