

A Hybrid FE–TCN–Transformer Framework for Water Quality Prediction in a Large Urban River System

Lei Huang^{a,d,*}, Zhihao Wang^a, Qiannan Duan^{b,*}, Zenghui Li^a, Fan Liu^a, Zihao Li^a, Chen

Ma^a, Xiaowei Lu^a, Gang Wang^d, Peng Wang^c, Ming Chen^e

^a State Key Laboratory of Water Pollution Control and Green Resource Recycling,
School of the Environment, Nanjing University, Nanjing 210023, China;

^b Shaanxi Key Laboratory of Earth Surface System and Environmental Carrying
Capacity, College of Urban and Environmental Sciences, Northwest University, Xi'an
710127, China;

^c Faculty of Civil Engineering and Mechanics, Jiangsu University, Zhenjiang 212013,
China;

^d Nanjing University (Suzhou) High-tech Institute, Suzhou 215123, China;

^e School of Environmental Engineering, Jiangxi University of Science and Technology,
Ganzhou, 341000, China;

Correspondence:

* Lei Huang, E-mail: huanglei@nju.edu.cn; * Qiannan Duan, E-mail:
qnduan@nwu.edu.cn.

Text S1 Algorithms

S1.1 Baseline models

To provide a broad and systematic evaluation of the proposed framework, seven representative baseline models were selected, covering classical statistical approaches, tree-based machine learning, recurrent neural networks, and Transformer-based time-series forecasting models.

S1.1.1 ARIMA

The autoregressive integrated moving average (ARIMA) model represents temporal dependence through autoregressive terms, differencing, and moving-average error terms (Box et al., 2015). It was included as a classical linear benchmark for evaluating whether short-term water-quality dynamics could be adequately captured from the historical behavior of the target variable alone.

S1.1.2 STL-ARIMA

STL-ARIMA combines seasonal-trend decomposition using locally estimated scatterplot smoothing (STL) with ARIMA forecasting (Cleveland et al., 1990; Box et al., 2015). STL separates the observed series into trend, seasonal, and remainder components. ARIMA is subsequently applied to the seasonally adjusted series, and the forecast is reconstructed by combining the predicted non-seasonal component with the corresponding seasonal component. This model provides a decomposition-based statistical benchmark.

S1.1.3 Temporal XGBoost

Extreme gradient boosting (XGBoost) constructs an additive ensemble of regression trees using a regularized objective and second-order gradient information (Chen and Guestrin, 2016). For time-series forecasting, lagged observations of the selected input variables were organized as a fixed-length predictor vector, allowing the model to represent nonlinear relationships and interactions among historical variables without recurrent operations.

S1.1.4 LSTM

Long short-term memory (LSTM) is a recurrent neural-network architecture designed to alleviate the vanishing-gradient problem and preserve information over extended sequences (Hochreiter and Schmidhuber, 1997). Its memory cell and input, forget, and output gates regulate the retention, removal, and transmission of temporal information. The final sequence representation was mapped to the one-step-ahead water-quality prediction.

S1.1.5 GRU

The gated recurrent unit (GRU) provides a compact gated recurrent architecture using reset and update gates to control the integration of historical and current information (Cho et al., 2014). Unlike LSTM, GRU does not maintain a separate memory-cell state and therefore uses fewer gating operations. It was included to evaluate recurrent temporal modeling with a comparatively compact architecture.

S1.1.6 Informer

Informer is a Transformer-based architecture developed for time-series forecasting (Zhou et al., 2021). It introduces ProbSparse self-attention to reduce the computational cost of long-sequence dependency modeling and employs attention distillation to retain dominant temporal representations. In this study, Informer was configured for single-step water-quality forecasting using the selected historical input variables.

S1.1.7 PatchTST

PatchTST represents time series as sequences of partially overlapping temporal patches and applies a Transformer encoder to the resulting patch representations (Nie et al., 2023). Patching increases the effective receptive field while reducing the number of tokens processed by the attention layers. Its channel-independent representation further enables shared temporal-pattern learning across input variables.

For models supporting multivariate inputs, including temporal XGBoost, LSTM, GRU, Informer, and PatchTST, the same task-specific input-variable sets used by the proposed FE–TCN–Transformer were adopted. The conventional ARIMA and STL–ARIMA models were implemented as univariate forecasting models and therefore used only the historical observations of the corresponding target indicator. All models were evaluated using the same

historical window, chronological data partitions, preprocessing procedure, forecasting horizon, and evaluation metrics. Detailed model architectures and hyperparameter settings are provided in Table 1.

S1.2 Fourier-EMA

The Fourier-EMA module represents the denoised target series y_t using a causal multi-scale trend and a periodic seasonal component. For each scale s , the causal EMA sequence e_t is updated using only current and previous observations (Eq. 1) and assembled into the multi-scale feature vector E_t (Eq. 2). Ridge-regression weights w are estimated over the training interval (Eq. 3) to obtain $Trend_t$ over the full series (Eq. 4). The seasonal component is fitted to the detrended residual r_t defined in Eq. 7, using multi-period, multi-harmonic Fourier basis functions (Eq. 5) assembled into the feature vector F_t (Eq. 6). The coefficients β are estimated by ridge regression over the training interval (Eq. 8) to generate $Seasonal_t$ over the full series (Eq. 9). All parameters are fitted exclusively on the training interval and extrapolated forward to the validation and test periods to prevent information leakage.

$$e_t^{(s)} = \alpha_s y_t + (1 - \alpha_s) e_{t-1}^{(s)}, \quad \alpha_s = \frac{2}{s+1}, \quad s \in P \quad (1)$$

$$E_t = [e_t^{(s_1)}, \dots, e_t^{(s_M)}] \quad (2)$$

$$w = \underset{m}{\operatorname{argmin}} \left\| y_{1:t_{\text{split}}} - E_{1:t_{\text{split}}} w \right\|_2^2 + \lambda \|w\|_2^2 \quad (3)$$

$$Trend_t = E_t \widehat{w}, \quad t = 0, \dots, N-1 \quad (4)$$

$$\phi_{p,k}^{\sin}(t) = \sin\left(\frac{2\pi kt}{p}\right), \quad \phi_{p,k}^{\cos}(t) = \cos\left(\frac{2\pi kt}{p}\right), \quad p \in P, k = 1, \dots, K \quad (5)$$

$$F_t = \left[\left\{ \phi_{p,k}^{\sin}(t), \phi_{p,k}^{\cos}(t) \right\}_{p \in P, k=1}^K \right], \quad M = 2|P|K \quad (6)$$

$$r_t = y_t - Trend_t \quad (7)$$

$$\beta = \underset{\beta}{\operatorname{argmin}} \left\| r_{1:t_{\text{split}}} - F_{1:t_{\text{split}}} \beta \right\|_2^2 + \lambda \|\beta\|_2^2 \quad (8)$$

$$Seasonal_t = F_t \widehat{\beta}, \quad t = 0, \dots, N-1 \quad (9)$$

Here, y_t is the denoised target series; N is the series length; t_{split} is the end of the training interval; P is the period set; s is the EMA scale; $e_t^{(s)}$ is the corresponding EMA sequence; E_t is the multi-scale EMA feature vector; $\widehat{w}$ is the ridge-regression weight vector; λ is the ridge

regularization coefficient ($\lambda = 10^{-6}$ in Eqs. 3 and 8); r_t is the detrended residual;

$\phi_{p,k}^{\sin}(t), \phi_{p,k}^{\cos}(t)$ are the Fourier basis functions; F_t is the M-dimensional Fourier feature vector;

β is the estimated seasonal-coefficient vector.

S1.3 TCN

The Temporal Convolutional Network (TCN) (Bai et al., 2018) extracts sequence features via one-dimensional convolutions along the time axis, employing a causal convolution form where outputs depend solely on historical inputs to prevent future information leakage (Eq. 10). To cover longer historical windows with small kernels, TCN incorporates dilation rates in convolutional layers for sparse historical sampling and receptive field expansion (Eq. 11). Typical architectures stack multiple convolutional layers as residual blocks, with skip connections stabilizing deep training and preserving information flow consistency across layers. Core TCN computations are shown in Eqs. 10–12.

$$y_t = \sum_{i=0}^{k-1} w_i x_{t-i} \quad (10)$$

$$y_t = \sum_{i=0}^{k-1} w_i x_{t-di} \quad (11)$$

$$R = 1 + (k-1) \sum_{\ell=1}^L \prod_{j=1}^{\ell} d_j \quad (12)$$

Here, x_t and y_t denote inputs and outputs at time t ; w_i the convolutional kernel weights, K the kernel size; d the dilation rate (Eq. 11), reducing to standard causal convolution (Eq. 10) when $d=1$; R the receptive field length, spanning historical inputs influencing output y_t (Eq. 12); L the number of convolutional layers, d_ℓ the dilation rate of layer ℓ .

S1.4 Transformer

Transformer (Vaswani et al., 2017) treats time series as timestep tokens and dynamically aggregates information across timesteps via self-attention to model long-range dependencies. In WQ–MET multivariate input scenarios, it captures both inter-variable couplings and non-adjacent temporal correlations. For the 4 h records, each L-step sliding window is linearly embedded and combined with positional encoding (Eq. 13). Scaled dot-product attention is defined in Eq. 14, and multiple attention heads are computed and concatenated in Eqs. 15–16.

A position-wise feed-forward network transforms each timestep representation (Eq. 17), while residual updates are expressed in Eqs. 18–19. The task output layer produces the H-step-ahead forecast (Eq. 20). Decoder-side masked attention is used when an encoder-decoder prediction head is applied to prevent access to future labels.

$$z_t = x_t W_{\text{emb}} + b_{\text{emb}} + p_t \quad (13)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (14)$$

$$\text{head}_h = \text{Attention}(QW_h^Q, KW_k^K, VW_k^V), \quad h = 1, \dots, H \quad (15)$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_H)W^O \quad (16)$$

$$\text{FFN}(u) = \phi(uW_1 + b_1)W_2 + b_2 \quad (17)$$

$$u_t = z_t + \text{MultiHead}(Q, K, V)_t \quad (18)$$

$$h_t = u_t + \text{FFN}(u_t) \quad (19)$$

$$\hat{y}_{t+1:t+H} = g(h_{t-L+1:t}) \quad (20)$$

Here, x_t is the input vector at timestep t , concatenated from selected WQ and MET subsets; W_{emb} , b_{emb} are the linear embedding matrix and bias; p_t is the positional encoding; z_t is the post-embedding timestep representation. Q , K , and V are the query, key, and value matrices; d_k is the key dimension; $\text{softmax}(\cdot)$ normalizes attention weights temporally. In Eqs. 15–16, H is the number of attention heads; W_h^Q , W_k^K , W_k^V are the linear mappings for head h ; W^O is the output projection; $\text{Concat}(\cdot)$ concatenates along the feature dimension. W_1, W_2, b_1, b_2 are the FFN parameters; $\phi(\cdot)$ is the nonlinear activation. u_t , h_t are the post-attention and post-feedforward representations; $g(\cdot)$ is the task output layer that maps the encoded sequence to the forecast. In Eq. 20, H is the forecast horizon and equals 1 in this study.

S1.5 TCN-Transformer

TCN-Transformer achieves joint modeling of local temporal patterns and long-range dependencies in time series via sequential cascading of TCN and Transformer. Specifically, the input sequence first undergoes convolutional modeling in the TCN frontend, comprising multiple residual-connected convolutional blocks to enhance local and meso-scale temporal

feature representations while preserving timestep alignment. TCN output features then directly serve as Transformer inputs, with encoder layers further modeling global inter-timestep dependencies temporally.

