Supplement of Geosci. Model Dev., 19, 10719–10734, 2026
https://doi.org/10.5194/gmd-19-10719-2026-supplement
© Author(s) 2026. CC BY 4.0 License.



Geoscientific Model Development
Open Access
EGU

Supplement of

A hybrid physics–ML framework for integrating groundwater dynamics into land surface modeling: a ParFlow-based demonstration in headwater watersheds of the Pearl River Basin, China

Chen Yang et al.

Correspondence to: Chen Yang (yangch329@mail.sysu.edu.cn)

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

Supplementary Information

Table S1. Correlations between segment-scale precipitation characteristics and absolute water table depth (WTD) error metrics across the 12 720-h test segments. At each hour, precipitation was first spatially averaged over the model domain; the precipitation mean and standard deviation for each segment were then calculated from the resulting 720-h domain-mean precipitation time series. At each hour, absolute WTD errors across the model domain were summarized using their spatial mean, median, 75th percentile (q75), and 90th percentile (q90). "Time mean" denotes the temporal average of the corresponding spatial error statistic over the 720-h segment, whereas "final" denotes its value at the final hour of the segment. Pearson correlation coefficients (r) and Spearman rank correlation coefficients (ρ), calculated across the 12 segments, are reported together with their corresponding p values.

Precipitation metric	Error metric	Pearson r	Spearman ρ	Pearson p	Spearman p
precipitation mean (mm/h)	time mean mean	0.3873	0.3916	0.2136	0.2081
	time mean median	0.3017	0.4266	0.3406	0.1667
	time mean q75	**0.8648**	**0.8811**	0.0003	0.0002
	time mean q90	**0.6319**	**0.7203**	0.0275	0.0082
	final mean	0.2846	0.3287	0.3700	0.2969
	final median	-0.1069	0.1329	0.7410	0.6806
	final q75	**0.7285**	**0.7273**	0.0072	0.0074
	final q90	**0.5558**	**0.5804**	0.0606	0.0479
precipitation std (mm/h)	time mean mean	0.5404	0.5035	0.0697	0.0952
	time mean median	0.5515	0.5734	0.0630	0.0513
	time mean q75	**0.9440**	**0.9021**	<0.0001	0.0001
	time mean q90	**0.7945**	**0.8112**	0.0020	0.0014
	final mean	0.4084	0.4755	0.1875	0.1182
	final median	-0.0277	0.0559	0.9319	0.8629
	final q75	**0.7035**	**0.6643**	0.0107	0.0185
	final q90	**0.7083**	**0.6294**	0.0099	0.0283

Text S1. Pressure heads were standardized separately for each vertical layer as $\tilde{p}_k = (p_k - \mu_k)/\sigma_k$, where μ_k and σ_k were calculated from the training period. The inverse transformation, $p_k = \sigma_k \tilde{p}_k + \mu_k$, restores pressure heads to physical units. We evaluated the vertical total-head gradient between adjacent cell centers as

$$G = \frac{p_{upper} - p_{lower}}{\Delta z} + 1,$$

where elevation increases upward and Δz is the actual distance between the cell centers. Figures S1 and S2 show that vertical hydraulic-gradient errors are relatively small in the early stage and generally increase as errors accumulate during the continuous hybrid simulation. Figure S3 shows some differences between dry and saturated conditions, but no distinct pattern of gradient deterioration that can be attributed to layer-wise standardization.

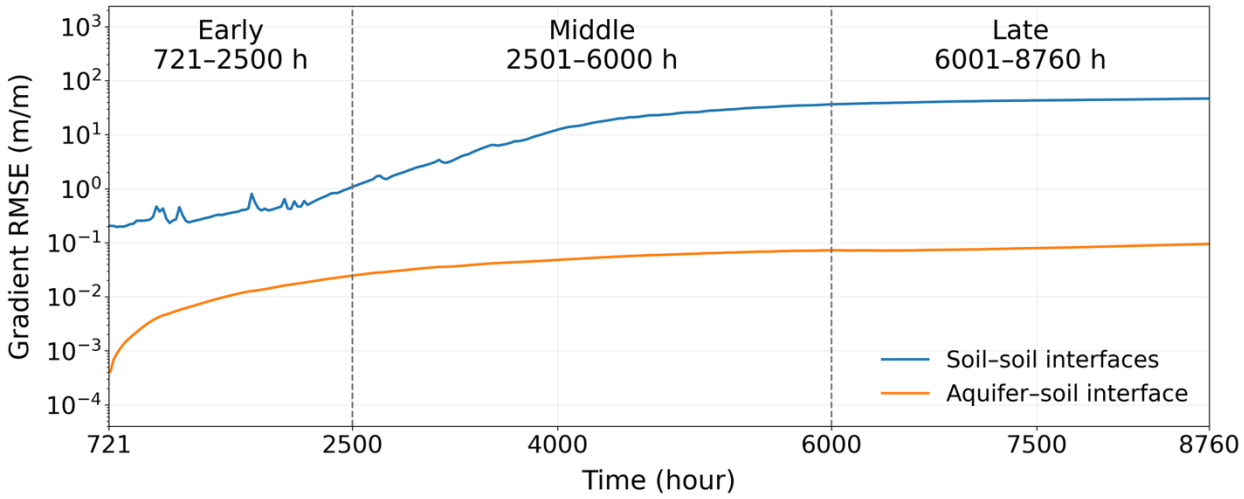


Figure S1. Temporal evolution of vertical hydraulic-gradient RMSE between the hybrid and physics-based simulations over hours 721–8760. The curves represent the nine soil–soil interfaces pooled together and the single aquifer–soil interface. Soil–soil RMSE assigns equal weight to all horizontal grid cells and all nine interfaces, whereas aquifer–soil RMSE is calculated across all horizontal grid cells at that interface. Daily values are obtained by averaging the hourly squared RMSE values and taking the square root. Vertical dashed lines at hours 2500 and 6000 separate the early, middle, and late periods; daily bins crossing these boundaries are split accordingly. Gradients are calculated from pressure heads in physical units using the actual distances between adjacent cell centers. The vertical axis is logarithmic.

