Dear Editor,

In response to the reviewer's remaining concerns, we have carefully revised the manuscript to clarify several aspects of the framework that were insufficiently explained in the previous version, including the scope of "data scarcity", the physical meaning and role of net inflow, the status of the implemented physical constraints, and the recurrent propagation of simulated storage. Corresponding revisions have been made throughout the manuscript and Supporting Information, and a detailed point-by-point response to each of the reviewer's comments is provided below.

In particular, we have addressed the reviewer's central concern that the use of observation-derived net inflow artificially anchors simulated storage to observations and therefore largely explains the reported storage performance and spatial-generalization results. As we clarify in our response and demonstrate through the governing error-propagation relationship, net inflow cancels from the storage-error equation; storage errors instead accumulate from errors in the simulated release. Thus, simulated storage is not constrained to follow the observed trajectory by construction, and good storage performance is not an automatic consequence of using net inflow. We believe this clarification resolves the principal issue underlying the reviewer's remaining concerns and more precisely communicates the validity and intended scope of the proposed framework.

Yi Zheng, on behalf of the author team

Response to Reviewer #4's comments

I thank the authors for their detailed and transparent responses and for updating the code repository. Having reviewed the authors' replies alongside both the original and revised code, I have several remaining concerns that I believe are fundamental to the paper's central claims.

Response: We thank the reviewer for carefully reviewing our revised manuscript and code. We address the remaining concerns point by point below and have revised the manuscript accordingly.

1. Net inflow as a dynamic input undermines the spatial generalization claim

The authors confirmed in their response to Comment 4 that net inflow is defined as:

$I_net,t = S_\{t+1\} - S_t + R_t$

This is not an independent hydrometeorological forcing. It is algebraically derived from observed storage and release — the very variables the model predicts and that would be unavailable for a truly data-scarce reservoir. For the spatial cross-validation (Experiments III and IV) to be a meaningful proxy for ungauged deployment, inflow must be obtainable without prior knowledge of S and R at the target reservoir. Net inflow does not satisfy this requirement.

This has a direct consequence for the performance metrics. When the model is initialized from observed storage and forced with net inflow — which encodes the observed storage and release trajectory by construction — the mass balance update becomes:

$$S_{\{t+1\}}^{\{sim\}} = S_t^{\{sim\}} - R_t^{\{sim\}} + I_{net},t = S_t^{\{sim\}} - R_t^{\{sim\}} + \left(S_{\{t+1\}}^{\{obs\}} - S_t^{\{obs\}} + R_t^{\{obs\}}\right)$$

The simulated storage trajectory is therefore strongly anchored to the observed one regardless of the quality of the release prediction. The low PBIAS values (<0.4%) in Fig. S5 are at least partially a structural artifact of this input definition, not a demonstration of learned operational skill. By extension, good storage simulation (Figure 7) is expected almost by construction in the data-rich experiments, and the strong spatial generalization results likely reflect the same artifact for the held-out reservoirs — whose net inflow must also have been computed from their own observed S and R records.

I would ask the authors to confirm explicitly how net inflow was computed for held-out reservoirs in Experiments III and IV. If it required those reservoirs' own observed records, the claimed "scalable pathway for representing human water management under widespread data scarcity" is not demonstrated by these experiments.

Response: We thank the reviewer for raising this concern. We agree that the source and role of net inflow should be stated more explicitly. However, we respectfully disagree that the use of net inflow either anchors simulated storage to observations or undermines the spatial-generalization experiments. We would also like to further clarify the rationale for using net inflow in this study.

First, we clarify that **the proposed PLSTM-Reg framework is not dependent on the use of net inflow**. Its primary objective is to learn transferable reservoir operation behavior for integration into large-scale hydrological models. In such applications, hydrological forcing is supplied externally by the host model, and PLSTM-Reg can therefore be driven by either net inflow or independently observed

or simulated reservoir inflow. This flexibility has already been demonstrated in Yu et al. (2025), where the local PLSTM reservoir model was driven directly by simulated or observed inflow.

