Dear editors and reviewers,

Thank you very much for taking the time to review our manuscript and for providing valuable and constructive feedback. We greatly appreciate your help in improving our work on PLSTM-Reg v1.0. We have carefully revised the manuscript and the supplement following your suggestions. Below, we provide point-by-point replies to each comment. The original comments are shown in blue, and our responses are shown in black. The line numbers are consistent with those in the clean (i.e., no changes tracked) version of the revised manuscript.

Yi Zheng, on behalf of the author team

Response to Reviewer #1's comments

[0] Greetings! This study presents PLSTM-Reg v1.0, a regional physics-encoded LSTM framework that effectively simulates reservoir operations by integrating mass-balance constraints with deep learning. The methodology demonstrates strong spatial and temporal generalizability, particularly through its synergy with remote sensing data. Overall, I think this is a very high-quality, well-organized study that addresses a critical challenge in hydrological modeling, which fully meets the requirements for publication in *Geoscientific Model Development* (GMD).

Response: Thank you for your positive and encouraging feedback. We have carefully addressed your comments and suggestions to further improve the clarity of the manuscript, as detailed below.

[1] I have two minor suggestions. First, I recommend that the authors move the descriptions of static attributes to the main text, as they are essential parts of the model structure.

Response: Thank you for your suggestion. We have revised the description of the 21 static attributes (previously Table S1 in the Supplement) and moved it to **Table A1 in Appendix A** of the manuscript. This improves their visibility and accessibility while preserving the flow of the Data and Methods section.

[2] Second, I would like to see a little discussion on which of these attributes have the most important impacts on the model results and which are less so.

Response: Following your suggestion, we conducted a feature attribution analysis using the Integrated Gradients (IG) method. Details of the IG methodology and the feature-importance results are provided in Supplementary **Text S3** and **Fig. S6**, respectively. In the main text (**lines 335-340**, Section 3.1), we now state:

"*Furthermore, we conducted a feature attribution analysis using the Integrated Gradients (Sundararajan et al., 2017) method (see Supplementary Text S3). The analysis shows that the model relies primarily on operational parameters (e.g., fill fraction, reservoir functions, and storage size ratio) together with regional topographic and hydroclimatic factors (e.g., elevation, precipitation, and temperature), whereas geometric variables (e.g., dam length, depth, and surface area) exert a comparatively minor influence (see Fig. S6).*"

The new Fig. S6 is also provided below:

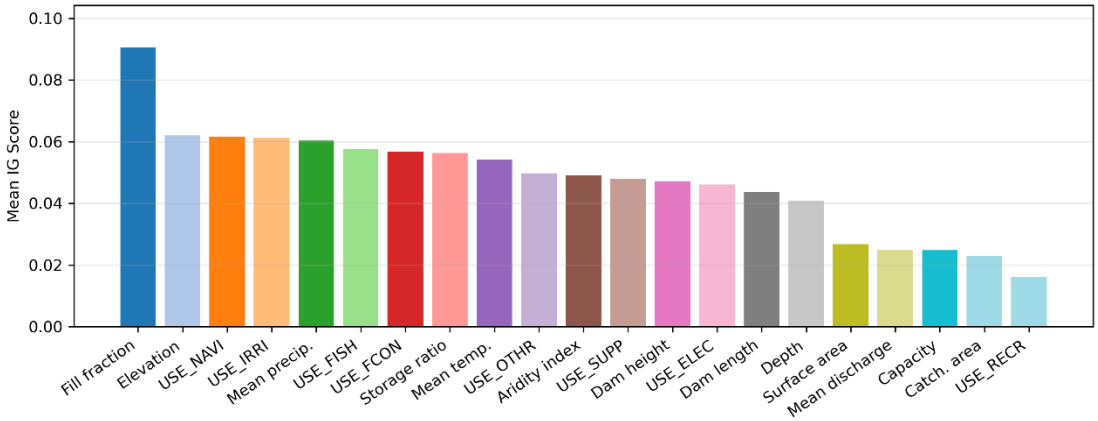


*Figure S6. Mean Integrated Gradients (IG) importance scores of the 21 static attributes (see **Table A1** in Appendix A) for reservoir release predictions. IG scores were normalized within each reservoir-day and then averaged across reservoirs, dates, and ensemble repeats. Higher values indicate greater relative contribution to model predictions.*

[3] I also have three technical questions that I hope the authors could clarify to improve reproducibility. 1) What is the specific sequence length used for the training and testing of the sequence-to-sequence regional model?

Response: Thank you for this question. Different sequence lengths were used for short-term prediction and long-term simulation experiments. To clarify this and ensure

reproducibility, we have added corresponding details in Section 2.3.

In **lines 219-221**, the original sentence was revised as: "*In short-term prediction (Experiment I), PLSTM models were trained in a sequence-to-one manner with a sequence length of 180 days to predict next-day release and storage.*"

In **lines 226-229**, we added: "*The sequence length for both model training and testing was set to 1460 days (4 years). A sensitivity analysis using 2-, 4-, and 6-year sequence lengths yielded similar performance across all tested settings (Supplementary Fig. S1), indicating that the results are not strongly sensitive to this choice.*"

[4] 2) Why did the authors choose a recursive sequence-to-one model instead of a direct sequence-to-sequence approach (e.g., outputting a 7-day vector) for the 7-day lead time forecast?

Response: Thank you for this insightful question. We adopted the recursive sequence-to-one architecture for two reasons. First, it is the standard approach widely used in deep learning hydrology (e.g., Feng et al., 2020; Jiang et al., 2022; Kratzert et al., 2018). Second, and more importantly, it is required by the model structure. As shown in Table 1, reservoir storage is used as a dynamic input feature in Experiments I and III. Therefore, predicting the state at day $t+1$ requires the storage at day t, making autoregressive recursive prediction necessary for multi-day prediction.

To clarify this design choice, we added the following text in **lines 222-224**: "*This structure follows standard hydrological prediction practice (Dong et al., 2023a; Kratzert et al., 2019) and is required because the model uses storage at day t to predict release and storage at day t+1.*"

[5] 3) Did the authors encounter gradient vanishing or truncation issues, given that the hard physical constraints in the PLSTM cell might truncate the gradient flow during backpropagation?

Response: Thank you for this insightful question. Preliminary experiments show that PyTorch's native clamp() function remained stable during backpropagation and did not hinder model optimization, despite the possibility of gradient truncation when constraints are exceeded. Similar implementations are also adopted in recent differentiable hydrology frameworks such as hydrodl2 (mhpi, 2026). We therefore used clamp() for its simplicity and exact enforcement of physical bounds.

To clarify this point, we added the following text in **lines 177-181** of Section 2.2: "*It is worth noting that although hard physical bounds implemented via the torch.clamp function may truncate gradients when constraints are exceeded, preliminary experiments showed stable optimization without requiring smoothing techniques (e.g., tanh smoothing). Similar implementations are also adopted in recent differentiable hydrology frameworks such as hydrodl2 (mhpi, 2026)*"

References

Dong, N., Guan, W., Cao, J., Zou, Y., Yang, M., Wei, J., Chen, L., and Wang, H.: A hybrid hydrologic modelling framework with data-driven and conceptual reservoir operation schemes for reservoir impact assessment and predictions, Journal of Hydrology, 619, 129246, https://doi.org/10.1016/j.jhydrol.2023.129246, 2023a.

Feng, D., Fang, K., and Shen, C.: Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales, Water Resources Research, 56, e2019WR026793, https://doi.org/10.1029/2019WR026793, 2020.

Jiang, S., Zheng, Y., Wang, C., and Babovic, V.: Uncovering Flooding Mechanisms Across the Contiguous United States Through Interpretive Deep Learning on Representative Catchments, Water Resources Research, 58, https://doi.org/10.1029/2021WR030185, 2022.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, Water Resour. Res., 55, 11344–11354, https://doi.org/10.1029/2019WR026065, 2019.

mhpi: hydrodl2, GitHub [code], available at: https://github.com/mhpi/hydrodl2 (last access: 5 May 2026), 2026.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic Attribution for Deep Networks, arXiv:1703.01365 [cs], 2017.

[0] The manuscript addresses a critical challenge in large-scale hydrological modeling: the generalization of reservoir operation rules to data-scarce regions. The authors propose PLSTM-Reg v1.0, a physics-encoded Long Short-Term Memory (LSTM) model that integrates reservoir storage directly into the LSTM cell to enforce mass conservation. The model is evaluated using 256 reservoirs across the Continental United States (CONUS) through five distinct experimental setups designed to test temporal generalization, spatial generalization, and historical data reconstruction. Overall, the manuscript presents an innovative approach to reservoir modeling. The initial results are promising; however, there are several major concerns regarding the experimental logic, the clarity of the evaluation metrics, and the positioning of the work within the existing literature that need to be addressed.

Response: Thank you for your positive and encouraging feedback. We have carefully considered and addressed your comments and suggestions to improve the manuscript, as detailed below.

Major Comments

[1] Logical Consistency in Experiment Comparison (IV vs. V): The authors conclude that incorporating remotely sensed (RS) surface area leads to superior model performance compared to local models (mentioned in places like lines 23-26 and lines 313-343). However, the logic underpinning this comparison in Table 1 is not fully sound.

Specifically, in Experiment IV, storage information is omitted. In Experiment V, remotely sensed surface area, which serves as a highly informative proxy for storage, is included. Given that Experiment V introduces a significant predictive variable that Experiment IV lacks, it is not a safe conclusion that V would outperform IV because of adding RS information. This comparison does not necessarily prove the superiority of RS integration in reconstructing historical records; rather, it highlights the value of the extra information provided. To make this claim robust, the authors must clarify if the "reconstruction" baseline is intended to represent a scenario where NO storage records exist. If storage data were available and utilized in IV, would the RS approach still be superior?

Response: Thank you for this insightful comment. Experiment V was intended to represent a historical reconstruction scenario in which no ground-based storage records are available, relying entirely on satellite-derived surface area as a proxy for storage information. We agree that directly comparing Experiment IV and Experiment

V is not fully appropriate, because Experiment V introduces an additional source of information that is absent in Experiment IV.

To clarify this point, we have revised the manuscript accordingly. Specifically,

In **Table 1**, we relabeled Experiment V as "Historical reconstruction" instead of "long-term simulation" to better reflect its intended application scenario.

In Section 3.3, we removed the direct benchmarking statements between Experiments IV and V to avoid overstating the advantage of incorporating the remote sensing proxy.

