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

Dear Editor,
Please find attached the second review of the manuscript.

Major comment 1
In sections 2.5.1 and 2.5.2, you explain that after the dense layer of both the LSTM and the GRU, you applied a sigmoid. Is there a reason why this was done? This is not a usual approach, and existing benchmarks such as Kratzert2019 for CamelsUS, Lees2022 for CamelsGB and Loritz2024 for CamelsDE, do not apply this. A sigmoid is going to remove negative values but is also going to restrict high values. This might be the reason why the min-max transformation was working better (mentioned in the first revision), because you were restricting the training data to a 0-1 range.
With the sigmoid you cannot go above 1. Therefore, even during validation and testing, there is a structural restriction that your model cannot produce values larger than the maximum value in training. Also, the sigmoid will saturate the output of the model for large values, so even if the lstm/gru is trying to go higher, the sigmoid will cut the value short, and will disproportionally constrict higher values to a smaller range (due to the gradient of the sigmoid in high values of x). This is a major structural deficiency for the models.

In the course of finalizing our modeling setup, we systematically evaluated multiple normalization and activation strategies. First, we tested both standardization (as applied by Kratzert et al., 2019) and MinMax scaling for each of the three models under investigation (CNN, LSTM, and GRU). As shown in Figure 1, MinMax scaling consistently yielded significantly higher average KGE values. While we did not conduct a detailed root-cause analysis to explain why standardization performed less favourably in our study, the superior performance of MinMax scaling, documented in Figure 1, provided a clear motivation to pursue this approach.

In parallel, we also evaluated four different output-layer activation functions for the LSTM. According to Table 1, the sigmoid activation produced the highest max KGE results, slightly outperforming alternative functions such as Leaky ReLU, Softplus and Linear. Given these findings and in the interest of consistency across the recurrent architectures, we employed sigmoid activations in both the LSTM and GRU models without conducting a separate activation function analysis for the GRU model. However, as can be observed in Table 1, the performance differences between sigmoid and leaky ReLU were relatively small, with leaky ReLU even performing slightly better for the GRU model. Based on the concerns you raised regarding the restrictive nature of the sigmoid function—such as its inherent limitation on output range and its saturation effect for larger values—we acknowledge that leaky ReLU might have been a more suitable choice. Nonetheless, our decision to use the sigmoid activation function was driven by the fact that our MinMax-scaled data ranged between 0 and 1. Additionally, when using leaky ReLU, we observed instances of negative discharge predictions, which are physically implausible and undesirable in our specific hydrological context. Given this, the sigmoid activation function appeared to be a reasonable choice despite its structural limitations.

We acknowledge, however, the deficiency when using sigmoid combined with a MinMax Scaling approach. Future research may wish to explore alternative approaches—such as reverting to Leaky ReLU or employing specialized output transformations—to enable the model

to capture very high discharge events more accurately. We have added a recommendation to this effect in the Discussion section, noting the trade-off between potentially negative predictions and the ability to represent the full dynamic range of discharges.

In summary, although we concur that a final sigmoid layer could impose structural limitations on predicted discharge extremes, our empirical evaluations showed that this configuration, in combination with MinMax scaling, resulted in the most robust performance across the training and test datasets used in this study. We have incorporated an explicit recommendation in the manuscript for future work to revisit these design decisions, especially concerning the choice of activation function.

Thank you for highlighting this point, as it underscores the importance of scrutinizing how normalization and activation function choices can impact model performance and interpretability.

