General comments

Feng et al tested the performance of a hybrid model in simulating daily streamflow at several thousand global watersheds. They benchmarked the performance of the hybrid model with a deep learning model and found the hybrid model has a satisfied performance. The topic is a good fit to the scope of Geoscientific Model Development, though I agree with previous review that this study uses the same model and design of experiments from the authors' previous study. The new insight of this work is to test the hybrid model framework in more watersheds locate from different continents. The idea of hybrid model is great and represents a significant contribution to hydrological modeling. It would be helpful for the authors to discuss the challenges of coupling the ML/DL model with a hydrological model. Specifically, since HBV model is relatively simple, is it possible to develop such a hybrid model with a more complicated hydrological model? Please also see my specific comments in the following.

Specific comments

As the author argued in Line 47 – Line 55 that DL and LTSM have been demonstrated with good performance of simulating hydrological variables from local to continental scales, I wonder what is the novelty of this work? Although this study extends previous application at CONUS to global scales, the selected dataset doesn't have a good coverage for the whole globe. Please find my detailed comments regarding the selection of dataset in the following.

Line 124: Why the authors select this dataset given there exist other global streamflow datasets that contain much more gauges and have a better spatial coverage? Since the selected dataset archives headwater catchments, the observed streamflow is approximately same as the runoff generation, with the less river routing impacts. Is this because there is not river routing component in HBV? If runoff is the target variable, there are multiple well validated global runoff datasets to be used. Then it is not reasonable for the author to justify that the application is at global scales. In addition, Figure 1 shows the selected catchments are mainly from certain regions. Except North America, the gauges over other continents do not cover the continent uniformly, thus they are not representative for the continent.

Line 95 – Line 97: There are several studies that calibrated the models to match monthly variations.

Section 2.3: In traditional hydrological model calibration, one needs to run the physical model many times with perturbed parameters. How many forward simulations are needed in the hybrid model to identify the best parameter? Please clarify.

Line 202 – Line 218: I agree with the authors that PUR experiment is more challenging spatial extrapolation than PUB experiment. However, in practice, PUB experiment is more useful than PUR. Streamflow is probably the most well observed hydrological variables. At global scales, there exists abundant streamflow gauges with a good spatial coverage for each continent, though some continents have relative less than others. Therefore, one doesn't need to assume a whole continent is ungauged. It is the selected dataset in this study that gives sparse spatial coverage.

Line 236: Do you mean a median KGE of 0.78?

Section 3.2: I expect dPL + evolved HBV with DP is always better than dPL + evolved HBV, because the former model is more flexible to capture the observation. If no better, it should not be worse than the latter one. But Figure 5 shows, dPL + evolved HBV with DP is worse than dPL + evolved HBV in arid region. It will be helpful for the authors to clarify such clarification.

Line 341: I don't think the median KGE = 0.58 is significantly better than median KGE = 0.52. They are pretty close performance to me. I suggest the authors to plot the cumulative density functions (CDFs) of the KGE for different model, and test if the CDFs are statistically different.

Line 350 – Line 351: Do you mean "cannot be obtained by straightforwardly training *physical* or *DL* models on data alone"? Figure 6 suggests the hybrid model is better than the traditional model calibration (Beck20). But it doesn't support that performance of hybrid model is statistically better than a purely data-driven model.

Line 371: Why not using KGE and NSE for the ET evaluation to be consistent with discharge evaluation?