We revised the abstract (**lines 26-28**) as: "*In historical storage reconstruction, incorporating monthly satellite-derived surface area as a proxy for missing ground-based records enables reconstruction accuracy comparable to local models.*"

We revised Fig. 7a and Fig. S10 by removing the direct comparison between Experiments IV and V, and now present only the results of Experiment V.

We also removed the original Fig. S6.

[2] Justification of Model Input for Long-Term Simulations: Following the point above, why is storage omitted from the long-term simulation setups? In many operational contexts, historical storage records (even if fragmented) are the primary benchmark. The authors should justify the decision to exclude storage in these specific experiments or discuss the implications of this omission for the model's practical utility.

Response: Thank you for this insightful question. Storage was intentionally excluded from the long-term simulation setups (Experiments II and IV) because these experiments are designed for integration with Large-scale Hydrological Models (LHMs) in future climate change projection scenarios. In such applications, future operational storage records are inherently unavailable and therefore cannot be used as model inputs.

By contrast, the short-term prediction experiments (Experiments I and III) are intended for real-time operational applications, where storage observations are available. The different experimental configurations were therefore designed to reflect the data availability and practical requirements of these distinct application scenarios.

To clarify this rationale, we added the following explanation to Section 2.3 (**lines 208-211**): "*Specifically, short-term prediction (Experiments I and III) supports real-time*

operational decision-making, where storage observations are typically available, while long-term simulation (Experiments II and IV) is designed for coupling with Large-scale Hydrological Models (LHMs) to assess future climate change impacts, where operational records do not exist."

[3] Review of Relevant Literature: The literature review (Lines 58-72) summarizes recent regionalization efforts (e.g., Turner et al., 2021; Steyaert et al., 2025) but misses several foundational or recent studies that utilize machine learning and remote sensing for parameter generalization. Including the following would provide a more comprehensive context:

- **DZTR Model:** Utilizing hydrological quantiles for straightforward generalization (Yassin et al., 2019).
    - https://doi.org/10.5194/hess-23-3735-2019
- **SBTS Model:** Generalizing storage-based piecewise rules using ML and satellite observations (Shen et al., 2025).
    - https://doi.org/10.1029/2024WR037620
- **MODROM Model:** A modular model using ML for generalization (Li & Villarini, 2026).
    - https://doi.org/10.1029/2025MS005180

Response: Thank you for recommending these important references. We have revised the Introduction to provide a more comprehensive discussion of recent advances in rule-based reservoir operation modeling and machine-learning-assisted parameter generalization. The revised text (**lines 47-77**) now reads:

"*Rule-based reservoir operation models (Hanasaki et al., 2008a, b; Wisser et al., 2010; Zajac et al., 2017) are widely used in large-scale hydrological studies because of their interpretability and modest data requirements. However, their fixed recommended parameters and simplified operating rules limit their ability to represent site-specific and nonlinear real-world operating behavior (Steyaert and Condon, 2024; Yang et al., 2021). Some progress has been made in recent studies. Yassin et al. (2019) proposed an improved target storage-and-release parameterization that better represents reservoir storage zones and seasonal release behavior, but it requires in-situ storage and release series (or distributions) to determine model parameters. Turner et al. (2021) developed a transferable rule-based*

framework that learns operating polices from data-rich reservoirs and extrapolate them to neighboring data-scarce reservoirs with similar purposes. However, directly transferring the same policy parameters from source reservoirs to target reservoirs may impose source-specific operating behavior and introduce significant bias.

Recently, machine learning has been introduced to improve the parameterization and transferability of rule-based reservoir operation schemes. Some studies have used machine learning to estimate parameters for existing operation schemes, such as reservoir-operation parameters inferred from reservoir attributes using Random Forest models (Steyaert et al., 2025) and flood storage capacities estimated using XGBoost to define storage-zone parameters in a target storage scheme (Shen et al., 2025). Researchers have also explored to use machine learning to develop hybrid models for improved reservoir representation. (Chen et al., 2022) developed a data-driven framework that extracts interpretable operation modules and associated application conditions from historical operation records. Building on this modular reservoir modeling concept, Li and Villarini (2026) further investigated its generalization using Random Forest models to transfer parameters associated with typical operation modules, such as constant, linear, and piecewise-linear release relationships, across reservoirs. Although these approaches improve parameter estimation and transferability while retaining interpretability, they generally represent reservoir operations through simplified modules or predefined rules. Moreover, because these modules or parameters are typically identified from individual reservoirs before being transferred, these approaches cannot fully leverage cross-site information during model training, which has been shown to improve the generalization of regional deep learning models (Fang et al., 2022; Kratzert et al., 2024). Consequently, their ability to learn complex operating behavior across diverse reservoirs may remain constrained."

[4] Clarity on Evaluation Metrics and Performance (Lines 214-229): The KGE scores presented in Figure 3 are exceptionally high, with medians near 1.0. This level of "perfection" suggests the evaluation might be influenced by the model's short-term configuration. Are these scores calculated on the training, testing, or overall dataset? If the model utilizes lagged information for next-day release predictions, the high performance is likely a reflection of persistence rather than operational rule learning. The authors need to provide a more detailed description of the evaluation: specifically, define the ground truth, the simulation period, and the exact lead time for the releases being scored.

Response:

Thank you for this insightful question. We agree that near-perfect KGE at a 1-day lead time may partly reflect short-term persistence.

To clarify, all KGE scores in Figure 3 are calculated on the unseen testing period (2010 onward), using observed daily releases as the ground truth. While the model uses the previous day's storage (S_{t-1}), **it does not use the previous day's release (R_{t-1}) as an input**.

We agree that persistence likely contributes to the extremely high skill at short lead times. However, persistence alone is unlikely to fully explain the model performance. If the forecasts were driven primarily by release persistence, a rapid decline in performance would be expected as lead time increases. Instead, Figure 3d shows that PLSTM-Reg maintains relatively high KGE across longer lead times, suggesting that the model exploits information contained in the predictor variables beyond simple persistence and captures aspects of reservoir operational behavior.

We added the following clarification in Section 3.1 (**lines 276-279**): "*While the superior performance at the 1-day lead time may partly reflect the natural short-term persistence of reservoir systems, the model's ability to maintain high KGE at longer lead times suggests that it captures information relevant to reservoir operations beyond simple persistence.*"

Minor Comments

Line 20: The term "local models" is used without a prior definition. Please provide a brief explanation at the first mention to ensure clarity for readers unfamiliar with the authors' specific terminology.

Response:

Thank you for pointing this out. To improve clarity, we revised the sentence in the abstract (**line 17-20**) as follows: "*Under temporal testing, the regional model improves 1-day-ahead release forecasts from a median Kling–Gupta Efficiency (KGE) of 0.83 to 0.96 relative to local counterparts (i.e., models trained exclusively on site-specific records), …*"

Line 27: The hybrid nature of the model (Physics + DL) is a highlight of the work but is under-represented in the Abstract. Please mention the specific physical constraints

(e.g., mass balance) earlier in the abstract to emphasize the "Physics-encoded" aspect.

Response: Thank you for this suggestion. To emphasize the hybrid nature, we revised sentence in the abstract (**lines 12-14**) as follows: "*Here, we develop PLSTM-Reg v1.0, a regional physics-encoded deep learning framework that explicitly enforces mass balance and operational constraints to simulate reservoir operations across diverse systems.*"

Line 110: Please add a formal citation for the Daymet dataset.

Response: Thank you for pointing this out. We added citations for the Daymet dataset (Thornton et al., 2022, 2021) in **line 125**.

Lines 119-120: When discussing "reconstructing records from remote sensing," please specify which variables are being reconstructed (e.g., inflow, storage, or release).

Response: Thank you for this suggestion. Our framework reconstructs both storage and release records. We clarified this in **lines 146-147** as follows: "*In addition, to evaluate the contribution of remote sensing data to reconstructing operation records (specifically storage and release) for data-scarce reservoirs*"

Line 123: Regarding the use of SARAH-CONUS to supplement GRSAD: Did the authors perform a consistency check or bias correction between these two datasets? Please clarify if there were systematic differences and how they were handled.

Response: Thank you for this important question. After re-examining our preprocessing workflow, we discovered that the reference to SARAH-CONUS in the manuscript was inaccurate. Although we initially considered using SARAH-CONUS to supplement post-2018 data, we ultimately decided not to use it in order to avoid potential cross-product biases between datasets, as you anticipate.

Instead, for periods beyond the GRSAD coverage (post-2018), we filled missing values using the long-term monthly mean surface area derived from the GRSAD climatology. Because the post-2018 period represents only a small fraction of the full record, we believe this approach provides a conservative and internally consistent treatment.

We therefore revised **lines 146-156** to remove the erroneous reference to SARAH-CONUS and clarify the actual gap-filling procedure, which reads "*In addition, to*

evaluate the contribution of remote sensing data to reconstructing operation records (specifically storage and release) for data-scarce reservoirs, we incorporated monthly satellite-derived surface area time series from the Global Reservoir Surface Area Dataset (GRSAD; Zhao and Gao, 2018), which provides data for 6,817 reservoirs listed in the GRanD database from 1984 to 2018. Prior to imputation, the raw monthly GRSAD time series across the evaluated reservoirs had an average missing-data ratio of 3.7%, with reservoir-level values ranging from 0% to 6.6%. For missing values within the GRSAD record, data were imputed according to gap length: gaps of six months or shorter were linearly interpolated, whereas longer gaps were filled with the long-term monthly mean surface area. For periods extending beyond the GRSAD coverage (post-2018), the same climatological monthly mean was used to maintain dataset consistency and avoid cross-product biases."

Line 183: The term "PLSTM-Loc" appears abruptly. Please define this counterpart model before using its abbreviation.

Response:

Thank you for pointing this out. We have revised Section 2.2 to define the abbreviation. The revised sentence (**lines 182-183**) reads: "*For comparison, we also implemented local PLSTM models (PLSTM-Loc) trained separately for each reservoir to assess the value of regional information.*"

Equations (1)-(3): There are inconsistencies in the fonts used for variables and operators. Please ensure all LaTeX formatting is uniform throughout the manuscript.

Response: Thank you for this suggestion. We updated Eqs. (1)-(3) to ensure consistent LaTeX formatting for variables and operators throughout the manuscript.

References

Chen, Y., Li, D., Zhao, Q., and Cai, X.: Developing a generic data-driven reservoir operation model, Advances in Water Resources, 167, 104274, https://doi.org/10.1016/j.advwatres.2022.104274, 2022.