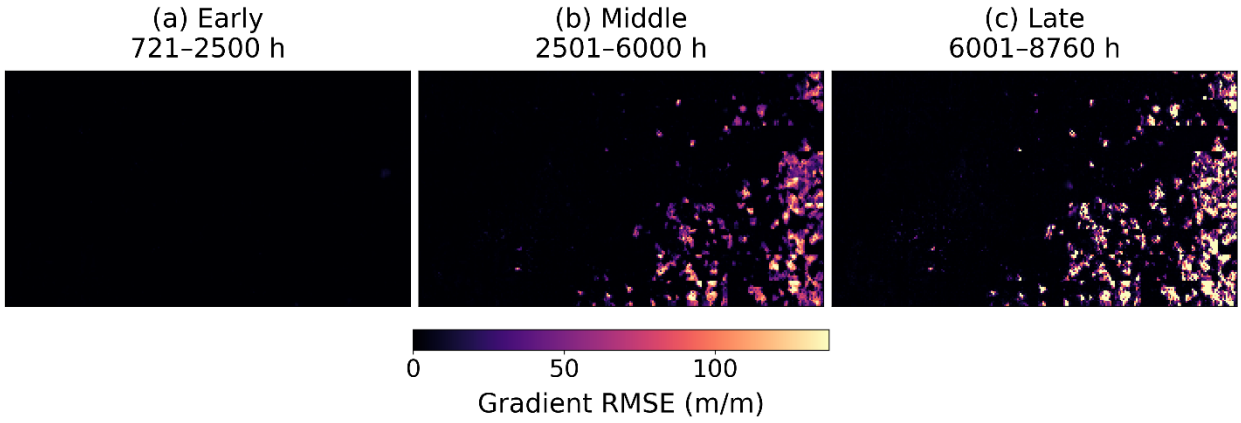


Figure S2. Spatial distributions of vertical hydraulic-gradient RMSE between the hybrid and physics-based simulations during the (a) early (hours 721–2500), (b) middle (hours 2501–6000), and (c) late (hours 6001–8760) periods. At each horizontal grid cell, RMSE is calculated over all hourly samples and the nine soil–soil interfaces within each period, with equal weighting. Gradients are evaluated from pressure heads in physical units using the actual distances between adjacent cell centers. All panels share the same color scale, with its upper limit set to the pooled 99th percentile of the three RMSE maps; larger values are displayed using the maximum color. The aquifer–soil interface is not included because its vertical hydraulic-gradient RMSE remains substantially smaller than that of the soil–soil interfaces throughout the simulation (Figure S1); Figure S2 therefore focuses on the nine soil–soil interfaces that dominate the spatial error pattern.

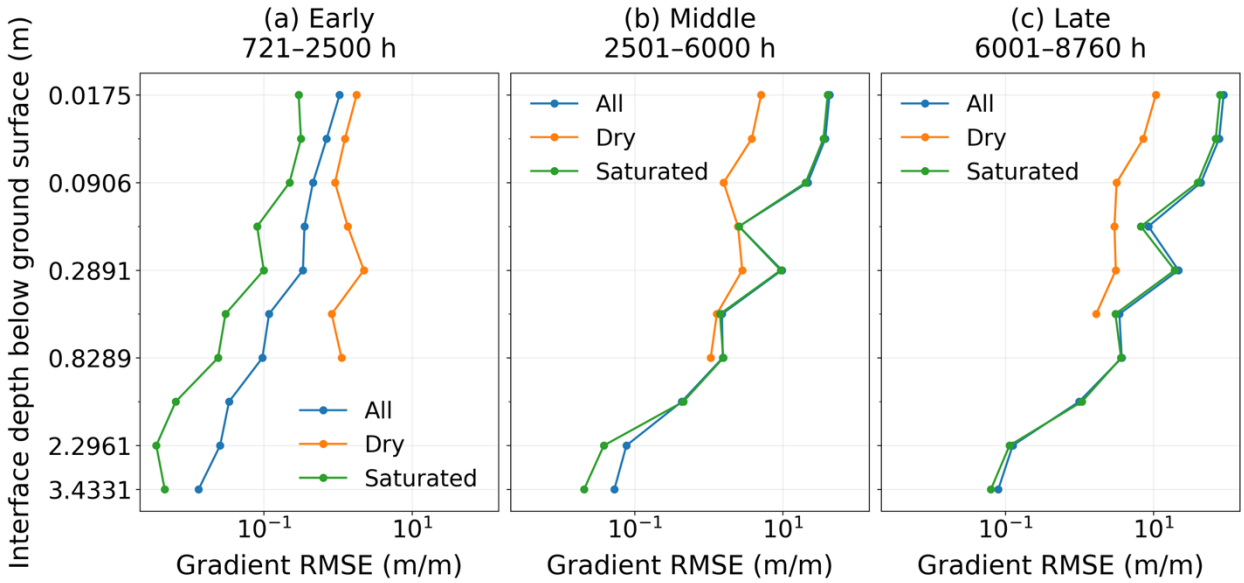


Figure S3. Vertical hydraulic-gradient RMSE between the hybrid and physics-based simulations during the (a) early (hours 721–2500), (b) middle (hours 2501–6000), and (c) late (hours 6001–8760) periods. Curves represent all samples, dry samples, and saturated samples. Moisture classes are defined using the physics-based simulation: dry samples require effective saturation ≤ 0.2 in both adjacent cells, whereas saturated samples require nonnegative pressure heads in both cells. RMSE is calculated separately at each interface over all corresponding grid-cell and hourly samples within each period. Gradients are calculated from pressure heads in physical units using the actual distances between adjacent cell centers. Interface depths are displayed at equal spacing for readability, and the horizontal axis is logarithmic. Missing points indicate the absence of qualifying samples.

Text S2. To further explore the potential for improving surrogate performance, we conducted additional experiments by introducing attention mechanisms into the model architecture. Inspired by recent spatiotemporal attention studies (Lim et al., 2021; Zheng et al., 2021; Xiong et al., 2024), we implemented a temporal self-attention module between the final hidden layer and the output layer. Prior to applying attention, the spatial dimensions of the hidden states were compressed while preserving channel-wise information. Two alternative modes were considered. In the pooling mode, global adaptive average pooling was applied to each hidden feature map, reducing the spatial grids to a channel-wise representation. In the xyconv mode, two depthwise convolutions with kernel sizes and strides equal to the patch size were used to compress the spatial dimensions to a single location—first along the y-direction and then along the x-direction—yielding a learnable spatial aggregation of each feature channel. Both modes were retained in the implementation to facilitate flexible adaptation to different applications.

The resulting channel embeddings across time steps were then fed into a temporal self-attention module, where the current timestep was used as the query and all preceding timesteps (including the current one) as keys and values. The attention weights were used to reweight the stacked hidden states across time, assigning a scalar importance to each timestep's spatial feature map. The weighted hidden states were then aggregated to produce an attention-enhanced spatiotemporal representation, which was subsequently passed through the output convolution layer for prediction. A learnable scaling factor was applied to modulate the contribution of the attention-enhanced features before being added to the original prediction.

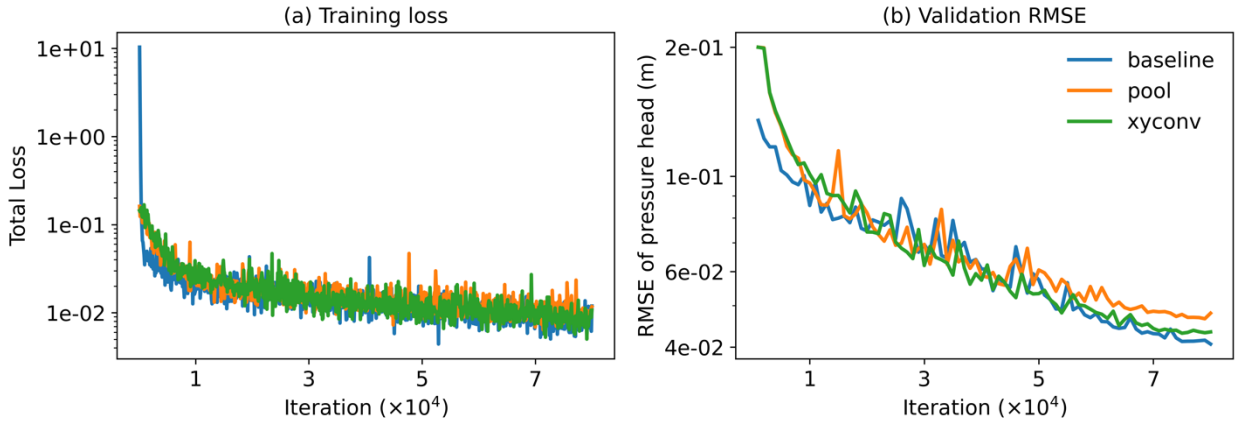


Figure S4. Training loss and validation RMSE of the ParFlow surrogate model for different architectural configurations. (a) Evolution of the training loss. (b) Validation RMSE of pressure head. The baseline model and variants with alternative spatial aggregation strategies (pooling and xyconv) exhibit similar convergence behavior and comparable validation performance.

