An argument for parsimony in differentiable hydrologic models

Sandeep Poudel and Scott Steinschneider

Corresponding author: Sandeep Poudel, sp2596@cornell.edu

Key

Black font: Reviewer comments

Bold blue font: Author responses

Acronyms: OOS-time = out-of-sample in time; OOS-space = out-of-sample in space

---

Dear editor, please find attached my review of the manuscript.

1. Scope

The scope of the article is inside the scope of HESS.

2. Summary

The authors test how hybrid models composed of a NN+HBV behave when the NN gives static parameterization. They test two approaches, MLP and LSTM. They show that the performance of static-parameterised hybrid models is competitive with dynamic parameterization.

3. General comments

In general, I really liked the paper. I think it is very well-written, the objectives and the methods used are clear. I just have two major comments.

We thank the reviewer for the thorough and constructive review. Below we provide a point-by-point response to each comment.

Comment 1:

Why didn't you use the same periods as Feng2022? You compare your model to his, but you used different training and testing periods. One of the advantages of running experiments in CAMELS-US is that you can benchmark against existing models, automatically placing your study in current literature. But to do this, you should reproduce as good as possible the other

study, and this includes training/testing splits. Even with different dates, you got similar performance, but then other factors could have played a role. I would suggest that you consider running the experiments for the same periods.

We agree that consistent experimental conditions facilitate direct comparison, though we note there is no consistent training/testing split even in reference studies. For instance, Feng et al. (2022) trained in 1980–1995 and tested out-of-sample in time for 1995–2010, while Feng et al. (2023) used a different training period (1989–1999) for out-of-sample in space testing. There are also additional differences between their study and what other researchers, including us, have adopted in basin selection (671 vs. 531), loss function (weighted NSE vs. MSE), and stopping criteria (fixed epochs vs. early stopping on validation loss). While it is debatable which choices are preferable, our experience with differentiable modeling suggests that overall performance is generally not that sensitive, and these differences in choices are acceptable.

That said, we agree that at minimum aligning training and evaluation periods would enable more direct comparison. We therefore re-trained our MLP-based ensemble hybrid model using the same period as Feng et al. (2022). Our results are similar with both our original findings and theirs, as shown in the NSE CDF plot below. We will discuss this in our revised manuscript.

Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. Water Resources Research, 58(10), e2022WR032404.

Feng, D., Beck, H., Lawson, K., & Shen, C. (2023). The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment. Hydrology and Earth System Sciences, 27(12), 2357-2373.

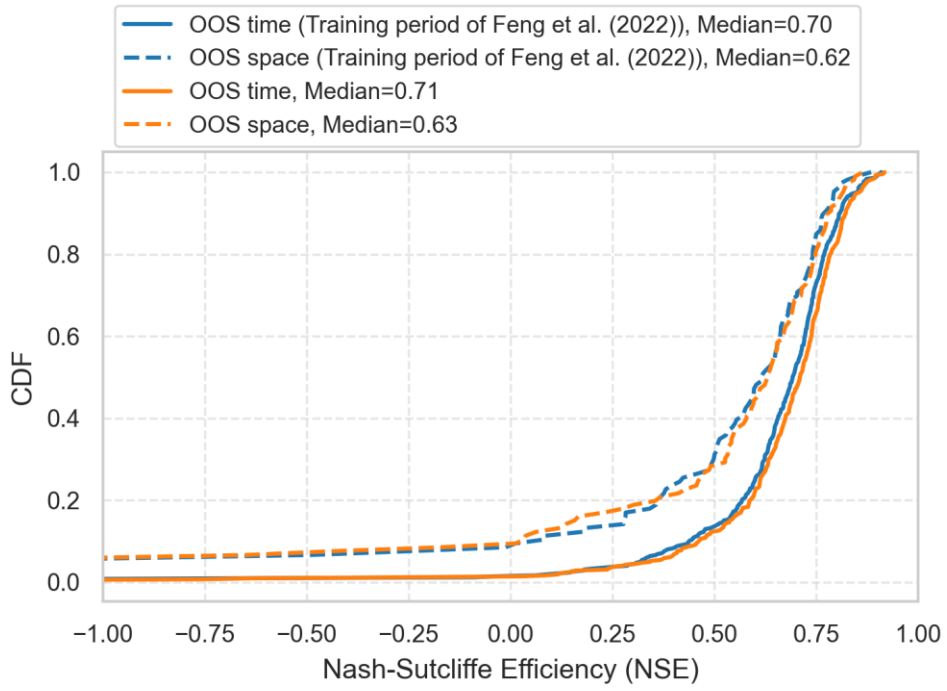

Figure R1: Cumulative distribution of NSE for differentiable MLP-based model trained on 1990-2005 and tested on 1980-1989 versus training and testing on the same time period as in Feng et. al (2022).

