

Article

Nonlinear Impacts of Air Pollutants and Meteorological Factors on PM_{2.5}: An Interpretable GT-iFormer Model with SHAP Analysis

Dong Li ^{1,2} , Mengmeng Liu ^{3,4,*}, Houzeng Han ¹ and Jian Wang ^{1,2}

- ¹ School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 102616, China; ld130173@gmail.com (D.L.); hanhouzeng@bucea.edu.cn (H.H.); wangjian@bucea.edu.cn (J.W.)
- ² Research Center for Urban Big Data Applications, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
- ³ School of Emergency Technology and Management, North China Institute of Science and Technology, Sanhe 065201, China
- ⁴ Multi-Scene Water Chain Accident Wisdom Emergency Technology Innovation Center of Hebei, Langfang 065201, China
- * Correspondence: lmm@ncist.edu.cn

Abstract

Accurate prediction of PM_{2.5} concentration is crucial for air quality management and public health protection. However, existing methods often struggle to capture and interpret the nonlinear relationships among multiple atmospheric variables. This study proposes GT-iFormer, a novel interpretable deep learning model that integrates graph convolutional networks (GCNs), Temporal Convolutional Networks (TCNs), and inverted Transformer (iTransformer) for PM_{2.5} concentration prediction. The model features a GTCN-Block that encapsulates GCN and TCN with residual-style fusion, preserving feature-level dependencies alongside temporal patterns to prevent information degradation. The Pearson correlation coefficients and KNN algorithm are innovatively integrated to build a data-driven graph structure, which allows GCNs to flexibly model the nonlinear relationships between pollutants and meteorological factors based on observed data. TCNs obtain multi-scale temporal patterns via causal dilated convolutions. Subsequently, the concatenated representations of GTCN-Block are input into iTransformer to model global inter-variable interactions using attention mechanisms along the axis of the variable. We incorporated SHAP (SHapley Additive exPlanations) analysis to expose feature importance and nonlinear relationships with PM_{2.5} predictions. Results on the hour-level data of Beijing (2020–2021) and Shenzhen (2021) show that our proposed GT-iFormer surpasses all baseline models, with an RMSE of 8.781 $\mu\text{g}/\text{m}^3$ and R^2 of 0.978 for Beijing, and an RMSE of 3.871 $\mu\text{g}/\text{m}^3$ and R^2 of 0.957 for Shenzhen on single-step prediction, equating to RMSE reductions of 15.75% and 17.92%, respectively, over the best baseline model. The SHAP analysis shows clearly distinct regional patterns, with combustion sources dominant in Beijing (represented by CO at 28.231%), and traffic emissions dominant in Shenzhen (represented by NO₂ at 25.908%). Crucial threshold effects are established for all variables, with significant cross-city differences that can serve as general forecasts and guidance for city-specific air quality management policies.

Keywords: PM_{2.5} prediction; GCN; TCN; inverted transformer; SHAP analysis



Academic Editor: William R. Stockwell

Received: 31 January 2026

Revised: 28 February 2026

Accepted: 2 March 2026

Published: 3 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

PM_{2.5} (particulate matter with aerodynamic diameter less than 2.5 µm) has emerged as a primary atmospheric pollutant affecting public health due to its ability to penetrate deep into the respiratory system and its strong association with cardiovascular and respiratory diseases [1]. Driven by rapid industrialization and urbanization in China, the air pollution problem has attracted more and more attention. Accurate forecasting of PM_{2.5} concentration has important practical applications for environmental early warning systems, public health decisions, and pollution prevention and control [2]. But pollutant concentration is co-affected by various atmospheric pollutants and meteorological factors, and complex nonlinear coupling among these factors also brings about challenges to accurate prediction and mechanism analysis [3,4].

In addition to surface meteorological variables, vertical atmospheric structure also plays a crucial role in pollutant dispersion and accumulation. The planetary boundary layer height (PBLH) and atmospheric stability conditions directly regulate the vertical mixing intensity and dilution capacity of urban pollutants [5,6]. A shallow boundary layer combined with stable stratification can significantly suppress vertical diffusion, leading to rapid accumulation of PM_{2.5} near the surface. Previous studies have demonstrated that variations in the PBLH and vertical stability strongly modulate urban air pollution episodes, especially during stagnant weather conditions [7]. Therefore, incorporating meteorological representations that reflect dispersion capacity is essential for improving prediction robustness and physical interpretability. This design concept suits the multivariate characteristics of PM_{2.5} prediction.

Most of the existing approaches for predicting PM_{2.5} can be divided into two categories. The first is the family of numerical simulation methods, including WRF-Chem and Community Multi-scale Air Quality (CMAQ), which make forecasts on physical and chemical processes of pollutant transport, transformation, and deposition within pollutant atmospheres [8,9]. Although these techniques yield a detailed mechanistic understanding of atmospheric processes, the need for extensive computational resources and their sensitivity to the accuracy of emission inventories limit their real-time forecasting applicability [10].

The second group of techniques include statistical learning-based approaches [11]. Traditional statistical models such as multivariate or autoregressive models can capture some basic structures of time series, but these models are not capable of modeling periodic patterns and sudden changes that are typically seen in air pollution time series [12]. Machine learning models have been free from the limitation of linear assumptions and have improved the ability to identify nonlinear patterns, but they are still shallow in depth and limited in terms of hierarchical feature learning [13,14].

The rise in deep learning has led to new advances in the prediction of PM_{2.5}. Recurrent neural networks (RNNs) and their variants LSTM and GRU can effectively model temporal dependencies, but have the problems of vanishing gradients and low sequential computing efficiency [15]. Temporal Convolutional Networks (TCNs) construct exponentially increasing receptive fields by a hierarchical architecture based on causal dilated convolutions, which enables concurrent learning of multi-scale temporal patterns with theoretical advantages in computational efficiency and gradient stability. However, these approaches are mainly focused on modeling the temporal dimension and thus poorly capture the temporal dynamics of inter-variable relationships [16].

