Pol. J. Environ. Stud. Vol. XX, No. X (XXXX), 1-21
DOI: *10.15244/pjoes/213858* ONLINE PUBLICATION DATE: 2026-02-03

Original Research

Environmental Assessment of $PM_{2.5}$ Concentration Patterns Through TC-MixerInformer Modeling: Cross-Regional Analysis of Shanghai and London

Zihan Wang, Changgui Gu*

Business School, University of Shanghai for Science and Technology, Shanghai, 200093, China

Received: 24 July 2025
Accepted: 02 November 2025

Abstract

Urban air quality prediction faces dual challenges: complex pollution processes and cross-regional diversity. Traditional methods and deep learning techniques often struggle with non-stationary time series data and fail to adapt to unique urban pollution patterns. Additionally, existing models face computational limitations when processing long sequences. Although the Informer model improves computational efficiency through ProbSparse self-attention mechanisms, its accuracy decreases with longer prediction horizons. This limitation stems from inadequate adaptability to non-stationary environmental changes and insufficient capture of temporal variations inherent in urban air quality data. Meanwhile, cities in developed and developing countries exhibit fundamentally different pollution mechanisms, challenging models' cross-regional generalization capabilities. To tackle these two challenges, this study proposes TC-MixerInformer, combining Reversible Instance Normalization (RevIN) with a Temporal-Channel Mixer (TCMixer) module. Validation using Shanghai (Jing'an) and London (North Kensington) monitoring stations demonstrates excellent performance in both short-term (1-12 h) and long-term (24-168 h) predictions, with 8%-54% error reductions compared to baseline models. The model effectively handles Shanghai's complex pollution patterns (2-378 μg/m³, mean 28.23 μg/m³) and London's lower concentrations (0-121.56 μg/m³, mean 8.08 μg/m³). RevIN addresses time series non-stationarity while TCMixer enhances multi-scale feature expression, maintaining stable performance across different time scales. The model shows particular advantages in predicting extreme pollution events, especially capturing substantial peaks approaching 350+ μg/m³ in Shanghai. Our research provides a new technical approach for addressing scale diversity and temporal non-stationarity in urban air quality prediction.

Keywords: urban air quality prediction, TC-MixerInformer model, $PM_{2.5}$ forecasting, time series non-stationarity

*e-mail: gu_changgui@163.com

Introduction

Urban air quality has emerged as one of the most pressing environmental challenges of the 21st century, with fine particulate matter ($PM_{2.5}$) serving as a critical indicator of atmospheric pollution and public health risks. $PM_{2.5}$ is defined as particulate matter with an aerodynamic diameter smaller than 2.5 micrometers, which poses significant threats to human health due to its ability to penetrate deep into the respiratory system and enter the bloodstream, leading to various diseases, including respiratory and cardiovascular disorders [1, 2]. The World Health Organization estimates that ambient air pollution, primarily driven by $PM_{2.5}$, causes approximately 7 million premature deaths globally each year [3]. This alarming statistic underscores the urgent need for accurate, reliable, and timely $PM_{2.5}$ concentration prediction systems to support environmental management and public health protection strategies.

However, $PM_{2.5}$ concentration prediction is far more complex than conventionally understood, rooted in multifaceted interactions between anthropogenic activities and natural systems. Recent research has revealed intricate relationships between land resource allocation and environmental pollution patterns. Studies on China's land resource misallocation demonstrate significant spatial spillover effects, where pollution in one region propagates to neighboring areas through atmospheric transport, creating complex interdependencies that traditional prediction models often fail to capture [4]. The dynamic impacts of ecosystem service supply and demand on air quality have been extensively documented in large watershed systems such as the Yellow River Basin, where industrial emissions, urbanization pressures, and natural ecological processes interact nonlinearly [5, 6]. Fine-scale analyses reveal that production-living-ecological function coupling varies significantly across spatial scales and temporal periods, fundamentally influencing local and regional air quality patterns [7].

This complexity is further compounded by fundamental atmospheric chemistry and physics mechanisms. $PM_{2.5}$ concentrations result from nonlinear interactions between primary emissions (directly emitted particles from combustion sources) and secondary aerosol formation through photochemical reactions involving precursor gases (SO_2, NOx, VOCs). The reaction rates and pathways vary significantly across climate zones: subtropical regions like Shanghai experience higher photochemical activity due to elevated solar radiation and temperature, leading to more rapid secondary aerosol formation, while temperate maritime climates like London exhibit slower photochemical processes but a stronger influence from synoptic-scale meteorological systems [8]. Moreover, meteorological conditions (temperature inversions, boundary layer height, wind patterns) exhibit nonlinear relationships with pollutant dispersion, where small changes in atmospheric stability can trigger order-of-magnitude variations in surface concentrations. These multiscale nonlinear interactions explain why traditional linear models fail and why effective $PM_{2.5}$ prediction models must transcend simple emission-concentration relationships to incorporate socioeconomic drivers, land use dynamics, ecosystem service flows, and atmospheric processes – a requirement that poses substantial challenges for conventional modeling approaches but creates opportunities for deep learning methods. Accurate prediction of air pollutant concentrations has thus become a critical challenge in environmental science. Traditional statistical models and numerical simulation methods often struggle to characterize the dynamic patterns of $PM_{2.5}$ due to its complex formation mechanisms and significant spatiotemporal variability [9].

Within this wave of technological development, the application of deep learning in air quality prediction has evolved significantly from early architectures to more sophisticated approaches. Initial studies primarily employed Recurrent Neural Network (RNN) architectures [10-12], with bidirectional LSTM networks achieving remarkable success in short-term prediction tasks [2, 13]. Subsequently, the integration of Convolutional Neural Networks (CNN) with LSTM enhanced feature extraction capabilities, enabling better capture of spatial and temporal patterns in air quality data. However, these models faced critical limitations when handling long-sequence predictions, including gradient vanishing and computational inefficiency. To overcome these limitations, the introduction of Transformers brought revolutionary breakthroughs to time series prediction tasks through self-attention mechanisms, enabling parallel processing of sequential data and first demonstrating significant success in natural language processing [14, 15]. When applied to air quality prediction, these models showed remarkable advantages in capturing long-term dependencies among pollutants [16]. Nevertheless, standard Transformers encountered severe challenges related to quadratic computational complexity when processing long sequences, making them computationally prohibitive for practical air quality prediction applications.

To address the computational bottleneck of Transformers, the Informer model proposed by Zhou et al. [17] effectively resolved this issue through the ProbSparse self-attention mechanism, reducing complexity from $O(L^2)$ to $O(LlogL)$. This innovative approach, combined with self-attention distillation to minimize memory consumption and a generative decoder architecture, makes Informer particularly suitable for air quality prediction tasks that require processing large amounts of historical data to capture seasonal and cyclical pollution patterns. Despite these significant computational advantages, preliminary experiments conducted in this study indicate that Informer still exhibits considerable accuracy degradation with increasing prediction horizons when handling non-stationary time series data – an inherent

characteristic of urban air quality patterns that fluctuate due to seasonal factors, economic activities, and policy implementations [18].

This limitation is not an isolated case, but rather reflects a common challenge faced by current deep learning architectures in time series prediction. Through systematic analysis of mainstream models, including LSTM, Transformer, and their variants, we have identified that the performance bottlenecks of these methods in cross-regional air quality prediction stem from three fundamental theoretical deficiencies. These deficiencies not only explain the accuracy degradation phenomenon observed in Informer, but also reveal the root causes why existing techniques struggle to achieve robust cross-domain generalization: (1) Mathematical Non-Stationarity and Distribution Shift [19, 20]: Urban $PM_{2.5}$ concentrations exhibit time-varying statistical properties – shifting mean $\mu(t)$ and variance $\sigma^2(t)$ – that violate the stationarity assumption underlying traditional neural network architectures. Standard normalization techniques (e.g., BatchNorm) compute global statistics across the training data and apply fixed transformations during inference, resulting in systematic prediction errors when the test data diverge from the training distributions. This manifests acutely in cross-regional applications: a model trained on high-concentration urban environments learns feature representations optimized for that distribution, systematically producing biased predictions when applied to cities with fundamentally different pollution characteristics. The mathematical root cause lies in the non-reversible nature of standard normalization: while it standardizes training data, it cannot adapt to new distributions without retraining, leading to substantial performance degradation when applied to cities with significantly different baseline concentrations and variability patterns.

(2) Attention Mechanism Limitations in Long-Sequence Modeling: While Transformers theoretically capture long-range dependencies through self-attention, the softmax-based attention mechanism produces over-smoothed representations for extended prediction horizons. This occurs because attention weights computed via $softmax\left(\frac{QK^T}{\sqrt{d}}\right)$ distribute probability mass across all time steps, causing distant historical information to receive near-zero weights that are numerically unstable during backpropagation. As sequence length L increases, the effective receptive field shrinks due to gradient dilution, preventing the model from learning long-term pollution evolution patterns. This theoretical limitation manifests as substantial accuracy degradation in long-term predictions: existing models show markedly increasing errors as prediction horizons extend from hours to days. The ProbSparse attention mechanism reduces computational complexity from $O(L^2)$ to $O(L\log L)$ but does not address the fundamental gradient dilution problem, explaining why computational efficiency improvements do not translate to accuracy gains in long-term predictions.

(3) Cross-Domain Generalization Paradox [21]: Deep learning models face a fundamental trade-off between specialization (overfitting to source domain characteristics) and generalization (underfitting to target domain patterns). This paradox is particularly severe in air quality prediction due to the heterogeneity of urban pollution mechanisms. Models must simultaneously learn: (a) universal temporal dynamics (diurnal cycles, meteorological influences) that transfer across cities, and (b) city-specific pollution signatures (emission source structures, topographical effects) that require local adaptation. Standard training procedures optimize for average performance across the training distribution, producing models that excel within their training domain but fail to extrapolate to new environments. The theoretical challenge lies in the absence of explicit mechanisms to disentangle universal patterns from domain-specific characteristics during representation learning. Without such disentanglement, learned features conflate transferable temporal dynamics with non-transferable regional baselines, causing systematic prediction biases when applied to new cities. This explains why existing models trained on single-city datasets exhibit limited cross-regional applicability despite achieving high accuracy within their training domains.

These theoretical insights reveal that improving cross-regional air quality prediction requires architectural innovations that explicitly address distribution shift (through adaptive normalization), long-sequence modeling limitations (through alternative attention mechanisms), and cross-domain generalization (through disentangled representation learning).

Based on a comprehensive understanding of the aforementioned challenges, this study proposes TC-MixerInformer, a novel deep learning architecture that combines advanced normalization techniques with efficient attention mechanisms for robust cross-regional $PM_{2.5}$ prediction.

To ensure the practicality and reliability of the model, this work conducts systematic cross-regional validation using data from Shanghai (Jing'an) and London (North Kensington), which represent different pollution mechanisms, with a comprehensive evaluation across multiple prediction horizons ranging from 1 h to 7 days. This comprehensive validation approach ensures the robustness and applicability of the model under different geographical and climatic conditions.

Specifically, the main innovations of this work lie in two key technical contributions that directly address the identified limitations. First, the integration of the Reversible Instance Normalization (RevIN) mechanism [20] specifically addresses the distribution shift challenges that limit cross-regional model transferability. RevIN performs instance-level normalization to eliminate regional baseline differences and seasonal

variations while preserving temporal dynamics through mathematical reversibility, enabling the model to maintain accuracy across different geographical regions with varying pollution characteristics. Second, the development of the Temporal-Channel Mixer (TCMixer) module [22] revolutionizes temporal and feature dependency modeling through a unified mixing strategy that captures cross-dimensional interactions. This innovation enhances multi-scale feature representation and improves long-term prediction accuracy. The integration of RevIN and TCMixer innovations produces a model that maintains high prediction accuracy across multiple time scales while demonstrating exceptional cross-regional adaptability, providing a robust solution for global air quality monitoring applications with significant implications for environmental management and public health protection.