In this hybrid structure, no additional temporal resampling, pooling, or linear mappings intervene between TCN and Transformer; temporal and feature dimensions remain consistent throughout, ensuring synergistic operation within the same timestep representation space. Transformer encodings are ultimately mapped via the task output layer to H-step ahead predictions.

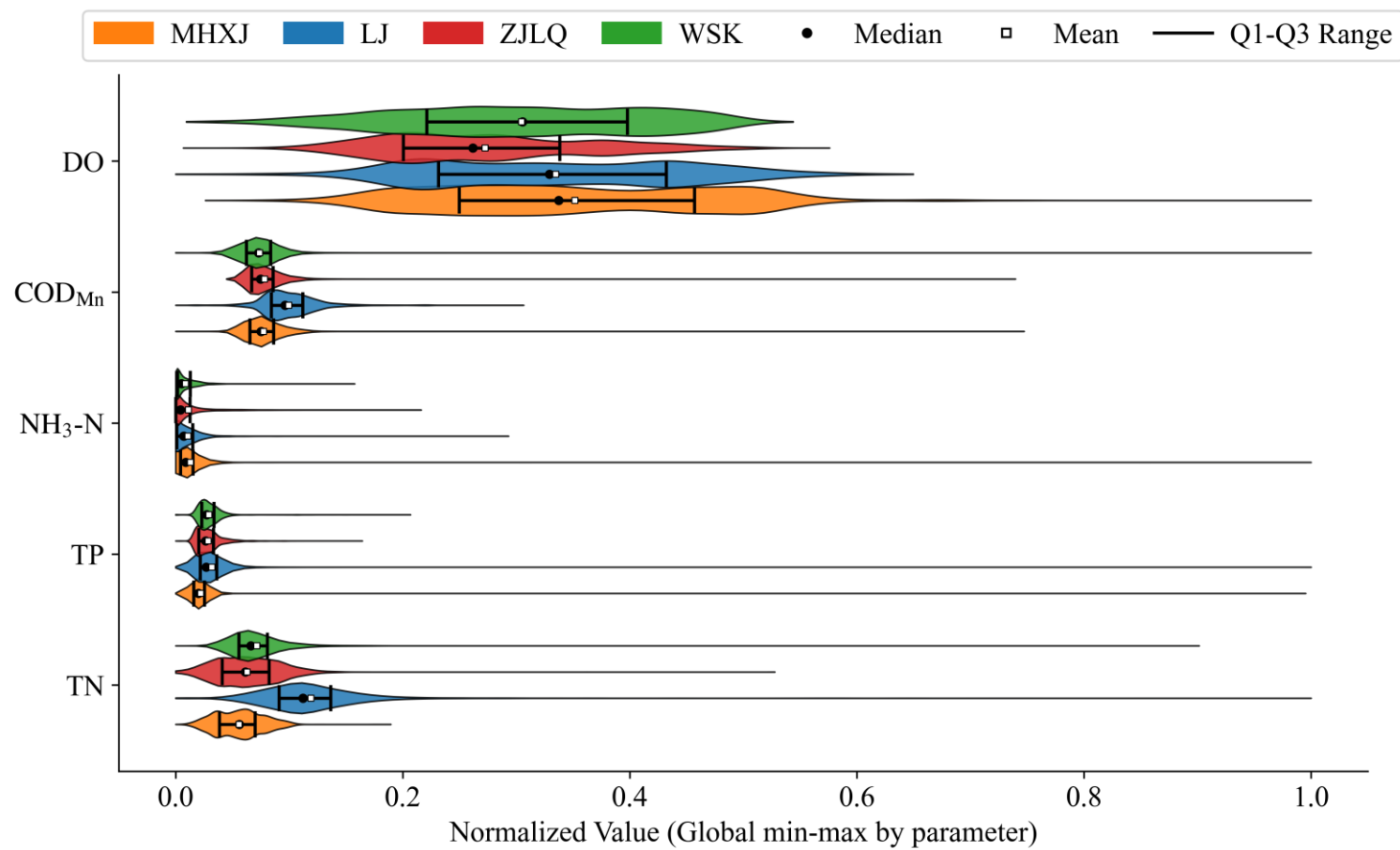


Fig. S1 Violin plots of normalized distributions for TN, TP, NH₃-N, COD_{Mn}, and DO across the four monitoring sites (MHXJ, LJ, ZJLQ, and WSK).

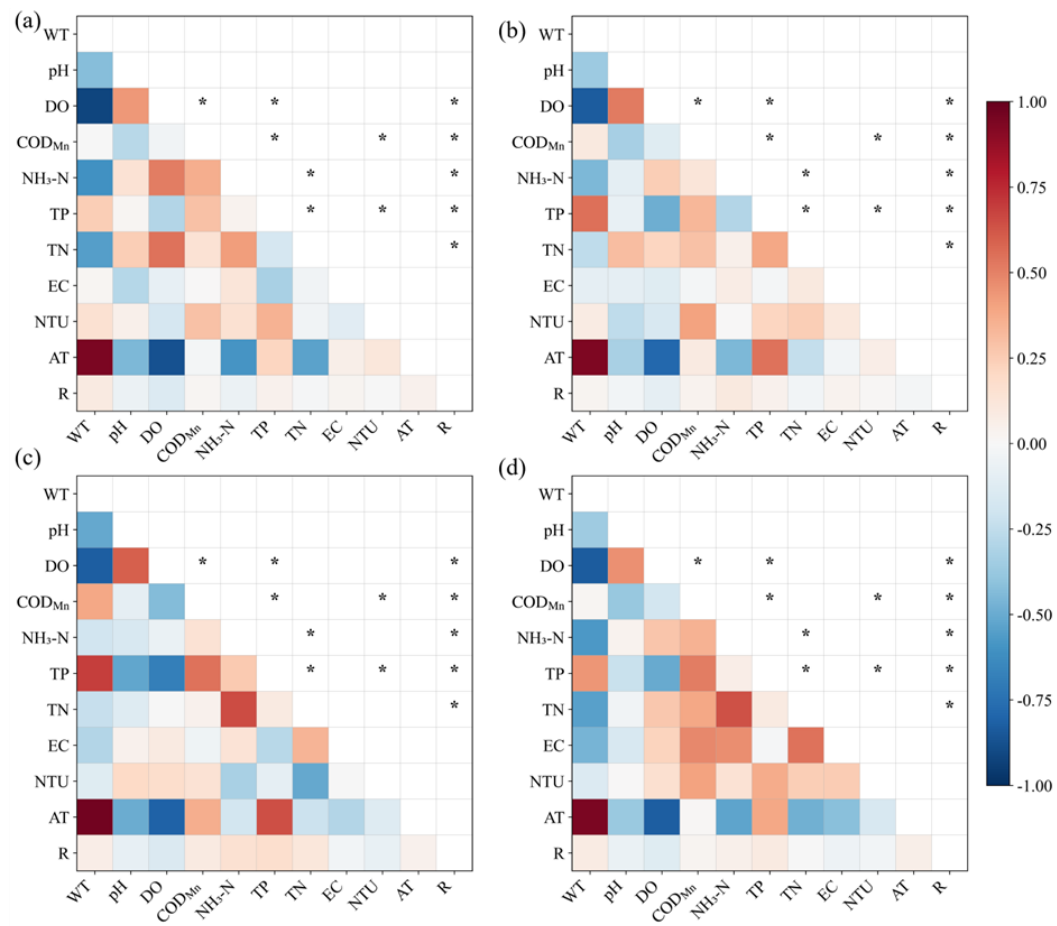


Fig. S2 Spearman rank correlation matrices among candidate variables at the four monitoring sites. Panels (a–d) correspond to MHXJ, LJ, ZJLQ, and WSK, respectively.

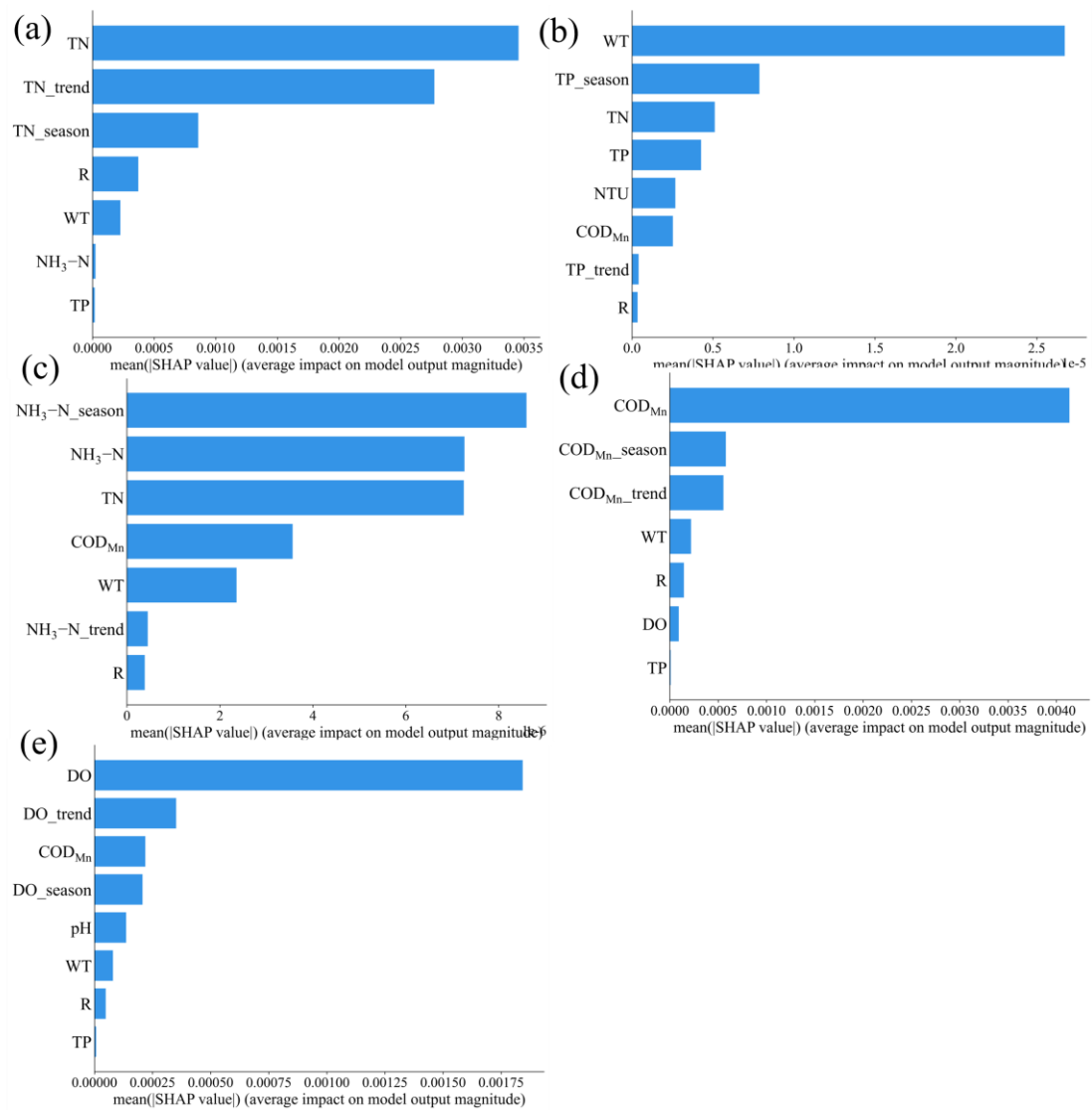


Fig. S3 Mean absolute SHAP-value bar plots for TN, TP, NH₃-N, COD_{Mn}, and DO at MHXJ. Panels (a–e)

correspond to each indicator.

