Response to Reviewer #2's comments

[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 & Gao, 2018), which provides data for 6,817 reservoirs listed in the GRanD database from 1984 to 2018. 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.