Response to report #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)

Thanks for authors‘efforts on replying the comments and making revisions. My few comments on the study are as follows.

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

1) In figure 6, authors compared the predictions from DKM, LSTM-rr and LSTM-q in A and B basins. Authors pointed that the different performances of LSTM-rr and LSTM-q in two basins were caused by the basin attributes. However, I have a doubt because the LSTM-q model considered the additional DKM simulations as inputs compared to LSTM-rr model, and both models have taken the basin DEM and slope as static inputs. The DKM got a good performance in base flow prediction in basin B, so I wonder if the poor performance of LSTM-q model was caused by nonoptimal model optimization.

Reply: The two stations shown in Figure 6 were not used for training, they are two stations from spatiotemporal split-sample experiments. The estimation of discharge at such two stations is estimated from their inputs and attributes. If the inputs and attributes are extreme cases among the training dataset, the LSTM prediction may give a low accuracy.

We agree with the argument that LSTM-q is nonoptimal. An optimal LSTM-q could figure out how to trust DKM simulated discharge. For example, ideally, the LSTM model follows DKM simulated discharge in the basins where DKM has good performance, but LSTM simulates discharge more based on precipitation and other variables in basins where DKM has poor performance. However, it seems the trained LSTM-q model cannot figure it out, and the simulations are poorer compared with DKM simulations in a few basins. There are many reasons (and we don’t think it solely can be linked to model optimization), such as the inputs are insufficient and the representativeness is low, the model design is imperfect. We have discussed this point in lines: 539-545.

2) In figure 7, I am curious of the correlations between different basin attributes and models. If authors select the positive correlated parameters as model inputs instead of all attributes, does it improve model predictive performance?.

Reply: Figure 7 shows linear correlation between catchment attributes and model performance. A negative value indicates the model simulates streamflow less accurately with rising attribute value, but the attribute likely remains important. For example, with deeper shallow groundwater level, the models generally struggle to accurately simulate streamflow.

Despite some low correlations in this simple linear correlation analysis, the respective covariates still might have informational value for the more complex LSTM model. Also, the hyperparameter search allowed for complex enough architectures to be able to ingest a row of static attributes, not having to limit ourselves to very few.

3) In figure 8 caption, there is no LSTM-qr shown in the figure?

Reply: We did not show the results of LSTM-qr, we have corrected the caption of figure 8, and we removed ‘LSTM-qr’ from the caption.