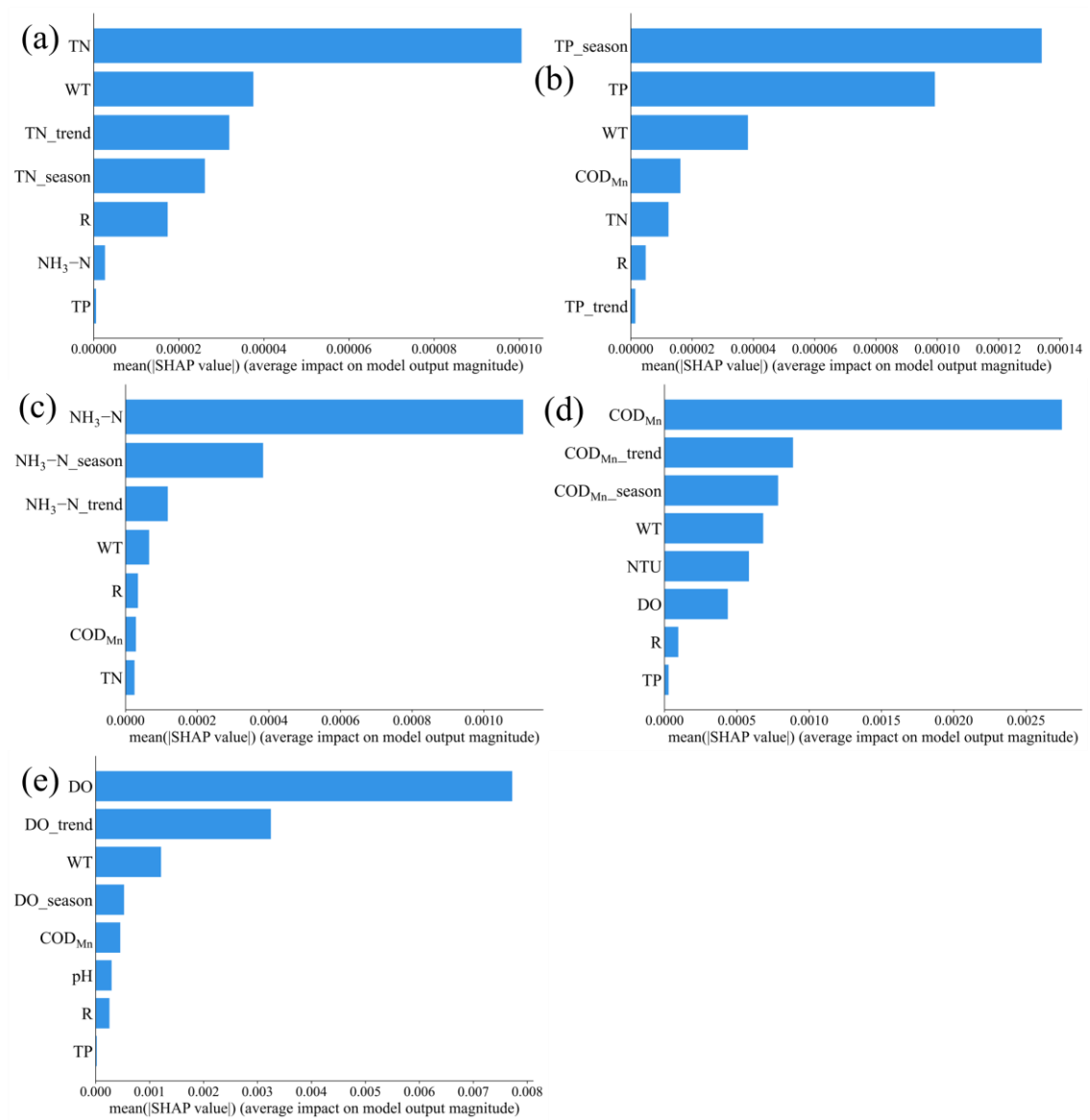


Fig. S4 Mean absolute SHAP-value bar plots for TN, TP, NH₃-N, COD_{Mn}, and DO at LJ. Panels (a–e) correspond to each indicator.

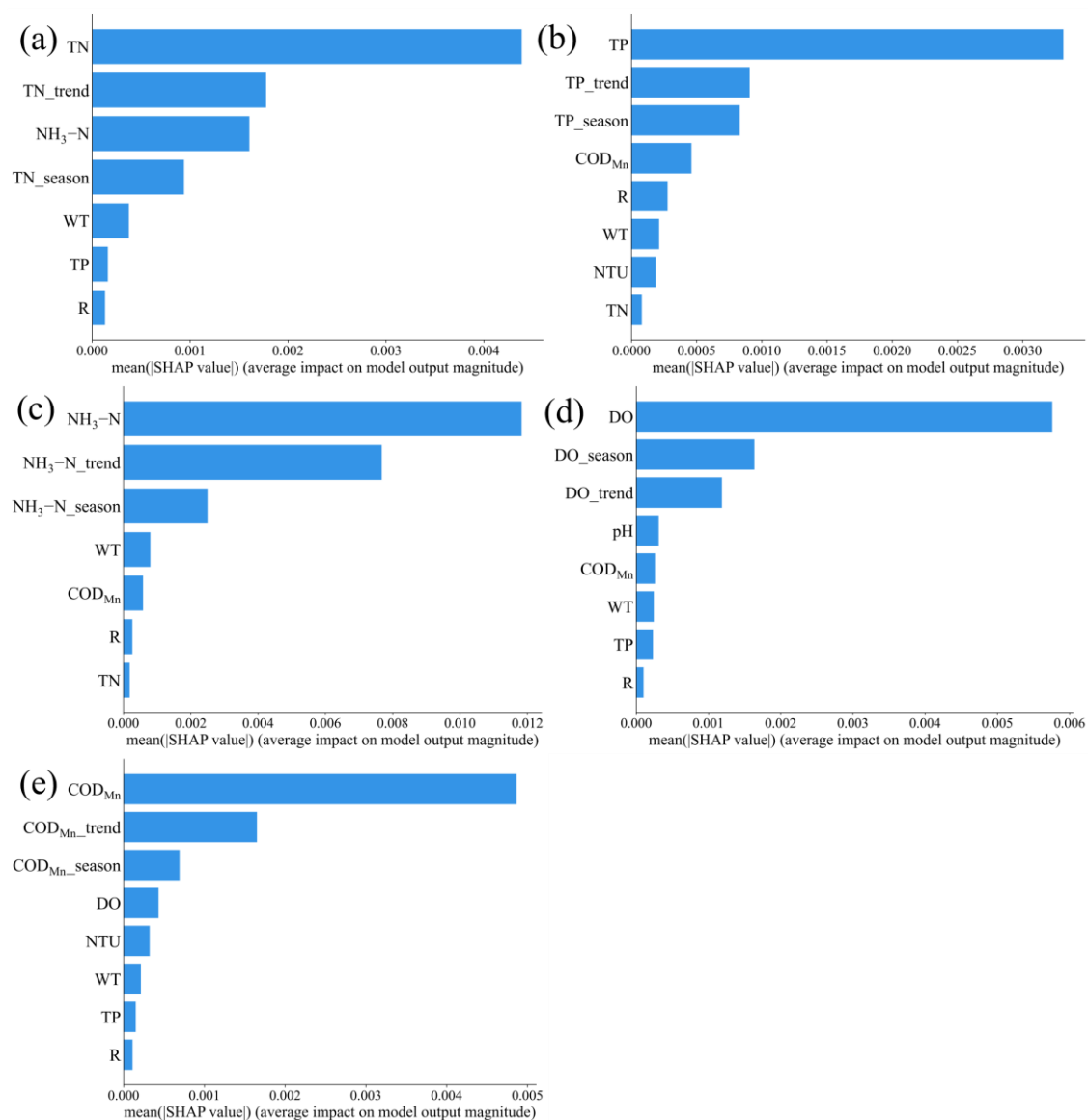


Fig. S5 Mean absolute SHAP-value bar plots for TN, TP, NH₃-N, COD_{Mn}, and DO at WSK. Panels (a–e) correspond to each indicator.

Table S1 Water quality and meteorological indicators and the corresponding Class I–V standards (GB 3838—2002)

Category	Indicator (abbr.)	English full name	Unit	Class I	Class II	Class III	Class IV	Class V
WQ	WT	Water temperature	°C	Weekly mean increase ≤ 1 ; weekly max increase ≤ 2				
WQ	pH	pH	dimensionless	6–9	6–9	6–9	6–9	6–9
WQ	DO	Dissolved oxygen	mg/L	≥ 7.5 or saturation $\geq 90\%$	≥ 6	≥ 5	≥ 3	≥ 2
WQ	EC	Electrical conductivity	$\mu\text{S}/\text{cm}$	-	-	-	-	-
WQ	NTU	Turbidity	NTU	-	-	-	-	-
WQ	COD _{Mn}	Permanganate index	mg/L	≤ 2	≤ 4	≤ 6	≤ 10	≤ 15
WQ	NH ₃ -N	Ammonia nitrogen	mg/L	≤ 0.15	≤ 0.5	≤ 1.0	≤ 1.5	≤ 2.0
WQ	TP	Total phosphorus	mg/L	≤ 0.02	≤ 0.10	≤ 0.20	≤ 0.30	≤ 0.40
WQ	TN	Total nitrogen	mg/L	≤ 0.20	≤ 0.50	≤ 1.00	≤ 1.50	≤ 2.00
MET	AT	Air temperature	°C	-	-	-	-	-
MET	R	Rainfall	mm	-	-	-	-	-

Table S2 Descriptive statistics of water quality variables at four monitoring stations in the Huangpu River (4-h observations, 2020–2024).											
Water quality monitor station	Water quality indicators	Sample size	Minimum value	Lower quartile (Q1)	Median (Q2)	Upper quartile (Q3)	Mean	Maximum value	Standard deviation	Raw missing (%)	Hampel outliers (%)
LJ	TN	6809	0.05	3.38	4.14	5.04	4.40	36.57	2.03	11.38	13.27
LJ	TP	6809	0.01	0.11	0.13	0.17	0.15	4.65	0.21	14.03	13.58
LJ	NH ₃ -N	6809	0.03	0.04	0.17	0.32	0.24	5.88	0.31	14.13	15.60
LJ	COD _{Mn}	6809	0.25	3.90	4.42	5.09	4.57	13.53	1.10	16.05	12.74
LJ	DO	6809	0.02	4.65	6.60	8.65	6.71	13.00	2.33	0.18	16.35
MHXJ	TN	7114	0.05	1.46	2.09	2.60	2.09	6.97	0.78	2.33	6.99
MHXJ	TP	7114	0.01	0.08	0.10	0.12	0.11	4.63	0.12	1.76	7.53
MHXJ	NH ₃ -N	7114	0.03	0.11	0.20	0.33	0.28	19.99	0.73	1.03	9.64
MHXJ	COD _{Mn}	7114	0.25	3.08	3.51	3.98	3.60	32.64	0.89	2.46	9.41
MHXJ	DO	7114	0.54	5.01	6.76	9.15	7.05	20.00	2.50	0.20	6.10
WSK	TN	7658	0.05	2.08	2.47	3.00	2.66	32.97	1.28	0.71	7.36
WSK	TP	7658	0.01	0.11	0.13	0.16	0.14	0.97	0.05	1.57	7.63
WSK	NH ₃ -N	7658	0.03	0.06	0.13	0.28	0.19	3.17	0.19	0.84	12.98
WSK	COD _{Mn}	7658	0.25	2.95	3.42	3.87	3.45	43.59	0.96	2.23	8.60
WSK	DO	7658	0.21	4.44	6.12	7.97	6.11	10.89	2.19	0.01	8.18
ZJLQ	TN	6907	0.05	1.55	2.29	3.05	2.34	19.32	1.11	0.87	5.87
ZJLQ	TP	6907	0.01	0.10	0.13	0.16	0.14	0.77	0.06	1.28	8.16
ZJLQ	NH ₃ -N	6907	0.03	0.03	0.11	0.28	0.25	4.34	0.39	1.40	8.07
ZJLQ	COD _{Mn}	6907	2.19	3.15	3.49	3.96	3.64	32.31	0.91	3.45	7.90
ZJLQ	DO	6907	0.16	4.02	5.25	6.78	5.47	11.53	1.86	0.21	7.18