Importantly, **net inflow remains a physically meaningful hydrological forcing**. From the complete reservoir balance,

$$S_{t+1} = S_t + I_t - R_t + (G_t - L_t),$$

net inflow can be expressed as

$$I_{net,t} = S_{t+1} - S_t + R_t = I_t + G_t - L_t,$$

where I_t is river inflow, and G_t and L_t represent other reservoir gains and losses, such as direct precipitation, evaporation, and seepage. Thus, although the net inflow provided by Chen et al. (2025) was derived using reservoir observations to close the water balance, it physically represents the net external water flux into the reservoir. Such water-balance-based reconstruction is a common practice in large-scale reservoir studies (Ford and Sankarasubramanian, 2023; Liu and He, 2023; Sadki et al., 2023).

In the present study, we used the net inflow provided by the Chen et al. (2025) dataset because consistent long-term records of directly observed reservoir inflow are not available for this large sample of reservoirs. Thus, **the use of net inflow reflects the data available for the present large-sample experiment**, rather than an inherent requirement of the modeling framework.

We next address the reviewer's concern that the use of net inflow may artificially constrain simulated storage to remain close to observations. We show that this dependence does not mathematically anchor simulated storage to the observed storage trajectory. Let $e_t^S = S_t^{sim} - S_t^{obs}$ and $e_t^R = R_t^{sim} - R_t^{obs}$ represent the simulation errors of storage and release at time t, respectively. Because the same net inflow enters both the simulated and observed water-balance equations,

$$S_{t+1}^{sim} = S_t^{sim} - R_t^{sim} + I_{net,t},$$

$$S_{t+1}^{obs} = S_t^{obs} - R_t^{obs} + I_{net,t},$$

their difference gives

$$e_{t+1}^S = S_t^{sim} - S_t^{obs} - (R_t^{sim} - R_t^{obs}) = e_t^S - e_t^R$$

and iterating this relation over time gives,

$$e_t^S = e_0^S - \sum_{\tau<t} e_\tau^R.$$

Net inflow cancels from the storage-error equation and therefore does not anchor simulated storage to observations. **Storage errors instead accumulate from release errors, so good long-term storage performance is not guaranteed by the way net**

inflow is defined.

Finally, we clarify that, in this study, **"data scarcity" specifically refers to the lack of historical operational records required for site-specific training or calibration, rather than the absence of hydrological forcing**. Experiments III and IV test whether reservoir operation behavior learned from other reservoirs can be transferred to such data-scarce reservoirs, rather than whether hydrological forcing itself can be independently reconstructed at those reservoirs. **For large-scale hydrological applications, the required forcing can be supplied by the host hydrological model**, either as routed reservoir inflow or as net inflow that combines routed inflow with other model-derived gains and losses.

To clarify the application scope of the framework, the role and interpretation of the net-inflow data, and its implications for simulation errors, we have made the following changes.

In the last paragraph of the Introduction (**lines 94–96**), we added the following text: "*Here "data scarcity" specifically refers to the lack of historical operational records required for site-specific training or calibration, rather than the absence of hydrological forcing.*"

In Section 4.2 (**lines 484–487**), we revised the following sentence: "*In addition, the regional model has a strong spatial generalization ability (Supplementary Fig. S7) to previously unseen reservoirs, offering a transferable baseline for reservoir operation prediction without site-specific model training and thereby supporting applications in data-scarce systems.*"

In Section 4.2 (**lines 500–511**), we also added the following paragraph: "*It is worth noting that the framework is not tied to the net inflow used in this study as a required forcing. Rather, depending on data availability and the configuration of the host hydrological model, it can be driven by either observed or simulated reservoir inflow, or by net inflow that incorporates other reservoir gains and losses (Yu et al., 2025). In the present large-sample experiments, we used the net inflow provided by Chen et al. (2025) because consistent long-term inflow records are unavailable across the evaluated reservoirs. Although this dataset-provided net inflow was derived from reservoir observations to close the water balance, it does not mathematically constrain simulated storage to follow observations; storage errors still accumulate from release errors (see Supplementary Text S4 for details). Accordingly, the spatial-generalization experiments evaluate the transferability of learned reservoir operation behavior under prescribed hydrological forcing, whereas in large-scale hydrological applications the required forcing can be supplied directly by the host model.*"

Supplementary Text S4 ("*Effect of net inflow on storage-error propagation*") was added to the Supporting Information, which provides the mathematical details explaining why the use of net inflow does not artificially constrain simulated storage to remain close to observations.

