

The following comments detail my review of the manuscript "From prediction improvement to uncertainty propagation structure: revisiting hybrid hydrological models," submitted to *Hydrology and Earth System Sciences*. The paper addresses a current and very active area of research regarding hybrid hydrological models. It is well written, and the results are supported by the evidence shown. However, I suggest ****major revisions**** before publication.

Response: Thanks for your recognition and valuable suggestions. We provide the response to the comment starting with "**Response**".

Major Comments

1. The authors need to better justify their selection of catchments. The initial selection of just HUCs 1, 4, and 17 seems arbitrary. Why not just filter for catchments with the described mean annual air temperatures regardless of the region? Furthermore, the authors need to describe the "substantial computational cost" to justify reducing their potential training set by 75%.

Response: We agree that the rationale for catchment selection and the computational requirements were insufficiently explained. Our study includes the replacement of precipitation partitioning and snowmelt processes, motivating a focus on relatively cold regions. These three regions meet the regional selection criteria of mean annual air temperature below 8 °C and more than 20 available catchments, and were previously used to evaluate snow-related process replacement in Li et al. (2023). The present study extends that investigation by examining whether improvements in predictive performance are accompanied by increased optimization-induced variability in internal hydrological processes and states. Retaining the same regional setting provides continuity between these research questions. We acknowledge that screening individual catchments by temperature across the whole CAMELS dataset would also be a valid approach and could broaden the study domain. We will clarify the rationale for our regional focus and acknowledge that the selected catchments do not comprehensively represent all cold-region catchments.

Regarding computational cost, the average runtime per independent optimization run was approximately 10 minutes, calculated across the three model configurations (EXP, NN_Q and NN_{QSM}). This represents an overall average across configurations, rather than an identical runtime for each model. The main ensemble experiment comprised 4650 independent optimization runs (31 catchments $\times$ 3 configurations $\times$ 50 realizations), corresponding to an estimated cumulative runtime of 775 hours. Extending the same experimental design to all 124 candidate catchments would require 18 600 runs, with an estimated cumulative runtime of 3100 hours. These estimates represent the summed runtimes of individual runs; actual elapsed time depends on parallel execution. We will include these quantitative estimates in the revised manuscript to clarify the computational considerations underlying the catchment subsampling.

2. Although the model hyperparameters and training configuration are likely detailed in previous work, I suggest including a summarized version in this paper for completeness, perhaps in the appendix.

Response: Thank you for this suggestion. We will include a summary of the model hyperparameters and training configuration in the appendix. We will also make the model implementation and training code publicly available and provide the repository link in the revised manuscript to support reproducibility.

3. Of the models evaluated, I would insist on including a data-driven baseline, such as an LSTM. Such a baseline provides a point of comparison that simply fits the data directly, alongside all evaluated metrics against the observed data. This suggestion is also motivated by the large increase in the IQR of some of the internal processes for the NN_{QSM} models (S, M, S0). This indicates that the internal trajectories of these states and fluxes are not truly constrained, meaning the NN_{QSM} model is essentially one small step away from being a fully data-driven solution.

Response: We will include an LSTM as a data-driven baseline, using the same 31 basins, meteorological inputs, training and evaluation periods, and observed runoff targets. Consistent with the existing experiments, the LSTM will be trained separately for each basin with 50 independent optimization runs. Because a standard LSTM does not provide physically defined counterparts to S , M and S_0 , the baseline comparison will focus on runoff predictions and their ensemble variability. We also acknowledge the concern regarding the increased IQR of S , M and S_0 in NN_{QSM} . This pattern suggests that calibration against runoff alone may provide insufficient constraints on internal processes and states. We will strengthen the discussion of this limitation and its potential connection to equifinality.

4. Section 3.2: Particularly for the neural network ENN_Q, I believe it is a missed opportunity to only analyze state variables in this study. Höge et al. (2022) in Sect. 4 show how much replacing a process with a neural network changes the expected behavior of that process. Showing a similar response surface for ENN_Q in NN_Q and NN_QSM would complement this section nicely.

Response: Thank you for this helpful suggestion. Following the approach in Sect. 4 of Höge et al. (2022), we will explore response-surface analysis of ENN_Q in NN_Q and NN_{QSM} , examining runoff as a function of liquid-water input (P_r+M) and basin water storage (S_1). To connect this analysis with the uncertainty focus of our study, we will explore the ensemble median and IQR of these responses across the 50 optimization runs, evaluated at common input combinations. This could help identify the conditions under which independently optimized ENNs exhibit greater disagreement in their learned runoff relationships. We will consider the input ranges supported by the simulations and assess how these diagnostics can complement the existing analysis in Sect. 3.2.