Table S3 Input features of model for different prediction targets

Covariate combination				
Predict index	MHXJ	LJ	ZJLQ	WSK
TN		WT, NH ₃ -N, TP, R		
TP	WT,TN,COD _{Mn} , NTU, R		WT,TN,COD _{Mn} ,R	
NH ₃ -N		WT, TN, COD _{Mn} , R		
COD _{Mn}	WT,TP,DO,R	WT,TP,DO,NTU, R	WT,TP,DO,R	WT,TP,DO,NTU, R
DO		WT, pH, TP, COD _{Mn} , R		

Table S4 Default parameters for the FE (Fourier–EMA) feature engineering module

Model module	Purpose	Default setting
Sampling interval	Define discrete time step	4-hourly time step
Trend baseline (EMA)	Capture low-frequency trend	EMA with adjust=False, min_periods=1
EMA scales computed	Provide multi-scale trend candidates	Daily (6), weekly (42), monthly (180), quarterly (540), semiannual (1080)
Trend weight fitting	Combine trend candidates	Ridge regression, $\alpha = 1e-6$ (training segment)
Final trend baseline	Deterministic trend component	Weighted daily + weekly EMA
Seasonal baseline (Fourier)	Capture periodic structure	Fourier series with K = 5 harmonics
Seasonal periods	Candidate seasonal cycles	Daily (6), weekly (42), monthly (180), quarterly (540), semiannual (1080)
Seasonal coefficient fitting	Estimate seasonal amplitudes	Ridge regression, $\alpha = 1e-6$ (training segment)
FE baseline output	Predictable baseline signal	trend + season
Supervised target	Learning objective	Residual = observation – baseline
FE features appended	Input covariates	<target>_trend, <target>_season
Scaling strategy	Stabilize optimization	Inputs: Min–Max; residual target: z-score

Table S5 FE-TCN-Transformer Model performance in predicting water quality in four study sites

WQ	Station	NSE	RMSE	MAE	MAPE
TN	MHXJ	0.990	0.059	0.043	0.030
	LJ	0.669	0.428	0.234	0.059
	ZJLQ	0.902	0.216	0.121	0.056
	WSK	0.871	0.158	0.115	0.045
TP	MHXJ	0.572	0.031	0.006	0.048
	LJ	0.741	0.020	0.012	0.077
	ZJLQ	0.888	0.018	0.009	0.056
	WSK	0.820	0.015	0.009	0.064
NH ₃ -N	MHXJ	0.895	0.022	0.016	0.332
	LJ	0.723	0.068	0.036	0.317
	ZJLQ	0.864	0.140	0.057	0.448
	WSK	0.754	0.053	0.028	0.410
COD _{Mn}	MHXJ	0.907	0.176	0.110	0.033
	LJ	0.773	0.224	0.148	0.036
	ZJLQ	0.870	0.221	0.138	0.038
	WSK	0.791	0.194	0.148	0.047
DO	MHXJ	0.991	0.164	0.115	0.020
	LJ	0.849	0.663	0.423	0.080
	ZJLQ	0.947	0.403	0.254	0.052
	WSK	0.963	0.340	0.240	0.062

Table S6 Performance of various models at the MHXJ site

Parameter	Model	NSE	RMSE	MAE	MAPE
TN	ARIMA	0.977	0.101	0.070	0.048
TN	STL-ARIMA	0.972	0.111	0.079	0.055
TN	XGBoost	0.968	0.119	0.086	0.061
TN	LSTM	0.975	0.104	0.075	0.054
TN	GRU	0.976	0.103	0.073	0.053
TN	Informer	0.976	0.103	0.072	0.050
TN	PatchTST	0.975	0.105	0.074	0.051
TN	Transformer	0.963	0.114	0.088	0.070
TN	TCN-Transformer	0.968	0.105	0.083	0.063
TN	FE-Transformer	0.989	0.061	0.045	0.031
TN	FE-TCN-Transformer	0.990	0.059	0.043	0.030
TP	ARIMA	0.453	0.054	0.010	0.111
TP	STL-ARIMA	0.325	0.059	0.030	0.351
TP	XGBoost	0.150	0.067	0.014	0.140
TP	LSTM	0.254	0.063	0.012	0.112
TP	GRU	0.323	0.060	0.012	0.117
TP	Informer	0.410	0.056	0.012	0.115
TP	PatchTST	0.064	0.070	0.020	0.193
TP	Transformer	0.355	0.061	0.032	0.337
TP	TCN-Transformer	0.140	0.056	0.026	0.238
TP	FE-Transformer	0.526	0.033	0.007	0.071
TP	FE-TCN-Transformer	0.572	0.031	0.006	0.048
NH3-N	ARIMA	0.544	0.102	0.025	0.219
NH3-N	STL-ARIMA	0.537	0.103	0.029	0.321
NH3-N	XGBoost	0.339	0.385	0.302	0.127
NH3-N	LSTM	0.605	0.095	0.029	0.309
NH3-N	GRU	0.594	0.097	0.029	0.318
NH3-N	Informer	0.317	0.125	0.046	0.414
NH3-N	PatchTST	0.576	0.099	0.033	0.294
NH3-N	Transformer	0.226	0.075	0.054	0.870
NH3-N	TCN-Transformer	0.261	0.059	0.053	1.290
NH3-N	FE-Transformer	0.801	0.030	0.023	0.585
NH3-N	FE-TCN-Transformer	0.895	0.022	0.016	0.332
COD _{Mn}	ARIMA	0.247	0.713	0.287	0.094
COD _{Mn}	STL-ARIMA	0.250	0.712	0.287	0.094
COD _{Mn}	XGBoost	0.263	0.705	0.296	0.097
COD _{Mn}	LSTM	0.301	0.687	0.281	0.091
COD _{Mn}	GRU	0.292	0.691	0.274	0.091
COD _{Mn}	Informer	0.271	0.701	0.285	0.093
COD _{Mn}	PatchTST	0.204	0.733	0.331	0.109

Parameter	Model	NSE	RMSE	MAE	MAPE
COD _{Mn}	Transformer	0.796	0.272	0.190	0.060
COD _{Mn}	TCN-Transformer	0.827	0.250	0.156	0.048
COD _{Mn}	FE-Transformer	0.905	0.179	0.111	0.033
COD _{Mn}	FE-TCN-Transformer	0.907	0.176	0.110	0.033
DO	ARIMA	0.964	0.300	0.211	0.036
DO	STL-ARIMA	0.964	0.301	0.211	0.036
DO	XGBoost	0.966	0.294	0.207	0.036
DO	LSTM	0.963	0.307	0.222	0.038
DO	GRU	0.963	0.303	0.216	0.037
DO	Informer	0.967	0.288	0.198	0.034
DO	PatchTST	0.969	0.280	0.202	0.035
DO	Transformer	0.943	0.410	0.332	0.060
DO	TCN-Transformer	0.973	0.281	0.204	0.037
DO	FE-Transformer	0.989	0.181	0.118	0.021
DO	FE-TCN-Transformer	0.991	0.164	0.115	0.020

Table S7 Performance of various models at the LJ site

Parameter	Model	NSE	RMSE	MAE	MAPE
TN	ARIMA	0.157	1.139	0.651	0.145
TN	STL-ARIMA	0.121	1.163	0.680	0.152
TN	XGBoost	0.219	1.097	0.644	0.148
TN	LSTM	0.253	1.072	0.653	0.150
TN	GRU	0.256	1.070	0.657	0.152
TN	Informer	0.246	1.077	0.665	0.155
TN	PatchTST	0.263	1.065	0.642	0.146
TN	Transformer	0.309	0.686	0.530	0.148
TN	TCN-Transformer	0.583	0.533	0.316	0.079
TN	FE-Transformer	0.601	0.471	0.233	0.059
TN	FE-TCN-Transformer	0.669	0.428	0.234	0.059
TP	ARIMA	0.040	0.058	0.032	0.183
TP	STL-ARIMA	0.186	0.061	0.036	0.204
TP	XGBoost	0.336	0.048	0.027	0.150
TP	LSTM	0.321	0.049	0.028	0.160
TP	GRU	0.332	0.048	0.028	0.164
TP	Informer	0.296	0.049	0.030	0.172
TP	PatchTST	0.355	0.047	0.028	0.164
TP	Transformer	0.277	0.037	0.026	0.171
TP	TCN-Transformer	0.379	0.035	0.022	0.132
TP	FE-Transformer	0.739	0.020	0.012	0.080
TP	FE-TCN-Transformer	0.741	0.020	0.012	0.077
NH3-N	ARIMA	0.149	0.207	0.097	1.050
NH3-N	STL-ARIMA	0.137	0.209	0.100	1.116
NH3-N	XGBoost	0.125	0.210	0.105	1.638
NH3-N	LSTM	0.176	0.204	0.091	0.953
NH3-N	GRU	0.167	0.205	0.100	1.203
NH3-N	Informer	0.117	0.211	0.098	1.176
NH3-N	PatchTST	0.158	0.206	0.095	1.067
NH3-N	Transformer	0.463	0.105	0.067	1.026
NH3-N	TCN-Transformer	0.465	0.105	0.072	1.124
NH3-N	FE-Transformer	0.718	0.068	0.037	0.357
NH3-N	FE-TCN-Transformer	0.723	0.068	0.036	0.317
COD _{Mn}	ARIMA	0.181	0.679	0.411	0.092
COD _{Mn}	STL-ARIMA	0.133	0.699	0.428	0.096
COD _{Mn}	XGBoost	0.228	0.485	0.393	0.154
COD _{Mn}	LSTM	0.139	0.696	0.487	0.115
COD _{Mn}	GRU	0.178	0.681	0.431	0.098
COD _{Mn}	Informer	0.183	0.678	0.424	0.096
COD _{Mn}	PatchTST	0.248	0.651	0.386	0.086

