LEARNING INTRINSIC WATER-QUALITY DYNAMICS WITH RAINFALL FOR DATA-DRIVEN FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Rainfall is an important environmental driver of water-quality variations through processes such as runoff, pollutant transport, dilution, and resuspension. Traditional mechanistic models can explicitly describe these processes but often require substantial process specification and site-specific calibration, limiting their flexibility under changing hydrological conditions. In this work, we explore a data-driven alternative by proposing RaiNet to jointly model multiscale water-quality dynamics and station-specific rainfall effects across relative lags and temporal scales. RaiNet employs `LocTrend` to capture irregular water-quality dynamics, constructs station-oriented rainfall events from gridded precipitation, and introduces `XGateFusion` for conditional lag-aware fusion across scales. We further release three real-world multimodal datasets comprising over 150,000 temporally aligned water quality observations and gridded precipitation raster images. Experiments show that RaiNet outperforms general time-series, water quality, diffusion-based, and spatiotemporal models by over 20%, while component-wise analyses confirm the distinct contribution of each module.

1 INTRODUCTION

Accurate water-quality forecasting is essential for environmental management and ecosystem protection, but remains challenging because hydrological, meteorological, and episodic influences vary over time and across stations. Mechanistic models such as SWAT (Krysanova & White, 2015), WASP (Wool et al., 2020), and QUAL2K (Zhang et al., 2012) provide interpretable predictions but require detailed process specification, extensive environmental inputs, and site-specific calibration, limiting their transferability under changing catchment conditions. These limitations are particularly restrictive when the effects of rainfall vary over space and time. This motivates more flexible data-driven approaches that can learn evolving environmental dependencies directly from observations.

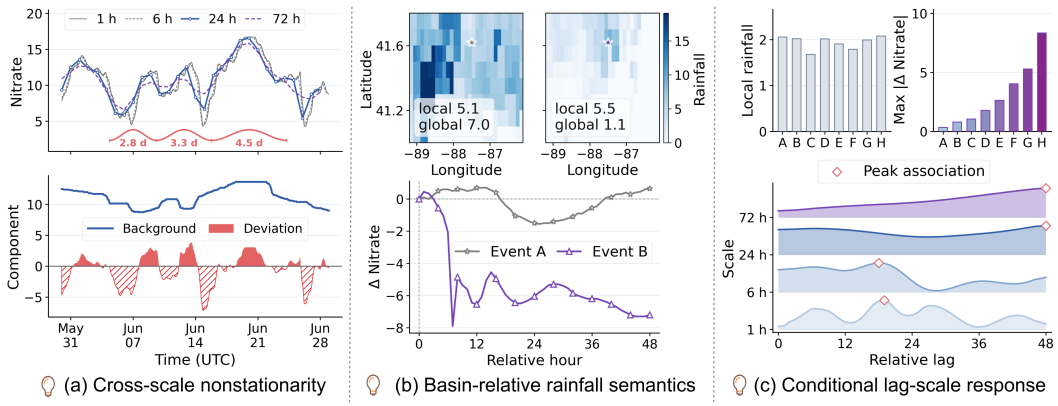


Figure 1: **Empirical motivations of this work**. (a) Water-quality dynamics vary across temporal resolutions, with irregular background-deviation patterns and no fixed periodicity. (b) Similar local rainfall can represent different events under different basin-wide contexts, leading to heterogeneous water-quality variations. (c) Comparable rainfall events can induce different water-quality responses, with strongest associations appearing at different relative lags across temporal scales.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recent time-series and spatiotemporal models can learn complex nonlinear dependencies from observations (Wu et al., 2023; Zhou et al., 2022; Ahmed et al., 2025; Zhong et al., 2025; Huang et al., 2025). However, general-purpose forecasting architectures often overlook the influence of external environmental conditions on water-quality dynamics. We focus on a specific yet challenging factor: *rainfall*. Its effects on water quality vary across stations through processes such as runoff, pollutant transport, and dilution. With increasingly available gridded precipitation products, rainfall can be integrated with water-quality observations as an additional modality for multimodal forecasting.

Effectively integrating the two modalities remains challenging because rainfall influence is intermittent, station-specific, lagged, and scale dependent. Existing multimodal fusion methods rarely model these conditional dependencies explicitly, limiting their ability to determine *when, at what lag, and at which temporal scale* rainfall should contribute to water-quality forecasting (Wang et al., 2024b; Papillon et al., 2025). As illustrated in Figure 1, this leads to three challenges. First, water-quality dynamics are multiscale and nonstationary, with short-lived fluctuations superimposed on slowly evolving backgrounds and no fixed periodic structure. Second, raw rainfall intensity does not directly represent a station-relevant event: similar local rainfall can have different meanings under different basin-wide contexts, while transient rainfall may not persist long enough to affect water quality. Third, rainfall and water-quality relationships are conditional, lagged, and scale dependent: similar rainfall events may correspond to different subsequent responses, and their strongest statistical associations can occur at different relative temporal offsets across scales.

To address these challenges, we propose **RaiNet**, a rainfall-informed water-quality forecasting network. It first models intrinsic multiscale water-quality dynamics, constructs station-oriented rainfall events from gridded precipitation, and then selectively uses rainfall to correct independently predicted water-quality baselines. Our main contributions are:

- We develop a multiscale background-deviation modeling mechanism for irregular water-quality dynamics. Multiscale downsampling exposes temporal structures at different resolutions, while `LocTrend` adaptively separates locally evolving structural backgrounds from dynamic deviations without relying on predefined temporal patterns.
- We introduce a rainfall-event encoder. By combining basin-to-local spatial contrast with temporal persistence, it transforms raw rainfall grids into station-aligned event representations that better characterize environmentally meaningful rainfall conditions.
- We propose `XGateFusion`, a conditional, lag-aware, and scale-dependent fusion mechanism. It learns statistical rainfall and water-quality relationships across relative temporal offsets and adaptively determines whether and to what extent rainfall should correct the independently predicted water-quality baseline at each temporal scale.

We establish three multimodal benchmarks with over 150,000 aligned water quality observations and gridded precipitation records, supporting research on multimodal environmental modeling.

2 Related Works

Water-Quality Forecasting. Existing water-quality forecasting methods broadly include mechanistic, statistical, and deep learning approaches. Mechanistic models, such as SWAT (Krysanova & White, 2015), WASP (Wool et al., 2020), and QUAL2K (Zhang et al., 2012), explicitly represent hydrological and biogeochemical processes but require detailed inputs, process assumptions, and site-specific calibration. Statistical methods simplify this process by extracting temporal regularities through autoregression and seasonality, including ARIMA (Shumway & Stoffer, 2017), SARIMA (Sathya et al., 2023), Holt-Winters (Chatfield, 1978), and STL (Yan et al., 2024). General-purpose deep time-series models extend this data-driven direction to nonlinear dynamics: TimesNet (Wu et al., 2023) models period-dependent variations, FEDformer (Zhou et al., 2022) combines decomposition with frequency-domain learning, and TimeMixer (Wang et al., 2024a) exchanges information across temporal scales. Water-quality studies have also coupled wavelet decomposition with long short-term memory (LSTM) to alleviate nonstationarity (Chen et al., 2024). However, their fixed temporal assumptions limit adaptation to evolving backgrounds and irregular deviations.

Spatiotemporal Modeling across Monitoring Stations. Graph-based models represent monitoring stations as nodes and use geographical, hydrological, or learned relations for cross-station aggregation (Ahmed et al., 2025; Ho et al., 2025; Bi et al., 2025b). STF-GNN (Wan et al., 2025) combines

graph convolution, gate recurrent unit, and self-attention to jointly model spatial and temporal dependencies. Such propagation is useful when stations have meaningful hydrological connectivity. However, in some real-world monitoring networks, stations are sparse, widely separated, and influenced by heterogeneous local environments (Zhang et al., 2023; Wang et al., 2026). Water-quality signals are therefore attenuated, dispersed, or transformed over long travel distances, weakening direct correspondence across monitoring sites. This limits explicit cross-station propagation and motivates station-wise forecasting that exploits spatial information through environmental context.

Rainfall-Informed and Multimodal Water-Quality Forecasting. Rainfall is incorporated into water-quality forecasting through process-based and data-driven approaches. Process-based models describe its effects through hydrological processes: SWAT (Krysanova & White, 2015) converts precipitation into runoff and transport fluxes, while WASP (Wool et al., 2020) simulates pollutant transport and biochemical reactions. Although physically interpretable, they require detailed inputs and site-specific calibration. Data-driven methods learn rainfall effects from observations using lag-selected predictors (Ortiz-Lopez et al., 2023) or recurrent networks with meteorological inputs (Huang et al., 2024), but commonly reduce rainfall to scalar covariates. Recent multimodal methods preserve spatial information through radar features (Baek et al., 2020) or direct rainfall-image fusion through CNN (Bi et al., 2025a). However, station-specific rainfall events and their duration, relative-lag, and scale-dependent effects remain underexplored.

3 METHODOLOGY

We consider multistep and multi-site forecasting with historical in-situ observations and gridded rainfall data. Let $\mathbf{X} \in \mathbb{R}^{T\times N}$ denote the historical measurements of a target water-quality indicator from N monitoring stations over T time steps, and let $\mathbf{I} \in \mathbb{R}^{T\times C_I\times H_I\times W_I}$ denote the corresponding gridded rainfall sequence, where C_I is the number of rainfall channels and $H_I \times W_I$ represents the spatial resolution. The directional relationships among monitoring stations are modeled by a directed topology $\mathcal{G}$. Given $\mathbf{X}$, $\mathbf{I}$, and $\mathcal{G}$, our objective is to predict the target indicator at all stations over the next H time steps: $\widehat{\mathbf{Y}} = f_\theta(\mathbf{X}, \mathbf{I}, \mathcal{G}) \in \mathbb{R}^{H\times N}$, where f_θ denotes the forecasting model.

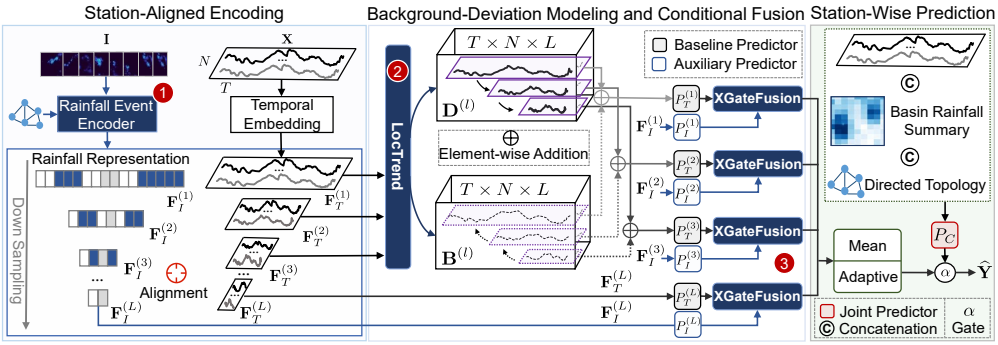


Figure 2: **The RaiNet framework.** The three key modules are the rainfall event encoder, `LocTrend`, and `XGateFusion`: (1) identifies station-relevant rainfall events; (2) separates evolving backgrounds from dynamic deviations without a fixed period; and (3) conditionally integrates rainfall according to its relative lag and temporal scale.

The RaiNet architecture is illustrated in Figure 2. It encodes station-oriented rainfall alongside multiscale water-quality dynamics. `LocTrend` decomposes the water-quality representations into locally evolving structural backgrounds and dynamic deviations. Scale-specific predictors then map both branches into a common forecast space, where `XGateFusion` captures relative-lag associations and applies rainfall-derived corrections. Finally, the multiscale forecasts are combined with joint prediction that incorporates basin-level rainfall context and directed inter-station relationships.

3.1 STATION-ALIGNED ENCODING

Multiscale water-quality encoding. Water-quality dynamics evolve at different temporal scales, ranging from short-term fluctuations caused by transient disturbances to long-term variations driven

by gradual environmental changes. To capture these heterogeneous patterns, we construct a temporal pyramid through successive average pooling:

$$\mathbf{X}^{(1)} = \mathbf{X}, \qquad \mathbf{X}^{(l)} = \mathrm{AvgPool}_q\left(\mathbf{X}^{(l-1)}\right) \in \mathbb{R}^{T_l \times N}, \qquad T_l = \left\lfloor \frac{T_{l-1}}{q} \right\rfloor, \quad l = 2, \ldots, L, \tag{1}$$

where T_l is the temporal length at scale l, L is the number of scales, and q is the pooling window size. Each scale is then normalized and mapped into a d-dimensional latent space as $\mathbf{F}_T^{(l)} = \mathrm{Emb}_T(\mathrm{Norm}_l(\mathbf{X}^{(l)})) \in \mathbb{R}^{T_l \times N \times d}$, where $\mathrm{Norm}_l(\cdot)$ denotes scale-specific normalization and $\mathrm{Emb}_T(\cdot)$ is a temporal embedding shared across scales. It combines convolutional value embedding with scale-aligned time features, and further details are provided in Appendix A.1.1.

Station-oriented rainfall-event encoding. Station locations from $\mathcal{G}$ are first used to align each monitoring station with the rainfall grid. Because station-local rainfall may reflect either a localized event or part of a basin-wide system, we interpret it relative to the simultaneous basin background:

$$r_t^{\mathrm{global}} = \frac{\sum_p v_{t,p} I_{t,p}}{\sum_p v_{t,p}}, \qquad r_{n,t}^{\mathrm{local}} = \frac{\sum_p \pi_{n,p} v_{t,p} I_{t,p}}{\sum_p \pi_{n,p} v_{t,p}}, \qquad r_{n,t}^{\mathrm{rel}} = r_{n,t}^{\mathrm{local}} - r_t^{\mathrm{global}}, \tag{2}$$

where t, n, and p index time steps, monitoring stations, and rainfall-grid cells, respectively; $I_{t,p}$ is the rainfall intensity and $v_{t,p} \in \{0, 1\}$ indicates whether it is valid. The normalized spatial prior $\pi_{n,p}$ assigns larger weights to cells closer to station n. Accordingly, r_t^{global}, $r_{n,t}^{\mathrm{local}}$, and $r_{n,t}^{\mathrm{rel}}$ denote basin-wide rainfall, station-local rainfall, and their signed spatial contrast. In this case, small contrasts may reflect minor spatial variability or measurement uncertainty, and we retain only contrasts exceeding the normal variability observed at each station:

$$e_{n,t} = \mathrm{sign}\left(r_{n,t}^{\mathrm{rel}}\right) \mathbb{I}\left(\left|r_{n,t}^{\mathrm{rel}}\right| > \epsilon_n\right), \qquad e_{n,t} \in \{-1, 0, +1\}. \tag{3}$$

where ϵ_n is a station-specific threshold estimated exclusively from the training set, $\mathrm{sign}(\cdot)$ returns the contrast direction, and $\mathbb{I}(\cdot)$ is the indicator function. Thus, $e_{n,t} \in \{-1, 0, +1\}$ represents locally weaker rainfall, no significant contrast, or locally stronger rainfall, respectively. Discretization reduces sensitivity to within-state amplitude fluctuations (shown in Appendix A.11), while continuous basin-level rainfall summaries retain basin-scale magnitude information for prediction (Section 3.3).