Fang, K., Kifer, D., Lawson, K., Feng, D., and Shen, C.: The Data Synergy Effects of Time-Series Deep Learning Models in Hydrology, Water Resources Research, 58,

e2021WR029583, https://doi.org/10.1029/2021WR029583, 2022.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An integrated model for the assessment of global water resources – Part 1: Model description and input meteorological forcing, Hydrology and Earth System Sciences, 12, 1007–1025, https://doi.org/10.5194/hess-12-1007-2008, 2008a.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An integrated model for the assessment of global water resources – Part 2: Applications and assessments, Hydrology and Earth System Sciences, 12, 1027–1037, https://doi.org/10.5194/hess-12-1027-2008, 2008b.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin, Hydrology and Earth System Sciences, 28, 4187–4201, https://doi.org/10.5194/hess-28-4187-2024, 2024.

Shen, Y., Yamazaki, D., Pokhrel, Y., and Zhao, G.: Improving Global Reservoir Parameterizations by Incorporating Flood Storage Capacity Data and Satellite Observations, Water Resources Research, 61, e2024WR037620, https://doi.org/10.1029/2024WR037620, 2025.

Steyaert, J. C. and Condon, L. E.: Synthesis of historical reservoir operations from 1980 to 2020 for the evaluation of reservoir representation in large-scale hydrologic models, Hydrology and Earth System Sciences, 28, 1071–1088, https://doi.org/10.5194/hess-28-1071-2024, 2024.

Steyaert, J. C., Sutanudjaja, E., Bierkens, M., and Wanders, N.: A data derived workflow for reservoir operations for simulating reservoir operations in a global hydrologic model, EGUsphere, 1–38, https://doi.org/10.5194/egusphere-2024-3658, 2025.

Thornton, M. M., Shrestha, R., Wei, Y., Thornton, P. E., and Kao, S.-C.: Daymet: Daily Surface Weather Data on a 1-km Grid for North America (4.1), https://doi.org/10.3334/ORNLDAAC/2129, 2022.

Thornton, P. E., Shrestha, R., Thornton, M., Kao, S.-C., Wei, Y., and Wilson, B. E.: Gridded daily weather data for North America with comprehensive uncertainty quantification, Sci Data, 8, 190, https://doi.org/10.1038/s41597-021-00973-0, 2021.

Turner, S. W. D., Steyaert, J. C., Condon, L., and Voisin, N.: Water storage and release policies for all large reservoirs of conterminous United States, Journal of Hydrology, 603, 126843, https://doi.org/10.1016/j.jhydrol.2021.126843, 2021.

Wisser, D., Fekete, B. M., Vörösmarty, C. J., and Schumann, A. H.: Reconstructing 20th century global hydrography: a contribution to the Global Terrestrial Network-Hydrology (GTN-H), Hydrology and Earth System Sciences, 14, 1–24,

https://doi.org/10.5194/hess-14-1-2010, 2010.

Yang, T., Zhang, L., Kim, T., Hong, Y., Zhang, D., and Peng, Q.: A large-scale comparison of Artificial Intelligence and Data Mining (AI&DM) techniques in simulating reservoir releases over the Upper Colorado Region, Journal of Hydrology, 602, 126723, https://doi.org/10.1016/j.jhydrol.2021.126723, 2021.

Yassin, F., Razavi, S., Elshamy, M., Davison, B., Sapriza-Azuri, G., and Wheater, H.: Representation and improved parameterization of reservoir operation in hydrological and land-surface models, Hydrology and Earth System Sciences, 23, 3735–3764, https://doi.org/10.5194/hess-23-3735-2019, 2019.

Zajac, Z., Revilla-Romero, B., Salamon, P., Burek, P., Hirpa, F. A., and Beck, H.: The impact of lake and reservoir parameterization on global streamflow simulation, Journal of Hydrology, 548, 552–568, https://doi.org/10.1016/j.jhydrol.2017.03.022, 2017.

Zhao, G. and Gao, H.: Automatic Correction of Contaminated Images for Assessment of Reservoir Surface Area Dynamics, Geophysical Research Letters, 45, 6092–6099, https://doi.org/10.1029/2018GL078343, 2018.

Response to Reviewer #3's comments

[0] This manuscript develops PLSTM-Reg, a regional LSTM model trained and evaluated using long-term daily records from 256 reservoirs across the CONUS to simulate reservoir release and storage under temporal and spatial generalization settings. The study is valuable because it uses a high-quality reservoir dataset, considers both short-term and long-term simulation performances, designs multiple experiments to test transferability, and tests the added value of remote-sensing surface area. The manuscript is generally well organized. However, the novelty relative to recent pooled or transferable reservoir-operation models is fully stated, and several aspects of the validation require clarification, including possible information leakage from static attributes, and the fairness of local-model comparisons. Overall, I find the study promising and potentially useful for large-scale hydrologic modeling, but I recommend major revision before publication.

Response: Thank you for the positive assessment and constructive comments. We have carefully revised the manuscript, as detailed below.

Major Comments

[1] Table S1 lists "Average discharge" as an input. This raises two concerns. First, it may partly explain the improvement of the regional model over local models, since local models do not use this static input. Providing an ablation excluding average discharge can be more convincing. Second, in the spatial-transfer experiment, average discharge for held-out reservoirs may create leakage. This would weaken the claim that the model is tested on truly "unseen" or data-scarce reservoirs.

Response:

Thank you for raising this important concern. The "average discharge" attribute is not derived from the reservoir operation records. It is the long-term (1971–2000) average discharge at the dam location, derived from the HydroSHEDS flow-routing scheme combined with downscaled WaterGAP runoff estimates at 15 arc-second resolution (Lehner et al., 2024). Therefore, this variable does not use historical release or storage records from the held-out reservoirs and does not introduce leakage from the target operation data. Moreover, long-term discharge descriptors are commonly used in reservoir operation modeling and parameterization. For example, the storage size ratio used by rule-based benchmark models is calculated from storage capacity and long-term inflow or discharge (Hanasaki et al., 2008a, b; Yassin et al., 2019). Thus, using a globally available long-term discharge descriptor is consistent with the information typically available for data-scarce reservoir modeling.

To clarify this point, we revised section 2.1 in **lines 133-135** as follows: "*For each reservoir, we compiled static attributes from the GDW database, including physical characteristics (e.g., dam height, storage capacity), primary management purpose, and hydrological context (i.e., long-term mean discharge at the dam location during 1971–2000).*"

We also performed an ablation analysis excluding the GDW mean discharge attribute from the static inputs. As shown in the figure included below, removing this attribute does not substantially change the overall performance of PLSTM-Reg, although a slight performance decrease is observed for some lower-performing reservoirs. These results indicate that the performance advantage of PLSTM-Reg over local models cannot be attributed solely to the inclusion of the mean discharge attribute. Instead, the attribute provides useful hydrological context in some challenging cases while having a limited influence on overall model skill. Because this analysis was conducted specifically in response to the reviewer's concern and does not materially affect the study conclusions, we present the results here for completeness rather than

incorporating them into the revised manuscript.

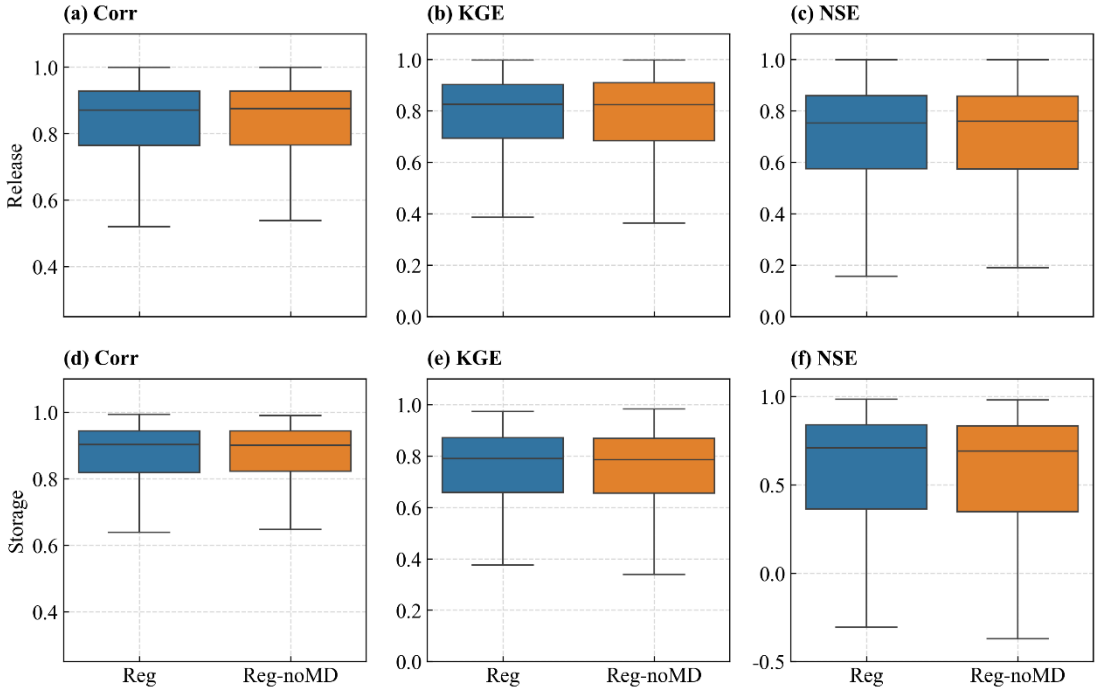


Figure R1. Ablation analysis of the GDW mean discharge attribute in long-term simulation. Boxplots compare release (a–c) and storage (d–f) simulation performance between PLSTM-Reg and PLSTM-Reg-noMD under Experiment II. PLSTM-Reg-noMD denotes the ablated regional model excluding the GDW mean discharge attribute. Removing this attribute causes no substantial change in overall performance, although some lower-performing reservoirs show slight degradation. This indicates that the GDW mean discharge attribute provides useful hydrological context for difficult cases.

[2] The model seems to assume that reservoir operation depends only on local basin forcings, local inflow/storage, and reservoir attributes. However, many real reservoirs are operated as part of cascade reservoir systems. The authors should clarify whether such cases are included and whether PLSTM-Reg can handle coupled upstream–downstream operation (such as the flexibility to add upstream reservoir information as static input in current framework). If not, the application boundary should be stated more clearly.

Response: Thank you for this insightful comment. The 256-reservoir dataset includes reservoirs located within cascading systems. In the current PLSTM-Reg framework, upstream operational effects are implicitly reflected in the inflow received by downstream reservoirs. Our study focuses on simulating the operation of individual

reservoirs rather than explicitly modeling coordinated cascade operations. For applications involving multi-reservoir systems, PLSTM-Reg can be coupled with a river routing module, whereby releases from upstream reservoirs are routed and used as inflows to downstream reservoirs. In this way, the framework can be readily extended to simulate cascade reservoir systems while retaining the same reservoir operation module for each reservoir.

