Qinling-GFFE: A Novel Station-based Benchmark and Graph-Frequency Fusion Enhancer for Precipitation Forecasting

Supplementary Material

8. The Details of Data

To support rigorous evaluation, Qinling integrates hourly observations from 1,528 authentic stations across the semi-arid-monsoon transition zone (103-113°E, 31-40°N). Through rigorous quality control, the benchmark ensures extensive continuous precipitation sequences and precise geolocation throughout 2024, enabling comprehensive assessment of spatiotemporal generalization and spatial accuracy.

Unlike existing precipitation datasets that primarily rely on gridded or image-based representations, Qinling establishes a rigorous benchmark by leveraging authentic measurements to evaluate model robustness against the inherent local abruptness and spatial heterogeneity of precipitation. This design makes it particularly valuable for assessing forecasting accuracy under complex terrain and diverse climatic influences. As the first high-density benchmark from a semi-arid-monsoon transition zone, Qinling's climatic diversity and spatial variability provide a novel platform for studying regional precipitation evolution mechanisms and evaluating the generalization capability of spatiotemporal forecasting models across multiple climate regimes. Additional visualizations of the benchmark's spatial distribution and structural characteristics are provided in Figs. 8.

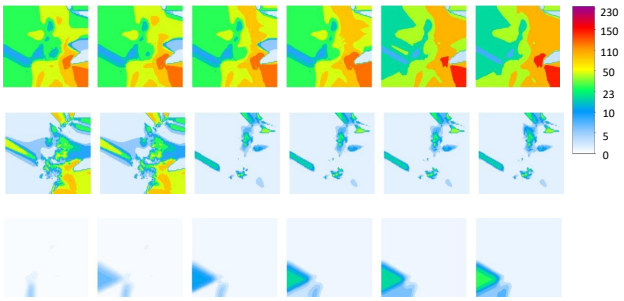


Figure 8. More samples from the Qinling dataset. We select several representative types of precipitation events, including sustained heavy rainfall, gradual increases in precipitation, and rapid decreases in precipitation.

9. Development of GFFE

As illustrated in Algorithm 1, we provide an intuitive summary of the GFFE pipeline in the form of pseudocode.

Given an input feature map $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$, GFFE first extracts its frequency-domain representation through the Fourier extractor, while simultaneously encoding spatial structures via the spatial encoder. A graph is then dynamically constructed based on the spatial features, enabling the GNN to model non-local dependencies. The spectral encoder further extracts frequency features. Subsequently, a cross-domain attention module adaptively fuses the spectral and graph-enhanced information. Finally, the enhanced output and residual projection are combined through a learnable fusion gate to produce the final output $\mathbf{y}$. This pipeline ensures that GFFE effectively integrates multi-domain information while preserving the original model's stability and computational efficiency.

Algorithm 1: GFFE Forward Pass

Input: Input feature map $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$
Output: Enhanced output $\mathbf{y}$

$\mathbf{x}_{ft} \leftarrow \text{FourierExtractor}(\mathbf{x})$;
$\mathbf{F}_{spa} \leftarrow \text{SpatialEncoder}(\mathbf{x})$;
$\mathbf{F}_{spe} \leftarrow \text{SpectralEncoder}(\mathbf{x}_{ft})$;
$\text{graph} \leftarrow \text{GraphConstructor}(\mathbf{F}_{spa})$;
$\mathbf{F}_{en} \leftarrow \text{GNN}(\text{graph})$;
$\mathbf{F}_{attn} \leftarrow \text{CrossDomainAttention}(\mathbf{F}_{en}, \mathbf{F}_{spe})$;
$\mathbf{Y}_{enhanced} \leftarrow \text{OutputProj}(\mathbf{F}_{attn})$;
$\mathbf{Y}_{residual} \leftarrow \text{ResidualProj}(\mathbf{x})$;
$G_{out} \leftarrow \text{FusionGate}(\text{Concat}(\mathbf{Y}_{enhanced}, \mathbf{Y}_{residual}))$;
$\mathbf{y} \leftarrow G_{out} \cdot \mathbf{Y}_{enhanced} + (1 - G_{out}) \cdot \mathbf{Y}_{residual}$;
return $\mathbf{y}$

