Response to Review #3 comments concerning HESS submission:

A National Scale Hybrid Model for Enhanced Streamflow Estimation - Consolidating a Physically Based Hydrological Model with Long Short-term Memory Networks

Jun Liu, Julian Koch, Simon Stisen, Lars Troldborg, Raphael J. M. Schneider

Department of hydrology, Geological Survey of Denmark and Greenland, Copenhagen, 1350, Denmark

Correspondence to: Jun Liu (juliu@geus.dk)

We have prepared a plan to address each point systematically, and the manuscript will be updated accordingly in the future. The original comments are highlighted in blue, while our responses are in black. We hope the responses fulfil the requested changes required to proceed with the publication of the updated manuscript.

The paper aims to investigate the advantages of utilizing a distributed Process-based Model (PBM) in implementing an LSTM representation for streamflow prediction. The researchers tested various traditional combinations to analyze the pros and cons of each configuration. They concluded that LSTM with the output of the PBM as input (LSTM-q) and an LSTM model learning the residual error of PBM (LSTM-qr) were the best models. One of the interesting findings of the study is that the hybrid model requires less memory (sequence length) than a simple LSTM (LSTM-rr). This indicates that by using PBM in LSTM, it can incorporate longer temporal dependences, which mitigates one of the issues of LSTM representation. Another notable finding related to this is that LSTM decreases performance in groundwater-dominated catchments, as suggested by other studies.

Reply: We appreciate the reviewer for reading our manuscript and providing valuable comments. We have planned to revise the manuscript based on their comments.

However, I have some major comments about the lack of clarity in defining the criteria for the best model and the explanation of some figures. Although the authors defined several metrics to evaluate the model, it was not clear which one or what combination of them was used to define the best model. Additionally, in many cases, the differences in performance are so small that they are probably not statistically significant.

Reply: We agree that the involvement of a group of metrics for the evaluation complicates the conclusions to be drawn from the results. We will use NSE as the basic index to evaluate the performance of different models. NSE is a comprehensive metric to measure overall fit of a model and how well the model captures the variability of observations. We will provide the results of NSE in the main part of the manuscript and provide the rest of the metrics in the Appendix. According to NSE, LSTM-q and LSTM-qr are comparable to each other, but better than the others (see table 1.). In case the differences in performance are not statistically significant, we will carry out a two sample Kolmogorov-Smirnov test https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html, to see if the metric distributions of two models are significantly different or not.

Moreover, some figures were presented without any further explanation. For instance, Figure 3 shows 16 subplots, but the text only mentions two lines about it. It’s crucial to present figures that support the story presented in the paper, and if a figure isn’t explained, it should go to the appendix. However, the authors should try to analyze each figure as much as possible because they will find more details that support their findings.

Reply: We will retain the crucial subplots (NSE) in the figure and move the others (KGE, NSE2, NSE_log, FHV, etc.) to the appendix. Overall, we will make sure to address each presented figure with sufficient detail in the revised manuscript.

Minor comments.

Line 29: If you mention extra/interpolation, please explain why it is important for your goal.

Reply: We are going to add one more sentence after this line for the explanation. The potential sentence could be as follows:

'They are used to supplement the missing streamflow at stations, transfer the parameters to basins showing high hydrological similarities, and predict streamflow under future conditions.

Line 59-60: The statement could mislead readers to believe that only DL methods experience a decline in performance. Please modify it.

Reply: We agree that the sentences are not rigorous. We will modify the statement as follows, some references will also by cited.

While these models often demonstrate higher performance, accuracy may decrease when attempting to transfer them from gauged basins to ungauged ones. This issue is a common concern in the context of physical models as well.

Line 97: Please use a software or a method to verify references as I found an incorrect citation. It should be De la Fuente et al. (https://doi.org/10.5194/hess-2023-252) instead of Fuente et al.

Reply: We will thoroughly review the references.

Line 98: I agree with the sentence, but you should provide references, considering the "limited attention" given to the topic.

Reply: The references, initially placed in the middle of the sentence, will be moved to the end. Additionally, we will explore more recent references to ascertain their potential to support our statement.

Table 1: It would be very useful to add some summary statistics, such as range and mean.

Reply: This is a nice suggestion; we will include the ranges and means of these attributes.

Line 291. Why did you change the loss function? Different loss functions emphasize different components of the error.

