Supplementary Materials

0.0.1. Statistical Significance Testing

To rigorously evaluate whether the performance improvements of the proposed ST-GNN model are statistically significant, we conducted paired statistical tests. For each watershed and each baseline model, we compared the performance metrics (RMSE and DC) obtained from five independent runs with different random seeds. We first assessed the normality of the performance differences using the Shapiro-Wilk test. When the differences followed a normal distribution ($p > 0.05$), we employed paired t-tests; otherwise, we used the non-parametric Wilcoxon signed-rank test. All tests were one-tailed, with the alternative hypothesis that ST-GNN performs better (lower RMSE, higher DC) than the baseline. Statistical significance was determined at the $\alpha = 0.05$ level.

Table S1. Statistical significance (p-values) of performance differences between ST-GNN and baseline models for 6-hour ahead streamflow prediction across four watersheds. P-values are from paired tests comparing metrics from 5 independent runs. An asterisk (*) in the main table denotes statistical significance ($p < 0.05$).

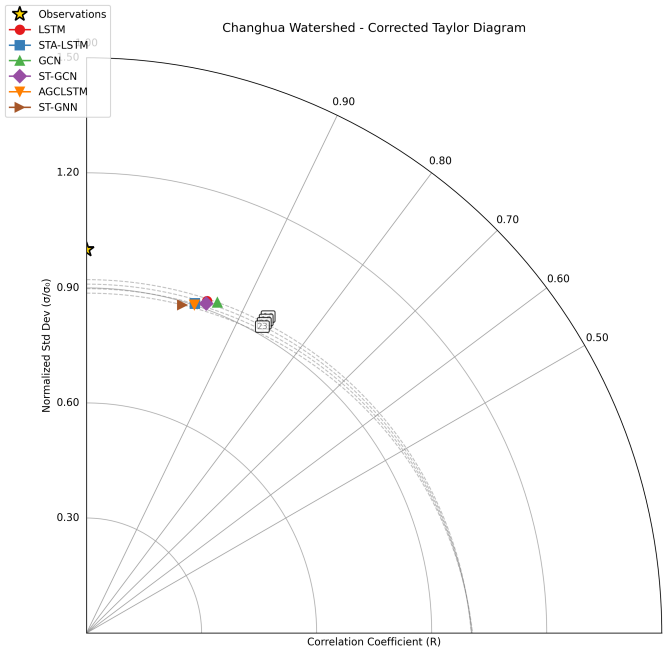
Watershed	Baseline Model	RMSE		DC (NSE)	
		p-value	Sig.	p-value	Sig.
Changhua	LSTM	1.2×10^{-4}	*	9.8×10^{-5}	*
	STA-LSTM	3.5×10^{-3}	*	4.1×10^{-3}	*
	GCN	8.7×10^{-5}	*	6.5×10^{-4}	*
	ST-GCN	2.1×10^{-3}	*	1.9×10^{-3}	*
	AGCLSTM	4.8×10^{-2}	*	3.9×10^{-2}	*
Chenhe	LSTM	6.3×10^{-5}	*	2.1×10^{-4}	*
	STA-LSTM	1.4×10^{-4}	*	3.0×10^{-4}	*
	GCN	5.5×10^{-5}	*	1.7×10^{-4}	*
	ST-GCN	2.8×10^{-3}	*	1.5×10^{-3}	*
	AGCLSTM	9.2×10^{-4}	*	2.2×10^{-3}	*
Pingyao	LSTM	$< 10^{-5}$	*	$< 10^{-5}$	*
	STA-LSTM	2.2×10^{-4}	*	3.3×10^{-4}	*
	GCN	$< 10^{-5}$	*	$< 10^{-5}$	*
	ST-GCN	1.1×10^{-3}	*	8.7×10^{-4}	*
	AGCLSTM	1.9×10^{-2}	*	2.4×10^{-2}	*
Daheba	LSTM	3.8×10^{-5}	*	7.9×10^{-5}	*
	STA-LSTM	6.1×10^{-4}	*	1.2×10^{-3}	*
	GCN	4.2×10^{-5}	*	1.1×10^{-4}	*
	ST-GCN	3.3×10^{-3}	*	4.5×10^{-3}	*
	AGCLSTM	2.7×10^{-2}	*	3.1×10^{-2}	*