To quantitatively evaluate the structural overhead introduced by GFFE across different models, we first examine its parameter scale on various methods. Because GFFE is designed as a plug-and-play module, it requires adaptation to the feature dimensions and structures of different backbones. Consequently, when integrated into different models or hierarchical levels, the additional parameters introduced by GFFE will vary accordingly. Table 6 reports the parameter statistics after integrating GFFE into a variety of representative backbone models. Across different architectural families, including CNNs, RNNs, Transformers, diffusion models, and hybrid frameworks, GFFE introduces only a moderate number of additional parameters. Specifically, the proportion ranges from merely 0.14% in large diffusion models such as MCVD [35] to around 9.90% in lightweight CNN architectures like TAU [30].

To gauge the computational overhead introduced by the GFFE module, we measure the average inference time on the Qinling test set. For 24-hour, hour-by-hour precipitation forecasting on the Qinling dataset, the baseline model, TAU [30], completes the full-sequence inference in 3.1971

Table 6. Parameter statistics of models before and after integrating GFFE.

Model Category	Model	Total Parameters	GFFE Parameters	GFFE Proportion
CNN	TAU [30]	7.17M	709.51K	9.90%
RNN	ConvLSTM [28]	12.39M	569.39K	4.60%
Transformer	Earthformer [11]	7.82M	638.83K	8.20%
Diffusion	MCVD [35]	103.51M	149.36K	0.14%
Hybrid framework	AlphaPre [22]	128.52M	557.34K	0.43%

seconds, whereas integrating GFFE increases the runtime only slightly to 3.3083 seconds—an overhead of merely 0.11 seconds. The other baseline, such as Earthformer [11] completes the inference in 3.8563 seconds, and its GFFE-enhanced version takes 4.4922 seconds, corresponding to an increase of 0.66 seconds. Considering the need to autoregressively generate 24 hourly predictions for each sample, such sub-second increments are extremely small and well within acceptable operational limits. These results demonstrate that GFFE introduces only minimal inference-time overhead while significantly improving predictive performance. In other words, GFFE enhances model capability without imposing substantial computational burden, making it suitable for a wide range of spatiotemporal forecasting architectures.

10. Additional Information on Metrics

This section provides additional information on all evaluation metrics used in our experiments, including pixel-level regression metrics (MAE, MSE, SSIM, PSNR) and threshold-based precipitation verification metrics (ETS, CSI, HSS). Note that ETS, CSI, and HSS operate under a predefined precipitation threshold τ, where each pixel is classified as "precipitation event" or "non-precipitation event" based on whether its value exceeds τ.

10.1. Mean Absolute Error (MAE)

MAE measures the average absolute deviation between the predicted values and the ground truth. A smaller MAE represents higher prediction accuracy. Given the prediction $\hat{x}_i$ and the corresponding ground truth x_i, the metric is defined as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\|\hat{x}_i - x_i\|, \tag{13}$$

where N denotes the total number of samples.

10.2. Mean Squared Error (MSE)

MSE measures the average squared difference between the predictions and the ground truth, placing larger penalties on larger errors. It is formulated as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2. \tag{14}$$

Due to the squared term, MSE is more sensitive to extreme deviations.

10.3. Structural Similarity Index Measure (SSIM)

SSIM assesses perceptual similarity by jointly comparing luminance, contrast, and structural information between two tensors. The index ranges from -1 to 1, with higher values indicating better structural similarity:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_y^2 + C_2)}, \tag{15}$$

where μ_x and μ_y denote the mean intensities of tensors x and y, σ_x^2 and σ_y^2 represent their variances, and σ_{xy} is the covariance between them. Constants C_1 and C_2 stabilize the division when the denominator is close to zero.