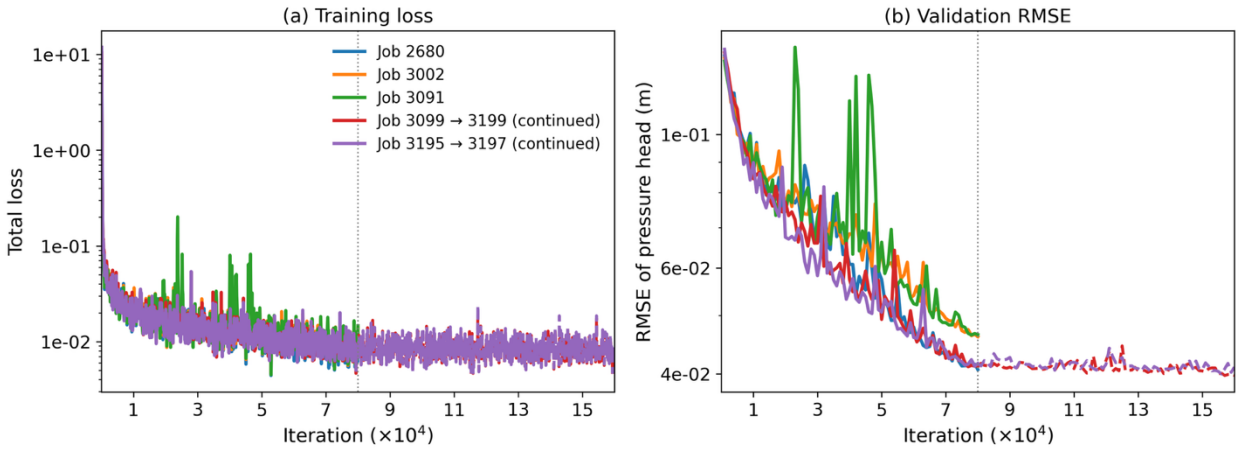


Figure S5. Training loss and validation performance across five repeated training runs of the ParFlow surrogate model conducted using the same random seed and otherwise identical training settings. (a) Evolution of the training loss. (b) Validation RMSE of pressure head. Job numbers identify the individual training runs. Jobs 3099 and 3195 were continued beyond 80,000 iterations as Jobs 3199 and 3197, respectively, and the vertical dotted line marks the original 80,000-iteration endpoint. The checkpoint from Job 2680 at 80,000 iterations was selected for the subsequent evaluation.

Text S3. For both earlier checkpoints (70,000 and 75,000), predictions remained close to those obtained with the selected checkpoint (80,000), with mean absolute differences of 0.024–0.031 m for WTD, 0.0020–0.0024 for soil moisture, and 0.475–0.491 W m^{-2} for latent heat flux across all evaluated grid cells and hours in Segment 9, indicating limited overall sensitivity to checkpoint selection within this late-training interval.

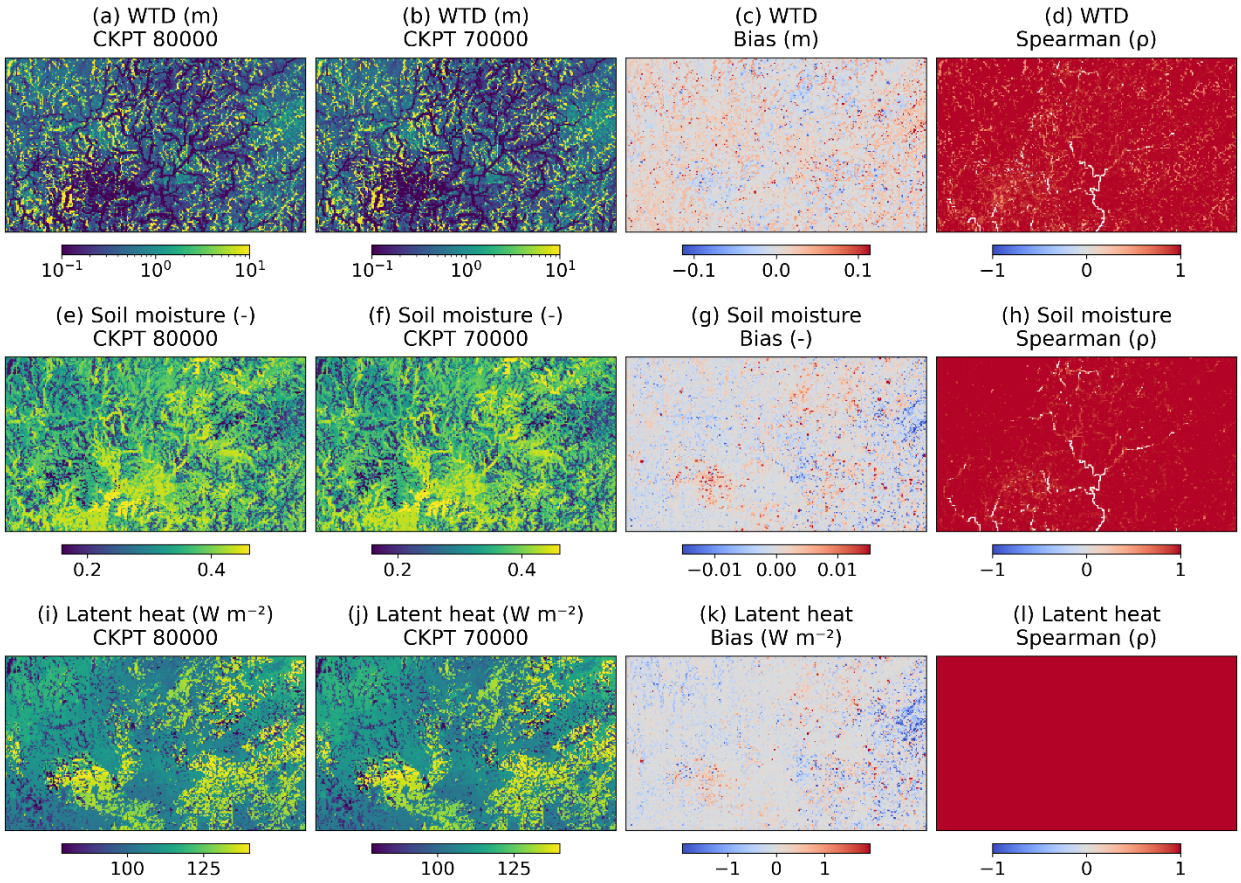


Figure S6. Spatial comparison of Segment 9 predictions using CKPT 70000 and the reference CKPT 80000. Rows show water table depth, soil moisture, and latent heat flux; columns show the reference and comparison temporal means, bias (CKPT 70000 minus CKPT 80000), and temporal Spearman correlation over hours 5761–6480. CKPT denotes checkpoint.

