Supplement of Geosci. Model Dev., 19, 10699–10717, 2026
https://doi.org/10.5194/gmd-19-10699-2026-supplement
© Author(s) 2026. CC BY 4.0 License.



Geoscientific Model Development
Open Access
EGU

Supplement of

PLSTM-Reg v1.0: a regional physics-encoded LSTM model for simulating reservoir operations without site-specific training

Bin Yu et al.

Correspondence to: Yi Zheng (zhengyi@westlake.edu.cn)

The copyright of individual parts of the supplement might differ from the article licence.

Text S1. Detailed mathematical formulations of the PLSTM cell

The PLSTM cell integrates physical reservoir storage dynamics directly into the recurrent gating mechanism. The mathematical formulation of the PLSTM cell is as follows:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i s_{t-1} + b_i) \quad \text{(S1)}$$

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + V_f s_{t-1} + b_f\right) \quad \text{(S2)}$$

$$\widetilde{c_t} = \tanh\left(W_g x_t + U_g h_{t-1} + V_g s_{t-1} + b_g\right) \quad \text{(S3)}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o s_{t-1} + b_o) \quad \text{(S4)}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad \text{(S5)}$$

$$h_t = o_t \odot \tanh(c_t) \quad \text{(S6)}$$

$$\widetilde{r_t} = FC(h_t) \quad \text{(S7)}$$

$$\widehat{r_t} = \min\left(\max(\widetilde{r_t}, Q_{min}), Q_{max}\right) \quad \text{(S8)}$$

$$\tilde{s}_t = s_{t-1} + x_t^{in} - \widehat{r_t} \quad \text{(S9)}$$

$$s_t = \min\left(\max(\tilde{s}_t, S_{min}), S_{max}\right) \quad \text{(S10)}$$

$$r_t = x_t^{in} + s_{t-1} - s_t \quad \text{(S11)}$$

where i_t, f_t, and o_t represent the input, forget, and output gates, respectively, controlled by logistic sigmoid (σ) and hyperbolic tangent ($tanh$) activation functions. The learnable parameters include bias vectors (b) and weight matrices (W, U, and V) corresponding to the input x_t, hidden state h_{t-1}, and explicitly tracked reservoir storage s_{t-1}. In the state update mechanism, $\odot$ denotes element-wise multiplication, while c_t and $\tilde{c}_t$ denote the cell state and candidate cell state, respectively.

Equations (S8–11) define the "Physical Knowledge" module. In Eq. (S8), the candidate release $\widetilde{r_t}$ is constrained within predefined operational limits (Q_{min}, Q_{max}) such as the minimum environmental flow and maximum turbine capacity. 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. In Eq. (S9), this adjusted release $\widehat{r_t}$ is used in the dynamic water balance equation along with the inflow component x_t^{in} to calculate a candidate storage $\widetilde{s_t}$ for the next time step. In Eq. (S10), this candidate storage is further bounded by the reservoir's physical limits (S_{min}, S_{max}) such as dead and maximum storage capacity. In Eq. (S11), the actual release r_t is recalculated via mass balance to account for any storage adjustments, ensuring the system strictly adheres to conservation laws.

In the model implementation, the hard physical bounds in Eqs. (S8) and (S10) are executed using PyTorch's native clamp() function. Although this operation may truncate gradients during backpropagation when constraints are exceeded, preliminary

experiments showed stable optimization without requiring continuous smoothing techniques (e.g., tanh smoothing). Similar implementations are also adopted in recent differentiable hydrology frameworks such as hydrodl2 (mhpi, 2026).

Text S2. Implementation details of the four benchmark models

The first benchmark is the HANA model (Hanasaki et al., 2008), which has been widely used in land surface models such as MATSIRO (Pokhrel et al., 2012) and large-scale hydrological models like VIC (Droppers et al., 2020). It controls reservoir release using two operational schemes based on a size ratio, defined as storage capacity divided by mean annual inflow. The model sets the operational year as beginning in the first month when climatological inflow drops below the mean annual inflow, with the initial storage determining the yearly target release. For reservoirs with a size ratio above 0.5, daily release is linked to mean annual inflow and adjusted according to the deviation of initial storage from a normal storage level, fixed at 85% of capacity. When the size ratio is 0.5 or lower, release is further modified to reflect daily inflow variability. Although an irrigation-specific variant exists, in which releases scale to downstream demand, this study applies the non-irrigation formulation consistently, as irrigation demand data are unavailable.

The second benchmark, referred to as the WISS model, determines reservoir release using a piecewise linear relationship between inflow and outflow, and is a core component of the WBM-plus global hydrological framework (Wisser et al., 2010). Two linear equations are used depending on whether the current inflow exceeds the long-term mean annual inflow. The corresponding slopes are governed by the parameters κ and $\lambda - 1$. Following Wisser et al. (2010), who calibrated the model against operational data from 30 reservoirs worldwide, we adopt the recommended coefficient values of 0.16 for κ and 0.6 for $\lambda - 1$ in this study.

The third benchmark is the ZAJC model, which operates on a target storage-release formulation and serves as the reservoir representation within the LISFLOOD hydrological framework (Zajac et al., 2017). This algorithm partitions reservoir capacity into three operational zones: the dead storage zone (S_{dead}), the conservation zone (S_{cons}), and the flood control zone (S_{fld}). Each zone corresponds to a release target: minimum (Q_{min}), normal (Q_{norm}), and non-damaging flood (Q_{ndf}) releases. Within an active zone, daily discharge varies linearly with storage and inflow. While these boundaries and targets are ideally calibrated using historical operational records, generalized empirical values are necessary in data-scarce regions. Following Zajac et al. (2017), we set Q_{min}, Q_{norm}, and Q_{ndf} to the 5th, 30th, and 97th percentiles of natural streamflow, and define the three storage zones at 10%, 30%, and 97% of capacity.