Comment 2:

You are selling hybrids as a way to have competitive performance with data-driven models while maintaining an interpretable model. But I think it is fair to ask the question, how much interpretability are we actually gaining? And I ask this as a person that have work a lot with hybrid models, and that also bought the interpretability argument.

Your models have 16 HBVs acting in parallel, for a total of 320 parameters. Can you really interpret this? If your HBV have 3 or 4 buckets, you are talking about 48 to 64 states. Moreover, you have basin-averaged quantities, so how useful is the information contained in the buckets? What can you tell a stakeholder based on this? That there is a lot of snow in the catchment and that the groundwater is high. This could be the case, but how useful is this? And I ask this as a genuine question, maybe this is useful, I just do not know how. In my opinion, basin-averaged conceptual models give interpretability mostly by association, and the physical principles are

quite weak. LSTMs also do not provide physical interpretability, but I think the method is quite honest about it and mostly focused on performance.
To be clear, this does not invalidate any of the paper experiment nor results, and hybrid models might have other advantages. Moreover, making the models more parsimonious by replacing the LSTM with a MLP would indeed be a good strategy. My only concern is that promising interpretability might be an oversell, because even with a simpler and more parsimonious model, is the conceptual head layer really interpretable?

We appreciate the reviewer's thoughtful comments on this point and agree with much of what has been raised. Lumped conceptual models aggregated at the catchment scale are only abstractions and do not represent the true complexity of real watershed systems. Ensembles of such models, combined with equifinality, further complicate the interpretation of individual parameters and internal states. We therefore agree that claiming full or unambiguous physical interpretability would be an oversell.

That said, our position is that a carefully designed hybrid model can still offer meaningful advantages over purely data-driven approaches, even if the interpretability is partial. These models are built on physically consistent equations reflecting how hydrologists conceptualize watershed functioning at a coarse scale, and they guarantee conservation of mass regardless of whether internal parameters are perfectly identified. Furthermore, simulated internal states such as soil moisture and evapotranspiration, despite being estimated solely by fitting to streamflow, often correlate well with independent satellite-based observations, providing additional credibility to the model's internal workings beyond streamflow prediction alone. We believe that this partial interpretability, along with the possibility of additional validation against independent observations and competitive predictive performance relative to LSTMs, does provide a meaningful edge for such hybrid models. This offers modelers additional trust in the utility of their model, as well as a story of the event (with use of internal states) to communicate with stakeholders. Feng et al. (2022) go into detail on the advantages of such hybrid models, noting that one could look into the internal states and see that high antecedent soil moisture or thawing snow primed the watershed for floods, which could be an important point for

communication with stakeholders. Another point is that if the hybrid model is developed using an MLP like we suggest in this work, we get truly static parameters, similar to the traditional approach of using some Genetic Algorithm to calibrate a process-based model. In this case, all the model diagnostics developed over the past few decades to assess internal states and parameters of process-based models (Clark et al., 2017) are directly applicable to MLP-based hybrid models, which directly connects the hybrid model to the long history of process-based model diagnostics and may provide additional ways for interpretability.

Overall, we agree with the reviewer's caution that physical structure in a hybrid setup does not automatically guarantee better interpretability or model behavior, a point also demonstrated in the reviewer's own work (Acuna Espinoza et al., 2024). We will revise the manuscript to discuss the challenges of interpreting these hybrid models, including those arising from the lumped and conceptual model structure, equifinality, and the ensemble setup which compounds these challenges further.

Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., Pauwels, V. R. N., Cai, X., Wood, A. W., and Peters-Lidard, C. D.: The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism, Hydrol. Earth Syst. Sci., 21, 3427–3440, https://doi.org/10.5194/hess-21-3427-2017, 2017.

