Our answers are highlighted in yellow

# Revision 1

**Major comments:**

Metrics:

The authors make their analysis by reporting the mean metrics. When reporting results for regional models usually the median is a more robust estimate, because it is influenced less by outliers. That is why studies like Kratzert et al (2019b) https://doi.org/10.5194/hess-23-5089-2019, Lees et al (2021) https://doi.org/10.5194/hess-25-5517-2021 and Feng et al (2022) https://doi.org/10.1029/2022WR032404 (among others) report both mean and median or only median. This property of being robust against outliers is especially important when one is comparing the different models because if one is reporting only the mean, one really bad-performing catchment can affect the overall mean value and the relative ranking between the models might change. Is there a reason why the authors prefer to use the mean? I am not saying that the authors need to change their reported metric, maybe they can just calculate the median for a couple of metrics to see if the relative ranking between the models stays the same. I leave it to the discretion of the authors and the editor to see if this change is necessary.

*This is a really good point, and there is no specific reason for choosing only the mean values within our study. We added the median to table 5 as a more robust metric in the revised version of the manuscript to support a better comparison of the models.*

Statistical-significant:

The differences between the overall model performance when considering static features (Figure 5) are quite small (CNN: 0.8 – LSTM:0.78 – GRU:0.77). You are reporting in line 317 that the CNN is slightly better than the others. Are these differences statistically significant? A quick test that can be made is to train a small ensemble (3-5 models) with different random initializations, to see if the differences are just due to the stochastical nature of the training process or are indeed statistically significant. I leave it to the discretion of the authors and the editor to see if this change is necessary.

The reviewer has a good point here, a test on significance would be a very good addition, not only to this manuscript, but too a lot of recent publications in this field. 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. However, in the spirit of the reviewers raised point, we decided to call for statistical tests in the discussion section of the manuscript.

The text reads: "Additionally, to better compare the performance differences among the models, multiple runs of each model with different random initializations should be considered, as small observed differences might result from the stochastic nature of model initialization and optimization processes." (L709-711)

Sensitivity analysis:

I do have a major concern during the sensitivity analysis. There is no reason why data-driven models should interpret static attributes as we do. When training/testing a regional model, the static attributes help the model contextualize the type of basin it is dealing with. However, I do not think that there is enough information during training to be able to do a proper sensitivity analysis with the static attributes. In this case, the authors have 11 static attributes. This means they have an 11-dimensional space filled with only 54 points. I would argue that it is difficult to infer/capture relevant patterns in these conditions, and this could be the reason why there are confusing patterns in the reported results. For example, in some models, a positive perturbation in a static attribute increases the discharge, while in other models the same effect decreases it. The authors mention this in several cases:

- Line 525: "The disparate effects of soil type classes on discharge are complex and do not appear to adhere to a discernible logic"
- Line 534: "However, the hierarchy of the categories does not follow a structured sequence."
- Line 569: "The causative factors behind this distinct behaviour remain ambiguous and are likely not solely attributable to soil type characteristics"
- Line 524: "Land use emerged as an important factor, with changes in dominant land use leading to notable variations in discharge prediction, albeit with some differences in magnitude and directionality between the models."
- Line 626: "Soil type and texture displayed varying impacts on discharge prediction across the models, with some categories showing consistent trends while others exhibited divergent behaviours"

Therefore, I am not sure that the patterns that are being reported as logical/significant, have an actual causal connection or are associated with other factors. I would suggest that the authors reduce the sensitivity analysis to only perturbations in the dynamic inputs, which I think are quite consistent for most of the cases, and are better justified. However, if there is a reason why the authors consider that the sensitivity analysis of the static attributes should be kept, I will gladly reconsider/discuss it.

Your concern is right. There is not enough data to fully understand the behavior of the models with regard to static features. We might got lost in finding patterns where there is no statistically robust evidence, and therefore decided to remove all parts of the sensitivity analysis with regard to static features.

**Minor comments:**

Line 110: The authors are saying that "encoding accommodates ordinal scales, which is better suited for hierarchical features such as permeability". However, it should also be considered that encoding gives a sequential order to classes in which a

sequential order does not make much sense (soil type, soil texture…). It is ok to use encoding, this does not need to be changed, but you should also indicate the disadvantages.

We agree with the reviewer. The text reads now “Moreover, label encoding accommodates ordinal scales, which is better suited for hierarchical features such as permeability. In contrast, categorical features without a meaningful order, such as soil type or soil texture, are better handled by one–hot–encoding, which treats each category independently” (L109-113)

Line 114: Is there a reason the min-max transformation was used instead of standardization?

In our previous trials, the min-max scaler showed better results, so we kept using it. As the focus of this manuscript is not on the transformation approach, we decided to inform the reader about the approach used and do not show any comparison.

Line 121: At first it is not clear the purpose of the moving window. If I understood correctly, it is to create the batches. If so, please indicate this at the beginning.