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. 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.

Text S3. Feature Attribution Analysis using Integrated Gradients

To quantify the relative importance of the 21 static attributes in the PLSTM-Reg framework, we employed the Integrated Gradients (IG) method (Sundararajan et al., 2017). IG is a widely used, robust attribution technique designed for deep neural networks. It calculates feature importance by accumulating the gradients of the model's output with respect to its inputs along a continuous, straight-line path from a neutral baseline state to the actual input values.

Mathematically, the Integrated Gradient $IG_i(x)$ for the i-th feature of input x is defined as:

$$IG_i(x) = (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad \text{(S12)}$$

where x is the actual input vector of static attributes for a given reservoir; x' is the baseline input vector (defined in this study as a zero tensor in normalized input space); α is the interpolation constant scaling from 0 to 1; $F(\cdot)$ represents the PLSTM-Reg model function mapping the static attributes to the target variable; $\frac{\partial F}{\partial x_i}$ is the gradient of the model output with respect to the i-th static attribute.

The IG attributions were computed using Captum (Kokhlikyan et al., 2020), a PyTorch model interpretability library. To determine the global ranking of static feature importance, we first calculated the raw IG scores for each daily time step. The relative importance of each feature on a given day was then computed by taking its absolute IG value and dividing it by the sum of the absolute values across all static features. Finally,

these daily relative importances were averaged to yield the overall global importance metric.

Text S4. Effect of net inflow on storage-error propagation

The reservoir water balance can be written as

$$S_{t+1} = S_t + I_t - R_t + (G_t - L_t) \tag{S13}$$

where S_t is reservoir storage, I_t is river inflow, R_t is reservoir release, and G_t and L_t represent other reservoir gains and losses, such as direct precipitation, evaporation, and seepage. Rearranging the water balance gives the net inflow

$$I_{net,t} = S_{t+1} - S_t + R_t = I_t + G_t - L_t \tag{S14}$$

Thus, the net inflow provided by Chen et al. (2025) physically represents the net external water flux entering the reservoir after accounting for otherwise unresolved gains and losses.

To examine whether the use of net inflow mathematically constrains simulated storage to follow the observed storage trajectory, let e_t^S and e_t^R denote the simulation errors of storage and release at time t, respectively:

$$e_t^S = S_t^{sim} - S_t^{obs} \tag{S15}$$

$$e_t^R = R_t^{sim} - R_t^{obs} \tag{S16}$$

Because the same net inflow enters both the simulated and observed water-balance equations,

$$S_{t+1}^{sim} = S_t^{sim} - R_t^{sim} + I_{net,t} \tag{S17}$$

$$S_{t+1}^{obs} = S_t^{obs} - R_t^{obs} + I_{net,t} \tag{S18}$$

their difference gives

$$e_{t+1}^S = S_t^{sim} - S_t^{obs} - (R_t^{sim} - R_t^{obs}) = e_t^S - e_t^R \tag{S19}$$

and iterating this relation over time yields

$$e_t^S = e_0^S - \sum_{\tau<t} e_\tau^R \tag{S20}$$

Thus, the net-inflow term cancels from the storage-error propagation equation and does not mathematically anchor simulated storage to the observed storage trajectory. Instead, storage errors depend on the initial storage error and the accumulated release errors.

Figure S1

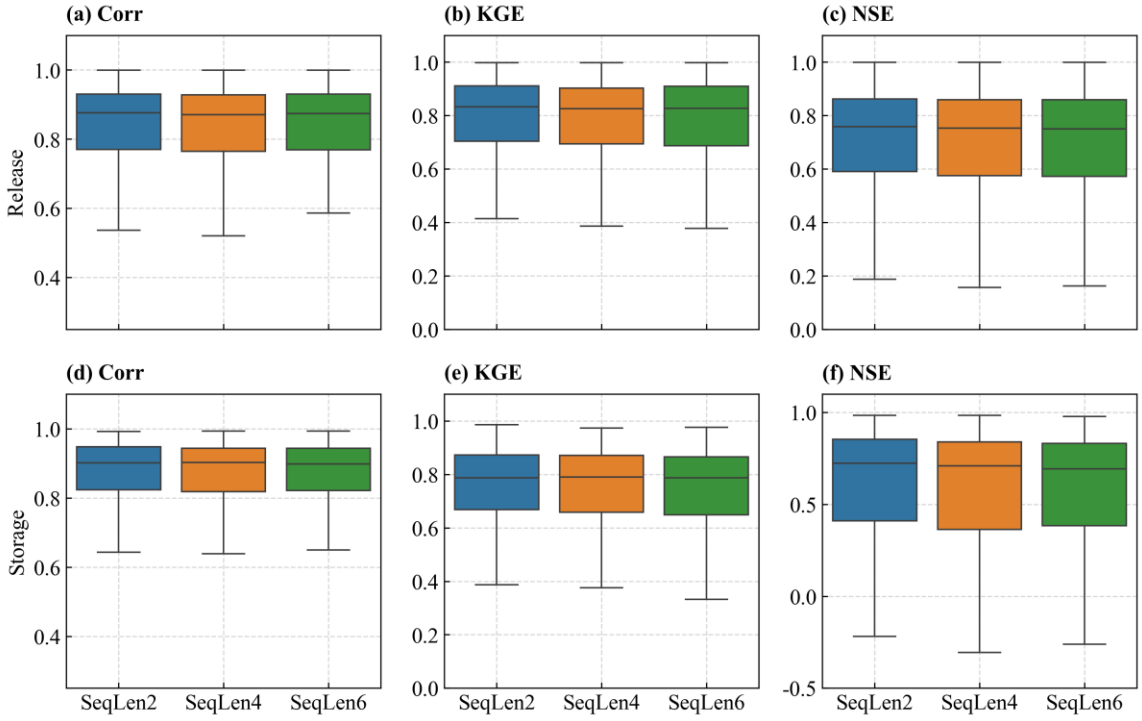


Figure S1. Sensitivity of long-term simulation performance to training sequence length. Boxplots compare release (a–c) and storage (d–f) simulation performance across training sequence lengths of 2, 4, and 6 years for all reservoirs. Evaluation metrics include Pearson Correlation (Corr), Kling-Gupta 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.