A station-relevant event should also persist beyond an isolated observation, so the time-indexed states are aggregated through signed voting using persistence setting h:

$$\widetilde{e}_{n,t} = \mathrm{sign}\left(\sum_{s=b(t)h+1}^{(b(t)+1)h} e_{n,s}\right), \qquad b(t) = \left\lfloor \frac{t-1}{h} \right\rfloor, \qquad \mathbf{F}_I = \mathcal{E}_I(\mathbf{I}) = [\widetilde{e}_{n,t}]_{t,n}, \tag{4}$$

where $b(t)$ identifies the block containing time step t, and $\widetilde{e}_{n,t}$ is assigned to every position in the corresponding block. The block construction is adapted to each dataset's sampling interval. The resulting $\mathbf{F}_I \in \mathbb{R}^{T \times N}$ is the station-related rainfall representation produced by encoder $\mathcal{E}_I$.

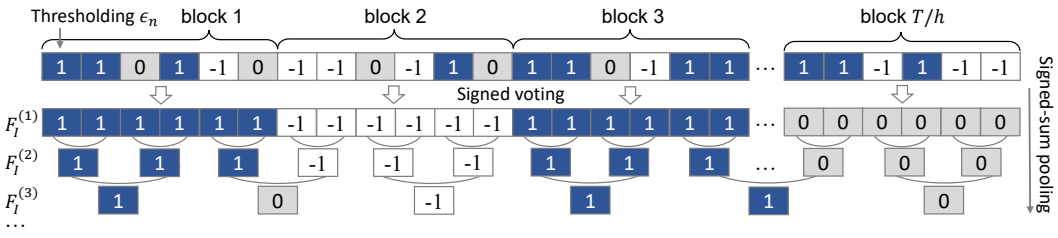


Figure 3: **Rainfall event encoder.** Signed voting with setting h stabilizes each rainfall event and reduces sensitivity to within-state amplitude fluctuations, while signed-sum pooling preserves its dominant direction over broader supports and aligns it with the corresponding water-quality scales.

Building on this stabilized event sequence, we further represent rainfall evidence over the temporal supports of the water-quality pyramid. Specifically, causal signed-sum pooling aggregates consecutive states over progressively broader intervals and retains the dominant event direction:

$\mathbf{F}_I^{(l)} = \mathcal{A}_l(\mathbf{F}_I) \in \mathbb{R}^{T_l \times N}$, where $\mathcal{A}_l(\cdot)$ uses the aggregation span of scale l. This operation both captures the persistence of rainfall evidence at different temporal scales and aligns $\mathbf{F}_I^{(l)}$ with $\mathbf{F}_T^{(l)}$ along the station and temporal dimensions. The overall process is illustrated in Figure 3, with implementation details provided in Appendix A.1.2.

3.2 Background-Deviation Modeling and Conditional Fusion

The two modalities play asymmetric roles in RaiNet. $\mathbf{F}_T^{(l)}$ captures intrinsic water-quality dynamics, whereas $\mathbf{F}_I^{(l)}$ provides temporally and station-aligned rainfall-event conditions. This module first decomposes water-quality dynamics into a locally evolving background and dynamic deviations, and then selectively applies lag-aware rainfall corrections to the resulting baseline.

3.2.1 LocTrend Background-Deviation Decomposition

Fixed-period decomposition imposes recurrent temporal structures that do not match the locally evolving patterns of water-quality dynamics. `LocTrend` reconstructs a locally evolving structural background from data-dependent patterns without specifying a daily, weekly, or annual period.

`LocTrend` is applied only before the temporal sequence becomes excessively compressed. Let L_{dir} denote the number of coarsest scales preserved through the direct temporal path. It is applied to the remaining scales ($l = 1, \ldots, L - L_{\text{dir}}$):

$$\left(\mathbf{B}^{(l)}, \mathbf{D}^{(l)}\right) = \text{LocTrend}\left(\mathbf{F}_T^{(l)}\right). \tag{5}$$

The remaining coarsest scales $l = L - L_{\text{dir}} + 1, \ldots, L$ are preserved without further decomposition, as their short sequences already summarize broad temporal contexts and may lose useful coarse-scale information under local decomposition.

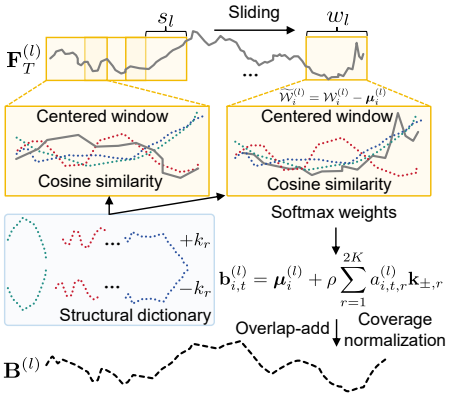


Figure 4: `LocTrend` reconstructs local structural backgrounds.

For each decomposed scale, we extract overlapping windows and remove their local levels:

$$\mathcal{W}_i^{(l)} = \mathbf{F}_{T,\, q_i:q_i+w_l-1}^{(l)}, \qquad \boldsymbol{\mu}_i^{(l)} = \frac{1}{w_l} \sum_{t=1}^{w_l} \mathcal{W}_{i,t}^{(l)}, \qquad \widetilde{\mathcal{W}}_i^{(l)} = \mathcal{W}_i^{(l)} - \boldsymbol{\mu}_i^{(l)}, \tag{6}$$

where w_l and s_l are the window length and stride, respectively, and $q_i = (i-1)s_l + 1$ is the starting position of window i. Centering shifts the dictionary from absolute levels to local patterns.

Training-fitted structural dictionary. We represent local structures using a shared learnable dictionary $\mathbf{K} = \{\mathbf{k}_r\}_{r=1}^K$. Its initial value $\mathbf{K}^{(0)}$ is obtained once from training histories using a two-pass structural initialization:

$$\left\{\widetilde{\mathcal{W}}_q\right\}_{q \in \mathcal{Q}_{\text{tr}}} \xrightarrow[\text{averaged kernel eigendecomposition}]{\text{RBF temporal similarities}} \{(\lambda_r, \mathbf{u}_r)\}_{r=1}^K \xrightarrow[\text{feature covariance eigendecomposition}]{\text{window projection}} \mathbf{K}^{(0)}, \tag{7}$$

where $\mathcal{Q}_{\text{tr}}$ is the collection of centered training windows, and K is the number of retained components and resulting dictionary atoms. In the first pass, a radial-basis-function (RBF) kernel converts pairwise distances within each window into nonlinear temporal similarities; averaging their centered Gram matrices and retaining the top eigenpairs $\{(\lambda_r, \mathbf{u}_r)\}_{r=1}^K$ identifies temporal directions recurring across training windows. In the second pass, the windows are projected along these directions, and the leading eigenvectors of their aggregated feature covariance form the initial atoms $\mathbf{k}_r^{(0)}$. The dictionary is shared across the decomposed scales and subsequently optimized jointly with the temporal encoder and prediction modules under the forecasting objective. Appendix A.2.1 details the temporal eigendirection extraction and their conversion into feature-space structural atoms.

Local background reconstruction. To represent centered patterns in either direction, we form the symmetric dictionary $\mathbf{K}_{\pm} = [\mathbf{K}; -\mathbf{K}]$ with $2K$ atoms. Each centered token is softly assigned using

temperature-scaled cosine similarity:

$$a_{i,t,r}^{(l)} = \mathrm{softmax}_r\left(\frac{\cos\left(\widetilde{\mathbf{w}}_{i,t}^{(l)}, \mathbf{k}_{\pm,r}\right)}{\vartheta}\right), \qquad \mathbf{b}_{i,t}^{(l)} = \boldsymbol{\mu}_i^{(l)} + \rho \sum_{r=1}^{2K} a_{i,t,r}^{(l)} \mathbf{k}_{\pm,r}, \tag{8}$$

where $\widetilde{\mathbf{w}}_{i,t}^{(l)}$ is token t of the centered window, ϑ is the assignment temperature, and ρ controls the centered reconstruction strength. This soft assignment allows the background to follow locally recurring shapes without imposing a predefined period. Finally, overlapping reconstructions are averaged according to their coverage to obtain $\mathbf{B}^{(l)}$ and $\mathbf{D}^{(l)} = \mathbf{F}_T^{(l)} - \mathbf{B}^{(l)}$. Thus, $\mathbf{B}^{(l)}$ captures the locally evolving structural background, while $\mathbf{D}^{(l)}$ preserves the exact dynamic remainder. The complete `LocTrend` forward procedure is detailed in Appendix A.2. By centering local windows and restoring their means, `LocTrend` absorbs local level shifts into the background while preserving dynamic deviations, without relying on fixed periodic structures (as proved in Appendix A.2.3).

3.2.2 Cross-Scale Modeling and Scale-Specific Prediction

To improve consistency across decomposed scales, RaiNet propagates dynamic deviations from fine to coarse resolutions (arrows between different scales in Figure 2), where short-term changes are more clearly resolved, and structural backgrounds from coarse to fine resolutions, where long-range context is more stable, i.e., $\{\overline{\mathbf{D}}^{(l)}\} = \mathcal{M}_D^{\mathrm{f2c}}(\{\mathbf{D}^{(l)}\})$ and $\{\overline{\mathbf{B}}^{(l)}\} = \mathcal{M}_B^{\mathrm{c2f}}(\{\mathbf{B}^{(l)}\})$. Each transition adaptively combines convolutional and temporal linear paths. Details are provided in Appendix A.3.

For each decomposed scale, each pair is recombined as $\widetilde{\mathbf{F}}_T^{(l)} = \mathbf{F}_T^{(l)} + \mathrm{FFN}(\overline{\mathbf{B}}^{(l)} + \overline{\mathbf{D}}^{(l)})$. The remaining coarsest scales bypass this operation. Scale-specific predictors then map the two modalities into the common water-quality prediction space:

$$\widehat{\mathbf{Y}}_T^{(l)} = P_T^{(l)}\left(\widetilde{\mathbf{F}}_T^{(l)}\right), \qquad \widehat{\mathbf{Y}}_I^{(l)} = P_I^{(l)}\left(\mathbf{F}_I^{(l)}\right), \qquad \widehat{\mathbf{Y}}_T^{(l)}, \widehat{\mathbf{Y}}_I^{(l)} \in \mathbb{R}^{H\times N}. \tag{9}$$

where $\widehat{\mathbf{Y}}_T^{(l)}$ is the predicted water-quality baseline. $P_I^{(l)}$ translates historical rainfall events into a candidate water-quality correction in the forecast space, yielding $\widehat{\mathbf{Y}}_I^{(l)}$. `XGateFusion` then determines how much of this correction should be applied for each station, forecast step, and scale.

3.2.3 XGateFusion

Rainfall correction should only be introduced when compatible with the current water-quality state. At scale l, `XGateFusion` compares the station-wise baseline $\mathbf{u}_{T,n}^{(l)} = \widehat{\mathbf{y}}_{T,n}^{(l)}$ with its rainfall-conditioned counterpart $\mathbf{u}_{I,n}^{(l)} = \widehat{\mathbf{y}}_{T,n}^{(l)} + \widehat{\mathbf{y}}_{I,n}^{(l)}$, where $\widehat{\mathbf{y}}_{T,n}^{(l)}$ and $\widehat{\mathbf{y}}_{I,n}^{(l)}$ are the station-n columns of the scale-level baseline and correction, respectively. Comparing them in the same prediction space allows the model to evaluate candidate corrections without directly matching heterogeneous representations.

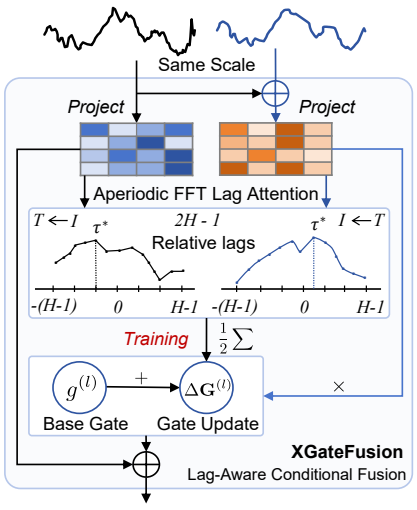


Figure 5: Bidirectional aperiodic FFT lag attention adapts corrections across scales, stations, and forecast steps.

Rainfall-related changes may align with the baseline at different relative forecast offsets. Bidirectional aperiodic lag attention therefore evaluates their statistical associations over the complete lag range, while avoiding artificial wraparound at the horizon boundaries. The interactions are computed by Fast Fourier Transform (FFT) with $\mathcal{O}(H \log H)$ temporal complexity, without explicitly constructing the full lag-interaction matrix. The resulting representation produces a bounded gate update $\Delta\mathbf{G}^{(l)} \in \mathbb{R}^{H\times N}$, allowing the correction strength to vary across stations and forecast steps. Together with the scale-level base gate $g^{(l)}$, the fused prediction is

$$\widehat{\mathbf{Y}}_{TI}^{(l)} = \widehat{\mathbf{Y}}_T^{(l)} + \left(g^{(l)}\mathbf{1}_{H\times N} + \Delta\mathbf{G}^{(l)}\right) \odot \widehat{\mathbf{Y}}_I^{(l)}. \tag{10}$$

where $\mathbf{1}_{H\times N}$ is an all-one matrix and $\odot$ denotes element-wise multiplication. The base gate $g^{(l)}$ controls the scale-level contribution of rainfall, whereas $\Delta\mathbf{G}^{(l)}$ provides station and horizon-specific adaptation. The dynamic gate-update head is initialized to zero for stable training. Detailed derivations of `XGateFusion` are provided in Appendix A.4.

3.3 Station-Wise Prediction

Because the informative temporal scale varies across samples, RaiNet uses forecast statistics to adaptively weight the scale-wise predictions, yielding $\widehat{\mathbf{Y}}_{\mathrm{w}} = \sum_{l=1}^{L} \omega_l \widehat{\mathbf{Y}}_{TI}^{(l)}$. It is balanced with their mean $\widehat{\mathbf{Y}}_{\mathrm{m}} = L^{-1} \sum_{l=1}^{L} \widehat{\mathbf{Y}}_{TI}^{(l)}$, i.e., $\widehat{\mathbf{Y}}_{TI} = \zeta_{\mathrm{m}} \widehat{\mathbf{Y}}_{\mathrm{m}} + \zeta_{\mathrm{w}} \widehat{\mathbf{Y}}_{\mathrm{w}}$, and $\zeta_{\mathrm{m}} + \zeta_{\mathrm{w}} = 1$, where the sample-dependent weights balance stable averaging against selective scale aggregation.

Beyond station-aligned rainfall corrections, water-quality prediction also depends on basin-level rainfall ($\mathbf{c}_I$) and spatial relations among stations. The Joint Predictor P_C concatenates the water-quality history, basin rainfall summaries, and topological features, allowing their complementary information to be modeled jointly (see Appendix A.5 for feature construction). It produces $\widehat{\mathbf{Y}}_C = P_C(\mathbf{X}, \mathbf{c}_I, \mathcal{G}) \in \mathbb{R}^{H\times N}$. The final prediction is $\widehat{\mathbf{Y}} = (\mathbf{1} - \boldsymbol{\alpha}) \odot \widehat{\mathbf{Y}}_{TI} + \boldsymbol{\alpha} \odot \widehat{\mathbf{Y}}_C, \boldsymbol{\alpha} \in [0, 1]^{H\times N}$.

4 Experiments

4.1 Experimental Setup

We evaluate RaiNet on three multimodal datasets. They contain five-year hourly turbidity records from 15 stations, three-year hourly nitrate records from five stations, and five-year 30-minute nitrate records from five stations, all aligned with gridded precipitation. We publicly release over 150,000 paired time steps as benchmarks. Each dataset is chronologically split at a 7:1:2 ratio. The training objective of RaiNet and the dataset details are provided in Appendices A.6 and A.7, respectively.

