Comments by TN in *blue*, our responses in black.

The manuscript was well written and the topic is very interesting. I really enjoy reading this manuscript. The discussion is thorough, providing a deep understanding of LSTM and good practice. I think there is still room for discussion (related to the questions below). Please feel free to include it if the authors think these points are relevant.

1. *Why is the performance of physical or conceptual models (e.g., mHM and VIC) when calibrated for a single basin ("basin model") better than the one calibrated for multiple basins ("regional model") while that is the opposite with LSTM?*
2. *"Never train an LSTM on a single basin". What makes LSTM so special that we should not train on a single basin? What if we want to model new processes with LSTM and we just have data for a single (or just a few) basin(s)? – we do not have other publicly available data (as for streamflow modeling).*

Response: Related to the first question, the reason for this particular behavior of conceptual models is more-or-less well known (it is related to the "uniqueness of place" problem in hydrology). ML is different because it can learn more general relationships (generally, what this reviewer describes in their comments below), and ML is almost always better (across almost all domains of application) when trained on more data. We will add some of that background to the introduction of the paper.

Regarding the second point, we would like to note that we don't see the LSTM as the swiss-army-knife that should be blindly applied to any problem. One always needs to carefully assess if this specific model is appropriate for the task (and data) at hand. That being said, there are various options that one could think about how DL models can still be used for a task with limited data availability. Probably the most obvious approach to us would be multi-task learning. That is, maybe there is another target with abundant data that is connected to the other task (e.g. discharge and soil moisture). In that case, one could train a model on both tasks together, hoping that learning about task 1 (with abundant data) helps the model to better understand task 2 (with limited data). That all being said, we see this discussion as somewhat off-topic for what we discuss in this manuscript, but hope that our answer helps the reviewer to explore this area.

Blow are possible discussions regarding these questions (just my opinion, please correct me if I am wrong)

Discussion of question 1:

Physically-based (or conceptual) model formulation: The physical processes are conceptualized using mathematical /empirical equations that were based on experimental results or the modeler's understanding of reality. In this sense, PB models already used some kinds of data (because the equations in PB models are based on experimental results

or the modeler's understanding of reality – so equations in PB models are also a kind of data).

PB models: "basin" and "regional" models are often calibrated for a few dozen (or a few hundreds) calibration parameters (I am not sure how many calibration parameters there are for the VIC and mHM that used in this study?). There are uncertainties in the data, model structure, and imprecise description of the physical process via mathematical equations in PB models, the calibration process will adjust calibraiton parameters to compensate for these errors/uncertainties. However, when applying to a large number of basins, the number of parameters might not be large enough to compensate for these errors/uncertainty => leading to a reduce model performance of the "regional" model compared to "basin" model.

ML models: "basin" and "regional" both have a high number of "trainable" parameters (similar to the term calibration parameters in PB) compared to PB models. For example, a basin (regional) model with 16 (256) hidden states, 1 layer of LSTM, and the model head consisting of 1 dense layer could have more than 1,400 (250,000) "trainable" parameters (these numbers were roughly calculated for a native LSTM network in PyTorch). With such models, it can easily overfit the train data and provide poor performance for test data. While the "regional" model with more data the overfitting problem might be avoided (but still possible) and provide a better performance for test data compared to "basin" model. This could be a reason why the performance of the "regional" model is better than the "basin" model with the LSTM network.

Now if we calibrate "basin" and "regional" PB model by adjusting parameters for individual model grid cells (or many groups of grid cells based on certain criteria), which results in a much larger number of calibration parameters – up to the level that the number of calibrated parameters of PB models and the number of trainable parameters of LSTM are the same, will PB behave similar to ML model? (I mean will the performance of the "basin" model is worse than the "regional" model?).

Response: Generally speaking, spatially distributed, grid-based hydrology models are worse than conceptual hydrology models. Our perspective is that hydrological processes are simply not understood well enough to be modeled effectively in explicit ways. There is quite a lot of evidence in hydrology literature that this is the case. An example are the Biosphere-2 hillslope experiments, which measured soil properties in a dense grid along a small (artificial) hillslope, and even with that level of information process-based models could not model saturated flow with meaningful accuracy.

Discussion of question 2:

What makes LSTM so special that we should not train on a single basin? Assume that we have one basin and train the model for daily streamflow simulation for 10 years. In this case, we have 3650 pairs of inputs and outputs– this is already "big data" compared to

other models that use a much lower number data points (e.g., Table 1 – in Zhu et al., 2022; https://doi.org/10.1016/j.eehl.2022.06.001). I would be curious to know if this could be because the number of trainable parameters of the LSTM networks is high or something related to the long-short term memory (that we need more data to "warm-up" the model?).

What if we want to model new processes with LSTM and we just have data for a single (or very few) basin? Could we impose our understanding of the process of interest on the structure of the LSTM model or somehow make LSTM applicable even for a single basin? For example in Figure 4 (page 5), if we know that there are seasonal and trend components, can we model these two components separately and then combine them which might improve the model prediction, etc.. I would very helful it if the authors could give some comments on this.

Response: Please see our response above on our opinion regarding the second question.

Thank you

Tam