Reply: NSE is a comprehensive metric for measuring the differences between simulated discharge and observed discharge. However, we don't consider it a suitable parameter for measuring the differences in residuals or error factor see Figure 2 This is why we use RMSE in the loss function for residual and error factor models.

Line 307. Line 307: The hyperparameter search could generate some inconsistencies because only 6.2% of the parameter space is being explored (100/1620). To address this, it is recommended to fix hyperparameters with low sensitivity such as LR, BC, NE, and DR. This way, a more detailed exploration of the hyperparameters that matter can be carried out.

Reply:

We defined the candidate values because we didn't have a good sense of the optimal values for the hyperparameters. We selected the values based on previous studies. However, we agree with this suggestion; some of the hyperparameters are less sensitive according to our results. This is a limitation of our study, and we will discuss it in the updated manuscript. To facilitate a discussion on the relationship between model performance and sensitive hyperparameters (i.e., Figure 9), we intend to conduct an additional hyperparameter search. We will use a limited set of values for the number of epochs (30) and hidden unit size (64, 128, 256), while fixing the dropout rate (0.3), learning rate (10^-3), batch size (128). and the sequence lengths (10, 30, 60, 90, 180, 270, 365, 730). The total number of hyperparameter combinations will be 24 (1*3*1*1*1*8) for each model. Subsequently, we will reproduce Figure 9.

Table 3. How much is the difference with the second best? It is a little strange that models with the simulated streamflow as input have more hidden cells than the baseline that is learning the entire dynamic of the system.

Reply: We will provide the range of NSE, including minimum, maximum, median, and mean values of the results in table 3. The models with more hidden cells not only simulate streamflow but also incorporate average water content in soil layers, actual evapotranspiration, and depth to the phreatic layer as dynamic inputs. Therefore, the LSTM model could be more complex and have more hidden cells.

Table 5. If the difference between the mean and median is not mentioned, it's recommended to either delete one of them or move the discussion to the appendix.

Reply: We will retain the mean NSE values in the updated manuscript and relocate the other metrics to the Appendix.

Line 348. LSTM-q outperforms LSTM-qr only for NSE2. Please, check your analyses.

Reply: We will revise the analysis of the sentence.

Line 353. In the text, it is mentioned that LSTM-rr has a lower NSE_log than DKM, but the table shows the opposite.

Reply: The argument behind this statement relies on the mean values of NSElog. The mean NSElog of LSTM-rr is 0.36, whereas the mean NSElog of DKM is 0.41 in the spatiotemporal split experiment. However, the median NSElog of the two models is contradictory, causing confusion in the statement. To address this, we plan to simplify Table 5 by including only the mean values of NSE, while the remaining metrics will be moved to the appendix. We will also rephrase the sentence in line 353.

Figure 3. You did not analyze the figure. Therefore, you should either delete or move it to the appendix. However, I believe that this figure is more informative than Table 5, so I encourage you to describe it.

Reply: We will keep NSE as the main metric in the revised manuscript, so Figure 3 and Table 5 will be kept short to only show the results of NSE. The results of the other metrics will be moved to the Appendix in the updated manuscript.

Figure 3. Cumulative distribution functions (CDFs) are typically displayed with metrics on the x-axis and probability on the y-axis. To improve clarity, it is suggested to limit the axis for KGE and NSE between [0,1]. This will provide a better visualization of the behavior of each line. Additionally, you may consider showing only some of the metrics in detail, while the rest could be placed in the appendix.

Reply: We will modify Figure 3 by following the suggestions and displaying the CDFs with metrics on the x-axis and probability on the y-axis. Considering that some of the models have negative KGE and NSE values, we initially limited the range to [-0.5, 1]. However, we will change these settings and restrict the range of NSE and KGE to [0, 1]. We will retain the first row of Figure 3, which only shows the results of NSE. The remaining rows of Figure 3 will be included in the appendix.

Line 360. A value greater than zero is still unsatisfactory. Please rephrase this sentence.

Reply: We will modify this sentence. According to Moriasi et al. 2007, NSE>0.5 indicates model performance is satisfactory, we will count the number of basins with NSE higher than 0.5.

Line 363. It can be difficult for someone from another country to identify specific areas. Adding a latitude-longitude grid can help with this issue.

Reply: We have the names mentioned in the manuscript in Figure 1(c), which may not be clear to see. To enhance clarity, we will add a latitude-longitude grid to Figure 1(c) and change the font size of the names.

