This study presents a simulation using a differentiable model at 3753 basins globally. It represents an advance in the field and is thus worthy of being published somewhere. However, the used model and the design of the experiments are almost the same as those published in the authors' previous papers. Readers of GMD would expect more progress in those aspects. Thus, I suggest a major revision.

Dear reviewer,

We thank you for your comments. We would like to note, although the model has not been changed, the dataset and training changed drastically. It expanded from data from only the USA to the whole globe, and this is no easy feat. Enormous effort went into cleaning the data, improving model efficiency, training and testing, and interpreting the results. It was highly uncertain if the model qualities would carry to such a large dataset. Many of the insights are new. Plus, we do want to ensure some consistency between the evaluation across datasets so we understand how the same model behaves on different datasets.

We will revise the introduction to better highlight the questions.

GMD does have model evaluation paper type and there are many papers on GMD which evaluate the same model under different scenarios. In fact, we can probably say that it is a minority case to be changing the model structure significantly in each paper. That is why we think the paper fits GMD and will be a valuable contribution.

Major comments

- The authors should introduce more details of the experimental design, such as the metrics used in the training, how many experiments (temporal generalization, PUB, and PUR; correct me if I am wrong), and the purpose of the experiments. Some details may have been presented somewhere else. I find this manuscript difficult to follow without reading the authors' previous publications.

Thanks for the comment. We will revise the manuscript to give some more background on earlier work, and add details like metrics, experimental design, etc.

- L124: what are the criteria for the selection? Can you describe the erroneous cases? Do the erroneous cases include the data processing error described in L350?

There are some data records that just seem totally impossible, for example, total discharge that are way larger than precipitation, yet unable to be linked to some simple unit conversation errors. We will clarify in the revised manuscript that, yes, L124 and L350 refer to the same group of gages.

- L306-L317: Can you discuss more about comparing the traditional regionalization method and PUR? The PUR regionalization method can utilize a large number of observations to

calibrate/train the differentiable model, whereas the traditional method can only use very limited samples. In other words, PUR may have a much higher chance of finding the optimal parameters than the traditional method.

We will add to the discussion to clarify this. It has been established in the past work that differentiable has a better performance than MPR (Tsai et al., 2021) and another regionalization scheme. MPR actually has a similar concept as dPL, but it only uses linear functions because the model is not differentiable and they cannot optimize a large number of weights (as in neural networks) like the differentiable models can.

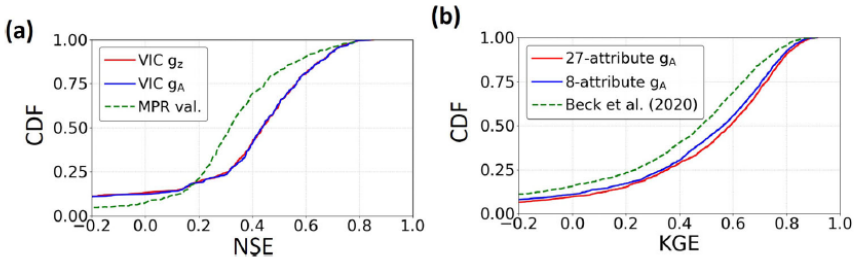
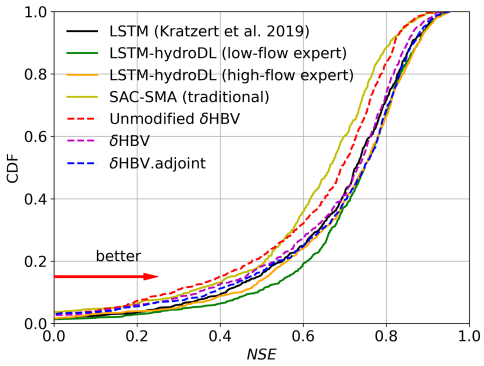



**Fig. 6 Comparison of dPL and regionalization schemes for streamflow calibration. a** Calibrating the VIC hydrologic model via a differentiable surrogate model on the CAMELS dataset over the conterminous United States, in comparison to Multiscale Parameter Regionalization (MPR). **b** Calibrating the HBV hydrologic model (not a surrogate) on the Beck20 global dataset, in comparison to the Beck20 regionalization scheme. We used NSE for the CAMELS case and KGE for the global case because these metrics were used by the respective papers, and the main purpose of the case studies was to compare with the existing literature. For both panels, curves on the right represent better models.

The above figure is from Tsai et al., 2021. Compare that with our model today:




(green line is the model benchmarked in this paper). You can see that we have made significant progress, so while we cannot run MPR directly, it would be logical to derive that the current version of the differentiable HBV is a lot stronger than MPR.

The biggest power of differentiable model is that it can learn robust relationships from big data. Hence what the reviewer said is true, but we would modify it as "*differentiable models may have a much higher chance of finding the most high-performing, robust and transferable parameters than the traditional method.*"

Minor comments

- Title: the phrase, global hydrologic simulations, is misleading. The simulations are conducted at 3753 basins across the globe. It represents a concept different from the "global hydrologic simulations."
Will revised to “Deep dive into hydrologic simulations at global scale”

- L127: is the classification from Beck et al., 2020b?
We will clarify this.

- L26 & L212: How is the subset of 1675 basins selected? What is the objective of the selection?
The selection criteria was long-term data availability. We will clarify this.

- L223: can you describe more about the structural issues? Why does the explicit solution scheme introduce numerical errors here?
Here we mentioned numerical error introduced by the explicit scheme. There are also issues related to how groundwater storage cannot feedback to the upper layers, which will be mentioned upon revision. However, we do not know all the issues (otherwise we would have fixed them).

- L291-L293: please rewrite the sentence. It is difficult to read.
Will revise to “*The sensitivity of model performance to missing processes is both good and bad news. It is good news because this means we can identify better processes from data...*”

- L398, "the underrepresentation of the processes...": this conclusion is too general. The difficulty of representing arid/polar/anthropogenic processes is known before reading this paper. The conclusion should be specific.

Well. There are many structural deficiencies with HBV, but it seems the parameterization approach can make up for some of the deficiencies. Rather here we are saying these two cases even the NN-based parameterization cannot make up for these missing processes. We will clarify upon revision.