Materials and Methods

Study Areas and Data Sources

This study utilizes air quality and meteorological data from two representative urban monitoring stations across different climate zones and pollution characteristics (Fig. 1). The Shanghai Jing'an station (31.23°N, 121.46°E) represents a typical subtropical megacity environment with complex industrial-urban mixed pollution sources. The London North Kensington station (51.52°N, 0.21°W) serves as a temperate climate reference with predominantly traffic-related emissions.

Research data include two parts: (1) Air Quality Monitoring Data: Air quality data for the Jing'an (Shanghai) monitoring station were sourced from Wang Xiaolei's research group and the National Urban Air Quality Real-time Publishing Platform of China's Environmental Monitoring Center. Data for the North Kensington (London) monitoring station were obtained from the London Air Quality Network official database (https://www.londonair.org.uk/). Monitored parameters include particulate matter ($PM_{2.5}$, PM_{10}) and gaseous pollutants (SO_2, NO_2, O_3, CO) concentrations, with hourly observations spanning 2020-2024. (2) Meteorological Data: Meteorological observations for London Heathrow Airport and Shanghai Hongqiao International Airport were acquired through the Meteostat open-source data platform (https://meteostat.net/), encompassing key parameters including temperature, dew point temperature, relative humidity, wind speed, and atmospheric pressure.

Data Preprocessing

This study employed a comprehensive multi-level data quality control methodology to ensure the reliability of subsequent analyses. The preprocessing workflow began with data standardization, where a standard time series spanning 2020-2024 was constructed, comprising 35,064 observations, with temporal alignment of air

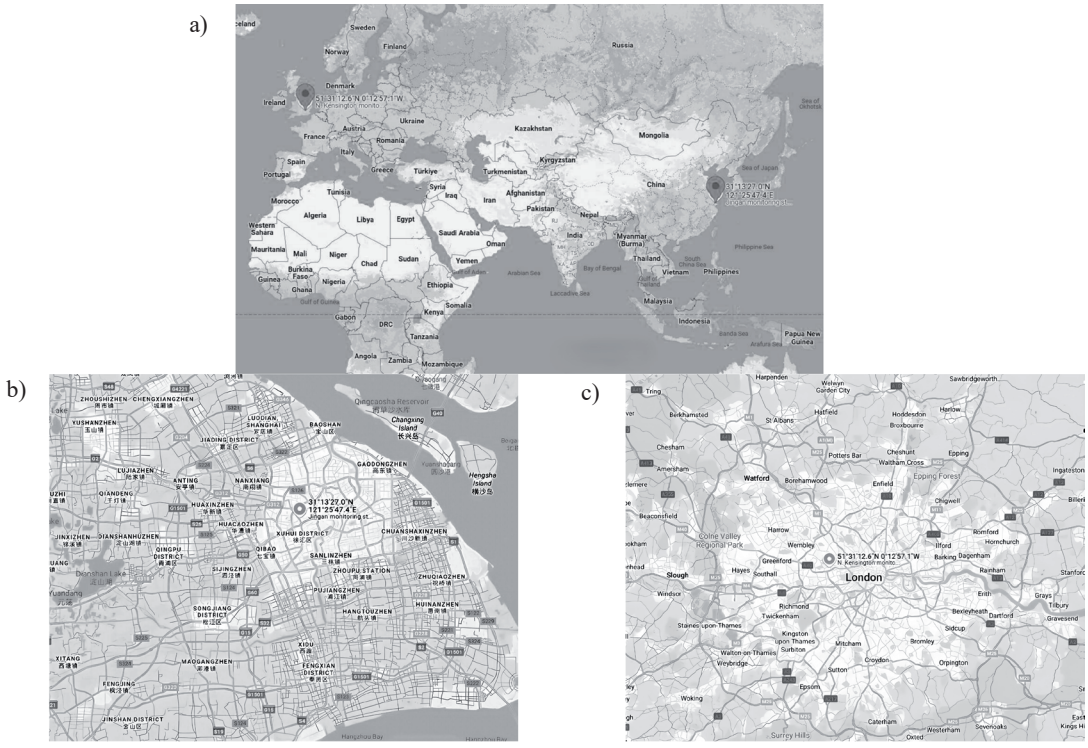

Fig. 1. Study area locations. a) Overview map showing Jing'an (Shanghai, 31.23°N, 121.46°E) and North Kensington (London, 51.52°N, 0.21°W) monitoring stations. b) Shanghai Station urban context. c) London station urban context.

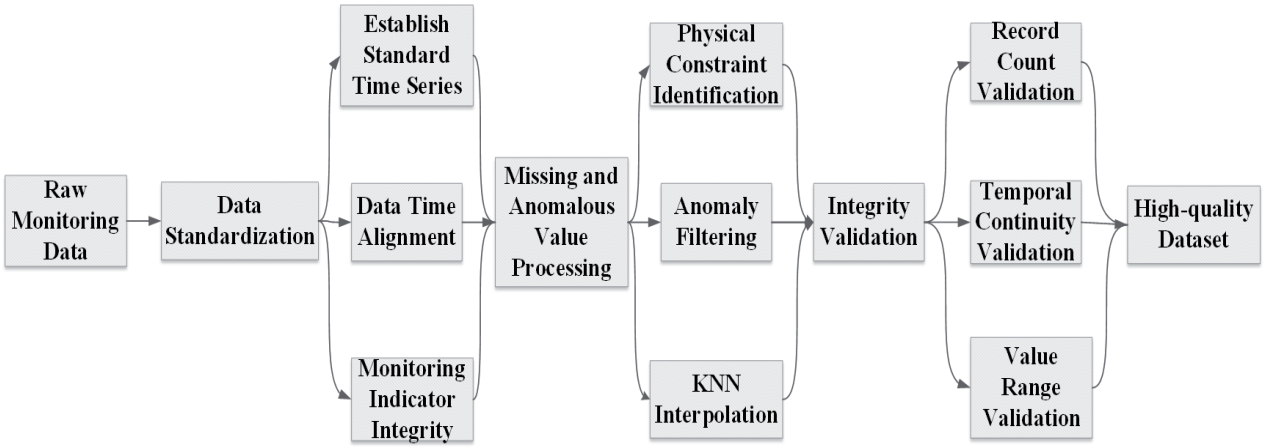

Fig. 2. Air Quality Data Preprocessing Framework Diagram.

quality indicators ($PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3, CO) and meteorological parameters. Outlier identification and processing were subsequently conducted based on established physical constraint ranges (e.g., temperature -10ºC to 45ºC, relative humidity 0-100%) combined with a time-grouped KNN algorithm [23]. Missing value imputation was implemented through a 24 h time-grouped KNN interpolation method, utilizing distance-weighted 5-nearest-neighbor interpolation for gap-filling [24]. The final step involved comprehensive completeness verification to ensure data record integrity (35,064 entries), indicator range reasonability, and time series continuity. The preprocessing workflow framework is shown in Fig. 2.

TC-MixerInformer Model Architecture

This research proposes an improved Informer model for air quality prediction (TC-MixerInformer), which builds upon the original Informer's long-sequence prediction capabilities by innovatively introducing the Reversible Instance Normalization (RevIN) mechanism and Time-Channel Mixer (TCMixer) module to enhance the model's adaptability to pollution characteristics across different cities [25].

Informer Model Foundation

The Informer model is a deep learning model specifically designed for long-sequence time series prediction tasks, incorporating significant improvements to the basic Transformer architecture. Traditional Transformers face challenges of high computational complexity, large memory consumption, and low prediction efficiency when handling long-sequence predictions. To address these challenges, Informer introduces three key innovations. First, the ProbSparse self-attention mechanism, which is based on the observation that most attention scores in traditional self-attention contribute minimally to the output. By defining the concept of dominant queries and selecting the Top-u queries for attention calculation, it successfully reduces computational complexity from $O(L^2)$ to O(LlogL). Second, the self-attention distillation mechanism, which introduces convolutional layers between encoder layers to distill features, progressively reducing sequence length and filtering redundant information while preserving important features, effectively addressing memory efficiency issues. Third, the design of a generative decoder, which simplifies the traditional Transformer decoding process by adopting a generative prediction strategy, improves the efficiency of long-sequence prediction. These innovations enable Informer to maintain high prediction accuracy while reducing computational complexity and memory consumption [13, 17]. The successful development of Informer provides new ideas and methods for solving long-sequence time series prediction problems, and its efficient computational characteristics and excellent prediction performance make it one of the important models in this field, as shown in Fig. 3.

RevIN Mechanism

To address distribution shifts caused by seasonal fluctuations and sudden events in time series data, RevIN dynamically normalizes input features and restores the original distribution during prediction through reversibility [20].

1. Normalization phase: At the input stage of the model, RevIN normalizes the input feature matrix $X \in R^{B \times L \times D}$, where B represents batch size, L represents sequence length, and D represents feature dimension (i.e., the number of pollutant indicators). The normalization operation on X includes the following steps:

(1) Calculate mean $\mu_{b,d}$ and standard deviation $\sigma_{b,d}$ along the time dimension:

$$\mu_{b,d} = \frac{1}{L}\sum_{t=1}^{L} X_{btd} \; \forall b \in [1,B], t \in [1,L], d \in [1,D], \tag{1}$$

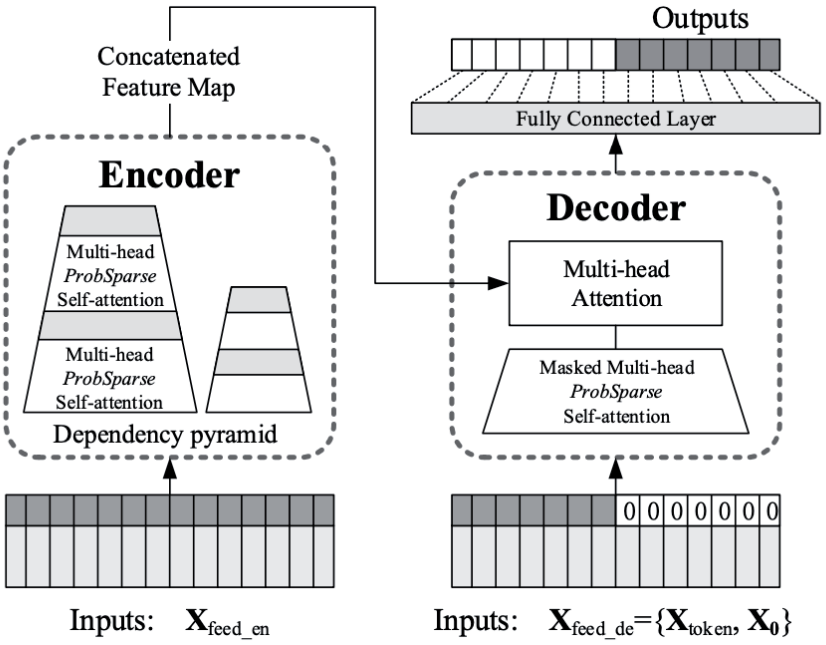

Fig. 3. Informer overall architecture.

$$\sigma_{b,d} = \sqrt{\frac{1}{L}\sum_{t=1}^{L}\left(X_{btd} - \mu_{b,d}\right)^2 + \epsilon} \quad \forall b \in [1,B], t \in [1,L], d \in [1,D], \tag{2}$$

where $\mu_{b,d} \in R^D$ is the mean of feature d in batch b, $\sigma_{b,d} \in R^D$ is the standard deviation of feature d in batch b; $\epsilon > 0$ is a small positive value used to avoid division by zero.