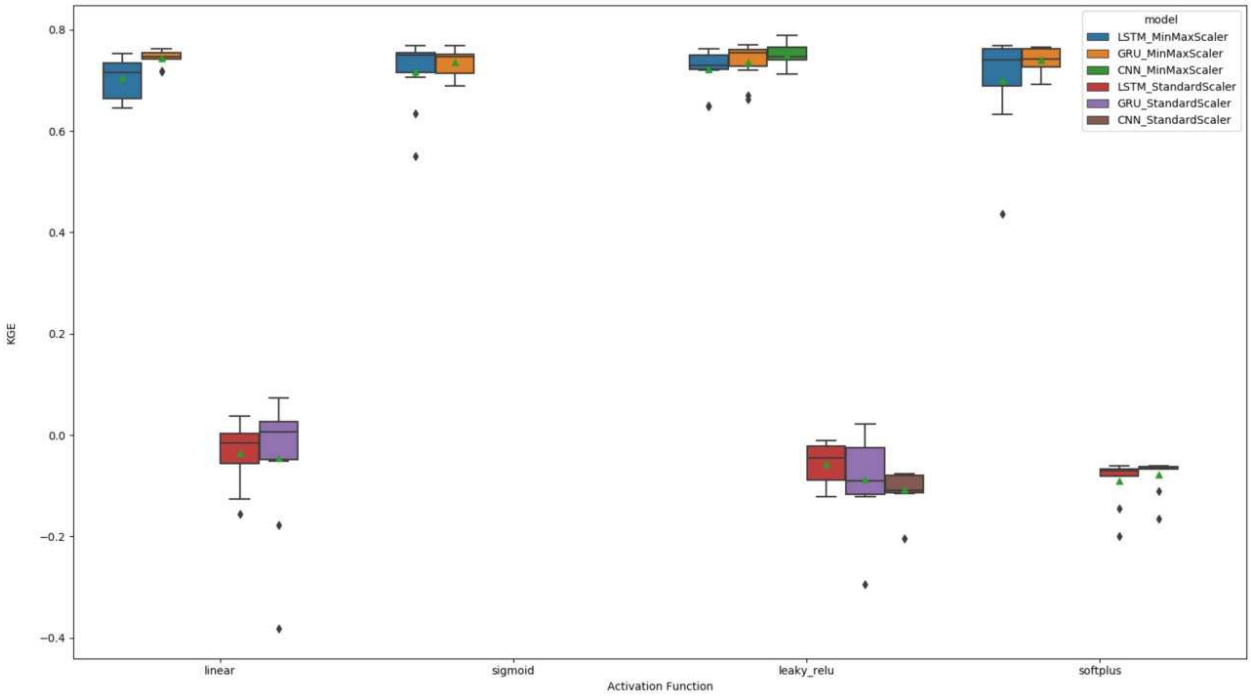

Figure 1 Comparison of Min-Max Scaling versus Standardization with Respect to Activation Function: Each boxplot represents the distribution of the mean Kling-Gupta Efficiency (KGE) across all catchments for 10 iterations.