Figure S2

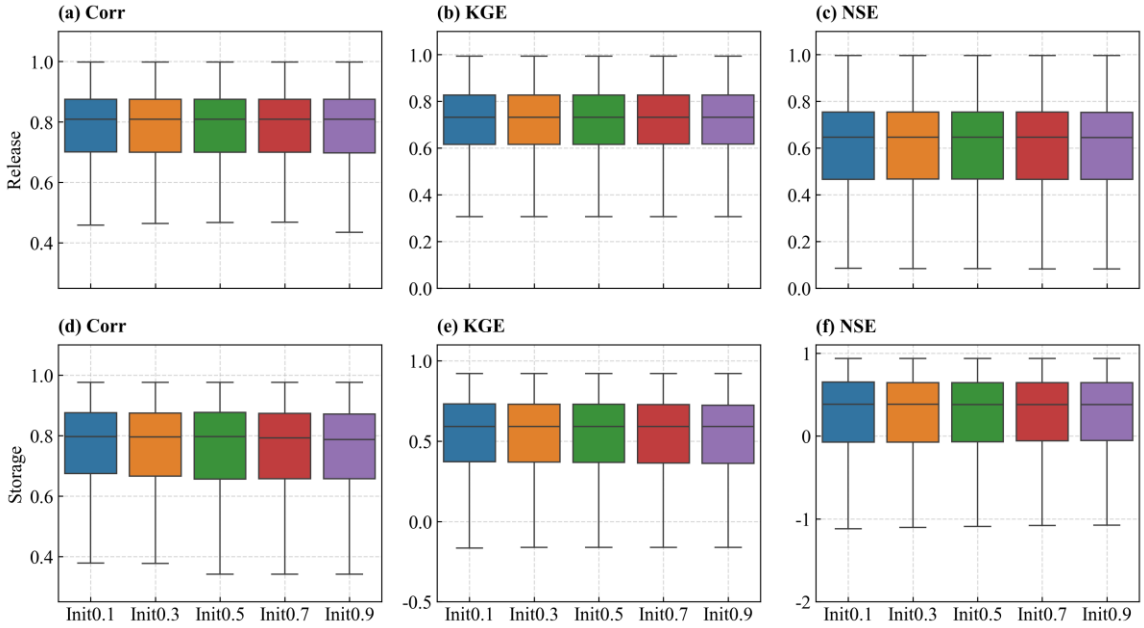


Figure S2. Sensitivity of spatial out-of-sample long-term simulation (Experiment IV) performance to initial storage conditions. Boxplots display evaluation metrics for release simulations (a–c) and storage simulations (d–f) when the physical storage state is initialized at 10%, 30%, 50%, 70%, and 90% 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.

Figure S3.