To clarify this, we added the follow discussion in **lines 490-494**: "*Alongside geographical expansion, future development may also extend the framework to cascading reservoir systems. The current PLSTM-Reg framework is designed for single-reservoir simulation, with upstream operational effects represented implicitly through downstream inflows. Extension to multi-reservoir systems could be achieved by coupling PLSTM-Reg with a river routing module that routes upstream releases to downstream reservoirs.*"

[3] The comparison with recent advances is incomplete. Comparing mainly with older rule-based models weakens the significance of the claimed innovation, especially because the key claim is multi-reservoir pooling/regional learning. The following studies should be compared with, or at least discussed:

Tran et al. (2025), "Improving the prediction of daily reservoir releases over the CONUS using conditioned LSTM." This is a closed predecessor: a pooled/conditioned LSTM across nearly 200 CONUS reservoirs using static reservoir attributes.

The following three are cited in this work, but only as background. They should be discussed more directly against PLSTM-Reg as baselines:

- Ford and Sankarasubramanian (2023), "Generalizing reservoir operations using a piecewise classification and regression approach."
- Turner et al. (2021), "Water storage and release policies for all large reservoirs of conterminous United States."
- Chen et al. (2022), "Developing a generic data-driven reservoir operation model."

Response: Thank you for this suggestion. We revised the manuscript to compare PLSTM-Reg more directly with recent regional or transferable reservoir-operation models, including Tran et al. (2025), Ford & Sankarasubramanian (2023), Chen et al.

(2022), and Turner et al. (2021).

We agree that Tran et al. (2025) is closely related to the present study. However, the two studies address fundamentally different objectives. Tran et al. (2025) focuses on next-day release prediction and relies on observed storage inputs without enforcing mass balance, whereas PLSTM-Reg is designed for both short-term prediction and long-term release–storage simulation through an explicit storage state and mass-balance constraint. Therefore, Tran et al. (2025) is discussed in the Introduction but is not used as a direct benchmark for the long-term simulation experiments.

Ford and Sankarasubramanian (2023) was not included as a direct benchmark in Experiment IV because its PLRT formulation requires reservoir-specific statistical information for storage and release standardization, which in turn requires historical observations from the target reservoir. This requirement is inconsistent with the held-out, data-scarce spatial-transfer setting considered in Experiment IV.

Similarly, Chen et al. (2022) was not included as a direct benchmark because its module extraction and application-condition classification rely on historical inflow, storage, and release records from the target reservoir. As such, it is not directly applicable to the Experiment IV setting, where no operational records are assumed to be available for the target reservoir.

Following the reviewer's suggestion, we added Turner et al. (2021) as an explicit benchmark in both temporal and spatial long-term simulation experiments. For Experiment II, ISTARF was calibrated using the training period of each reservoir and evaluated on the chronological test period; this comparison is now included in **Supplementary Fig. S4** and mentioned in the main text. For Experiment IV, ISTARF was implemented in a spatial-transfer setting: parameters were fitted only using reservoirs in the training folds, and donor selection for each held-out reservoir was restricted to the training pool to avoid information leakage. The implementation details for both ISTARF settings are provided in **Supplementary Text S2**, and **Fig. 5** has been updated to include ISTARF.

To better contextualize our model and differentiate our approach from existing studies, we substantially revised the Introduction. The revised text (**lines 47-95**) now reads:

"*Rule-based reservoir operation models (Hanasaki et al., 2008a, b; Wisser et al., 2010; Zajac et al., 2017) are widely used in large-scale hydrological studies because of their interpretability and modest data requirements. However, their fixed recommended parameters and simplified operating rules limit their ability to*

represent site-specific and nonlinear real-world operating behavior (Steyaert and Condon, 2024; Yang et al., 2021). Some progress has been made in recent studies. Yassin et al. (2019) proposed an improved target storage-and-release parameterization that better represents reservoir storage zones and seasonal release behavior, but it requires in-situ storage and release series (or distributions) to determine model parameters. Turner et al. (2021) developed a transferable rule-based framework that learns operating policies from data-rich reservoirs and extrapolates them to neighboring data-scarce reservoirs with similar purposes. However, directly transferring the same policy parameters from source reservoirs to target reservoirs may impose source-specific operating behavior and introduce significant bias.

Recently, machine learning has been introduced to improve the parameterization and transferability of rule-based reservoir operation schemes. Some studies have used machine learning to estimate parameters for existing operation schemes, such as reservoir-operation parameters inferred from reservoir attributes using Random Forest models (Steyaert et al., 2025) and flood storage capacities estimated using XGBoost to define storage-zone parameters in a target storage scheme (Shen et al., 2025). Researchers have also explored the use of machine learning to develop hybrid models for improved reservoir representation. Chen et al. (2022) developed a data-driven framework that extracts interpretable operation modules and associated application conditions from historical operation records. Building on this modular reservoir modeling concept, Li and Villarini (2026) further investigated its generalization using Random Forest models to transfer parameters associated with typical operation modules, such as constant, linear, and piecewise-linear release relationships, across reservoirs. Although these approaches improve parameter estimation and transferability while retaining interpretability, they generally represent reservoir operations through simplified modules or predefined rules. Moreover, because these modules or parameters are typically identified from individual reservoirs before being transferred, these approaches cannot fully leverage cross-site information during model training, which has been shown to improve the generalization of regional deep learning models (Fang et al., 2022; Kratzert et al., 2024). Consequently, their ability to learn complex operating behavior across diverse reservoirs may remain constrained.

Another line of research uses machine learning directly for reservoir operation modeling in an end-to-end way. Site-specific models based on decision trees (Dong et al., 2023; Yang et al., 2016, 2021), recurrent neural networks (Cheng et al., 2024; Yang et al., 2019), LSTM variants (Longyang and Zeng, 2023; Zhang et al., 2018), and physics-encoded deep learning (Yu et al., 2025; Zheng et al., 2022) have shown

strong ability to learn release behavior directly from historical records without prescribing explicit operating rules. However, these models require local operational data for training and therefore have limited applicability to data-scarce reservoirs. To address this limitation, recent studies have begun to explore regional end-to-end models that learn release behavior across multiple reservoirs. Ford and Sankarasubramanian (2023) proposed a regionalized piecewise linear regression tree framework for reservoir release prediction, but its reliance on reservoir-specific release and storage statistics for normalization limits direct transfer to unseen reservoirs without historical operation records. Tran et al. (2025) developed a conditioned LSTM for daily reservoir release prediction across CONUS reservoirs using static reservoir attributes, demonstrating the potential of pooled deep learning for reservoir operations. However, the model focuses on next-day release prediction, relies on observed storage as an input, and does not explicitly enforce mass balance. In contrast, long-term reservoir simulation requires the model to internally evolve storage states while maintaining mass balance, which remains largely unexplored."

In section 2.3, we have added (**lines 233-236**): "*In Experiment II, we also compared PLSTM-Reg with the calibrated ISTARF model (Turner et al., 2021), using parameters fitted from the training period of each reservoir and evaluated on the chronological test period. The implementation details are provided in Supplementary Text S2.*"

In section 3.1, we have added (**lines 303-305**): "*Both PLSTM-Reg and PLSTM-Loc outperform the calibrated ISTARF benchmark, which achieves median KGE values of 0.64 for release and 0.47 for storage (Supplementary Fig. S4).*"

In Supplementary Text S2, we have added: "*The fourth benchmark is the ISTARF model developed by (Turner et al., 2021), which represents reservoir storage and release policies using a 19-parameter formulation. ISTARF employs harmonic regression to define the upper and lower bounds of a seasonally varying Normal Operating Range (NOR) as percentages of storage capacity. During normal operations within this range, target release is computed as a fractional deviation from long-term mean annual inflow and adjusted dynamically according to the current storage state and inflow. When storage falls below or exceeds the NOR, releases are constrained by predefined minimum or maximum bounds, respectively.*

We implemented ISTARF in two ways, corresponding to the temporal and spatial long-term simulation experiments. In Experiment II, ISTARF was calibrated separately for each reservoir using the chronological training period and evaluated on the corresponding test period. This calibrated configuration represents the

performance of ISTARF when historical operation records are available for the target reservoir during model fitting. In Experiment IV, ISTARF was implemented in a data-scarce spatial-transfer setting. For each fold, ISTARF parameters were fitted using only reservoirs in the training folds. Donor selection for each held-out reservoir was then restricted to the training pool, following the donor-selection hierarchy based on hydrologic-region similarity, operational-purpose alignment, and geographic proximity. The transferred parameters were used to simulate daily release and storage for the held-out reservoirs."

The revised **Fig. 5** and **Fig. S4** are provided below:

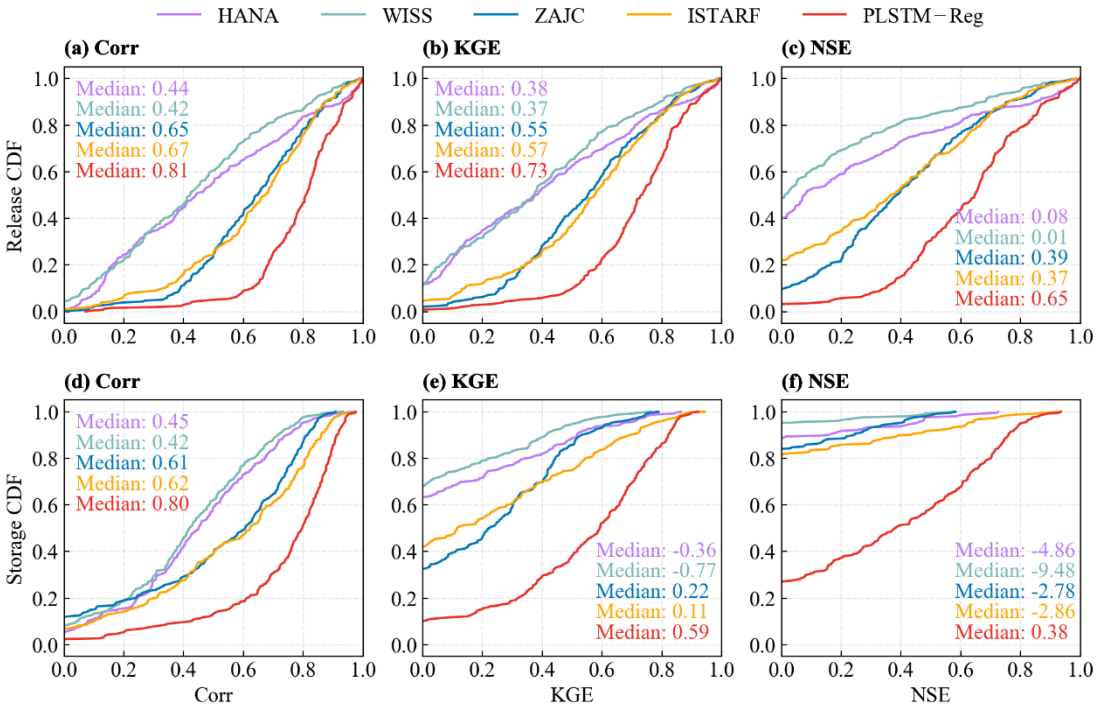


Figure 5. Comparison of release and storage long-term simulation performance (Experiment IV) between the regional model and four benchmark models. CDFs of three metrics (Corr, KGE, and NSE) are displayed for 256 reservoirs over the full evaluation period. The upper row (a–c) illustrates release performance, and the lower row (d–f) shows storage performance. Median metric values for each model are annotated within the respective panels.