We agree with the reviewer. The text reads now "To transform the data sets into training batches a two–dimensional moving window, characterized by dimensions T × D, was subsequently implemented, where T represents the moving window size, also known as look–back period or sequence length (Figure 2).” (L123-125)

Line 142: What do you mean by “with the exception of the recurrent layer”?

We meant that the actual LSTM and GRU layer within our models, which inherits the recurrent calculation, is the only difference between the two models. The text reads now “Because the employed LSTM and GRU models possess an identical layer structure, both models share an equivalent set of hyperparameters.”(L145-147)

Line 221: “The loss function is regulated by an algorithm known as the optimizer.” Can you further elaborate here? Because I would argue that the opposite case is true. The optimizer is regulated by the loss function.

The reviewer is right and this error has been corrected in the revised version of the manuscript. The text reads now “The optimizer, an algorithm designed to minimize the loss, regulates the process of updating the model’s parameters. This optimizer strives to enhance the model performance by iteratively determining the loss and then adjusting the model parameter to reduce this loss” (L225-226)

Line 224: “Thus, by minimising the loss, the machine learning model can improve its predictive accuracy and thereby enhance its capacity to generalise from the training data to unseen instances”. I do not agree with the second part of this sentence. By minimizing the loss, you indeed improve the predictive accuracy, but this action by itself does not assure you good generalization capabilities. If you overfit your model,

you will reduce the training loss as much as possible, but you will lose the generalization capabilities of the model.

The reviewer is right, we changed the sentence to “Thus, by minimizing the loss, the machine learning model can improve its predictive accuracy” (L229-230).

Line 226. “The optimizer used for all utilised models in this study is the Adam–optimizer.” It would read better if you removed utilized. Also, would be better to cite the paper here instead of the next line.

Changed as proposed.

Line 237: In line 115 you indicate that the target is also scaled. If you scale your target data, then the direct evaluation of certain facets of the discharge time series by the KGE is no longer applicable (see section 6.3 of Santos et at., 2018). If this is the case, please modify this sentence.

The study of Santos et al. 2018 states that when using logarithmic transformation it can come to issues. But since we use a linear scaling method (min-max scaler) and reversed the scaling before calculating the KGE, there should be no issues.

Line 272: What is the purpose of increasing the learning rate in the first 3 epochs?

The purpose of the warmup period is to allow the model to explore the parameter space with a smaller learning rate initially, before transitioning to the main learning rate schedule. This strategy helps to prevent large fluctuations in the loss function during the early stages of training and facilitates a smoother optimization process.

https://doi.org/10.48550/arXiv.1812.01187
https://doi.org/10.48550/arXiv.1706.02677

Line 278: “The analysis depicted in Figure 5 delineates a comparative evaluation of model efficiency…” Figure 5 indicates model performance, not model efficiency.

Changed as proposed.

Line 339: It does not make much sense to compare to Nguyen et al. (2023a) if the conditions are so different. If I understood correctly, they trained single basins LSTM and are not evaluating PUB. This is different from what you are doing. The other comparison by Kratzert makes more sense.

A model that is calibrated to only a single basin tends to give better results than a model that is generalized over several catchments, especially for PUBs. We wanted to show that the model trained in this study shows even better results than the specialized model. We clarified our argumentation: “In the context of existing literature, Nguyen et al. (2023a) reported an NSE of 0.66 for an LSTM model calibrated across three distinct catchments, each with its own separate calibration and not extending to ungauged scenarios. While models calibrated to individual basins often perform better than those generalised across multiple catchments,

particularly in PUB, our results demonstrate that the generalised models trained here achieves even better results than these specialized model" (L351-355)

Line 346: Would be good to write the PBIAS equation. Because depending on how you write it, a positive/negative value can be associated with under/over estimation.

Changed as proposed. For clarity, we added the formula of all used metrics. (L345-348)

Line 507: You state, "The daily forcing evapotranspiration showed a positive impact of 0.4%. The observation that daily evapotranspiration increases with discharge is seemingly counterintuitive. However, daily evapotranspiration derived from Jehn et al. (2021) represents actual evapotranspiration, which can increase with wetter conditions and therefor also correlate positively with discharge." This is the argument that you use to justify the result. On the other hand, when analyzing the other two models, you indicate that the LSTM and GRU produce lower discharge with higher evapotranspiration. In line 550, you mention "all meteorological features of the LSTM model align with anticipated behavior, consistent with conventional understanding of hydrological processes". Similarly, in line 589, you state that for the GRU model "All meteorological features in the GRU model exhibited expected behaviors, aligning with established hydrological principles, a consistency observed in the LSTM model as well." Given that you are using the same input data for all models, the argument for the CNN case becomes invalid. The ET-Q relationship cannot be consistent in the CNN if it is positive and also consistent in the LSTM and GRU if it is negative.

We agree that if the input data is the same, the response of the models should also be the same. However, it appears that the different modelling approaches interpret the input data differently. From our point of view, there is no objective, unambiguous reason to judge one or the other interpretation of the models as 'correct'. Our argumentation therefore explains the possible different behaviours of the models without evaluating them. We have made this clearer in the revised version of the manuscript "Although this may offer a plausible explanation for the observed anomalous behavior, it is unlikely within the context of this study. Given that all models share the same input features, both the LSTM and GRU models should exhibit similar behavior, which is not observed (see Figure 9)." (536-539).