With the rise in Graph Neural Networks (GNNs), a new angle emerges to model the relationships between variables [17]. Graph convolutional networks (GCNs) generalize convolution to graphs by aggregating neighbor information, extend multi-layer perceptron (MLP) and recurrent neural network (RNN) models with a parameterized spatial graph convolutional filter in order to capture the local structural characteristics of graphs, and

employ a message-passing mechanism to incorporate neighbor node information [18]. In predicting air quality, pollutants and meteorological factors, such as result factors, the results can be represented as nodes of a graph, which are connected to each other through graph structures that signify the latent relationships among features. The trend learned from the predefined adjacency matrix is as follows: traditional GCNs usually depend on given adjacency matrices and are unable to capture the dynamic correlations among variables directly from data. In addition, traditional GCNs treat all neighbor nodes equally when aggregating information, which impairs the effects of node features with different significance [19,20].

Transformer architectures, which form the backbone of modern natural language processing (NLP) models, are able to capture long range dependencies via self-attention and have also recently led to state-of-the-art results in time series forecasting [21,22]. Instead of performing attention along the temporal dimension, the iTransformer framework introduces attention mechanisms along the variable dimension, considering each feature channel as a single token that can be used to explicitly model global inter-variable exchanges for the overall time series [23,24]. This design concept suits the multivariate characteristics of PM_{2.5} prediction. Recent studies have begun to explore the application of iTransformer and related inverted-attention architectures in environmental and geoscience time series forecasting tasks, including air quality prediction, hydrological runoff modeling, and renewable energy forecasting [25]. These studies highlight the advantages of variable-wise attention mechanisms in capturing cross-feature interactions and improving multivariate forecasting stability. Such evidence further supports the suitability of the iTransformer framework for complex atmospheric systems characterized by strong inter-variable coupling [26].

To sum up, the existing methods have two basic limitations. Single models have difficulty modeling inter-variable dependencies, as well as local and global temporal features at three levels of information. The GCN is good at learning variable correlations but ignores temporal patterns, the TCN is good at learning temporal patterns but cannot learn variable correlations, and iTransformer is good at learning global interactions but needs high-quality feature inputs. The question of how to somehow combine the strengths of these three types of models for joint extraction of multi-level features remains an open problem to be addressed. Third, the deep learning model is difficult to interpret, which limits its application in environmental management. Decision-makers need not only accurate prediction results, but also clear insights into the main driving factors of changes in PM_{2.5} concentrations to design effective measures for controlling pollution.

To this end, GT-iFormer is introduced by fusing GCNs, Temporal Convolutional Networks (TCNs) and inverted Transformer, which is a model that is capable of capturing spatial-temporal dependencies while preserving interpretability. The main novelty lies in the graph construction strategy by using Pearson correlation coefficients with the KNN algorithm to construct a data-driven graph structure, which allows GCNs to adaptively explore nonlinear relationships among pollution and meteorological variables from raw observational data. The model implements a GTCN-Block that combines GCNs and TCNs with residual-style fusion to retain feature-level dependencies in conjunction with temporal information to avoid information loss in temporal aggregation. The iTransformer module also discovers global inter-variable relations through an attention mechanism along the variable dimension as a side effect. In addition, SHAP analysis is embedded to provide transparent model interpretability in terms of the influence of air pollutants and meteorological factors on PM_{2.5} predictions. This work makes the following contributions:

- (1) A new GTCN-Block design: This study proposed the GTCN-Block, which enables the residual-style fusion of GCNs and TCNs. This architecture is designed to capture inter-variable dependencies via GCNs as well as the multi-scale temporal patterns

via TCNs simultaneously, and the residual connections ensure that both types of information are well preserved for the next processing stage.

- (2) Hierarchical feature complementarity: The fused output of GTCN-Block is fed into the iTransformer module that models global feature interactions via attention along the variable dimension. Therefore, it implies hierarchical complementarity, i.e., the iTransformer projects the local, refined features extracted by the GTCN-Block into a global space.
- (3) Adaptive graph construction: This study proposes a novel, data-driven graph construction method based on the fusion of Pearson correlation coefficients and the KNN algorithm. This approach adaptively captures statistical dependencies among variables from observational data without prior knowledge, and outputs a meaningful and dynamic topological structure for GCN message passing.
- (4) Better model interpretability: This study adds SHAP analysis for post hoc interpretability. This offers a way to measure each input feature's share on the final PM_{2.5} prediction, exposing nonlinear influence patterns and making the model interpretable for actionable decision support in air quality management.