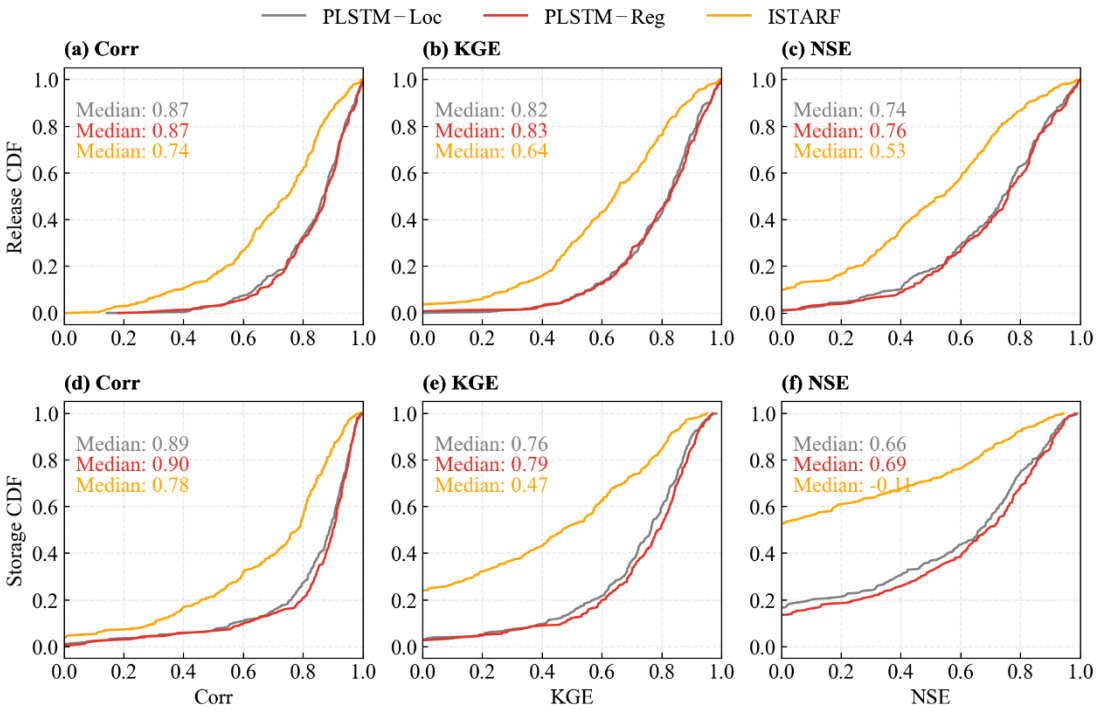


Figure S4. Comparison of release and storage long-term simulation performance under Experiment II across all reservoirs during the test period. CDFs display release simulation performance in the upper row (a–c) and storage simulation performance in the lower row (d–f). Median metric scores are annotated within the respective panels.

[4] The source of improvement should be interpreted more carefully. The gains over rule-based methods may come from both regional pooling and the flexible neural-network architecture. Simply attributing the improvement to the regional setting is not fully fair. The authors should distinguish the effects of regional training, static attributes, nonlinear LSTM architecture, and physical constraints.

Response: Thank you for pointing this out. We agree that the improvement over the rule-based benchmarks should not be attributed solely to regional training. The contributions of individual components are discussed separately in the manuscript: Section 3.1 evaluates regional training, Section 4.1 evaluates static attributes, and the benefits of the physics-encoded PLSTM architecture have been demonstrated in our previous study (Yu et al., 2025). To clarify this, we revised Section 3.2 and the Conclusions, as follow:

In Section 3.2, we have revised **lines 353-358** to: "*These results highlight the advantages of the proposed framework. The outperformance over generic and*

uncalibrated rule-based schemes reflects the combined effects of the nonlinear PLSTM architecture, regional data pooling, static reservoir attributes, and embedded physical constraints. Together, these components enable more realistic predictions in data-scarce settings where simplified empirical rules often fail."

In the Conclusions section, we have revised **lines 516-520** to: "*This result indicates that the combination of a physics-encoded neural network architecture, regional training, and static reservoir attributes can capture shared behavioral patterns more effectively than simplified empirical rules, providing a more reliable basis for simulating reservoirs with limited or no operational records.*"

Minor Comments

Line 14. "Representative reservoirs" needs clearer definition in the data section to avoid selection bias. Is data length the only selection criteria?

Response: Thank you for pointing this out. We clarified that data length was not the only selection criterion. In Chen et al. (2025), data length and missing-data rate were the main filtering criteria. The term "representative" refers to the broad geographic, hydroclimatic, and operational diversity of the resulting 256-reservoir dataset, as described in Section 2.1.

In Section 3.2, we have revised **lines 117-119** to: "*The reservoirs were selected based on record availability and data quality, requiring at least 25 years of continuous records spanning 1990–2021 (end dates ranged from 31 December 2014 to 30 April 2021) and <10% missing data.*"

Line 50. Previous models require water-demand data for parameterization, while this model does not use demand data but still performs well. The authors should explain why demand information may be implicitly captured, for example through storage, release history, seasonality, reservoir purpose, and static attributes.

Response: Thank you for this suggestion. We added the following explanation in Section 2.2 (**lines 161-167**): "*Although explicit water-demand data are not used, demand-related operational patterns may be implicitly represented through the model inputs. Seasonal indicators, hydrometeorological forcings, and static attributes such as reservoir purpose provide information related to the timing and magnitude of water demands. Furthermore, historical storage and release records reflect the*

cumulative effects of reservoir operation decisions driven by water demands and management objectives, thereby providing indirect information on demand-related operating behaviors."

Line 124-126. It would be helpful to know the missing ratios of the remote sensing data.

Response: Thank you for this suggestion. We have added the missing ratios of the raw monthly GRSAD surface-area data in Section 2.1 (**lines 150-156**): "*Prior to imputation, the raw monthly GRSAD time series across the evaluated reservoirs had an average missing-data ratio of 3.7%, with reservoir-level values ranging from 0% to 6.6%. For missing values within the GRSAD record, data were imputed according to gap length: gaps of six months or shorter were linearly interpolated, whereas longer gaps were filled with the long-term monthly mean surface area. For periods extending beyond the GRSAD coverage (post-2018), the same climatological monthly mean was used to maintain dataset consistency and avoid cross-product biases.*"

Line 165. The authors use the term "forecasting." However, if future forcings/inflows are assumed to be known when running the model recursively, this is closer to conditional simulation or hindcast prediction than true operational forecasting. Please clarify.

Response: Thank you for pointing this out. We agree that the setup is not an operational forecast because future meteorological forcings and inflows are taken from observed records. We have updated the manuscript to replace all instances of "short-term forecasting" with "short-term prediction" to avoid ambiguity.

Line 176. Long-term simulation capability depends on the training window length. Longer sequence windows may help the model learn to avoid error accumulation. This does not require a new experiment, but it is worth mentioning in the discussion.

Response: Thank you for this suggestion. We agree that sequence length may affect long-term free-running simulations because storage errors can accumulate over time. We have clarified this in Section 2.3. We have also added a sensitivity analysis using 2-, 4-, and 6-year training sequences. The results (Supplementary Fig. S1) show similar performance across the tested sequence lengths, indicating that the long-term

simulation results are not strongly sensitive to this hyperparameter within the tested range.

We revised Section 2.3 (**lines 226–229**) as follows: "*The sequence length for both model training and testing was set to 1460 days (4 years). A sensitivity analysis using 2-, 4-, and 6-year sequence lengths yielded similar performance across all tested settings (Supplementary Fig. S1), indicating that the results are not strongly sensitive to this choice.*"

Line 197. The authors should clarify how the monthly surface-area time series is used at daily steps. If the current month's value is used for all days in that month, it may include future information for earlier days.

Response: Thank you for this question. Monthly surface-area data were only used in Experiment V, which is designed for historical reconstruction rather than short-term prediction. Monthly values were assigned to the midpoint of each month and linearly interpolated to daily time steps. The interpolated series may incorporate information from later dates within the same month; however, this is consistent with the retrospective reconstruction objective, as all surface-area observations used in Experiment V are historical data.

We have clarified this point in Section 2.3 (**lines 253–257**): "*To align with the model's daily temporal resolution, monthly surface area values were assigned to the midpoint of each month and linearly interpolated to daily time steps. The interpolated daily series may use within-month information from both earlier and later dates, which is appropriate for the reconstruction setting because the complete historical surface-area record is available during reconstruction.*"

Lines 227–229. It is interesting that release performance differs clearly between PLSTM-Reg and PLSTM-Loc, while storage performance is nearly the same. The authors should explain this.

Response: Thank you for this insightful question. This contrast mainly reflects the strong persistence of reservoir storage and the short 1–7 day prediction horizon. Both models are initialized with observed storage, and storage predictions are updated over only a few days. Because daily inflows and releases are generally small relative to total storage, differences in release prediction have limited impact on storage over such a short period. As a result, storage performance remains similarly high for both

PLSTM-Reg and PLSTM-Loc, whereas release predictions more clearly reveal the benefits of regional training.

We revised Section 3.1 (**lines 286–291**) as follows: "*As for short-term storage prediction, both approaches achieve similarly high accuracy because storage is strongly constrained by the observed initial condition and exhibits substantial persistence over the 1–7 day prediction horizon (Supplementary Fig. S3). As daily inflows and releases are generally much smaller than total storage, differences in release prediction accuracy have limited influence on simulated storage over this short period.*"

Figure 3. Since all statistical tests appear significant, the asterisks could be removed and the caption can simply state that all comparisons pass the significance test. Same for Figure 6.