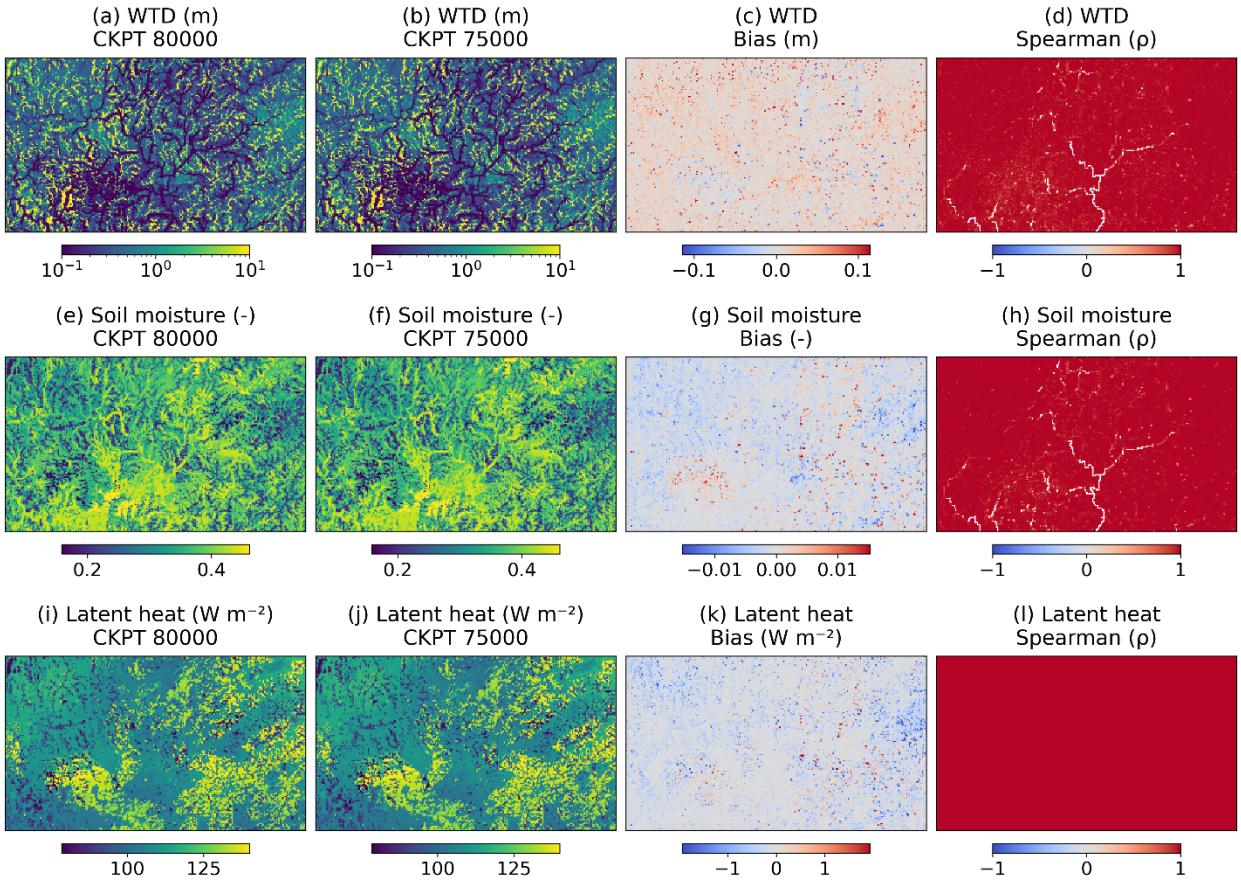


Figure S7. Same as Figure S6, but comparing CKPT 75000 with the reference CKPT 80000. Bias is calculated as CKPT 75000 minus CKPT 80000.