2. Materials and Methods

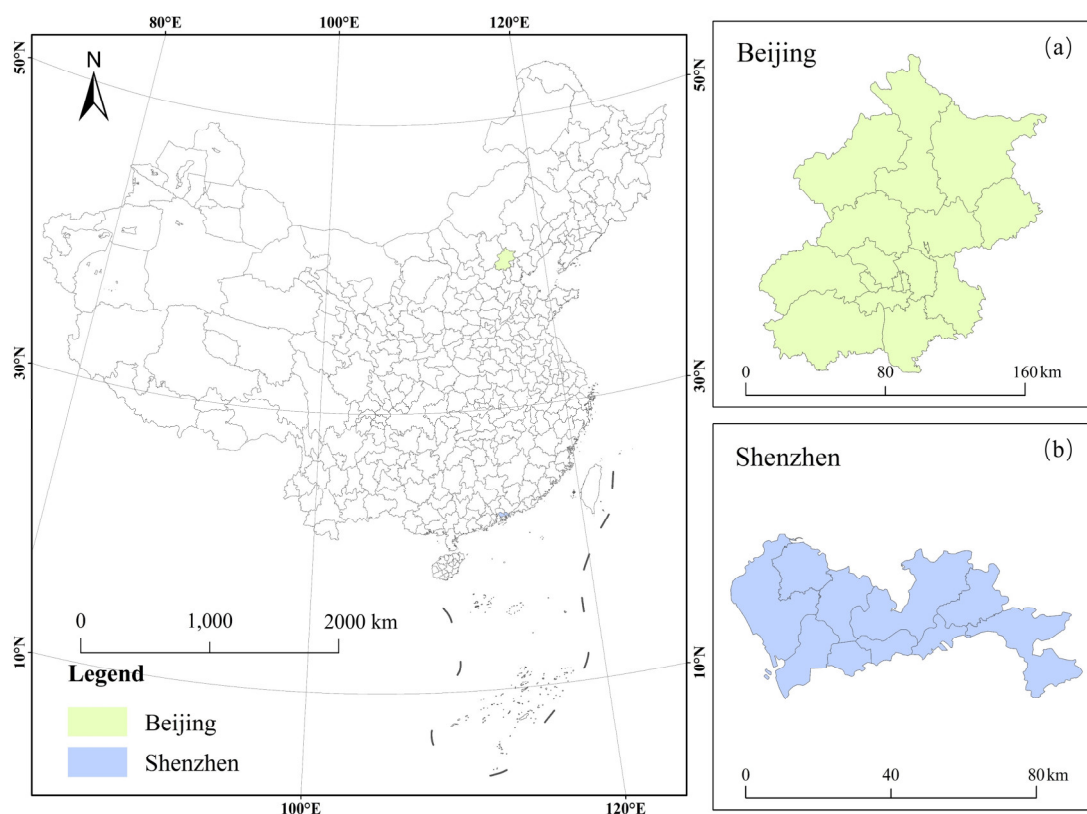
2.1. Study Area and Dataset

This study is based on two typological megacities in China: Beijing and Shenzhen (Figure 1). These cities were chosen because of their different geographical locations, climatic conditions and regional socioeconomic development modes, which represent different cases for a test on the generalization of the proposed model. Located on the northern part of the North China Plain, Beijing is subject to high PM_{2.5} levels in autumn and winter as a result of adverse meteorological conditions, including prevailing stagnant air masses or thermal inversions [27]. Shenzhen, located in southern China and influenced by a subtropical maritime climate, has a different seasonal variation in pollution from those observed in northern cities, and, being adjacent to the Pearl River Delta (PRD) industrial region [28], is surrounded by a more complex atmospheric environment. The analysis of pollution dynamics in these two cities provides crucial empirical information for formulating tailored environmental policies that could be adapted to a broader range of urban areas within China.

This study employed hourly pollution concentrations and meteorological data in the two cities under study. The Beijing dataset covers from 1 January 2020, to 31 December 2021, a full two years of data. The Shenzhen datasets are from 1 January 2021, to 31 December 2021, with one full year of hourly data in each. Table 1 provides descriptive statistics for both datasets in detail. The method of analysis deals with four air pollutants—sulfur dioxide (SO₂), nitrogen dioxide (NO₂), ozone (O₃), and carbon monoxide (CO)—and five meteorological variables and one categorical weather condition indicator—temperature, pressure, relative humidity, wind speed, wind direction, and weather. Measurements of pollution were taken from the air quality database (<https://quotsoft.net/air/> [29], accessed on 1 March 2023); however, meteorological data and one categorical weather condition indicator were obtained from the OpenWeather source (<https://openweathermap.org/> [30], accessed on 1 March 2023).

Table 1. Basic data information from Beijing and Shenzhen.

Parameters	Beijing			Shenzhen		
	Max	Min	Missing (%)	Max	Min	Missing (%)
PM _{2.5} ($\mu\text{g} \cdot \text{m}^{-3}$)	742	1	1.31	203	1	1.45
SO ₂ ($\mu\text{g} \cdot \text{m}^{-3}$)	43	1	1.53	35	1	1.68
NO ₂ ($\mu\text{g} \cdot \text{m}^{-3}$)	144	1	1.48	194	1	1.91
O ₃ ($\mu\text{g} \cdot \text{m}^{-3}$)	322	1	3.16	300	1	3.05
CO ($\text{mg} \cdot \text{m}^{-3}$)	4	0.1	1.62	1.9	1	1.71
Temperature (°C)	313.7	254.52	0	309.29	279.76	0
Pressure (Pa)	1048	986	0	1031	994	0
Relative humidity (%)	100	6	0	99	24	0
Wind speed (m/s)	8.59	0.02	0	9.7	0	0
Wind direction (°)	360	0	0	360	0	0
Weather	11	1	0	13	1	0

**Figure 1.** Study area. (a) Beijing; (b) Shenzhen.

Data preprocessing went through several steps to maintain analytical soundness. The raw air quality and meteorological observations were originally recorded at an hourly resolution by national monitoring stations. Therefore, the temporal sampling frequency used in this study is one hour, consistent with the original data acquisition system. As the percentage of missing data in both datasets was less than 4% of the total number of records, we filled the missing values by linear interpolation. In this study, a 4% threshold was adopted as the exclusion criterion. If the missing data proportion for a specific variable or time segment had exceeded 4%, the corresponding records would have been removed

from the analysis to avoid introducing excessive interpolation bias. However, since the missing rates in both datasets were below this threshold, no records were excluded. Outlier detection was further conducted to ensure data reliability. Observations with physically implausible values (e.g., negative pollutant concentrations) were removed. Identified outliers were replaced using linear interpolation to preserve temporal continuity. Descriptions of the weather needed to be converted into numbers for the computer to process. As shown in Table 2, the precipitation intensity levels were encoded as follows, light rain is 1, moderate rain is 2, and heavy-intensity rainfall is 3, and other weather activities have codes in the same manner [31]. Then, in order to eliminate units among variables and speed up model convergence, min–max normalization was performed on all variables to scale to the [0, 1] range [30]. This uniform processing scheme for the two cities guarantees fairness of comparison and demonstrates the robustness of the model under different atmospheric situations. The air quality data were obtained from the official national monitoring platform. However, the publicly available dataset does not disclose the exact installation heights of the monitoring stations. Therefore, station elevation information could not be incorporated into the present analysis.

