Response to Review #2 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)

This manuscript proposes the use of different combinations of long short-term memory (LSTM) with a physical model-Danish Water Resources Model (DKM) to improve the accuracy of streamflow prediction. It is suggested that the hybrid model improved the model accuracy in ungauged and gauged basins. The authors further pointed out that the hybrid models could enhance the accuracy of extreme events. The knowledge gap is convincing, and the paper is clear. However, I have some comments that should be addressed before publication.

Reply: We thank the referee for his/her time and effort in reviewing our manuscript. We have revised the manuscript and responded to the comments point-by-point accordingly. The font color of the original comments is blue, while the font color of our replies is automatically black.

Specific comments:

1) In the introduction, the authors claimed that the DKM model is a well-established groundwater modeling system (Line 100). However, Nevertheless, this paper lacks observational evidence, particularly in results to support this claim. References are suggested.

Reply: We cited the evaluation reports of the DKM in lines 41-45. Detailed evaluation reports of DKM performance can be found in Henriksen et al., 2021 and Stisen et al., 2020, which also has been cited in the manuscript. We will also provide more publications of applications based on the DKM in the updated version of the manuscript.

2) Sections 2.1 and 2.2.4 are overlapped and can be merged into a whole section.

Reply: We will modify the structure of section 2 and move sections 2.1 to sections 2.2.4.

3) A table is suggested to compare the inputs and outputs of each type of model in section 2.3.

Reply: We will provide a figure to show the inputs and outputs of each type of LSTM model in the updated version of the manuscript.

4) The caption of Figure 2 is not clear. Please explain each of the subplots.

Reply: The caption of Figure 2 will be modified in the updated manuscript, here is a copy.

5) In section 3.2, the operational forecasting framework uses only the observations of meteorological factors as model inputs. However, many studies try to combine the historical simulations or observations as model inputs which contributes to model forecasting. In the authors' cases, please comment on the impacts if considering the history series as model inputs.

Reply: In section 3.2, we evaluated the performance of LSTM-q and LSTM-qr in the operational forecasting framework, due to their higher performance. The inputs are climate forecasting and DKM simulations. Our plan for the operational forecasting framework is to use climate forecasting and DKM simulations as inputs to get a higher accuracy of streamflow. We are aware that some studies have promoted methods involving the use of LSTM with data integration to enhance streamflow forecasting (Feng et al. 2020). However, this topic is beyond the scope of the current manuscript, and it is part of our future work. We plan to include sentences discussing how we incorporate newly available observations into the operational framework in the updated manuscript, for example, after line 530.

6) In the discussion part, it is interesting to get the conclusion of model performance: LSTM-qr ≈ LSTM-q > LSTM-rr > LSTM-qf > LSTM-pf ≈ DKM. Could authors explain why the LSTM-rr performed better than the LSTM-pf, as the LSTM-pf is the pretraining and finetuning LSTM-rr?

Reply: Regarding the LSTM-pf model, it was pretrained based on all ID15 catchments, totalling 2830. The input data consist of climate forcing, and the target variable is DKM simulated discharge. The fine-tuning process took place at 276 gauged basins, employing the same climate forcings, but with the target variable being observed discharge. The main reason for the failure of the LSTM-pf model is primarily attributed to differences in data between the pretraining and fine-tuning phases. The DKM model has been calibrated in gauged basins, but its performance in ungauged basins remains unknown. Notably, basins with larger areas exhibit higher accuracy than smaller basins, but the majority of ID15 catchments is small basins and the representation of DKM simulations of real world is low. Despite using numerous basins for pretraining, significant data discrepancies persist between the pretraining and fine-tuning datasets. It can be hypothesized that the imbalance of basins used for pretraining and finetuning (2830 sim vs. 276 obs) is the reason why the LSTM pf model performs poorer than the LSTM rr model. Future work should address such trade-offs and the role of epochs in both training rounds, which will certainly affect the overall result. We will extend the discussion of these results in the revised manuscript.