10.4. Critical Success Index (CSI)

CSI (also known as Threat Score) evaluates precipitation event detection performance under a given threshold τ. It is computed based on the contingency table, where hits H denote the number of correctly predicted precipitation events, misses M denote the number of precipitation events that are not predicted, and false alarms F denote the number of predicted precipitation events that do not actually occur:

$$\text{CSI} = \frac{H}{H + M + F}. \tag{16}$$

10.5. Heidke Skill Score (HSS)

The Heidke Skill Score (HSS) extends CSI by incorporating correct negatives CN, thereby evaluating event detection performance relative to random chance. Unlike CSI, HSS can penalize models that achieve high CSI merely due to a large number of non-events, providing a more balanced assessment of detection accuracy. The score is defined as:

$$\text{HSS} = \frac{2(H \cdot CN - M \cdot F)}{(M + F)(F + CN) + (H + M)(H + CN)}. \tag{17}$$

10.6. Equitable Threat Score (ETS)

The Equitable Threat Score (ETS) further adjusts for hits that would be expected purely by random chance (H_{random}). Compared with CSI and HSS, ETS offers a fairer evaluation of event detection by removing the influence of chance, making it particularly suitable when event frequencies are low or imbalanced. The score is defined as:

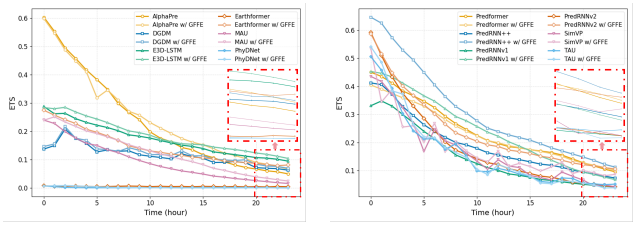
$$\text{ETS} = \frac{H - H_{random}}{H + M + F - H_{random}}, \tag{18}$$

where

$$H_{random} = \frac{(H + M)(H + F)}{H + M + F + CN}. \tag{19}$$

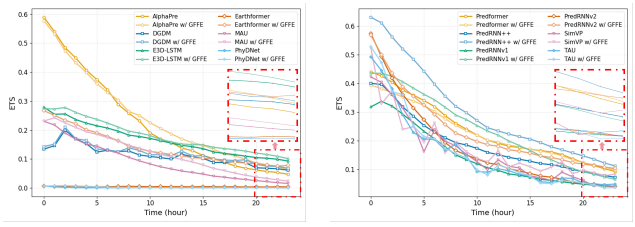
11. More Experiments

In the main text, we present the overall performance of our GFFE on the Qinling benchmark. Here, we further present and compare the model performance across different forecasting thresholds at each hourly lead time. To avoid clutter in visual presentation, we divide the models into two groups and display the same metric for each group in separate subplots of a single figure.



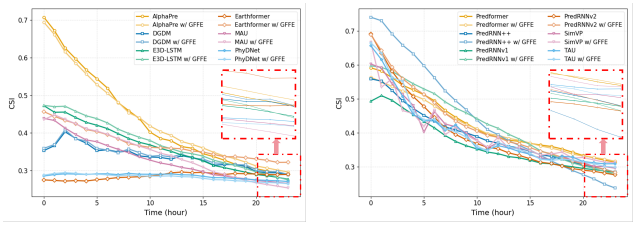
(a) (b)

Figure 9. ETS scores of GFFE at different forecasting lead times for the 200-unit threshold.



(a) (b)

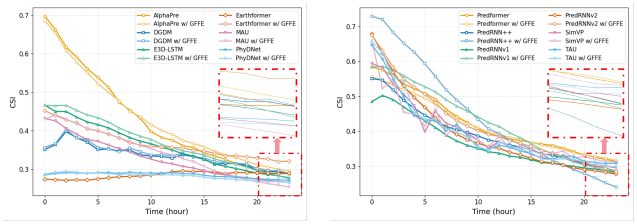
Figure 10. ETS scores of GFFE at different forecasting lead times for the 150-unit threshold.



