Reviewer #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.

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:

1. test-train splitting

   We appreciate the reviewer's suggestion. We will include a new subsection clearly describing our train-test splitting approach. This subsection will detail temporal splits to prevent information leakage and spatial cross-validation to evaluate model generalization.

2. did you do hyperparameter-tuning or how did you decide on the hyperparameters, resp. on the number of hidden layers, nodes etc.

   We appreciate the reviewer's comment. In our revision, we will clarify that the hyperparameters for the LSTM models were selected based on previously published studies. For the Transformer-based models, we performed hyperparameter tuning using the validation subset of the CAMELS dataset, following prior recommendations.

3. 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?
4. how is the data extracted? Average over whole catchment or all pixels per catchment added to model?
5. what is the temporal resolution of the input data (resampled to daily or weekly or used as is?)

   We appreciate the reviewer's questions. We plan to clarify these points in the revised manuscript as follows:

   - For point-scale observations (e.g., ISMN soil moisture and SWE), the observed values are directly used as model targets. Corresponding predictor variables (dynamic meteorological variables or static attributes) are extracted from gridded datasets based on their geographic coordinates, without interpolation to raster format.
   - For basin-scale datasets (CAMELS, Global Streamflow, and DO), observed values at basin outlets serve as model targets. Predictor variables are derived by averaging gridded inputs over each catchment.

- Latitude and longitude coordinates are used solely for the spatial extraction of predictor variables and are not directly input into the models.
- All input data have a daily temporal resolution, and no resampling was performed.

6. how are the LLMs used for hydrologic modelling? Just providing the data and asking them for the target variable?

   We appreciate the reviewer's question. To address this, we will clarify in the manuscript how LLMs were applied to hydrologic modeling. Specifically, historical streamflow data and basin attributes will be formatted as textual prompts, and numerical predictions will be derived from the LLM responses.

Some more minor points to address are the following:

7. 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.

   We appreciate the reviewer's suggestion. To clarify this point, we will revise the abstract to describe the prediction tasks evaluated (regression, forecasting, autoregression, spatial cross-validation, and zero-shot forecasting), specify the hydrologic variables estimated (e.g., runoff, soil moisture, snow water equivalent, dissolved oxygen), and identify the models employed.

8. line 122: how did you decide on the thresholds to exclude basins larger than 5000 and smaller than 50km2?

   We appreciate the reviewer's question. In our revision, we will clarify that these thresholds (50 km^2 and 5,000 km^2) were adopted following Beck et al. (2020). Specifically, basins smaller than 50 km^2 were excluded to ensure sufficient spatial resolution for gridded meteorological inputs, while basins larger than 5,000 km^2 were excluded to minimize the influence of channel routing at daily timescales.

9. line 123: what type of manual quality checks did you do?

   We appreciate the reviewer's question. We will clarify in the manuscript that our manual quality checks involved visually inspecting streamflow time series plots to identify problems such as prolonged flat lines, abrupt discontinuities, or insufficient data length.

10. line 135: a table with all predictors and datasets would help.

    We will add a table summarizing all predictors and datasets used in this study.

11. 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).

    We appreciate the reviewer's question. To clearly address this point, we will include a detailed comparison of computational times for all models in the Supplementary Information. Additionally, we will discuss how the computational efficiency of PatchTST and TimesNet decreases as the number of catchments increases.

12. line 165: for me, the term forecasting would be more intuitive. Or why did you choose the term data integration?

We appreciate the reviewer's comment. We initially used the term "data integration" to maintain consistency with our previous studies. However, we agree that "forecasting" is clearer and more intuitive. Therefore, we will replace "data integration" with "forecasting" throughout the manuscript.

13. line 180: 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.

We appreciate the reviewer's question. In the revised manuscript, we will clarify that our benchmark model is an LSTM with a setup closely matching that of Kratzert et al. (2021). Prior to conducting our main experiments, we confirmed that our benchmark LSTM achieves similar performance (median KGE $\approx$ 0.80), thereby indirectly enabling comparison with the results reported by Kratzert et al.

14. line 190: would it be possible to finetune the LLMs to the task of hydrologic modelling?