Acuña Espinoza, E., Loritz, R., Álvarez Chaves, M., Bäuerle, N., & Ehret, U. (2024). To bucket or not to bucket? Analyzing the performance and interpretability of hybrid hydrological models with dynamic parameterization. Hydrology and Earth System Sciences, 28(12), 2705-2719.

4. Specific comment

Line 290-296: Why do LSTMs outperform MLP for the single HBV case, if both are providing static parameterization? Is there any additional flexibility that the LSTM can provide? Because during testing, there is only one set of parameters for the whole testing series. Are the parameters

"better informed" because of the dynamic series? Why does this difference disappear for the ensemble case? The performance gain from 1 to 16 HBVs is expected (more model flexibility), but why does the gap between MLP and LSTM disappear? Further explanation is necessary.

This is an interesting observation. While we do not have a definitive mechanistic explanation, we speculate that under the single HBV setup, the LSTM's capacity to process dynamic input sequences provides some additional representational flexibility over the MLP – this advantage may help the LSTM produce a better-calibrated single parameter set. However, when the model is expanded to an ensemble of 16 HBVs, the diversity in parameters generated by the MLP-based ensemble appears sufficient to match the performance gain of the LSTM's architectural flexibility. We will revise the manuscript to provide this explanation.

Line 303-304: What about out-of-sample in time? Are these metrics similar?

Yes, out-of-sample in time performance is similarly competitive. We previously omitted this result for brevity, we have now included the result for both cases, which will be updated in the manuscript (see figure below).

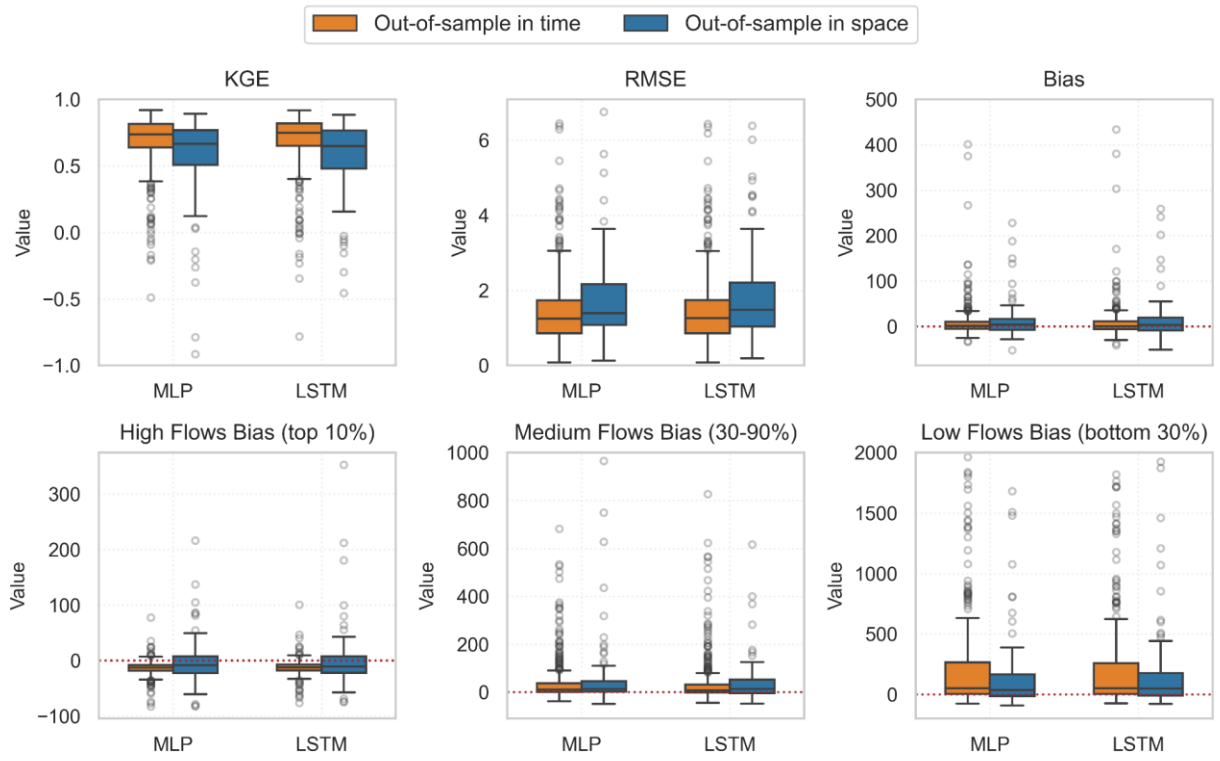

Figure R2: Performance comparison across both OOS time and OOS space cases. (We will update Figure 4 in main manuscript with this figure)

Line 329-333: I do not share this point.