Parameter	Model	NSE	RMSE	MAE	MAPE
COD _{Mn}	Transformer	0.466	0.385	0.281	0.070
COD _{Mn}	TCN-Transformer	0.534	0.360	0.248	0.062
COD _{Mn}	FE-Transformer	0.759	0.230	0.158	0.039
COD _{Mn}	FE-TCN-Transformer	0.773	0.224	0.148	0.036
DO	ARIMA	0.434	1.367	1.009	0.565
DO	STL-ARIMA	0.449	1.349	0.993	0.548
DO	XGBoost	0.436	1.365	0.962	0.886
DO	LSTM	0.432	1.369	0.991	0.620
DO	GRU	0.444	1.354	0.995	0.573
DO	Informer	0.291	1.530	1.146	0.955
DO	PatchTST	0.499	1.286	0.874	0.687
DO	Transformer	0.725	0.956	0.628	0.437
DO	TCN-Transformer	0.730	0.948	0.598	0.308
DO	FE-Transformer	0.846	0.669	0.397	0.078
DO	FE-TCN-Transformer	0.849	0.663	0.423	0.080

Table S8 Performance of various models at the ZJLQ site

Parameter	Model	NSE	RMSE	MAE	MAPE
TN	ARIMA	0.881	0.310	0.187	0.084
TN	STL-ARIMA	0.880	0.311	0.194	0.088
TN	XGBoost	0.861	0.334	0.212	0.100
TN	LSTM	0.873	0.319	0.194	0.089
TN	GRU	0.874	0.318	0.204	0.097
TN	Informer	0.863	0.332	0.200	0.089
TN	PatchTST	0.838	0.360	0.232	0.110
TN	Transformer	0.668	0.427	0.320	0.173
TN	TCN-Transformer	0.737	0.380	0.278	0.147
TN	FE-Transformer	0.899	0.222	0.117	0.057
TN	FE-TCN-Transformer	0.902	0.216	0.121	0.056
TP	ARIMA	0.795	0.026	0.014	0.091
TP	STL-ARIMA	0.795	0.026	0.014	0.091
TP	XGBoost	0.771	0.027	0.015	0.096
TP	LSTM	0.765	0.028	0.015	0.093
TP	GRU	0.798	0.026	0.014	0.088
TP	Informer	0.767	0.028	0.016	0.101
TP	PatchTST	0.782	0.027	0.016	0.099
TP	Transformer	0.818	0.025	0.014	0.093
TP	TCN-Transformer	0.770	0.028	0.017	0.117
TP	FE-Transformer	0.887	0.018	0.010	0.059
TP	FE-TCN-Transformer	0.888	0.018	0.009	0.056
NH3-N	ARIMA	0.786	0.207	0.084	0.311
NH3-N	STL-ARIMA	0.782	0.208	0.095	0.631
NH3-N	XGBoost	0.702	0.244	0.109	0.624
NH3-N	LSTM	0.815	0.192	0.083	0.503
NH3-N	GRU	0.822	0.188	0.080	0.457
NH3-N	Informer	0.719	0.237	0.112	0.812
NH3-N	PatchTST	0.732	0.231	0.123	1.248
NH3-N	Transformer	0.718	0.224	0.111	0.955
NH3-N	TCN-Transformer	0.699	0.232	0.118	1.510
NH3-N	FE-Transformer	0.863	0.141	0.060	0.393
NH3-N	FE-TCN-Transformer	0.864	0.140	0.057	0.448
COD _{Mn}	ARIMA	0.613	0.368	0.224	0.063
COD _{Mn}	STL-ARIMA	0.613	0.368	0.224	0.063
COD _{Mn}	XGBoost	0.606	0.371	0.255	0.075
COD _{Mn}	LSTM	0.643	0.354	0.226	0.065
COD _{Mn}	GRU	0.657	0.346	0.218	0.061
COD _{Mn}	Informer	0.667	0.341	0.222	0.063
COD _{Mn}	PatchTST	0.573	0.387	0.259	0.075

Parameter	Model	NSE	RMSE	MAE	MAPE
COD _{Mn}	Transformer	0.724	0.346	0.254	0.074
COD _{Mn}	TCN-Transformer	0.741	0.335	0.211	0.059
COD _{Mn}	FE-Transformer	0.841	0.245	0.140	0.039
COD _{Mn}	FE-TCN-Transformer	0.870	0.221	0.138	0.038
DO	ARIMA	0.874	0.628	0.432	0.083
DO	STL-ARIMA	0.866	0.649	0.454	0.087
DO	XGBoost	0.805	0.783	0.557	0.101
DO	LSTM	0.848	0.690	0.490	0.091
DO	GRU	0.857	0.670	0.477	0.089
DO	Informer	0.851	0.685	0.484	0.093
DO	PatchTST	0.863	0.656	0.454	0.086
DO	Transformer	0.885	0.614	0.421	0.082
DO	TCN-Transformer	0.892	0.594	0.383	0.077
DO	FE-Transformer	0.947	0.406	0.251	0.051
DO	FE-TCN-Transformer	0.947	0.403	0.254	0.052

Table S9 Performance of various models at the WSK site

Parameter	Model	NSE	RMSE	MAE	MAPE
TN	ARIMA	0.648	0.302	0.236	0.094
TN	STL-ARIMA	0.532	0.305	0.247	0.148
TN	XGBoost	0.721	0.268	0.205	0.085
TN	LSTM	0.640	0.305	0.236	0.094
TN	GRU	0.635	0.307	0.240	0.096
TN	Informer	0.586	0.327	0.254	0.102
TN	PatchTST	0.723	0.268	0.201	0.080
TN	Transformer	0.550	0.321	0.244	0.102
TN	TCN-Transformer	0.070	0.461	0.384	0.146
TN	FE-Transformer	0.849	0.171	0.114	0.044
TN	FE-TCN-Transformer	0.871	0.158	0.115	0.045
TP	ARIMA	0.522	0.024	0.018	0.167
TP	STL-ARIMA	0.531	0.024	0.017	0.163
TP	XGBoost	0.631	0.021	0.015	0.159
TP	LSTM	0.535	0.024	0.018	0.167
TP	GRU	0.528	0.024	0.018	0.171
TP	Informer	0.502	0.024	0.018	0.170
TP	PatchTST	0.644	0.021	0.014	0.137
TP	Transformer	0.624	0.024	0.017	0.154
TP	TCN-Transformer	0.679	0.022	0.014	0.129
TP	FE-Transformer	0.796	0.016	0.009	0.065
TP	FE-TCN-Transformer	0.820	0.015	0.009	0.064
NH3-N	ARIMA	0.275	0.113	0.074	0.944
NH3-N	STL-ARIMA	0.288	0.131	0.068	1.064
NH3-N	XGBoost	0.466	0.097	0.054	0.656
NH3-N	LSTM	0.299	0.111	0.075	1.035
NH3-N	GRU	0.292	0.112	0.076	1.014
NH3-N	Informer	0.406	0.102	0.060	0.714
NH3-N	PatchTST	0.453	0.098	0.052	0.566
NH3-N	Transformer	0.296	0.108	0.091	1.617
NH3-N	TCN-Transformer	0.552	0.086	0.044	0.393
NH3-N	FE-Transformer	0.629	0.065	0.028	0.461
NH3-N	FE-TCN-Transformer	0.754	0.053	0.028	0.410
COD _{Mn}	ARIMA	0.278	0.459	0.376	0.127
COD _{Mn}	STL-ARIMA	0.255	0.466	0.372	0.126
COD _{Mn}	XGBoost	0.613	0.336	0.255	0.088
COD _{Mn}	LSTM	0.256	0.466	0.370	0.128
COD _{Mn}	GRU	0.270	0.461	0.378	0.128
COD _{Mn}	Informer	0.644	0.322	0.241	0.079
COD _{Mn}	PatchTST	0.583	0.349	0.269	0.089

Parameter	Model	NSE	RMSE	MAE	MAPE
COD _{Mn}	Transformer	0.550	0.343	0.268	0.083
COD _{Mn}	TCN-Transformer	0.611	0.319	0.244	0.080
COD _{Mn}	FE-Transformer	0.770	0.203	0.146	0.047
COD _{Mn}	FE-TCN-Transformer	0.791	0.194	0.148	0.047
DO	ARIMA	0.718	0.838	0.689	0.152
DO	STL-ARIMA	0.723	0.830	0.664	0.146
DO	XGBoost	0.889	0.525	0.346	0.080
DO	LSTM	0.731	0.818	0.668	0.148
DO	GRU	0.738	0.807	0.648	0.143
DO	Informer	0.891	0.521	0.356	0.080
DO	PatchTST	0.897	0.507	0.348	0.080
DO	Transformer	0.895	0.603	0.436	0.136
DO	TCN-Transformer	0.908	0.564	0.400	0.109
DO	FE-Transformer	0.961	0.353	0.238	0.063
DO	FE-TCN-Transformer	0.963	0.340	0.240	0.062

Table S10a Distribution of relative error (RE) across sites and water-quality indicators

Site	Indicator	Sample size (n)	RE ≤ 1%	RE ≤ 5%	RE ≤ 10%	RE > 10%
LJ	COD _{Mn}	1688	22.2%	74.5%	93.1%	6.9%
LJ	DO	1688	15.9%	55.2%	78.2%	21.8%
LJ	NH ₃ -N	1688	2.5%	11.6%	22.6%	77.4%
LJ	TN	1688	14.7%	53.9%	83.5%	16.5%
LJ	TP	1688	10.7%	49.2%	75.1%	24.9%
WSK	COD _{Mn}	1693	16.2%	62.7%	88.2%	11.8%
WSK	DO	1693	19.6%	63.7%	84.7%	15.3%
WSK	NH ₃ -N	1693	2.6%	14.6%	27.8%	72.2%
WSK	TN	1693	15.1%	63.8%	92.3%	7.7%
WSK	TP	1693	11.9%	53.0%	82.4%	17.6%
ZJLQ	COD _{Mn}	1602	18.7%	72.7%	92.4%	7.6%
ZJLQ	DO	1602	20.3%	64.2%	83.0%	17.0%
ZJLQ	NH ₃ -N	1602	3.4%	15.7%	29.0%	71.0%
ZJLQ	TN	1602	18.0%	63.9%	86.6%	13.4%
ZJLQ	TP	1602	18.4%	65.7%	85.7%	14.3%
MHXJ	COD _{Mn}	1706	24.0%	80.8%	95.4%	4.6%
MHXJ	DO	1706	37.2%	92.9%	98.9%	1.1%
MHXJ	NH ₃ -N	1706	3.5%	17.0%	33.1%	66.9%
MHXJ	TN	1706	27.0%	81.3%	96.9%	3.1%
MHXJ	TP	1702	4.4%	25.8%	64.8%	35.2%

Table S10b Concentration-stratified prediction errors for NH₃-N and TP at the four monitoring sites.

