Review of HESS Manuscript Titled “To Bucket or not to Bucket? Analyzing the performance and interpretability of hybrid hydrological models with dynamic parameterization”

Discussion Paper: https://doi.org/10.5194/egusphere-2023-1980

Reviewer: Grey Nearing

Summary of Paper

The paper addresses the following two questions:

1. Do conceptual models serve as an effective regularization mechanism for the dynamic parameterization of LSTMs?
2. Does the data-driven dynamic parameterization compromise the physical interpretability of the conceptual model?

Neither question is answered directly in the paper, but as far as I can tell, the authors intend to convey that the answer to both questions is “No”.

Additionally, the authors conclude that while it does not help increase model skill to add a bucket model to an LSTM, doing so does allow the model to predict variables other than the training target (which here is streamflow).

Summary of Review

In general I think these are interesting and informative experiments. Also, I want to express how refreshing it is to review an ML-based hydrology paper that uses best practices. Specifically, the authors are careful to train the models in a way that matches prior publications, and they build on existing community benchmarks (although some improvement on using the benchmarks more carefully would help).

I have two questions/concerns with the conclusions:

1) The lesson that I take from the quantitative results in this paper is that there is no value in adding bucket models to LSTMs. In other words, there appears to be a clear and direct answer to the question in the title. Bucket models do not increase skill, and there are no examples in this paper of something that the hybrid modeling structure can predict that the LSTM alone cannot. It appears that the bucket model (unless intentionally configured to be nonsense) is a transparent “head layer” in the deep learning model stack that contributes no information.

   I suspect that the authors might respond to my interpretation of their results by saying that the bucket models allow for estimating non-target states and fluxes. However, what is conspicuously missing from this paper is a quantitative benchmark against the results from Lees er al (2022), who show that the LSTM can estimate (unobserved) soil moisture and snowpack (the two non-target variables explored here).

   I’m searching for anything in this paper that provides evidence that there is any value to using bucket models. Otherwise, I suggest that the authors answer the titular question directly.

2) I’m not 100% sure that I understand the reasoning behind concluding that the bucket models cannot regularize the LSTM. The 0.71 median NSE from the LSTM+NonSense model seems to indicate that this regularization is possible, in principle. Interpreting the bucket models as head layers on the LSTM, this experiment seems to indicate that a really bad head layer can result in information loss (which makes sense to me).

   I am not 100% sure what is going on in these experiments, but here is one hypothesis. Bucket models (I would prefer to think of them as head layers, because the lesson generalizes beyond bucket models) can eat information (presumably by poorly regularizing the loss response), but “reasonable” bucket models used as head layers don’t add any information. Whether or not the "reasonable" bucket models are regularizing the loss response surface is unknown, since it appears that the LSTM is finding a similar solution either way. However, since the LSTM is finding similar solutions either way – there is no useful regularization from these head layers, and also no information loss.

   I would suggest that there are probably ways that we could look more closely at the functionals that the LSTM and hybrid models discover. Are they really finding similar relationships? Sensitivity analyses, input/output response surfaces, integrated gradients, etc. I’m not sure what the best approach would be (one of many would probably work), but I think that if you wanted, you could be more rigorous about understanding how these head layers are interacting with the LSTM component of the model to create functional mappings. I don’t really care whether the authors do that for this paper or not.

Specific Comments

Line 155: “...selected the one that performed the best for each basin.” Was this selection done with train-period, evaluation-period, or test-period data?

Line 215: I’m not sure that I understand (or agree with) the distinction being made in this paragraph. The LSTM has a state vector, just like the SHM – there is no difference here. The LSTM requires consecutive predictions in exactly the same way as SHM – again, there is no difference. We use a sequence-to-one training procedure when training the LSTM because this provides more diversity in the minibatch and accelerates training. You could do exactly the same thing for training SHM and/or the hybrid model. You are using a 2-year period for the hybrid model, and you could accomplish the same thing by using a 2-year sequence length for the hybrid model while still training sequence-to-one. You can think about the LSTM as having a 180-day spinup while the hybrid model has a 365 day spinup, and the LSTM is trained seq2one while the hybrid model is trained seq2seq. It’s fine if you’ve found that training the hybrid model using a sequence-to-sequence approach is better, but the way that is motivated (and the way this distinction is framed) in this paragraph is not correct.

Line 230: Do these median statistics come from ensembles (as is used by Kratzert et al), or from single LSTM/Hybrid models?

