Revision of: "From RNNs to Transformers: benchmarking deep learning architectures for hydrologic predictions" by Jiangtao Liu et al.

June 12, 2025

1 General comments

The manuscript "From RNNs to Transformers: benchmarking deep learning architectures for hydrologic predictions" by Jiangtao Liu et al. compares various deep learning models (from LSTM over transformers to LLMs) for estimating and forecasting various hydrological variables like runoff, soil moisture, snow water equivalent, and dissolved oxygen. The models are tested for five different tasks: regression, data integration, autoregression, spatial cross-validation, and zero-shot. The results show that LSTM often performs best on regression tasks, while attention-based models are getting better for more complex tasks.

This manuscript provides very useful insights for hydrological modelling using deep learning algorithms, presents the results in a comprehensive way, and is well-written. However, the methods are not yet sufficiently described to make it clear how useful the results are. Thus, I recommend reconsidering the manuscript for publication after the methods section has been adjusted. The points that need to be addressed are described below.

2 Specific comments

As mentioned above, the only major point to address is a more elaborate description of the methods used in this manuscript. Without a clear description of the methods, it is hard to evaluate the usefulness of the results and reproducibility is not given. Specifically, I would need more information about:

- test-train splitting
- did you do hyperparameter-tuning or how did you decide on the hyperparameters, resp. on the number of hidden layers, nodes etc.
- how are the point datasets used? Are they interpolated to raster format or are they used as predictors as points? Is the lat/lon information of the points used in the model too?
- how is the data extracted? Average over whole catchment or all pixels per catchment added to model?
- what is the temporal resolution of the input data (resampled to daily or weekly or used as is?)
- how are the LLMs used for hydrologic modelling? Just providing the data and asking them for the target variable?

Some more minor points to address are the following:

- line 16-21: the different model setups do not get clear from the abstract. This part needs to be reformulated to clarify which parameters are estimated, which tasks are done (regression, zero-shot etc.) and which models are used for these tasks.
- line 122: how did you decide on the thresholds to exclude basins larger than 5000 and smaller than 50km2?

- line 123: what type of manual quality checks did you do?
- line 135: a table with all predictors and datasets would help.
- line 148: are all other DL tools more efficient? A comparison of the calculation time of all models would be interesting (maybe in the supplementary files).
- line 165: for me, the term forecasting would be more intuitive. Or why did you choose the term data integration?
- line 180ff: Is this the same setup as Kratzert et al 2021 used? If so, it would be interesting to show how your model results perform in comparison to theirs.
- line 190ff: would it be possible to finetune the LLMs to the task of hydrologic modelling?
- line 220: Is it fine with the journal to have a combined results and discussion section?
- line 226: justify the statement that LSTM reach their performance ceiling
- Fig. 2: has wrong y axis lables. The scores should be on the x axis. Also, what is it the CDF from? Please also add the mean for an easier comparison.
- line 302: why do you use for every validation type another variable? It would be easier to focus on one variable.
- Fig. 3: Same as for Fig. 2. These results are hardly comparable like this.

3 Technical corrections

- line 22ff: please also mention in brackets the performance metrics (e.g. NSE) after mentioning which model performs best and write how much better this is than the other models.
- line 110f: you write 5 datasets but mention only 4.
- line 126: please also cite the ISMN paper from Dorigo et al..
- Table 1: include also LSTM
- line 165ff: Also mention the lead times for which you do the forecasting here.
- line 218: add citations
- line 247: ISMN has not been mentioned before, please do so with the citation mentioned above.
- line 305: Please add Fig S3 and Tab S3 in the paper and not in the supplementary files.
- line 314: please add R2 or other metric to text here when mentioning that Pyraformer performs best.
- Fig. 1: Please take LSTM as first bar in the plots, since it is the baseline.
- Fig. 4: Please add NSE or other metric to plot

4 Review Criterias

Scientific significance: Does the manuscript represent a substantial contribution to scientific progress within the scope of Hydrology and Earth System Sciences (substantial new concepts, ideas, methods, or data)?

(1) Yes the manuscript represents a substantial contribution

Scientific quality: Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)?

This point cannot be validated yet, more information about the methods are required, as pointed out in my comments.

Presentation quality: Are the scientific results and conclusions presented in a clear, concise, and well-structured way (number and quality of figures/tables, appropriate use of English language)? (2) The text is well written and the conclusions are clear and concise. The figures can still be improved, as pointed out in my comments.

4.1 Further review points

1. *Does the paper address relevant scientific questions within the scope of HESS?* Yes
2. *Does the paper present novel concepts, ideas, tools, or data?* Yes
3. *Are substantial conclusions reached?* Yes
4. *Are the scientific methods and assumptions valid and clearly outlined?* No, this is not clear yet, as pointed out above.
5. *Are the results sufficient to support the interpretations and conclusions?* Yes, although, this is also dependent on the last point, so further info on methods required to validate this.
6. *Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)?* No, not yet.
7. *Do the authors give proper credit to related work and clearly indicate their own new/original contribution?* Yes
8. *Does the title clearly reflect the contents of the paper?* Yes
9. *Does the abstract provide a concise and complete summary?* Some improvements could be done to make it easier understandable. I left some comments for that in the the minor comments
10. *Is the overall presentation well structured and clear?* Yes
11. *Is the language fluent and precise?* Yes
12. *Are mathematical formulae, symbols, abbreviations, and units correctly defined and used?* Yes
13. *Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated?* Yes, as indicated above.
14. *Are the number and quality of references appropriate?* Yes
15. *Is the amount and quality of supplementary material appropriate?* Yes