2. Equation S8 should be removed from the main text

The authors confirmed that Eq. S8 is intentionally inactive. Adding a clarifying sentence to Supplementary Text S1 is insufficient. The main text of Section 2.2 presents Eqs. S8–S11 as the Physical Knowledge Module, giving the impression that all four constraints are enforced.

I recommend restructuring Section 2.2 to clearly distinguish between constraints that are always active (Eqs. S9–S11) and those that are architecturally supported but not activated in this study (Eq. S8), with the explanation given in the main text, not the supplement.

Response: Thank you for pointing this out. We agree that the active and inactive constraints should be distinguished explicitly in the main text, while the technical details remain in the Supplementary Information. We have therefore added the following clarification in Section 2.2 (**lines 175–178**): "*Note that while the architecture supports operational release bounds (Eq. S8), this constraint is not activated in this study because reservoir-specific release limits are unavailable for the large-sample reservoir dataset.*"

3. The storage capacity constraint is not enforced at the reservoir level

This concern follows directly from point 2. The authors defend Eqs. S9–S11 as the primary physics enforcement in the absence of Eq. S8. However, inspection of the code reveals that this enforcement is not physically meaningful at the individual reservoir level.

In rnncell, storage, release, and inflow are all denormalized using the same global training set statistics:

s2_denorm = s2 * self.target_scale[1] + self.target_center[1] # global scaler

release_denorm = release_factor * self.target_scale[0] + self.target_center[0] # global scaler

inflow_d_denorm = inflow_d_raw * self.feature_scale[0] + self.feature_center[0] # global scaler

new_s2_denorm = s2_denorm - release_denorm + inflow_d_denorm

cstred_s2_denorm = new_s2_denorm.clamp(min=0, max=1)

The [0, 1] clamp cannot be in raw MCM — that would drain every reservoir above 1 MCM capacity to near-zero after one timestep. But if storage is expressed as S/S_capacity (making the clamp physically meaningful), it is in a dimensionless ratio while release and inflow remain in volumetric units, making the mass balance dimensionally inconsistent. The only self-consistent interpretation is that all variables are in a globally z-score-transformed space, in which case the clamp has no well-defined physical meaning for any individual reservoir — let alone one at the extremes of a dataset spanning 0.1 to 31,200 MCM.

The authors should clarify the actual units and normalization applied to each variable, and demonstrate that the storage capacity constraint is functioning as physically described for individual reservoirs. The code comment "storage_MCM" is inconsistent with either interpretation and should be corrected.

The good storage performance in Figure 7 does not resolve this concern. As argued in point 1, good storage simulation is largely expected by construction when net inflow is used. Furthermore, Figure 7 shows only reservoirs with above-median KGE, systematically excluding half the dataset. A random selection, or examples from the lower performance tail, would give a more complete and honest picture of model behavior. I would encourage the authors to revise this figure accordingly.

Response: Thank you for carefully inspecting the code. We apologize for the confusion caused by misleading inline comments referring to variables as being in MCM. The apparent dimensional inconsistency arises because an earlier preprocessing step was not clearly described. Before regional standardization, storage, release, and inflow were normalized by the storage capacity of each reservoir. The _denorm operations shown in the code reverse only the subsequent regional standardization; the variables therefore remain capacity-normalized when the physical constraint is applied. Accordingly, the [0,1] clamp enforces $0 \leq S_t/S_{\text{capacity}} \leq 1$, i.e., $0 \leq S_t \leq S_{\text{capacity}}$, separately for each reservoir.