(a) (b)

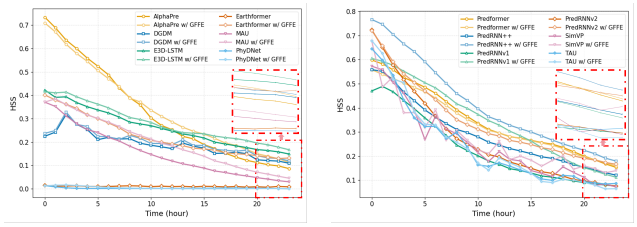
Figure 11. CSI scores of GFFE at different forecasting lead times for the 200-unit threshold.

As evidenced in Figs. 9 to 14, the integration of GFFE yields consistent and substantial gains in ETS, CSI, and HSS across all evaluated models. Compared with their respective baselines, these gains remain consistent under different precipitation thresholds, indicating that GFFE provides effective corrective information across various spatial and intensity scales of precipitation. More importantly,



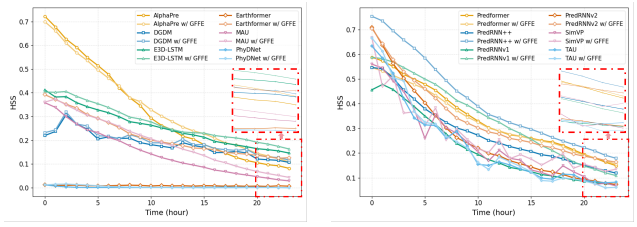
(a) (b)

Figure 12. CSI scores of GFFE at different forecasting lead times for the 150-unit threshold.



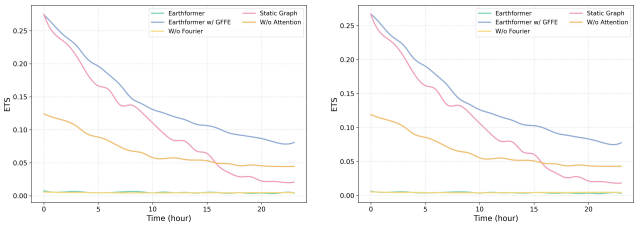
(a) (b)

Figure 13. HSS scores of GFFE at different forecasting lead times for the 200-unit threshold.



(a) (b)

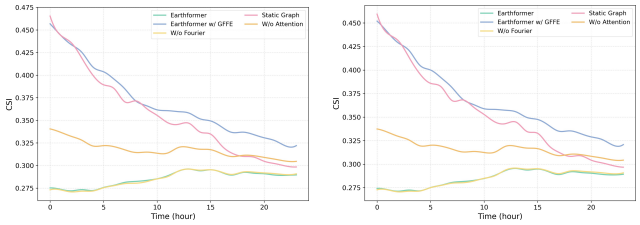
Figure 14. HSS scores of GFFE at different forecasting lead times for the 150-unit threshold.



(a) (b)

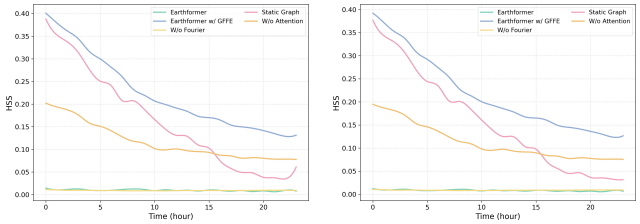
Figure 15. ETS of the GFFE ablation experiments across different thresholds on the Qinling dataset.

as the forecasting lead time extends, this improvement remains evident, indicating that GFFE plays a crucial role in mitigating the accumulation of prediction errors. To further substantiate the effect of GFFE, we also supplement more visualizations of the experimental results, as shown in Figs. 18 and 19. Overall, these results demonstrate that



(a) (b)

Figure 16. CSI of the GFFE ablation experiments across different thresholds on the Qinling dataset.



(a) (b)

Figure 17. HSS of the GFFE ablation experiments across different thresholds on the Qinling dataset.

GFFE strengthens the models' ability to characterize localized precipitation patterns and significantly enhances the robustness of mid to long range forecasts.