RaiNet uses $L = 4$ temporal scales, constructed by successive average pooling with window size $q = 2$. We set $L_{\mathrm{dir}} = 1$, so the last coarsest scale follows the direct temporal path. Rainfall-event states $h = 6$ captures short sub-daily rainfall persistence without merging separate events. For `LocTrend`, we use $w_l = 24$, $s_l = 12$, $K = 4$, $\vartheta = 1$, and $\rho = 1$. We set the hidden dimension $d = 16$, the feedforward dimension to 32, and dropout to 0.1. Models are trained with a batch size of 64 and a learning rate of 10^{-3}. Sensitivity analysis of hyperparameters is provided in Appendix A.8.

We compare RaiNet with four groups of baselines. Time-series models include TimeKAN (Huang et al., 2025), FilterTS (Wang et al., 2025b), GTR (Cao et al., 2026), and MixLinear (Ma et al., 2026). Spatio-temporal models include HyperD (Shao et al., 2026), RSTIB-MLP (Chen et al., 2025), and ST-SSDL (Gao et al., 2026), representing periodicity decoupling, robust spatio-temporal learning, and deviation-aware self-supervision. Water-quality-specific baselines include PIWON (Tang et al., 2026), D-TGCN (Zhang et al., 2026), CMLIP (Bi et al., 2025a), RGCN (Smith et al., 2024), and XGBR (Ali et al., 2025). We also compare with diffusion-based SimDiff (Ding et al., 2026) and MA-TSD (Wang et al., 2025a). Detailed descriptions of these models are provided in Appendix A.9.

4.2 Main Results

As shown in Tables 1 and 2, RaiNet consistently achieves the best MSE across all datasets and horizons. It obtains the best MAE in eight of nine settings; the only exception is 170900 at 192 steps, where RSTIB-MLP is marginally better. The gains remain consistent across different sampling intervals and basin configurations, demonstrating consistent robustness. Conventional spatiotemporal models are unstable, particularly on 071200 and 051201. Water-quality stations are sparsely distributed and exhibit heterogeneous, asynchronous dynamics, making generic cross-station propagation prone to introducing irrelevant spatial information. RaiNet avoids this mismatch by forecasting station-wise dynamics while using spatial rainfall patterns as conditional environmental context, exploiting basin information without making direct cross-station propagation the primary forecasting pathway. It also consistently outperforms water-quality-specific and diffusion-based models. The main tables report results obtained with a fixed random seed, while five-seed stability, full-horizon results, detailed analyses, and computational costs are provided in Appendix A.10.

Table 1: Forecasting results with representative time-series and spatiotemporal baselines. The best and second-best results are highlighted in **bold** and underline, respectively.

Model		TimeKAN		FilterTS		GTR		MixLinear		HyperD		RSTIB-MLP		ST-SSDL		RaiNet	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
170900	24	0.2628	0.1667	0.2692	0.1722	0.2763	0.1726	0.3790	0.2324	0.3786	0.1909	0.2960	0.1871	0.2792	0.2107	**0.2185**	**0.1560**
	96	0.4778	0.2623	0.4771	0.2626	0.4834	0.2648	0.5739	0.3208	0.5518	0.2729	0.4994	0.2676	0.5410	0.3177	**0.4096**	**0.2604**
	192	0.6053	0.3204	0.5974	0.3199	0.5975	0.3207	0.7047	0.3786	0.6442	0.3183	0.5937	**0.3061**	0.6603	0.3691	**0.5283**	0.3102
071200	24	0.1084	0.1905	0.1105	0.1885	0.1162	0.1967	0.2458	0.2848	0.6530	0.4605	0.1630	0.2555	0.1516	0.2535	**0.0872**	**0.1747**
	96	0.4412	0.3802	0.4473	0.3784	0.4591	0.3894	0.5818	0.4424	0.7663	0.5291	0.4613	0.3912	0.7145	0.5240	**0.3225**	**0.3344**
	192	0.6771	0.4918	0.6879	0.4944	0.7010	0.5041	0.8131	0.5463	0.8453	0.5759	0.6597	0.4930	1.0558	0.7019	**0.5415**	**0.4443**
051201	24	0.0074	0.0328	0.0075	0.0337	0.0081	0.0353	0.0194	0.0599	1.3482	0.7703	0.7583	0.6066	0.0217	0.0821	**0.0072**	**0.0318**
	96	0.0462	0.0889	0.0468	0.0903	0.0491	0.0938	0.0609	0.1089	1.3572	0.8016	0.7660	0.6449	0.0756	0.1626	**0.0414**	**0.0867**
	192	0.1129	0.1524	0.1113	0.1505	0.1151	0.1536	0.1378	0.1763	1.3842	0.8280	0.9683	0.7014	0.1491	0.2120	**0.0992**	**0.1455**

4.3 Ablation Studies

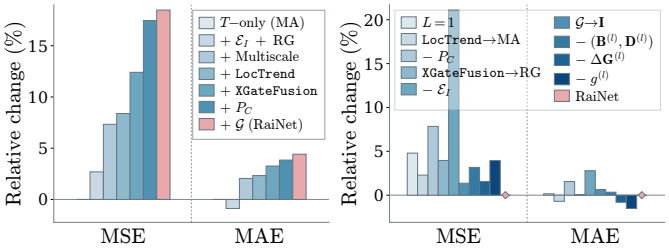


Figure 6: Cumulative and standalone ablations of RaiNet.

Figure 6 summarizes cumulative additions (left) and standalone variants (right) across all three basins. Relative changes are computed within each basin and then averaged to account for different targets and error scales. MA and RG denote moving average and residual gate, respectively; $L = 1$ removes multiscale modeling, while '−" and '→" denote removal and replacement. Starting from the temporal-only baseline, progressively introducing the rainfall-event branch, multiscale modeling, `LocTrend`, `XGateFusion`, P_C, and $\mathcal{G}$ reduces average MSE and MAE by 18.5% and 4.4%, respectively. The MSE gain is retained across all three basins. Standalone variants provide complementary evidence: removing $\mathcal{E}_I$ causes the largest degradation, while removing P_C or replacing `LocTrend` and `XGateFusion` also increases average MSE. MAE changes are smaller and occasionally basin-dependent because MSE emphasizes consequential rainfall-associated excursions, whereas MAE is dominated by routine observations. This pattern is consistent with heterogeneous basin responses, including rapid turbidity pulses and more dispersed nitrate transport. Further analyses are provided in Appendix A.11.

4.4 Qualitative Analysis

Figure 7 analyzes `LocTrend` at the 1, 2, and 4-hour scales. The line plots show that the same temporal representation is decomposed into different locally evolving backgrounds and exact dynamic deviations as the resolution changes: finer scales retain more local structural variation, whereas the background becomes progressively smoother at coarser scales. The lollipop plots further explain this adaptation. *Pair use* measures the relative utilization of each positive-negative dictionary pair, while *Sign bias* indicates the preference between its two signs. Although all scales share the same structural dictionary, their distinct assignment patterns show that `LocTrend` selectively recombines shared structural atoms according to the temporal resolution. This shared-dictionary yet scale-adaptive design enables flexible background-deviation decomposition without introducing scale-specific dictionaries or predefined temporal

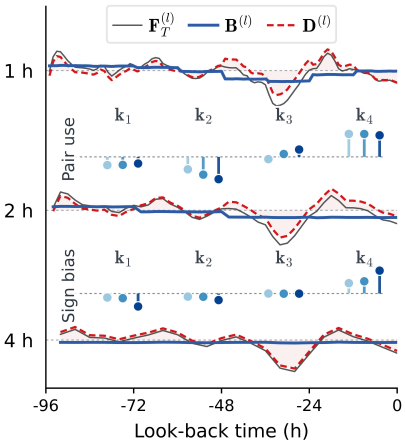


Figure 7: `LocTrend` analysis.

Table 2: Forecasting results with water-quality-specific and diffusion-based models. The best and second-best results are highlighted in **bold** and underline, respectively.

Model		PIWON		D-TGCN		CMLIP		RGCN		XGBR		SimDiff		MA-TSD		RaiNet	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
170900	24	0.2523	0.2013	0.2719	0.2111	0.2796	0.1794	0.2976	0.2113	0.3924	0.1913	0.2773	0.1641	0.2843	0.2023	**0.2185**	**0.1560**
	96	0.5303	0.3326	0.4633	0.3072	0.4943	0.2723	0.4795	0.3179	0.8964	0.3228	0.5062	0.2643	0.5011	0.3314	**0.4096**	**0.2604**
	192	0.9224	0.5161	0.5618	0.3661	0.6139	0.3338	0.5669	0.3673	1.4016	0.4131	0.6387	0.3250	0.6263	0.4098	**0.5283**	**0.3102**
071200	24	0.0961	0.1936	0.1166	0.2043	0.1272	0.2011	0.1784	0.2460	0.1227	0.1999	0.1387	0.2068	0.1280	0.2168	**0.0872**	**0.1747**
	96	0.4115	0.3869	0.4347	0.4086	0.4734	0.3919	0.4476	0.3912	0.4779	0.3951	0.4708	0.3911	0.4824	0.4257	**0.3225**	**0.3344**
	192	0.5877	0.4698	0.6435	0.5066	0.7240	0.5107	0.6450	0.4971	0.8040	0.5385	0.7305	0.5194	0.7246	0.5501	**0.5415**	**0.4443**
051201	24	0.0090	0.0461	0.0093	0.0448	0.0081	0.0343	0.0980	0.1883	0.0121	0.0551	0.0098	0.0401	0.0081	0.0367	**0.0072**	**0.0318**
	96	0.0555	0.1233	0.0569	0.1180	0.0516	0.0951	0.1394	0.2101	0.0565	0.1102	0.0530	0.0930	0.0482	0.1009	**0.0414**	**0.0867**
	192	0.1316	0.1981	0.1241	0.1884	0.1206	0.1558	0.2055	0.2689	0.1272	0.1702	0.1260	0.1593	0.1130	0.1642	**0.0992**	**0.1455**

periods. This enables irregular but recurrent local structures to be absorbed into the background while preserving transient departures in the deviation branch. The nearly flat background at the coarsest scale further supports bypassing decomposition after additional downsampling.

Figure 8 shows how RaiNet progressively transforms rainfall into conditional forecast corrections. The rainfall event encoder distinguishes similar station-local rainfall under different basin-wide contexts and produces different station-aligned event states instead of relying on absolute rainfall intensity alone (a). Given the same rainfall-event representation but different water-quality histories, `XGateFusion` learns distinct bidirectional relative-lag profiles across temporal scales, demonstrating context-dependent statistical alignment (b). These profiles are subsequently translated into signed corrections whose direction and magnitude vary across scales and stations (c). Additional qualitative analyses are provided in Appendix A.12.

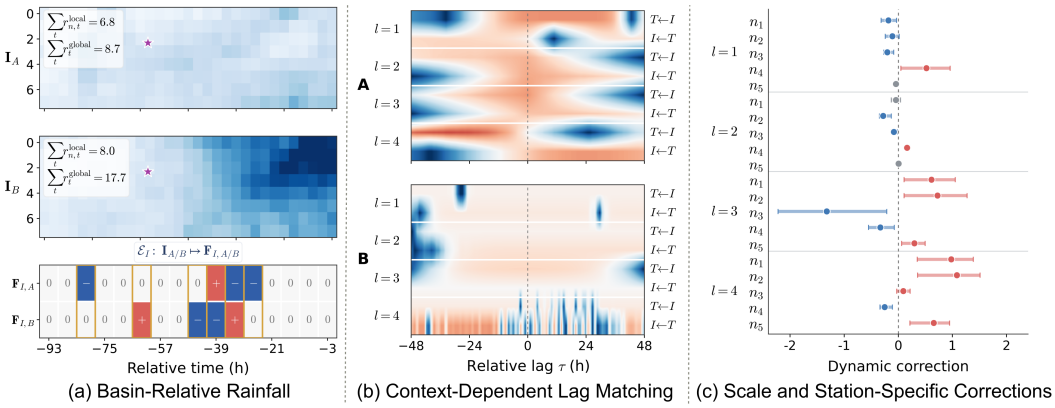


Figure 8: **Analysis of the rainfall event encoder and `XGateFusion`.** (a) $\mathcal{E}_I$ distinguishes station-aligned events under different basin contexts. (b) `XGateFusion` learns context and scale-dependent relative-lag profiles. (c) The resulting corrections to the baseline predictions vary across scales and stations; points and intervals denote their means and 10-90th percentiles.

5 Conclusion

In this work, we presented RaiNet, a rainfall-informed multimodal framework for long-term water-quality forecasting. RaiNet models multiscale nonstationary dynamics through `LocTrend`, converts gridded precipitation into basin-relative station-aligned rainfall events, and selectively integrates them using lag-aware `XGateFusion`. Experiments on three real-world datasets demonstrate consistent improvements over competitive forecasting and multimodal baselines, while qualitative analyses verify the distinct roles of each module. These results provide a data-driven approach toward accurate future water-quality forecasting.

AI USE STATEMENT

We did not use generative AI tools for tasks requiring disclosure, including research ideation, methodological or experimental design. Other required disclosure tasks, including synthetic-data generation and the development of theoretical claims or proofs, are not applicable to this work. Generative AI tools were used solely to edit author-written manuscript text for grammar, clarity, concision, and readability. All AI-assisted edits were reviewed and revised by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses publicly available environmental monitoring records and gridded precipitation products and does not involve human participants, personal data, or sensitive attributes. The study is intended to support water-quality forecasting research. Its predictions should be used as decision support rather than as a substitute for operational monitoring or expert judgment.

REPRODUCIBILITY STATEMENT

The main paper and the Appendix describe the model architecture, data construction, chronological splits, training protocol, hyperparameters, and evaluation metrics. The supplementary material includes the implementation code.

REFERENCES

Nourhan Ahmed, Vijaya Krishna Yalavarthi, and Lars Schmidt-Thieme. Motif-aware graph neural networks for networked time series imputation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11409–11417, 2025.

Salou Moumouni Ali, Syed Usama Imtiaz, Nasr Azadani Mitra, and Alamdari Nasrin. Near real-time and next-day prediction for escherichia coli (e. coli) concentrations in highly urbanized watersheds. *Water Research*, pp. 125030, 2025.

Sang-Soo Baek, Jongcheol Pyo, and Jong Ahn Chun. Prediction of water level and water quality using a cnn-lstm combined deep learning approach. *Water*, 12(12):3399, 2020.

Jing Bi, Yibo Li, Haitao Yuan, Mengyuan Wang, Ziqi Wang, Jia Zhang, and Mengchu Zhou. Hybrid water quality prediction with multimodal low-rank fusion and localized attention. *IEEE Internet of Things Journal*, 12(12):21158–21169, 2025a.

Jing Bi, Ziqi Wang, Haitao Yuan, Xiangxi Wu, Renren Wu, Jia Zhang, and MengChu Zhou. Long-term water quality prediction with transformer-based spatial-temporal graph fusion. *IEEE Transactions on Automation Science and Engineering*, 22:11392–11404, 2025b.

Fanpu Cao, Lu Dai, Jindong Han, and Hui Xiong. Enhancing multivariate time series forecasting with global temporal retrieval. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QUJBPSfyui.

Chris Chatfield. The holt-winters forecasting procedure. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 27(3):264–279, 1978.

Min Chen, Guansong Pang, Wenjun Wang, and Cheng Yan. Information bottleneck-guided mlps for robust spatial-temporal forecasting. In *Forty-second International Conference on Machine Learning*, 2025.