We have corrected the misleading code comments and clarified this two-step normalization procedure in Section 2.2 (**lines 205–210**): "*Before model training, storage, release, and inflow were first normalized by the storage capacity of each reservoir. These capacity-normalized variables were then standardized using regional scaling factors computed from the training set, rather than per-reservoir statistics. The physical storage constraint is applied after reversing the regional standardization*

but without reversing the capacity normalization, such that the bound [0,1] corresponds to $0 \leq S_t \leq S_{\text{capacity}}$ *for each reservoir.*"

We also agree with the reviewer's suggestion regarding Fig. 7. We have revised the reservoir selection using KGE-quartile-stratified random sampling to represent the full performance distribution, including lower-performing cases. The revised **Fig. 7** is provided below:

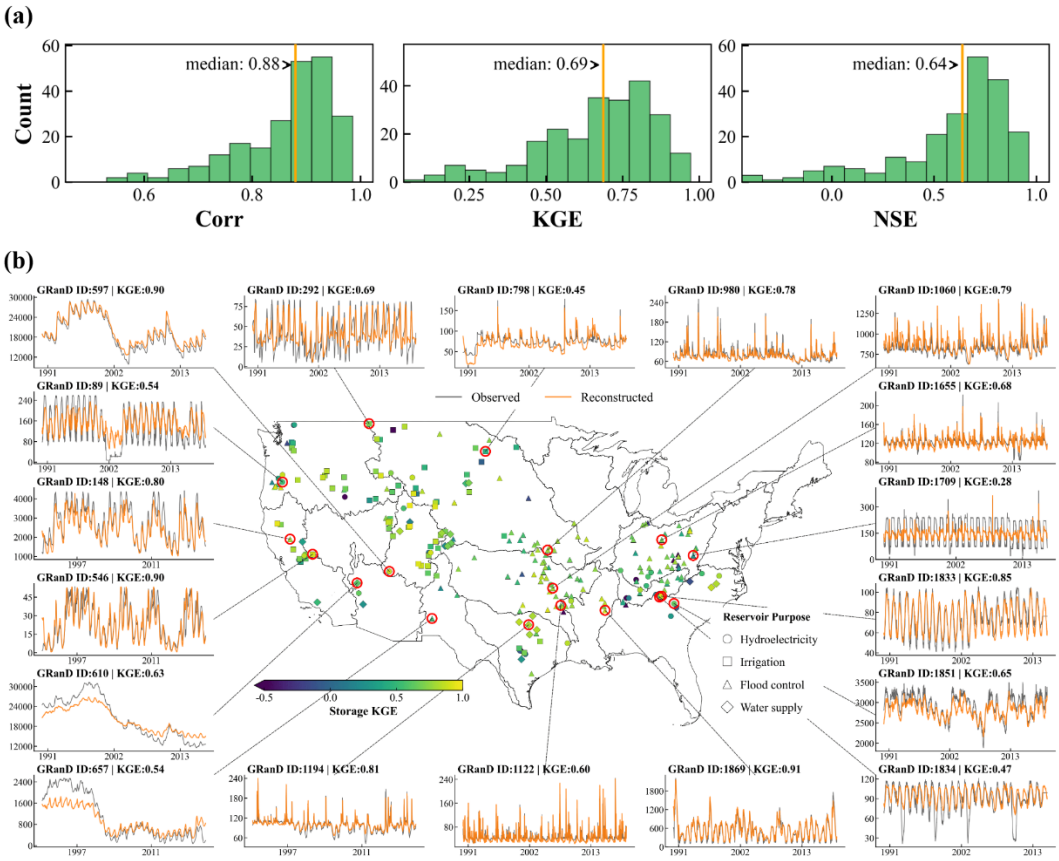


Figure 7. Evaluation of remote-sensing-augmented historical storage reconstruction (PLSTM-Reg+RS). (a) Overall reconstruction performance. Histograms show the distribution of Correlation, KGE, and NSE across the 256 U.S. reservoirs over the full data period. Orange vertical lines mark the median values. For visualization clarity, statistical outliers (values outside 1.5 × the interquartile range) are excluded. (b) Spatiotemporal analysis of the PLSTM-Reg+RS reconstruction. The central map illustrates the spatial distribution of KGE scores (color). Surrounding panels display storage time series for 18 reservoirs selected from reservoirs with storage size ratio > 0.2 using KGE-quartile-stratified random sampling covering 14 HUC2 regions. Orange and gray lines represent model estimates and in situ measurements, respectively (units: million cubic meters).

4. The storage feedback does not prevent drift independently of net inflow

The authors argue that feeding simulated storage back into the LSTM at each timestep prevents persistent drift. This is the model's own simulated output, not an observation, so it does not anchor the trajectory to reality. The model is trained end-to-end in free-run mode, which does incentivize stable trajectories within the training distribution — but as argued above, the apparent absence of drift in the reported experiments is primarily a consequence of net inflow encoding the observed storage and release, not of the recurrent storage feedback.

