## Review of "Ensembling Differentiable Process-based and Data-driven Models with Meteorological Forcing Datasets to Advance Streamflow Simulation"

### General comments

This paper comprehensible evaluates the performance of different ensembles based on LSTM and δHBV models, and three different forcing datasets, across CAMELS catchments. The ensembles are evaluated in terms of a temporal test, a prediction in ungauged basins test (PUB), and a prediction in ungauged regions test (PUR). The main conclusion is that the data-driven LSTM and process-based δHBV ensemble improves NSE, particularly for PUB and PUR tests.

Overall, the manuscript is well-structured and clearly conveys its main point. Please find below some comments and suggestions.

### Specific comments

1. L150 and L240-L244: Please explicitly indicate which features (static and dynamic) are used by the LSTM model or at least refer to Appendix C. Are static characteristics of the catchment also used during the PUB and PUR tests? Is it the case that for PUB the model does not use previous streamflow observations to generate the predictions? Or does PUB only refer to the model being tested at basins not used during training?

2. Table 2: Is it possible to draw any conclusions about the skill of the models to extrapolate to a warmer climate based on the temporal test? I assume that the period 1995-2010 is warmer than 1980-1995.

3. Table 2: What would happen if you were to train δHBV using only the same 531-basin subset as for the LSTM instead of the 671 basins?

4. It could be useful to also provide similar plots to Fig. 3 in the appendix where $\delta HBV^2$ and $\delta HBV^3$ are used instead of $\delta HBV^1$.

5. Fig. 4 and L439-L454: Can you expand on potential reasons for the lower model-skill in midwestern and western basins? Is human management of streamflow an important factor here, despite being CAMELS basins?

6. Fig. 5: Please clarify. L399-L401 says there is a small difference when using 3 or 10 seeds, but Fig. 5 shows the difference between individual seeds and using 10 seeds. It is interesting to note in Fig. 5 that $LSTM^{multi}$ with 10 seeds achieves a similar skill as $(LSTM + \delta HBV)^{123}$ at least for the temporal test. This could also be an important conclusion.

7. Fig. 6: Suggestion to have LSTM models with shades of one color and δHBV models with shades of a different color to better highlight the differences between LSTM and δHBV mentioned in L461-L464.

8. L467-L469: Suggestion to extend the sentence to clarify what is meant here.

### Minor comments and technical corrections

1. L13: Replace “while” for “however”.

2. L20: Suggest deleting “utilized in two ways”. Here it just raises the question which two ways? Also in L526 it would be good to explicitly mention the “two ways”.

3. L177-L179: Are all modifications to δHBV1.0 of similar importance? Or can it be said which of them are more important?

4. Table 2: Isn’t there more recent data for PUB and PUR? Why are they trained only until 1999?

5. Figure B1. Xlabel on right panels should be temperature (C), correct?