Table 2. Numerical mapping representation of weather conditions.

Description of Weather Conditions	Numerical Conversion
Light rain	1
Moderate rain	2
Heavy-intensity rain	3
Very-heavy rain	4
Light snow	5
Snow	6
Sky is clear	7
Few clouds	8
Scattered clouds	9
Broken clouds	10
Overcast clouds	11
Mist	12
Haze	13

2.2. Research Framework

The SHAP-based interpretable PM_{2.5} prediction model is depicted in Figure 2. The overall framework consists of three stages, which together enable accurate prediction with prompt interpretability.

The first stage is data collection and preprocessing. The PM_{2.5} concentration is the output variable, and the input includes six meteorological parameters and four atmospheric pollutants. The raw observational data are processed through a series of preprocessing steps, such as filling missing values, converting categorical variables, and using min–max normalization to meet data quality and model requirements.

With the second stage at the heart of our proposed scheme, we introduce GT-iFormer, the novel architecture that achieves high-accuracy and interpretable PM_{2.5} prediction. There are two main parts to the architecture: the GTCN-Block and the iTransformer module. In the GTCN-Block, the GCN first builds a correlation-based graph on the feature space through Pearson coefficients and K-Nearest Neighbors and selection, then the GCN performs its

spectral graph convolution to capture nonlinear dependencies among input features in the form of message-passing mechanisms.

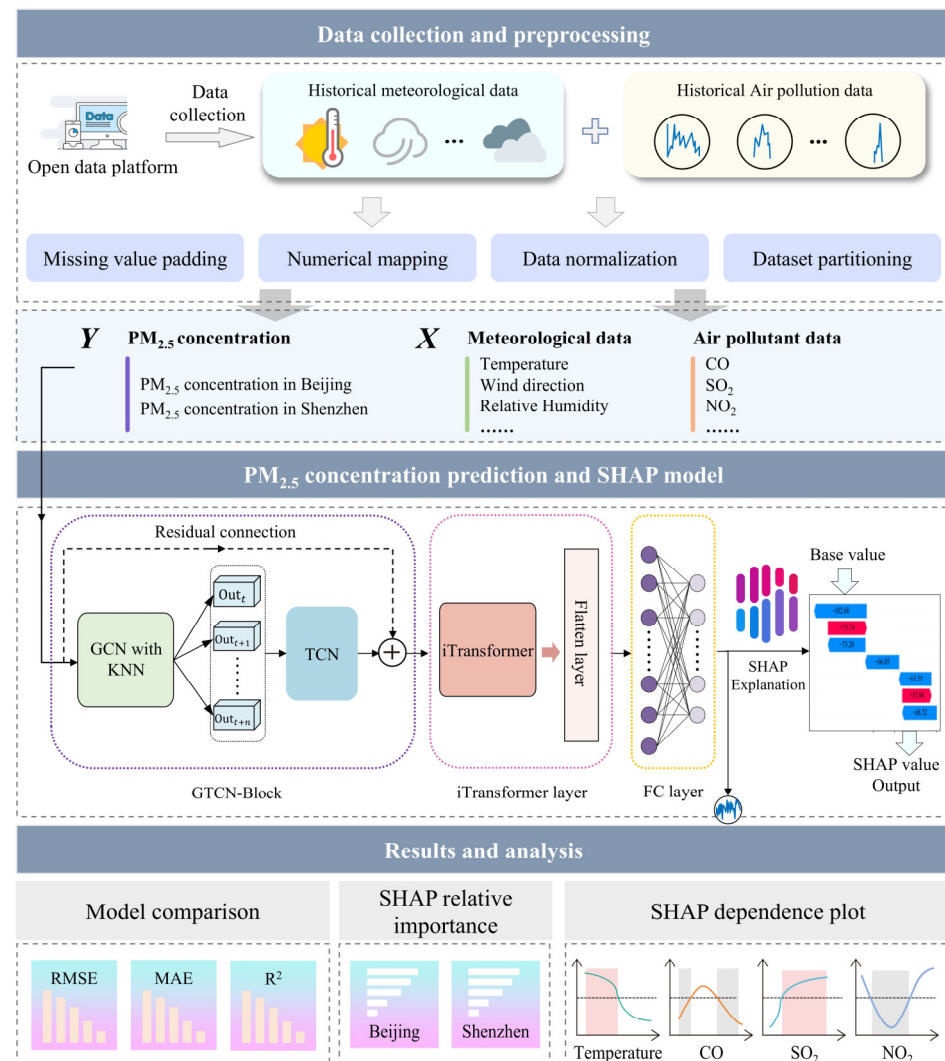


Figure 2. Research framework.

Then, the temporal TCN applies causal dilated convolutions on the GCN-enhanced representations, capturing multi-scale temporal patterns with exponentially growing receptive fields. The outputs of the two parts are concatenated by a residual-type fusion method that maintains the feature-level dependencies as well as temporal dynamics. The resulting fused representations are then passed to the iTransformer module, which utilizes multi-head self-attention over the variable dimension to directly capture global inter-variable interactions among all variables of the time series. Subsequently, the attended representations are passed into fully connected layers to give the predictions of PM_{2.5}. Furthermore, we incorporate SHAP analysis to measure the contribution of each feature.