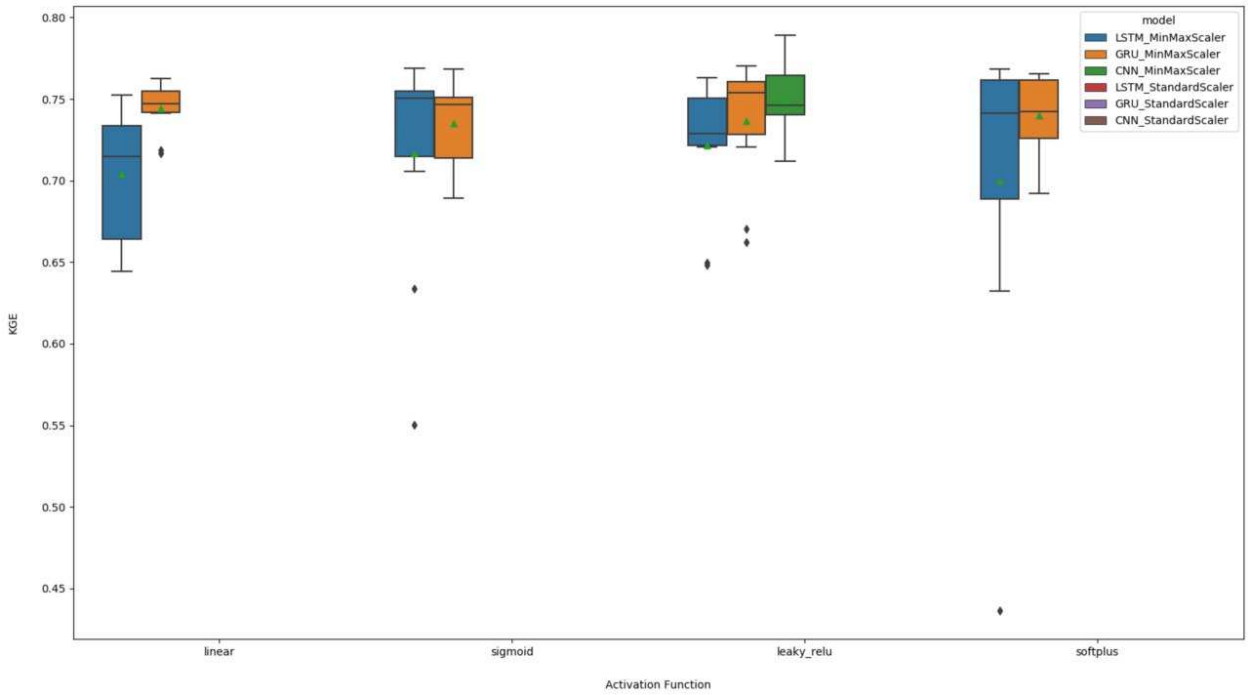

Figure 2 Comparison of different activation functions Each boxplot represents the distribution of the mean Kling-Gupta Efficiency (KGE) across all catchments for 10 iterations.

Table 1 Optimal Activation Functions for Various Metrics Across Different Models

	Median		Mean		Max	
CNN	leaky_relu	0.7460	leaky_relu	0.7500	leaky_relu	0.7893
GRU	leaky_relu	0.7537	linear	0.7442	leaky_relu	0.7703
LSTM	sigmoid	0.7503	leaky_relu	0.7214	sigmoid	0.7688

Major comment 2

About the statistical-significance tests suggested in the first review process, the authors indicated in their response that:

"…However, as it is not our intention to claim at this point that any of tested models is better or worse, we decided to leave this test for future work."

I do not agree with this point. The title of the paper states "…A comparative study of different algorithms..." and section 3.1 directly compares the models. Moreover, you state conclusions as "Despite the general use of LSTM models, this study demonstrated that CNN models offer advantages in terms of performance and runtime for time series prediction." Therefore, I think you are testing which models are better or worse, and to test if the reported differences are significant or just due to random initialization, would greatly benefit the paper. The ensembles are created after all the hyperparameter tunning steps, and given the reported times to train each model, I think would be feasible to do this.

As suggested by the reviewer, we have added a section demonstrating the statistical robustness of the values in line 387-400

Minor comments

Line 8: “dynamic input features” is a more common term than “non-static input features”.

Changed as proposed

Line 26. Cite LSTM paper, (Hochreiter and Schmidhuber, 1997).

Changed as proposed

Line 44: Change “one model fits all” to “regional”.

Changed as proposed

Line 62: I would not say that LSTM incurs at tremendous computational costs. One can train for the whole CAMELS-US or CAMELS-GB in around 5 hours, if one has a normal GPU. It is for sure more than a FFNN, but considering other deep learning applications, it is quite affordable, and even comparable to training ~600 conceptual models.

Changed as proposed

Line 60-70: This is more of a personal style, but as a suggestion, in a technical report one should avoid adjectives such as “distinguished” and “renowned”. Also, at the beginning of the introduction, you mention “paramount importance”. It is better to just state facts. Again, this is personal style.

Changed as proposed

Line 112: Change Moreover to Furthermore, because the previous connection in line 109 was Moreover.

Changed as proposed

Line 134: I would remove “remarkably”.