(2) Normalize the data:

$$X_{norm,btd} = \frac{X_{btd} - \mu_{b,d}}{\sigma_{b,d}}, \forall b, t, d. \tag{3}$$

2. Denormalization phase: Use the mean and standard deviation saved during the input phase to restore the normalized results to the original distribution:

$$Y_{denorm,btd} = Y_{norm,btd} \cdot \sigma_{b,d} + \mu_{b,d}, \forall b, t, d. \tag{4}$$

TCMixer Module Design and Principles

The TCMixer module employs a dual-branch architecture to handle complex temporal and feature interactions [26]:

Temporal Processing Branch:

$$H_{temporal}(X) = X + MLP_{temporal}(X^T)^T. \tag{5}$$

Channel Processing Branch:

$$MLP_{temporal}(X) = W_2 \cdot Relu(W_1 X + b_1) + b_2, \tag{6}$$

where $W_1 \in R^{d_{time} \times L}$, $W_2 \in R^{L \times d_{time}}$, d_{time} = L×2, b_1, b_2 are corresponding bias terms.

Feature Fusion:

$$H_{out} = H_{channel}\left(H_{temporal}(X)\right). \tag{7}$$

Final output through linear transformation:
When there is no time encoding (X_{mark}):

$$Y = Linear(H_{out}) = W_{linear} H_{out} + b. \tag{8}$$

When there is time encoding (X_{mark}):

$$Y = Linear(H_{out}; X_{mark}) = W_{linear}[H_{out}; X_{mark}] + b. \tag{9}$$

The Mixer architecture is shown in the Fig. 4.

This architectural design enables the model to simultaneously capture temporal dependencies and interactions between pollutants, effectively improving prediction performance. Particularly when handling long-sequence prediction tasks, the model demonstrates significant advantages, providing reliable technical support for air quality prediction.

Model Evaluation and Experimental Design

Experimental Setup and Comparison Models

The experiments were implemented based on Python 3.8 and the PyTorch 1.8.0 framework. Model input features include historical data of co-pollutants such as PM_{10}, SO_2, NO_2, CO, O_3, and meteorological elements, including temperature, humidity, wind speed, and wind direction, to capture the chemical transformation relationships between pollutants and the influence of meteorological conditions, thereby improving the prediction accuracy for $PM_{2.5}$.

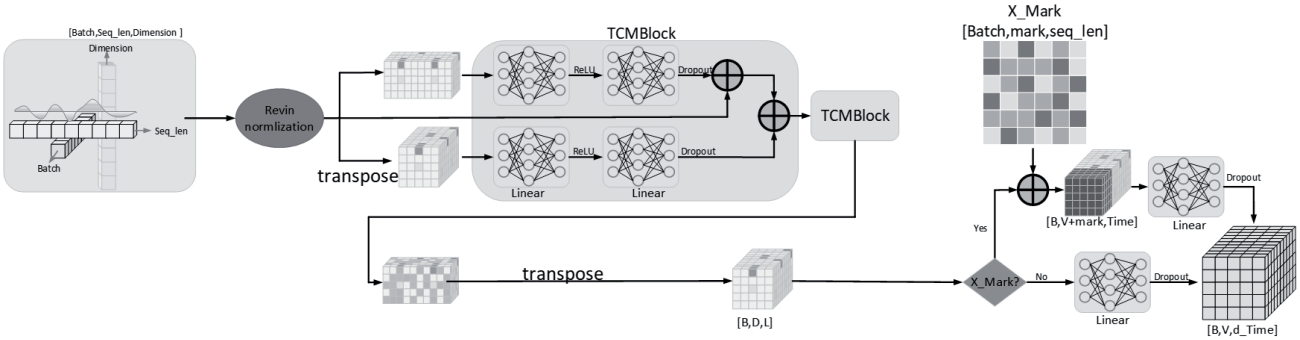

Fig. 4. TC-Mixer architecture.

The dataset was divided into training, validation, and testing sets at a ratio of 7:1:2. An Adam optimizer with a learning rate of 0.001 was employed, and a dynamic batching strategy was designed based on prediction length: short-term prediction (≤12 h) with a batch size of 1024, medium-term (≤24 h) with 512, longer-term (≤72 h) with 128, and long-term (>72 h) with 64. To ensure experimental reproducibility, each set of experiments was repeated 5 times, and the average values were taken. The following typical models were selected for comparison, as shown in Table 1:

The experiment employed a sliding window for data sampling, with random shuffling only applied to the training set while maintaining the temporal continuity of the validation and test sets to ensure the authenticity of the evaluation.

Model Evaluation Metrics

This research uses Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), and Mean Absolute Percentage Error (MAPE) as indicators for evaluating model prediction performance.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}}, \tag{13}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{14}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}, \tag{15}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|. \tag{16}$$

Where y_i represents the true values, $\hat{y}_i$ represents the predicted values, $\bar{y}$ represents the mean of all observed values, and n represents the sample size. RMSE reflects the overall level of prediction error, MAE represents the average magnitude of prediction bias, R^2 measures the degree to which the model explains the variability in the data, and MAPE provides a percentage measure of relative error.

Results and discussion

Data Characteristics and Regional Comparison

The comparative analysis of $PM_{2.5}$ concentration patterns between the Shanghai (Jing'an) and London (North Kensington) monitoring stations reveals fundamental differences in urban air quality characteristics that directly inform

Table 1. Comparison Model Introduction.

Model Name	Description	Parameter Settings
Informer [17]	Long sequence prediction model based on a sparse self-attention mechanism	Number of attention heads = 8, Hidden layer dimension = 512
TSMixer [22]	MLP-based temporal feature mixing model	Number of mixing layers = 2, Hidden layer dimension = 512
LightTS [27]	Lightweight self-attention mechanism for time series prediction	Number of attention heads = 4, Hidden layer dimension = 256, Number of encoder layers = 2
Pyraformer [28]	Prediction model with pyramid attention mechanism	Number of pyramid layers = 2, Attention dimension = 512
GRU [29]	Classic gated recurrent unit network as a classic RNN variant	Hidden layer dimension = 512, Number of layers = 2

the TC-MixerInformer model development. These regional variations provide the empirical foundation for understanding the complexity of cross-regional air quality prediction challenges.

Statistical Analysis of Data

Shanghai demonstrates significantly higher $PM_{2.5}$ concentrations with a mean of 28.23 μg/m³ (range: 2-378 μg/m³) compared to London's mean of 8.08 μg/m³ (range: 0.19-121.56 μg/m³) (Table 2, Table 3). The coefficient of variation reveals substantial temporal variability in both cities (Shanghai: 78.26%, London: 89.21%), indicating the inherent non-stationary nature of urban air quality time series. Shanghai's 95th percentile concentration (72 μg/m³) exceeds WHO guidelines by a factor of five, while London's (21.98 μg/m³) approaches the recommended threshold, reflecting the distinct pollution regimes between developing and developed urban environments.

The extreme concentration events present particular modeling challenges. Shanghai experiences frequent pollution episodes exceeding 100 μg/m³ with maximum values reaching 378 μg/m³, characteristics of complex industrial-urban mixed pollution sources typical of rapidly developing megacities. Conversely, London's pollution profile exhibits lower baseline concentrations with occasional moderate peaks, primarily attributed to traffic emissions and meteorological accumulation effects. These contrasting patterns necessitate adaptive

Table 2. Statistical Characteristics of Air Quality and Meteorological Parameters at the Jing'an (Shanghai) Monitoring Station.

	Mean	Min	Max	50%	cv(%)	95th_percentile
$PM_{2.5}$	28.23	2	378	22	78.26	72
CO	0.66	0.3	2.42	0.61	30.62	1.04
NO_2	32.15	4	140	27	59.01	72
PM_{10}	43.71	1	554	36	72.11	100
O_3	69.14	2	267	65	55.48	142
SO_2	6.44	3	33	6	29.28	10
Temperature_c	18.63	-7	40	19	49.3	33
Dew point_c	12.55	-15	29.1	13	78.73	26
Relative_humidity_percent	70.15	15	100	72	26.09	94
Wind_speed_kmh	13.8	1.8	75.6	14.4	48.11	25.2
Wind_direction	166.47	0	360	150	65.58	350
Pressure_hpa	1016.2	982	1044	1016	0.92	1031

Table 3. Statistical Characteristics of Air Quality and Meteorological Parameters at the North Kensington (London) Monitoring Station.

	Mean	Min	Max	50%	cv(%)	95th_percentile
$PM_{2.5}$	8.08	0.19	121.56	5.9	89.21	21.98
CO	0.17	0.01	2.49	0.13	73.62	0.36
NO_2	18.5	0.19	191.66	13.58	80.35	50.3
PM_{10}	13.17	0.4	160.45	10.8	71.27	31
O_3	52.99	0.13	216.53	54.48	47.93	91.4
SO_2	1.26	0.07	15.97	0.8	104.41	3.5
Temperature_c	12.23	-8.3	40.2	11.9	50.5	22.7
Dew point_c	7.62	-14.3	20.1	7.9	64.57	15.2
Relative_humidity_percent	75.86	14	100	80	21.76	96
Wind_speed_kmh	14.67	1.8	68.4	13	54.78	29.5
Wind_direction	193	10	360	210	48.51	330
Pressure_hpa	1015.17	955.7	1049.6	1016.1	1.11	1031.8

modeling approaches capable of handling both high-amplitude variations and subtle fluctuation patterns.

Data Correlation Analysis

Correlation analysis employs the Spearman rank correlation coefficient [30], a non-parametric correlation measure, to evaluate the degree of linear correlation between different pollutants in Shanghai and London.

Correlation analysis using Spearman rank coefficients reveals distinct meteorological influences on $PM_{2.5}$ concentrations across regions (Fig. 5). In Shanghai, $PM_{2.5}$ exhibits weak negative correlations with temperature (r = -0.24), relative humidity (r = -0.15), and wind speed (r = -0.14), consistent with photochemical reaction dynamics and atmospheric dispersion mechanisms in subtropical climates [31]. The strongest positive correlation occurs with PM_{10} (r = 0.76), followed by moderate positive correlations with CO (r = 0.57) and NO_2 (r = 0.56), indicating significant contributions from combustion processes and vehicular emissions.

London demonstrates a notably stronger $PM_{2.5}$-PM_{10} correlation (r = 0.93), reflecting more uniform particulate matter sources consistent with London's urban characteristics, where traffic-related emissions constitute a dominant contribution to fine particulate matter. The correlation with NO_2 (r = 0.48) remains substantial but is lower than in Shanghai, attributable to more stringent vehicular emission controls under European standards and differences in fuel composition [32]. Most notably, the relationship with wind speed (r = -0.40) is significantly stronger than in Shanghai, which aligns with the pronounced influence of Atlantic weather systems and the dynamic westerly wind patterns characteristic of temperate maritime climates.

These region-specific correlation patterns provide empirical foundations for the TC-MixerInformer's adaptive feature interaction mechanisms. The differential correlation strengths directly informed the design of temporal-channel mixing strategies, enabling the model to apply region-appropriate weightings to meteorological and chemical features.

These contrasting pollution signatures provide crucial guidance for TC-MixerInformer architecture design, particularly in feature interaction modeling and temporal processing strategies, enabling enhanced adaptability to diverse urban pollution scenarios.

Model Performance Evaluation

Multi-City Prediction Performance Evaluation

Experimental results presented in Table 4 and Table 5 demonstrate that TC-MixerInformer exhibits significant predictive advantages across both urban environments. At the Shanghai Jing'an station, the model maintains exceptional stability from short-term (1-12 h) to long-term predictions (24-168 h). In 168 h predictions, TC-MixerInformer achieves RMSE = 8.000 compared to Informer's 17.492, representing a 54% improvement. At the London North Kensington station, performance remains robust across all horizons (1 h RMSE = 1.938 to 168 h RMSE = 2.600), substantially outperforming GRU (1 h RMSE = 2.424 to 168 h RMSE = 5.871).

