I thank the authors for the detailed response as they have correctly addressed most of my comments (i.e., Eqs S8, normalization, storage feedback). However, my primary concern about the use of the net inflows remains in this revised version.

The mechanism I intended to raise is separate from the one analyzed in Text S4. In addition to entering the mass balance, $I_{net,t}$ is supplied to the network as a dynamic input (Table 1, which should list 'Net Inflows' rather than 'Inflows'). I know the authors claim that PLSTM is not tied to net inflows, but I return to this point later in my review.

Since $I_{net,t} = \left(S_{t+1}^{obs} - S_t^{obs}\right) + R_t^{obs}$, the observed release at the same timestep is recoverable from that input up to the storage increment, and a model trained to minimize release loss has a direct incentive to recover it. Small release errors then imply small storage errors through the authors' own Eq. (S20). Text S4 concerns the storage-error propagation but does not address what the network can learn from its inputs.

I accept that net inflow is not a requirement of the framework in principle. My concern is that, in the experiments reported, it plausibly accounts for a substantial share of the performance, and this has not been tested. I agree that one hydrological model could supply time series of net inflows. However, hydrology models that can predict net inflow are uncommon, and difficult to calibrate given the lack of observations of the additional components involved (e.g., reservoir evaporation, losses, unaccounted withdrawals from the reservoirs and so on).

The authors' statement that their observation-closed residual net inflows can be replaced by a host model waves away from the difficulty the framework is meant to address. An observation-closed residual is not the same quantity as a net inflow supplied either by a host model or an offline model: it has its own error structure and does not contain the observed release as a component.

The authors referred to their previous works to support their current approach, in particular Yu et al. (2025). I have read Yu et al. (2025) and found the context of this study to be very different. Because that model is not constrained to close the mass balance across the watershed, the simulated inflows are free to adjust to whatever the reservoir module needs in order to reproduce the observed releases. Note also that simulated storage does not enter the loss function in Yu et al. (2025). The cited work therefore does not demonstrate that the present framework performs comparably when driven by an independently derived inflow. Obtaining a net inflows time series offline (i.e., produced by a hydrology model calibrated or trained independently of the reservoir module) is likely even harder, since a model calibrated in that way has no observed release series to constrain it.

The revised scoping of "data scarcity" is inappropriate given the importance it takes in this research: the title itself suggests the model is designed to tackle data-scarce region. However, training a model using net inflows derived from historical operations as input does not qualify as such. I asked the authors to confirm explicitly how net inflow was obtained for the held-out reservoirs in Experiments III and IV. The response defines data scarcity as the absence of operational records rather than of forcing, but I_{net} at a held-out reservoir is computed from that reservoir's own observed storage and release, the records stipulated to be unavailable. The forcing is a function of the withheld data. In my view, this dependence likely explains a large part of the report out-of-sample performance. Again, the authors pushed back claiming that using hydrology model simulating net inflows could be used, and I agree with this statement, with the caveat I mentioned above. Net inflows models are difficult to calibrate and validate, and their simulation is not the same as an observation closed residual. If the authors are convinced that hydrology simulations of net inflows would provide similar performance, I certainly encourage them to demonstrate this claim.

In a new paragraph in Section 4.2, the authors mentioned that “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.” This is not accurate. The ResOpsUS dataset provides such a dataset (Steyaert al. 2022). In ResOpsUS dataset, 241, 158, 94 and 48 reservoirs have respectively at least 10, 20, 30 or 50 years of continuous time series of inflows, outflow and storage. Note that ResOpsUS was used by Chen et al. (2025), in addition to Li et al. (2024), to create the net inflows dataset used in this study. Chen et al. (2025) selected reservoirs with at least 25 years of data, which leaves over 120 reservoirs with continuous inflows, outflows and storage.

Steyaert, J.C., Condon, L.E., W.D. Turner, S., Voisin, N., 2022. ResOpsUS, a dataset of historical reservoir operations in the contiguous United States. Sci. Data 9. https://doi.org/10.1038/s41597-022-01134-7

Li, D., Chen, Y., Lyu, L., & Cai, X. (2024). Uncovering historical reservoir operation rules and patterns: Insights from 452 large reservoirs in the contiguous United States. Water Resources Research, 60(8), e2023WR036686. https://doi.org/10.1029/2023WR036686

Something the authors fail to discuss in their manuscript is that using net inflows input other than the observed-closed residual used in this study is likely to yield significantly lower performance than advertised here. This is a very important issue this revision has not addressed. Along those lines, I find it very misleading that all elements in the abstract, discussion and conclusion sections gloss over the difficulty at i) obtaining net inflows that are not derived from operational records and ii) applying PLSTM with forcings that do not embed the target variables.

Last, regarding the Figure R1: what I asked for was a baseline in which the physical knowledge layer is disconnected: a standard LSTM taking the same input as PLSTM (i.e., net inflow and the meteorological forcings as inputs) and predicting release only, with no storage input and no storage correction during the run. Storage is then reconstructed offline by integrating the water balance from the simulated release series. The description given for LSTM-AR does not allow me to determine how close it is to this. It is not stated whether net inflow was retained as an input, whether the storage supplied during training was observed or simulated, or whether any bound was applied to storage at inference. These details determine what the comparison shows, and I would ask that they be specified.

Setting that aside, the result itself is surprising. In PLSTM, the clamp in Eq. (S10) forces release to be recomputed through Eq. (S11) whenever the candidate storage would leave the physical bounds, so the emitted release is the one mass balance requires rather than the one the network would otherwise produce. A model carrying that correction should be *less* accurate on release than one optimizing release without it, and the expected signature of the physics encoding is therefore a modest loss in release accuracy purchasing a substantial gain in storage. Figure R1 instead shows PLSTM identical release medians, or even better, with a distribution that appears somewhat better across much of the range; and significantly better on storage. The physical knowledge layer appears to cost nothing anywhere, which is not what the mechanism predicts. I would ask the authors to explain what release accuracy the physical constraint costs in their framework, and why Fig. R1 does not show it.

To conclude, I find the framework novel and the regional learning approach genuinely promising, and I hope the authors continue in this direction. As presented, however, the manuscript does not meet the challenge named in its own title. The forcing contains the target variables, which a trained network can readily decode; the difficulty of obtaining a net inflow series without operational records is not acknowledged; and the claimed applicability to data-scarce reservoirs therefore rests on data those reservoirs would not have. Since resolving this requires re-running the central experiments with an independently derived inflow, the outcome of which would determine the paper's principal conclusions, I do not think it can be settled within the scope of a revision, and I recommend rejection. I would encourage the authors to resubmit once those experiments are complete: the ResOpsUS records identified above appear to make them feasible with the existing reservoir sample.