The last stage covers model testing and interpretability investigation. The predictive performance is evaluated with various measures (RMSE, MAE and R²) on several baseline models. Interpretability analysis uses the SHAP methodology: a global importance analysis sorts input variables by their contributions, and dependence plots explore nonlinear patterns and threshold effects on PM_{2.5} concentration. These results verify the reliability of the model and offer a scientific basis for air pollution control policies.

2.3. GTCN-Block: Graph Temporal Feature-Extraction Module

In order to capture both complicated relationships between input features and temporal dynamics in air pollution data at the same time, we propose the GTCN-Block, a unified unit that combines GCNs and TCNs within residual-fashion feature fusion. The GTCN-Block is divided into three steps when processing input data: (1) graph convolution for modeling the dependency of the features, (2) temporal convolution for capturing multi-scale patterns, and (3) residual fusion for maintaining the information at the feature level and temporal level.

2.3.1. Graph Convolutional Network

To capture high-order dependencies among the input features, we use GCNs that treat each feature as a node on a graph and infer the latent relations between features by message-passing mechanisms. A crucial aspect of GCNs is how to construct a suitable graph adjacency matrix, which contains the inherent correlations among features [32]. In this paper, we employ a doublets-based approach utilizing Pearson correlation coefficients along with KNN selection. Given the input feature matrix $\mathbf{X} \in \mathbb{R}^{T \times N}$, where T denotes the number of time steps and N represents the number of features (including four pollutants, SO_2 , NO_2 , O_3 , and CO , and six meteorological variables), we first compute the correlation matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$:

$$C_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sigma_{X_i} \cdot \sigma_{X_j}} \quad (1)$$

where X_i and X_j represent the i -th and j -th feature columns, $\text{Cov}(\cdot, \cdot)$ denotes the covariance function, and σ represents the standard deviation. To build a sparse yet meaningful graph topology, we perform KNN selection to identify the adjacency relationships. For each node i , we identify the K nodes with the highest absolute correlation values and establish edges accordingly:

$$A_{ij} = \begin{cases} 1, & \text{if } j \in \mathcal{N}_K(i) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where $\mathcal{N}_K(i)$ denotes the set of K nearest neighbors of node i based on the absolute correlation values, with $K = 5$ in this study. This implies that for every feature variable, an edge is created with the five variables most highly correlated to it, which are the first, second, third, fourth, and fifth most highly correlated variables. Self-connections are added to retain the individual's own features of a node in the aggregation process (i.e., $A_{ij} = 1$ for all i). This correlation-based graph construction approach offers clear physical interpretability: features with stronger statistical dependencies are connected, enabling information propagation along meaningful pathways.

As shown in Figure 3, on the constructed graph, we perform spectral graph convolution, which aggregates neighbor information by a localized first-order approximation. The layer-wise propagation rule is defined as follows:

$$\mathbf{H}^{(l+1)} = \sigma(\tilde{\mathbf{D}}^{-\frac{1}{2}} \cdot \tilde{\mathbf{A}} \cdot \tilde{\mathbf{D}}^{-\frac{1}{2}} \cdot \mathbf{H}^{(l)} \cdot \mathbf{W}^{(l)}) \quad (3)$$

where $\mathbf{H}^{(l)} \in \mathbb{R}^{N \times d_l}$ represents the node's feature matrix at layer l , with $\mathbf{H}^{(0)} = \mathbf{X}$ being the input features. $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N$ is the adjacency matrix with added self-loops, and $\tilde{\mathbf{D}}$ is the corresponding degree matrix with $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. $\tilde{\mathbf{D}}$ is the corresponding degree matrix defined at the feature node level, where $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$, and i, j are index feature nodes rather than temporal steps. The graph convolution is applied independently at each time step, meaning that aggregation is performed across feature dimensions while the temporal dimension is preserved for subsequent TCN processing. $\mathbf{W}^{(l)} \in \mathbb{R}^{d_l \times d_{l+1}}$

denotes the trainable weight matrix, and $\sigma(\cdot)$ represents the nonlinear activation function (ReLU in our implementation). The symmetric normalization $\tilde{\mathbf{D}}^{-\frac{1}{2}} \cdot \tilde{\mathbf{A}} \cdot \tilde{\mathbf{D}}^{-\frac{1}{2}}$ ensures that the aggregation operation preserves the scale of feature vectors, preventing numerical instability during training. With this propagation process, the representation of each node is updated by aggregating the information from its neighboring nodes, which enables it to capture the nonlinear relations between pollutant concentrations and meteorological conditions. The output of the GCN module is denoted as $\mathbf{Z} \in \mathbb{R}^{B \times T \times d_g}$, where B is the batch size and d_g represents the GCN's hidden dimension.