The superior performance stems from three architectural innovations addressing fundamental modeling challenges. First, the dual-branch TCMixer architecture simultaneously processes temporal dependencies (token-mixing) and inter-variable relationships (channel-mixing) [33], whereas recurrent models like GRU sequentially process only temporal patterns. This architectural difference explains GRU's severe degradation in long-term Shanghai predictions (168 h RMSE = 23.762 vs. 8.000), as recurrent

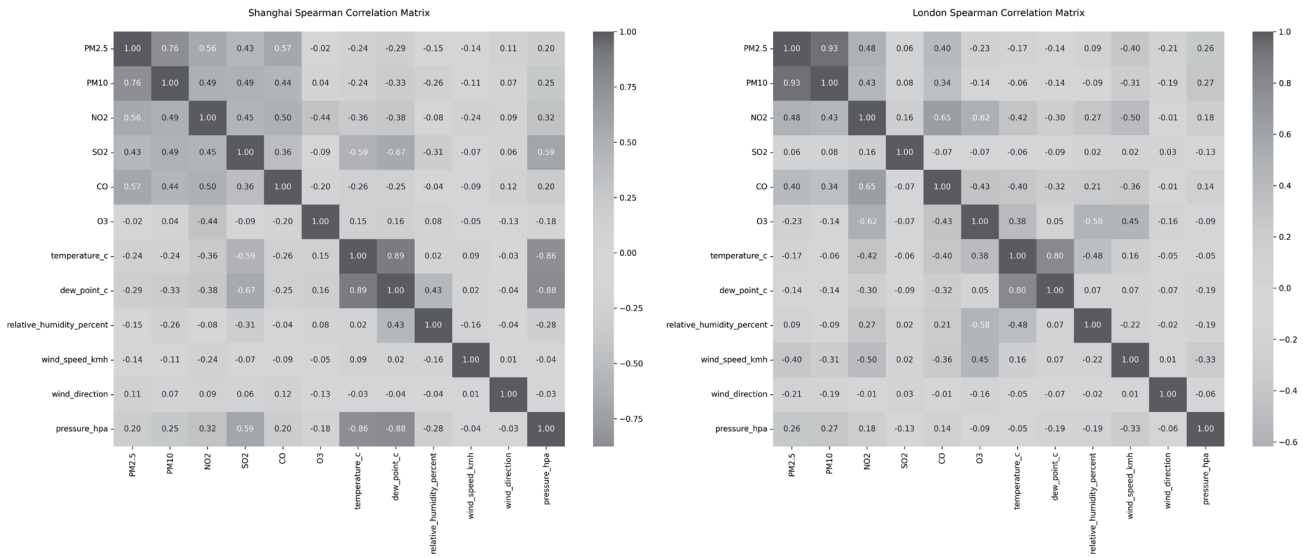

Fig. 5. Multivariate correlation analysis for Shanghai and London.

Table 4. Jing'an Shanghai Model Prediction Performance comparison.

Model	TS-MixerInformer		Informer		LightTS		TSMixer		Pyraformer		GRU	
Predicted duration	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
1 h	6.074	2.856	8.359	4.095	10.623	6.191	9.225	5.753	7.923	3.535	7.488	4.093
6 h	6.391	3.045	7.882	3.828	8.012	4.588	11.460	7.293	8.728	3.985	13.987	8.560
12 h	6.387	3.130	8.757	4.553	9.979	5.954	14.651	9.716	9.275	4.548	16.653	10.455
24 h	6.836	3.231	8.217	4.130	7.832	4.446	11.837	7.686	10.614	5.485	22.042	14.550
48 h	8.545	5.012	9.387	5.987	9.691	5.618	12.660	8.239	10.199	5.856	23.509	15.518
72 h	7.184	4.264	12.804	7.986	8.365	4.829	11.558	7.777	10.508	5.607	24.100	15.993
168 h	8.000	4.822	17.492	10.394	11.255	7.324	12.395	8.224	10.788	5.974	23.762	15.830

Table 5. North Kensington (London) Model Prediction Performance comparison.

Model	TS-MixerInformer		Informer		LightTS		TSMixer		Pyraformer		GRU	
Predicted duration	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
1 h	**1.938**	**1.149**	2.275	1.438	3.432	2.304	2.848	1.935	2.061	1.306	2.424	1.597
6 h	**1.989**	**1.209**	2.169	1.387	2.367	1.541	3.416	2.374	2.217	1.441	4.143	2.858
12 h	**1.993**	**1.208**	2.271	1.493	3.047	2.093	4.134	3.073	2.456	1.673	4.691	3.279
24 h	**2.150**	**1.320**	2.517	1.736	2.515	1.614	3.418	2.351	2.717	1.885	5.406	4.002
48 h	**2.329**	**1.485**	4.132	3.092	2.648	1.740	3.611	2.413	3.111	2.081	5.614	4.191
72 h	**2.358**	**1.538**	2.793	1.935	2.430	1.596	3.667	2.508	2.952	1.956	5.718	4.304
168 h	**2.600**	**1.799**	5.567	3.499	3.514	2.415	4.272	3.066	3.521	2.350	5.871	4.505

Note: The best experimental results are highlighted in bold.

connections suffer gradient vanishing beyond 24-48 h, preventing effective capture of weekly pollution cycles. Second, RevIN's dynamic instance-level normalization adapts to distribution shifts through reversible transformations that preserve temporal dynamics while eliminating regional baseline differences. In contrast, LightTS employs static normalization optimized for training distributions, causing systematic errors when test conditions diverge. This manifests in London's stable environment, where LightTS achieves comparable 168 h performance (RMSE = 3.514 vs. 2.600), but deteriorates significantly in Shanghai's high-variability conditions (RMSE = 11.255 vs. 8.000) – a 40.7% performance gap demonstrating inability to handle non-stationary distributions. Third, ProbSparse attention identifies critical pollution events through selective focus on dominant queries, achieving a 27.3% improvement over standard Informer in Shanghai. Fixed-weight approaches in TSMixer and Pyraformer distribute attention uniformly across all time steps, failing to prioritize extreme concentration periods and resulting in systematic underestimation during rapid pollution transitions, as evidenced by their 168 h RMSE values (12.395 and 10.788, respectively) exceeding TC-MixerInformer by 55% and 35%.

Comparative Analysis of Short-term and Long-term Prediction Performance Under Urban Differences

The scatter plot analysis (Fig. 6 and Fig. 7) provides an intuitive visualization of prediction accuracy characteristics across different models through the distribution relationship between predicted and actual values. TC-MixerInformer demonstrates optimal linear fitting performance in both 1 h and 168 h predictions, with prediction points tightly clustered around the ideal diagonal line. For the Shanghai Jing'an station, the 1 h prediction achieves an R^2 of 0.940, maintaining excellent consistency even in high concentration regions (>100 μg/m³); for the London North Kensington station, it exhibits the tightest linear relationship (R^2 = 0.885) with the most uniform distribution of prediction points.

In contrast, other models exhibit significant prediction deviations. The Informer model already shows noticeable bias in high concentration regions during short-term predictions, with performance further

deteriorating in 168 h forecasts where scatter points deviate markedly from the diagonal line. The GRU model demonstrates substantial errors across the entire concentration range, with long-term prediction scatter points exhibiting a "fan-shaped" distribution, indicating that prediction uncertainty increases dramatically with concentration levels.

The point density distribution in the scatter plots further confirms TC-MixerInformer's superiority in prediction consistency. Near the ideal diagonal line, TC-MixerInformer displays the highest point density concentration, while other models show relatively dispersed scatter distributions, particularly exhibiting greater prediction variability in extreme value regions.

Time Series Prediction Visualization: Comparison of Multi-model Short-term and Long-term Prediction Performance

The time series comparisons further validate the model stability (Fig. 8 and Fig. 9). Shanghai data exhibits extensive concentration variability (0-350+ μg/m³), presenting significant forecasting challenges (Fig. 8). TC-MixerInformer demonstrates exceptional trend-tracking capability in both short-term and long-term predictions, with prediction curves highly consistent with actual measurements, maintaining good alignment even during sudden pollution peak periods. In contrast, Informer shows obvious deviations in high-concentration intervals, with particularly significant performance degradation in long-term predictions.

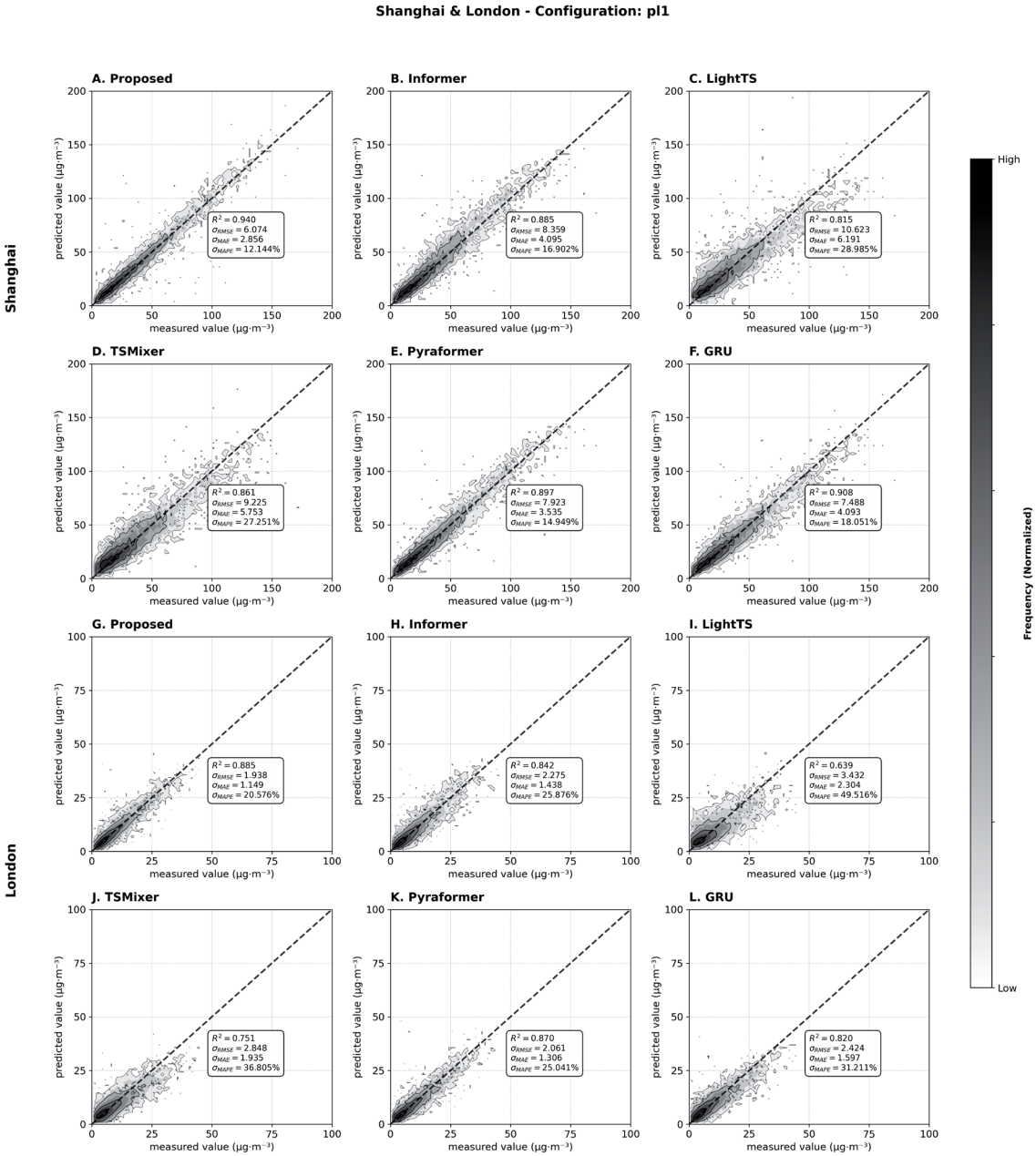

Fig. 6. Comparison of scatter plots of different models at the Jing'an (Shanghai) and North Kensington (London) stations (1 h prediction).