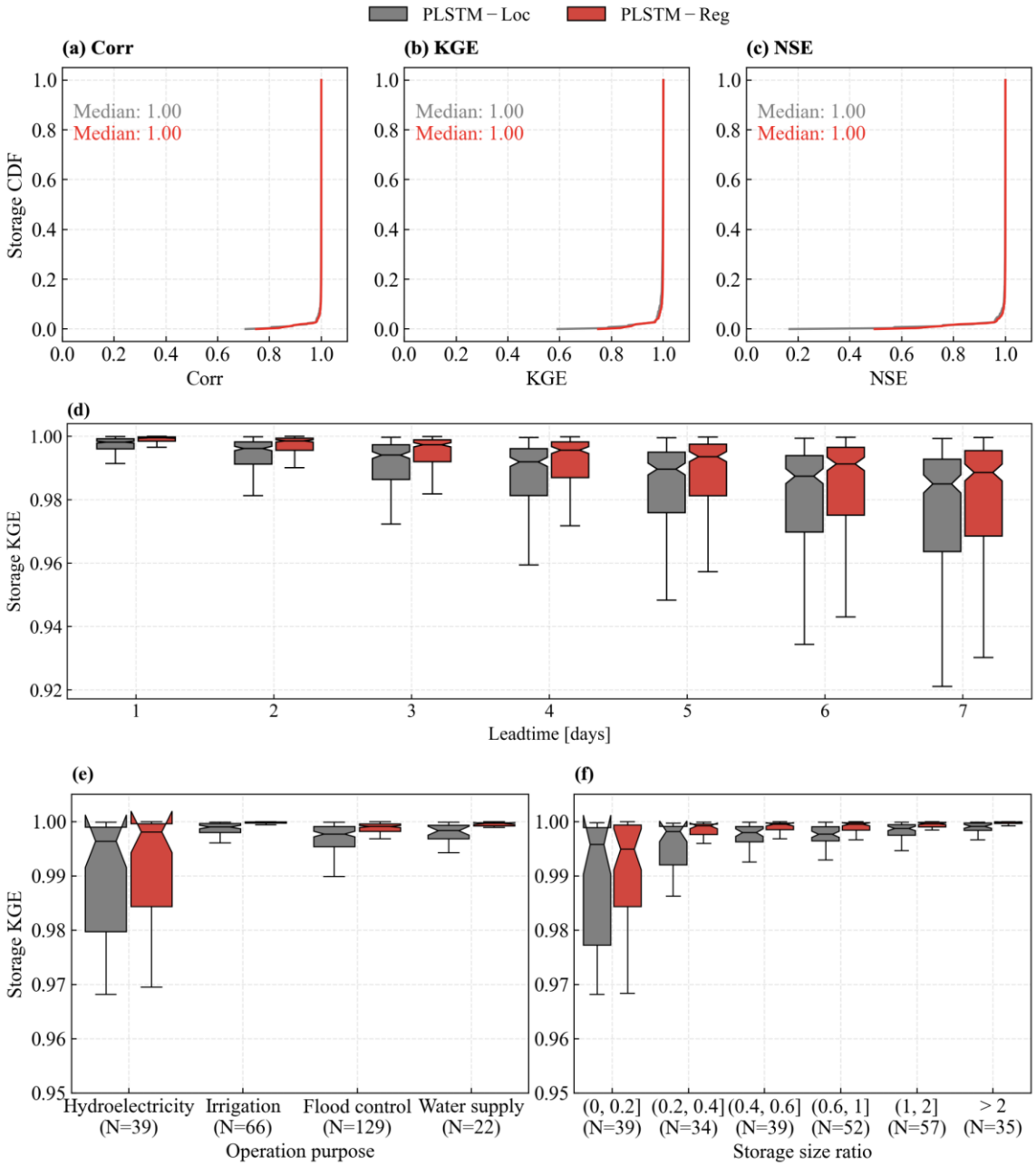


Figure S3. Evaluation of short-term storage prediction performance (Experiment I) across all reservoirs during the test period. (a–c) CDFs of one-day-ahead storage metrics comparing regional and local models. Median scores are annotated. (d) Evolution of storage KGE scores across lead times (1–7 days). (e–f) Performance stratified by (e) primary operational purpose and (f) storage size ratio. All paired comparisons shown in panels (d–f) are statistically significant based on the paired Wilcoxon signed-rank test ($p < 0.05$). N denotes the number of reservoirs in each category.

Figure S4.

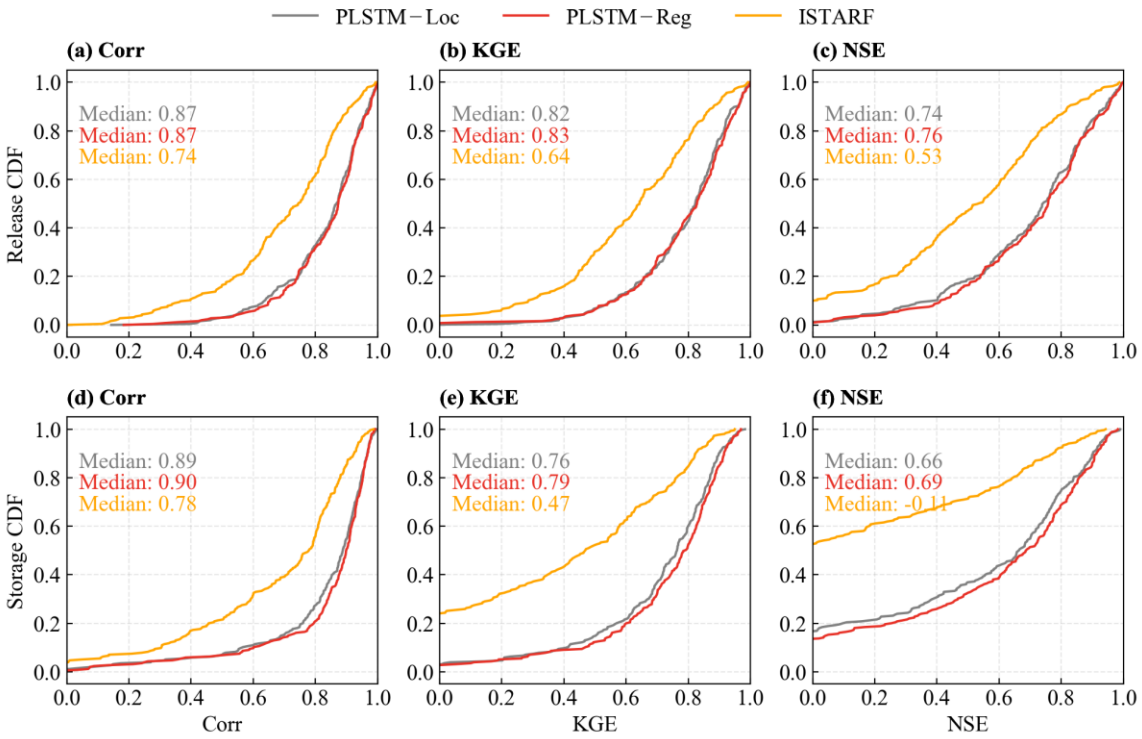


Figure S4. Comparison of release and storage long-term simulation performance under Experiment II across all reservoirs during the test period. CDFs display release simulation performance in the upper row (a–c) and storage simulation performance in the lower row (d–f). Median metric scores are annotated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

Figure S5

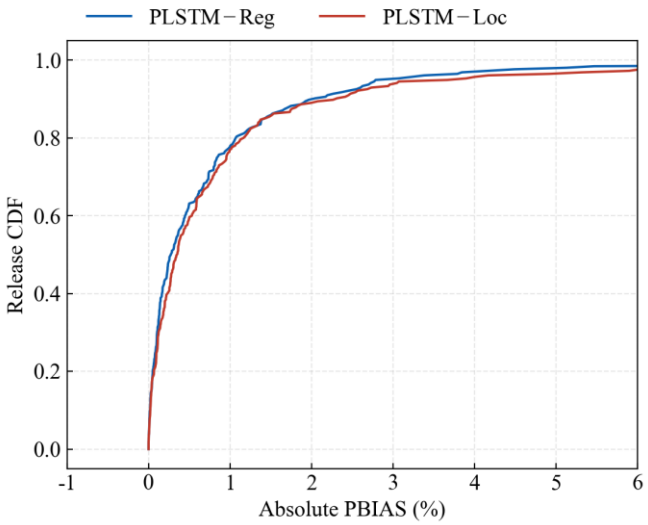


Figure S5. Empirical validation of long-term mass balance consistency across all reservoirs during the testing period. CDFs display the absolute Percent Bias (|PBIAS|) of simulated release for the regional (PLSTM-Reg) and local (PLSTM-Loc) models. Both models exhibit very low long-term release bias across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

Figure S6.