0.0.2. Taylor Diagrams for 6-hour Ahead Prediction

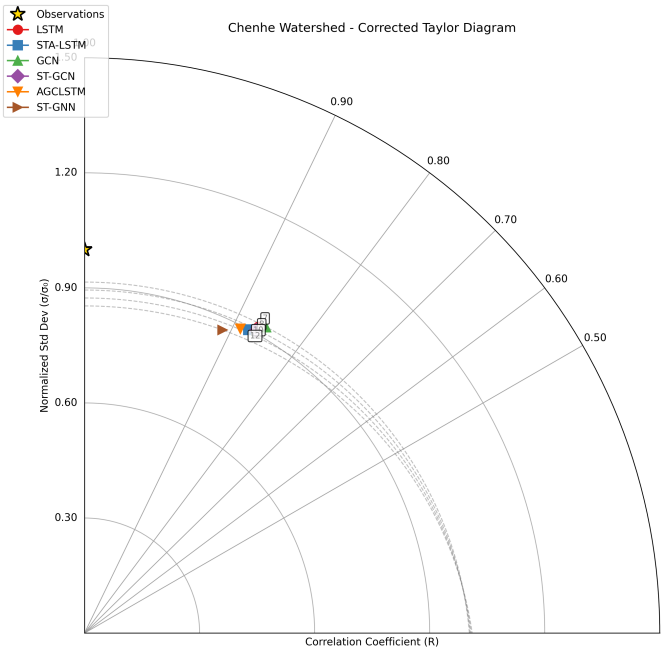
0.0.3. Supplementary Figures: Ablation Study Visualization

To complement the quantitative analysis in Figure 15 that visually represent the performance distributions of all model variants across the four watersheds. The combined plot displays both RMSE (top panel) and DC (bottom panel) metrics, with each point representing one of five independent runs per model-watershed combination.

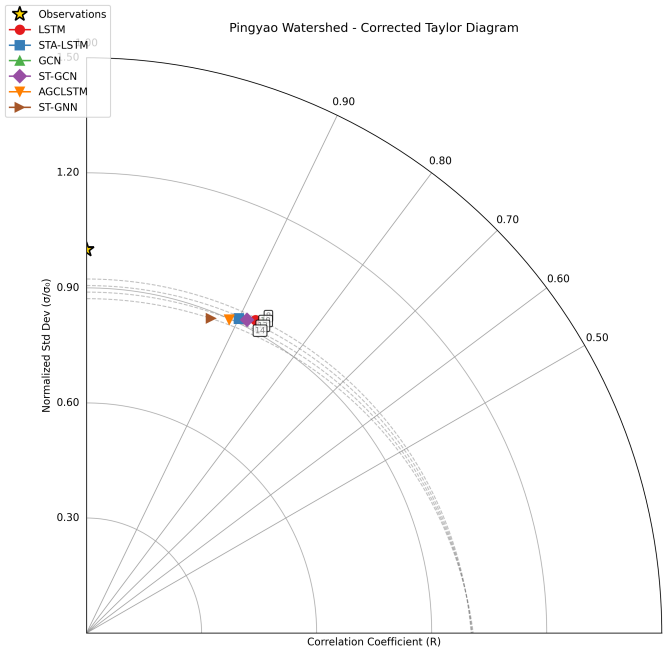
Three key patterns emerge from the visualization: First, the complete ST-GNN model consistently clusters at the optimal performance regions (lowest RMSE, highest DC) with tight point distributions, indicating stable performance across multiple runs. Second, a clear performance degradation hierarchy is visible: removing the feature enhancement module (ST-GNN-rfe, rightmost) causes the largest performance shift, followed sequentially by the future rainfall input (ST-GNN-rfi), Graph-LSTM (ST-GNN-rglstm), and spatial GCN (ST-GNN-rgcn). Third, the color-coded points reveal consistent performance patterns across watersheds—each model variant shows similar relative positioning regardless of watershed type (humid: Changhua and Pingyao; semi-arid: Chenhe and Daheba). This visual consistency strengthens the conclusion that the importance ranking of modules is robust to hydrological variations.