Response: Thank you for this suggestion. We have made changes accordingly.

Lines 273–274. In the "unseen reservoir" experiment, it is unclear how PLSTM-Loc is trained. Does it use local data from the held-out reservoir, or random initial weights? This needs clarification.

Response: Thank you for this question. In Experiment III, PLSTM-Reg was evaluated on reservoirs that were excluded from regional training. The PLSTM-Loc values reported in this section are not models trained for the held-out reservoirs under Experiment III. Instead, they are the site-specific models from Experiment I, trained using each reservoir's local historical records under a chronological split. They are included as a reference benchmark to illustrate that the regional model can achieve performance comparable to, or better than, locally trained models even when transferred to previously unseen reservoirs.

We revised Section 3.2 (**lines 343–349**) as follows: "*In short-term prediction (Experiment III), PLSTM-Reg achieves a median 1-day-ahead KGE of 0.95 (Supplementary Fig. S7b) for reservoirs excluded from regional training. This performance exceeds that of the site-specific PLSTM-Loc models from Experiment I, which achieved a median KGE of 0.83 when trained using local historical records. The advantage of PLSTM-Reg persists across increasing lead times (Supplementary Fig. S7 g–h), indicating that the regional model can successfully transfer learned operating behaviors to previously unseen reservoirs.*"

Line 280. The improvement over rule-based methods may come from both regional pooling and flexible black-box modeling. The manuscript should avoid attributing the gain only to regional learning.

Response: Thank you for pointing this out. We agree that the performance gains over rule-based benchmarks should not be attributed solely to regional learning. This issue is addressed in our response to Major Comment 4, where we clarify that the observed improvements arise from the combined effects of regional data pooling, model architecture, static reservoir attributes, and physical constraints.

Figure 5. Please clarify why the metric values are not consistent with Figure S2. Do they correspond to different experimental settings or evaluation periods?

Response: Thank you for pointing this out. The metric values differ because the two figures correspond to different experiments and evaluation periods. Figure 5 presents Experiment IV, which evaluates spatial generalization to unseen reservoirs over the full available evaluation period. In contrast, Figure S4 (formerly Figure S2) presents Experiment II, which evaluates temporal generalization during the test period only (2010 onward).

We have revised both figure captions to clarify this distinction:

"*Figure 5. Comparison of release and storage long-term simulation performance (Experiment IV) between the regional model and four benchmark models. CDFs of three metrics (Corr, KGE, and NSE) are displayed for 256 reservoirs over the full evaluation period. The upper row (a–c) illustrates release performance, and the lower row (d–f) shows storage performance. Median metric values for each model are annotated within the respective panels.*"

"*Figure S4. Comparison of release and storage long-term simulation performance under Experiment II across all reservoirs during the test period. CDFs display release simulation performance in the upper row (a–c) and storage simulation performance in the lower row (d–f). Median metric scores are annotated within the respective panels.*"

Text S2. Rule-based benchmarks use default parameters, not calibrated ones. It does not necessarily imply that PLSTM-Reg outperforms these schemes.

Response: Thank you for raising this concern. The objective of Experiment IV is to evaluate performance under a spatial-transfer, data-scarce setting in which operational records are unavailable for the target reservoirs. Accordingly, benchmark implementation was designed to reflect the information that would realistically be available under this setting. HANA, WISS, and ZAJC were therefore implemented using their recommended parameters, consistent with their intended use as generic reservoir-operation parameterizations in large-scale hydrological and land-surface models. Calibrating these schemes separately for each held-out reservoir would require target-reservoir operation records and would therefore be inconsistent with the design of Experiment IV.

We also clarified the ISTARF setup. In Experiment IV, ISTARF was implemented using a fold-restricted parameter-transfer procedure, so that held-out reservoir records were not used during parameter fitting or donor selection. As a complementary comparison under data-rich conditions, calibrated ISTARF was evaluated in Experiment II using each reservoir's chronological training period.

To clarify the rationale behind the benchmark implementation and the intended scope of the comparison, we added the following text in Supplement Text S2: "*Because Experiment IV evaluates spatial transfer to held-out reservoirs under data-scarce conditions, benchmark parameterization was designed to reflect information availability in large-scale hydrological modeling. HANA, WISS, and ZAJC were implemented using their recommended parameters, consistent with their typical use as generic reservoir-operation schemes when local operation records are unavailable. Calibrating these models separately for the held-out reservoirs would require target-reservoir operation records and would therefore be inconsistent with the spatial-transfer setting. ISTARF was evaluated using the fold-restricted parameter-transfer procedure described above, ensuring that information from held-out reservoirs was not used during parameter fitting or donor selection. As a complementary comparison under data-rich conditions, calibrated ISTARF was also evaluated in Experiment II using each reservoir's chronological training period.*"

References

Chen, Y., Li, D., Zhao, Q., and Cai, X.: Developing a generic data-driven reservoir operation model, Advances in Water Resources, 167, 104274, https://doi.org/10.1016/j.advwatres.2022.104274, 2022.

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.

Cheng, H., Wang, T., and Yang, D.: Quantifying the Regulation Capacity of the Three Gorges Reservoir on Extreme Hydrological Events and Its Impact on Flow Regime in a Changing Climate, Water Resources Research, 60, e2023WR036329, https://doi.org/10.1029/2023WR036329, 2024.

Dong, N., Guan, W., Cao, J., Zou, Y., Yang, M., Wei, J., Chen, L., and Wang, H.: A hybrid hydrologic modelling framework with data-driven and conceptual reservoir operation schemes for reservoir impact assessment and predictions, Journal of Hydrology, 619, 129246, https://doi.org/10.1016/j.jhydrol.2023.129246, 2023.

Fang, K., Kifer, D., Lawson, K., Feng, D., and Shen, C.: The Data Synergy Effects of Time-Series Deep Learning Models in Hydrology, Water Resources Research, 58, e2021WR029583, https://doi.org/10.1029/2021WR029583, 2022.

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.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An integrated model for the assessment of global water resources – Part 1: Model description and input meteorological forcing, Hydrology and Earth System Sciences, 12, 1007–1025, https://doi.org/10.5194/hess-12-1007-2008, 2008a.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An integrated model for the assessment of global water resources – Part 2: Applications and assessments, Hydrology and Earth System Sciences, 12, 1027–1037, https://doi.org/10.5194/hess-12-1027-2008, 2008b.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin, Hydrology and Earth System Sciences, 28, 4187–4201, https://doi.org/10.5194/hess-28-4187-2024, 2024.

Lehner, B., Beames, P., Mulligan, M., Zarfl, C., De Felice, L., van Soesbergen, A., Thieme, M., Garcia de Leaniz, C., Anand, M., Belletti, B., Brauman, K. A., Januchowski-Hartley, S. R., Lyon, K., Mandle, L., Mazany-Wright, N., Messager, M. L., Pavelsky, T., Pekel, J.-F., Wang, J., Wen, Q., Wishart, M., Xing, T., Yang, X., and Higgins, J.: The Global Dam Watch database of river barrier and reservoir information for large-scale applications, Sci Data, 11, 1069, https://doi.org/10.1038/s41597-024-03752-9, 2024.

Longyang, Q. and Zeng, R.: A Hierarchical Temporal Scale Framework for Data-

Driven Reservoir Release Modeling, Water Resources Research, 59, e2022WR033922, https://doi.org/10.1029/2022WR033922, 2023.

Shen, Y., Yamazaki, D., Pokhrel, Y., and Zhao, G.: Improving Global Reservoir Parameterizations by Incorporating Flood Storage Capacity Data and Satellite Observations, Water Resources Research, 61, e2024WR037620, https://doi.org/10.1029/2024WR037620, 2025.

Steyaert, J. C. and Condon, L. E.: Synthesis of historical reservoir operations from 1980 to 2020 for the evaluation of reservoir representation in large-scale hydrologic models, Hydrology and Earth System Sciences, 28, 1071–1088, https://doi.org/10.5194/hess-28-1071-2024, 2024.

Steyaert, J. C., Sutanudjaja, E., Bierkens, M., and Wanders, N.: A data derived workflow for reservoir operations for simulating reservoir operations in a global hydrologic model, EGUsphere, 1–38, https://doi.org/10.5194/egusphere-2024-3658, 2025.

Tran, H., Zhou, T., Tan, Z., Fang, Y., and Ruby Leung, L.: Improving the prediction of daily reservoir releases over the CONUS using conditioned LSTM, Journal of Hydrology, 661, 133750, https://doi.org/10.1016/j.jhydrol.2025.133750, 2025.

Turner, S. W. D., Steyaert, J. C., Condon, L., and Voisin, N.: Water storage and release policies for all large reservoirs of conterminous United States, Journal of Hydrology, 603, 126843, https://doi.org/10.1016/j.jhydrol.2021.126843, 2021.

Wisser, D., Fekete, B. M., Vörösmarty, C. J., and Schumann, A. H.: Reconstructing 20th century global hydrography: a contribution to the Global Terrestrial Network-Hydrology (GTN-H), Hydrology and Earth System Sciences, 14, 1–24, https://doi.org/10.5194/hess-14-1-2010, 2010.

Yang, S., Yang, D., Chen, J., and Zhao, B.: Real-time reservoir operation using recurrent neural networks and inflow forecast from a distributed hydrological model, Journal of Hydrology, 579, 124229, https://doi.org/10.1016/j.jhydrol.2019.124229, 2019.

Yang, T., Gao, X., Sorooshian, S., and Li, X.: Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme, Water Resources Research, 52, 1626–1651, https://doi.org/10.1002/2015WR017394, 2016.

Yang, T., Zhang, L., Kim, T., Hong, Y., Zhang, D., and Peng, Q.: A large-scale comparison of Artificial Intelligence and Data Mining (AI&DM) techniques in simulating reservoir releases over the Upper Colorado Region, Journal of Hydrology, 602, 126723, https://doi.org/10.1016/j.jhydrol.2021.126723, 2021.

Yassin, F., Razavi, S., Elshamy, M., Davison, B., Sapriza-Azuri, G., and Wheater, H.: Representation and improved parameterization of reservoir operation in hydrological and land-surface models, Hydrology and Earth System Sciences, 23, 3735–3764, https://doi.org/10.5194/hess-23-3735-2019, 2019.

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.

Zajac, Z., Revilla-Romero, B., Salamon, P., Burek, P., Hirpa, F. A., and Beck, H.: The impact of lake and reservoir parameterization on global streamflow simulation, Journal of Hydrology, 548, 552–568, https://doi.org/10.1016/j.jhydrol.2017.03.022, 2017.