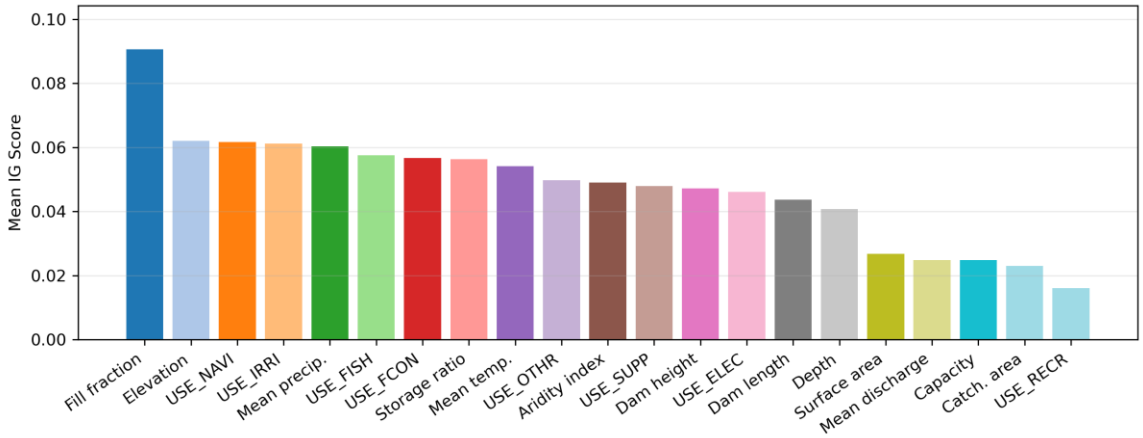


Figure S6. Mean Integrated Gradients (IG) importance scores of the 21 static attributes (see Table A1 in Appendix A) for reservoir release predictions. IG scores were normalized within each reservoir-day and then averaged across reservoirs, dates, and ensemble repeats. Higher values indicate greater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

Figure S7.

