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.

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:

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

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.

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.

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.

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.

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.