I would also note that this storage feedback pathway — simulated S_{t-1} concatenated into the LSTM input vector at every timestep — is not described in the main text. The paper states that in long-term simulation "only the initial storage is prescribed," which does not convey that simulated storage is also an explicit LSTM input at each subsequent step. This should be described in the methods, particularly since storage is not listed as an input variable in Table 1 for Experiments II and IV.

Response: Thank you for allowing us to clarify this mechanism. Recurrently feeding simulated storage back into the PLSTM provides an endogenous state-dependent feedback pathway. Because the simulated storage state influences the subsequent release prediction, a deviation in storage can induce corresponding changes in simulated release that may counteract further error accumulation. For example, an overestimated storage state may lead the model to predict a larger release, which in turn reduces the storage estimate at the next timestep. This feedback enables the model to respond to its evolving internal state and can mitigate the accumulation of storage errors during long-term simulation. This state-dependent feedback emerges from the recurrent PLSTM architecture and has also been demonstrated in Yu et al. (2025), where the local PLSTM model (i.e., a model trained for an individual reservoir) was driven by independently simulated inflow. The feedback mechanism therefore does not depend on the use of net inflow in the present study. Moreover, as shown in our response to Point 1, net inflow does not provide a restoring force toward observed storage and therefore cannot artificially constrain simulated storage from drifting away from observations.

We also clarify that simulated storage is not an external input at subsequent timesteps, but an internally propagated physical state of the PLSTM, analogous to the recurrent hidden and cell states h_t and c_t in a standard LSTM. Its concatenation with the external input vector is an implementation detail within the recurrent cell and does not mean that storage is externally prescribed. Thus, only the initial storage is provided in Experiments II and IV, and storage is correctly omitted from Table 1.

To avoid potential ambiguity, the following changes were made:

In **lines 178–181**, we added the following text: "*At each timestep, the updated simulated storage state* s_t *is propagated recurrently to the subsequent PLSTM cell and concatenated internally with the recurrent state* h_t*, as well as external input vector* x_t *(Fig. 2b). As* s_t *is generated by the model rather than externally prescribed, it is not treated as a dynamic input.*"

In **line 181-187**, we revised the original text and added a reference to Yu et al. (2025): "*This architecture provides an endogenous mechanism for maintaining physically feasible behavior and mitigating long-term mass-balance drift, as demonstrated in Yu et al. (2025). Because the simulated storage state influences the subsequent release prediction, a deviation in storage can induce corresponding changes in simulated release that may counteract further error accumulation. For example, an overestimated storage state may lead the model to predict a larger release, which in turn reduces the storage estimate at the next timestep.*"

Overall assessment

Taken together, these concerns point to a consistent pattern: the paper claims to enforce physical knowledge through architectural constraints, but the actual implementation does not deliver this in a physically meaningful way. The storage constraint operates in a globally-normalized space that does not correspond to per-reservoir capacity bounds. Equation S8, presented as part of the Physical Knowledge Module, is never activated. The strong performance reported across experiments is largely attributable to the use of net inflow — a quantity derived from the observed storage and release records — rather than to the physics-encoded architecture. Standard LSTMs trained on release have been shown to provide good prediction skill (Chen et al., 2026); if trained with net inflows, offline calculation of storage from simulated release would yield very similar storage performance to that reported here, but for the same structural reasons rather than through genuine physics enforcement.

The framework is genuinely interesting, and the regional learning approach has real merit. However, in its current form the paper significantly overstates both the degree of physics enforcement and the practical applicability of the results to data-scarce ungauged reservoirs. I recommend that the manuscript be substantially revised to: (i) honestly characterize the role of net inflow and its implications for the spatial generalization claim, including a justification of how net inflow could be obtained in practice by potential users of this model; (ii) remove or clearly qualify Eq. S8 in the main text; (iii) clarify the normalization scheme and demonstrate that the storage constraint is physically meaningful at the reservoir level; and (iv) assess performance using inflow that does not require observed S and R, or reframe the contribution

accordingly. Until these points are addressed, I do not believe the manuscript is ready for acceptance

Response: Thank you for the overall assessment and for recognizing the merit of the regional learning framework. The major concerns summarized here have been addressed individually in Points 1–4 above.