Zhang, D., Lin, J., Peng, Q., Wang, D., Yang, T., Sorooshian, S., Liu, X., and Zhuang, J.: Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm, Journal of Hydrology, 565, 720–736, https://doi.org/10.1016/j.jhydrol.2018.08.050, 2018.

Zheng, Y., Liu, P., Cheng, L., Xie, K., Lou, W., Li, X., Luo, X., Cheng, Q., Han, D., and Zhang, W.: Extracting operation behaviors of cascade reservoirs using physics-guided long-short term memory networks, Journal of Hydrology: Regional Studies, 40, 101034, https://doi.org/10.1016/j.ejrh.2022.101034, 2022.

Response to community comments CC2&CC3:

I would like to congratulate the authors for this original and well-structured contribution. The PLSTM-Reg model is an elegant way to embed physical consistency directly into the recurrent loop, and the results across the diverse set of reservoirs are compelling. I have several question regarding the model training and evaluation that I hope the authors can address:

Response: Dear Dr. Francois, thank you very much for your thorough review and encouraging comments. We have carefully addressed your comments, as detailed below.

Major Comments

1. Estimation of operational release limits Q_min and Q_max

Initial Comment1:

The Physical Knowledge Module (Equations S8–S10) requires per-reservoir estimates of minimum and maximum allowable release, Q_min and Q_max. The manuscript and supplement do not describe how these values were derived. Could the authors clarify: Were they computed from the observed outflow record (e.g., obs min and max values? empirical percentiles?) or from engineering/regulatory sources?

Follow-up Comment1:

After reviewing the code shared by the authors, I would like to ask a few follow-up questions regarding Equation S8 and the role of Q_min and Q_max in the Physical Knowledge Module.

In the released code (ReservoirLSTM.rnncell), the physical module applies Equations S9–S11 as described in the supplement, but Equation S8 does not appear to be implemented. Instead, the candidate release predicted by the linear output head is denormalized directly using the per-reservoir empirical mean and standard deviation of observed outflow:

release_denorm = release_factor * target_scale[0] + target_center[0]

This denormalization step does act as a soft, data-driven constraint: when the LSTM output is near zero, the predicted release is close to the observed mean, and the per-reservoir scaling discourages predictions far outside the observed distribution. However, this is a statistical prior derived from the training data rather than a hard operational bound. It cannot be directly interpreted as Q_min and Q_max, and it provides no guarantee that releases remain within physically or operationally meaningful bounds, particularly for out-of-distribution conditions.

I understand that the storage clamp (Eq S10) combined with Eq S11 provides an implicit physical floor and ceiling on release (you cannot release more than the available water, and you must spill when the reservoir is full). This is a meaningful hard constraint. However, it does not serve the same purpose as Q_min (minimum environmental or operational flow requirement) and Q_max (maximum release capacity).

Could the authors clarify: (a) Was Equation S8 intentionally omitted from the implementation, with the normalization scheme serving as the intended proxy for release bounds? (b) If so, how are Q_min and Q_max defined in the manuscript, and

are they used anywhere in the training or evaluation pipeline? (c) Does the absence of an explicit release clamp affect the physical interpretability of the model, particularly the claim that "physical knowledge" is enforced?

Response: Thank you for these questions regarding the implementation of Eq. (S8). Q_{min} and Q_{max} are intended to represent reservoir-specific operational limits (e.g., downstream environmental flow or navigation requirements for Q_{min}; turbine or spillway capacity for Q_{max}) rather than empirical percentiles derived from the observed outflow records. In our previous work (Yu et al., 2025), these constraints were explicitly enforced because reliable operational data were available for the two studied reservoirs. However, such data were not consistently available across the 256 reservoirs analyzed here. Therefore, although Eq. (S8) was retained in the manuscript for methodological completeness, it was intentionally not activated as a hard clamp in the present large-sample implementation. Instead, the statistical denormalization served as a soft data-driven prior.

We believe the omission of Eq. (S8) would not undermine the core physical interpretability of the PLSTM-Reg model. The primary physical knowledge enforced by the architecture is mass conservation. While Eq. (S8) reflects human-defined operational preferences, Eqs. (S9)-(S11) enforce fundamental physical constraints on storage and release dynamics. Specifically, the model cannot release more water than it is available in storage and must spill when storage capacity is exceeded. Therefore, the mass-balance consistency of the model remains preserved.

To clarify this point, we have added the following text to **Text S1** of the Supplement: "*It is important to note that, although Eq. (S8) is included for theoretical completeness, reliable estimates of the operational and infrastructure limits* Q_{min} *and* Q_{max} *are generally unavailable for large-sample reservoir datasets. Therefore, to avoid introducing potentially unrealistic operational constraints, Eq. (S8) was intentionally not enforced in the current large-sample implementation*"

Furthermore, to better align our published Zenodo repository (https://doi.org/10.5281/zenodo.18265198; Yu, 2026) with the theoretical framework, we have updated the reference code to include the Q_{min} and Q_{max} constraints. These constraints remain deactivated by default for the current large-sample experiments, but can be enabled when reliable reservoir-specific operational metadata are available.

2. LSTM warmup period and handling of the physical storage state

Initial Comment2:

Standard LSTM implementations use a warmup period to bring the hidden state from zero-initialization to a realistic operating regime before the loss is computed. In PLSTM-Reg, the physical storage state s_t is simultaneously updated at every timestep via the water-balance equation (Eq. S9), creating a tighter coupling between the recurrent state and the physical state than in a conventional LSTM. Could the authors clarify: (a) Was a warmup period used, and if so, what length? (b) During warmup, was the physical knowledge module (Eqs. S8–S11) active, or was the physical state held at the observed initial storage? Else?

Follow-up Comment3:

I also noticed that the initial physical storage state is hard-coded to 0.5 (i.e., half of the normalized capacity) regardless of the actual observed storage at the start of each sequence:

```
# initialize reservoir storage using 0.5
device = x_d_embedded.device
temp_vars = torch.ones((batch,2)).to(device=device,
dtype=x_d_embedded.dtype)*0.5
```

Could the authors comment on the sensitivity of the model to this initialization choice? In particular, for short sequences or reservoirs that are frequently at extreme storage levels (near empty or near full), a fixed 0.5 initialization may introduce a systematic spin-up bias. Was this choice evaluated against alternatives such as initializing from the observed storage at the start of each sequence, or using a learned initial state? Is this initial storage used as the start of the warm-up period during training only?

Response: We appreciate this important question regarding state initialization and warmup.

Yes, 1-year (365 days) warmup period was used for all long-term sequence simulations. During warmup, the physical knowledge module (Eqs. S8–S11) remained fully active, allowing both the LSTM hidden state and the physical storage state to evolve towards a realistic mass-balanced regime before the loss computation and evaluation.

Regarding the initialization of the physical storage state, the strategy depended on the experimental setting. In data-rich settings (Training & Experiment II), the observed

initial storage was used. In ungauged or data-scarce settings (Experiment IV & Regional Benchmarks), where initial storage was assumed unavailable, the model used 50% capacity (0.5) as a neutral prior.

To evaluate the sensitivity to the 0.5 initialization setting, we conducted an additional sensitivity analysis for the spatial generalization model (Experiment IV), initializing storage at 10%, 30%, 50%, 70%, and 90% of reservoir capacity. As shown in **Supplementary Fig. S2** (included below), model performance remained highly consistent across all initialization settings, indicating that the 1-year active warmup period effectively removes the influence of initial condition on long-term simulation.

We have clarified this point in the revised manuscript (**lines 245-249**): "*Furthermore, for all simulations in data-scarce settings (Experiment IV and benchmarks), storage was initialized at 50% of reservoir capacity, whereas observed initial storage was used during training. A 1-year warm-up period was applied to equilibrate model states. Sensitivity analysis confirming the robustness of the model to this initial storage assumption is provided in Supplementary Fig. S2.*"

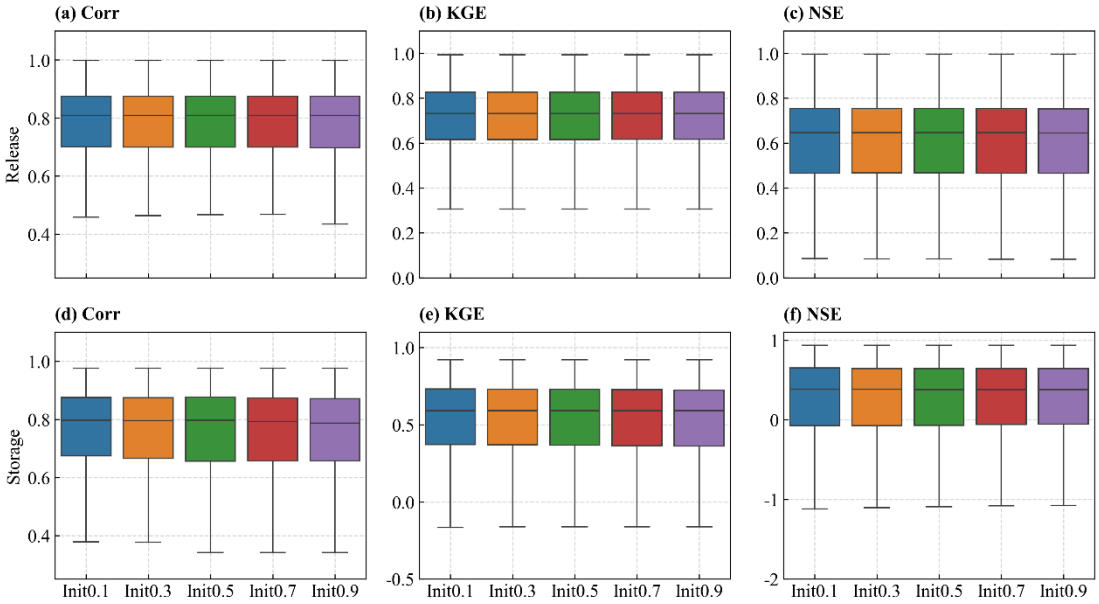


Figure S2. Sensitivity of spatial out-of-sample long-term simulation (Experiment IV) performance to initial storage conditions. Boxplots display evaluation metrics for release simulations (a–c) and storage simulations (d–f) when the physical storage state is initialized at 10%, 30%, 50%, 70%, and 90% of reservoir capacity. The similar performance distributions across initialization settings demonstrate that the 1-year warmup period effectively minimizes sensitivity to initial storage conditions in long-term simulations.