Changed as proposed

Line 166: The (Hochreiter and Schmidhuber, 1997) citation in this line is not consistent. What part of the phrase are you citing? If it is “have been extensively discussed in prior research” then cite it after part, otherwise you should remove it.

Changed as proposed

Line 259: Change to “a multiple of…”.

Changed as proposed

Line 353: Rephrase “a testament to the proficiency of these artificial model”

Changed as proposed

Line 527: Which notable differences are you referring to?

This sentence was a fragment of a previous revision, where the sensitivity analysis included static features. It is now removed.

Appendix A: Units of discharge are mm/day. Also, all the figures in this appendix have the same name! The names should be different and indicate the details of the figure.

Changed as proposed

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

Dear Editor,
Please find attached the second review of the manuscript.

Major comment 1

For the metrics provided in Table 5, it seems the differences are not obvious among the three methods. For example, the mean of KGE in the case of with batch size of 256 and +SF features is 0.8, 0.78 and 0.77 for CNN, LSTM and GRU respectively. In addition, some parameters such as window size shown in Table 6 are different. It is difficult to interpret the models' performance based on their algorithms and make a comparison. Could authors explain more the different behaviours from three methods performance like the model structure?

We added a paragraph analysing the influence of the window size towards the model performance. (line 384-400)

Major comment 2

All models failed to make good prediction of lower flows (Q1, Q2 and even Q3 in Figure 7). Authors tried to avoid a 'smooth' prediction using RMSE which capture the mean variability in the training dataset but instead using the KGE as lose function. Although the correlations, deviations and means of observations and simulations are considered in the KGE, the predictions produced high variability ratios. Authors explained that "This phenomenon may be attributed to a bias in the KGE towards elevated flows, thereby inadequately penalising inaccuracies in lower flow predictions in Lines 467-468". Could authors make a clearer explanation for that? In addition, I was wondering whether these behaviours are related to the training datasets of the 54 catchments, which may include a large part of high flows but limited with low flows.

A clearer explanation of this issue has been provided in lines 501–505.
Regarding the second part of the question, Figure 1 presents a histogram comparing the discharge distributions of the training and test datasets. The test dataset exhibits a slightly higher proportion of very low discharge values than the training dataset, suggesting a potential underrepresentation of low-flow conditions in the training set. In contrast, the distributions for moderate and high discharge values are largely similar between the two datasets. Although the highest discharge values are cut off in the histogram, the QQ plot (Figure 2) shows that the test dataset contains higher absolute discharge values than the training dataset for highest flows.

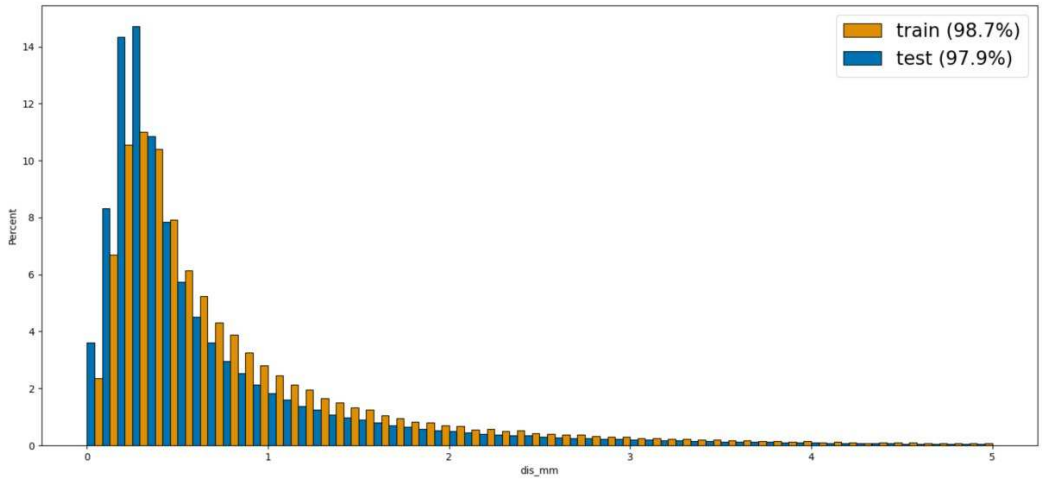

Figure 3 Comparison of Discharge Distributions in Training and Test Datasets.