Shengyue Chen, Jinliang Huang, Peng Wang, Xi Tang, and Zhenyu Zhang. A coupled model to improve river water quality prediction towards addressing non-stationarity and data limitation. *Water Research*, 248:120895, 2024.

Hang Ding, Xue Wang, Tian Zhou, and Tao Yao. Simdiff: Simpler yet better diffusion model for time series point forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20781–20789, 2026.

Haotian Gao, Zheng Dong, Jiawei Yong, Shintaro Fukushima, Kenjiro Taura, and Renhe Jiang. How different from the past? spatio-temporal time series forecasting with self-supervised deviation learning. *Advances in Neural Information Processing Systems*, 38:146647–146675, 2026.

Thi Kieu Khanh Ho, Ali Karami, and Narges Armanfard. Graph anomaly detection in time series: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Sheng Huang, Yueling Wang, and Jun Xia. Which riverine water quality parameters can be predicted by meteorologically-driven deep learning? *Science of The Total Environment*, 946:174357, 2024.

Songtao Huang, Zhen Zhao, Can Li, and Lei Bai. Timekan: Kan-based frequency decomposition learning architecture for long-term time series forecasting. *arXiv preprint arXiv:2502.06910*, 2025. URL https://arxiv.org/abs/2502.06910.

Valentina Krysanova and Mike White. Advances in water resources assessment with swat-an overview. *Hydrological Sciences Journal*, 60(5):771–783, 2015.

Aitian Ma, Dongsheng Luo, and Mo Sha. Mixlinear: Extreme low resource multivariate time series forecasting with $0.1k$ parameters. In *International Conference on Learning Representations*, volume 2026, pp. 77764–77789, 2026.

Christian Ortiz-Lopez, Andres Torres, Christian Bouchard, and Manuel Rodriguez. A methodology for integrating time-lagged rainfall and river flow data into machine learning models to improve prediction of quality parameters of raw water supplying a treatment plant. *Journal of Hydroinformatics*, 25(6):2406–2426, 2023.

Olivier Papillon, Rafik Goubran, James Green, Julien Larivière-Chartier, Caitlin Higginson, Frank Knoefel, and Rébecca Robillard. Sleep stage classification using multimodal embedding fusion from electrooculography and pressure-sensitive mats. In *2025 IEEE Medical Measurements and Applications (MeMeA)*, pp. 1–6, 2025. doi: 10.1109/MeMeA65319.2025.11068039.

R. Sathya, Leander P Steve, Nitish Narayan, Abhishek Upadhye, and M. Aravindharaj. Covid wave prediction using sarima machine learning algorithm. In *2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pp. 1–5, 2023. doi: 10.1109/SCEECS57921.2023.10062979.

Minlan Shao, Zijian Zhang, Yili Wang, Yiwei Dai, Xu Shen, and Xin Wang. Hyperd: Hybrid periodicity decoupling framework for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15689–15697, 2026.

Robert H Shumway and David S Stoffer. Arima models. In *Time series analysis and its applications: with R examples*, pp. 75–163. Springer, 2017.

Jared D Smith, Lauren E Koenig, Margaux J Sleckman, Alison P Appling, Jeffrey M Sadler, Vincent T DePaul, and Zoltan Szabo. Predictive understanding of stream salinization in a developed watershed using machine learning. *Environmental Science & Technology*, 58(42):18822, 2024.

Sijie Tang, Jiping Jiang, Shuo Wang, Yi Zheng, Dragan Savic, and Aijie Wang. Non-point source pollution prediction and dynamics simulation in urban runoff: a physics-informed neural network approach. *Water Research*, pp. 126379, 2026.

Hang Wan, Long Xiang, Yanpeng Cai, Yulei Xie, and Rui Xu. Temporal and spatial feature extraction using graph neural networks for multi-point water quality prediction in river network areas. *Water Research*, 281:123561, 2025.

Chenxi Wang, Linxiao Yang, Zhixian Wang, Liang Sun, and Yi Wang. A non-isotropic time series diffusion model with moving average transitions. In *Forty-second International Conference on Machine Learning*, 2025a.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and JUN ZHOU. Timemixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=7oLshfEIC2.

Weichen Wang, Guowangchen Liu, Mingjing Wang, Yan Pan, Yukun Ma, Lu Yang, Jing Sang, Zhenyao Shen, and Lei Chen. Improving watershed-scale daily nutrient simulation using a process-model-informed graph attention network with multi-source data integration. *Water Research*, pp. 125532, 2026.

Xixi Wang, Xiao Wang, Bo Jiang, Jin Tang, and Bin Luo. Mutualformer: Multi-modal representation learning via cross-diffusion attention. *International Journal of Computer Vision*, 132(9): 3867–3888, 2024b.

Yulong Wang, Yushuo Liu, Xiaoyi Duan, and Kai Wang. Filterts: Comprehensive frequency filtering for multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 35438–35445, 2025b. doi: 10.1609/aaai.v39i20.35438.

Tim Wool, Robert B Ambrose Jr, James L Martin, and Alex Comer. Wasp 8: The next generation in the 50-year evolution of usepa's water quality model. *Water*, 12(5):1398, 2020.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Duojin Yan, Chunjie Yang, Shuming Sun, Siwei Lou, Liyuan Kong, and Yuyan Zhang. One-sided relational autoencoder with seasonal-trend decomposition to extract process correlations for molten iron quality prediction. *IEEE Transactions on Instrumentation and Measurement*, 73:1–13, 2024. doi: 10.1109/TIM.2024.3353844.

Ruibin Zhang, Xin Qian, Xingcheng Yuan, Rui Ye, Bisheng Xia, and Yulei Wang. Simulation of water environmental capacity and pollution load reduction using qual2k for water environmental management. *International journal of environmental research and public health*, 9(12):4504–4521, 2012.

Xuanran Zhang, Yufeng He, Zikuan Zhou, Weihao Li, Lizhi Tao, Jie Chen, Haibo Zou, and Hui Lin. Incorporating flow direction as a physical prior in graph networks: A directed temporal graph convolutional network for enhanced water quality forecasting in lake basins. *Transactions in GIS*, 30(4):e70309, 2026.

Yingyin Zhang, Xianqiang He, Gang Lian, Yan Bai, Ying Yang, Fang Gong, Difeng Wang, Zili Zhang, Teng Li, and Xuchen Jin. Monitoring and spatial traceability of river water quality using sentinel-2 satellite images. *Science of The Total Environment*, 894:164862, 2023.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. *arXiv preprint arXiv:2502.04395*, 2025.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022. URL https://arxiv.org/abs/2201.12740. arXiv:2201.12740.

A Appendix

A.1 Multiscale Embedding Strategy

A.1.1 Temporal (Sensor) Modality.

At scale l, the water-quality sequence is denoted by $\mathbf{X}^{(l)} \in \mathbb{R}^{T_l \times N}$, where T_l and N are the temporal length and number of monitoring stations, respectively. After scale-wise normalization, a

convolutional value embedding is applied independently to each station sequence:

$$\mathbf{e}_t^{\mathrm{val}} = \mathrm{Conv1D}_3\left(\mathrm{Norm}_l(\mathbf{X}^{(l)})\right)_t, \qquad \mathbf{e}_t^{\mathrm{time}} = \mathbf{m}_t \mathbf{W}_{\mathrm{time}}, \qquad \mathbf{e}_t^{\mathrm{val}}, \mathbf{e}_t^{\mathrm{time}} \in \mathbb{R}^d, \tag{11}$$

where $\mathrm{Conv1D}_3(\cdot)$ is a one-dimensional convolution with kernel size 3 and circular padding, and $\mathrm{Norm}_l(\cdot)$ denotes normalization at scale l. The timestamp vector $\mathbf{m}_t \in \mathbb{R}^4$ contains the normalized hour of day, day of week, day of month, and day of year at time step t, while $\mathbf{W}_{\mathrm{time}} \in \mathbb{R}^{4\times d}$ is a learnable linear projection. The value and temporal-feature embeddings are summed to produce

$$\mathbf{F}_T^{(l)} = \left[\mathrm{Dropout}\left(\mathbf{e}_t^{\mathrm{val}} + \mathbf{e}_t^{\mathrm{time}}\right)\right]_{t=1}^{T_l} \in \mathbb{R}^{T_l \times d}. \tag{12}$$

The embedding parameters are shared across temporal scales, and the timestamp features are subsampled consistently with the water-quality sequence at each scale.

A.1.2 Rainfall-Event Encoder Details

Station-centered spatial prior. The spatial prior of station n is constructed by evaluating a Gaussian kernel over the rainfall grid:

$$\pi_{n,p} = \frac{\exp\left(-\|\mathbf{z}_p - \mathbf{c}_n\|_2^2/(2\sigma^2)\right)}{\sum_{p'} \exp(-\|\mathbf{z}_{p'} - \mathbf{c}_n\|_2^2/(2\sigma^2))}, \tag{13}$$

where $\mathbf{z}_p$ is the coordinate of grid cell p, $\mathbf{c}_n$ is the geographic location of station n, and σ controls the spatial extent of the station neighborhood. At each time step, invalid rainfall cells are excluded and the remaining prior weights are renormalized, ensuring that the station-local rainfall is computed only from valid observations.

Training-only event threshold. The station-specific threshold ϵ_n is fitted using only rainfall histories associated with the training origins:

$$\epsilon_n = \sqrt{\frac{1}{|\mathcal{S}_{\mathrm{tr}}|} \sum_{t \in \mathcal{S}_{\mathrm{tr}}} \left(r_{n,t}^{\mathrm{rel}}\right)^2}, \tag{14}$$

where $\mathcal{S}_{\mathrm{tr}}$ denotes the collection of historical rainfall time steps contained in the training samples. This root-mean-square threshold characterizes the typical magnitude of the local-to-basin contrast at each station. The fitted thresholds remain fixed during validation and testing. The ternary state encodes the direction of salient local-to-basin contrast and reduces sensitivity to within-state amplitude fluctuations; continuous basin-level rainfall summaries retain complementary basin-scale magnitude information for prediction.

Persistence aggregation and scale alignment. The time-indexed states $e_{n,t}$ are partitioned into non-overlapping persistence blocks using $h = 6$. Signed voting assigns one state to each block, which is repeated over all positions in the corresponding block to construct $\mathbf{F}_I \in \mathbb{R}^{T\times N}$. The block construction is adapted to the sampling interval of each dataset. For scale l, the aligned event sequence is computed as

$$\left[\mathbf{F}_I^{(l)}\right]_{k,n} = \mathrm{sign}\left(\sum_{s=(k-1)q_l+1}^{kq_l} \left[\mathbf{F}_I\right]_{s,n}\right), \qquad k = 1, \ldots, T_l, \tag{15}$$

where q_l is the aggregation factor shared with the corresponding water-quality scale, k indexes the aligned temporal positions, and $T_l = T/q_l$ is the resulting sequence length. The summation accumulates event evidence over the temporal support of scale l, while the sign preserves its dominant local-to-basin direction. All blocks contain only historical rainfall observations available at the forecast origin.

A.2 Detailed Derivation and Theoretical Guarantees for LocTrend

A.2.1 Training-Only Structural Initialization and Joint Learning

The initialization procedure is performed before gradient-based training and uses neither validation nor test samples. Historical water-quality inputs are processed by the temporal encoder at its initial

parameter values in evaluation mode, with gradient recording disabled. This encoder is not separately pretrained. Its parameters remain unchanged throughout both initialization passes and are subsequently learned jointly with the dictionary. The initialization contains two passes: the first extracts temporal directions shared by centered local windows, and the second converts these directions into feature-space atoms used for background reconstruction. This two-pass construction separates *when* recurrent structures occur within a local window from *how* they are represented in the latent feature space: Pass 1 identifies shared temporal organizations, while Pass 2 converts them into reusable feature-space atoms. Let $\mathcal{Q}_{\mathrm{tr}}$ denote the collection of overlapping windows extracted from all training origins and decomposed temporal scales. Each centered window is written as $\widetilde{\mathcal{W}}_q = [\widetilde{\mathbf{w}}_{q,1}, \ldots, \widetilde{\mathbf{w}}_{q,w}] \in \mathbb{R}^{w \times d}$, where $q \in \mathcal{Q}_{\mathrm{tr}}$ indexes a training window, w is its length, and d is the feature dimension.

Pass 1: shared temporal eigendirections. For each window, we first compute the pairwise squared distances

$$d^2_{q,t,u} = \|\widetilde{\mathbf{w}}_{q,t} - \widetilde{\mathbf{w}}_{q,u}\|_2^2, \qquad t, u \in \{1, \ldots, w\}. \tag{16}$$

The RBF bandwidth is adapted to the local feature scale using the median pairwise squared distance over distinct temporal positions (including zero distances):

$$\gamma_q = \left(\mathrm{median}_{t<u}\, d^2_{q,t,u}\right)^{-1}. \tag{17}$$

Numerically, the denominator is lower-bounded by a small positive constant. The corresponding RBF Gram matrix is

$$[\mathbf{G}_q]_{t,u} = \exp\left(-\gamma_q d^2_{q,t,u}\right). \tag{18}$$

Unlike a linear similarity matrix, $\mathbf{G}_q$ emphasizes nonlinear proximity between temporal positions whose feature states exhibit similar local structures. Each Gram matrix is centered as

$$\mathbf{G}_{q,c} = \mathbf{J}\mathbf{G}_q\mathbf{J}, \qquad \mathbf{J} = \mathbf{I}_w - \frac{1}{w}\mathbf{1}_w\mathbf{1}_w^\top, \tag{19}$$

where $\mathbf{I}_w$ is the $w \times w$ identity matrix and $\mathbf{1}_w$ is an all-one vector of length w. We then average the centered matrices over all training windows:

$$\overline{\mathbf{G}} = \frac{1}{|\mathcal{Q}_{\mathrm{tr}}|} \sum_{q \in \mathcal{Q}_{\mathrm{tr}}} \mathbf{G}_{q,c}. \tag{20}$$

The top K eigenpairs of this shared matrix satisfy

$$\overline{\mathbf{G}}\mathbf{u}_r = \lambda_r \mathbf{u}_r, \qquad r = 1, \ldots, K, \tag{21}$$

where λ_r and $\mathbf{u}_r \in \mathbb{R}^w$ are the r-th retained eigenvalue and temporal eigendirection, respectively. Because they are extracted from the average kernel geometry, these directions summarize nonlinear temporal organizations recurring across the training windows instead of patterns from a single sample.

Pass 2: feature-space atom construction. The temporal eigendirections are next applied to every centered training window. For component r, we compute

$$\mathbf{z}_{q,r} = \frac{1}{\sqrt{\lambda_r}} \sum_{t=1}^{w} u_{t,r}\widetilde{\mathbf{w}}_{q,t} \in \mathbb{R}^d, \qquad \overline{\mathbf{z}}_{q,r} = \frac{\mathbf{z}_{q,r}}{\|\mathbf{z}_{q,r}\|_2}, \tag{22}$$

where $u_{t,r}$ is element t of $\mathbf{u}_r$. Thus, each temporal direction produces a normalized direction in the feature space of a training window. In other words, $u_{t,r}$ acts as a temporal weighting pattern over the window, projecting the w temporal tokens into a d-dimensional feature direction $\mathbf{z}_{q,r}$. Components associated with larger kernel eigenvalues receive greater weight: $\omega_r = \frac{\lambda_r}{\max_j \lambda_j}$. Their weighted feature covariance is accumulated over all windows:

$$\mathbf{C}_F = \frac{1}{K|\mathcal{Q}_{\mathrm{tr}}|} \sum_{q \in \mathcal{Q}_{\mathrm{tr}}} \sum_{r=1}^{K} \omega_r \overline{\mathbf{z}}_{q,r}\overline{\mathbf{z}}_{q,r}^\top \in \mathbb{R}^{d \times d}. \tag{23}$$

Finally, the top K eigenvectors of $\mathbf{C}_F$ initialize the structural dictionary:

$$\mathbf{C}_F\mathbf{k}_r^{(0)} = \xi_r \mathbf{k}_r^{(0)}, \qquad \mathbf{K}^{(0)} = \{\mathbf{k}_r^{(0)}\}_{r=1}^K. \tag{24}$$

where ξ_r is the corresponding feature-covariance eigenvalue, and $\mathbf{k}_r^{(0)} \in \mathbb{R}^d$ is an initial structural atom. Eigenvector signs are fixed deterministically for reproducibility.