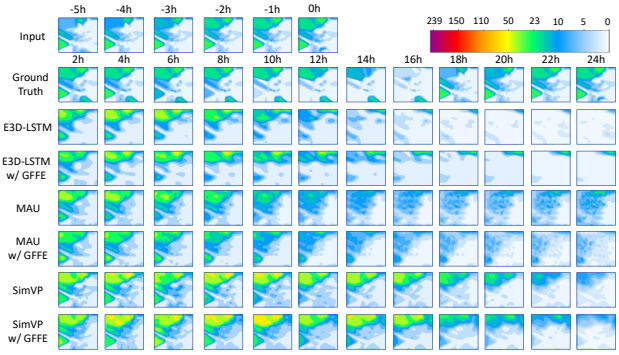


Figure 18. Supplementary visualization of precipitation forecasting on the Qinling dataset, using models different from those in the main text.

Moreover, we conduct a comprehensive comparison of the predictive performance of the ablated models across different forecasting lead times. As illustrated in Figs. 15 to 17, all three ablation strategies, including w/o Fourier, static graph and w/o attention, lead to substantial degradation in precipitation forecasting performance, although the degree of degradation varies across different components. Removing the Fourier feature processing module causes the model performance to regress almost to the baseline level, indicating that this module plays a pivotal role in capturing high-frequency precipitation structures, modeling complex local surface variations, and enhancing the model's spatiotemporal representation capacity. Similarly, eliminating the cross-domain attention mechanism results in a pronounced performance decline, demonstrating the importance of cross-domain feature interaction for effectively integrating multi-source cues and improving the model's sensitivity to localized extreme rainfall signals. In addition, replacing dynamic graph construction with a static graph structure induces a relatively smaller yet consistent performance drop, with the degradation becoming increasingly severe at longer lead times. This pattern underscores the irreplaceable value of dynamic graph construction in modeling time-varying neighborhood dependencies and capturing the evolving structure of precipitation fields over extended forecasting horizons. As illustrated in Fig. 20, these findings provide strong evidence that each component within GFFE is essential for improving precipitation forecasting accuracy, enhancing model generalization, increasing robustness under complex meteorological conditions, and maintaining stability in longer-range predictions.

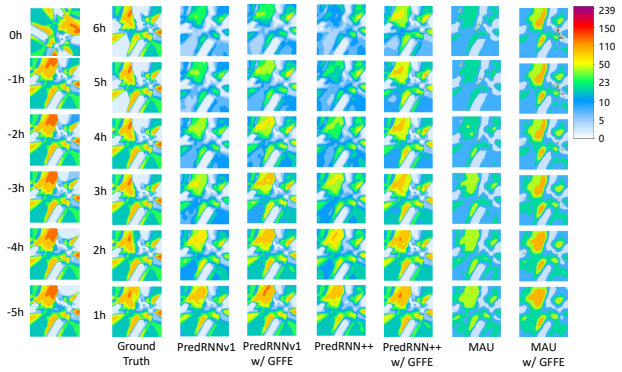


Figure 19. Visualization of 6-hour precipitation forecasting on the Qinling dataset.

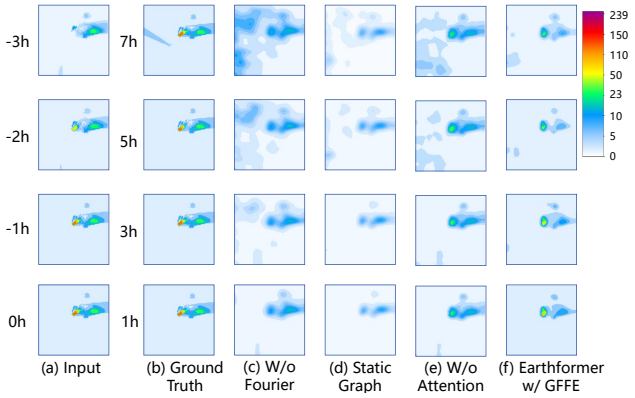


Figure 20. Additional visualizations of the ablation experiments on the Qinling dataset.

The ablation trends further indicate that the effectiveness of GFFE does not stem from a single dominant module. Instead, it emerges from the complementary contributions of multiple components operating at reflecting different physical mechanisms, enabling a more comprehensive, fine-grained, and physically consistent representation of precipitation processes.