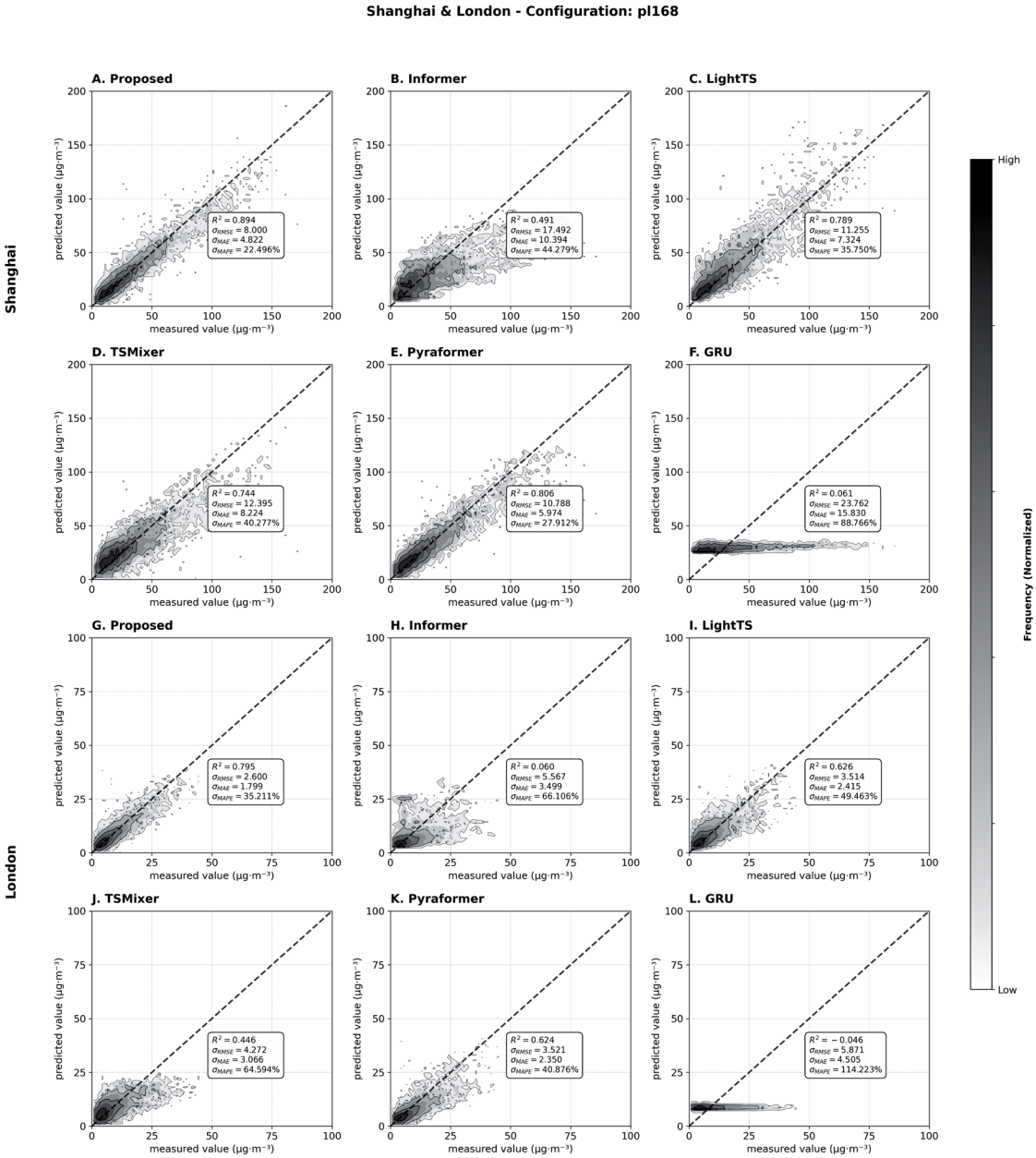

Fig. 7. Comparison of scatter plots of different models at the Jing'an (Shanghai) and North Kensington (London) stations (168 h prediction).

While the GRU model can track basic trends, it still exhibits certain gaps in prediction accuracy.

London data presents a contrasting profile, with lower concentrations (0-20+ μg/m³) characterized by frequent minor fluctuations (Fig. 9). TC-MixerInformer demonstrates superior adaptability in both short-term and long-term predictions, with prediction curves capable of precisely tracking subtle measurement variations while maintaining good forecasting stability.

Alternative models exhibit obvious limitations: Informer shows over-prediction at several time points, with this deviation further amplified in long-term predictions; TSMixer and Pyraformer display temporal inconsistencies in responding to minor fluctuations; the GRU model shows reduced sensitivity to small-scale changes, with prediction curves being overly smooth.

Comprehensive Analysis of Cross-Regional Adaptability and Generalization Capability

(1) Temporal Scale Adaptability Characteristics

TC-MixerInformer demonstrates significant adaptability characteristics across different prediction time horizons (as shown in Table 4 and Table 5). The model exhibits exceptional stability in short-term predictions (1-12 h), with performance degradation of only 5.2% and 2.8% in Shanghai and London, respectively, while medium-to-long-term predictions (24-168 h) show more pronounced performance

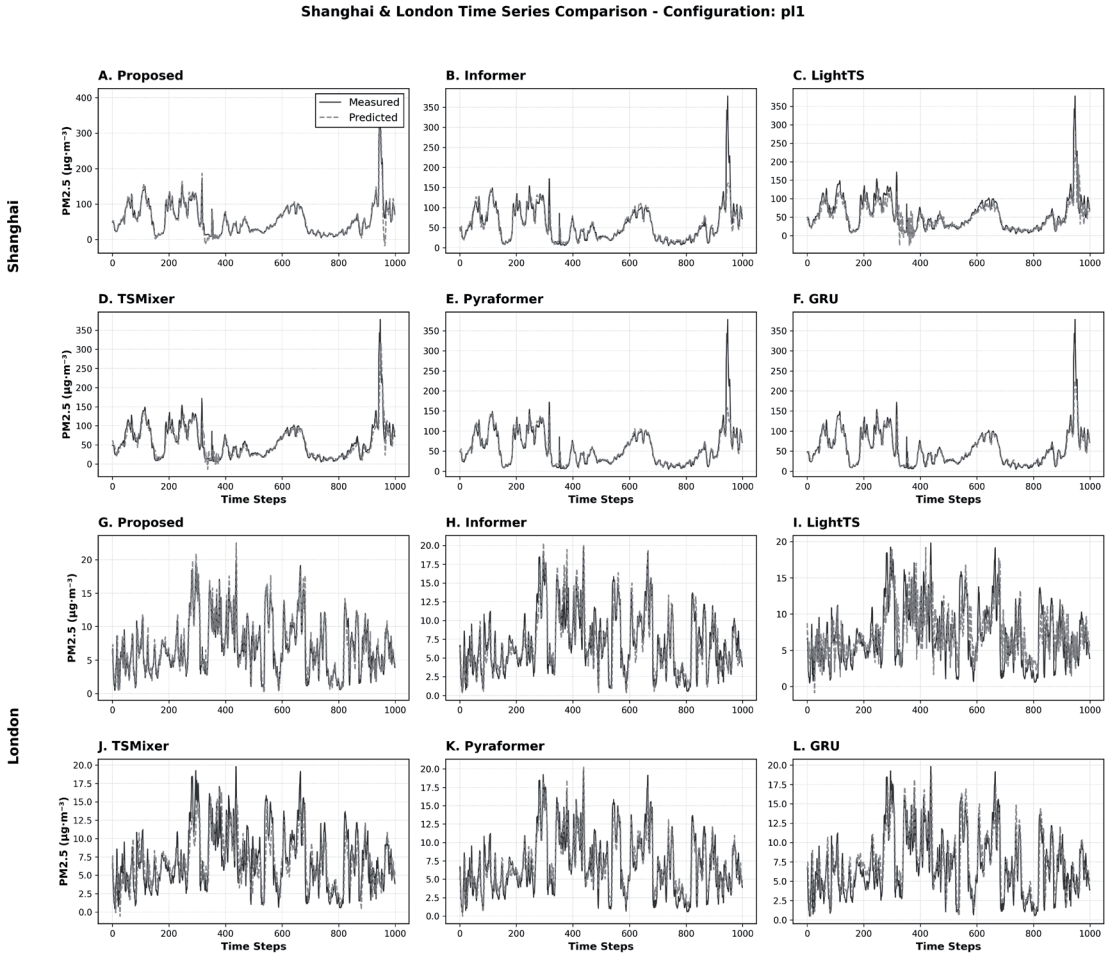

Fig. 8. Time series comparison of different models at the Shanghai and London stations for 1 h prediction.

decline, with RMSE increases reaching 17.0% and 20.9%, respectively [34]. Detailed analysis reveals that performance degradation follows a progressive pattern: short-term internal decay (1 h→12 h) remains at the lowest levels of 5.2% in Shanghai and 2.8% in London; medium-term decay (12 h→24 h) increases to 7.0% and 7.9%; long-term decay (24 h→168 h) reaches 17.0% and 20.9%. The data indicate that the model maintains relatively stable performance in the 12-24 h range, which is closely related to the diurnal cycle characteristics of urban pollution.

Performance degradation stems from three interrelated fundamental factors. First, nonlinear cumulative error effects cause slight prediction errors to gradually amplify into systematic biases in long-term predictions. Short-term predictions are primarily controlled by local meteorological conditions and near-source emissions, making pollutant concentration changes relatively predictable, while long-term predictions need to consider complex processes, including regional transport and chemical transformation, whose nonlinear characteristics and randomness increase prediction difficulty [35]. Second, the complexity of large-scale meteorological systems exceeds the model's expressive capacity, as long-term predictions need to handle complex atmospheric processes, including frontal passages and air mass transitions. Additionally, dynamic changes in emission source patterns pose challenges for long-term modeling, including weekday-weekend emission patterns and seasonal variations, which are difficult to fully capture.

TC-MixerInformer's hierarchical architecture design enables it to adopt differentiated processing strategies across different time scales. In short-term predictions, the model primarily utilizes the token-mixing branch of TCMixer to capture local temporal dependencies within sequences, while the RevIN mechanism ensures statistical stability of input features. As the prediction window extends, the model gradually shifts to relying on the channel-mixing branch to handle long-term interaction patterns among multiple variables. However, limited by the expressive capacity of linear transformations, the model still faces difficulties in fully characterizing complex nonlinear long-term dependencies. In contrast, the baseline Informer model exhibits more severe performance degradation in long-term predictions, demonstrating the critical role of the RevIN and TCMixer modules in maintaining long-term prediction stability.