I agree with what you said before (lines 329-330) that temporal generalization is worse with 2 static inputs because 2 static inputs are not enough to "fine-tune" the basin´s specific behaviour. On the other hand, 30 static attributes allow the MLP and LSTMs (I will say NN to refer to both) to further tune the output to the basin response. So for sure the NN have some memory of which basin they are treating.

What I do not agree with is saying that there are no meaningful relationships learn (line 329) with the static parameters. You actually showed this with the second experiment! You showed that if static attributes are permutated, the performance drops (with respect to the reference) for out-of-sample in space, proving that there are some meaningful relationships that were being learn. Otherwise, why will the performance drops?

The fact that permutation does not drop performance for out-of-sample in time can be explained due to the fact that, if the static-feature space is large enough (30 dimensions) the NN can still use this space to fine-tune basin-specific behaviours. So it basically uses it to memorize where it

is and give the best parameters. However, because the static attributes are not consistent with the basin characteristics (because they were permutated), out-of-sample in space drops with respect to the reference (0.42 instead of 0.62). This showed that the reference did interpret the non-permutated parameters in a way that allows it to better generalize in space.
The NN do not have to interpret the static parameters the same as us, we are actually not training them to do that. Moreover, they should not be used for contrafactual checking (e.g. what would happen to runoff if we double the area) because again, we are not training them to do that. But they can learn that basins at certain latitudes are more arid or that bigger areas show bigger discharges.
This comment also extend to the second point in Discussion and Conclusions.

We note that the reviewer agrees with our interpretation that static attributes help the network memorize (or fine-tune) basin-specific rainfall–runoff behavior, supporting OOS-time generalization. The main point the reviewer raises is that if nonsensical (permuting) static attributes cause a drop in OOS-space performance, this suggests that some meaningful physical relationships are being learned from the correct attributes.

We agree that the true static attributes contain meaningful information the network can use. However, we do not think the current results clearly show that the network is learning *physically* meaningful process relationships in an interpretable sense (e.g., that high clay content leads to faster runoff response). Rather, our results suggest that the primary role of the static attributes is to help the network infer each basin's geographic context, which in turn supports OOS-space generalization. The strongest evidence for this is that lat/lon alone yields OOS-space performance comparable to the full set of 30 physical attributes, whereas permuted attributes substantially reduce OOS-space performance (0.42 vs. 0.62). This implies that the network is mainly using the static attributes to determine where a basin sits geographically, where many attributes are similar across relatively homogeneous regions, rather than to learn physically meaningful hydrologic process relationships. A secondary role of the static attributes is to provide a high-dimensional input space that supports memorization of basin-specific behavior, as shown by the fact that OOS-time performance is retained even when the 30 static attributes are nonsensical.

There are therefore two possible interpretations for how a neural network learns from static attributes. Under Interpretation A (Physical relationship learning), the network learns meaningful physical mappings between static basin attributes and hydrological processes. Under this interpretation, a network using 30 physical attributes should outperform a network using 2 geographic coordinates in OOS-space generalization. Under Interpretation B (Geographic coordinate learning), the network uses static attributes mainly to infer a basins geographic context, learning that certain combinations of catchment properties are associated with particular regions and their characteristic rainfall–runoff behavior. Under this interpretation, latitude and longitude should perform comparably to the 30 physical attributes for OOS-space generalization.

Our results are more consistent with Interpretation B, as summarized in the table below:

Static attributes used as NN input	OOS Time generalization	OOS Space generalization
30 physical attributes (no lat and lon)	Yes, because the high-dimensional input supports basin memorization	Yes, because the attributes are used to learn geographic positioning and the model can locate unseen basins geographically
Lat and lon only	No, because only 2 inputs are insufficient for basin memorization	Yes, because geographic positioning is directly encoded with lat and lon and spatial generalization is retained
30 nonsensical physical attributes	Yes, because the high-dimensional input still supports memorization despite carrying no physical meaning	No, because nonsensical inputs prevent geographic learning and spatial generalization is lost

We acknowledge that the original manuscript did not explain these points with sufficient clarity, and we will carefully revise the relevant sections to reflect this discussion.