After the two passes, $\mathbf{K}$ is initialized to $\mathbf{K}^{(0)}$ and included in the forecasting optimizer as a trainable parameter shared across all decomposed scales. Gradients reach the atoms through both cosine-softmax assignment and background reconstruction, allowing the dictionary and the temporal encoder to co-adapt to the forecasting objective. The temporal eigenvectors and eigenvalues used for initialization are stored as buffers and are not recomputed during ordinary training. Thus, training-only initialization does not imply a frozen reconstruction dictionary.

During each forward pass, the current dictionary is expanded as $\mathbf{K}_\pm = [\mathbf{K}; -\mathbf{K}]$. Atom directions are normalized for cosine matching, while their current, unnormalized values are used for synthesis. Orthogonality of the initial eigenvectors is not enforced after optimization. The trained dictionary is held fixed during validation and inference; local means and assignment weights still depend on the input window. This input-dependent reconstruction does not by itself guarantee robustness to unseen distribution shifts.

A.2.2 LocTrend Forward Decomposition

The procedure is applied independently to each station and each decomposed scale $l = 1, \ldots, L - L_{\mathrm{dir}}$. For scale l, the inputs are the temporal representation $\mathbf{F}_T^{(l)}$, the fitted dictionary $\mathbf{K} = \{\mathbf{k}_r\}_{r=1}^K$, the window length w_l, stride s_l, assignment temperature ϑ, and reconstruction gain ρ. The outputs are the structural background $\mathbf{B}^{(l)}$ and dynamic deviation $\mathbf{D}^{(l)}$.

Window coverage. The window starting positions are collected in

$$\mathcal{Q}_l = \{1,\, 1 + s_l,\, 1 + 2s_l, \ldots\} \cup \{T_l - w_l + 1\}\,, \tag{25}$$

where only positions not exceeding $T_l - w_l + 1$ are retained. The final position is included when necessary so that every temporal position is covered by at least one window. For a window beginning at $q_i \in \mathcal{Q}_l$, we extract and center

$$\mathcal{W}_i^{(l)} = \mathbf{F}^{(l)}_{T,\, q_i:q_i+w_l-1}, \qquad \boldsymbol{\mu}_i^{(l)} = \frac{1}{w_l}\sum_{t=1}^{w_l} \mathcal{W}_{i,t}^{(l)}, \qquad \widetilde{\mathbf{w}}_{i,t}^{(l)} = \mathcal{W}_{i,t}^{(l)} - \boldsymbol{\mu}_i^{(l)}. \tag{26}$$

This removes the local level before structural matching, so dictionary assignment depends on local shape rather than absolute magnitude.

Signed soft reconstruction. The current shared dictionary is symmetrically expanded as $\mathbf{K}_\pm = [\mathbf{K}; -\mathbf{K}]$. The symmetric expansion allows the same structural atom to represent locally recurring patterns with opposite feature directions, avoiding the need to learn separate atoms for both signs. For each centered token, the assignment weights are

$$a_{i,t,r}^{(l)} = \frac{\exp\Big(\cos\Big(\widetilde{\mathbf{w}}_{i,t}^{(l)}, \mathbf{k}_{\pm,r}\Big)/\vartheta\Big)}{\sum_{j=1}^{2K}\exp\Big(\cos\Big(\widetilde{\mathbf{w}}_{i,t}^{(l)}, \mathbf{k}_{\pm,j}\Big)/\vartheta\Big)}. \tag{27}$$

The resulting centered structural reconstruction is then combined with the local level to obtain the window-wise background:

$$\widetilde{\mathbf{b}}_{i,t}^{(l)} = \sum_{r=1}^{2K} a_{i,t,r}^{(l)}\mathbf{k}_{\pm,r}, \qquad \mathbf{b}_{i,t}^{(l)} = \boldsymbol{\mu}_i^{(l)} + \rho\widetilde{\mathbf{b}}_{i,t}^{(l)}. \tag{28}$$

Overlap aggregation. Each local reconstruction is added to its original temporal position. Let $\Omega_\tau^{(l)}$ denote the set of windows covering global position τ. The structural background and exact dynamic remainder are

$$\mathbf{B}_\tau^{(l)} = \frac{1}{|\Omega_\tau^{(l)}|}\sum_{i\in\Omega_\tau^{(l)}} \mathbf{b}^{(l)}_{i,\,\tau-q_i+1}, \qquad \mathbf{D}^{(l)} = \mathbf{F}_T^{(l)} - \mathbf{B}^{(l)}. \tag{29}$$

Coverage normalization prevents positions appearing in more overlapping windows from receiving systematically larger background values. The full forward procedure is summarized in Algorithm 1.

Algorithm 1 `LocTrend` Forward Decomposition

Input: temporal representation $\mathbf{F}_T^{(l)}$; current shared dictionary $\mathbf{K}$; window length w_l; stride s_l; temperature ϑ; gain ρ
Output: structural background $\mathbf{B}^{(l)}$; dynamic deviation $\mathbf{D}^{(l)}$
Form the signed dictionary $\mathbf{K}_{\pm} \leftarrow [\mathbf{K}; -\mathbf{K}]$
Construct window starts $\mathcal{Q}_l$, including the final start $T_l - w_l + 1$ when required
Initialize $\mathbf{B}_{\text{sum}}^{(l)} \leftarrow \mathbf{0}_{T_l \times d}$ and coverage counts $\mathbf{m}^{(l)} \leftarrow \mathbf{0}_{T_l}$
for each starting position $q_i \in \mathcal{Q}_l$ **do**
  Extract $\mathcal{W}_i^{(l)} \leftarrow \mathbf{F}_{T,\, q_i:q_i+w_l-1}^{(l)}$
  Compute $\boldsymbol{\mu}_i^{(l)} \leftarrow w_l^{-1} \sum_{t=1}^{w_l} \mathcal{W}_{i,t}^{(l)}$
  Center $\widetilde{\mathcal{W}}_i^{(l)} \leftarrow \mathcal{W}_i^{(l)} - \boldsymbol{\mu}_i^{(l)}$
  for $t = 1$ to w_l **do**
    Compute cosine similarities between $\widetilde{\mathbf{w}}_{i,t}^{(l)}$ and all atoms $\mathbf{k}_{\pm,r}$ in $\mathbf{K}_{\pm}$
    Obtain assignment weights $\{a_{i,t,r}^{(l)}\}_{r=1}^{2K} \leftarrow \mathrm{softmax}_r\left(\cos\left(\widetilde{\mathbf{w}}_{i,t}^{(l)}, \mathbf{k}_{\pm,r}\right)/\vartheta\right)$
    Reconstruct $\widetilde{\mathbf{b}}_{i,t}^{(l)} \leftarrow \sum_{r=1}^{2K} a_{i,t,r}^{(l)} \mathbf{k}_{\pm,r}$
    Restore the local level and scale the structural reconstruction: $\mathbf{b}_{i,t}^{(l)} \leftarrow \boldsymbol{\mu}_i^{(l)} + \rho \widetilde{\mathbf{b}}_{i,t}^{(l)}$
    Set the global position $\tau \leftarrow q_i + t - 1$
    Accumulate the local reconstruction: $\mathbf{B}_{\text{sum},\tau}^{(l)} \leftarrow \mathbf{B}_{\text{sum},\tau}^{(l)} + \mathbf{b}_{i,t}^{(l)}$
    Update the coverage count: $m_\tau^{(l)} \leftarrow m_\tau^{(l)} + 1$
  end for
end for
Normalize overlapping reconstructions: $\mathbf{B}_\tau^{(l)} \leftarrow \mathbf{B}_{\text{sum},\tau}^{(l)} / m_\tau^{(l)}$, for $\tau = 1, \ldots, T_l$
Compute the exact dynamic deviation: $\mathbf{D}^{(l)} \leftarrow \mathbf{F}_T^{(l)} - \mathbf{B}^{(l)}$
return $\mathbf{B}^{(l)}, \mathbf{D}^{(l)}$

A.2.3 Local Adaptation of LocTrend

We formalize how `LocTrend` adapts to locally shared level shifts in the temporal representation without relying on a predefined period. The scale and station indices are omitted for clarity. The following property holds for any given dictionary, including the jointly learned dictionary, with the same parameters used for both inputs. For temporal position t, let Ω_t denote the set of overlapping windows covering t, and define its effective local neighborhood as

$$\mathcal{N}_t = \bigcup_{i \in \Omega_t} \{q_i, \ldots, q_i + w - 1\}, \tag{30}$$

where q_i and w are the starting position and length of window i, respectively.

Theorem 1 (Local Level Equivariance of `LocTrend`)**.** *Let $\mathbf{F}$ be an input representation with `LocTrend` decomposition $\mathbf{F} = \mathbf{B} + \mathbf{D}$. For any constant feature offset $\boldsymbol{\delta} \in \mathbb{R}^d$, construct a locally shifted input*

$$\mathbf{F}'_\tau = \begin{cases} \mathbf{F}_\tau + \boldsymbol{\delta}, & \tau \in \mathcal{N}_t, \\ \mathbf{F}_\tau, & \tau \notin \mathcal{N}_t. \end{cases} \tag{31}$$

Then the `LocTrend` decomposition at position t satisfies

$$\mathbf{B}'_t = \mathbf{B}_t + \boldsymbol{\delta}, \qquad \mathbf{D}'_t = \mathbf{D}_t. \tag{32}$$

Proof. For every window $i \in \Omega_t$, all tokens in that window belong to $\mathcal{N}_t$ and are shifted by the same vector $\boldsymbol{\delta}$. Its local mean therefore becomes

$$\boldsymbol{\mu}'_i = \boldsymbol{\mu}_i + \boldsymbol{\delta}. \tag{33}$$

For every local position $j = 1, \ldots, w$, the centered token remains unchanged:

$$\widetilde{\mathbf{w}}'_{i,j} = (\mathbf{w}_{i,j} + \boldsymbol{\delta}) - (\boldsymbol{\mu}_i + \boldsymbol{\delta}) = \widetilde{\mathbf{w}}_{i,j}. \tag{34}$$

Consequently, its cosine similarities to the signed dictionary, soft assignments, and centered reconstruction remain unchanged. Restoring the shifted local mean therefore gives

$$\mathbf{b}'_{i,j} = \mathbf{b}_{i,j} + \boldsymbol{\delta}. \tag{35}$$

Coverage-normalized averaging over all windows in Ω_t yields

$$\mathbf{B}'_t = \frac{1}{|\Omega_t|} \sum_{i \in \Omega_t} \mathbf{b}'_{i,\, t-q_i+1} = \mathbf{B}_t + \boldsymbol{\delta}. \tag{36}$$

Since $\mathbf{F}'_t = \mathbf{F}_t + \boldsymbol{\delta}$ and $\mathbf{D}_t = \mathbf{F}_t - \mathbf{B}_t$, we obtain

$$\mathbf{D}'_t = \mathbf{F}'_t - \mathbf{B}'_t = \mathbf{F}_t - \mathbf{B}_t = \mathbf{D}_t. \tag{37}$$

□

The theorem shows that `LocTrend` absorbs a locally shared representation-level shift into the structural background while leaving the dynamic deviation unchanged. This property follows directly from local centering and overlap reconstruction and does not require the shift to recur at a predefined temporal period.

A.3 Component-Specific Cross-Scale Mixing

Let $L_{\text{dec}} = L - L_{\text{dir}}$ denote the number of scales processed by `LocTrend`. Cross-scale mixing is performed only among these decomposed scales.

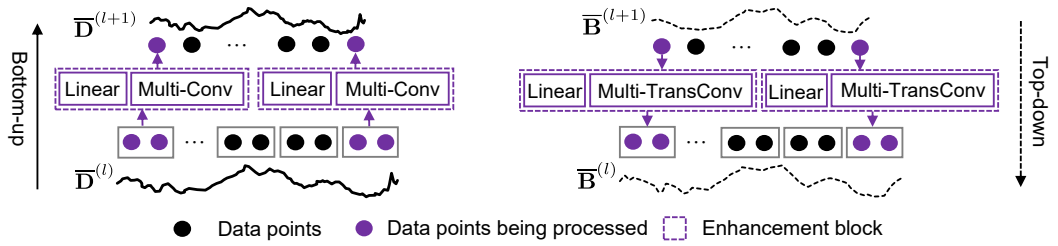


Figure 9: Component-specific cross-scale mixing with opposite propagation directions. Dynamic deviations are enhanced bottom-up from fine to coarse scales, whereas structural backgrounds are enhanced top-down from coarse to fine scales. At each transition, a temporal linear path and a multi-kernel convolutional or transposed-convolutional path are adaptively combined through learnable softmax fusion.

Fine-to-coarse deviation mixing. Dynamic deviations are propagated sequentially from scale l to the coarser scale $l+1$. We initialize $\overline{\mathbf{D}}^{(1)} = \mathbf{D}^{(1)}$ and, for $l = 1, \ldots, L_{\text{dec}} - 1$, compute

$$\mathbf{U}_{D,\text{conv}}^{(l\to l+1)} = \text{Conv}_{\{3,5,7\},\downarrow}^{(l)} \left(\overline{\mathbf{D}}^{(l)} \right), \tag{38}$$

$$\mathbf{U}_{D,\text{lin}}^{(l\to l+1)} = \text{MLP}_{\downarrow}^{(l)} \left(\overline{\mathbf{D}}^{(l)} \right), \tag{39}$$

$$\overline{\mathbf{D}}^{(l+1)} = \mathbf{D}^{(l+1)} + \text{SF}_{D}^{(l)} \left(\mathbf{U}_{D,\text{conv}}^{(l\to l+1)}, \mathbf{U}_{D,\text{lin}}^{(l\to l+1)} \right). \tag{40}$$

where $\text{Conv}_{\{3,5,7\},\downarrow}^{(l)}$ denotes a strided temporal convolutional path containing kernels of sizes 3, 5, and 7, while $\text{MLP}_{\downarrow}^{(l)}$ maps the temporal length from T_l to T_{l+1}. The former extracts local variations over multiple receptive fields, whereas the latter preserves a direct global transformation between the two scales.