Figure 4. If you are not going to describe the other maps, move them to the appendix and present only the histogram. The current color scheme makes it difficult to identify patterns. To address this, Figure a) could benefit from a traffic light palette (green for good, yellow for regular, and red for bad). Meanwhile, the other figures could use a palette that transitions from white in the center to either red or blue on the extremes, respectively. This approach will allow readers to better focus on relevant changes.

Reply: We will modify the figure accordingly. Our plan is to retain the results of DKM, LSTM-q, and LSTM-qr in the figure while relocating the others to the Appendix. We will also adjust the colors of the legend as suggested.

Figure 5. It is difficult to distinguish between the colours of LSTM-rr and LSTM-qf. Adding the KGE or NSE of each model to the legend would provide an additional comparison.

Reply: We will change the colour of LSTM-qf, for example, to brown. NSE will be appended to the end of the model name in the legend.

Line 392. Please provide references to support the fact that this finding is not surprising. As far as I know, lumped models tend to perform poorly in predicting larger areas when there is a non-uniform distribution of precipitation. However, larger catchments are usually influenced more by baseflow which should be easier to predict. Hence, to validate this conclusion, it would be helpful to compare it with other studies conducted in the region.

Reply: The DKM model performance report documents higher accuracy over large basins. We will provide the references in the revised manuscript. We also need to mention that most of our catchments are still quite small compared with larger rivers globally, and the spatial variations of precipitation is not significant and spatial aggregation is not that important.

Line 397. To gain a better understanding of the region, please add the range of groundwater levels.

Reply: We will provide the general ranges of groundwater depth in Table 1, and the corresponding sentence.

Line 402. I would like to draw your attention to the fact that the correlation analysis was conducted using only one variable. This means that any interaction between attributes has been overlooked. Additionally, the Pearson correlation only represents linear correlations, which underestimates more complex relationships. To address these issues, I suggest using the Spearman correlation and discarding low correlation values. Alternatively, you could build a random forest model to examine how the combination of attributes affects performance.

Reply: We will recalculate the correlation between model performance and basin attributes with Spearman correlation. However, we won't go deep to analysis the interaction between attributes and model performance. We agree that two or more variables coherently contribute to model performance, but that is not the main objective.

Figure 6. Please double-check the color bar and ensure that the white color is set to zero.

Reply: We will modify this figure, changing the colour bar to ensure that white represents zero in the updated manuscript.

Line 428. The sentence "chosen based on their superior performance" is not accurate as the performance of the LSTM models is similar. Additionally, the models were not statistically compared.

Reply: We selected the two LSTM models based on the mean NSE (0.63) from the spatiotemporal split experiment. Their mean NSE is slightly higher than that of the other models.

Figure 7. The figure contains too much information with the density function and histogram. It would be better to only show the density function and widen the figure.

Reply: We will remove the time series in the subplots and change the size of the figure to make it wide and fits the width of the paper.

Line 445-446. Could you describe the locations of these regions?”

Reply: We will carefully check the names and locations in the manuscript and make sure they are properly shown in the study area figure (Figure 1).

Line 448. Figure 7 exhibits an underestimation of the streamflow. Can you explain the source of this statement?

Reply: Figure 7 shows the streamflow for specific events. For instance, Figure 7a illustrates the average streamflow of all stations over a 30-day period starting from December 20, 2011. However, in line 448, we discuss the long-term changes in streamflow from 2011 to 2019.When compared with the DKM shown in Figure 8b, the two LSTM models demonstrate lower values in the west of Denmark but positive values in the east for high flow (Figure 8b, e, h). Both LSTM-q and LSTM-qr exhibit better performance than DKM, resulting in simulations that are closer to the observations. Consequently, when DKM indicates higher discharge values, there is a higher probability of overestimation. The expression in this sentence is misleading, and we will rewrite it in the updated manuscript.

Line 460. Could you please clarify on which results this conclusion is based?

Reply: we will rewrite the sentence and clarify the results:

The results revealed that utilizing LSTM models, especially the hybrid schemes that were coupled with physically based simulations, exhibited better performance for both overall performances spanning a decade (section 3.1) and specific hydrograph extreme events (section 3.2).

Figure 9. Why aren't the values from Table 5 included in this figure? Also, why is LSTM-rr showing lower values than the others?