Parameter	Site	Level	Observed range (mg/L)	n	MdAPE (%)	P90 APE (%)	RE >10% (%)
DO	All sites	Low	<5.00	2341	4.0	16.5	21.7
DO	All sites	Medium	≥5.00 to <6.00	1785	2.5	9.7	9.1
DO	All sites	High	≥6.00	2563	1.9	9.8	9.7
DO	LJ	Low	<5.00	396	5.5	23.0	26.8
DO	LJ	Medium	≥5.00 to <6.00	449	3.1	14.0	14.0
DO	LJ	High	≥6.00	843	4.5	17.4	23.6

DO	WSK	Low	<5.00	766	5.2	19.5	26.6
DO	WSK	Medium	≥5.00 to <6.00	495	2.7	9.5	7.7
DO	WSK	High	≥6.00	432	1.9	6.4	3.9
DO	ZJLQ	Low	<5.00	732	4.7	17.6	26.0
DO	ZJLQ	Medium	≥5.00 to <6.00	406	3.3	12.0	13.5
DO	ZJLQ	High	≥6.00	464	2.0	7.3	5.8
DO	MHXJ	Low	<5.00	447	2.0	5.5	1.6
DO	MHXJ	Medium	≥5.00 to <6.00	435	1.8	4.8	1.4
DO	MHXJ	High	≥6.00	824	1.1	3.1	0.7
COD_{Mn}	All sites	Low	≤4.00	5449	2.7	8.4	6.5
COD_{Mn}	All sites	Medium	>4.00 to ≤6.00	1234	3.2	11.6	12.8
COD_{Mn}	All sites	High	>6.00	6	19.7	30.9	83.3
COD _{Mn}	LJ	Low	≤4.00	932	2.4	7.2	4.1
COD _{Mn}	LJ	Medium	>4.00 to ≤6.00	755	2.9	10.1	10.3
COD _{Mn}	LJ	High	>6.00	1	32.6	32.6	100.0
COD _{Mn}	WSK	Low	≤4.00	1615	3.6	10.7	12.0
COD _{Mn}	WSK	Medium	>4.00 to ≤6.00	78	4.5	9.6	7.7
COD _{Mn}	WSK	High	>6.00	0	—	—	—
COD _{Mn}	ZJLQ	Low	≤4.00	1380	2.7	7.5	4.3
COD _{Mn}	ZJLQ	Medium	>4.00 to ≤6.00	217	4.5	16.8	26.3
COD _{Mn}	ZJLQ	High	>6.00	5	18.2	25.9	80.0
COD _{Mn}	MHXJ	Low	≤4.00	1522	2.3	6.9	4.1
COD _{Mn}	MHXJ	Medium	>4.00 to ≤6.00	184	2.8	9.9	9.2
COD _{Mn}	MHXJ	High	>6.00	0	—	—	—
NH₃-N	All sites	Low	≤0.50	6359	21.0	80.7	72.6
NH₃-N	All sites	Medium	>0.50 to ≤1.00	220	13.1	62.4	58.6
NH₃-N	All sites	High	>1.00	110	10.9	38.1	53.6
NH ₃ -N	LJ	Low	≤0.50	1652	23.5	77.8	77.7
NH ₃ -N	LJ	Medium	>0.50 to ≤1.00	36	29.4	66.0	63.9
NH ₃ -N	LJ	High	>1.00	0	—	—	—
NH ₃ -N	WSK	Low	≤0.50	1675	20.9	75.4	71.9
NH ₃ -N	WSK	Medium	>0.50 to ≤1.00	18	43.4	62.2	100.0
NH ₃ -N	WSK	High	>1.00	0	—	—	—
NH ₃ -N	ZJLQ	Low	≤0.50	1326	24.1	107.6	74.7
NH ₃ -N	ZJLQ	Medium	>0.50 to ≤1.00	166	10.9	52.6	53.0
NH ₃ -N	ZJLQ	High	>1.00	110	10.9	38.1	53.6
NH ₃ -N	MHXJ	Low	≤0.50	1706	16.9	64.9	66.9
NH ₃ -N	MHXJ	Medium	>0.50 to ≤1.00	0	—	—	—
NH ₃ -N	MHXJ	High	>1.00	0	—	—	—
TP	All sites	Low	≤0.10	1605	7.5	17.3	34.5
TP	All sites	Medium	>0.10 to ≤0.20	4526	4.6	13.3	17.9
TP	All sites	High	>0.20	558	6.2	20.8	32.8

TP	LJ	Low	≤ 0.10	111	5.0	13.8	24.3
TP	LJ	Medium	>0.10 to ≤ 0.20	1374	4.9	16.4	23.4
TP	LJ	High	>0.20	203	6.9	20.2	35.5
TP	WSK	Low	≤ 0.10	203	4.6	12.5	15.8
TP	WSK	Medium	>0.10 to ≤ 0.20	1417	4.5	12.8	16.7
TP	WSK	High	>0.20	73	7.5	23.0	39.7
TP	ZJLQ	Low	≤ 0.10	346	2.7	7.7	4.9
TP	ZJLQ	Medium	>0.10 to ≤ 0.20	980	3.1	12.4	13.7
TP	ZJLQ	High	>0.20	276	5.0	19.3	28.3
TP	MHXJ	Low	≤ 0.10	945	10.1	19.3	50.6
TP	MHXJ	Medium	>0.10 to ≤ 0.20	755	6.0	11.2	15.8
TP	MHXJ	High	>0.20	6	54.5	282.5	66.7
TN	All sites	Low	≤ 0.50	0	—	—	—
TN	All sites	Medium	>0.50 to ≤ 1.00	434	3.9	9.9	9.7
TN	All sites	High	>1.00	6255	3.2	10.0	10.2
TN	LJ	Low	≤ 0.50	0	—	—	—
TN	LJ	Medium	>0.50 to ≤ 1.00	0	—	—	—
TN	LJ	High	>1.00	1688	4.5	12.5	16.5
TN	WSK	Low	≤ 0.50	0	—	—	—
TN	WSK	Medium	>0.50 to ≤ 1.00	0	—	—	—
TN	WSK	High	>1.00	1693	3.7	9.2	7.7
TN	ZJLQ	Low	≤ 0.50	0	—	—	—
TN	ZJLQ	Medium	>0.50 to ≤ 1.00	84	5.8	12.9	16.7
TN	ZJLQ	High	>1.00	1518	3.2	12.0	13.2
TN	MHXJ	Low	≤ 0.50	0	—	—	—
TN	MHXJ	Medium	>0.50 to ≤ 1.00	350	3.5	9.4	8.0
TN	MHXJ	High	>1.00	1356	2.0	6.1	1.8

Table S10c Early-warning performance at the GB 3838—2002 Class II and Class III thresholds for all sites

pooled.

Parameter	Warning threshold	Limit (mg/L)	n	Observed events	Precision	Recall	F1
DO	Below Class II limit	6.00	6689	4126	0.954	0.957	0.956
DO	Below Class III limit	5.00	6689	2341	0.908	0.927	0.917
CODMn	Exceed Class II limit	4.00	6689	1240	0.843	0.842	0.842
CODMn	Exceed Class III limit	6.00	6689	6	0.333	0.167	0.222
NH ₃ -N	Exceed Class II limit	0.50	6689	330	0.864	0.791	0.826

NH ₃ -N	Exceed limit	Class	III	1.00	6689	110	0.849	0.718	0.778
TP	Exceed limit	Class	II	0.10	6689	5084	0.946	0.976	0.961
TP	Exceed limit	Class	III	0.20	6689	558	0.788	0.731	0.758
TN	Exceed limit	Class	II	0.50	6689	6689	1.000	1.000	1.000
TN	Exceed limit	Class	III	1.00	6689	6255	0.990	0.995	0.993

Table S11 Summary of one-sided paired Wilcoxon signed-rank tests for ablation gains relative to the baseline Transformer across 20 site-by-indicator tasks.

Metric	Variant	n	median	p	sig
dNSE	TCN-Transformer	20	0.0237	0.0318613	*
RMSE_gain	TCN-Transformer	20	0.0139	0.001576	**
dNSE	FE-Transformer	20	0.172	9.54e-07	***
RMSE_gain	FE-Transformer	20	0.097	9.54e-07	***
dNSE	FE-TCN-Transformer	20	0.207	9.54e-07	***
RMSE_gain	FE-TCN-Transformer	20	0.110	9.54e-07	***

Table S12 One-sided paired Wilcoxon signed-rank tests for the marginal contribution of TCN conditional on FE (FE–TCN–Transformer vs. FE–Transformer) across 20 site-by-indicator tasks.

Metric	n	median	q25	q75	improved	improved_ratio	p	sig
ddNSE = NSE(FE-TCN) - NSE(FE)	20	0.00412	0.00191	0.0255	20	1	9.54e-07	***
RMSE_gain = RMSE(FE) - RMSE(FE-TCN)	20	0.00564	0.00149	0.0120	20	1	9.54e-07	***

Table S13. Model complexity, computational cost, and predictive performance across different forecasting models.

Model	Model class	Median complexity	Total training time (s)	Median training time (s/task)	Median inference time (ms/sample)	Mean NSE	Mean RMSE	Mean MAE	Mean MAPE
ARIMA	Statistical	2	1.818	0.079	0.688	0.527	0.400	0.257	0.233
STL-ARIMA	Statistical hybrid	2	1.687	0.075	0.689	0.517	0.404	0.261	0.279
XGBoost	Tree ensemble	165	5.071	0.238	0.001	0.555	0.389	0.255	0.276
LSTM	Recurrent neural network	4,513	55.733	2.322	0.029	0.541	0.398	0.262	0.251
GRU	Recurrent neural network	3,393	189.619	8.452	0.065	0.550	0.394	0.258	0.258
Informer	Attention model	33,474	2278.958	114.537	3.924	0.548	0.387	0.247	0.278
PatchTST	Attention model	17,345	74.391	3.067	0.043	0.570	0.372	0.233	0.270
Transformer	Attention model	25,189	818.160	47.663	0.135	0.613	0.307	0.220	0.336
TCN-Transformer	Hybrid neural network	25,509	893.141	52.000	0.134	0.626	0.288	0.194	0.307
FE-Transformer	Feature-engineered attention model	25,639	882.653	50.239	0.135	0.816	0.189	0.113	0.132
FE-TCN-Transformer	Feature-engineered hybrid model	26,255	834.645	49.514	0.109	0.839	0.181	0.113	0.116

Table S14. Prediction performance of FE–TCN–Transformer with and without Hampel filtering and linear interpolation.

Station Indicator		NSE with preprocessing	NSE without preprocessing	RMSE with preprocessing	RMSE without preprocessing	MAE with preprocessing	MAE without preprocessing	MAPE with preprocessing	MAPE without preprocessing
LJ	TN	0.669	0.544	0.428	0.639	0.234	0.367	0.059	0.085
MHXJ	TN	0.990	0.985	0.059	0.071	0.043	0.051	0.030	0.036
WSK	TN	0.871	0.753	0.158	0.241	0.115	0.140	0.045	0.053
ZJLQ	TN	0.902	0.922	0.216	0.205	0.121	0.125	0.056	0.057
LJ	TP	0.741	0.543	0.020	0.032	0.012	0.018	0.077	0.106
MHXJ	TP	0.572	0.601	0.031	0.032	0.006	0.008	0.048	0.066
WSK	TP	0.820	0.775	0.015	0.017	0.009	0.011	0.064	0.078
ZJLQ	TP	0.888	0.861	0.018	0.023	0.009	0.011	0.056	0.065
LJ	NH ₃ –N	0.723	0.466	0.068	0.137	0.036	0.061	0.317	0.855
MHXJ	NH ₃ –N	0.895	0.597	0.022	0.051	0.016	0.022	0.332	0.734
WSK	NH ₃ –N	0.754	0.681	0.053	0.069	0.028	0.038	0.410	0.588
ZJLQ	NH ₃ –N	0.864	0.827	0.140	0.185	0.057	0.077	0.448	0.308
LJ	COD _{Mn}	0.773	0.570	0.224	0.382	0.148	0.222	0.036	0.051
MHXJ	COD _{Mn}	0.907	0.391	0.176	0.633	0.110	0.202	0.033	0.056
WSK	COD _{Mn}	0.791	0.724	0.194	0.234	0.148	0.172	0.047	0.054
ZJLQ	COD _{Mn}	0.870	0.817	0.221	0.285	0.138	0.162	0.038	0.045
LJ	DO	0.849	0.799	0.663	0.754	0.423	0.503	0.080	0.096
MHXJ	DO	0.991	0.988	0.164	0.183	0.115	0.121	0.020	0.021

WSK	DO	0.963	0.964	0.340	0.325	0.240	0.218	0.062	0.054
ZJLQ	DO	0.947	0.925	0.403	0.490	0.254	0.311	0.052	0.061
Mean	All tasks	0.839	0.737	0.181	0.249	0.113	0.142	0.116	0.173

Table S15. Feature-subset sensitivity of the FE–TCN–Transformer across 20 site–target tasks. F1 represents the current feature set, F2 target history only, F3 F1 without rainfall, and F4 all available variables. Values are reported using NSE, RMSE, MAE, and MAPE.