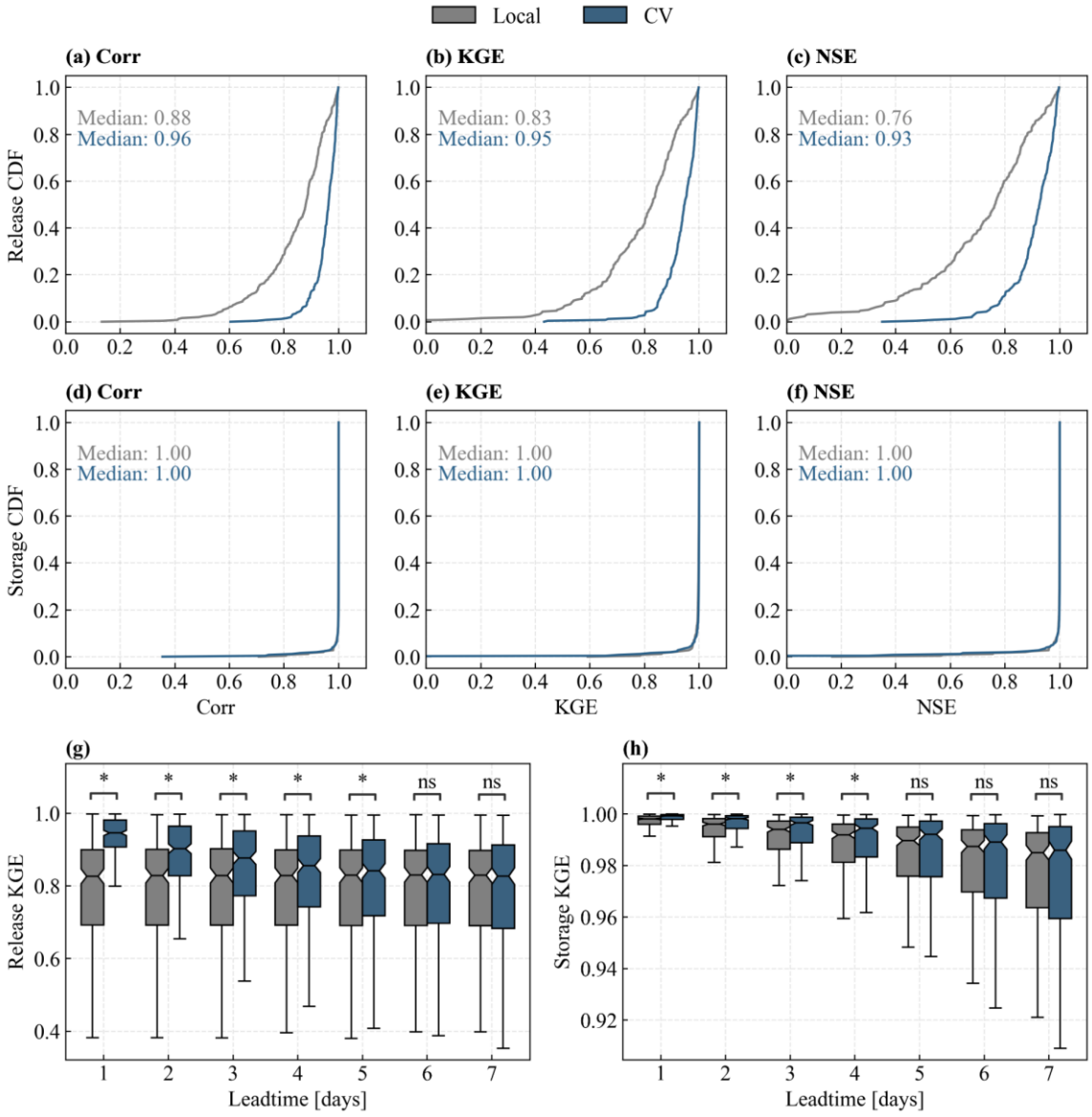


Figure S7. Performance comparison of short-term release and storage prediction among Cross-Validation (CV) and Local models during the test period. The CV model simulates data-scarce conditions via 5-fold cross-validation. CDFs display 1-day ahead performance for release in the upper row (a–c) and storage in the middle row (d–f) over the test period. Median values are annotated within the panels. The final row (g, h) illustrates KGE performance across lead times of 1 to 7 days. Asterisk (*) indicates a statistically significant difference (paired Wilcoxon signed-rank test, $p < 0.05$), while "ns" denotes not significant.

Figure S8.

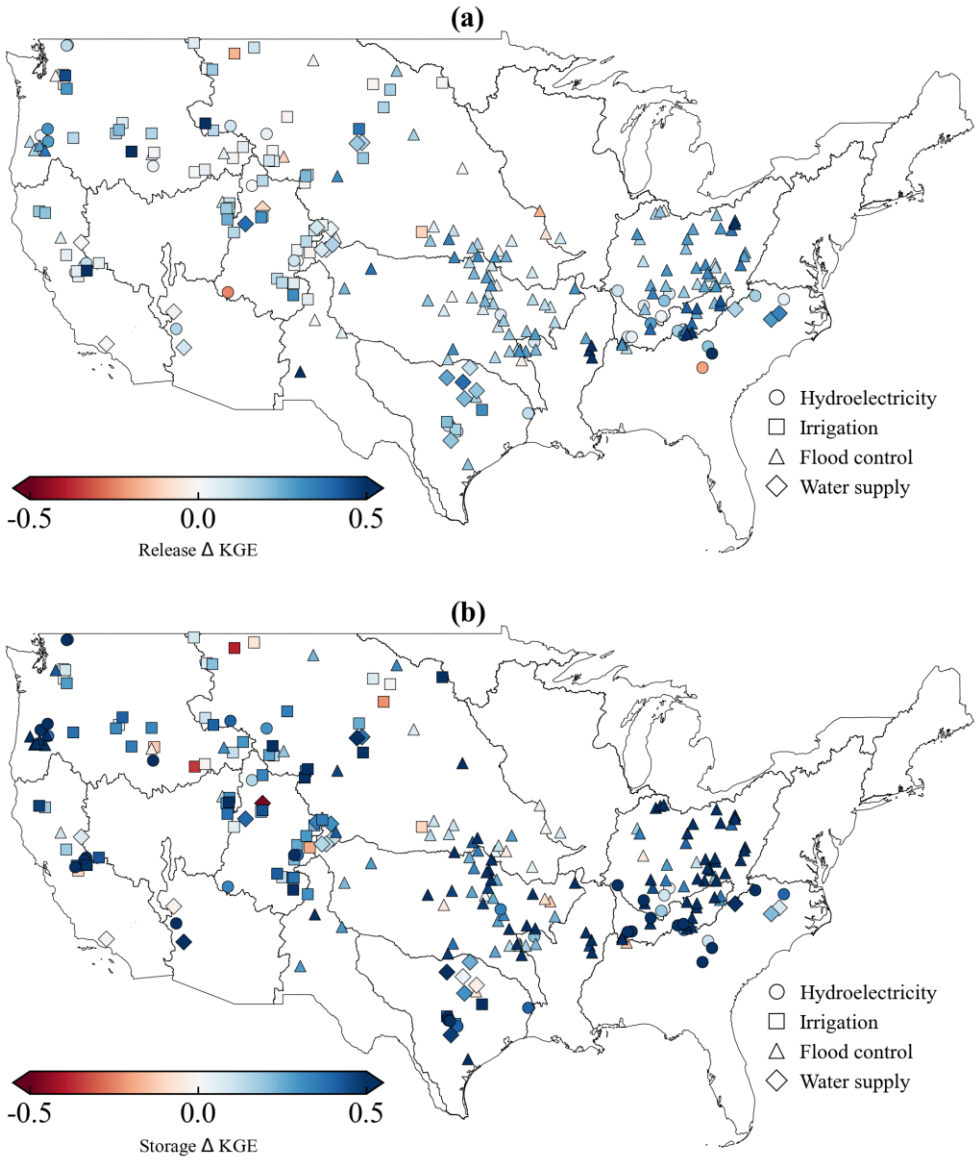


Figure S8. Spatial distribution of release (a) and storage (b) performance differences between the regional (PLSTM-Reg) and ZAJC models. Colors show the change in KGE, and marker shapes indicate primary operational purpose.

Figure S9.