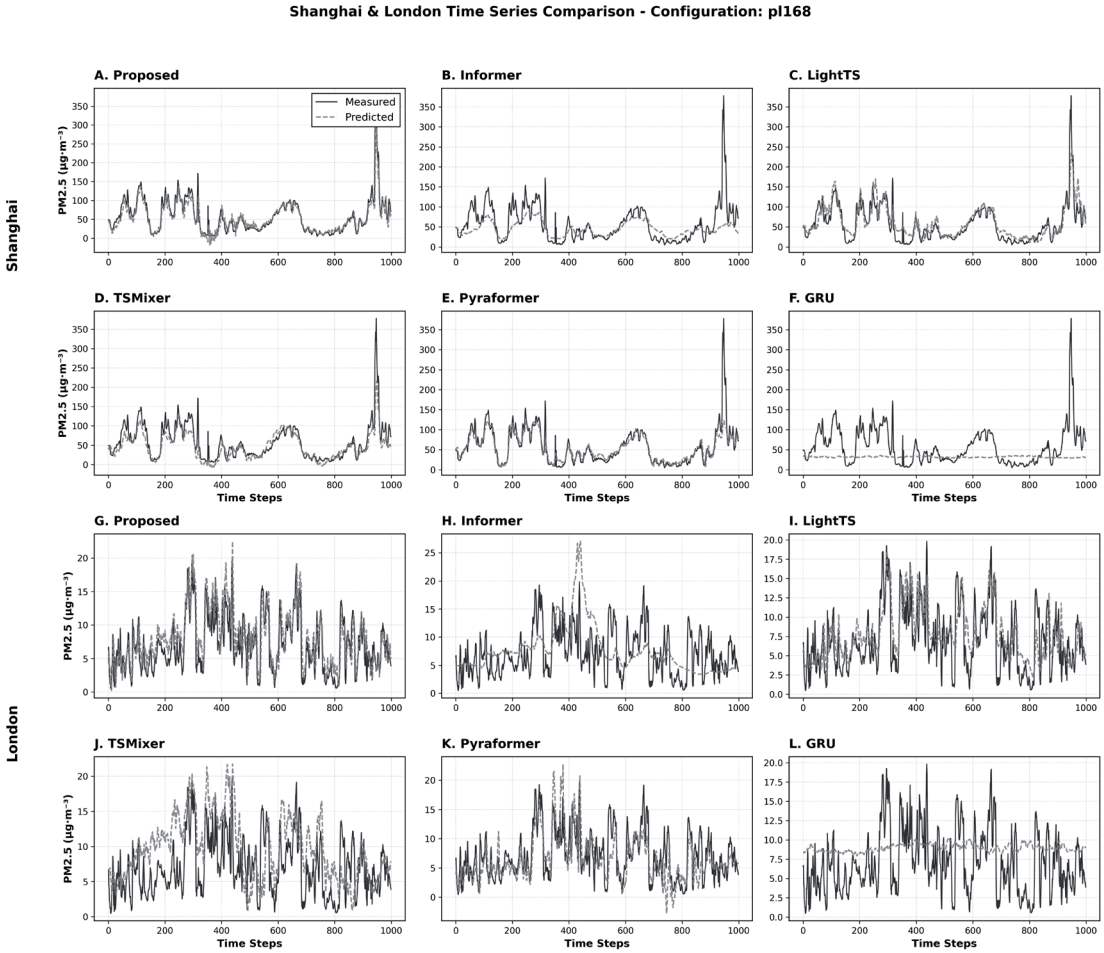

Fig. 9. Time series comparison of different models at the Shanghai and London stations for 168 h prediction.

(2) Cross-Regional Performance Mechanisms and Adaptation Challenges

Through comprehensive comparative analysis of the TC-MixerInformer performance in Shanghai and London, we systematically examined the model's cross-regional adaptability characteristics. Although Shanghai's 1 h prediction RMSE of 6.074 significantly exceeded London's 1.938 (approximately 3-fold difference), R^2 coefficient analysis revealed that TC-MixerInformer maintained exceptional trend-tracking capability in Shanghai's complex pollution environment (1 h prediction R^2 = 0.940 vs. Informer's 0.885; 168 h prediction R^2 = 0.894 vs. Informer's 0.491), demonstrating the model's strong adaptability to complex nonlinear pollution systems, as illustrated in Fig. 6 and Fig. 7 [36].

Despite higher absolute errors, TC-MixerInformer exhibited superior relative performance and model-environment synergy in Shanghai, revealing amplified architectural advantages under complex environmental conditions. Performance enhancement analysis demonstrated that the model's architectural innovations produced more significant improvements in Shanghai's challenging environment: 1 h prediction RMSE decreased by 27.3% (from 8.359 to 6.074), while London showed only a 14.8% reduction (from 2.275 to 1.938). This differential improvement pattern indicates that TC-MixerInformer's complex architecture identifies greater optimization potential in high-complexity environments, where traditional models encounter greater difficulties, thereby amplifying the relative advantages of advanced feature learning mechanisms.

Root cause analysis revealed three interconnected factors driving regional performance differences. Fundamental differences in pollutant concentration distribution characteristics constitute the primary challenge. Shanghai's $PM_{2.5}$ concentration ranged from 2-378 μg/m³ (mean 28.23 μg/m³), characterized by high concentrations, large variations, and frequent extreme events, while London's concentration ranged from 0-121.56 μg/m³ (mean 8.08 μg/m³), displaying relatively moderate and stable variations as detailed in Table 2 and Table 3. Correlation analysis shows that Shanghai's $PM_{2.5}$ exhibits moderate positive correlations with multiple pollutants (PM_{10}: r = 0.76, CO: r = 0.57, NO_2: r = 0.56), reflecting complex multi-source interactions, while London demonstrates a dominant $PM_{2.5}$-PM_{10} correlation (r = 0.93), indicating more uniform emission patterns. Shanghai's nearly 200-fold extreme concentration range and frequent pollution events exceeding 100 μg/

m³ created optimal conditions for TC-MixerInformer's dual-branch architecture in nonlinear pattern recognition and extreme value prediction, while London's limited concentration range provided restricted opportunities for the model's complex architecture to demonstrate its full predictive capabilities.

Meteorological condition complexity constitutes the second critical factor. Shanghai's subtropical monsoon climate, influenced by multiple weather systems, exhibits significantly higher variability in meteorological variables compared to London's temperate oceanic climate, substantially increasing model learning complexity. Due to complex urban topography, Shanghai exhibits weaker wind speed effects (r = -0.14), while London demonstrates stronger atmospheric dispersion relationships (r = -0.40), as shown in Fig. 5. Emission source structural differences constitute the third critical challenge. Shanghai, as an industrial port city, features diversified pollution sources (industrial emissions, vehicle exhaust, ship emissions, construction dust) with significant spatiotemporal variations in source contributions, providing rich multi-source interaction learning opportunities for the model's channel mixing branch, while London, following extensive environmental management, exhibits relatively singular and well-controlled traffic-dominated emission sources with stable emission patterns [37].

Based on these findings, the complexity-adaptability matching principle emerges as a key insight: TC-MixerInformer's architectural design is inherently more suitable for high-complexity, multi-variable interactive time series prediction tasks. In Shanghai's complex industrial-urban mixed pollution environment, the model fully leverages its design advantages through dual-branch mixing mechanisms for handling complex variation patterns across multiple temporal scales, RevIN dynamic normalization for processing high-variability distribution characteristics, and attention mechanisms for capturing long-range dependencies. Conversely, in London's relatively simple traffic-dominated pollution environment, the complex architecture's advantages cannot be fully realized, potentially leading to relative "over-engineering" phenomena.

Cross-regional application validation further confirmed TC-MixerInformer's remarkable environmental adaptability. In London's relatively simple traffic-dominated pollution environment, the model achieved stable high-precision predictions through effective capture of regular temporal patterns and dominant variable relationships. In Shanghai's complex industrial-urban mixed pollution source environment, despite higher absolute errors, the model successfully captured pollution peaks approaching 350+ μg/m³ while maintaining excellent trend-tracking capability in complex nonlinear relationships, showing more significant relative improvements compared to baseline models. Particularly noteworthy is that the model demonstrated stronger relative improvement advantages in Shanghai's more challenging environment, validating TC-MixerInformer as an adaptive framework capable of automatically adjusting feature learning strategies according to environmental complexity, providing reliable technical solutions for air quality prediction in diverse urban environments.

Several additional factors contribute to the observed performance disparities. First, regional transport contributions differ substantially: Shanghai's $PM_{2.5}$ is significantly influenced by long-range transport from the Yangtze River Delta [38], resulting in higher background concentrations, whereas London's pollution is more locally dominated. Second, pollution episode characteristics vary: Shanghai experiences more frequent extreme events (>150 μg/m³) with longer persistence (multi-day episodes), creating highly nonlinear dynamics that challenge prediction, while London's events are typically shorter and less extreme [39]. Third, the 2013-2023 period encompasses different policy trajectories: Shanghai underwent rapid emission control changes, introducing stronger non-stationarity, whereas London experienced gradual improvements [40]. These factors collectively increase Shanghai's prediction complexity, explaining higher absolute errors despite robust relative performance improvements.

Technical Discussion: Model Mechanisms, Limitations, and Future Improvements

RevIN Mechanism and Its Operational Principles

The introduction of the Reversible Instance Normalization (RevIN) mechanism effectively addresses the non-stationarity issues inherent in air quality time series data. By dynamically preserving and restoring distributional characteristics of the data, RevIN enables the model to better handle distributional variations across different time periods and abrupt pollution events. This mechanism demonstrates excellent adaptability when processing the distinctly different pollution patterns observed in Shanghai and London, providing crucial support for the model's cross-regional generalization capabilities.

Innovation Value of the TCMixer Module

The Time-Channel Mixer module achieves effective modeling of complex temporal dependencies and feature interactions through its dual-branch architecture. The temporal mixing branch focuses on extracting variation patterns across different temporal scales ranging from hourly to multi-day intervals, while the channel mixing branch models the interrelationships between $PM_{2.5}$ and other pollutants as well as meteorological variables. This parallel processing mechanism allows the model to simultaneously consider temporal evolution patterns and multivariate interaction effects, thereby demonstrating excellent adaptability when handling two cities with distinct pollution characteristics – Shanghai and London.

Model Limitations and Challenge Analysis

Despite TC-MixerInformer's excellent performance in air quality prediction, it still faces several important limitations when handling complex environmental conditions.

Root Cause Analysis of Sudden Event Prediction Limitations

The model's limited capability in predicting sudden pollution events stems from three fundamental architectural constraints that are well-documented in time series forecasting literature.

First, training data distribution imbalance creates systematic bias toward common pollution patterns. As shown in Table 2, Shanghai's $PM_{2.5}$ concentrations exhibit extreme variability (range: 2-378 μg/m³, CV = 78.26%), with the 95th percentile (72 μg/m³) representing only a small fraction of observations. This severe class imbalance causes the model to optimize primarily for frequent moderate pollution patterns rather than rare extreme events. The loss function's equal weighting across all samples means that extreme pollution episodes, despite their critical importance for public health warnings, contribute minimally to the total training loss due to their low frequency [39].

Second, the absence of external trigger information fundamentally limits the model's ability to anticipate sudden events. The current input features comprise only historical pollution concentrations and routine meteorological parameters (temperature, humidity, wind speed, pressure), lacking critical indicators of sudden pollution triggers. Research has demonstrated that extreme pollution events are often triggered by factors outside the scope of conventional monitoring data, including industrial accidents, biomass burning, dust storms, and regional transport from upstream sources. Without access to such trigger signals, the model can only react to concentration increases after they manifest in monitoring data, resulting in delayed predictions.

Third, temporal context window limitations constrain the model's ability to capture precursor signals of extreme events. While the current 168 h input window captures weekly cycles and short-term meteorological patterns, it may be insufficient for detecting synoptic-scale atmospheric circulation changes that precede major pollution episodes. Studies have shown that extreme pollution events are often preceded by identifiable atmospheric circulation pattern shifts several days in advance, including weakening of cold front intensity, establishment of stable atmospheric stratification, and reduction in boundary layer height.

The model primarily relies on historical data patterns, which limits its predictive capability for sudden pollution events such as industrial accidents or dust storms. For extreme pollution events, including industrial accidents, dust storms, and forest fires, the model's reliance on historical pattern learning makes it difficult to predict anomalies that exceed the distribution range of training data. Experimental data show that Shanghai's $PM_{2.5}$ concentrations can reach extreme levels of 378 μg/m³, and when facing such sudden high-concentration pollution, the model's prediction accuracy faces challenges. This limitation stems from the model's dependence on historical data patterns, while historical data itself lacks sufficient representation of rare extreme events [41].

Mechanistic Analysis of Data Quality Dependence

The model's sensitivity to input data quality manifests through three distinct propagation pathways documented in environmental monitoring literature.

Systematic sensor bias introduces persistent directional errors that accumulate across prediction horizons. Research on air quality monitoring equipment has shown that $PM_{2.5}$ sensors typically exhibit calibration drift over time, with bias magnitude often correlating with local pollution levels. When such biased data enters the model's training process, the RevIN normalization mechanism, designed to handle distribution shifts, may inadvertently learn and perpetuate these systematic errors. The reversible nature of RevIN means that any systematic bias in the normalized space will be faithfully restored during denormalization, potentially amplifying prediction errors in long-term forecasts.