Line 235: I think it is too strong to claim that the “consistency and proximity” with Lees “validate the reliability of findings.” This isn’t a rigorous way to build on a community benchmark. A better approach is to actually recreate the existing benchmark exactly, which demonstrates that the models are built and trained correctly, then transition the experiment to the basins / time periods that you want to use for this study. This is, for example, what we did in the Frame et al. papers that required different training/test sets than what were used by previous community benchmarks.

Line 240: “ The LSTM network has the capability to account for biases in the forcing variables (e.g. precipitation or evapotranspiration) because mass conservation is not enforced.” We have a paper that demonstrates this explicitly:

> Frame, Jonathan M., et al. "On strictly enforced mass conservation constraints for modelling the Rainfall‐Runoff process." *Hydrological Processes* 37.3 (2023): e14847.

Line 270: What does the word “overwrite” mean in this sentence? I think that being very specific here about what you want to test is important because how the bucket model influences the LSTM is the central question. There is no such thing as “overwriting”. I suggest being crystal clear and precise about what you are envisioning when you ask this question.

Line 300: I agree that the results from section 3.2 don’t indicate that a conceptual model can’t be used, but again, I want *some* type of evidence that this has value. I don’t care about subjective arguments or what members of “the community” might think (e.g., line 35), I want some type of real, quantitative, scientific evidence that this is a useful thing to do. Showing strong correlation with soil moisture is interesting, but isn’t anything new for these ML-based rainfall-runoff models.

Line 315: "...suggesting that our model effectively utilizes the well-structured conceptual part to get better predictions of the untrained variables." I disagree – this is one possible explanation for these results, but there is another possibility (that SHM is simply doing nothing), and these experiments can't differentiate between these two possibilities.. These results show that the poorly-structured head layers cause the LSTM to lose information about soil moisture, but there is no indication that the "well-strucutred" model is being used in any way. The LSTM could be (and I suspect is) providing all of the information about soil moisture here, and the SHM model is doing nothing. A way to test this would be to use the Lees methodology and see what the LSTM can do alone to predict soil moisture. You probably will need to check both the LSTM standalone and the LSTM component of the hybrid model, since the method by which SHM would add value would be through regularization during training, and it is likely (although not guaranteed by the data processing theorem, due to the meteorological inputs to SHM) that the LSTM in the hybrid model will contain all or most of the information about soil moisture that is present in that whole model.

Line 345: The su_max and beta parameters being higher *might* result in lower water availability for down-stream buckets, but really what they are doing is reducing the outflow *rates*. More total water could compensate for this. Is all that is happening here is that the model is pulling a lever to reduce the total output, or is it more complicated – does the rate need to reduce for some reason. What do total inputs (P - ET) look like over these seasonal cycles? If P - ET is lower for the low-flow seasons, why is it necessary that the flow rates from the unsaturated zone also must be lower? Anyway, it's a little simplistic to just say that higher parameter values result in lower water availability in the downstream buckets.

Line 350: Is there any reason to believe that "noisy" parameter values indicate parameter interaction? Could it just be that there is no low-frequency signal that is needed in this basin to compensate for lack of information in seasonal precipitation signal? Also, what does "noisy" mean? Why do we think this high-frequency variability is noise?

This whole set of experiments is leading me to hypothesize that the problem with the bucket models is that there aren't enough buckets. More buckets allow for more flexibility in the mixing of residence times. We could use hundreds of buckets wired in parallel and series, with skip connections (to account for mass transfers between buckets on timescales shorter than the number of timesteps represented by the distance between two buckets in series). We could train this much bigger bucket model and look at residence time mixing ratios over seasons. And really, this would not be very much different than using a set of fully connected layers with linear coefficients, which – surprise, surprise – is exactly what the head layer on the standalone LSTM is.

Line 415: Referring back to my main criticism, I would like to take issue with this sentence from the conclusion section: *"This test addressed one of the main benefits of hybrid models over purely data-driven ones, which is their ability to predict untrained variables."* I do not believe that it has been shown that this is a benefit of the hybrid modeling approach. LSTMs do this by themselves, and the authors even cited papers in their introduction that demonstrated this. I can imagine thinking that this might be a benefit, but that was not demonstrated in this paper (or any other that is currently published).

One question I have after reading all this is about why the parameters should change over time (i.e., why dynamic parameters work and static ones don’t). Of course, the answer appears to be simply that the bucket model isn’t providing any information and since the LSTM is doing the predicting, of course that prediction needs to be dynamic. But I wonder whether the dynamics in the conductivity parameters in particular might vary with moisture content in a some way that could be interpreted as a characteristic curve?