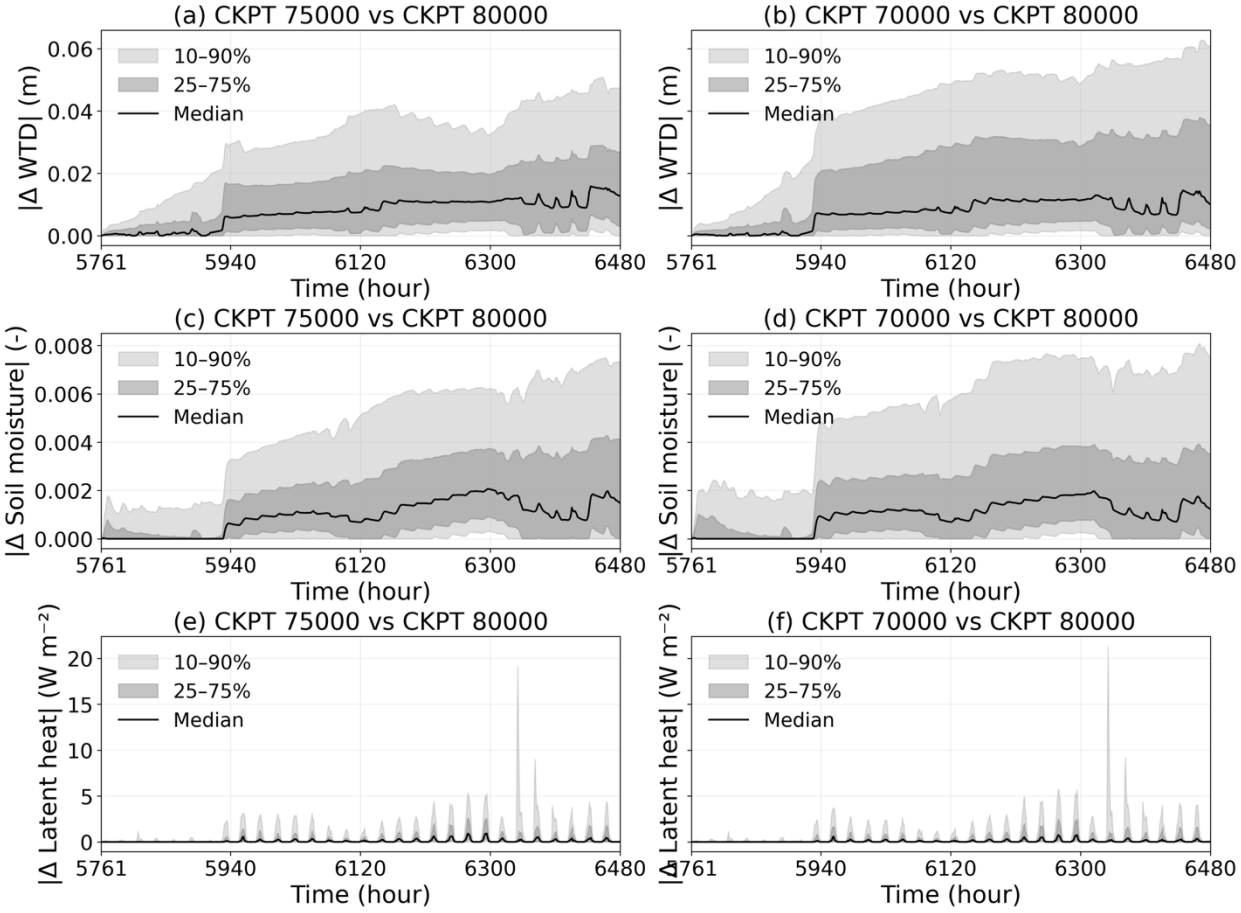


Figure S8. Temporal evolution of absolute differences in water table depth, soil moisture, and latent heat flux between each earlier checkpoint (CKPT 75000 and CKPT 70000) and the reference CKPT 80000 during Segment 9. Lines show spatial medians, and shaded bands indicate the spatial 25th–75th and 10th–90th percentile ranges. Hours indicate their positions within the simulation year.

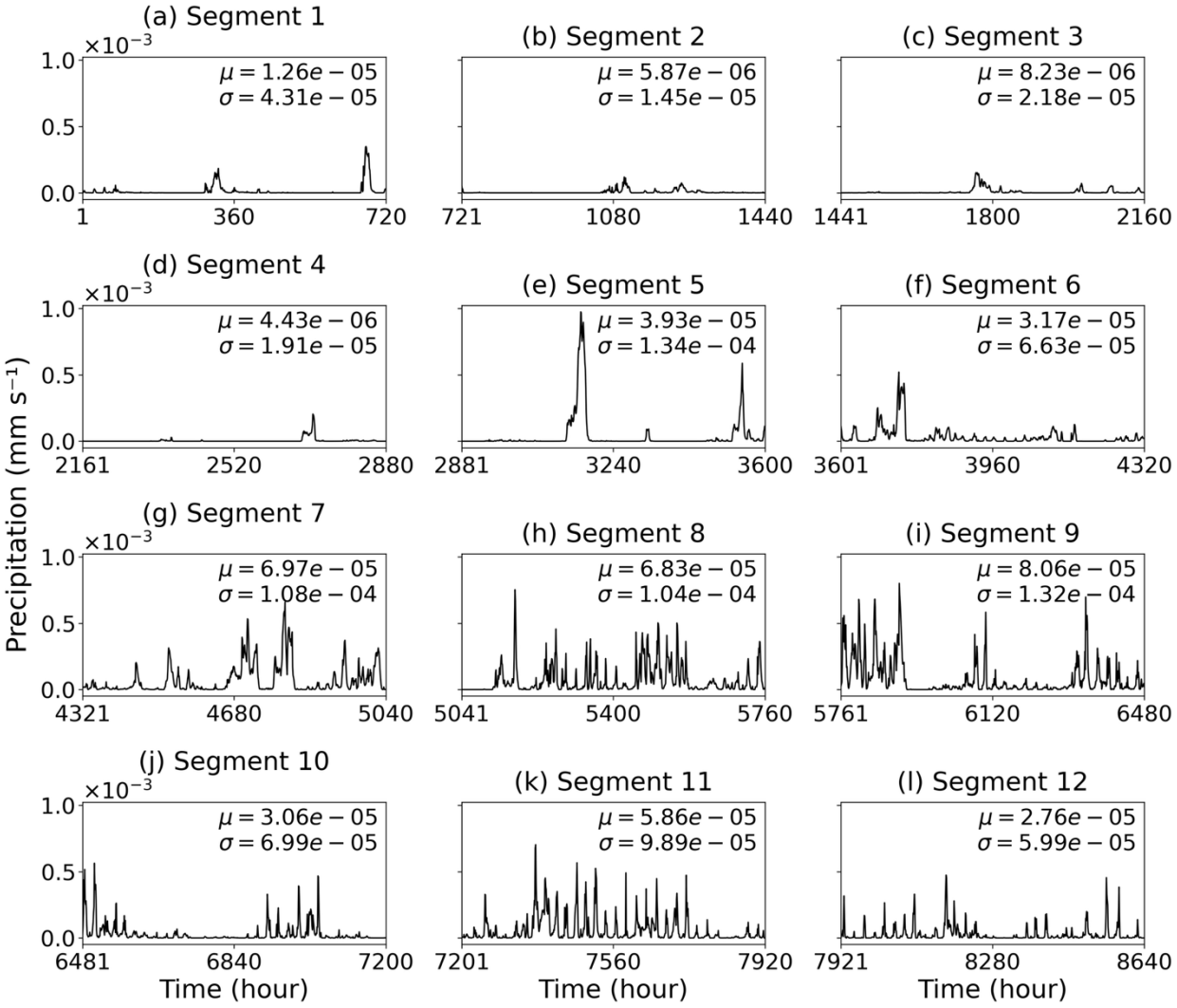


Figure S9. Segmented time series of precipitation over the simulation period (WY2021). The hourly precipitation series is divided into 12 consecutive 720-h segments (a–l), each spanning an equal duration. For each segment, the temporal variability is illustrated, and the mean (μ) and standard deviation (σ) are annotated to quantify differences in precipitation intensity and variability across periods. The segmentation highlights the intermittent and highly heterogeneous nature of precipitation, with distinct periods of low activity and intense events.

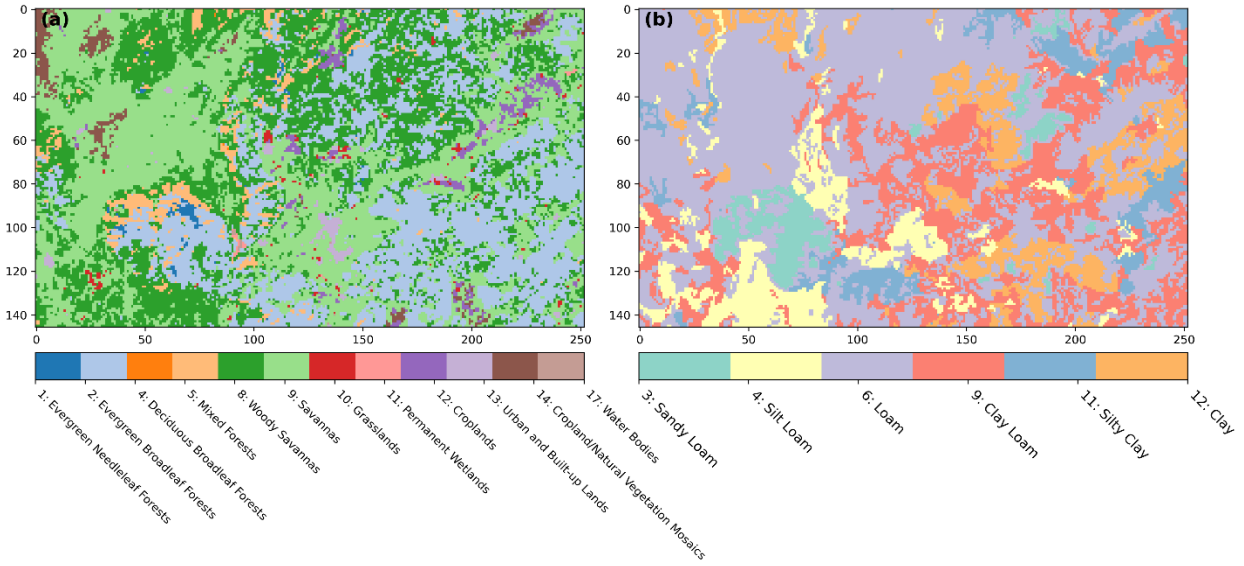


Figure S10. Spatial distributions of land cover and soil texture in the study domain. (a) Land cover classification. (b) Soil texture classes.