Reply: Figure 9 is based on the evaluation results of the spatiotemporal split experiment, so it should include the values from Table 5. We will carefully check Figure 9. LSTM-rr has lower performance in the spatiotemporal split experiment compared to other hybrid models, which is one of the conclusions of the manuscript stating that hybrid models have improved accuracy compared to benchmarks, i.e., LSTM-rr and DKM.

Line 481. If there are not many studies that have made this comparison, you should include references to those studies.

Reply: We will add the references which used the spatiotemporal hold-out experiments in the updated manuscript.

Line 498. Can we attribute this improvement to the use of longer memory, as shown in Figure 9?

Reply: We can attribute the improvement of LSTM-rr to longer memory, which requires a longer sequence length to achieve better performance, as shown in Figure 9. However, the improvement in the hybrid model is attributed to the provision of more information about complex hydrological processes during the training of LSTM models. Hybrid LSTM models do not show an improvement with longer input sequences in Figure 9.

Line 500. You must explain your ranking process as you are using multiple metrics. Are you using one of them, their average, or some combination? Additionally, many of the models may not be statistically significant.

Reply: We plan to use mean NSE as the main metric to rank the models. According to the mean NSE values in Table 5, the order is LSTM-qr ≈ LSTM-q > LSTM-qf > LSTM-rr > LSTM-pf = DKM for the spatiotemporal split experiment. Additionally, the order is LSTM-qr > LSTM-q = LSTM-rr > LSTM-qf > LSTM-pf > DKM for the spatiotemporal split experiment. Considering that the performance of LSTM-q is higher than LSTM-rr in the spatiotemporal split experiment, we have concluded that LSTM-qr and LSTM-q are the two best models.

Line 525. Could you please clarify what you mean by shallow structures? It seems contradictory to say that 256 hidden cells in parallel with 365 days of sequence length is a shallow structure. Many studies have shown that using more than 2 layers in a series does not result in significant improvement. This suggests that a single layer has sufficient complexity to capture the necessary processes.

Reply: The term 'shallow structure' here implies that we only trained one type of LSTM model encompassed the entire study area and its complex hydrological processes. In contrast to the approach of training distinct LSTM models for each sub-region, a single LSTM model lacks versatility. We acknowledge that the phrase may be misleading, and as a result, we will remove the term 'shallow structure' from the sentence.

Line 526. I do not believe that the features and attributes are deeply hidden. The problem lies in the representation and inputs used to extract information.

Reply: We will modify the sentence,

Line 529-530. Instead of making your representation deeper to increase its complexity, you should try using a multi-representation approach. Different representations or architectures can capture different pieces of information. Using local models may alleviate the issue, but it does not solve it. This is like trying to approximate a high-degree polynomial by using only order 2 polynomial segments. Adding more order 2 polynomials segments does not increase the complexity; it only segments the extraction of information.

Reply: we are not very sure about the multi-representation approach. It will be great if the reviewer can explain more about it on our case Subsequently, we will make the necessary modifications to our manuscript.

Line 542. Some researchers have used graph neural networks with or without routing parameters in training. Mention them.

Reply: This comment is very informative and expands our knowledge in this area. We will explore graph neural networks and cite them properly here.

Line 559. Are these values significantly different from the LSTM-rr? for this reason, you must be more specific about what was your final multi-objective criteria.

Reply: We plan to use NSE to rank different models. The remaining indices will be included in the appendix for discussion.

Line 561-562. Why were LSTM-qr and LSTM-q not able to beat always to the DKM model despite using its outputs?

Reply: There are two reasons the LSTM model failed to improve streamflow estimation:

(1) The DKM model has been calibrated and demonstrates good performance in some basins (NSE > 0.8), indicating limited room for further improvement by the LSTM.

(2) We trained a single LSTM for all basins, integrating all inputs spatially, leading to the omission of some information. So, despite the complex internal structures of LSTM models, they remain catchment-lumped models, whereas the DKM is a fully distributed model.

Line 567-568. Do you have any suggestions or ideas?

Reply: We will propose suggestions to potentially enhance the prediction of extreme events. For instance, we can employ objective functions that emphasize extreme values, such as NSE^2, to train the LSTM model.

Line 570. It would be helpful if you could mention the complex hydrological processes.

Reply: We will rewrite the sentence and mention the complex hydrological processes in the updated version. For instance, *'Previous studies concentrated on gauged basins, whereas this study makes contributions to the topic through a national-scale analysis, highlighting areas characterized by intricate hydrological processes.'*