

1 This manuscript examines how replacing selected hydrological processes with
2 embedded neural networks affects both streamflow prediction and variability in
3 internal fluxes and states. The topic is interesting, and the 50 stochastic
4 realizations across 31 basins provide useful information beyond conventional
5 performance metrics. My main concern is that several interpretations are broader
6 than what the current experiments directly support. The study mainly quantifies
7 variability arising from random initialization and stochastic optimization, while
8 the manuscript often discusses general “uncertainty propagation.” In addition, the
9 proposed buffering and temporal accumulation mechanisms, and the attribution
10 of increased variability specifically to ENN embedding, require stronger evidence.
11 I therefore recommend major revision.

12 **Response:** Thanks for your recognition and valuable suggestions. We provide the
13 response to the comment starting with “**Response**”.

14 **1. Scope and definition of uncertainty.** The manuscript should distinguish
15 optimization-induced/stochastic-training variability from broader hydrological or
16 predictive uncertainty. Because forcing, observations, and model structure within
17 each configuration are fixed, the ensemble does not represent forcing,
18 observational, structural, or complete predictive uncertainty. Different
19 uncertainty sources can affect physical and machine-learning models differently
20 (Li et al., 2020; Li et al., 2025), and probabilistic uncertainty estimation generally
21 requires more than stochastic-training spread alone (Klotz et al., 2022).
22 Meteorological forcing products can also introduce substantial variability
23 (Kratzert et al., 2021; Li et al., 2025). The title, abstract, and conclusions should
24 therefore either use more specific terminology or better justify the broader
25 framing of “uncertainty propagation.”

26 **Response:** We agree that the ensemble variability examined in this study need to be
27 more clearly distinguished from broader hydrological and predictive uncertainty.
28 References to internal uncertainty propagation should be qualified as concerning

optimization-induced variability across hydrological processes and states. Although Sect. 2.2 defined the scope as optimization-induced uncertainty, this distinction was not consistently maintained in the title, abstract, introduction, and conclusions. In the revised manuscript, we will consistently use the term “optimization-induced uncertainty” to describe variability arising from random parameter initialization and stochastic optimization. We will also strengthen the limitations discussion to explicitly acknowledge that forcing, observational, and structural uncertainty are not comprehensively assessed by the present ensemble framework. Furthermore, we will clarify that the ensemble percentile bands and their coverage statistics serve as diagnostics of the stochastic optimization ensemble, rather than providing a complete probabilistic characterization of predictive uncertainty.

2. Evidence for temporal uncertainty accumulation. The manuscript states that internal state variables act as “uncertainty reservoirs” in which uncertainty accumulates over time. However, a larger state-variable IQR demonstrates greater ensemble disagreement, but does not by itself establish temporal accumulation. A time-varying analysis of ensemble spread would provide more direct evidence of whether variability introduced by ENN-replaced processes persists or grows in storage states. Otherwise, the conclusion should be moderated to state that the results are consistent with memory-related uncertainty accumulation rather than directly demonstrating it.

Response: We agree that the current results do not directly demonstrate that internal state variables act as “uncertainty reservoirs” in which uncertainty accumulates over time. 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 acknowledge

that this is an interpretation of the observed patterns, rather than direct evidence of temporal accumulation.

We will therefore rename Sect. 3.3-3.4 and revise the corresponding statements in the abstract, discussion, and conclusions. Specifically, we will remove the “uncertainty reservoirs” metaphor and replace claims of temporal accumulation with descriptions of greater ensemble variability in the relevant internal states. The potential role of storage memory will be discussed as a possible explanation for these patterns, rather than as a mechanism demonstrated by the present analysis.

3. Model flexibility, identifiability, and training strategy. It is unclear whether the larger ensemble spread is caused specifically by ENNs or more generally by the greater flexibility of the substituted model structure. Replacing a constrained physical formulation with an ENN changes both the representation type and the model’s effective degrees of freedom. A more flexible structure may admit more similarly performing solutions under limited calibration information, leading to weaker identifiability or equifinality and greater seed sensitivity. A more flexible physics-based representation could potentially show similar behavior. This issue may be amplified by site-by-site training. Each ENN is constrained by data from a single basin and mainly by the runoff objective, allowing multiple internal representations to achieve similar streamflow performance. Multi-basin training may provide stronger constraints through more diverse hydrological information and could reduce non-identifiability and seed sensitivity (Kratzert et al., 2024).

Response: We agree that the present experiments do not separate the effects of neural-network representation from those of increased model flexibility and changes in process constraints. Replacing a physical formulation with an ENN changes these aspects simultaneously. The larger ensemble spread should therefore be interpreted as a finding associated with the specific configurations investigated, rather than an intrinsic property of ENNs. More flexible physics-based formulations could potentially exhibit similar behavior, although this possibility was not evaluated in our experiments.

We also agree that basin-specific training and calibration against runoff alone may contribute to the observed variability. All three configurations were trained separately for each basin using the same observational information. This provides a consistent basis for comparison within the basin-specific setting, but does not necessarily constrain configurations with different flexibility equally. Different parameterizations and internal process representations may achieve similar runoff performance while producing different internal fluxes and storage trajectories.

We will explicitly discuss limited identifiability and equifinality as possible explanations for the observed ensemble spread, while clarifying that the present analysis does not directly establish these mechanisms. We will also discuss multi-basin training as a potential approach to introducing more diverse hydrological information, drawing on the benefits reported for LSTM models by Kratzert et al. (2024). Whether this approach reduces seed sensitivity and improves internal-process identifiability in the present hybrid configurations remains an open question requiring dedicated evaluation.

Our study highlights that improvements in runoff prediction can coexist with greater optimization-induced variability in internal processes and states. We will expand Sects. 4.1 and 4.2 and qualify the relevant conclusions to distinguish this observed pattern from its possible explanations, and to acknowledge the limitations imposed by model flexibility, runoff-only supervision, and basin-specific training.

Minor comments

1. In Eqs. (5)–(6), N appears to represent time steps although it was previously defined as the number of ensemble members.

Response: Thanks for your suggestion. We will replace N with T in Eqs. (5)–(6), where T denotes the number of evaluation time steps. N will remain the notation for the number of ensemble members, as defined in Eq. (4).

2. Around lines 325–327, EXP appears to be incorrectly described as having an ENN runoff replacement.

113 **Response:** We will correct “EXP and NN_Q ” to “ NN_{QSM} and NN_Q ” in this sentence.

114 **3. The reference to Sect. 3.1.3 around line 331 appears incorrect.**

115 **Response:** We will correct “Sect. 3.1.3” to “Sect. 3.3”.

116 **4. In Appendix Eq. (A1), the definitions of PsT appear unclear or malformed.**

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

119 **5. Numerous additional grammatical and typographical issues remain; a careful
120 language edit is recommended.**

121 **Response:** We will carefully edit the manuscript to correct grammatical and
122 typographical errors and improve clarity.

123

124

125 Kratzert, F., Gauch, M., Klotz, D., & Nearing, G. (2024). HESS Opinions: Never train
126 a Long Short-Term Memory (LSTM) network on a single basin. Hydrology and Earth
127 System Sciences, 28, 4187–4201.