Benchmarking LSTM

Overall, this paper stands at the forefront of hydrology. There are three aspects of the paper that I like. First, this work show state-of-the-art performance in terms of large-scale streamflow prediction accuracy. This would serve to push hydrologic science forward. Second, the authors implemented a novel LSTM structure to enable a static layer through which they could examine the impacts of different static catchment attributes. Third, they investigated network internal embeddings which is the first time in hydrology which I have seen, and provided some insights (not so perfect, as I would expand on later). These are all novel and I believe the paper should eventually be accepted.

Upon deeper examination I indeed found some issues related to potentially un-robust analysis, points of confusion and lack of clarity, need for more hydrologic insights, and somewhat superficial discussion in the exploration of embeddings. Some relevant citations are also missing. Thus I rate the manuscript a moderate revision. The comments below are not to cast the paper in a negative way, but they are in the hope of helping the authors improve the paper to a strong state before publication.

Major comments:

1. Hydrologic understanding: the discussion of the clustering and embeddings was, shall I say, not entirely satisfying. I liked the novelty of the visualization and the construct of the LSTM to enable this. It helped us understand a bit more about how LSTM works. However, I craved for a bit more hydrologic understanding. The discussion in section 3.4 was a bit sporadic and not so memorable. The take-home message appears to be "the EA-LSTM is able to learn complex interactions between catchment attributes that allows for grouping different basins". Stopping here does not help with the long-standing criticism of machine learning as a blackbox. I had hoped to gain some deeper hydrologic insights, e.g., why different basins were grouped together? What is the characteristic of each cluster and how are these clusters different from previous catchment clustering schemes, e.g., (Berghuijs et al., 2014; Carrillo et al., 2011; Fang & Shen, 2017; Sawicz et al., 2011; Toth, 2013; Troch et al., 2013)? To go deeper it may not need additional work, but more thoughts about the results.
2. More robustness: I'm afraid many of the attributes in Table 4 are correlated in space and it may be not very robust to draw conclusions from them especially for attributes that are not the highest ranking. For example, does geological permeability really stand position #9? Can we take it that permeability is the second important factor amongst non-climatic factors? This is somewhat surprising and is worth more discussion, but I'm afraid it might just be due to coincidence. To see so the authors could remove some basins (randomly or removing a spatial cluster) or attributes (as the factors tend to have interaction in these kinds of factor analysis) and train again and see how this table react to the perturbation.
3. Details for reproducibility: one of the selling points of the paper was the high performance. Hence it imperative that the results are reproducible. Are the transformations applied for input and output? How many layers of LSTM were used (in comparison with authors' HESS 2018 paper, this choice seemed ad hoc?)? How was the ranking for Table 4 done indeed? This was a local method, so what is the origin for perturbation?
4. Share more experience please: there are many choices which were unexplained, and the community would benefit from the authors providing more discussion of what worked and what did not during their experiments. How did other objective functions do? What if you don't do

ensemble averaging? How large are the impacts of hyperparameters, e.g., hidden layers and learning rates? These do not necessarily need figures and could be answered by a couple of sentences. Some minor points below are related to this.

5. The authors should also expand on why climatic factors showed up on top of table 4. It appears other static basin physical attributes were not important at all. Does this suggest catchment co-evolution? A potential indication of overfitting (to climatic factors that obviously vary), and more discussion is begging to be done here.

Minor points:

1. I'm at a loss to understand the opening statement about streamflow being an out-standing problem. At what point is this problem solved vs not solved? Is there a hard threshold? Did the present work solve this problem?
2. L73, "which part of the network are used for a given basin"—this sentence is difficult to interpret at this point. What does "used for" mean here.
3. L76, "similarly behaving". Is this referring streamflow responses or attributes? (only the former would be called a behavior, but this work didn't seem to include streamflow response in the clustering part)
4. L78, "embedding". This is a natural language processing jargon. Quite difficult for hydrologists to comprehend. I think it would be reader friendly if the authors spend two sentences explaining this word. My understanding is that embeddings are not just hidden layer activations, but a mapping of inputs to an ordered hidden space that has meanings. For example, the hidden layers of machine translation layers form an embedding. Each ranked item in the embedding in NLP can be related to a linguistic concept.
5. L117 "some amount of information" is fuzzy. Is it about catchment attributes or about streamflow responses? This is critically important as the two have very different meaning regarding what would be done. From reading the later parts, here you seem to refer to static attributes.
6. L122, regarding using static attributes as a constant array. It would be relevant to cite (Fang et al., 2017) which used this setup and was already distinguishing different landscapes using static attributes as inputs to LSTM. It occurs this paper should at least be mentioned in the present one.
7. L134-135. This is an interesting setup. It's worth mentioning that, from Eq 9 & 11, what was selected by the input gate were not only x_d but also h from the last step.
8. L158 – what happened when you used other loss functions?
9. L171 "25,000 km2" – is it really appropriate to model those with an area of 25,000 km2 the same as other smaller basins?
10. L194 – "favor of"
11. L222, regarding the ensemble averaging, readers deserve to know, how big is the spread? What if you don't take the average? Sometimes the ensemble mean gets a better R2 but it misses peaks.
12. It is unclear what "six different settings and eight different models" are.
13. L261. might be useful to say you extracted gradients from the learned network after training (correct?), as some readers are unfamiliar with how this is done. However, these gradients are time-step dependent.

14. Also, why is it called global sensitivity test? It is also local, around a origin for perturbation.
15. L264 better say "the average of absolute gradients across all basins and all time steps", and---- why absolute?
16. L267-268 “represent xxx into xxx"? the sentence does not make grammar sense. please fix. This is obviously an expansion of from 27 to 256. Why would this be really necessary?
17. Table 2. this value is indeed the highest I have seen. Good work!
18. L380. Why 447 basins now? What are missing?
19. L410 Unsure how this answers the question if the network just remembers. The logic is confusing.
20. L414, mean precipitation, etc --- aren' these supposed to be climatic inputs rather than static? (can we not let the network generalize it from the forcing data)?
21. Table 4. Echoing a major point raised above. What further conclusions can be drawn from the fact that climatic attributes take the most important positions? catchment Co-evolution theory?
22. L454 “before vs. after the transformation into the embedding layer”. This is a good comparison, although later there didn’t seem to be much comment on this comparison
23. UMAP—might be good to briefly explain what it does. Is it just PCA?
24. L479 Honestly, it’s not that easy to see which cluster you are talking about. could use some annotation on the plot.
25. L489 I found this discussion, as a take-home message, to be somewhat superficial, and unsurprising. I'd appreciate somewhat more in-depth discussion about the hydrology.
26. Figure 11 these colors do not mean anything. It is a bit confusing. Why not use a at least partially consistent color scheme?
27. Figure 12 better annotate axes even if they don't mean much
28. L524. I am confused why this is called regional, as the LSTM was trained with all basins over CONUS. What would constitute a model that is not regional?
29. L529-530. It either goes against a belief or it does not. Can't go "somewhat against". And, the logic here is not quite clear. This paper is not about parameter identification. The fact that the network works does not imply that parameters can be identified. First the LSTM parameters cannot be interpreted. Second, even very different parameters could give you similar predictions.