Feature set	Target	Site code	NSE	RMSE	MAE	MAPE
F1	COD _{Mn}	LJ	0.773	0.224	0.148	0.036
F1	COD _{Mn}	MHXJ	0.907	0.176	0.110	0.033
F1	COD _{Mn}	WSK	0.791	0.194	0.148	0.047
F1	COD _{Mn}	ZJLQ	0.870	0.221	0.138	0.038
F1	DO	LJ	0.849	0.663	0.423	0.080
F1	DO	MHXJ	0.991	0.164	0.115	0.020
F1	DO	WSK	0.963	0.340	0.240	0.062
F1	DO	ZJLQ	0.947	0.403	0.254	0.052
F1	NH ₃ –N	LJ	0.723	0.068	0.036	0.317
F1	NH ₃ –N	MHXJ	0.895	0.022	0.016	0.332
F1	NH ₃ –N	WSK	0.754	0.053	0.028	0.410
F1	NH ₃ –N	ZJLQ	0.864	0.140	0.057	0.448
F1	TN	LJ	0.669	0.428	0.234	0.059
F1	TN	MHXJ	0.990	0.059	0.043	0.030
F1	TN	WSK	0.871	0.158	0.115	0.045
F1	TN	ZJLQ	0.902	0.216	0.121	0.056
F1	TP	LJ	0.741	0.020	0.012	0.077
F1	TP	MHXJ	0.572	0.031	0.006	0.048
F1	TP	WSK	0.820	0.015	0.009	0.064
F1	TP	ZJLQ	0.888	0.018	0.009	0.056

Feature set	Target	Site code	NSE	RMSE	MAE	MAPE
F2	COD _{Mn}	LJ	0.541	0.353	0.235	0.058
F2	COD _{Mn}	MHXJ	0.825	0.249	0.154	0.047
F2	COD _{Mn}	WSK	0.629	0.311	0.236	0.076
F2	COD _{Mn}	ZJLQ	0.730	0.338	0.211	0.058
F2	DO	LJ	0.734	0.939	0.598	0.202
F2	DO	MHXJ	0.976	0.271	0.182	0.032
F2	DO	WSK	0.915	0.551	0.384	0.108
F2	DO	ZJLQ	0.899	0.568	0.360	0.075
F2	NH ₃ -N	LJ	0.523	0.099	0.056	0.666
F2	NH ₃ -N	MHXJ	0.829	0.053	0.032	0.596
F2	NH ₃ -N	WSK	0.540	0.087	0.047	0.482
F2	NH ₃ -N	ZJLQ	0.738	0.215	0.080	0.539
F2	TN	LJ	0.581	0.555	0.348	0.088
F2	TN	MHXJ	0.980	0.093	0.067	0.046
F2	TN	WSK	0.762	0.249	0.182	0.072
F2	TN	ZJLQ	0.831	0.317	0.173	0.081
F2	TP	LJ	0.485	0.031	0.019	0.123
F2	TP	MHXJ	0.435	0.043	0.009	0.084
F2	TP	WSK	0.709	0.021	0.013	0.117
F2	TP	ZJLQ	0.803	0.025	0.014	0.090
F3	COD _{Mn}	LJ	0.548	0.351	0.234	0.057
F3	COD _{Mn}	MHXJ	0.823	0.251	0.155	0.048
F3	COD _{Mn}	WSK	0.632	0.310	0.235	0.076

Feature set	Target	Site code	NSE	RMSE	MAE	MAPE
F3	COD _{Mn}	ZJLQ	0.742	0.331	0.210	0.058
F3	DO	LJ	0.732	0.943	0.595	0.187
F3	DO	MHXJ	0.977	0.263	0.178	0.031
F3	DO	WSK	0.916	0.547	0.377	0.109
F3	DO	ZJLQ	0.897	0.572	0.363	0.076
F3	NH ₃ -N	LJ	0.513	0.100	0.056	0.679
F3	NH ₃ -N	MHXJ	0.796	0.058	0.036	0.657
F3	NH ₃ -N	WSK	0.566	0.084	0.048	0.542
F3	NH ₃ -N	ZJLQ	0.736	0.216	0.082	0.545
F3	TN	LJ	0.569	0.563	0.344	0.087
F3	TN	MHXJ	0.980	0.093	0.067	0.046
F3	TN	WSK	0.763	0.249	0.182	0.072
F3	TN	ZJLQ	0.833	0.315	0.171	0.080
F3	TP	LJ	0.506	0.031	0.019	0.123
F3	TP	MHXJ	0.404	0.044	0.009	0.093
F3	TP	WSK	0.697	0.022	0.013	0.117
F3	TP	ZJLQ	0.805	0.025	0.014	0.090
F4	COD _{Mn}	LJ	0.564	0.344	0.231	0.057
F4	COD _{Mn}	MHXJ	0.829	0.246	0.152	0.047
F4	COD _{Mn}	WSK	0.632	0.310	0.235	0.076
F4	COD _{Mn}	ZJLQ	0.739	0.332	0.213	0.059
F4	DO	LJ	0.735	0.936	0.593	0.201
F4	DO	MHXJ	0.977	0.265	0.183	0.032

Feature set	Target	Site code	NSE	RMSE	MAE	MAPE
F4	DO	WSK	0.919	0.538	0.378	0.109
F4	DO	ZJLQ	0.898	0.569	0.361	0.076
F4	NH ₃ -N	LJ	0.515	0.099	0.057	0.687
F4	NH ₃ -N	MHXJ	0.728	0.066	0.045	0.895
F4	NH ₃ -N	WSK	0.557	0.085	0.049	0.583
F4	NH ₃ -N	ZJLQ	0.746	0.211	0.084	0.619
F4	TN	LJ	0.527	0.589	0.344	0.087
F4	TN	MHXJ	0.980	0.093	0.068	0.047
F4	TN	WSK	0.769	0.246	0.179	0.071
F4	TN	ZJLQ	0.836	0.313	0.169	0.077
F4	TP	LJ	0.521	0.030	0.019	0.123
F4	TP	MHXJ	0.344	0.046	0.010	0.105
F4	TP	WSK	0.705	0.022	0.013	0.120
F4	TP	ZJLQ	0.807	0.025	0.013	0.089

Table S16. Robustness of the FE–TCN–Transformer under additional missing-data and sensor-outage scenarios across 20 site-by-indicator forecasting tasks.

Indicator	Station	Baseline NSE	MCAR 10% NSE	MCAR 20% NSE	24-h block NSE	72-h block NSE	Baseline RMSE	MCAR 10% RMSE	MCAR 20% RMSE	24-h block RMSE	72-h block RMSE
TN	MHXJ	0.990	0.979	0.977	0.976	0.977	0.059	0.095	0.100	0.101	0.099
TN	LJ	0.669	0.472	0.433	0.455	-0.368	0.428	0.623	0.646	0.633	1.004
TN	ZJLQ	0.902	0.828	0.800	0.783	0.809	0.216	0.319	0.345	0.359	0.337
TN	WSK	0.871	0.737	0.714	0.740	0.700	0.158	0.262	0.273	0.261	0.280
TP	MHXJ	0.572	0.264	0.267	0.264	0.259	0.031	0.048	0.049	0.048	0.049
TP	LJ	0.741	0.487	0.489	0.469	0.390	0.020	0.031	0.031	0.032	0.034
TP	ZJLQ	0.888	0.791	0.777	0.690	0.646	0.018	0.026	0.027	0.032	0.034
TP	WSK	0.820	0.626	0.653	0.696	0.690	0.015	0.024	0.023	0.022	0.022
NH3N	MHXJ	0.895	0.795	0.778	0.785	0.792	0.022	0.058	0.060	0.059	0.058
NH3N	LJ	0.723	0.509	0.499	0.414	0.429	0.068	0.100	0.101	0.109	0.108
NH3N	ZJLQ	0.864	0.694	0.675	0.631	0.576	0.140	0.232	0.239	0.255	0.273
NH3N	WSK	0.754	0.512	0.451	0.464	0.453	0.053	0.089	0.095	0.093	0.094
COD _{Mn}	MHXJ	0.907	0.804	0.802	0.803	0.588	0.176	0.264	0.265	0.264	0.382
COD _{Mn}	LJ	0.773	0.540	0.487	0.523	0.441	0.224	0.354	0.374	0.360	0.390

Indicator	Station	Baseline NSE	MCAR 10% NSE	MCAR 20% NSE	24-h block NSE	72-h block NSE	Baseline RMSE	MCAR 10% RMSE	MCAR 20% RMSE	24-h block RMSE	72-h block RMSE
COD _{Mn}	ZJLQ	0.870	0.731	0.694	0.690	0.648	0.221	0.337	0.360	0.362	0.386
COD _{Mn}	WSK	0.791	0.560	0.537	0.538	0.567	0.194	0.339	0.348	0.348	0.337
DO	MHXJ	0.991	0.975	0.974	0.971	0.959	0.164	0.274	0.280	0.295	0.350
DO	LJ	0.849	0.721	0.676	0.683	0.490	0.663	0.961	1.036	1.024	1.300
DO	ZJLQ	0.947	0.885	0.886	0.884	0.893	0.403	0.606	0.602	0.607	0.582
DO	WSK	0.963	0.905	0.885	0.890	0.888	0.340	0.583	0.641	0.625	0.633

Robustness of FE–TCN–Transformer under additional missing-data scenarios. Baseline represents the original test results. MCAR denotes randomly missing inputs, while 24 h and 72 h blocks simulate continuous sensor outages. Missing-condition results are the means of three random-seed runs without model retraining. Higher NSE and lower RMSE indicate better performance.

Table S17. Decomposition configurations and common forecasting protocol.

Scope	Item	Prespecified setting
Decomposition	Fourier–EMA	Deterministic Fourier seasonal terms and a causal shifted EMA trend; the main-model implementation was unchanged.
Decomposition	STL	Causal multi-seasonal STL; rolling history 168; minimum history 96; periods 6 and 42; seasonal windows 7 and 13; one outer iteration; 12-point local linear trend extrapolation.
Decomposition	VMD	Rolling history 168; five modes; $\alpha = 2000$; $\tau = 0$; no fixed DC mode; uniform initialization; tolerance 10^{-6} ; 12-point trend extrapolation.
Decomposition	Wavelet	Daubechies-4 wavelet; level 3; symmetric boundary extension; D3 and D2 oscillatory band; ridge-stabilized AR(6) endpoint forecast.
Tasks	Site–target combinations	Four sites $\times$ five water-quality variables = 20 paired tasks.
Partition	Chronological split	60% training, 20% validation, and 20% test, using identical label-index boundaries for all methods.
Forecast	Input and horizon	48 observations (8 days at 4-h resolution) to predict the next observation (4 h).
Inputs	Covariates and channels	Identical task-specific covariates; each decomposition supplied one trend and one seasonal or oscillatory channel, with the innovation residual as the prediction target.
Model	Architecture	Identical TCN–Transformer with model dimension 32, four attention heads, two encoder layers, one decoder layer, feed-forward dimension 32, and dropout 0.1.
Optimization	Training	Batch size 64; Adam learning rate 0.001; at most 80 epochs; validation-RMSE checkpoint selection; early-stopping patience 12.
Causality	Rolling origin	Features at time t used observations only through $t-1$; scaling parameters were estimated from the training interval only.
Statistics	Comparison protocol	Friedman tests across four methods and two-sided paired Wilcoxon tests with Holm correction; each of the 20 site–target tasks contributed one result per method; 10,000 task-level bootstrap resamples.