Missing meteorological variables create critical information gaps that force the model to rely on incomplete feature representations. As demonstrated in Fig. 5, different meteorological variables exhibit varying correlations with $PM_{2.5}$ across regions: wind speed shows a weak correlation in Shanghai ($r = -0.14$) but a stronger correlation in London ($r = -0.40$), while humidity and temperature effects differ substantially between subtropical and temperate climates. When key meteorological drivers are missing, the TCMixer module's channel-mixing branch cannot properly model the multivariate interactions that govern pollution dynamics, leading to degraded prediction accuracy.

Temporal data gaps disrupt the model's ability to track pollution evolution continuity. The TCMixer module's temporal-mixing branch relies on continuous temporal patterns to extract meaningful features across different time scales. When data gaps are filled using simple interpolation methods (as described in "Data Preprocessing" Section), the interpolated values lack the natural variability and correlation structures present in actual measurements. This artificial smoothness can mislead the temporal pattern recognition mechanisms, particularly for the self-attention components in the Informer encoder that depend on authentic temporal dependencies.

The model's performance exhibits high dependence on input data quality, constituting vulnerability in actual deployment. Missing key meteorological variables affect

prediction accuracy, and calibration deviations or sensor failures in pollution monitoring equipment may lead to systematic biases in model outputs, with cumulative effects propagating throughout the entire prediction time range [42]. Additionally, the dual-branch structure of the TCMixer module requires more computational resources compared to traditional single architectures, increasing the system's computational burden and memory requirements to some extent.

Computational Resource Requirements Analysis

The computational demands of TC-MixerInformer stem from its dual-branch TCMixer architecture and the integration of multiple advanced components. As described in "TC-MixerInformer Model Architecture" Section, the model combines the Informer's ProbSparse self-attention mechanism with the TCMixer's parallel temporal and channel mixing branches, plus the RevIN normalization layers.

Architectural complexity arises from several sources. The dual-branch structure of TCMixer processes both temporal dependencies (through token-mixing operations) and feature interactions (through channel-mixing operations) in parallel, effectively doubling the computational load compared to single-branch architectures. The temporal-mixing branch performs linear transformations across the time dimension with complexity $O(L \times d_{time})$, while the channel-mixing branch operates across the feature dimension with complexity $O(L \times d_{channel})$, where d_time and d_channel are typically set to 2×L and 2×D respectively as per the TSMixer design [22].

Memory requirements increase due to the need to maintain intermediate activations for both mixing branches throughout the forward pass, as well as storing gradients for both branches during backpropagation. The Informer encoder's multi-head attention mechanism, despite using ProbSparse attention to reduce complexity from $O(L^2)$ to $O(L \log L)$ still requires substantial memory for attention score matrices, particularly when processing long sequences (168 h in our implementation) [17].

Inference latency considerations become critical for real-time prediction applications. While the model achieves superior accuracy, the sequential nature of the encoder-decoder architecture and the dual-branch processing in TCMixer introduce computational overhead compared to simpler recurrent architectures like GRU. This trade-off between accuracy and computational efficiency is well-documented in deep learning literature and represents a fundamental challenge in deploying sophisticated models for operational air quality forecasting systems [34, 42].

The model's cross-regional transferability faces additional challenges, with uncertain effectiveness when directly applied to untrained cities. From experimental results, Shanghai and London show significant differences in $PM_{2.5}$ concentration characteristics (Shanghai: 2-378 μg/m³, mean 28.23 μg/m³; London: 0-121.56 μg/m³, mean 8.08 μg/m³). Differences in climate types, emission source structures, and monitoring networks among different cities all affect model applicability, requiring careful consideration of domain adaptation strategies. The current framework assumes relatively stable emission source characteristics and meteorological patterns, which may not adequately consider the impacts of rapid urban development or climate change, factors that could alter fundamental pollution dynamics over time.

Comprehensive Improvement Framework and Future Prospects

Based on the detailed root cause analysis of model limitations, we propose a systematic improvement framework comprising multiple interconnected technical strategies, each grounded in established methodologies from recent literature.

To address the challenge of sudden pollution event prediction, we propose a multi-component enhancement framework based on established deep learning techniques. Concentration-aware weighted loss functions can be implemented to assign higher weights to extreme pollution events, following approaches successfully applied in imbalanced time series prediction [41]. The weighted loss can be formulated as $L_{weighted} = \Sigma\, w(y_i) \cdot L(\hat{y}_i, y_i)$, where $w(y_i)$ increases exponentially with concentration levels to emphasize rare extreme events. This approach has been shown to improve model sensitivity to minority classes without sacrificing overall performance. Additionally, expanding the input feature space to incorporate sudden event precursor signals through multi-source data fusion would be valuable. Research has demonstrated the benefits of integrating external trigger information such as satellite observations, industrial activity monitoring, meteorological warnings, and traffic flow data, which can provide early signals of pollution events [31, 35]. Technical implementation would employ heterogeneous feature encoders that process diverse data types before fusion with the main architecture, following successful multi-modal fusion approaches in environmental prediction. Furthermore, addressing temporal receptive field limitations through hierarchical architectures that process both long-range context (capturing synoptic-scale patterns) and high-resolution recent dynamics would enable the model to detect atmospheric circulation changes that precede extreme pollution events by several days, inspired by multi-scale temporal modeling in weather forecasting [36].

To mitigate sensitivity to input data quality issues, several robustness enhancement strategies can be implemented. Meta-learning-based calibration correction approaches that learn to detect and compensate for systematic sensor biases using historical calibration data would involve training auxiliary networks to predict and correct measurement biases based on sensor

metadata and historical performance patterns. Replacing deterministic interpolation with probabilistic approaches (e.g., Gaussian Process regression) that preserve natural variability and uncertainty would improve missing data handling, particularly when combined with hybrid approaches that integrate spatial interpolation from nearby stations with temporal modeling for longer data gaps. Augmenting training with adversarial examples that simulate realistic data quality issues would force the model to learn robust features invariant to small input perturbations, a technique that has proven effective in improving model robustness across various domains.

Computational resource requirements can be addressed through established model compression techniques. Implementing city-specific model configurations that allocate computational resources proportionally to environmental complexity would allow cities with simpler pollution patterns (like London) to achieve adequate performance with reduced model capacity, while complex environments (like Shanghai) benefit from full architectural sophistication. Training lightweight student models that mimic full model predictions through knowledge distillation while requiring fewer computational resources has been successfully applied in deploying complex models for real-time applications [43].

Cross-regional adaptability can be enhanced through domain adaptation techniques. Integrating static city characteristics (climate type, emission source profiles, geographical features) as auxiliary inputs helps the model adapt to different urban contexts [18]. Implementing differentiated normalization strategies based on city pollution characteristics – where high-variability cities benefit from segmented normalization approaches while low-variability cities perform better with global normalization – would improve adaptation. Dividing model parameters into general layers (processing universal time series patterns) and city-specific layers (capturing local pollution characteristics) enables rapid adaptation to new cities by freezing general parameters and fine-tuning only city-specific layers with limited local data. Employing multi-task learning approaches that treat predictions for different cities as related tasks enables the model to learn shared representations while maintaining city-specific adaptations.

Beyond these immediate improvements, several promising research directions emerge for advancing urban air quality prediction capabilities. Incorporating satellite observations and real-time emission inventory data would enhance responsiveness to sudden environmental changes. Developing probabilistic prediction frameworks that provide reliable confidence intervals is crucial for risk-based decision-making in air quality management. Researching online learning systems that continuously update model parameters as new data become available would enable the model to adapt to evolving urban characteristics and climate patterns. Implementing attention visualization and feature importance analysis techniques would provide interpretable insights into model predictions, facilitating trust and adoption by environmental management agencies. These systematic improvements, grounded in established methodologies and recent advances in deep learning for environmental applications, will significantly enhance TC-MixerInformer's robustness, efficiency, and cross-regional adaptability to meet diverse urban air quality prediction needs globally.

Having established the technical capabilities and limitations of TC-MixerInformer, we now examine how these advancements translate into practical applications for urban environmental governance and public health protection.

Practical Implications for Environmental Management and Public Health

The TC-MixerInformer model's technical capabilities translate into three critical practical applications for urban environmental governance and public health protection, validated through the Shanghai and London case studies.

Early Warning Systems for Pollution Episodes

The model's 24-168 h prediction horizon with maintained accuracy (Shanghai 168 h: RMSE = 8.000 μg/m³, R^2 = 0.894) enables authorities to issue pollution alerts 1-7 days in advance, providing sufficient lead time for implementing mitigation measures. For Shanghai's extreme pollution events approaching 350+ μg/m³ (as captured in Fig. 8 and 9), 48-72 h advance warnings allow implementation of emergency response protocols, including temporary traffic restrictions, industrial emission controls, and construction activity suspensions. In London's context, the model's sensitivity to moderate pollution peaks (20-40 μg/m³) supports preemptive public health advisories for vulnerable populations (children, elderly, individuals with respiratory conditions). The cross-regional validation demonstrates that a single model framework can serve cities with fundamentally different pollution regimes, reducing the need for city-specific model development [44].

Quantitative Health Impact Assessment

Accurate extreme pollution prediction directly supports quantitative health risk assessment and intervention planning. Shanghai's frequent episodes exceeding 100 μg/m³ (95^{th} percentile: 72 μg/m³, Table 2) are associated with acute respiratory health impacts. The model's capability to predict these episodes 24-48 h in advance enables targeted interventions: (1) proactive distribution of protective masks to vulnerable populations in affected districts; (2) rescheduling of outdoor activities in schools and elderly care facilities; (3) pre-positioning of medical resources in hospitals, anticipating increased respiratory

emergency visits. For London, where concentrations remain generally lower (mean 8.08 μg/m³) but occasionally exceed WHO guidelines (95th percentile: 21.98 μg/m³, Table 3), the model's 168 h predictions support weekly air quality planning for outdoor events and urban green space management. While direct health outcome validation requires longitudinal epidemiological studies beyond this work's scope, the demonstrated prediction accuracy provides the technical foundation for integrating air quality forecasts into public health early warning systems, potentially reducing pollution-related health burdens through timely preventive actions [45].

Baseline Establishment for Policy Impact Assessment

The model's accurate representation of baseline pollution dynamics (evidenced by consistently high R^2 values across 24-168 h horizons in both cities, Figs 6 and 7) provides a critical foundation for evaluating emission control policy effectiveness [46]. By establishing expected concentration trajectories under business-as-usual conditions, the model enables post-implementation assessment of policy interventions through comparison of observed versus predicted concentrations. For example, if Shanghai implements emergency traffic restrictions or industrial emission controls during a predicted pollution episode, deviations between actual measurements and model forecasts can quantify the intervention's immediate effectiveness. The model's 24-168 h prediction window aligns with policy evaluation timescales, allowing assessment of both short-term emergency responses (24-48 h traffic restrictions) and sustained impacts of regulatory measures (weekly industrial emission adjustments). The cross-regional validation framework established in this study – Shanghai representing developing megacities with high pollution variability and London representing developed urban areas with lower baseline concentrations – demonstrates that consistent baseline modeling across different urban contexts is achievable [47]. This consistency enables comparative policy analysis, allowing cities to evaluate the transferability of emission control strategies across different pollution regimes and urban development stages. Furthermore, the model's ability to maintain accuracy across extreme events and moderate fluctuations ensures that policy impact assessments remain reliable across the full spectrum of pollution conditions encountered in real-world urban environments.