Text S4. We evaluated two complementary surface-water quantities: total surface water storage over the model domain and streamflow at selected locations. Streamflow was evaluated at the model outlet (Point A) and a nearby grid cell approximately 3.47 km away (Point B; Figure S11). Surface-water storage and streamflow were derived consistently from the pressure fields of the physics-based and hybrid simulations. NSE and KGE were calculated using the physics-based simulation as the reference for each corresponding time series.

Agreement depended on both the quantity evaluated and the simulation period (Figure S12). In Segment 1, the hybrid model closely reproduced surface storage and streamflow at both points, with NSE values exceeding 0.98. In Segment 9, surface storage remained comparatively well reproduced (NSE = 0.910; KGE = 0.770), whereas streamflow agreement was weaker, particularly in reproducing the magnitude of the peak flows. During the early stage of the continuous simulation, some broad temporal features were retained, although quantitative agreement was limited. Both storage and streamflow showed substantial deterioration during the middle and late stages, consistent with the progressive divergence of the predicted pressure fields. The contrast between storage and streamflow reflects their different sensitivities to pressure-field errors. Total surface storage integrates positive surface pressure over the model domain, which can reduce the influence of localized discrepancies when their contributions offset one another. Streamflow at a selected grid cell depends more directly on local surface-water depth and its nonlinear relationship with discharge. Consequently, reproducing broad spatial pressure patterns does not necessarily ensure accurate flow at a particular location. Spatial aggregation also offers no protection against widespread systematic errors, as illustrated by the deterioration of storage estimates during the later continuous simulation. The weaker point-scale streamflow agreement underscores the importance of local surface-pressure errors for discharge prediction. Nevertheless, the agreement in surface storage across both segments and in streamflow during Segment 1 indicates that the hybrid model captures key surface-water responses. This is consistent with the framework's primary focus on spatial groundwater patterns and their interactions with land-surface processes.

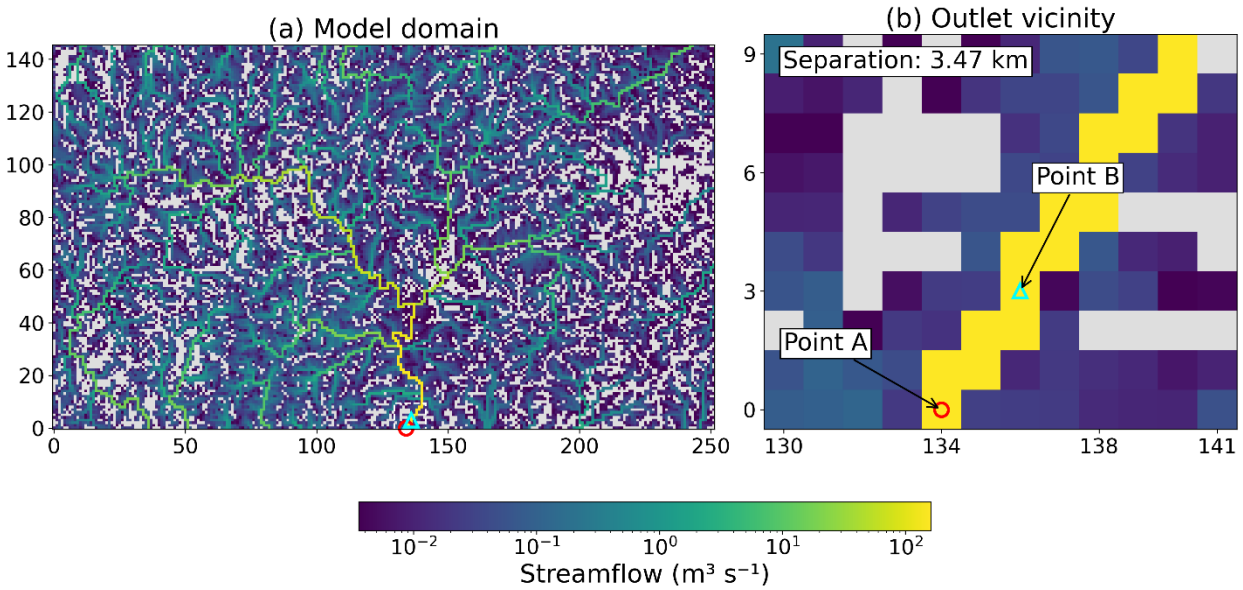


Figure S11. Location map. Locations of the two streamflow evaluation points: (a) the model domain and (b) an enlarged view near the outlet. Point A denotes the outlet, and Point B is a nearby grid cell approximately 3.47 km away. Colors indicate mean streamflow from the physics-based simulation.

