Review of HESS Manuscript
"Neural networks in catchment hydrology: A comparative study of different algorithms in an ensemble of ungauged basins in Germany"

Report 1

Thanks for authors' efforts on replying to the comments and making revisions. My comments are as follows.

Specific comments:

Authors made a detailed analysis among the three models in terms of different metrics, segment assessment, and model sensitivity. However, it is unclear for readers to understand the significant mechanisms in the three methods which contribute to varied performances. The conclusions stated that CNN model offered superior performance, LSTM model exhibited superior generalization capabilities across the entire spectrum of flow data, but the GRU model showed a promising balance between predictive accuracy and computational demand. Are these conclusions consistent with other studies, or only valid in this study area? I suggest to add a part to discuss the reasons causing the differences and limitations in the three methods as well as the suggestions for future research.

We thank the reviewer for their insightful comment. In response, we have thoroughly revised the conclusions to address the concerns raised. The updated version (L607-673) now explicitly discusses the underlying mechanisms that contribute to the differences in model performance. Furthermore, we have incorporated a critical comparison with existing literature to assess the generality of our findings. We also added a part, discussing the reasons behind the observed performance variations, the limitations of each method, and suggestions for future research directions.

Report 2

Dear Editor,
I am attaching the third review of the manuscript.

Comment 1: About the Min-Max scaler.
To clarify, in my previous comment I was not criticizing the Min-Max transformation, I just indicated that probably this is the reason the sigmoid activation yields to better results, because both the target and the simulated values are mapped to a 0-1 space.
If you want to keep using Min-Max that is your decision. However, I believe there is a problem with the test you are conducting to justify this decision (Fig 1 on the author´s response), where you show that the Min-Max works much better than the StandardScaler. There is no logical reason to have such a significant drop in performance when you use the StandardScaler. There are multiple benchmarks (Kratzert2019, Less2021, Loritz2024) that have used the StandardScaler, and none of them present such bad performances.
Moreover, in the CAMELS-DE study (Loritz et al., 2024) the authors run a LSTM that was trained on daily data, using the StandardScaler transformation and without a sigmoid activation. The authors report the results for 93 catchments in Hesse (the same region you are studying) and according to their results, the median NSE for the basins in Hesse was close to 0.88. Therefore, I do not think the results you presented in Fig1 of the author´s response are correct.
You can still use the min-max scaler, and keep the results of the articles as you did. This would not change the message of the paper. However, you should eliminate the parts where you indicate that the MinMax scaler was chosen because it gave better results, because the results you presented in Fig1 of the author´s response are not consistent with existing literature. For example, I would suggest eliminating this sentence:
Line 120: Subsequently, the two data sets were normalised by employing the a min–max scaling method, with a range of [0,1] chosen as the boundaries. This method was favoured over the standardization approach employed by Kratzert et al. (2019a), as it consistently yielded superior predictive performance across all models utilized in the study.
Reference:
Loritz, R., Dolich, A., Acuña Espinoza, E., Ebeling, P., Guse, B., Götte, J., Hassler, S. K., Hauffe, C., Heidbüchel, I., Kiesel, J., Mälicke, M., Müller-Thomy, H., Stölzle, M., & Tarasova, L. (2024). CAMELS-DE: Hydro-meteorological time series and attributes for 1582 catchments in Germany. Earth System Science Data, 16(12), 5625–5642. https://doi.org/10.5194/essd-16-5625-2024

We thank the reviewer for their thoughtful and detailed comments.

In contrast to our study, Loritz et al. (2024) and Kratzert et al. (2019a) did not incorporate categorical static features requiring explicit encoding. Therefore while using label encoding standardization might produce misleading scaled values, since the values are not normally distributed. We observed that two input features (precipitation , catchment size), as well as the target variable (discharge), exhibit strong positive skewness. Standardization assumes symmetric distributions and can be destabilized by extreme values, whereas MinMax scaling bounds all inputs within [0,1], promoting training stability, especially in networks using sigmoid activations. We have revised the manuscript to clarify that the choice of MinMaxScaler was empirical and dataset-specific, and have removed any generalized claims regarding its general superiority.

According to the reviewers suggestion we removed “This method was favoured over the standardization approach employed by Kratzert et al. (2019a), as it consistently yielded superior predictive performance across all models utilized in the study.” and added Line 118-119 “The choice of this scaling method was made empirically based on observed performance in our dataset and model configuration.”

Comment 2. About using a sigmoid at the end of the pipeline.
You are justifying using a sigmoid based on Fig1 of the response you gave. Again, if you want to keep using a sigmoid that is your decision, and probably the message of the article will not be affected. However, I do not think that the KGE metric, in which you based your decision, is the best for this case. The sigmoid will saturate in high values, and therefore the difference between the models with different activations will be seen in the highest peaks. Therefore, a metric as the KGE that gives an overview of the overall performance will probably not summarize the saturation problem caused by the sigmoid. If you want to see differences, you should focus on the highest peaks. Again, this will probably not change the points made on the paper, but you should consider that the explanation you are given can be biased by the metric you are reporting.
Even though in line 610 you are speaking clearly about the limitation given by the sigmoid:

“While the sigmoid activation function provided stable performance, its combination with Min–Max scaling constrained discharge predictions. Employing LeakyReLU could allow for greater flexibility in discharge predictions, albeit with the trade–off of potential negative values.”
I would suggest emphasizing this point a bit further. The sigmoid activation is artificially decreasing high flows and imposing a structural constraint that you cannot go above what you saw in training. You can keep using the sigmoid, and the message of the paper will probably still be the same, but you should state that other configurations should be preferred in practical cases, especially if one is interested in predicting extreme discharges (e.g, flood forecasting).

We thank the reviewer for highlighting the potential bias introduced by using KGE to evaluate model performance under sigmoid activation. We agree that sigmoid activations can theoretically induce saturation effects at both low and high flow extremes. However, as shown in our flow-segment analysis (Figure 8), the models demonstrated robust predictive skill primarily for the highest flow quartile (Q4), while performance for low and mid-range flows (Q1–Q3) was consistently poor across all architectures (KGE-metrics). This indicates that, despite the bounded nature of the sigmoid output, our models retained the ability to capture peak flow dynamics effectively, whereas low-flow conditions presented a greater challenge. Given that our analysis explicitly addressed the highest peaks, the reviewer's concern is not entirely clear to us.

We added Line 656-661:
"Certain design choices and limitations must be acknowledged. Both recurrent models (LSTM and GRU) constrained outputs to non–negative discharges within the training data range using sigmoid activation and min–max normalization. This constraint ensures physically plausible predictions but restricts extrapolation beyond maximum observed flows. This saturation effect may attenuate extreme flood peaks, limiting the model’s extrapolation capacity. For practical applications requiring accurate flood forecasting (primarily focusing on high discharge), alternative activation functions such as LeakyReLU, which allow unbounded outputs, may offer greater flexibility and should be considered in future model designs.”

Other minor comments for the article:
Line 26: Modify to: As demonstrated by Kratzert et al. (2019a), an artificial neural network (ANN) model, namely Long Short–Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997) , has shown unprecedented accuracy in PUB.

Changed as proposed by the reviewers. The Text reads now: “As demonstrated by Kratzert et al. (2019a), an artificial neural network (ANN) model, namely Long Short–Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997), has shown unprecedented accuracy in PUB”

Line 34: DOI to Ghimire et al. (2021) is not working.

We could not find any issues with this DOI.

Line 120: I would suggest removing this sentence (see first comment).

Removed as proposed by the reviewer.