Section 3.3 and Figure 7: Is there a reason you are doing these experiments in the training period? I think keeping consistency in doing everything in the testing period is better. Further down in this section (line 374 and forward) you change to results in the testing period. Why not do everything in testing?

Figure 7 must be shown for the training period because the LSTM-based differentiable model produces dynamic (sequence-varying) parameters only during training; during testing, a single parameter set is applied across the full period.

Line 390-393: I do not agree with this conclusion.
The deficiency I see in this set of experiments is that, you are approximating dynamic parameters as static parameters that can vary each 60 days. 2 months is not a short period, you cannot react to specific events, only to seasonal changes. The flexibility this might give you is a seasonal model that has a specific behavior every two months. On the other hand, the dynamic models from Feng2022 or Acuna2024 can react to specific events.
For example, the models Delta_1 and Delta_n from Feng 2022 show that when only one HBV is used, the median goes from 0.64 to 0.71, so there is a big performance gain when going dynamic. Now, Feng also shows that when you have 16 HBVs, the gain due to dynamic parameterization is much less , goes from 0.71 to 0.73. This, I would argue, is an indication that what gives improved performance is the additional model flexibility. That this flexibility is physically consistent is another point, and maybe you can argue that it is not. You also showed the same thing: 16 HBVs performed better than 1, because you have a more flexible model.
Alvarez2026 shows this in a really nice sets of experiments. I highly recommend this reading. If one has a perfect model, the parameters will be static, and as the model is a worse representation of what is happening, then the dynamic parameterization stars to compensate for structural deficiencies.
All this to say that I do not think your experiments can show that “Consequently, the observed

temporal variability in LSTM-predicted parameters may be better interpreted as a manifestation of equifinality rather than evidence of dynamically evolving process", because the 60-day windows you are using do not allow you to focus on dynamically evolving processes.
This comment extends to point 3 of Discussion and Conclusions.

We first clarify a point about Feng et al. (2022). The performance gain from Delta_1 to Delta_n (NSE 0.64 to 0.71) results entirely from increasing the ensemble size from 1 to n HBVs, not from dynamic parameterization. Both configurations use the same sequence-to-one LSTM with sequence-varying parameters. The additional gain from dynamic parameterization (0.71 to 0.73) only appears when a sequence-to-sequence LSTM with time-step-varying parameters is used, and whether this small gain reflects real process dynamics is arguable, as the reviewer also points out.

Now turning to the reviewer's main concern, we agree that a 60-day stride may limit the model's ability to react to individual events. However, our claim about equifinality should not be impacted by this, as even in this formulation the model has the ability to capture hydrological process dynamics occurring at the seasonal scale. If the sequence-varying parameters were physically meaningful, one would expect parameter sets tuned to different seasons (e.g., summer vs. winter) to produce better streamflow simulations than a single fixed parameter set taken from the final time step. Figure 8 in our manuscript clearly shows that this is not the case; performance is the same whether we use a single parameter set or allow parameters to vary across sequences.

Further, to directly address the concern about stride length, we ran an additional experiment with a stride of 1, allowing parameters to update at every time step. Results with stride 1 are consistent with those from stride 60, both in terms of dynamicity of HBV parameters (Figure R3 below), and also in terms of not showing improvement in simulation performance when using time-varying parameters relative to using the static parameter from the last time step (Figure R4). We also note here that the 60-day stride was originally chosen because reducing it below 60 days showed no improvement in validation performance while only increasing computational cost.

Overall, this additional result further supports our conclusion that temporal variation in predicted parameters within the sequence-to-one LSTM formulation reflects equifinality rather than true process dynamics. However, we do not extend this claim to time-step-varying sequence-to-sequence formulations such as those in Feng et al. (2022) and Acuna Espinoza et al. (2024), and we will make this distinction explicit in the revised manuscript. We will also discuss our result from the 1-day stride in the revised manuscript.