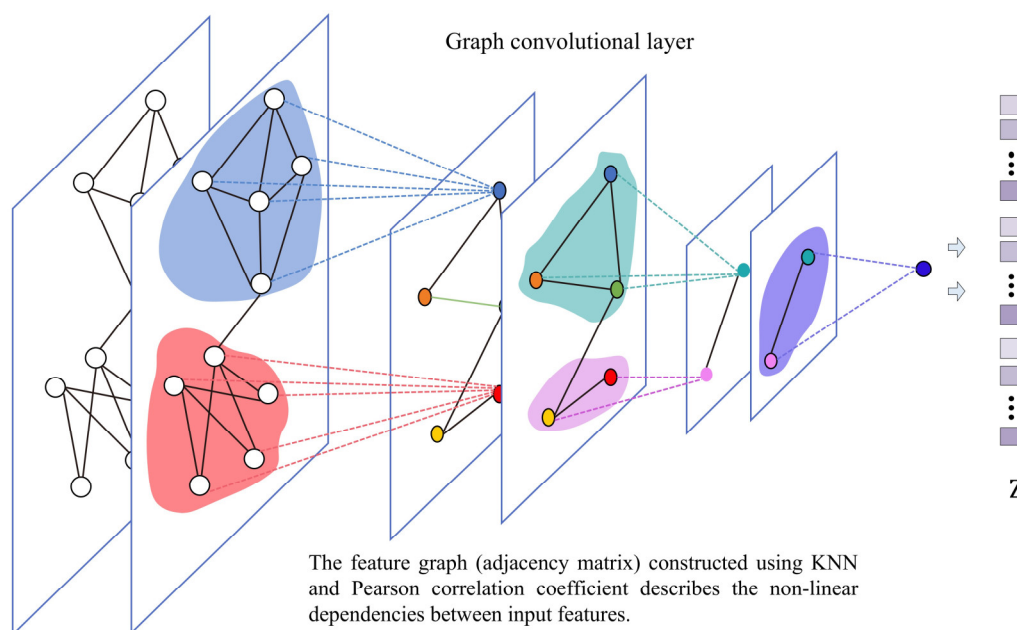


Figure 3. Architecture of graph convolutional network.

2.3.2. Temporal Convolutional Network

Based on the feature-enhanced representations from the GCN, the TCN module captures multi-scale temporal patterns via causal dilated convolutions. As shown in Figure 4, the TCN takes $\mathbf{Z}_{gcn}$ as the input, forming a sequential pipeline within the GTCN-Block. The TCN architecture is composed of stacked residual blocks, each of which is made up of dilated causal convolutions that exponentially increase the size of the receptive field while keeping computation efficient [33]. For the input sequence, the dilated causal convolution operation at block m with dilation factor d_m is defined as follows:

$$F(s) = \sum_{i=0}^{k-1} f(i) \cdot \mathbf{Z}_{in}(s - d_m \cdot i) \quad (4)$$

where k is the kernel size, $f(i)$ is the filter weights and s is the current time step. d_m is the dilation factor that grows exponentially with the network depth (usually $d_m = 2^m$ for block m); thus, it allows the receptive field to grow exponentially while maintaining a linear number of parameters. This structure enables the TCN to extract short-term variances and middle-term trends of $\text{PM}_{2.5}$ concentration without vanishing gradient problems which plague recurrent architectures.

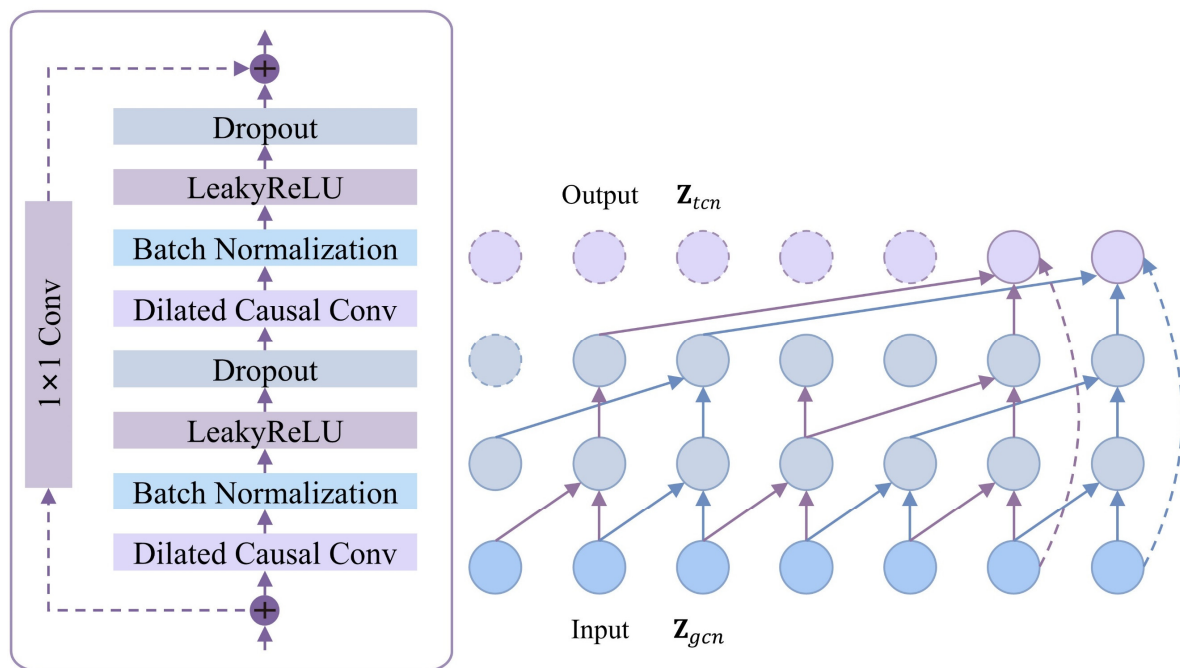


Figure 4. Architecture of TCN.

Two layers of dilated causal convolutions, weight normalization, ReLU activation, and dropout for regularization are employed in each residual block of the TCN. The residual connection helps facilitate gradient flow and allows the network to easily learn identity mappings when necessary.

$$\mathbf{Z}_{block}^{(m)} = \text{ReLU}\left(\mathbf{Z}_{in}^{(m)} - \mathcal{F}\left(\mathbf{Z}_{in}^{(m)}\right)\right) \quad (5)$$

where $\mathcal{F}(\cdot)$ represents the transformation applied by the convolutional layers within the block, and $\mathbf{Z}_{in}^{(m)}$ denotes the input to the m -th residual block. When the input and output dimensions differ, a 1×1 convolution is applied to the residual path to ensure dimensional compatibility. The causal nature of TCN convolutions ensures that predictions at time t depend only on observations from time t and earlier, preserving the temporal causality essential for time series forecasting. The output of the TCN module is denoted as $\mathbf{Z}_{tcn} \in \mathbb{R}^{B \times T \times d_t}$, where d_t represents the TCN's hidden dimension.