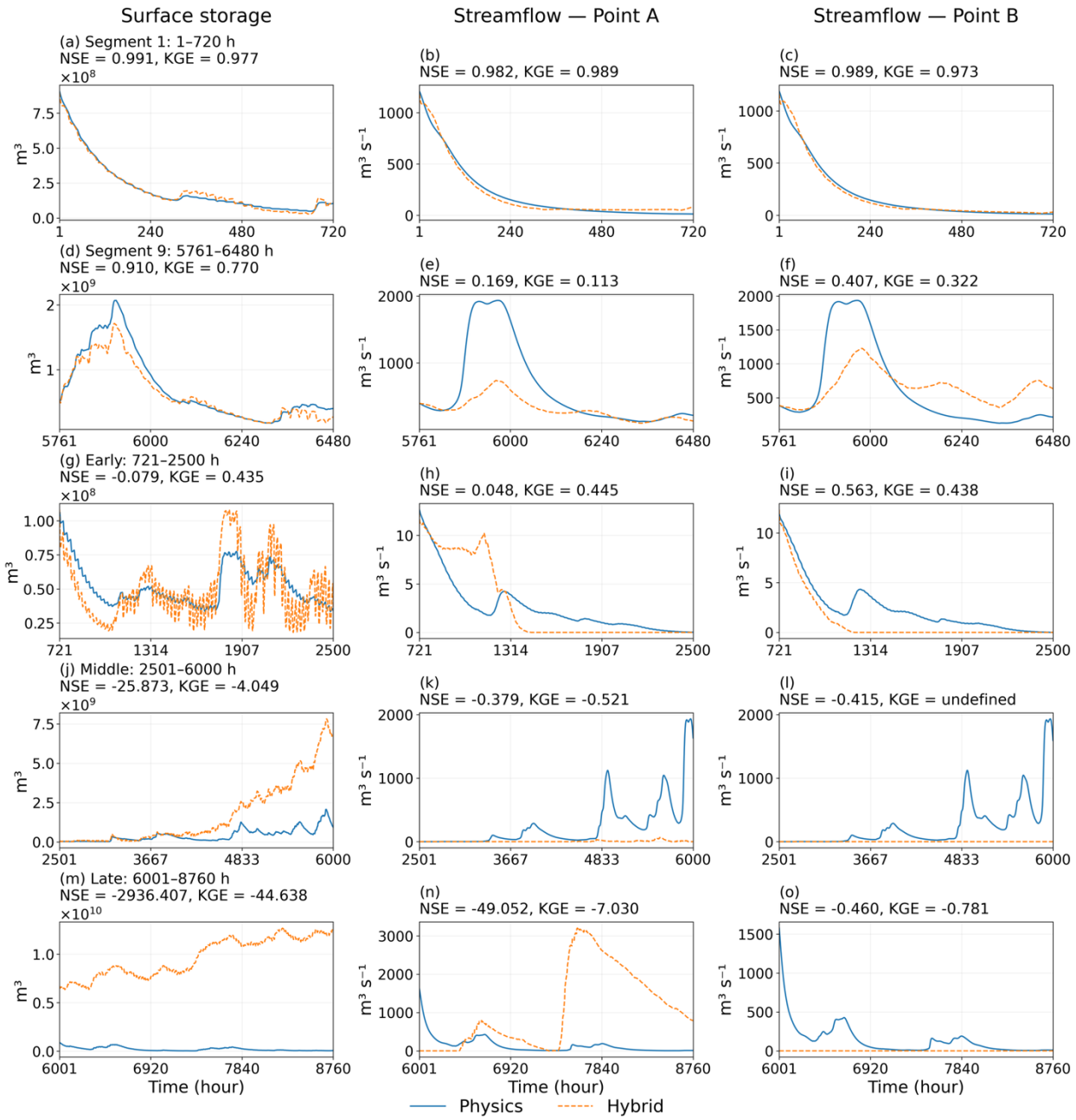


Figure S12. Time-series comparison. Comparison of surface water storage and streamflow at Points A and B between the physics-based and hybrid simulations for Segments 1 and 9 and three periods of the continuous 721–8760 h simulation. NSE and KGE are reported for each comparison.

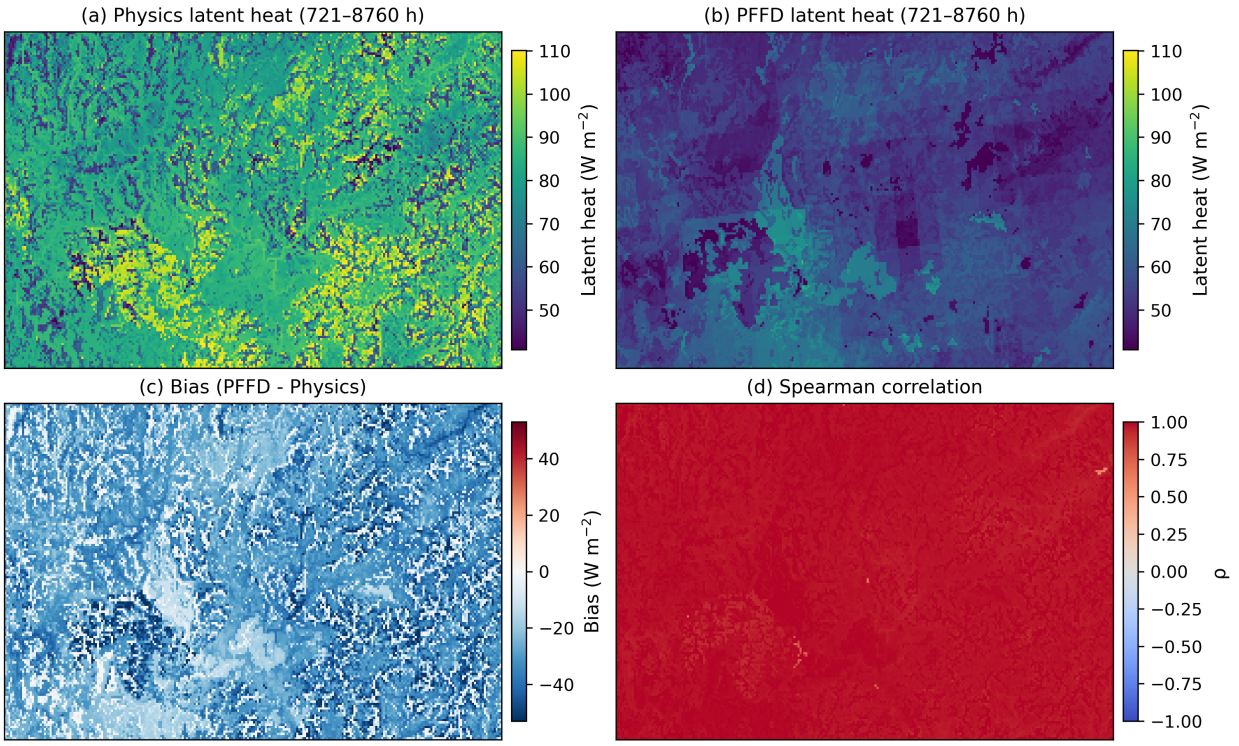


Figure S13. Spatial comparison of mean latent heat flux and model agreement between the physics-based and free-drainage (PFFD) configurations over the 8040-h evaluation period (hours 721–8760). (a) Mean latent heat flux from the physics-based model. (b) Mean latent heat flux from the PFFD configuration. (c) Bias (PFFD − Physics). (d) Spearman rank correlation coefficient (ρ) computed from the temporal series at each grid cell.