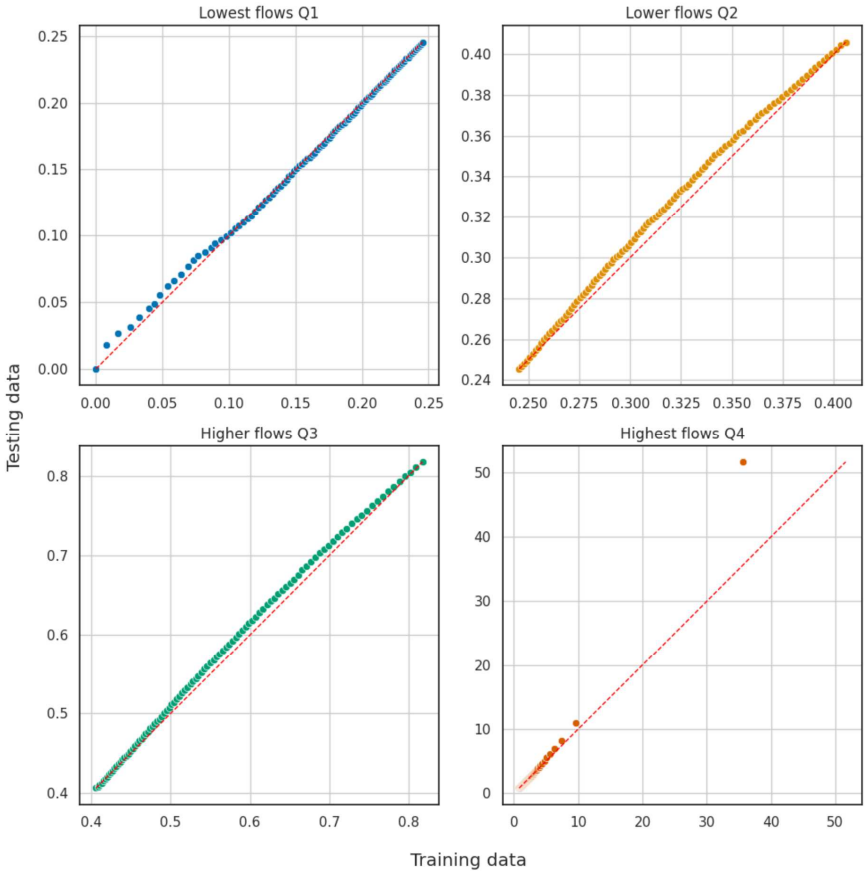

Figure 4 Quantile-Quantile (QQ) Plot Comparing Training and Test Discharge Distributions.

3) For the run time in Section 3.2, does the measured runtime including the training time or only the online prediction time?

The runtime of the model specifically refers to the training duration. This has been clarified in line 417.

4) It is not clear that whether discharge changes (%) in different scenarios were found in all test catchments or the averaged changes of the test catchments.

In line 548 we stated “The newly predicted discharge values were then systematically averaged over both time and all catchments” However, to provide further clarification, we have added the following sentence in line 552 – 553 “The results of this analysis are shown in Figure 10 representing the mean percentage change in discharge, calculated by averaging over all daily predictions and across all 35 catchments.”

5) Authors made a detailed analysis among the three models in terms of different metrics, segment assessment, and model sensitivity. However, it is not clear for readers to understand the differences in results and the more detailed explanations (e.g., comparative tests or references) and discussions are suggested to include in results.

We do not fully understand the reviewer’s question and would appreciate additional clarification to ensure we address the concern appropriately.