Coarse-to-fine background mixing. Structural backgrounds follow the reverse direction. Starting from $\overline{\mathbf{B}}^{(L_{\mathrm{dec}})} = \mathbf{B}^{(L_{\mathrm{dec}})}$, we compute, for $l = L_{\mathrm{dec}} - 1, \ldots, 1$,

$$\mathbf{U}_{B,\mathrm{conv}}^{(l+1\to l)} = \mathrm{TConv}_{\{3,5,7\},\uparrow}^{(l)}\left(\overline{\mathbf{B}}^{(l+1)}\right), \tag{41}$$

$$\mathbf{U}_{B,\mathrm{lin}}^{(l+1\to l)} = \mathrm{MLP}_{\uparrow}^{(l)}\left(\overline{\mathbf{B}}^{(l+1)}\right), \tag{42}$$

$$\overline{\mathbf{B}}^{(l)} = \mathbf{B}^{(l)} + \mathrm{SF}_{B}^{(l)}\left(\mathbf{U}_{B,\mathrm{conv}}^{(l+1\to l)}, \mathbf{U}_{B,\mathrm{lin}}^{(l+1\to l)}\right). \tag{43}$$

where $\mathrm{TConv}_{\{3,5,7\},\uparrow}^{(l)}$ denotes the multi-kernel transposed-convolutional path, and $\mathrm{MLP}_{\uparrow}^{(l)}$ expands the temporal length from T_{l+1} to T_l.

For either direction, the two parallel paths are combined by learnable softmax fusion:

$$\mathrm{SF}\left(\mathbf{U}_{\mathrm{conv}}, \mathbf{U}_{\mathrm{lin}}\right) = \alpha_{\mathrm{conv}}\mathbf{U}_{\mathrm{conv}} + \alpha_{\mathrm{lin}}\mathbf{U}_{\mathrm{lin}}, \qquad [\alpha_{\mathrm{conv}}, \alpha_{\mathrm{lin}}] = \mathrm{softmax}(\boldsymbol{\gamma}), \tag{44}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^2$ contains learnable path logits. This allows each transition to determine the relative contributions of local multi-receptive-field patterns and global temporal transformations. Finally, the enhanced components are recombined with the original temporal representation:

$$\widetilde{\mathbf{F}}_{T}^{(l)} = \mathbf{F}_{T}^{(l)} + \mathrm{FFN}\left(\overline{\mathbf{B}}^{(l)} + \overline{\mathbf{D}}^{(l)}\right), \qquad l = 1, \ldots, L_{\mathrm{dec}}, \tag{45}$$

where $\mathrm{FFN}(\cdot)$ denotes the feed-forward transformation within the residual update. The direct scales $l > L_{\mathrm{dec}}$ bypass this operation and preserve $\widetilde{\mathbf{F}}_{T}^{(l)} = \mathbf{F}_{T}^{(l)}$.

A.4 FFT Lag Interaction and Dynamic Gate Construction

For scale l and station n, `XGateFusion` receives the baseline and rainfall correction $\widehat{\mathbf{y}}_{T,n}^{(l)}, \widehat{\mathbf{y}}_{I,n}^{(l)} \in \mathbb{R}^H$. It constructs

$$\mathbf{u}_{T,n}^{(l)} = \widehat{\mathbf{y}}_{T,n}^{(l)}, \qquad \mathbf{u}_{I,n}^{(l)} = \widehat{\mathbf{y}}_{T,n}^{(l)} + \widehat{\mathbf{y}}_{I,n}^{(l)}, \tag{46}$$

so that the interaction is evaluated between the baseline and its rainfall-conditioned counterpart. Independent input projections first produce

$$\mathbf{H}_{T,n}^{(l)} = \phi_T\left(\mathbf{u}_{T,n}^{(l)}\right), \qquad \mathbf{H}_{I,n}^{(l)} = \phi_I\left(\mathbf{u}_{I,n}^{(l)}\right) \in \mathbb{R}^{H\times d}, \tag{47}$$

where ϕ_T and ϕ_I are learnable projections and d is the hidden dimension.

Bidirectional aperiodic lag interaction. For readability, the station index n is omitted within the lag-interaction equations. For the full relative-lag domain $\mathcal{T}_H = \{-(H-1), \ldots, H-1\}$, the two directional cross-correlation scores are

$$\mathbf{C}_{T\leftarrow I}^{(l)}(\tau) = \sum_{t} \mathbf{Q}_{T}^{(l)}(t) \odot \mathbf{K}_{I}^{(l)}(t-\tau), \tag{48}$$

$$\mathbf{C}_{I\leftarrow T}^{(l)}(\tau) = \sum_{t} \mathbf{Q}_{I}^{(l)}(t) \odot \mathbf{K}_{T}^{(l)}(t-\tau), \qquad \tau \in \mathcal{T}_H, \tag{49}$$

where $\mathbf{Q}$ and $\mathbf{K}$ denote query and key projections, and positions outside the forecast horizon are zero padded. Using an FFT length of at least $2H-1$ implements aperiodic linear correlation and avoids circular wraparound. The correlation scores are converted into bounded signed lag weights:

$$\mathbf{A}_{a\leftarrow b}^{(l)}(\tau) = \mathrm{Dropout}\left[d_h^{-1/2}\tanh\left(\mathbf{C}_{a\leftarrow b}^{(l)}(\tau)\right)\right], \tag{50}$$

where $(a,b) \in \{(T,I),(I,T)\}$ and d_h is the dimension of each attention head. The corresponding values are aggregated through linear lag convolution:

$$\mathbf{O}_{a\leftarrow b}^{(l)}(t) = \sum_{\tau\in\mathcal{T}_H} \mathbf{A}_{a\leftarrow b}^{(l)}(\tau) \odot \mathbf{V}_{b}^{(l)}(t-\tau), \tag{51}$$

where $\mathbf{V}_{b}^{(l)}$ denotes the value projection of branch b. Both correlation and aggregation are evaluated by FFT with temporal complexity $\mathcal{O}(H\log H)$ Wu et al. (2021). The two directions are combined symmetrically:

$$\mathbf{H}_{X,n}^{(l)} = \frac{1}{2}\left[\psi\left(\mathbf{O}_{T\leftarrow I,n}^{(l)}\right) + \psi\left(\mathbf{O}_{I\leftarrow T,n}^{(l)}\right)\right], \tag{52}$$

where ψ is the shared output projection. The relative lags describe learned statistical offsets between the two prediction sequences and should not be interpreted as causal rainfall-response delays.

Dynamic gate construction. Each scale has a learned scalar base gate $g^{(l)}$, which represents its overall reliance on the rainfall correction. The interaction representation further produces a bounded station and horizon-dependent update:

$$\Delta \mathbf{g}_n^{(l)} = \delta_{\max} \tanh\Big[W_o \,\mathrm{LN}\left(\mathbf{H}_{X,n}^{(l)}\right)\Big] \in \mathbb{R}^H, \tag{53}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization, W_o is a learnable output projection, and $\delta_{\max}$ bounds the update magnitude. Collecting all station-wise updates gives $\Delta \mathbf{G}^{(l)} = [\Delta \mathbf{g}_1^{(l)}, \ldots, \Delta \mathbf{g}_N^{(l)}] \in \mathbb{R}^{H \times N}$. The scale-wise fused prediction is therefore

$$\widehat{\mathbf{Y}}_{TI}^{(l)} = \widehat{\mathbf{Y}}_{T}^{(l)} + \left(g^{(l)} \mathbf{1}_{H \times N} + \Delta \mathbf{G}^{(l)}\right) \odot \widehat{\mathbf{Y}}_{I}^{(l)}. \tag{54}$$

The projection producing $\Delta \mathbf{G}^{(l)}$ is initialized to zero, making the initial output $\widehat{\mathbf{Y}}_{T}^{(l)} + g^{(l)} \widehat{\mathbf{Y}}_{I}^{(l)}$. The `XGateFusion` modules follow the same architecture across scales but use independent parameters.

A.5 Joint Predictor and Directed Topology

The scale-wise temporal-rainfall pathway focuses on station-aligned rainfall corrections. P_C incorporates basin-level rainfall conditions and directed inter-station information.

Basin-level rainfall context. For each historical time step t, we summarize the valid rainfall-grid cells using four statistics:

$$\mathbf{r}_t^{\mathrm{basin}} = \left[\sum_p v_{t,p} I_{t,p},\ \max_{p: v_{t,p}=1} I_{t,p},\ \frac{\sum_p v_{t,p} I_{t,p}}{\sum_p v_{t,p}},\ \sum_p v_{t,p} \mathbb{I}(I_{t,p} > 0)\right], \tag{55}$$

corresponding to total rainfall, maximum intensity, mean intensity, and the number of wet grid cells. We aggregate these statistics over recent windows $\mathcal{W}_C = \{1, 4, 12, 24, 48, 96\}$ and retain both their raw and logarithmic values:

$$\mathbf{c}_I = \Big\|_{w \in \mathcal{W}_C} \left[\sum_{t=T-w+1}^{T} \mathbf{r}_t^{\mathrm{basin}},\ \log\left(1 + \sum_{t=T-w+1}^{T} \mathbf{r}_t^{\mathrm{basin}}\right)\right]. \tag{56}$$

The window lengths are adjusted according to the sampling interval, so that they represent the same temporal durations across datasets.

Directed topological features. The river topology is constructed from directed upstream-downstream station pairs and their along-river distances. After removing redundant transitive edges when the graph is acyclic, an upstream-to-downstream adjacency matrix $\mathbf{A}$ is defined as

$$A_{j,i} = \begin{cases} \exp(-d_{i,j}/d_{\mathrm{med}}), & i \to j, \\ 0, & \text{otherwise}, \end{cases} \tag{57}$$

where $d_{i,j}$ is the river distance and d_{med} is the median distance of the retained edges. We construct row-normalized upstream and downstream propagation operators

$$\mathbf{P}^{\rightarrow} = \mathrm{RowNorm}(\mathbf{I} + \beta \mathbf{A}), \qquad \mathbf{P}^{\leftarrow} = \mathrm{RowNorm}(\mathbf{I} + \beta \mathbf{A}^{\top}), \tag{58}$$

where β controls the topology strength. For propagation order k, the topological differences are

$$\mathbf{D}^{\rightarrow,k} = (\mathbf{P}^{\rightarrow})^k - \mathbf{I}, \qquad \mathbf{D}^{\leftarrow,k} = (\mathbf{P}^{\leftarrow})^k - \mathbf{I}. \tag{59}$$

Applying these operators to the historical water-quality sequence produces the station-specific topology feature

$$\mathbf{g}_n = \Big\|_{k=1}^{K_G} \left[(\mathbf{D}^{\rightarrow,k} \mathbf{X})_{:,n},\ (\mathbf{D}^{\leftarrow,k} \mathbf{X})_{:,n}\right]. \tag{60}$$

We use $K_G = 1$ and $\beta = 2$ in the default configuration. The reverse operator supplies downstream-conditioned predictive context and does not imply reverse physical transport.

Station-wise joint prediction. For station n, the Joint Predictor receives its own historical observations, the shared basin-level rainfall context, and its corresponding topology features:

$$\mathbf{q}_n = [\mathbf{X}_{:,n} \parallel \mathbf{c}_I \parallel \mathbf{g}_n]\,. \tag{61}$$

Each feature is standardized using statistics fitted exclusively on the training split. An independent linear prediction head is used for each station:

$$\widehat{\mathbf{y}}_{C,n} = \mathbf{W}_{C,n}\widetilde{\mathbf{q}}_n + \mathbf{b}_{C,n} \in \mathbb{R}^H, \tag{62}$$

where $\widetilde{\mathbf{q}}_n$ denotes the standardized joint feature vector. Collecting all station-wise outputs gives $\widehat{\mathbf{Y}}_C \in \mathbb{R}^{H\times N}$. Finally, $\widehat{\mathbf{Y}}_C$ is combined with the scale-aggregated temporal-rainfall prediction $\widehat{\mathbf{Y}}_{TI}$ through a learnable station and horizon-specific gate.

A.6 Training Objective

RaiNet is trained using a combination of MSE and MAE. For a prediction $\widehat{\mathbf{Y}}$ and ground truth $\mathbf{Y}$, we define

$$\mathcal{L}_{\text{pred}}(\widehat{\mathbf{Y}}, \mathbf{Y}) = \text{MSE}(\widehat{\mathbf{Y}}, \mathbf{Y}) + \lambda_{\text{MAE}}\,\text{MAE}(\widehat{\mathbf{Y}}, \mathbf{Y}). \tag{63}$$

In addition to supervising the final prediction, we directly supervise the rainfall-conditioned multi-scale prediction $\widehat{\mathbf{Y}}_{TI}$ before the Joint Predictor. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}}(\widehat{\mathbf{Y}}, \mathbf{Y}) + \lambda_{TI}\mathcal{L}_{\text{pred}}(\widehat{\mathbf{Y}}_{TI}, \mathbf{Y}), \tag{64}$$

where the auxiliary term prevents the temporal-rainfall pathway from being trained only through the final joint prediction. We set $\lambda_{\text{MAE}} = 1$ and $\lambda_{TI} = 1$ in all experiments.

A.7 Dataset Details

Data construction. Three multimodal datasets are derived from USGS monitoring stations in basins 170900, 071200, and 051201[1]. Each dataset comprises monitoring stations within a shared drainage basin that are connected through the river network and provide concurrent records over the study period. The selected networks contain 15, 5, and 5 stations, respectively. For each station, missing values are forward-filled using the most recent earlier observation. Backward filling is applied only to leading missing values at the beginning of the full record.

Table 3 summarizes the resulting datasets. We study turbidity and nitrate because they represent complementary rainfall-related processes: turbidity reflects rapid sediment mobilization and runoff, whereas nitrate captures nutrient transport and leaching over broader timescales. Forecasting turbidity provides early indication of runoff-driven particle loads that can challenge water-treatment operations, while nitrate forecasting supports nutrient-pollution surveillance, source-water management, and eutrophication risk assessment. Basin 170900 provides five years of hourly turbidity observations, while basins 071200 and 051201 provide three and five years of nitrate observations, respectively. The 15-minute records of 051201 are aggregated to 30-minute intervals to match the native temporal resolution of the adopted IMERG product and avoid repeatedly pairing multiple water-quality observations with the same precipitation grid. Together, the datasets comprise 157,824 aligned multimodal time steps.

Table 3: Statistics of the three multimodal water-quality datasets.

Basin	Target	Period	Interval	Stations	Paired steps	Raster size
170900	Turbidity	2020/10–2025/09	1 h	15	43,824	7×18
071200	Nitrate	2022/04–2025/03	1 h	5	26,304	8×31
051201	Nitrate	2019/10–2024/09	30 min	5	87,696	37×9

[1] https://waterdata.usgs.gov/nwis/

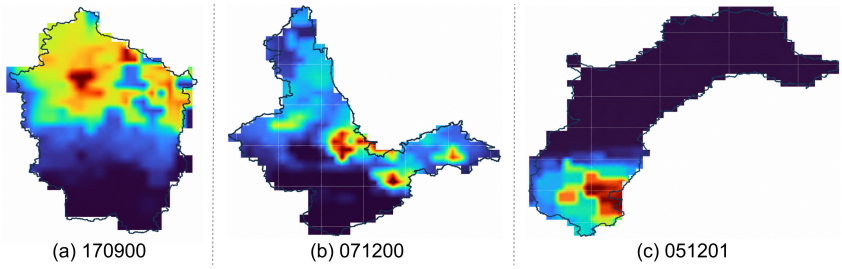


Figure 10: Example of gridded IMERG precipitation.

Table 4: Twelve-hour nitrate observations from five stations in basin 071200 ($\mathrm{mg\,L^{-1}}$).