Furthermore, we have updated the initialization logic in the public Zenodo repository (https://doi.org/10.5281/zenodo.18265198) so that the initialization state is

automatically determined from the experimental setting, avoiding ambiguity associated with hard-coded priors.

3. Training sequence length

Initial Comment3:

The choice of sequence length is particularly important for PLSTM-Reg because the model performs a free-run simulation over the full sequence, meaning error in storage accumulates over time. Could the authors indicate the sequence length used during training and whether they observed sensitivity to this hyperparameter? Specifically, did longer sequences lead to better long-term simulation performance at the cost of slower training convergence?

Response: We used a sequence length of 1460 days (approximately 4 years) for both training and testing in long-term simulations. This choice was based on our previous study (Yu et al., 2025), which showed a 4-year sequence effectively capture inter-annual variability while maintaining long-term mass balance.

To further evaluate the sensitivity to sequence length, we conducted an additional sensitivity analysis using 2-, 4-, and 6-year training sequences. As shown in the newly added **Supplementary Fig. S1** (included below), extending the training sequence from 4 to 6 years did not improve simulation performance, despite increased computational cost.

We have clarified this point in the revised manuscript (**lines 226-229**): "*The sequence length for both model training and testing was set to 1460 days (4 years). A sensitivity analysis using 2-, 4-, and 6-year sequence lengths yielded similar performance across all tested settings (Supplementary Fig. S1), indicating that the results are not strongly sensitive to this choice.*"

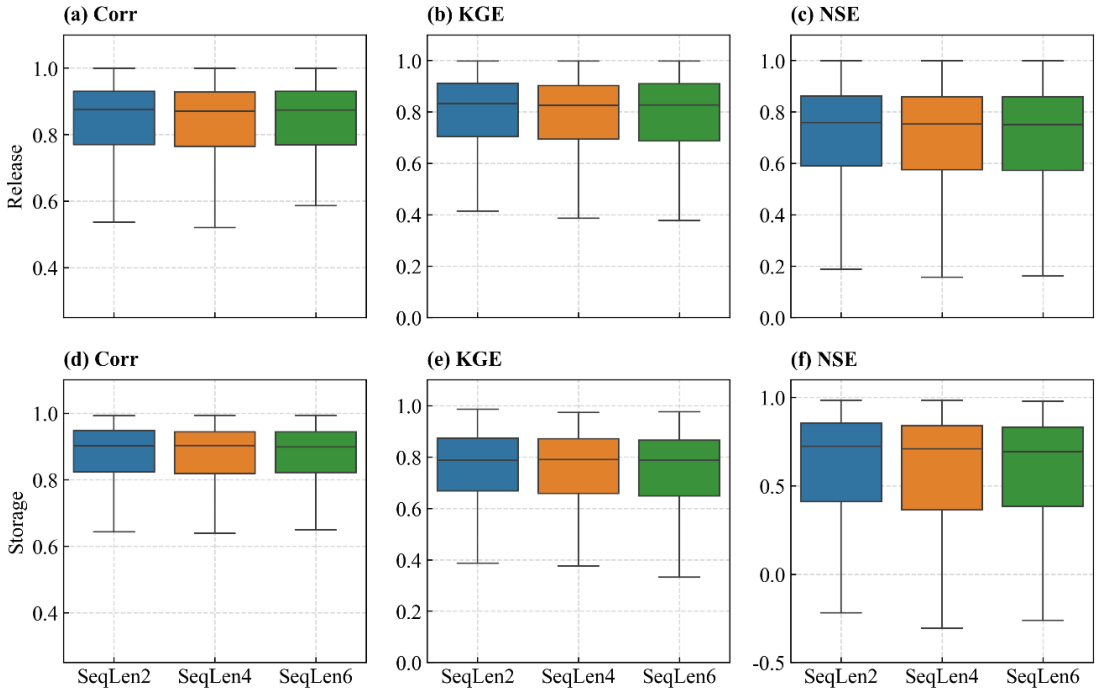


Figure S1. Sensitivity of long-term simulation performance to training sequence length. Boxplots compare release (a–c) and storage (d–f) simulation performance across training sequence lengths of 2, 4, and 6 years for all reservoirs. Evaluation metrics include Pearson Correlation (Corr), Kling-Gupta Efficiency (KGE), and Nash-Sutcliffe Efficiency (NSE). The largely consistent performance distributions across sequence lengths indicate that extending the training sequence beyond 4 years provides limited additional benefit for long-term simulations.

4. Long-term water balance consistency

Initial Comment4:

While the physical knowledge module enforces local mass balance at every timestep (Eq. S11) by construction, this does not guarantee that the long-term water balance is respected at the reservoir scale. We identify several potential sources of systematic bias:

First, the linear output head predicts a candidate release r~_t without any architectural constraint that enforces long-term mass conservation (i.e., there is no mechanism ensuring that the mean predicted release equals the mean inflow minus the long-term storage change). The LSTM may systematically over- or under-predict releases, creating a persistent bias in the physical storage trajectory.

Second, even if the candidate release were unbiased on average, the physical

knowledge module can introduce a systematic offset by activating the storage and flow constraints. For instance, frequent clamping at S_max (spilling excess water via Eq. S11) or at Q_max (capping release and accumulating storage) will alter the long-term mean release relative to what the LSTM predicted. The direction and magnitude of this effect depend on the distribution of reservoir states relative to the constraint bounds and is not self-correcting.

Third, and more fundamentally, the observed inflow and release records in datasets such as ResOpsUS may themselves not close the water balance, due to measurement uncertainty, unobserved fluxes (direct lake evaporation, groundwater exchange, water withdrawals from the reservoir), or data gaps. In such cases, the LSTM may partially learn to compensate for these residuals. This raises the question of whether the model is learning physically meaningful release dynamics or partly fitting an artifact of the input data.

Did the authors evaluate the long-term water balance of the PLSTM-Reg simulations, for example by comparing the multi-year mean simulated release to the mean inflow minus the observed long-term storage trend? A systematic evaluation of this property across the 259 reservoirs, and a discussion of how unobserved fluxes in the training data may affect the model's behavior, would significantly strengthen the physical interpretability of the results.

Response: We thank you for this insightful comment regarding long-term mass conservation and potential systematic drift in free-run simulations.

Regarding the concern about unobserved fluxes and residual water-balance errors in the training data, the reservoir dataset used in this study (Chen et al., 2025) provides net inflow rather than raw inflow. Specifically, net inflow is derived from the observed mass balance equation ($I_{net,t} = S_{t+1} - S_t + R_t$), such that unobserved gains and losses (e.g., evaporation, groundwater exchange, or withdrawals) are implicitly incorporated into the inputs. Therefore, the model is trained on a closed mass-balance system.

In addition, because the physical storage state is explicitly updated and fed back into the model at every timestep, persistent drift in release or storage cannot accumulate freely during long-term simulations.

To evaluate this empirically, we analyzed the long-term release bias across all 256 reservoirs using the absolute percent bias (|PBIAS|) over the testing period (Experiment II). As shown in the newly added **Supplementary Fig. S5** (included below), both the locally and regional PLSTM models maintain very low long-term

volume bias, with median |PBIAS| values below 0.4%, indicating that systematic long-term drift is minimal.

To clarify the issues of data closure and long-term mass conservation, we have added the following text to the revised manuscript:

In Section 2.1 (**lines 114-117**): "*Notably, the dataset provides net inflow derived from the observed storage–release balance ($I_{net,t} = S_{t+1} - S_t + R_t$), providing an internally closed water balance formulation that implicitly incorporates unobserved fluxes such as evaporation and seepage.*"

In Section 3.1 (**lines 305-307**): "*Crucially, both models maintain long-term water balance with minimal systematic drift (median absolute release percent bias (|PBIAS|) below 0.4%; Supplementary Fig. S5).*"

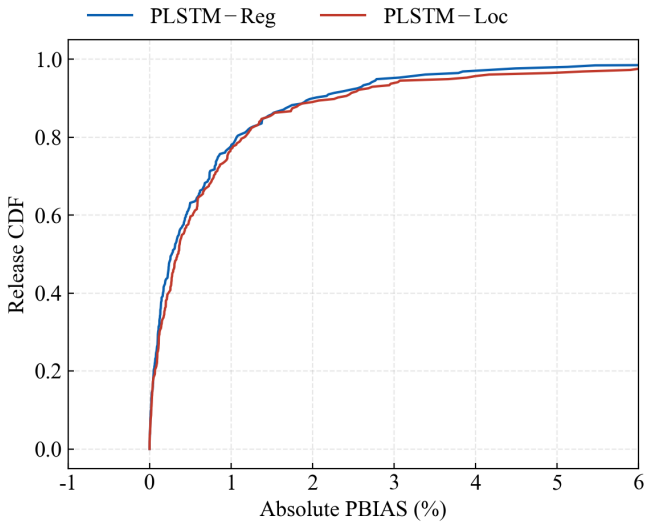


Figure S5. Empirical validation of long-term mass balance consistency across all reservoirs during the testing period. CDFs display the absolute Percent Bias (|PBIAS|) of simulated release for the regional (PLSTM-Reg) and local (PLSTM-Loc) models. Both models exhibit very low long-term release bias across reservoirs, indicating minimal systematic drift.

5. Spatial Out-of-Sample Generalization and Denormalization

Follow-up Comment2:

In addition, I am curious about the spatial out-of-sample evaluation. The model's linear output head predicts a normalized release factor that must be denormalized using the per-reservoir mean and standard deviation of observed release (target_center

and target_scale in the code). For reservoirs that were withheld from training, this implies that observed release statistics are still required at inference time. If that is the case, the model is not fully independent of observed data for unseen reservoirs, which would partially undermine the out-of-sample spatial generalization claim. Could the authors clarify how denormalization is handled for reservoirs not seen during training — specifically, whether per-reservoir observed statistics are used?

Response: We appreciate your careful inspection of the codebase. To clarify, we do not use per-reservoir observed statistics for denormalization in any experiments, including the spatial out-of-sample evaluations.

Specifically, the target_center and target_scale parameters are regional statistics computed exclusively from the aggregate training set. By using these global scaling factors, no local observed outflow information from unseen reservoirs is required during inference. Therefore, the spatial out-of-sample evaluation remain fully independent of local observational data.

To clarify this point, we have added the following sentence to Section 2.2 (**lines 194-195**): "*Data normalization was performed using regional scaling factors computed from the training set, rather than per-reservoir statistics.*"

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.

Yu, B.: PLSTM-Reg v1.0: A regional physics-encoded LSTM model for simulating reservoir operations under data scarcity, , https://doi.org/10.5281/zenodo.18265198, 2026.

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.