Review of *Better continental-scale streamflow predictions for Australia: LSTM as a land surface model post-processor and standalone hydrological model* by Shokri et al.

June 5, 2025

Overview

This manuscript compares a standalone LSTM (LSTM-C) and an LSTM that includes simulated streamflow from a land surface model (AWRA-L) as an additional dynamic input (LSTM-QC). The two LSTM-based models are additionally benchmarked against AWRA-L and GR4J, a conceptual hydrological model widely used in an Australian context. They compare the models using three different cross-validation strategies, evaluating the ability of the models to predict temporally out-of-sample, spatially out-of-sample, and spatiotemporally out-of-sample. Overall, they show that the two LSTM models outperform AWRA-L and GR4J in most catchments under all three cross-validation strategies, although there are some notable exceptions. The authors discuss the potential relevance of their study in three real-world applications, namely historical reconstruction, predictions in ungauged basins, and simulating hydrological change under climate change projections.

Main comments

1. As far as I can tell, LSTM-C and LSTM-QC are identical in every respect except for the inclusion of streamflow from AWRA-L as an additional dynamic input in LSTM-QC. I therefore question whether this paper is really testing whether the LSTM can correct the AWRA-L output, or whether the AWRA-L output provides any additional information content that can be leveraged by the LSTM architecture. In effect these two possibilities amount to the same thing, but the paper would benefit from emphasising one over the other. In my opinion, the latter characterisation more accurately reflects what the LSTM is actually doing.

2. The results of the comparison with LSTM-C (i.e. the LSTM model that does not include the AWRA-L streamflow as a dynamic input) is undermined by the limited number of static catchment attributes that are supplied to the model (Table 1). A large number of studies have shown that LSTM models perform best when they are trained across many catchments at once using catchment descriptors that adequately describe the physical diversity of catchments in the training set. For example, (Kratzert et al., 2019) train an LSTM on static catchment attribtues that include soils, climate, vegetation, topography, and geology. Here, the authors have selected attributes that broadly cover climate and geomorphology, but discard a large number of attributes from CAMELS-AUS that are potentially highly influential in determining the hydrological behaviour of Australian catchments (e.g. geology, land cover). Presumably, in common with most land surface models, AWRA-L is parameterised using land cover and geological data. Therefore, I believe it is at least a possibility that LSTM-QC is utilising the information on catchment diversity that is encoded in the AWRA-L output but which has been arbitrarily excluded from LSTM-C. It seem to me that LSTM-C is trained in a way that is inconsistent with our current understanding of how best to use this class of model for hydrological simulation, raising doubts about whether it is a fair comparison.

Minor comments

1. L27: "The ubiquity of these model predictions..." - are you referring to the spatial coverage or widespread use? Please clarify.

2. L33: It's worth pointing out that most land surface models were not originally designed to predict streamflow, but rather to provide the lower boundary condition to Earth system models.

3. L39-48: I agree that the lack of channel routing and calibration scheme are weaknesses of AWRA-L with respect to streamflow simulation, but is a lack of process understanding not also a weakness?

4. L62: Punctuation needed.

5. L73: A third advantage is that they are unconstrained by physical laws such as mass balance, so they are better able to implicitly correct biases in the input data. In land surface models, uncertainty in the input will propagate to the output.

6. L85-93: This passage is not particularly relevant to the topic in hand. As the introduction is already quite long it could be safely removed.

7. L98: "... as we show in the current study..." - This would seem to pre-empt the results.

8. L99-100: Arguably deficiencies in routing and bias in individual catchments amount to the same thing. Perhaps you could clarify what you mean here?

9. L122: I'm not sure it is particularly easy to test the ability of the model to perform well under climate change projections, because it is likely that the range of input values in the climate projections will exceed those the LSTM would encounter in the training set. Thus what you really ought to be testing is the ability of the model to extrapolate, but I'm not sure the experimental design achieves this at present.

10. L156: You could acknowledge here that using multiple precipitation datasets can enhance LSTM performance (Kratzert et al., 2021).

11. L175: Please clarify that you are referring to the hidden state size here.

12. L221: You describe the static and dynamic predictors, but not the target (i.e. streamflow). Please could you describe your treatment of the target variable (e.g. do you normalize by catchment area)?

13. L225: Please could you confirm that the two LSTM models are identical in every respect except for the inclusion of AWRA-L streamflow in LSTM-QC?

14. L240: You say this important but not that you actually do it. We later find out that you have, although this information is in the results section. Please consider moving 3.1.3 to 2.5.2.

15. L245: In general I think the training approach for LSTMs is well established and so you don't need to go into so much detail here. The text could also be shortened by using scientific notation (e.g. Section 2.3 of (Lees et al., 2021)). Typically when training an LSTM there will be a training period, a validation period (that is used during training to test each parameter set) and a hold-out test period. However, Table 2 only details a training and validation period. Please could you clarify whether the model is tested on an unseen dataset?

16. L265: This needs some clarification. I think it is feasible (i.e. it could be done under the experimental setup) but not meaningful, because in a real out-of-sample situation you would not have any data to conduct fine-tuning.

17. L285: I can see the argument for including GR4J in the model comparison, but I wonder whether it would be better to only use it in the TooS test. I would argue that by including GR4J in the SooS and TSooS tests you are really testing the parameter regionalization scheme, which is not really the main focus of the manuscript.

18. Figure 4/5/6: Your description of the results would benefit from using subplot labels, so the reader knows what they should be looking at.

19. L373: Notwithstanding my previous point about climate projections, I'm not sure why this is categorised as TSooS rather than TooS?

20. L390: I'm not sure it is meaningful to compare with LSTM-C at short sequence lengths, as we already know that LSTMs require long sequence lengths to make good predictions.

21. L481: This could arise because the LSTM training is suboptimal, as it has not been exposed to catchment attributes that may help it learn the hydrological behaviour in these regions.

References

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. *Water Resour. Res.*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Kratzert, F., Klotz, D., Hochreiter, S., & Nearing, G. S. (2021). A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. *Hydrol. Earth Syst. Sci.*, *25*(5), 2685–2703. https://doi.org/10.5194/hess-25-2685-2021

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., & Dadson, S. J. (2021). Benchmarking data-driven rainfall–runoff models in Great Britain: A comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. *Hydrol. Earth Syst. Sci.*, *25*(10), 5517–5534. https://doi.org/10.5194/hess-25-5517-2021