Time (UTC)	**05515500**	**05518000**	**05536356**	**05537980**	**05543010**
Latitude	41.390°N	41.183°N	41.619°N	41.536°N	41.300°N
Longitude	86.706°W	87.340°W	87.500°W	88.084°W	88.614°W
2022/04/01 00:00	1.720	2.135	5.400	4.500	4.200
2022/04/01 01:00	1.750	2.135	5.400	4.350	4.200
2022/04/01 02:00	1.765	2.120	5.400	4.300	4.200
2022/04/01 03:00	1.795	2.120	5.400	4.300	4.200
2022/04/01 04:00	1.840	2.125	5.350	4.200	4.200
2022/04/01 05:00	1.900	2.135	5.350	4.150	4.400
2022/04/01 06:00	1.955	2.140	5.600	4.250	4.400
2022/04/01 07:00	1.995	2.135	5.700	4.350	4.400
2022/04/01 08:00	2.015	2.120	5.800	4.400	4.400
2022/04/01 09:00	2.065	2.130	5.800	4.400	4.500
2022/04/01 10:00	2.095	2.130	5.800	4.300	4.500
2022/04/01 11:00	2.100	2.130	5.900	3.900	4.500

Gridded precipitation. Precipitation fields are obtained from NASA GPM IMERG through Google Earth Engine.[2] Each single-band GeoTIFF records precipitation intensity in $\mathrm{mm\,h^{-1}}$ on a $0.1° \times 0.1°$ grid. The raster extent is derived from the monitoring-station boundary with additional spatial padding, preserving both station-local rainfall and its surrounding basin context. The resulting spatial coverages are 123.6-121.8°W and 45.0-45.7°N for 170900, 89.2-86.1°W and 41.0-41.8°N for 071200, and 88.2-87.4°W and 37.4-41.0°N for 051201. Figure 10 presents an example of the gridded precipitation input.

Multimodal alignment. Water-quality observations and precipitation grids are matched by UTC timestamp. Hourly water-quality records are paired with hourly precipitation grids, while basin 051201 is evaluated at the native 30-minute IMERG interval. This produces a one-to-one correspondence between the two modalities. Table 4 provides 12 consecutive hourly nitrate observations from the five stations in basin 071200 as an example.

A.8 Hyperparameter Sensitivity

We evaluate the sensitivity of RaiNet to four parameters associated with its main designs: the number of temporal scales L, rainfall-event persistence setting h, LocTrend window configuration (w_l, s_l), and dictionary size K. We vary one parameter at a time while keeping the remaining settings at their default values. For (w_l, s_l), the stride is set to half of the window length to maintain a consistent overlap ratio. The same default configuration is used for all three datasets without dataset-specific tuning. The number of scales L controls the temporal support available to RaiNet, whereas h determines how consecutive station-relative rainfall states are consolidated into persistent events. Table 5 evaluates $L \in \{2, 3, 4, 5\}$ and $h \in \{3, 6, 9, 12\}$. These experiments examine the balance

[2] https://developers.google.com/earth-engine/datasets/catalog/NASA_GPM_L3_IMERG_V07

between multiscale coverage and temporal detail, as well as between suppressing isolated rainfall fluctuations and retaining event timing.

Table 5: Sensitivity to the number of temporal scales L and rainfall-event persistence setting h. Default settings are shown in bold in the column headers.

Dataset	$L = 2$		$L = 3$		$\mathbf{L = 4}$		$L = 5$	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
170900	0.4337	0.2573	0.4711	0.2971	0.4067	0.2578	0.4144	0.2575
071200	0.3255	0.3373	0.3279	0.3347	0.3238	0.3338	0.3275	0.3347
051201	0.0439	0.0901	0.0417	0.0857	0.0422	0.0868	0.0412	0.0868
Dataset	$h = 3$		$\mathbf{h = 6}$		$h = 9$		$h = 12$	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
170900	0.4129	0.2638	0.4066	0.2572	0.4106	0.2561	0.4240	0.2675
071200	0.3274	0.3355	0.3227	0.3330	0.3159	0.3310	0.3296	0.3357
051201	0.0416	0.0863	0.0417	0.0866	0.0412	0.0867	0.0422	0.0869

The results show that a moderate number of temporal scales is sufficient for capturing multiscale water-quality dynamics. In particular, $L = 4$ achieves the lowest MSE on 170900 and both the lowest MSE and MAE on 071200, while remaining close to the best results on 051201. Increasing the number of scales further does not consistently improve performance, suggesting that excessive downsampling provides limited additional benefit once sufficiently broad temporal supports have been covered. For rainfall-event persistence, moderate values generally perform better than an overly long aggregation interval. The strongest results are mainly observed around $h = 6$-9, whereas $h = 12$ tends to degrade performance. This behavior is consistent with the role of persistence aggregation: short aggregation intervals retain transient fluctuations, while excessively long intervals may merge temporally distinct rainfall events and blur their timing. We therefore retain $h = 6$ as a common short-range persistence setting across datasets rather than tuning the event duration separately for each dataset. We set $L_{\mathrm{dir}} = 1$, so the last coarsest scale follows the direct temporal path. As shown in Fig. 7, the structural background becomes nearly flat by the third decomposed scale, indicating that further decomposition after additional downsampling provides limited structural benefit. We preserve the fourth and coarsest scale without further decomposition.

Table 6: Sensitivity to the LocTrend window configuration (w_l, s_l) and shared dictionary size K. Default settings are shown in bold in the column headers.

Dataset	$(12, 6)$		$(18, 9)$		$(\mathbf{24}, \mathbf{12})$		$(30, 15)$		$(36, 18)$	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
170900	0.4060	0.2542	0.4213	0.2680	0.4092	0.2599	0.4149	0.2653	0.4590	0.2940
071200	0.3227	0.3331	0.3190	0.3334	0.3219	0.3332	0.3223	0.3318	0.3269	0.3340
051201	0.0406	0.0855	0.0422	0.0869	0.0411	0.0857	0.0420	0.0861	0.0417	0.0861
Dataset	$K = 1$		$K = 2$		$\mathbf{K = 4}$		$K = 6$		$K = 8$	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
170900	0.4025	0.2567	0.4008	0.2541	0.4072	0.2609	0.4076	0.2646	0.4051	0.2606
071200	0.3232	0.3336	0.3235	0.3332	0.3243	0.3343	0.3242	0.3340	0.3219	0.3345
051201	0.0416	0.0869	0.0414	0.0868	0.0422	0.0869	0.0415	0.0860	0.0410	0.0855

Table 6 shows that LocTrend remains reasonably stable over a range of moderate window configurations. Although the exact optimum varies across datasets, $(24, 12)$ remains close to the best-performing settings in all three cases. In contrast, excessively long windows can reduce accuracy, as most clearly observed for $(36, 18)$ on 170900. This suggests that overly broad temporal neighborhoods may weaken the local adaptivity required for background reconstruction. We therefore use $(24, 12)$ as a common moderate-size configuration across datasets. The effect of dictionary size is similarly non-monotonic. Increasing K does not consistently improve either MSE or MAE, and different datasets favor different dictionary sizes. This indicates that LocTrend does not rely on a large structural dictionary to obtain accurate forecasts; a small number of shared atoms is already

sufficient to represent the recurring structures needed for background reconstruction. We therefore use $K = 4$ as a fixed compact dictionary size throughout the experiments.

The sensitivity results show that RaiNet does not depend on a single dataset-specific hyperparameter optimum. The adopted configuration, $L = 4$, $h = 6$, $(w_l, s_l) = (24, 12)$, and $K = 4$, provides a consistent operating point across the three datasets and is kept unchanged throughout the main experiments.

A.9 Comparative Models

We compare RaiNet with four groups of representative forecasting models.

Time-series models. TimeKAN (Huang et al., 2025) models multiscale temporal patterns through frequency decomposition and Kolmogorov–Arnold Networks. FilterTS (Wang et al., 2025b) learns adaptive frequency-domain filters. GTR (Cao et al., 2026) retrieves informative global patterns beyond the local input window, while MixLinear (Ma et al., 2026) combines temporal and frequency-domain linear modeling.

Spatio-temporal models. HyperD (Shao et al., 2026) decouples periodic and residual dynamics for spatio-temporal forecasting. RSTIB-MLP (Chen et al., 2025) introduces an information-bottleneck objective for robust spatio-temporal representation learning. ST-SSDL (Gao et al., 2026) explicitly models deviations from historical patterns through self-supervised learning.

Water-quality-specific models. PIWON (Tang et al., 2026) models pollutant accumulation and wash-off processes. D-TGCN (Zhang et al., 2026) models directed dependencies among monitoring stations according to river-network topology. CMLIP (Bi et al., 2025a) jointly models water-quality observations and rainfall information. RGCN (Smith et al., 2024) combines graph and recurrent modeling for stream-quality prediction, and XGBR (Ali et al., 2025) is a tree-based water-quality prediction model.

Diffusion-based models. SimDiff (Ding et al., 2026) formulates forecasting as an end-to-end diffusion process. MA-TSD (Wang et al., 2025a) introduces moving-average-guided diffusion transitions for nonstationary temporal dynamics.

All methods are evaluated under the same chronological splits, forecasting horizons, and multi-station prediction protocol. Input modalities follow the modeling scope of each method: water-quality-specific methods incorporate aligned historical rainfall and other available environmental covariates where applicable, whereas general temporal and diffusion models operate on water-quality histories. Graph-based methods additionally use river-network topology when required by their architectures. All inputs are restricted to observations available at the forecast origin.

A.10 Full Comparative Result and Computational Cost

Tables 7 and 9 report the complete results over five forecasting horizons. RaiNet achieves the lowest MSE in all 15 settings and the lowest MAE in 14 of them. On 170900, it reduces MSE by 5.7-13.6% relative to the strongest baseline across horizons. The sole exception in MAE occurs at 192 steps, where RSTIB-MLP obtains 0.3061 compared with RaiNet's 0.3102. On 071200, RaiNet maintains clear advantages at every horizon, including a 21.6% MSE reduction at 96 steps. On the finer-resolution 051201 dataset, it consistently improves over strong time-series, water-quality-specific, and diffusion-based models, despite the smaller absolute errors and consequently narrower margins.

The spatiotemporal results reveal an important mismatch between generic cross-station modeling and basin-level water-quality forecasting. Monitoring stations are sparsely distributed, while their observations are affected by heterogeneous local environments and asynchronous responses. Consequently, directly propagating station features through a predefined spatial structure may introduce variance without providing consistently useful predictive information. This is reflected by the unstable performance of HyperD, RSTIB-MLP, and ST-SSDL: they remain competitive in a few 170900 settings, but fall behind on 071200, while HyperD and RSTIB-MLP deteriorate substantially on 051201. Thus, their limitations arise not from spatial modeling itself, but from treating cross-station dependence as a uniformly reliable forecasting signal.

Table 7: Forecasting results with representative time-series and spatio-temporal baselines. The best and second-best results are highlighted in **bold** and underline, respectively.

		Time-series Models								Spatio-temporal Models						Ours	
Model		TimeKAN		FilterTS		GTR		MixLinear		HyperD		RSTIB-MLP		ST-SSDL		RaiNet	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
170900	24	0.2628	0.1667	0.2692	0.1722	0.2763	0.1726	0.3790	0.2324	0.3786	0.1909	0.2960	0.1871	0.2792	0.2107	**0.2185**	**0.1560**
	48	0.3687	0.2137	0.3697	0.2131	0.3757	0.2150	0.4640	0.2704	0.4554	0.2290	0.3958	0.2227	0.3739	0.2431	**0.3162**	**0.2001**
	96	0.4778	0.2623	0.4771	0.2626	0.4834	0.2648	0.5739	0.3208	0.5518	0.2729	0.4994	0.2676	0.5410	0.3177	**0.4096**	**0.2604**
	144	0.5494	0.2998	0.5433	0.2946	0.5458	0.2963	0.6449	0.3530	0.6081	0.3009	0.5538	0.2895	0.6011	0.3261	**0.4904**	**0.2840**
	192	0.6053	0.3204	0.5974	0.3199	0.5975	0.3207	0.7047	0.3786	0.6442	0.3183	0.5937	**0.3061**	0.6603	0.3691	**0.5283**	0.3102
071200	24	0.1084	0.1905	0.1105	0.1885	0.1162	0.1967	0.2458	0.2848	0.6530	0.4605	0.1630	0.2555	0.1516	0.2535	**0.0872**	**0.1747**
	48	0.2425	0.2805	0.2421	0.2723	0.2529	0.2851	0.3506	0.3429	0.6092	0.4694	0.2822	0.3126	0.5788	0.5108	**0.1707**	**0.2414**
	96	0.4412	0.3802	0.4473	0.3784	0.4591	0.3894	0.5818	0.4424	0.7663	0.5291	0.4613	0.3912	0.7145	0.5240	**0.3225**	**0.3344**
	144	0.5820	0.4450	0.5952	0.4480	0.6019	0.4572	0.7111	0.5014	0.8402	0.5601	0.5813	0.4499	1.1075	0.7135	**0.4504**	**0.4010**
	192	0.6771	0.4918	0.6879	0.4944	0.7010	0.5041	0.8131	0.5463	0.8453	0.5759	0.6597	0.4930	1.0558	0.7019	**0.5415**	**0.4443**
051201	24	0.0074	0.0328	0.0075	0.0337	0.0081	0.0353	0.0194	0.0599	1.3482	0.7703	0.7583	0.6066	0.0217	0.0821	**0.0072**	**0.0318**
	48	0.0182	0.0530	0.0185	0.0543	0.0196	0.0564	0.0288	0.0734	1.3237	0.7886	0.7933	0.6167	0.0319	0.1066	**0.0166**	**0.0518**
	96	0.0462	0.0889	0.0468	0.0903	0.0491	0.0938	0.0609	0.1089	1.3572	0.8016	0.7660	0.6449	0.0756	0.1626	**0.0414**	**0.0867**
	144	0.0785	0.1214	0.0789	0.1230	0.0807	0.1242	0.0963	0.1415	1.3682	0.8210	0.8104	0.6486	0.1314	0.2041	**0.0680**	**0.1171**
	192	0.1129	0.1524	0.1113	0.1505	0.1151	0.1536	0.1378	0.1763	1.3842	0.8280	0.9683	0.7014	0.1491	0.2120	**0.0992**	**0.1455**

RaiNet avoids this failure mode by modeling water-quality dynamics station-wise and introducing spatial information asymmetrically through station-related rainfall events and basin context. Its consistent performance across different targets, temporal resolutions, station counts, and basin configurations demonstrates that conditional environmental fusion generalizes more reliably than direct cross-station propagation for these datasets. Table 9 further shows that this advantage extends beyond generic forecasting models: RaiNet also consistently outperforms water-quality-specific and diffusion-based alternatives, indicating that neither specialized spatial structures nor generative forecasting alone adequately resolves heterogeneous rainfall and water-quality dynamics.

Table 8 reports the mean and sample standard deviation over five random seeds. RaiNet exhibits consistently small run-to-run variation across all datasets and forecasting horizons. The largest relative standard deviations are approximately 5.3% for MSE and 4.0% for MAE, while most settings show substantially lower variation. This low dispersion across different targets, temporal resolutions, and prediction lengths demonstrates the stable and robust optimization behavior of RaiNet under different random initializations. Therefore, RaiNet is stable across random initializations and that its overall performance is not determined by a particular random seed.

Table 8: Run-to-run stability of RaiNet over five random seeds. Results are reported as mean $\pm$ sample standard deviation.