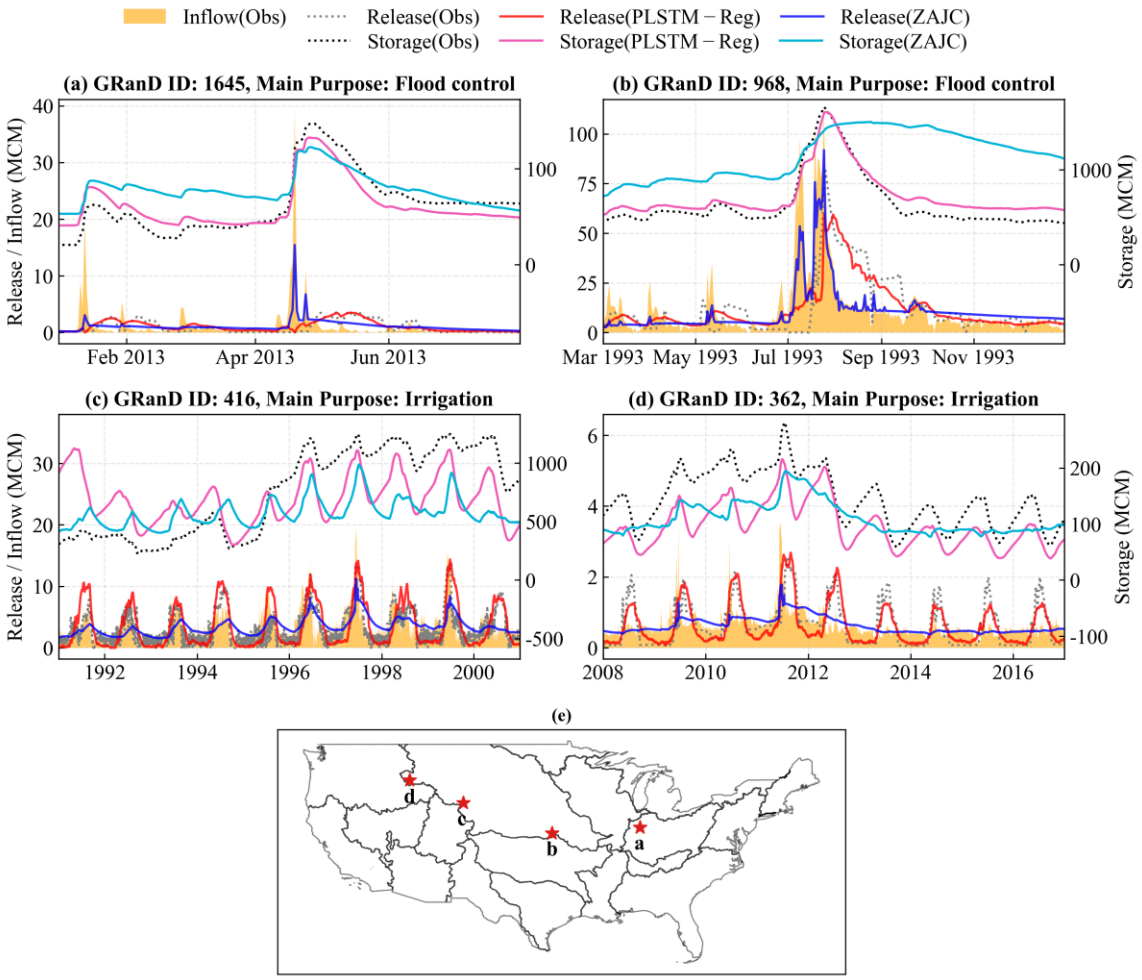


Figure S9. Comparison of operational behaviors between the regional model (PLSTM-Reg) and the generic rule-based benchmark (ZAJC) for selected ungauged reservoirs. (a)–(b) Flood Control Reservoirs, illustrating the regional model's ability to capture peak attenuation and storage buffering compared to the benchmark; (c)–(d) Irrigation Reservoirs, highlighting the regional model's capacity to infer seasonal release patterns; (e) The geographic locations of the selected reservoirs.

Figure S10.

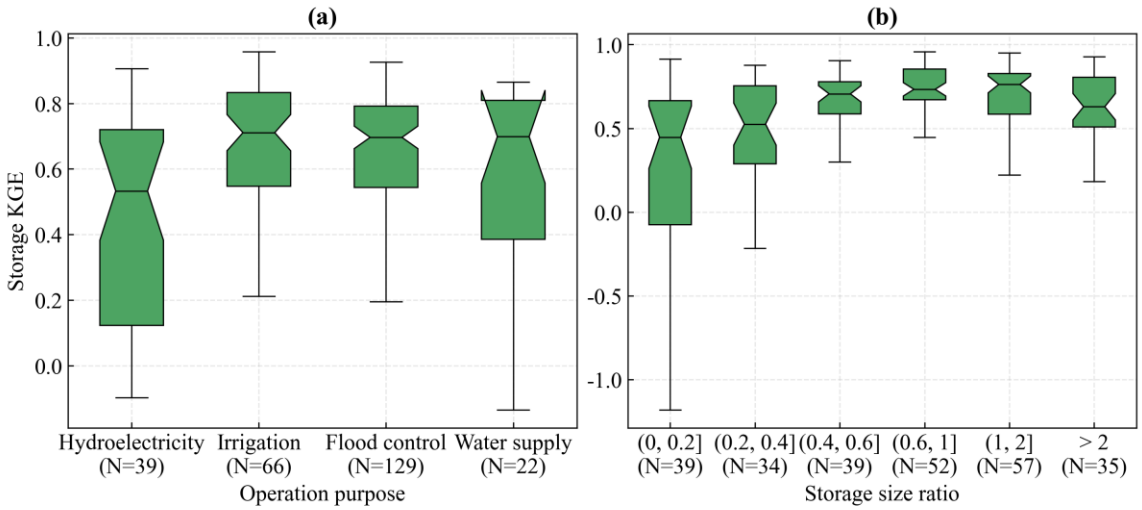


Figure S10. Storage performance of the remote-sensing-augmented regional model (PLSTM-Reg+RS) for historical reconstruction, disaggregated by (a) primary operational purpose and (b) storage size ratio. N denotes the number of reservoirs in each category.

Figure S11.

