Response to Review 1 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)

The authors compare the performance of different types of hybrid models based on LSTM networks and a physically-based model, DKM, in estimating streamflow for Denmark. Generally, the hybrid models outperform LSTM rainfall-runoff model (LSTM-rr) in ungauged basins. They find the hybrid dynamic inputs LSTM model (LSTM-q) and the LSTM residual error model (LSTM-qr) have the overall best performance. The hybrid models also improve streamflow estimates in groundwater dependent basins. The study is interesting, and the authors provide a comprehensive discussion. My concerns on the study are as follows.

Reply: we thank the referee of reading the manuscript and providing helpful comments and suggestions. Here we reply to the comments point-by-point. The font color of the original comments is blue, while the font color of our replies is black. We hope that these changes satisfy the requirements for proceeding with the publication of the updated manuscript.

In this study, LSTM-q and LSTM-qr are the optimal streamflow models for Denmark. Will the models be optimal in other regions in the world? I am worried that this might be a local conclusion.

Reply: We hold the opinion that LSTM models, particularly those utilizing simulations from physically based hydrological models, exhibit better performance also in other regions. We discussed this point in the manuscript and referenced previous studies (see lines 65-71 and lines 509-511). Although the accuracy of hydrological models varies from region to region, we expect that their simulations will contribute to improving the performance of hybrid LSTM models when being benchmarked against hydrological simulations and LSTM rainfall-runoff models Information from physically based models will and should always provide additional insights to deep learning models.

The authors discuss it in the Discussion section (Line 504-512) by comparing their conclusion with that from Konapala et al., (2020). They argue that the different best hybrid models in the two studies is due to the higher accuracy of DKM. Is it possible that the difference is also due to the different study domains, i.e., Denmark in this study and CONUS in Konapala et al., (2020)? In general, CONUS has a much deeper groundwater table depth than Denmark. The authors may evaluate the performance of different hybrid models in various groundwater table depth ranges to check whether the conclusion will be changed.

Reply: The performance of LSTM-q and LSTM-qr is comparable with LSTM-qr being slightly better. The inputs of LSTM-q and LSTM-qr are the same, the only difference is the target variables. The target variable of LSTM-q is discharge and the target variable of LSTM-qr is the residual between simulated discharge and observed discharge. As stated in

Konapala et al., (2020), LSTM-q is more correlated to the rainfall-runoff LSTM (figure 1(c)), while LSTM-qr is more correlated to the PB. Our conclusion is that LSTM-qr outperforms LSTM-q because DKM is better than the PBM applied by Konapala et al., (2020).

We showed that the correlation between model performance with basin attributes (including groundwater table depth) and the correlation patterns of LSTM-qr and LSTM-q are very similar. We argue that the different conclusions of which hybrid setup is best between our study and Konapala et al., (2020) mainly relates to the diverging physical based model performances. It should be noted that other study, for example, Cho and Kim, (2022) used a well-calibrated model WRF-Hydro (NSE = 0.72 and R = 0.88) to predict residuals and they share our conclusion that the residual model performs better. We will present this point more clearly in the revised manuscript and will therefore revisit our presentation and discussion of the results.

The common practice is to separate the data into training, validation and testing sets in the time order. Why do the authors choose 2011-2019 as the validation period and 1990-1999 as the testing period? Does the selected testing period have fewer human impacts on streamflow?

Reply: DKM model was calibrated in 2000-2010 and tested in 1990-1999, so we kept this routine to develop the LSTM models. DKM considers human activities during testing, training, and validation period, for example, water extractions and wastewater flow modelling. We will explain the reasons of how we decided period for testing, training and validation in the updated manuscript.

In addition, Section 3.2 compares the event performance of LSTM hybrid models during the validation period (2011-2019). At least 80% of study data used in the section are validation data, which have been observed by the hybrid models during the hyperparameter tuning. The performance of the hybrid models is expected to be good, but may not deliver reliable information.

Reply: In section 3.2, the model was retrained with data from all stations and the training period is 1990-2010. We did not use validation data during the training period. To make this point clear, we plan to add some sentences of the setting in the updated manuscript, see lines 427-430.

In Table 1, why does the hybrid models include as input phreatic depths at both 100 and 500 m resolutions?

Reply: We have two versions of the DKM model distinguished by different resolutions: 100m and 500m. The finer resolution provides more details and a more accurate representation of the phreatic depth. However, at basin scale the high-resolution spatial patterns of phreatic depth are not relevant since they will be averaged across the entire basin. Therefore, we will remove the 100m phreatic depth in the updated manuscript to avoid complexity and just stick to the 500 m.

Please improve the quality of the figures particularly Figure 8. If possible, please also increase the fontsizes in the figures.

Reply: Figure 8 was compressed so the quality is low. We will provide the original figures in the future.

Specific comments:

Line 129-130: Might change “hidden unit sizes” to “hidden neurons”.

Reply: The phrase will be changed in the updated manuscript.

Section 2.5: “Table 3” in Line 314 and 330 should be “Table 4”.

Reply: the errors will be corrected in the updated manuscript.

Table 5: Maybe only write the best evaluation scores in bold for better visualization.

Reply: We marked the values with the best evaluation scores in bold.

Figure 8: The word “bias” seems not be a right word to use in the caption, which suggests the error. Please consider replacing it with the word like “difference”.

Reply: We will change the words in the caption of Figure 8. Here is a copy of the modified caption.