Review:

Manuscript: To Bucket or not to Bucket? Analyzing the performance and interpretability of hybrid hydrological models with dynamic parameterization.

General Comments

The authors introduce and analyse a hybrid hydrological model consisting of a conceptual hydrological model and a LSTM data-driven model to estimate time dependent model parameter dependent on the same inputs as used to drive the conceptual model. The intension is to keep the excellent performance of data driven approaches that have been demonstrated in recent years, but also to keep or improve the interpretability of such data driven approaches.

In general, I am in favour of an intensive analysis of such approaches, and think the manuscript is well suited for the readership of HESS, in continuation of a significant number of important papers in this area in the same journal.

It is in general well written and figures support the understanding and flow of the text! However, I have a number of major and minor comments/suggestion that I believe would improve the manuscript and should be addressed before final publication.

- The authors motivate they work by a paper of Feng et al. who propose a general framework of hybrid dPL modelling. They use the HBV model as a basis and estimate static and dynamically HBV parameters using Catchment parameters and meteorological input (as used do force HBV). This paper extends and slightly varies the this approach by analysing simple bucket based models as well as (what they call) NonSense model. Dynamic parameters are estimated with an LSTM DL. Research question 1 is "do conceptual models serve as a regionalization mechanism for thwe dynamic parameterization? I do think this is an important question (and I miss the reference of Frame et al, 2022 in this context), however, I believe it is not addressed in such a rigorous way as would be needed here. Conceptual models can range over a large range of complexity. Wha,t if we would just apply a simple equation relating Rainfall to runoff (Q = c(x,t) * P) and allow c to be estimated by a LSTM as suggested. This is the simplest model I can think of, and then I would systematically increase the complexity of the conceptual models.
  (Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall–runoff predictions of extreme events, Hydrol. Earth Syst. Sci., 26, 3377–3392, https://doi.org/10.5194/hess-26-3377-2022, 2022.)
  In that procedure I would suggest to use a much wider set of catchments and characteristics in order to see under what physio-geographical properties and climate conditions (as has been used of plenty other previous application) to answer research question 1 in a more general way!
- Research question 2 addresses the physical interpretability of conceptual models and whether it is comprised by data driven dynamic parameterization. Fig. 8 shows some of the parameters for 2 catchments and how they vary in time. I am missing a few points that should be discussed: i) Are the variations of parameters du to structural imitations of the conceptual model component, or is it just needed because of averaging non-linear processes over spatial variable catchment characteristics, or is it compensating for biases in the ERA5

input data? Or all three? What do I learn from Fig. 8? Which weight is assigned to each individual input for driving the variation? ii) How does the methodology compare to “more classical/statistical approaches” such as state and time dependent parameter estimation techniques. iii) How does the methodology compare in philosophy and potential to approaches that have been introduced by e.g. Feigl et al. (2022), what do we learn here in this approach from mistakes?
(Feigl et al., 2022, Learning from mistakes-Assessing the performance and uncertainty in process-based models. Hydrological Processes 36).

- Overall, I miss a kind of “surprise” concerning the analysis – could that be more emphazided.

Specific/technical Comments

The following minor comments/suggestions I would like to make:

- L9ff: The last part of the abstract is hard to understand/follow – I read it before the rest of text and did not know what is meant.
- L20: Reference needed.
- L136: how is ETp calculated (may one short sentence)
- L161: how you calculate the gradiants for if/then and iterative loops with state updates?
- L214: is 855 batches true hen you consider tat one data point consideres 180 previous days as input?
- L216: Why not optimizing the initial conditions?
- L232: this refers to one major comment – when is the model complex enough so that the LSTM is able to produce the full output space just by varying parameters!? Is this already possible with the structure I suggested). When an I see limitations/restictions?
- L265: what is the criterium for overfitting! Have you used ensembles of optimized networks to see how robust results are?
- Fig. 6: it is hard to see any differences, perhaps you can enlarge an interesting part of the time seies!
- L309: I would guess that ERA5-Land data are also computed and not observed quantities. So it is a model state intercomparison!
- L329: why looking at average values and not show the distribution?
- L385: what has this paper contributed to a better understanding in this context! Be specific!
- L402: What s new compared to Feng et al., what are different findings!
- L417: States (instead of variables!?
- L421: correlation is a very weak measure-of -goodness-of-fit especially when dealing with cyclic data/processes)

Overall, I feel, the manuscript has in general the potential to be a valuable contribution to HESS, however, questions and issues raised in the general comments would need to be addressed and discussed to a significant part before final acceptance.