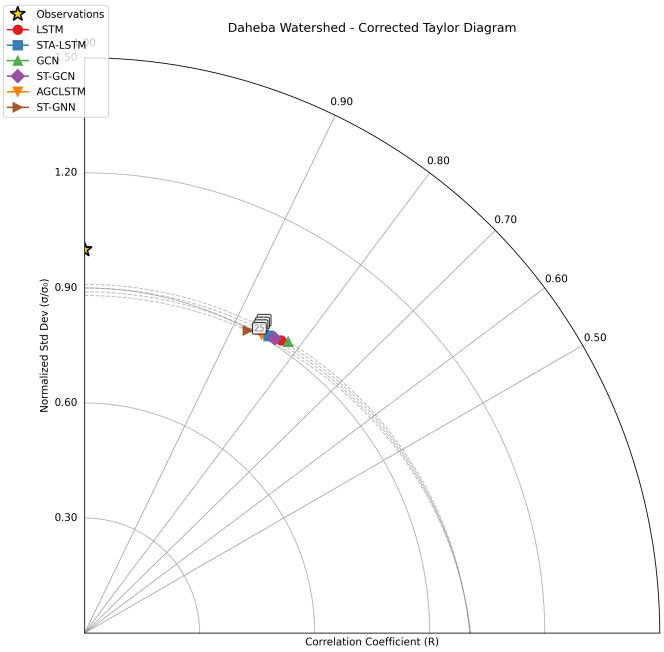
(a) Changhua watershed



(b) Chenhe watershed



(c) Pingyao watershed



(d) Daheba watershed

Figure S1. Taylor diagrams comparing the performance of six hydrological forecasting models for 6-hour ahead streamflow prediction across the four study watersheds... Each diagram visualizes three key performance metrics simultaneously: Correlation coefficient (R) on the angular axis (higher values toward 1.0 indicate better temporal alignment), Normalized standard deviation (σ/σ_0) on the radial axis (values near 1.0 indicate appropriate variance scaling), and Centered Root-Mean-Square Error (CRMSE) represented by dashed, gray contours (lower values indicate smaller errors; contours are labeled with their CRMSE values). The yellow star denotes the reference observations. Models positioned closer to the reference point demonstrate superior overall performance.