Conclusions

This study successfully developed TC-MixerInformer, a novel deep learning framework that addresses key challenges in cross-regional urban air quality prediction. The integration of Reversible Instance Normalization (RevIN) with Temporal-Channel Mixer (TCMixer) enables superior adaptability to diverse pollution characteristics across Shanghai and London monitoring stations. Key achievements include: (1) Consistent performance improvements with 8-54% error reductions compared to baseline models across all prediction horizons; (2) Effective handling of both Shanghai's high-concentration pollution events (up to 350+ μg/m³) and London's lower-concentration patterns (mean 8.08 μg/m³); (3) Maintained stability from short-term (1-12 h) to long-term (24-168 h) predictions; (4) Demonstrated cross-regional generalization capabilities across different urban pollution regimes.

The proposed architecture provides a practical solution for environmental monitoring systems requiring both temporal stability and geographical adaptability, supporting more effective public health protection strategies in smart city frameworks.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 12275179. The authors gratefully acknowledge the National Urban Air Quality Real-time Publishing Platform of China's Environmental Monitoring Center and the London Air Quality Network for providing the air quality monitoring data. We also thank the Meteostat platform for the meteorological data used in this study.

Conflict of Interest

The authors declare no conflict of interest.

References

1. LI T., SHEN H., YUAN Q., ZHANG X., ZHANG L. Estimating Ground-Level $PM_{2.5}$ by Fusing Satellite and Station Observations: A Geo-Intelligent Deep Learning Approach. Geophysical Research Letters. **44**, (23), **2017**.
2. LIU B., YAN S., LI J., QU G., LI Y., LANG J., GU R. A Sequence-to-Sequence Air Quality Predictor Based on the n-Step Recurrent Prediction. IEEE Access. **7**, **2019**.
3. IZUCHUKWU PRECIOUS O. Air pollution and public health: examining the correlation between $PM_{2.5}$ levels and respiratory diseases in major cities in Nigeria. Journal of Theory, Mathematics and Physics. **4**, (4), 1, **2025**.
4. ZHANG M., TAN S., PAN Z., HAO D., ZHANG X., CHEN Z. The spatial spillover effect and nonlinear relationship analysis between land resource misallocation and environmental pollution: Evidence from China. Journal of Environmental Management. **321**, 115873, **2022**.
5. CAO J., ZHANG M., CHEN E. The Dynamic Effects of Ecosystem Services Supply and Demand on Air Quality: A Case Study of the Yellow River Basin, China. Polish Journal of Environmental Studies. **34** (6), 8043, **2025**.
6. LU Z., ZHANG M., HU C., MA L., CHEN E., ZHANG C., XIA G. Spatiotemporal changes and influencing factors of the coupled production–Living–Ecological functions

in the Yellow River Basin, China. Land. **13** (11), 1909, **2024**.
7. PENG H., LOU H., YANG Y., HE Q., LIU Y., CHEN E., ZHANG M. Spatial and temporal heterogeneity of human-air-ground coupling relationships at fine scale. Polish Journal of Environmental Studies. **2025**.
8. BAKER K.R., FOLEY K.M. A nonlinear regression model estimating single source concentrations of primary and secondarily formed $PM_{2.5}$. Atmospheric Environment. **45** (22), 3758, **2011**.
9. BAI L., WANG J., MA X., LU H. Air pollution forecasts: An overview. International Journal of Environmental Research and Public Health. **15** (4), 780, **2018**.
10. WU N., BRADELY G., XUE B., SHAWN O.B. Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case. ArXiv. **2020**.
11. DING W., ZHU Y. Prediction of $PM_{2.5}$ Concentration in Ningxia Hui Autonomous Region Based on PCA-Attention-LSTM. Atmosphere. **13** (9), **2022**.
12. QI Y., LI Q., KARIMIAN H., LIU D. A hybrid model for spatiotemporal forecasting of $PM_{2.5}$ based on graph convolutional neural network and long short-term memory. Science of The Total Environment. **664**, 1, **2019**.
13. VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A.N., KAISER Ł., POLOSUKHIN I. Attention is all you need. ArXiv. **2017**.
14. YANG J., YAN R., NONG M., LIAO J., LI F., SUN W. $PM_{2.5}$ concentrations forecasting in Beijing through deep learning with different inputs, model structures and forecast time. Atmospheric Pollution Research. **12** (9), **2021**.
15. CHEN J., LU J., AVISE J.C., DAMASSA J.A., KLEEMAN M.J., KADUWELA A.P. Seasonal modeling of $PM_{2.5}$ in California's San Joaquin Valley. Atmospheric Environment. **92**, **2014**.
16. YU T., WANG Y., HUANG J., LIU X., LI J., ZHAN W. Study on the regional prediction model of $PM_{2.5}$ concentrations based on multi-source observations. Atmospheric Pollution Research. **13** (4), **2022**.
17. ZHOU H., ZHANG S., PENG J., ZHANG S., LI J., XIONG H., ZHANG W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. ArXiv. **2021**.
18. WANG L., GENG X., MA X., LIU F., YANG Q. Cross-city transfer learning for deep spatio-temporal prediction. ArXiv. **2019**.
19. LIU Y., WU H., WANG J., LONG M. Non-stationary transformers: Exploring the stationarity in time series forecasting. Advances in Neural Information Processing Systems. **35**, 9881, **2022**.
20. KIM T., KIM J., TAE Y., PARK C., CHOI J.H., CHOO J. Reversible instance normalization for accurate time-series forecasting against distribution shift. 10th International Conference on Learning Representations, **2022**.
21. ORESHKIN B.N., CARPOV D., CHAPADOS N., BENGIO Y. Meta-learning framework with applications to zero-shot time-series forecasting. ArXiv. **2021**.
22. EKAMBARAM V., JATI A., NGUYEN N., SINTHONG P., KALAGNANAM J. TSMixer: Lightweight MLP-Mixer Model for Multivariate Time Series Forecasting. The 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, **2023**.
23. GUO G., WANG H., BELL D., BI Y., GREER K. KNN model-based approach in classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). **2888**, 986, **2003**.
24. KIM H.S., PARK I., SONG C.H., LEE K., YUN J.W., KIM H.K., JEON M., LEE J., HAN K.M. Development of a daily PM_{10} and $PM_{2.5}$ prediction system using a deep long short-term memory neural network model. Atmospheric Chemistry and Physics. **19** (20), **2019**.
25. SHIH S.Y., SUN F.K., LEE H.Y. Temporal pattern attention for multivariate time series forecasting. Machine Learning. **108** (8-9), **2019**.
26. ZHANG T., CUI Z., WANG B., REN Y., YU H., DENG P., WANG Y. PreMixer: MLP-Based Pre-training Enhanced MLP-Mixers for Large-scale Traffic Forecasting. ArXiv. **2024**.
27. CAMPOS D., ZHANG M., YANG B., KIEU T., GUO C., JENSEN C.S. LightTS: Lightweight Time Series Classification with Adaptive Ensemble Distillation. Proceedings of the ACM on Management of Data. **1** (2), **2023**.
28. LIU S., YU H., LIAO C., LI J., LIN W., LIU A.X., DUSTDAR S. Pyraformer: low-complexity pyramidal attention for long-range time series modeling and forecasting. The Tenth International Conference on Learning Representations, **2022**.
29. YANG G., ZHANG Q., YUAN E., ZHANG L. GAT-EGRU: A Deep Learning Prediction Model for $PM_{2.5}$ Coupled with Empirical Modal Decomposition Algorithm. Journal of Systems Science and Systems Engineering. **32** (2), **2023**.
30. HAMILTON M.A., RUSSO R.C., THURSTON R.V. Trimmed Spearman-Karber Method for Estimating Median Lethal Concentrations in Toxicity Bioassays. Environmental Science and Technology. **11** (7), **1977**.
31. WANG R., YE X., HUANG W., LV Z., YAO Y., YANG F., LIU Y., HUO J., DUAN Y. Long-term Trends of $PM_{2.5}$ Composition during Cold Seasons in Shanghai after Releasing Clean Air Action Plan. Aerosol and Air Quality Research. **24** (11), 240085, **2024**.
32. CURLEY L., HOLLAND R., KHAN M.A.H., SHALLCROSS D.E. Investigating the Effect of Fine Particulate Matter ($PM_{2.5}$) Emission Reduction on Surface-Level Ozone (O_3) during Summer across the UK. Atmosphere. **15** (6), 733, **2024**.
33. TOLSTIKHIN I.O., HOULSBY N., KOLESNIKOV A., BEYER L., ZHAI X., UNTERTHINER T., YUNG J., STEINER A., KEYSERS D., USZKOREIT J. Mlp-mixer: An all-mlp architecture for vision. Advances inNeural Information Processing Systems. **34**, 24261, **2021**.
34. JALALI M.W., SAIDI B., FARAHMAND H., PANAH M.A.R., SARUHAN E.N. Scalable AI-driven air quality forecasting and classification for public health applications. Discover Atmosphere. **3** (1), 25, **2025**.
35. ISLAM F.S. A Comprehensive Analysis of Air Pollution in Dhaka City, Bangladesh, and the Application of Artificial Intelligence and Machine Learning for Enhanced Management and Forecasting. International Journal of Applied and Natural Sciences. **3** (1), 131, **2025**.
36. WU Y., CHEN Y., SU X., LIU Z. PolluVCCT: Multi-Scale Hybrid Learning for Robust Air Pollution Forecasting Across Diverse Climate Zones. SIGKDD. **2025**.
37. POELZL M., KERN R., KECORIUS S., LOVRIĆ M. Exploration of transfer learning techniques for the prediction of PM_{10}. Scientific Reports. **15** (1), 2919, **2025**.
38. QIAN W., CHEN J. Regional transport of PM $_{2.5}$ and O_3 based on complex network method and chemical transport model in the Yangtze River Delta, China. Journal of Geophysical Research: Atmospheres. **127** (5), **2022**.

39. CHEN G., WANG Y., TAO C., ZHANG Z., ZHOU M., YAN R., HUANG D.D., WANG H., ZHANG H. Nitrate-driven extreme winter $PM_{2.5}$ pollution in Shanghai, China. npj Clean Air. **1** (1), 28, **2025**.
40. LIU B., WANG L., ZHANG L., BAI K., CHEN X., ZHAO G., YIN H., CHEN N., LI R., XIN J. Evaluating urban and nonurban $PM_{2.5}$ variability under clean air actions in China during 2010-2022 based on a new high-quality dataset. International Journal of Digital Earth. **17** (1), 2310734, **2024**.
41. PANDA S., SINHA A. Advanced AI-Driven Approaches for Predicting Air Quality: A Comprehensive Review. Journal of Computational Analysis & Applications. **33** (6), **2024**.
42. DIVYA J., JAISON B. Enhancing Air Quality Prediction with Hybrid Deep Learning Techniques: A Review. IEEE, **2024**.
43. MA Z., WANG B., LUO W., JIANG J., LIU D., WEI H., LUO H. Air pollutant prediction model based on transfer learning two-stage attention mechanism. Scientific Reports. **14** (1), 7385, **2024**.
44. ZHANG Y., VU T.V., SUN J., HE J., SHEN X., LIN W., ZHANG X., ZHONG J., GAO W., WANG Y. Significant changes in chemistry of fine particles in wintertime Beijing from 2007 to 2017: impact of clean air actions. Environmental Science & Technology. **54** (3), 1344, **2019**.
45. ORGANIZATION W.H. WHO global air quality guidelines: particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization, **2021**.
46. GENG G., XIAO Q., LIU S., LIU X., CHENG J., ZHENG Y., XUE T., TONG D., ZHENG B., PENG Y. Tracking air pollution in China: near real-time $PM_{2.5}$ retrievals from multisource data fusion. Environmental Science & Technology. **55** (17), 12106, **2021**.
47. ZHANG Q., ZHENG Y., TONG D., SHAO M., WANG S., ZHANG Y., XU X., WANG J., HE H., LIU W. Drivers of improved $PM_{2.5}$ air quality in China from 2013 to 2017. Proceedings of the National Academy of Sciences. **116** (49), 24463, **2019**.