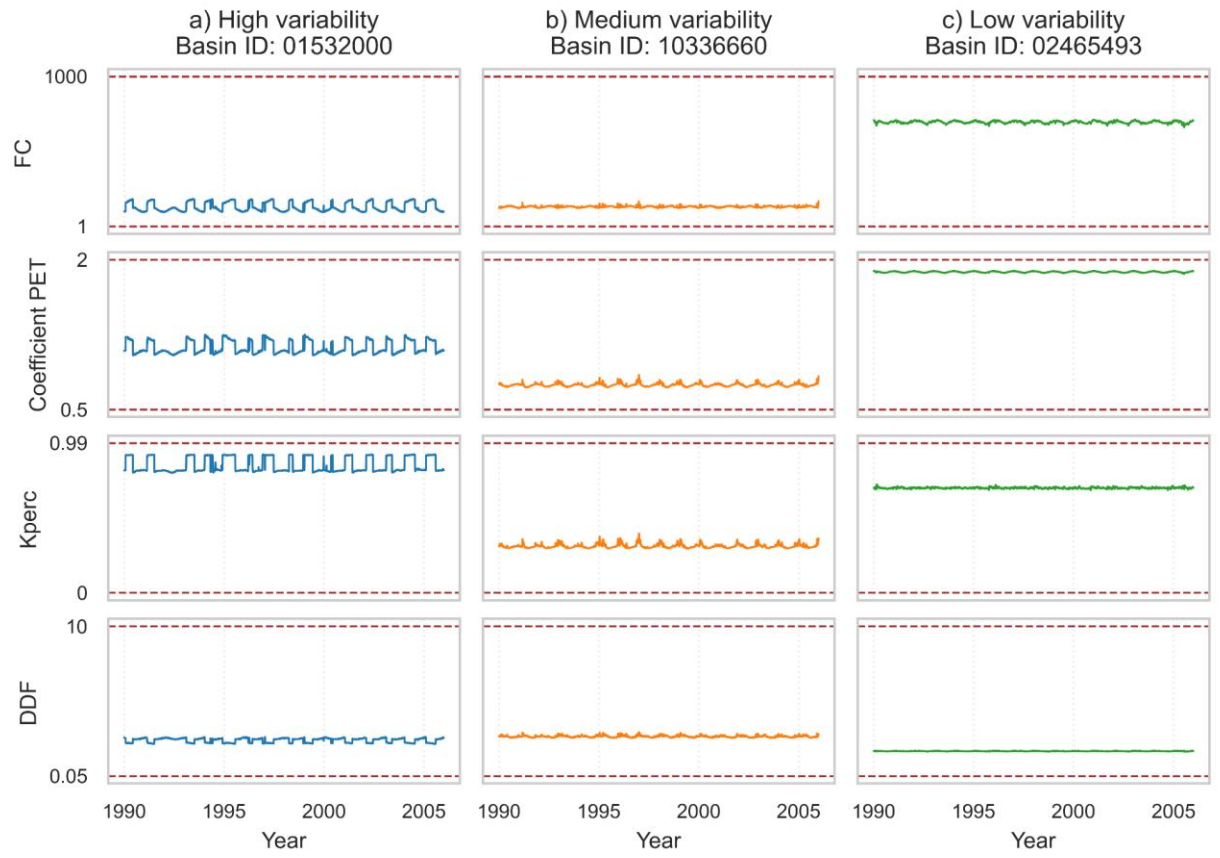

Figure R3: Temporal variability of HBV parameters for a stride of 1 day, meaning parameters vary each day. This figure is the same as Figure 7 in our manuscript which was for stride of 60 days but now this is for stride of 1 day.