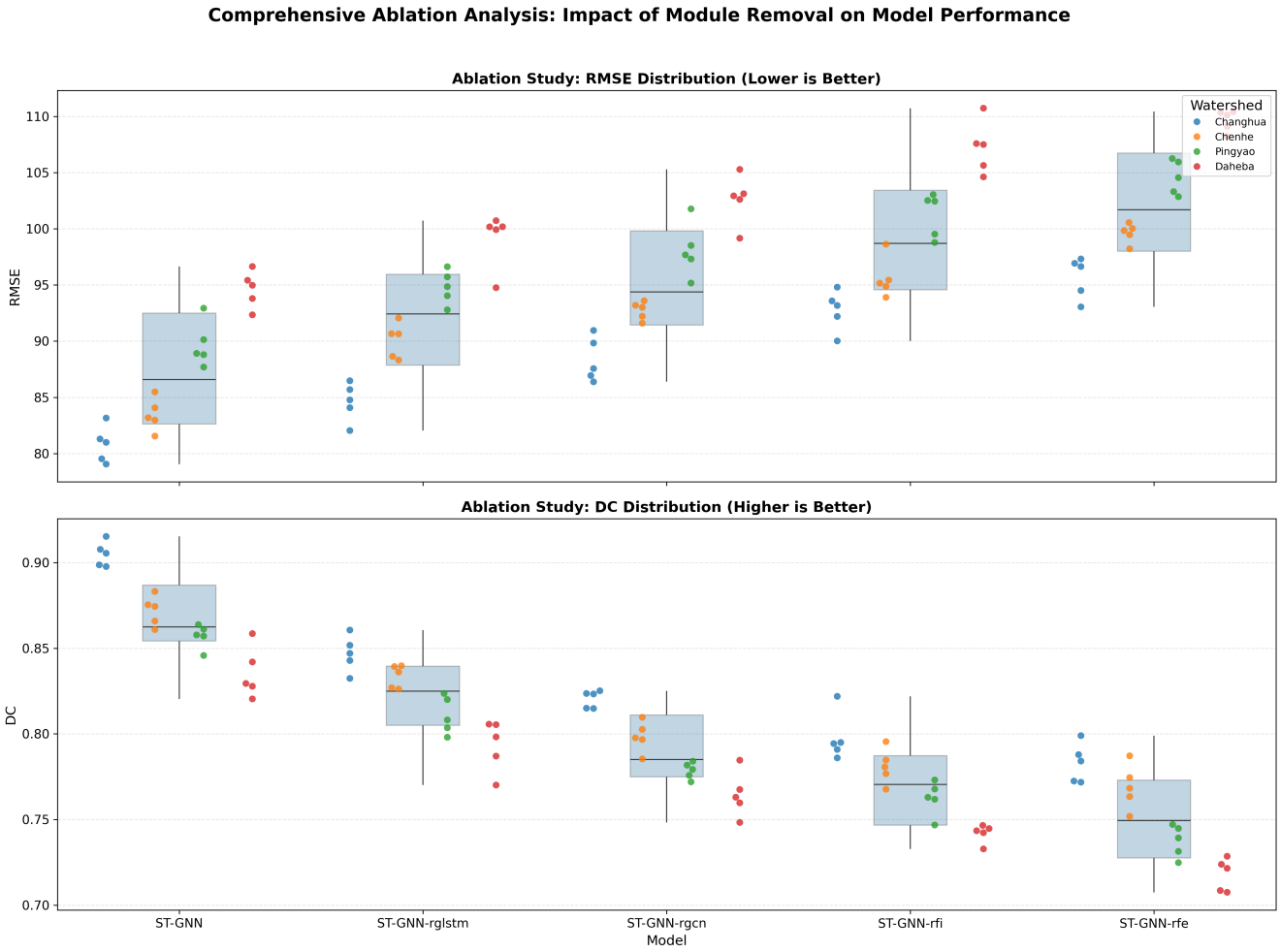


Figure S2. Combined beeswarm plots visualizing the impact of module removal in the ablation study across all watersheds. **(Top)** RMSE distribution: Lower values indicate better performance. **(Bottom)** DC distribution: Higher values indicate better performance. Each point represents one of five independent runs for a specific model-watershed combination. The box plots overlay show statistical summaries: median (central line), interquartile range (box), and data range (whiskers). Model variants (left to right): **ST-GNN** (full model, reference), **ST-GNN-rglstm** (without Graph-LSTM), **ST-GNN-rgcn** (without GCN), **ST-GNN-rfi** (without future rainfall input), and **ST-GNN-rfe** (without feature enhancement). Colors distinguish watersheds as shown in the legend. Key observations: (1) The full ST-GNN achieves the optimal performance with minimal variability across runs, (2) Removing the feature enhancement module causes the most significant performance degradation, (3) All module removals negatively impact both RMSE and DC metrics, and (4) Performance patterns are consistent across different watershed types (humid vs. semi-arid), demonstrating the robustness of the architectural design.

0.0.4. Joint Ablation Study Results

To further investigate the structural dependencies and interactions between the core modules, we conducted joint ablation experiments by removing multiple components simultaneously. Three additional variant models were evaluated:

- **ST-GNN-rgcn_rfe**: Removes both the spatial GCN and feature enhancement fusion modules
- **ST-GNN-rgcn_rglstm**: Removes both the spatial GCN and Graph-LSTM modules
- **ST-GNN-rfe_rglstm**: Removes both the feature enhancement fusion and Graph-LSTM modules

Table S2 presents the complete results of the joint ablation experiments described in Section 4.4. The performance degradation patterns observed across all four watersheds consistently demonstrate: (1) the feature enhancement fusion module has the strongest individual impact on model performance; (2) the combination of feature enhancement and Graph-LSTM removal causes the most severe degradation, indicating their synergistic relationship; and (3) the spatial GCN module, while important, can be partially compensated for by the other components when removed in combination with them.

Table S2. Performance of joint ablation variants across four watersheds (6-hour prediction). RMSE values are shown with percentage increase relative to full ST-GNN in parentheses.

Model Variant	Changhua	Chenhe	Pingyao	Daheba
ST-GNN (full)	82.67	61.37	46.03	107.47
ST-GNN-rgcn_rfe	94.22 (+14.0%)	75.89 (+23.6%)	56.41 (+22.5%)	128.35 (+19.4%)
ST-GNN-rgcn_rglstm	90.15 (+9.1%)	72.43 (+18.0%)	53.27 (+15.7%)	122.18 (+13.7%)
ST-GNN-rfe_rglstm	97.85 (+18.4%)	75.98 (+23.8%)	58.92 (+28.0%)	131.42 (+22.3%)