2.3.3. Residual Feature Fusion

One of the design principles of the GTCN-Block is residual-type feature fusion, which retains the information of both the GCN and TCN layers. To do so, we concatenate the outputs from the two modules in the feature dimension.

$$\mathbf{Z}_{fused} = [\mathbf{Z}_{gcn}; \mathbf{Z}_{tcn}] \in \mathbb{R}^{B \times T \times (d_g + d_t)} \quad (6)$$

Such residual-style fusion performs a skip connection over the TCN, which allows the inter-variable dependencies captured by the GCN to be directly exposed to the downstream layers. The design has two benefits: (1) it avoids possible information loss during temporal convolution, and (2) it allows the model to utilize the original feature correlations and temporally enriched information. The fused representation combines complementary information: $\mathbf{Z}_{gcn}$ encodes the static relational structure among features, while $\mathbf{Z}_{tcn}$ captures the dynamic temporal evolution of these features.

The fused representation is then projected to the model's hidden dimension through a linear transformation followed by layer normalization:

$$\mathbf{Z}_{proj} = \text{LayerNorm}(\mathbf{Z}_{fused} \cdot \mathbf{W}_{proj} + \mathbf{b}_{proj}) \quad (7)$$

where $\mathbf{W}_{proj} \in \mathbb{R}^{(d_s+d_t) \times d_f}$ and $\mathbf{b}_{proj} \in \mathbb{R}^{d_f}$ are learnable parameters, and d_f denotes the iTransformer's hidden dimension.

2.4. Inverted Transformer

To capture global dependencies across the feature channels, we employ the iTransformer architecture, which applies attention mechanisms along the variable dimension rather than the temporal dimension [34]. As shown in Figure 5, unlike conventional Transformers that treat each time step as a token, iTransformer treats each variable (feature channel) as a token, enabling the model to learn complex interactions among different pollutants and meteorological factors across the entire time series.

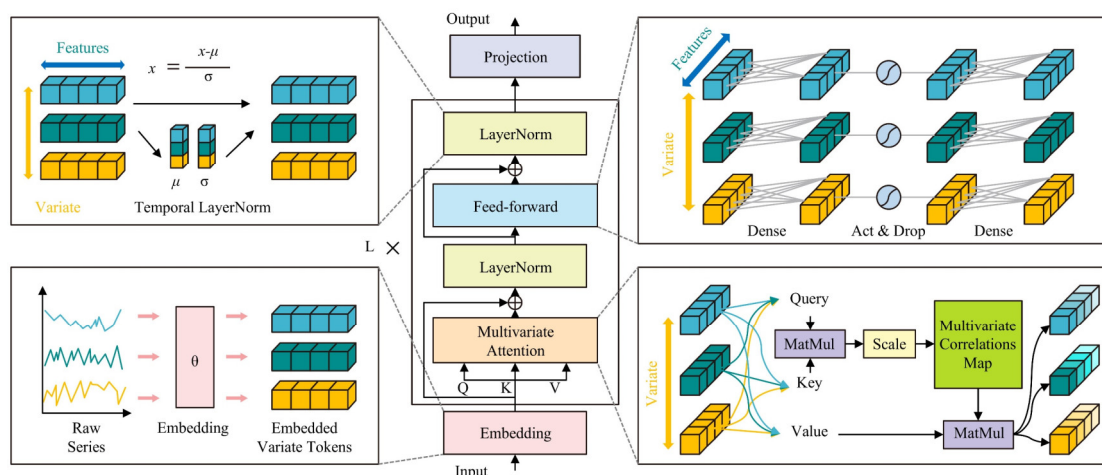


Figure 5. Architecture of iTransformer.

Given the projected representation $\mathbf{Z}_{proj} \in \mathbb{R}^{B \times T \times d_f}$, iTransformer first transposes the sequence to obtain $\mathbf{Z}_{var} \in \mathbb{R}^{B \times d_f \times T}$, where each of the d_f feature channels become a token with T -dimensional embeddings. The multi-head self-attention mechanism is then applied across the variable dimension:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}}\right) \cdot \mathbf{V} \quad (8)$$

where $\mathbf{Q} = \mathbf{Z}_{var} \cdot \mathbf{W}_Q$, $\mathbf{K} = \mathbf{Z}_{var} \cdot \mathbf{W}_K$, and $\mathbf{V} = \mathbf{Z}_{var} \cdot \mathbf{W}_V$ are the query, key, and value matrices, respectively, with $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{T \times d_k}$ being learnable projection matrices. The scaling factor $\sqrt{d_k}$ prevents the dot products from growing too large.

For multi-head attention with h heads, the outputs from each head are concatenated and projected:

$$\text{MultiHead}(\mathbf{Z}_{var}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \cdot \mathbf{W}_O \quad (9)$$

where $\text{head}_i = \text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i)$ and $\mathbf{W}_O \in \mathbb{R}^{h \cdot d_k \times d_f}$ is the output projection matrix.

Each iTransformer encoder layer consists of the multi-head attention sublayer followed by a position-wise feed-forward network (FFN), with residual connections and layer normalization applied to each sublayer:

$$\mathbf{Z}_{attn} = \text{LayerNorm}(\mathbf{Z}_{var} + \text{MultiHead}(\mathbf{Z}_{var})) \quad (10)$$

$$\mathbf{Z}_{out} = \text{FFN}(\mathbf{Z}_{attn} + \text{MultiHead}(\mathbf{Z}_{attn})) \quad (11)$$

The FFN consists of two linear transformations with a ReLU activation:

$$\text{FFN}(\mathbf{z}) = \text{ReLU}(\mathbf{z} \cdot \mathbf{W}_1 + \mathbf{b}_1) \cdot \mathbf{W}_2 + \mathbf{b}_2 \quad (12)$$

where $\mathbf{z}$ denotes the input to the feed-forward network (i.e., the output from the attention sublayer $\mathbf{Z}_{attn}$), $\mathbf{W}_1 \in \mathbb{R}^{d_f \times d_{ff}}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_{ff} \times d_f}$ are learnable weight matrices, $\mathbf{b}_1$ and $\mathbf{b}_2$ are bias terms, and d_{ff} is the feed-forward hidden dimension.