In addition, we directly tested the reviewer's hypothesis that a standard LSTM combined with the same net inflow and water-balance-based storage updating would yield similar storage performance. Specifically, we trained a standard LSTM using storage as an input variable to predict release, and then ran it autoregressively, updating storage at each timestep through the same reservoir water-balance equation. In the spatial cross-validation experiment, the two models showed similar overall release performance, with the same median KGE of 0.73, whereas storage performance was substantially lower for the standard LSTM baseline (median KGE = 0.46) than for PLSTM (0.59; Fig. R1). This provides direct evidence that the stronger storage performance of PLSTM cannot be explained solely by the use of net inflow and water-balance-based storage updating.

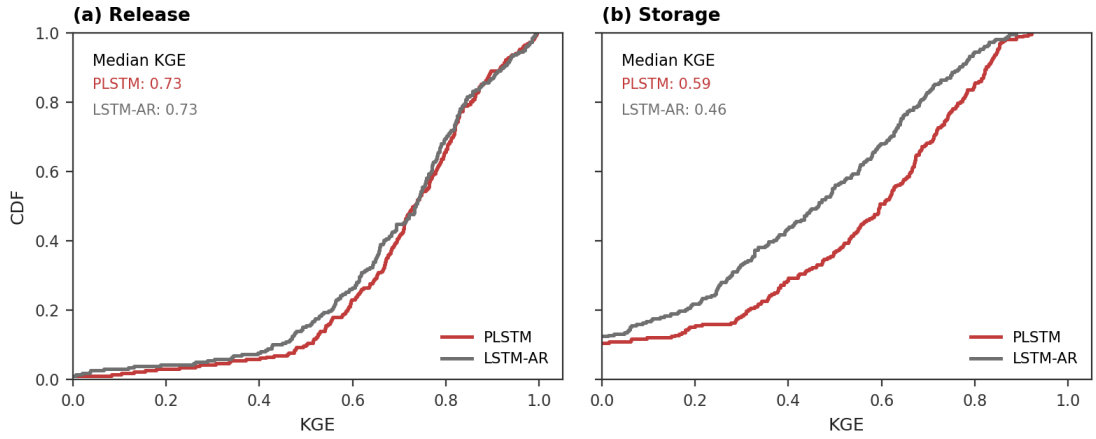


Figure R1. Comparison of PLSTM and the autoregressive standard LSTM baseline (LSTM-AR) in the spatial cross-validation experiment. Cumulative distribution functions (CDF) show KGE for (a) release and (b) storage across the held-out reservoirs. LSTM-AR predicts release and recursively updates storage using the same reservoir water-balance equation and net-inflow forcing as PLSTM.

References

Chen, Y., Cai, X., and Li, D.: Historical Operation Data of 256 Reservoirs in Contiguous United States, https://doi.org/10.4211/hs.092720588e2e4524bf2674235ff69d81, 2025.

Ford, L. and Sankarasubramanian, A.: Generalizing Reservoir Operations Using a

Piecewise Classification and Regression Approach, Water Resources Research, 59, e2023WR034890, https://doi.org/10.1029/2023WR034890, 2023.

Liu, Z. and He, X.: Balancing-oriented hydropower operation makes the clean energy transition more affordable and simultaneously boosts water security, Nat Water, 1, 778–789, https://doi.org/10.1038/s44221-023-00126-0, 2023.

Sadki, M., Munier, S., Boone, A., and Ricci, S.: Implementation and sensitivity analysis of the Dam-Reservoir OPeration model (DROP v1.0) over Spain, Geoscientific Model Development, 16, 427–448, https://doi.org/10.5194/gmd-16-427-2023, 2023.

Yu, B., Zheng, Y., He, S., Xiong, R., and Wang, C.: Physics-encoded deep learning for integrated modeling of watershed hydrology and reservoir operations, Journal of Hydrology, 657, 133052, https://doi.org/10.1016/j.jhydrol.2025.133052, 2025.