Author´s response

Dear Dr. Viviroli,

Please find in attached with this document the review version of our manuscript, which incorporates the changes proposed by Grey Nearing.

Major Changes:

1. With respect to the comment by referee#2 Grey Nearing:

I don't agree with the reasoning for performing the experiment on a subset of basins.

The reason that the authors state is that they expect conceptual models to not be able to model human-influenced catchments. If there is this (or another) limitation of one type of model in the study, then it seems that this limitation is part of any meaningful comparison. Instead, the approach I would take would be to do the full experiment (on the whole benchmark dataset), and then – if there is a conceptual data split that makes sense – report analyses of results on the full data set and also on that split that you want to highlight.

Even if you restrict your analysis to "near-natural" basins, – which, to reiterate, I think is somewhat artificial – you probably should train the LSTM models on the full dataset. 60 catchments is probably not enough for training (see Kratzert et al 2024).

Additionally, there is an unfortunate consequence of only using a subset of catchments in that, moving forward, if someone wants to benchmark against or build on your results, they only have a subset of the community benchmark to work with.

We retrained all our models to consider the whole benchmark dataset (CAMELS-GB). Consequently, the stand-alone LSTM, LSTM+SHM, LSTM+Bucket, LSTM+NonSense, stand-alone SHM, stand-alone Bucket and stand-alone NonSense are now trained in all 669 catchments of CAMELS-GB.

We included section 3.4 in which we tested the models in the original subset of 60 basins, justifying the reason of the subset and its importance there. However, the outcomes reported in the other sections correspond to the results obtained for the whole dataset (669 catchments).

2. We moved explanation about the SHM model and about the training details for the hybrid model to the appendix section. This is intended to make the article easier to read, by moving important but not essential information to the appendix section. The information included in the appendix was already contained in the original manuscript, we only modified the location.

Minor changes:

1. With respect to the comment by referee#2 Grey Nearing

Line 265: It seems like the experiments in this paper could/should be run with ensembles. We do not know whether the conceptual models and hybrid models benefit the same way from ensembles as the pure ML models.

Using ensembles in hybrid models is indeed an interesting idea. There are multiple ways in which one can do ensembles. For example, different initializations (as done with LSTM) or using multiple conceptual models in parallel and weigh their output. We are currently investigating this effect for future research; however, we including ensembles is out of the scope for this paper and do not align directly with the main points we want to make.

2. With respect to the comment by referee#2 Grey Nearing

Line 290: Hoge (2022) and Kraft (2022) do not show that fusing deep learning models with hydrological mechanistic models can reach state-of-the-art. The reason for this is that they did not test against current state-of–the-art, and instead tested against handicapped LSTM models that perform significantly worse than current state-of-the-art. As an example, Hoge (2022) used LSTM values from a different study (Jiang, 2022), but they did not use even the best-performing LSTM from that paper, let alone the current SOTA LSTM (which is not from Jiang). Kraftonly tested against physically based models, not ML models, and there is no physically-based hydrology model that is anywhere close to the current SOTA. These papers should not be referenced in the way that they are here.

We modified the references accordingly.

3. With respect to the comment by referee#2 Grey Nearing

Line 293: I am not sure that Mendoza et al discussed dynamic vs. fixed parameters. When they talk about fixed parameters, they mean parameters that are hard-coded in the source code and therefore cannot be calibrated. This is not the same thing as parameters that vary during time series prediction.

We modified the references accordingly.

Even with these changes, the main points of the paper did not change. We believe the modifications made cover the changes proposed by the referee. We would like to thank both referees, one who remained anonymous and Grey Nearing, as their input in the review process allowed us to produce a better manuscript.

Kind regards,
Eduardo Acuña on behalf of the co-authors.