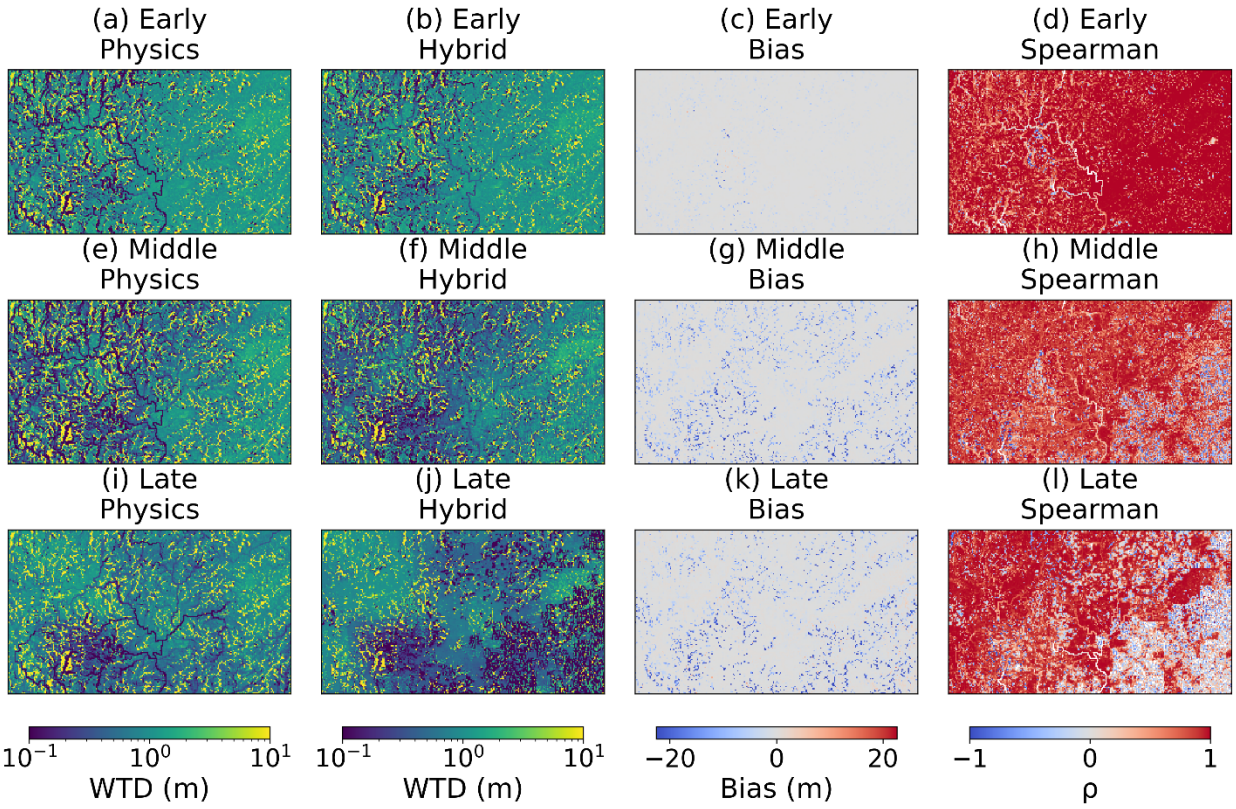


Figure S14. Temporal evolution of spatial patterns of water table depth (WTD) and model agreement across different prediction stages. Three periods are shown: early (hours 721–2500), middle (hours 2501–6000), and late (hours 6001–8760) stages of the simulation. For each period, columns represent the physics-based model, the hybrid model, the bias (Hybrid − Physics), and the Spearman correlation coefficient (ρ), respectively.

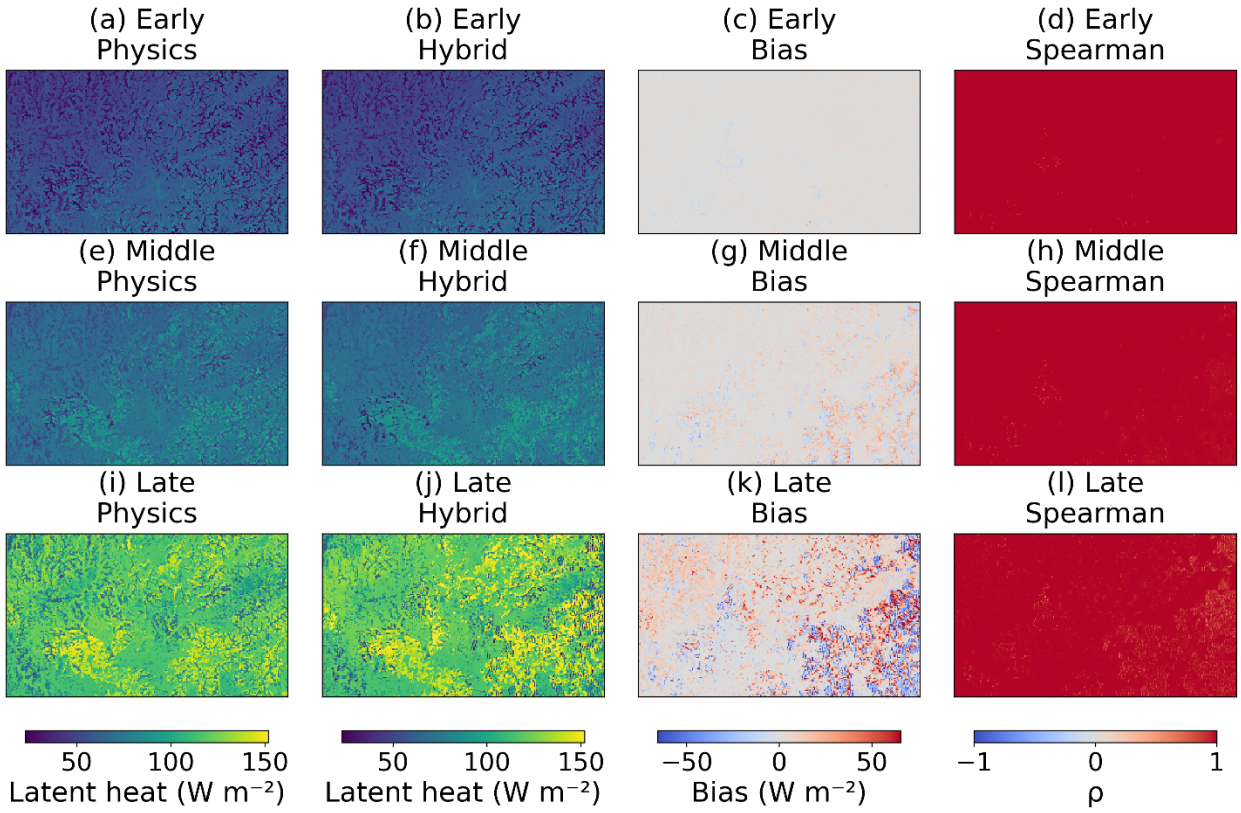


Figure S15. Temporal evolution of spatial patterns of latent heat flux and model agreement across different prediction stages. Three periods are shown: early (hours 721–2500), middle (hours 2501–6000), and late (hours 6001–8760) stages of the simulation. For each period, columns represent the physics-based model, the hybrid model, the bias (Hybrid − Physics), and the Spearman correlation coefficient (ρ), respectively.

References

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T.: Temporal Fusion Transformers for interpretable multi-horizon time series forecasting, International Journal of Forecasting, 37, 1748-1764, https://doi.org/10.1016/j.ijforecast.2021.03.012, 2021.

Xiong, Z., Zhang, Z., Gui, H., Zhu, P., Sun, Y., Zhou, X., Xiao, K., and Xin, Q.: Predicting time series of vegetation leaf area index across North America based on climate variables for land surface modeling using attention-enhanced LSTM, International Journal of Digital Earth, 17, 2372317, https://doi.org/10.1080/17538947.2024.2372317, 2024.

Zheng, H., Lin, F., Feng, X., and Chen, Y.: A Hybrid Deep Learning Model With Attention-Based Conv-LSTM Networks for Short-Term Traffic Flow Prediction, IEEE Transactions on Intelligent Transportation Systems, 22, 6910-6920, https://doi.org/10.1109/tits.2020.2997352, 2021.