We appreciate the reviewer's suggestion. Although fine-tuning LLMs specifically for hydrologic modeling was beyond the original scope of this study, we recognize its potential value. In the revised manuscript, we will highlight fine-tuning approaches, such as prompt-based fine-tuning and parameter-efficient methods, as promising directions for future research.

15. line 220: Is it fine with the journal to have a combined results and discussion section?

We appreciate the reviewer's suggestion. We will verify the journal's guidelines and check recent HESS publications to confirm whether a combined "Results and Discussion" section is acceptable.

16. line 226: justify the statement that LSTM reach their performance ceiling

We appreciate the reviewer's comment. We will provide a more detailed explanation by referencing recent studies, which showed minimal performance differences between LSTM and Transformer models, suggesting inherent data limitations rather than model constraints.

17. 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.

We appreciate the reviewer's suggestion. In our revision, we will correct Fig. 2 by placing the evaluation metrics (scores) on the x-axis and the cumulative density functions (CDF) on the y-axis. Additionally, we will clarify the meaning of the CDF in the figure caption and include mean values for easier comparison. Detailed numerical values will also be provided in the Supplementary Information.

18. line 302: why do you use for every validation type another variable? It would be easier to focus on one variable.

We appreciate the reviewer's question. In the revised manuscript, we will state that only the regression task utilized multiple hydrological variables (runoff, soil moisture, snow water equivalent, and dissolved oxygen) to assess model generalization. In contrast, all other tasks were conducted using the CAMELS dataset for streamflow prediction. We will highlight this distinction to clarify our methodological choices.

19. Fig. 3: Same as for Fig. 2. These results are hardly comparable like this.

    We appreciate the reviewer's suggestion. To address this concern, we will improve the comparability of Fig. 3 by adding horizontal dashed reference lines at CDF = 0.5 and including numerical median values in the Supplementary Material.

3 Technical corrections

20. 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.

    We appreciate the reviewer's suggestion. We will revise the abstract to include performance metrics (e.g., KGE) in parentheses when indicating which model performs best, and we will specify how much better it is compared to the other models.

21. line 110f: you write 5 datasets but mention only 4.

    We appreciate the reviewer's suggestion. We will revise the manuscript to clearly list all five datasets: two runoff datasets (CAMELS and Global Streamflow), as well as datasets for soil moisture (ISMN), snow water equivalent (SWE), and dissolved oxygen (DO).

22. line 126: please also cite the ISMN paper from Dorigo et al..

    We appreciate the reviewer's suggestion and will include the ISMN reference (Dorigo et al.) in the manuscript.

23. Table 1: include also LSTM

    We appreciate the reviewer's suggestion and will revise Table 1 to include the LSTM model.

24. line 165ff: Also mention the lead times for which you do the forecasting here.

    We appreciate the reviewer's suggestion. In our revision, we will specify that forecasts are evaluated at lead times of 1, 7, 30, and 60 days.

25. line 218: add citations

    We appreciate the reviewer's suggestion. We will add appropriate citations in the revised manuscript.

26. line 247: ISMN has not been mentioned before, please do so with the citation mentioned above.

    We appreciate the reviewer's suggestion. We will revise the manuscript to introduce the International Soil Moisture Network (ISMN), including the appropriate citation, before its first mention.

27. line 305: Please add Fig S3 and Tab S3 in the paper and not in the supplementary files.

    We appreciate the reviewer's suggestion. We will move Fig. S3 to the main manuscript to improve readability. However, due to space limitations and the large size of Table S3, we prefer to keep it in the supplementary materials.

28. line 314: please add R2 or other metric to text here when mentioning that Pyraformer performs best.

We appreciate the reviewer's suggestion. We will revise the manuscript to include performance metrics (e.g., KGE) in the text when mentioning Pyraformer's results.

29. Fig. 1: Please take LSTM as first bar in the plots, since it is the baseline.

We appreciate the reviewer's suggestion and will revise Fig. 1 to place LSTM as the first bar in each plot.

30. Fig. 4: Please add NSE or other metric to plot

We appreciate the reviewer's suggestion and will add the R^2 metric to Fig. 4.