Review of Manuscript

‘A GNN Routing Module Is All You Need for LSTM Rainfall–Runoff Models’

by H. Mosaffa et al.

Dear Editor,

I have reviewed the manuscript. My conclusions and comments are as follows:

1. Scope

The article is within the scope of HESS.

2. Summary

In their manuscript, the authors test a series of data-based hydrological models combining an LSTM for catchment rainfall-runoff with a Graph Neural Network (GNN) for river routing. They test four GNN alternatives: Graph Convolutional Network (GCN), Graph Attention Network (GAT), Graph SAmple and aggreGatE (GraphSAGE), and Chebyshev Spectral Graph Convolutional Network (ChebNet). As a benchmark, they use a standard LSTM-only model. Neither of the models receives any observed streamflow as input. They apply their models to a 30-year daily data set of meteorological drivers (precipitation, air temperature, soil moisture), static attributes and river gauge observations of 530 catchments in the upper Danube catchments taken from the LamaH-CE data set. The catchment sizes range from tens to thousands of square kilometers. Analyzing the time-wise out-of-sample testing results, the authors find that all LSTM-GNN models outperform the LSTM-only, that the GAT is the best among the GNN alternatives, and that performance improvements by the GNN addition increase with catchment size and number of upstream connecting nodes.

3. Evaluation

Overall, this is a well-designed study on a relevant topic, with conclusions that are supported by the data and that that will have an impact on data-based hydrological modeling. There are only a few adjustments necessary for clarification and to increase readability of the paper.

- In the LSTM-GNN approaches, each subcatchment has a unique place in the river network graph. Is each catchment LSTM then jointly trained with all catchments, but each as a single-catchment LSTM, or are they jointly trained as regional LSTM? Please clarify. Also, the LSTM-only, is it trained as a regional model? Please clarify.
- Are the GNN nodes placed only at gauge locations, or also at river confluence points? Please clarify.
- How are the catchment areas for each gauge determined for the LSTM-GNN and the LSTM-only? I assume for the LSTM-GNN it is either the upstream catchment (for a headwater gauge), or the intermediate catchment between the closest upstream gauge(s) and the current gauge, correct? And for the LSTM-only it is the total upstream catchment, independent from the presence of any upstream gauges. If my assumptions are correct, then the catchment-averaged drivers and static attributes will differ between the LSTM-GNN and the LSTM-only. Please clarify. Also, consider including two histograms or cdf's showing the distribution of catchment sizes for both the LSTM-GNN and LSTM-only case.
- Up to Vienna, the Danube catchment already covers more than 100.000 km², but the largest catchment included in this study is only 2.500 km² (Line 97). Why? The advantage of LSTM-GNNs over LSTM-only should be even more evident for very large catchments. Please clarify.

- Sequence length is 180 days (Line 130). Why not 1 year, especially if there are snow-dominated catchments included?
- For the LSTM-GNN, is the LSTM-output of the individual catchments rescaled to [m^3/s], before it is fed into the GNN part? In other words, how is it assured in the workflow that the relative runoff contribution of each catchment is correctly represented? Please clarify.
- Fig. 3: Unclear for which hop the results are shown. Also, in subplot d the color-coding is missing.
- All Figures: The dashed lines are hard to see. Use color-coding only.
- Fig. 5: Is this really a plot of all testing timesteps and all gauges? Please clarify. Visually this is dominated by the floods in the few largest catchments. Maybe it is more illustrative to show plots of scaled streamflow, where separately for each gauge, streamflow is scaled [0,1].
- Fig. 6: This is not very helpful as it does not reveal any distinct pattern. Consider replacing it with a plot of dNSE vs. catchment size (similar to Fig. 7d, but more detailed).
- Fig. 7 d: Here it seems that catchments > 100.000 km^2 are included, but only few. How does this match with the statements in Line 97?
- Fig. 9: Use same color-coding as in previous figures.
- As the topic is closely related, the authors may wish to take a look at a recent preprint by Kraft et al. (2025).

Yours sincerely,
Uwe Ehret

References

Kraft, B., Kauzlaric, M., Aeberhard, W., Zappa, M., and Gudmundsson, L.: DROP: A scalable deep learning approach for runoff simulation and river routing, 10.22541/au.176410929.91946608/v1, 2025.