Dataset	Horizon	MSE	MAE
170900	24	0.2157 ± 0.0026	0.1550 ± 0.0011
	48	0.3108 ± 0.0046	0.1995 ± 0.0013
	96	0.4371 ± 0.0232	0.2633 ± 0.0106
	144	0.4988 ± 0.0119	0.2995 ± 0.0089
	192	0.5291 ± 0.0046	0.3098 ± 0.0047
071200	24	0.0872 ± 0.0008	0.1741 ± 0.0008
	48	0.1710 ± 0.0007	0.2423 ± 0.0013
	96	0.3256 ± 0.0050	0.3337 ± 0.0041
	144	0.4521 ± 0.0059	0.3989 ± 0.0029
	192	0.5469 ± 0.0118	0.4451 ± 0.0057
051201	24	0.0071 ± 0.0001	0.0314 ± 0.0003
	48	0.0168 ± 0.0004	0.0517 ± 0.0003
	96	0.0418 ± 0.0007	0.0875 ± 0.0011
	144	0.0684 ± 0.0013	0.1167 ± 0.0016
	192	0.1000 ± 0.0016	0.1473 ± 0.0022

Figure 11 compares forecasting accuracy, computational cost, and peak GPU memory under a unified setting. All models are evaluated on basin 170900 using an input length of 96, a prediction length of 96, and a batch size of 64 on a single NVIDIA GeForce RTX 5080. Training FLOPs are measured for one complete forward-and-backward iteration, while memory denotes the peak GPU memory allocated during execution. Despite incorporating an additional gridded-rainfall modality, RaiNet requires only 19.03 GFLOPs per training iteration and 1.32 GB of peak allocated memory. Its computational cost is 55.2% lower than HyperD, 77.1% lower than CMLIP, 91.4% lower than TimesNet, and 96.4% lower than ST-SSDL, while remaining slightly below RSTIB-MLP. Although several lightweight temporal baselines require fewer operations, they produce consistently higher forecasting errors. RaiNet therefore occupies a favorable region of the accuracy–efficiency space, achieving the lowest MSE without the excessive computation required by several complex forecasting models. The modest absolute memory requirement also leaves substantial capacity for larger batches or concurrent forecasting tasks on a single RTX 5080. Combined with its station-wise prediction structure and efficient conditional integration of rainfall information, these results indicate that RaiNet is practical for scheduled basin-scale forecasting services without requiring multi-GPU infrastructure.

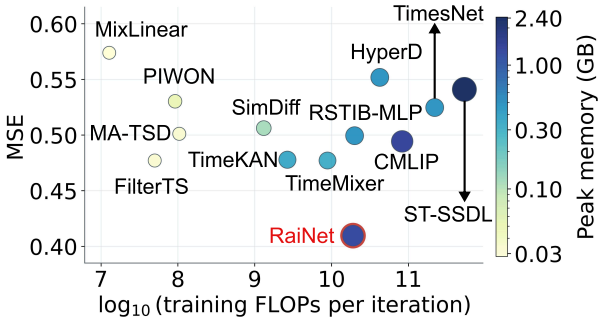


Figure 11: Accuracy-efficiency trade-offs on 170900 with an input and prediction length of 96. Bubble size and color denote peak allocated GPU memory.

Table 9: Forecasting results with water-quality-specific and diffusion-based models. The best and second-best results are highlighted in **bold** and underline, respectively.

		Water-quality-specific Models										Diffusion-based Models				Ours	
Model		PIWON		D-TGCN		CMLIP		RGCN		XGBR		SimDiff		MA-TSD		RaiNet	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
170900	24	0.2523	0.2013	0.2719	0.2111	0.2796	0.1794	0.2976	0.2113	0.3924	0.1913	0.2773	0.1641	0.2843	0.2023	**0.2185**	**0.1560**
	48	0.3960	0.2634	0.3658	0.2568	0.3875	0.2247	0.3897	0.2617	0.5593	0.2467	0.3982	0.2142	0.3877	0.2582	**0.3162**	**0.2001**
	96	0.5303	0.3326	0.4633	0.3072	0.4943	0.2723	0.4795	0.3179	0.8964	0.3228	0.5062	0.2643	0.5011	0.3314	**0.4096**	**0.2604**
	144	0.6787	0.4255	0.5201	0.3449	0.5500	0.2992	0.5312	0.3496	1.2130	0.3769	0.5781	0.2993	0.5739	0.3750	**0.4904**	**0.2840**
	192	0.9224	0.5161	0.5618	0.3661	0.6139	0.3338	0.5669	0.3673	1.4016	0.4131	0.6387	0.3250	0.6263	0.4098	**0.5283**	**0.3102**
071200	24	0.0961	0.1936	0.1166	0.2043	0.1272	0.2011	0.1784	0.2460	0.1227	0.1999	0.1387	0.2068	0.1280	0.2168	**0.0872**	**0.1747**
	48	0.1868	0.2676	0.2278	0.2864	0.2631	0.2840	0.2855	0.3049	0.2514	0.2810	0.2544	0.2765	0.2643	0.3105	**0.1707**	**0.2414**
	96	0.4115	0.3869	0.4347	0.4086	0.4734	0.3919	0.4476	0.3912	0.4779	0.3951	0.4708	0.3911	0.4824	0.4257	**0.3225**	**0.3344**
	144	0.4967	0.4261	0.5533	0.4683	0.6243	0.4640	0.5665	0.4565	0.6684	0.4805	0.6278	0.4680	0.6267	0.4962	**0.4504**	**0.4010**
	192	0.5877	0.4698	0.6435	0.5066	0.7240	0.5107	0.6450	0.4971	0.8040	0.5385	0.7305	0.5194	0.7246	0.5501	**0.5415**	**0.4443**
051201	24	0.0090	0.0461	0.0093	0.0448	0.0081	0.0343	0.0980	0.1883	0.0121	0.0551	0.0098	0.0401	0.0081	0.0367	**0.0072**	**0.0318**
	48	0.0235	0.0797	0.0217	0.0698	0.0200	0.0561	0.1193	0.2011	0.0251	0.0754	0.0197	0.0553	0.0197	0.0609	**0.0166**	**0.0518**
	96	0.0555	0.1233	0.0569	0.1180	0.0516	0.0951	0.1394	0.2101	0.0565	0.1102	0.0530	0.0930	0.0482	0.1009	**0.0414**	**0.0867**
	144	0.1031	0.1763	0.0911	0.1549	0.0881	0.1287	0.1718	0.2461	0.0913	0.1413	0.0930	0.1335	0.0797	0.1287	**0.0680**	**0.1171**
	192	0.1316	0.1981	0.1241	0.1884	0.1206	0.1558	0.2055	0.2689	0.1272	0.1702	0.1260	0.1593	0.1130	0.1642	**0.0992**	**0.1455**

A.11 DETAILED ABLATION STUDIES

Cumulative construction. Figure 6 summarizes the cumulative construction and standalone variants across the three basins. Because turbidity and nitrate have different error scales, relative changes are computed within each basin before averaging. Starting from the temporal-only moving-average baseline, we successively add the rainfall-event branch with RG, multiscale modeling, `LocTrend`, `XGateFusion`, the Joint Predictor P_C, and the directed topology $\mathcal{G}$. The complete model reduces average MSE and MAE by 18.5% and 4.4%, respectively. In particular, the final MSE gain is retained in all three basins, indicating that the components jointly improve the prediction of sub-

stantial water-quality deviations. The construction is physically motivated. Multiscale modeling retains rainfall and water-quality evidence over different temporal supports, whereas `LocTrend` replaces fixed smoothing with a locally evolving structural background and its dynamic deviation. The rainfall-event branch then provides station-related event evidence, and `XGateFusion` uses it as a conditional correction rather than applying rainfall uniformly. P_C further integrates the station-wise predictions, while $\mathcal{G}$ adds directed distance and orientation context among monitoring sites.

Standalone architectural variants. The standalone variants separate these effects. Setting $L = 1$, replacing `LocTrend` with MA, or replacing `XGateFusion` with RG each increases average MSE, supporting the roles of multiscale temporal structure, adaptive background-deviation decomposition, and conditional fusion. Removing $\mathcal{E}_I$ causes the largest degradation (21.1% MSE and 2.8% MAE), showing that station-related rainfall events provide information not contained in water-quality history alone. Removing P_C also yields a clear penalty, while replacing $\mathcal{G}$ by identity context produces a smaller but complementary loss. Thus, the model benefits from using rainfall as basin-aware environmental context without treating it as an unconditional peer of the temporal baseline. The remaining variants diagnose the fusion mechanism. Removing either the base gate $g^{(l)}$ or its dynamic update $\Delta\mathbf{G}^{(l)}$ increases average MSE, indicating that stable scale-level rainfall weighting and station and horizon-dependent adjustment serve distinct roles. Removing $(\mathbf{B}^{(l)}, \mathbf{D}^{(l)})$ degrades MSE, showing that subsequent modeling benefits from retaining locally reconstructed background and dynamic deviation separately.

MAE changes are smaller and occasionally basin-dependent. This is consistent with the physical heterogeneity of the task: MSE emphasizes comparatively rare rainfall-associated excursions, such as rapid runoff-driven turbidity pulses or pronounced nitrate transport events, whereas MAE is dominated by routine observations. Since the frequency and attenuation of such events differ across basins, a component that reduces large event errors can slightly rebalance small baseline errors in an individual basin. The accompanying MAE results complement the consistent MSE evidence for event-sensitive conditional forecasting.

Effect of rainfall-event discretization. Table 10 compares the station-oriented rainfall-event encoding with a continuous station-to-basin rainfall contrast. The continuous variant removes only the thresholding and sign encoding, while retaining the same event aggregation, multiscale alignment, model architecture, and training protocol. Positive values denote error reductions from discretization. Discretization trades fine-grained contrast magnitude for a more stable directional event cue: weak but genuine local deviations may be attenuated, whereas large amplitude fluctuations cannot disproportionately dominate event conditioning. Empirically, discretization improves MSE on 170900 and 051201, with the largest reduction of 6.19% on 170900, and improves both metrics on 051201. Although its MAE effect is basin-dependent and 071200 shows small regressions, the MSE gains indicate that the stabilizing effect is beneficial for the rainfall-informed forecasting tasks.

Table 10: Effect of rainfall-event discretization. The continuous variant replaces event states with continuous station-to-basin rainfall contrast while retaining the same aggregation, architecture, and training protocol. Positive changes indicate lower error with discretization.

Dataset	Continuous contrast MAE / MSE	Discrete event MAE / MSE	Relative change MAE / MSE
170900	0.254449 / 0.436682	0.261818 / 0.409673	-2.896% / $+6.185\%$
071200	0.330376 / 0.323141	0.333473 / 0.323517	-0.937% / -0.116%
051201	0.086932 / 0.042277	0.086103 / 0.041852	$+0.954\%$ / $+1.004\%$

A.12 Qualitative Analysis

We further examine whether `LocTrend` realizes the intended background-deviation decomposition while remaining adaptive to temporal resolution and local dynamics. Figure 12(a) first visualizes the decomposition at the 1-, 2-, and 4-hour scales. At each scale, `LocTrend` satisfies $\mathbf{F}_T^{(l)} = \mathbf{B}^{(l)} + \mathbf{D}^{(l)}$, where $\mathbf{B}^{(l)}$ captures the locally evolving structural background and $\mathbf{D}^{(l)}$ is its exact complementary deviation. The resulting background becomes progressively smoother as the temporal resolution becomes coarser, while the deviation retains the remaining local variations.

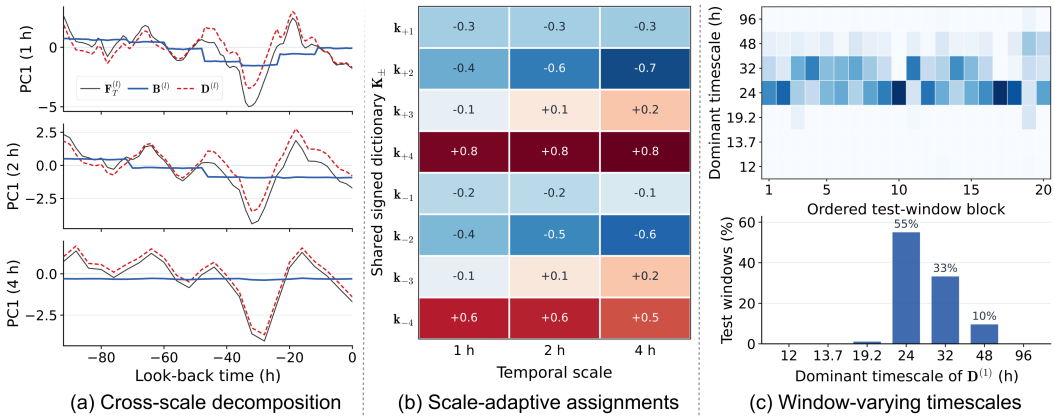


Figure 12: **Analysis of `LocTrend`.** (a) At different temporal scales, `LocTrend` separates the input $\mathbf{F}_T^{(l)}$ into a locally evolving structural background $\mathbf{B}^{(l)}$ and an exact dynamic deviation $\mathbf{D}^{(l)}$. (b) The three scales exhibit different soft-assignment patterns over the shared signed dictionary $\mathbf{K}_\pm$. (c) Post-hoc spectral analysis shows window-varying dominant timescales in $\mathbf{D}^{(1)}$. The heatmap and bars show their block-wise and overall distributions.

This scale-dependent behavior indicates that `LocTrend` does not impose an identical decomposition pattern across resolutions.

Figure 12(b) further reveals the mechanism underlying this adaptation. All decomposed scales share the same signed structural dictionary $\mathbf{K}_\pm$, but their soft-assignment patterns differ substantially. In particular, both the utilization of different atom pairs and the preference between their positive and negative directions vary across scales. Since the reconstructed background is determined by $\widetilde{\mathbf{b}}_{i,t}^{(l)} = \sum_{r=1}^{2K} a_{i,t,r}^{(l)} \mathbf{k}_{\pm,r}$, the scale dependence therefore arises from the assignment coefficients rather than from introducing separate dictionaries for different resolutions. This shared-dictionary yet scale-adaptive reconstruction allows common structural patterns to be reused while their contributions are adjusted according to the temporal scale.

Finally, Figure 12(c) examines whether the resulting decomposition is tied to a fixed periodicity. Even at the same 1-hour resolution, the dominant timescale of $\mathbf{D}^{(1)}$ varies considerably across test windows, with prominent occurrences around 24, 32, and 48 hours. These timescales are obtained only through post-hoc spectral analysis of the decomposed deviation; they are neither predefined in `LocTrend` nor used during dictionary fitting or forward reconstruction. Their variation across windows therefore provides empirical evidence that the decomposition is not anchored to a prescribed seasonal period. The three analyses show that `LocTrend` achieves scale adaptation through data-dependent structural assignment while accommodating locally varying temporal organizations without assuming fixed periodicity.

A.13 Basin-Specific Evaluation and Practical Deployment

Basins can differ in rainfall-runoff responses, pollutant sources, monitoring configurations, and water-quality dynamics. These differences motivate fitting and validating the forecasting model using observations from the target basin. Accordingly, our experiments on three basins train and evaluate RaiNet separately within each basin using chronological data splits. This protocol evaluates the applicability of the framework across different basin-specific forecasting tasks.

We use the same model and training hyperparameters across all three datasets, while fitting model parameters separately for each basin. No dataset-specific hyperparameter tuning is performed. Under this shared configuration, the ablation results show that component effects vary across basins and evaluation metrics. The complete framework achieves average MSE improvements over the evaluated ablated configurations, whereas some MAE comparisons show smaller gains or regressions. These results support the effectiveness of the overall framework across the evaluated basins without dataset-specific hyperparameter adjustments, while showing that individual components do

not benefit every basin and metric equally. We report both improvements and regressions, together with basin-specific results, to make this variability explicit. The associated system is designed to integrate water-quality monitoring and rainfall information to support water-quality forecasting and the identification of potential deterioration, providing information for monitoring and management decisions.