5. Section 3.4: The idea of uncertainty "accumulating" is misleading, as the analysis does not aggregate over time and only overall statements can be made. Uncertainty "accumulating" implies that the model remains able to make

accurate predictions over time despite greater variability in its internal states. Furthermore, the fact that this occurs (NN_QSM performs better than NN_Q, which performs better than EXP) points to an issue of equifinality that is not addressed at all in the paper.

Response: We agree that the current analysis does not directly demonstrate temporal accumulation of uncertainty. Our use of “uncertainty reservoirs” and “uncertainty accumulates” implied a temporal mechanism that was not established by the analysis. The ensemble comparisons show that greater optimization-induced variability in ENN-replaced processes is accompanied by greater variability in their corresponding state variables, whereas unreplaced processes show comparatively little change. Our intention was to suggest that increased optimization-induced uncertainty in ENN-replaced processes may propagate primarily to their corresponding state variables, with comparatively limited effects on unreplaced hydrological processes. We will revise the headings and relevant text in Sects. 3.3–3.4 and the corresponding statements in the abstract, discussion, and conclusions. We will remove the “uncertainty reservoirs” metaphor and replace claims of temporal accumulation with descriptions of greater ensemble variability in the relevant internal states. Any discussion of storage memory will be framed as a possible interpretation, rather than a mechanism demonstrated by the present analysis.

We also agree that the potential role of equifinality deserves explicit discussion. When calibration relies on runoff observations, different parameterizations and internal process representations may achieve similar runoff performance while producing different internal fluxes and storage trajectories. Improved runoff performance does not necessarily imply better-constrained internal processes. The coexistence of improved predictive performance and greater internal ensemble variability highlights this concern, although it does not by itself establish equifinality. We will expand the discussion to address this possible explanation and emphasize the importance of evaluating internal variability alongside predictive accuracy.

6. This uncertainty framework captures only optimization-induced uncertainty (random initialization/stochastic training), not parameter, forcing, or structural uncertainty. This is acknowledged as a limitation (L382–384), but given that the abstract and introduction frame the study broadly around "internal uncertainty propagation", it should be clarified in those sections too.

Response: We agree that the scope of the uncertainty assessment should be stated explicitly throughout the manuscript. We will revise the title, abstract, introduction, and conclusions to consistently use “optimization induced uncertainty,” referring to variability arising from random parameter initialization and stochastic optimization. References to “internal uncertainty propagation” will be explicitly qualified as concerning this source of variability.

Minor comments

1. L103: the process 'partitioning S' is not labeled in Figure 1a.

Response: We will add the label “S” for the precipitation partitioning process in Fig. 1.

2. L104: the process 'evapotranspiration E' is labeled in Figure 1a as ET.s

Response: We will replace “E” with “ET” at L104 and use “ET” consistently to denote evapotranspiration throughout the manuscript and figures.

3. L127: the use of 'day length' seems a bit unusual. Is it used only by the ENNs?

Response: Day length is used together with air temperature to calculate potential evapotranspiration using the Hamon formulation in all three model configurations (EXP, NN_Q , and NN_{QSM}). It is not a direct input to the ENNs used in this study. We will clarify its role in Sect. 2.3.

4. L203: this normalization seems important, yet there is little detail provided. Do the authors mean normalized by the 'total' precipitation in a basin? Or, if it is a daily process, how were zero-precipitation days handled?

Response: Thank you for pointing out this ambiguity. Snowfall, snowmelt, and evapotranspiration were normalized by the annual precipitation total for each basin. Basins receiving more precipitation may exhibit larger absolute IQR values partly

because their hydrological fluxes have larger magnitudes. The normalization expresses ensemble variability relative to each basin's annual precipitation input, reducing the influence of differences in water-input magnitude on IQR comparisons across basins. We will clarify both the normalization procedure and its purpose in the revised manuscript.

5. Figure 3: outliers are barely visible in these plots. Additionally, the overall quality of the figure is poor.

Response: We will redraw Fig. 3 with larger, higher-contrast outlier markers and improve the figure resolution, font sizes, and line clarity to enhance readability.

6. L226: typo: "compraison"

Response: We will correct "compraison" to "comparison".

7. L285: typo: "processess"

Response: We will correct "processess" to "processes".

8. L331: is this meant to refer to Sect. 3.3?

Response: Yes, we will correct "Sect. 3.1.3" to "Sect. 3.3".

9. L416: the variables and what they denote are incorrect here.

Response: We will correct the variable definitions: where P_s , P and T denote snowfall, precipitation, and mean air temperature, respectively.

Li, B., Sun, T., Tian, F. and Ni, G. 2023. Enhancing process-based hydrological models with embedded neural networks: A hybrid approach. Journal of Hydrology, 130107.