Line 508: Typo: therefore

Changed as proposed

Line 517: I do not agree that it is coherent that a higher percentage of sand and a lower percentage of clay should produce higher discharges. Sand has a higher infiltration capacity, and clay has a lower permeability. Therefore, depending on the case, you will have higher discharges with higher content of clay, because less water will infiltrate, and more direct runoff will be produced.

All soil types such as ‘sandy loam’, ‘silty loam’, ‘loam to sandy loam’ and ‘silty clay’ led to a reduction in runoff of between -0.7 % and -3.6 %. The decrease in runoff from ‘sandy loam’ to ‘silty clay’ could be explained by a reduction in the sand content and thus increasing water holding capacity of the soil (Easton and Bock, 2021). However, the lower infiltration capacity of clayey soils compared to sandy soils could also lead to a lower infiltration rate and increased surface runoff. OVerall, the catchments we investigated in Hesse are typically humid mountainous catchments in which surface runoff plays a minor role in runoff generation processes (Jehn et al. 2021, Breuer et al. 2009) and subsurface stormflow dominates (Chifflard et al. 2019). Due to the shortening of the sensitivity analysis, this point is now redundant.

Line 648: “revealing that all employed model architectures predominantly provide an authentic representation of the influence of input features”. I am not sure if your data supports this for the static attributes. The different models present different behaviors toward changes in the static attributes.

We changed the statement to cover daily feature only. The text reads now: “The sensitivity analysis provided valuable insights into the interpretability of the models, demonstrating that all model architectures accurately capture the impact of the non-static input features, with the exception of daily evapotranspiration in the CNN model. Precipitation emerged as the most significant driver of discharge predictions across all models. “ (L689-697)

# Revision 2

This paper compared three neural networks: CNN, LSTM and GRU, in discharge prediction in 54 catchments in Hesse, Germany. Detailed model comparison were conducted according to batch sizes, in/ex-cluding static attributes, computational efficiency, as well as model sensitivity. Generally, the structure of the paper is clear and well-organized, however, there are some concerns.

Specific comments:

1) For each catchment, it is unclear that authors used only one time series or the spatial averaged time series discharge data and three meteorological factors, as well as for the 11 static catchment features.

We are not sure, what you mean. But we added additional information about the discharge data. The text reads now “For each catchment, daily sum of precipitation [mm], daily sum of evapotranspiration [mm] and soil temperature in 5 cm soil depths [°C] are available along with the corresponding discharge [mm]. The discharge data is obtained from a gauging station located within the respective catchment.” (L89-91)

2) In the model training section, it seems that the parameters are set to be same in the three models. However, the CNNs generally need more epochs to converge than recurrent neural networks. Could authors list the specific/optimized parameter values for each model?

The specific hyperparameter of each model are presented in Table 6.

3) Could authors elaborate why static catchment attributes improve the overall model performance in line 284?

The text reads now " This aligns with the findings presented by Kratzert et al. (2019b), who assert that static catchment attributes enhance overall model performance by improving the distinction between different catchment-specific rainfall–runoff behaviors" (L288-290)

4) We can not get a conclusion that smaller batch sizes contributed to better predictive performance based on the only comparison between 256 and 2048. More batch sizes should be tested.

In our paper we compared only two batch sizes (small vs large), of which the smaller one showed better performance, hence the assumption. We reformulated the results to "the smaller batch size of 256 contributes to better model performance with regard to mean KGE values." (L322)

5) For GRU model, authors explained that the computational cost was increased with no static features due to the window size increasing from 87 to 298. How about window sizes in other models? Does it remains same in with/without static features?

This question relates to the previous question 2. All Window sizes are listed in table 6. To answer your question, each window size happens to be different. But for the other 2 model with a batch size of 256 the window size is smaller when static features are used.

6) It is incorrect to express "Across all models, meteorological features, particularly precipitation and average precipitation, consistently exhibited significant positive impacts on discharge prediction in lines 620." The temperature shows a negative relationship with discharge variation.

The soil temperature was grouped within the soil attribute class and not within meteorology, hence the sentence is correct within the context of our paper. However, since we shortend the sensitiviity analsysis with regard to the other reviewer, the issue is solved.

7) It is suggested to compare the predictive discharge from three models in one gauge station of one figure in the Appendix. It seems that predictive performance of three models are different in the same gauge. Does it relate to the spatial distribution of gauges?

We added hydrographs for 3 different gauges in Appendix A4. The following sentence was added " Appendix A4 presents a comparison of the simulated hydrographs for the same basin. Consistent performance trends are observed across all models, with either poor or high performance in the same basin. However, one plot exhibits mixed performance, where both LSTM and GRU models perform well, while the CNN model shows poor performance. Notably, this is the only validated catchment where such a strong discrepancy is observed."(L725-728)