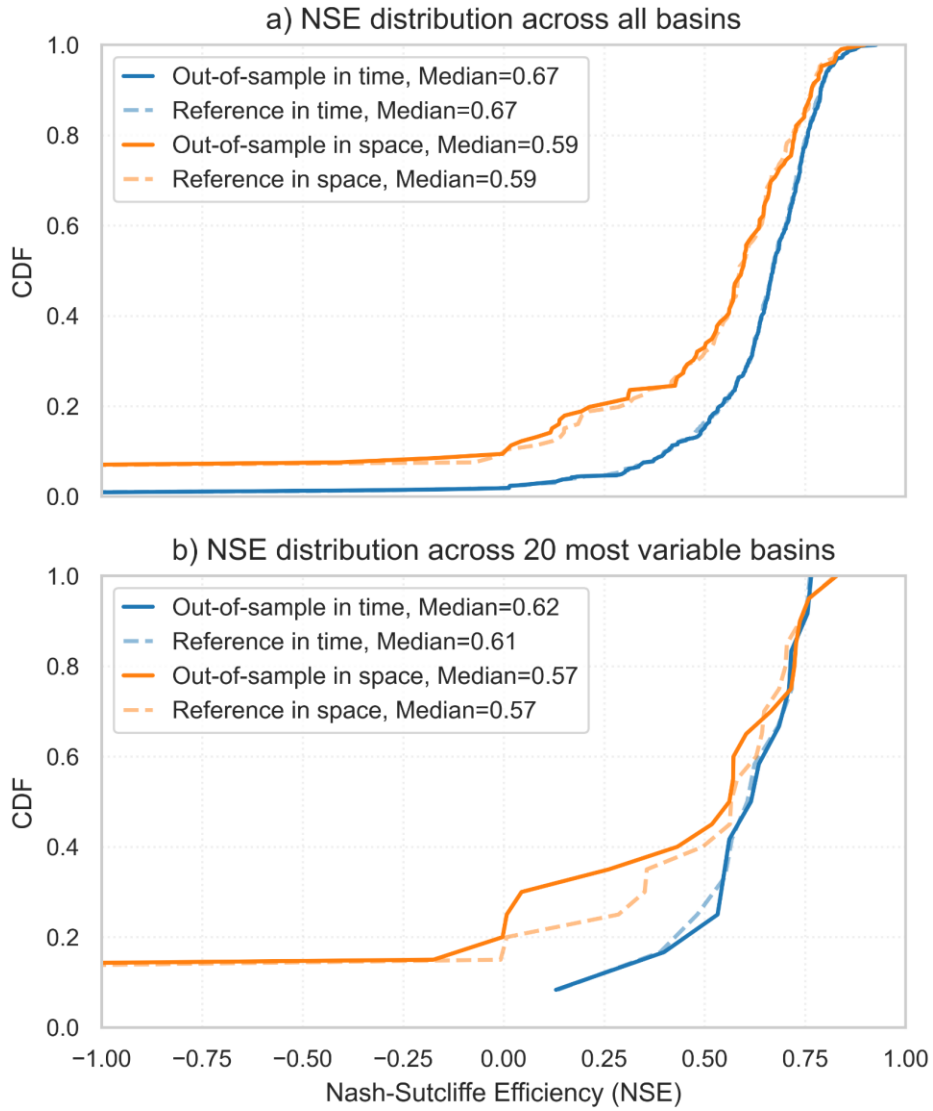

Figure R4: NSE distribution using stride of 1 day (solid line) versus using a single static parameter from last time step (dotted). This figure is the same as Figure 7 in our manuscript which was for stride of 60 days but now this is for stride of 1 day.

5. Recommendation

Based on the comments above, I recommend acceptance subject to major revisions. As I indicated, I think the document is really well done. I was planning to suggest minor revisions, but I think the timeline could be a bit short for the authors. However, if the editor feels that they can be done as minor revisions, I would also support it.

Kind regards,
Eduardo

Thank you Eduardo.

References:

Acuña Espinoza, E., Loritz, R., Álvarez Chaves, M., Bäuerle, N., & Ehret, U. (2024). To bucket or not to bucket? Analyzing the performance and interpretability of hybrid hydrological models with dynamic parameterization. Hydrology and Earth System Sciences, 28(12), 2705–2719. https://doi.org/10.5194/hess-28-2705-2024

Álvarez Chaves, M., Acuña Espinoza, E., Ehret, U., & Guthke, A. (2026). When physics gets in the way: An entropy-based evaluation of conceptual constraints in hybrid hydrological models. Hydrology and Earth System Sciences, 30(3), 629–658. https://doi.org/10.5194/hess-30-629-2026

Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, Learnable, Regionalized Process-Based Models With Multiphysical Outputs can Approach State-Of-The-Art Hydrologic Prediction Accuracy. Water Resources Research, 58(10), e2022WR032404. https://doi.org/10.1029/2022WR032404