Note. All implementations used two output channels and no test-set tuning. During the first 95 origins, the shared warm-up used the previous observation for trend and zero for seasonality.

Table S18. Overall predictive performance of the four decomposition methods across 20 site–target tasks.

Method	NSE	RMSE	MAE	MAPE
Fourier–EMA	0.714	0.270	0.169	0.188
STL	0.536	0.337	0.233	0.333
VMD	0.691	0.285	0.183	0.193
Wavelet	0.665	0.289	0.181	0.179

Note. Values are means across 20 paired site–target tasks. MAPE is reported as a decimal fraction.

Bold values indicate the best result for each metric. Higher NSE and lower RMSE, MAE, and MAPE indicate better performance. Because all four methods were rerun under the common optimization protocol in Table S17, the absolute values in Tables S18–S20 are specific to this controlled comparison and are not directly comparable with the primary benchmark results in Tables S5–S13.

Table S19. Task-level predictive performance for the four decomposition methods across 20 site–target forecasting tasks.

Method	Target	Site	NSE	RMSE	MAE	MAPE
Fourier–EMA	TN	LJ	0.492	0.612	0.354	0.090
Fourier–EMA	TN	MHXJ	0.980	0.093	0.068	0.047
Fourier–EMA	TN	WSK	0.764	0.248	0.183	0.073
Fourier–EMA	TN	ZJLQ	0.836	0.313	0.170	0.079
Fourier–EMA	TP	LJ	0.519	0.030	0.019	0.122
Fourier–EMA	TP	MHXJ	0.266	0.048	0.011	0.117
Fourier–EMA	TP	WSK	0.717	0.021	0.013	0.117
Fourier–EMA	TP	ZJLQ	0.805	0.025	0.014	0.091
Fourier–EMA	NH ₃ –N	LJ	0.512	0.100	0.057	0.680
Fourier–EMA	NH ₃ –N	MHXJ	0.794	0.058	0.036	0.664
Fourier–EMA	NH ₃ –N	WSK	0.572	0.084	0.047	0.535
Fourier–EMA	NH ₃ –N	ZJLQ	0.737	0.215	0.078	0.495
Fourier–EMA	COD _{Mn}	LJ	0.555	0.348	0.230	0.056
Fourier–EMA	COD _{Mn}	MHXJ	0.827	0.248	0.154	0.047
Fourier–EMA	COD _{Mn}	WSK	0.628	0.312	0.238	0.077
Fourier–EMA	COD _{Mn}	ZJLQ	0.736	0.334	0.212	0.059
Fourier–EMA	DO	LJ	0.737	0.934	0.595	0.195
Fourier–EMA	DO	MHXJ	0.977	0.265	0.180	0.031
Fourier–EMA	DO	WSK	0.920	0.535	0.373	0.107
Fourier–EMA	DO	ZJLQ	0.898	0.568	0.359	0.075
STL	TN	LJ	0.248	0.743	0.469	0.120
STL	TN	MHXJ	0.966	0.121	0.092	0.065
STL	TN	WSK	0.672	0.292	0.216	0.085
STL	TN	ZJLQ	0.703	0.420	0.282	0.145
STL	TP	LJ	0.295	0.037	0.024	0.160
STL	TP	MHXJ	-0.323	0.066	0.016	0.185
STL	TP	WSK	0.524	0.027	0.018	0.154
STL	TP	ZJLQ	0.636	0.034	0.022	0.144

Method	Target	Site	NSE	RMSE	MAE	MAPE
STL	NH ₃ -N	LJ	0.235	0.125	0.082	1.119
STL	NH ₃ -N	MHXJ	0.547	0.085	0.041	0.527
STL	NH ₃ -N	WSK	0.330	0.104	0.065	0.822
STL	NH ₃ -N	ZJLQ	0.558	0.279	0.158	2.129
STL	COD _{Mn}	LJ	0.333	0.426	0.311	0.077
STL	COD _{Mn}	MHXJ	0.669	0.343	0.232	0.071
STL	COD _{Mn}	WSK	0.461	0.375	0.290	0.093
STL	COD _{Mn}	ZJLQ	0.594	0.414	0.287	0.081
STL	DO	LJ	0.617	1.126	0.810	0.407
STL	DO	MHXJ	0.972	0.289	0.216	0.038
STL	DO	WSK	0.883	0.646	0.467	0.134
STL	DO	ZJLQ	0.808	0.782	0.560	0.112
VMD	TN	LJ	0.531	0.588	0.360	0.091
VMD	TN	MHXJ	0.979	0.096	0.070	0.047
VMD	TN	WSK	0.749	0.256	0.189	0.075
VMD	TN	ZJLQ	0.809	0.337	0.199	0.095
VMD	TP	LJ	0.508	0.031	0.018	0.119
VMD	TP	MHXJ	0.260	0.049	0.010	0.101
VMD	TP	WSK	0.672	0.023	0.014	0.125
VMD	TP	ZJLQ	0.770	0.027	0.015	0.101
VMD	NH ₃ -N	LJ	0.448	0.106	0.059	0.704
VMD	NH ₃ -N	MHXJ	0.867	0.047	0.018	0.213
VMD	NH ₃ -N	WSK	0.501	0.090	0.051	0.548
VMD	NH ₃ -N	ZJLQ	0.695	0.232	0.107	0.956
VMD	COD _{Mn}	LJ	0.534	0.356	0.238	0.058
VMD	COD _{Mn}	MHXJ	0.798	0.268	0.167	0.052
VMD	COD _{Mn}	WSK	0.579	0.332	0.251	0.081
VMD	COD _{Mn}	ZJLQ	0.665	0.376	0.236	0.065
VMD	DO	LJ	0.691	1.012	0.653	0.199
VMD	DO	MHXJ	0.976	0.267	0.184	0.032

Method	Target	Site	NSE	RMSE	MAE	MAPE
VMD	DO	WSK	0.899	0.599	0.416	0.119
VMD	DO	ZJLQ	0.884	0.606	0.397	0.082
Wavelet	TN	LJ	0.468	0.626	0.388	0.099
Wavelet	TN	MHXJ	0.979	0.095	0.068	0.046
Wavelet	TN	WSK	0.762	0.249	0.185	0.072
Wavelet	TN	ZJLQ	0.819	0.328	0.183	0.086
Wavelet	TP	LJ	0.273	0.037	0.021	0.136
Wavelet	TP	MHXJ	0.059	0.055	0.009	0.104
Wavelet	TP	WSK	0.671	0.023	0.014	0.125
Wavelet	TP	ZJLQ	0.784	0.026	0.014	0.097
Wavelet	NH ₃ -N	LJ	0.326	0.117	0.061	0.646
Wavelet	NH ₃ -N	MHXJ	0.835	0.052	0.021	0.266
Wavelet	NH ₃ -N	WSK	0.592	0.082	0.047	0.513
Wavelet	NH ₃ -N	ZJLQ	0.691	0.233	0.096	0.677
Wavelet	COD _{Mn}	LJ	0.513	0.364	0.243	0.060
Wavelet	COD _{Mn}	MHXJ	0.779	0.280	0.171	0.053
Wavelet	COD _{Mn}	WSK	0.604	0.322	0.243	0.078
Wavelet	COD _{Mn}	ZJLQ	0.714	0.348	0.219	0.061
Wavelet	DO	LJ	0.669	1.047	0.655	0.233
Wavelet	DO	MHXJ	0.963	0.335	0.218	0.038
Wavelet	DO	WSK	0.913	0.558	0.387	0.104
Wavelet	DO	ZJLQ	0.887	0.601	0.382	0.081

Table S20. Non-parametric statistical comparison of decomposition methods across 20 paired tasks.

Panel A. Friedman tests across the four methods

Metric	χ^2	df	p value	Kendall's W
NSE	44.10	3	1.44×10^{-9}	0.735
RMSE	44.10	3	1.44×10^{-9}	0.735
MAE	43.02	3	2.44×10^{-9}	0.717
MAPE	35.94	3	7.71×10^{-8}	0.599

Panel B. Fourier–EMA versus alternative paired Wilcoxon comparisons

Metric	Comparison	Median effect (95% bootstrap CI)	W	Holm-adjusted p	Fourier–EMA/tie/alternative wins
NSE	Fourier–EMA vs STL	-0.1684 (-0.2229, -0.1261)	0.0	5.72×10^{-6}	20/0/0
NSE	Fourier–EMA vs VMD	-0.0239 (-0.0437, -0.0123)	32.0	0.00486	18/0/2
NSE	Fourier–EMA vs Wavelet	-0.0228 (-0.0461, -0.0129)	19.0	0.00117	18/0/2
RMSE	Fourier–EMA vs STL	-0.0539 (-0.0876, -0.0247)	0.0	5.72×10^{-6}	20/0/0
RMSE	Fourier–EMA vs VMD	-0.0071 (-0.0199, -0.0019)	26.0	0.00199	18/0/2
RMSE	Fourier–EMA vs Wavelet	-0.0139 (-0.0178, -0.0045)	10.0	1.64×10^{-4}	18/0/2

Note. The independent unit was the site–target task ($n = 20$). For NSE, the paired effect is alternative minus Fourier–EMA; for RMSE, it is Fourier–EMA minus alternative. Thus, a negative effect favors Fourier–EMA for all reported contrasts. Confidence intervals used 10,000 task-level bootstrap resamples. Holm correction was applied separately within each metric.

References

- Bai, S., Kolter, J.Z., Koltun, V., 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. <https://doi.org/10.48550/arXiv.1803.01271>
- Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time Series Analysis: Forecasting and Control, 5th ed. Wiley, Hoboken, NJ.
- Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. <https://doi.org/10.1145/2939672.2939785>
- Cho, K., Merriënboer, B. van, Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. <https://doi.org/10.48550/arXiv.1406.1078>
- Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I., 1990. STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *J. Off. Stat.* 6, 3–73.
- De Livera, A.M., Hyndman, R.J., Snyder, R.D., 2011. Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. *J. Am. Stat. Assoc.* 106, 1513–1527. <https://doi.org/10.1198/jasa.2011.tm09771>
- Gardner, E.S., 2006. Exponential smoothing: The state of the art—Part II. *Int. J. Forecast.* 22, 637–666. <https://doi.org/10.1016/j.ijforecast.2006.03.005>
- Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. *Neural Comput.* 9, 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J., 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. <https://doi.org/10.48550/arXiv.2211.14730>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need. <https://doi.org/10.48550/arXiv.1706.03762>
- Zhou, H., Zhang, Shanghang, Peng, J., Zhang, Shuai, Li, J., Xiong, H., Zhang, W., 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. <https://doi.org/10.48550/arXiv.2012.07436>