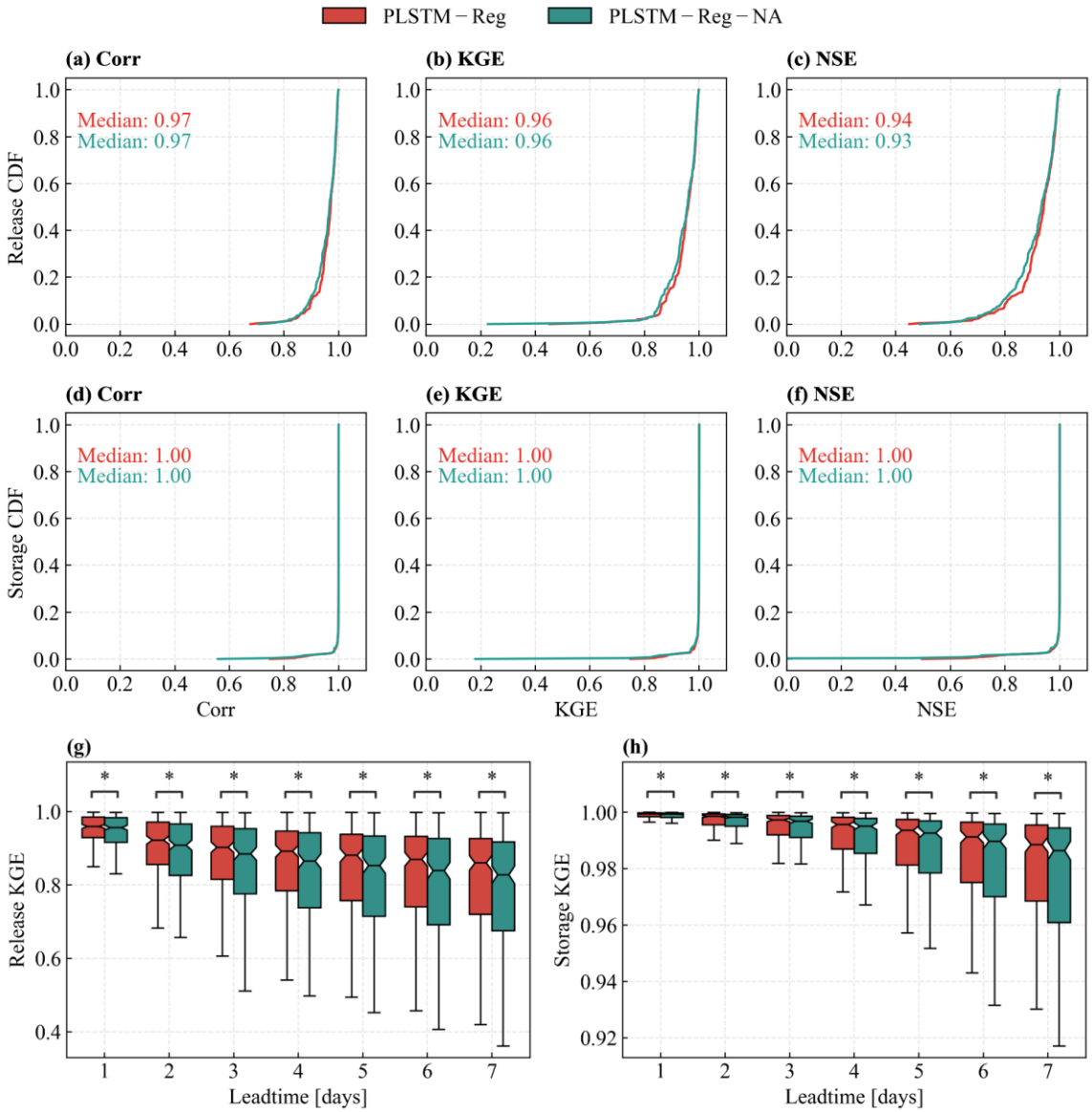


Figure S11. Performance comparison of short-term release and storage prediction between regional models with and without static attributes during the test period. Subplots (a–f) display CDFs of performance metrics for 1-day ahead predictions, where the upper row (a–c) represents release and the middle row (d–f) represents storage. Median performance values are annotated within these panels. The bottom row (g, h) illustrates the evolution of KGE scores across lead times ranging from 1 to 7 days. Asterisk (*) indicates a statistically significant difference (paired Wilcoxon signed-rank test, $p < 0.05$), while "ns" denotes not significant.

Table S1. Hyperparameters of network architecture for the PLSTM-Reg and PLSTM-Loc models.

Component	Hyperparameter	Regional	Local
Dynamic Embedding Network	Layer 1 neurons	16	8
	Layer 2 neurons	32	16
Static Embedding Network	Layer 1 neurons	16	N/A
	Layer 2 neurons	32	N/A
PLSTM cell	Hidden neurons	128	32
	Dropout rate	0.2	0.2

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.

Droppers, B., Franssen, W. H. P., van Vliet, M. T. H., Nijssen, B., and Ludwig, F.: Simulating human impacts on global water resources using VIC-5, Geosci. Model Dev., 13, 5029–5052, https://doi.org/10.5194/gmd-13-5029-2020, 2020.

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, Hydrol. Earth Syst. Sci., 12, 1007–1025, https://doi.org/10.5194/hess-12-1007-2008, 2008.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., and Reblitz-Richardson, O.: Captum: A unified and generic model interpretability library for PyTorch, https://doi.org/10.48550/arXiv.2009.07896, 16 September 2020., 2020.

mhpi: hydrodl2, https://github.com/mhpi/hydrodl2 (last access: 5 May 2026), 2026.

Pokhrel, Y., Hanasaki, N., Koirala, S., Cho, J., Yeh, P. J.-F., Kim, H., Kanae, S., and Oki, T.: Incorporating Anthropogenic Water Regulation Modules into a Land Surface Model, J. Hydrometeorol., 13, 255–269, https://doi.org/10.1175/JHM-D-11-013.1, 2012.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic Attribution for Deep Networks, ArXiv170301365 Cs, 2017.

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, J. Hydrol., 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), Hydrol. Earth Syst. Sci., 14, 1–24, https://doi.org/10.5194/hess-14-1-2010, 2010.

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, J. Hydrol., 548, 552–568, https://doi.org/10.1016/j.jhydrol.2017.03.022, 2017.