The iTransformer module complements the GTCN-Block by providing a global perspective on inter-variable interactions. While the GTCN-Block captures local feature dependencies (via the GCN) and sequential temporal patterns (via the TCN), the iTransformer learns how different pollutants (SO_2 , NO_2 , O_3 , and CO) and meteorological factors (temperature, humidity, wind speed, etc.) interact with each other across the entire observation period through its attention mechanism.

After processing through the iTransformer encoder layers, we extract the representation at the last time step for the final prediction:

$$\mathbf{Z}_{last} = \text{Transpose}(\mathbf{Z}_{out})(:, -1, :) \in \mathbb{R}^{B \times d_f} \quad (13)$$

The prediction head consists of a multi-layer perceptron (MLP) that maps the final representation to the $\text{PM}_{2.5}$ concentration:

$$\hat{y} = \text{MLP}(\mathbf{Z}_{last}) \in \mathbb{R}^{B \times 1} \quad (14)$$

The entire GT-iFormer model is trained end-to-end using Mean Squared Error (MSE) loss:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^B (y_i - \hat{y}_i)^2 \quad (15)$$

where $\mathcal{L}$ denotes the loss function, and y_i and $\hat{y}_i$ represent the ground truth and predicted $\text{PM}_{2.5}$ concentrations for sample i , respectively.

2.5. Shapley Additive Explanations

While the proposed GT-iFormer model demonstrates strong predictive performance for $\text{PM}_{2.5}$ concentration forecasting, understanding the contribution of each input feature to the predictions is essential for model validation and practical decision-making. To meet this requirement for interpretability, we utilize SHAP, a unified approach for interpreting model predictions that uses the game-theoretic Shapley values.

The SHAP approach attributes an importance value to each feature by calculating its Shapley value, which can be interpreted as the fair allocation of the prediction among the features [35,36]. For a predictive model f with N input features, the Shapley value ϕ_i for feature i is defined as follows:

$$\phi_i = \sum_{S \subseteq \{1, \dots, N\} / \{i\}} \frac{|S|!(n - |S| - 1)!}{N!} \cdot [f(S \cup \{i\}) - f(S)] \quad (16)$$

where S is a subset of features not containing feature i , $|S|$ is the size of S , $f(S \cup \{i\})$ is the model output when feature i is included, and $f(S)$ is the model output when only features in S are present and all other features are replaced by background values. In this study, the expected model output over the training dataset, $E[f(X)]$, is used as the baseline

(null model) for SHAP computation. The SHAP framework guarantees that the sum of all feature attributions equals the difference between the model output and this baseline value.

$$f(x) = \phi_0 + \sum_{i=1}^N \phi_i \quad (17)$$

where ϕ_0 represents the average model output across the training set.

Here, SHAP analysis in our study offers two major implications for PM_{2.5} forecasting. First, the average of absolute SHAP values over all samples gives the global feature importance.

$$\bar{\phi}_i = \frac{1}{M} \sum_{j=1}^M |\phi_i^{(j)}| \quad (18)$$

where M is the number of samples. This measurement shows what meteorological variables and air pollutants have the largest impact on PM_{2.5} estimations. Secondly, SHAP dependence plots reveal the nonlinear relationship between the input features and the predictions. By plotting feature values against associated SHAP values, we can observe more intricate patterns, such as threshold effects or saturation effects, which linear correlation analysis can never detect.

2.6. Evaluation Metrics for Model Performance

To quantitatively measure the prediction performance of the proposed GT-iFormer, three commonly used evaluation metrics are adopted: Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and the Coefficient of Determination (R^2) [37,38]. These indices are complementary in the assessment of prediction accuracy.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (19)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (20)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (21)$$

where y_i is the observed PM_{2.5} concentration, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the average over all observed values, and n is the number of samples. RMSE expresses the standard deviation of residuals and is more sensitive to larger errors in prediction. MAE calculates the average magnitude of the errors between observations and forecasts, offering a well-understood measure of typical forecast accuracy. The R^2 can be viewed as the explained variance in the observed data by the model, as a metric of overall goodness-of-fit. Lower values are better for both RMSE and MAE in terms of prediction accuracy, while better model fitting is indicated by how close the R^2 is to one.

2.7. Parameter Setting

For each city, the dataset was maintained in strict chronological order to preserve temporal dependency. All experiments were implemented using Python 3.9 and the TensorFlow framework. The full time series dataset was divided into three mutually exclusive subsets: training (60%), validation (20%), and testing (20%). The training set was used for model parameter learning, the validation set was employed for hyperparameter tuning and early model selection, and the testing set was reserved exclusively for final performance evaluation.

Optimal hyperparameter settings for GT-iFormer were determined through systematic iterative experimentation. Model performance on the validation subset was evaluated after each training epoch by computing RMSE and MAE metrics. The training process comprised 100 epochs in total. Network weights achieving superior validation performance were dynamically preserved and updated throughout the optimization procedure.

Dropout regularization with a rate of 0.2 was applied in both the TCN and iTransformer modules to mitigate overfitting risks. The Adam optimizer with a learning rate of 0.001 was employed to minimize the MSE loss function. Unless otherwise specified, neural network weights were initialized using the default Glorot uniform initialization provided by TensorFlow, and bias terms were initialized to zeros. Training concluded when the model achieved optimal predictive performance on the validation set. The complete hyperparameter configuration is summarized in Table 3. All experiments were conducted under the same computational environment to ensure consistency and reproducibility.

Table 3. GT-iFormer model parameter.

Layer Name	Parameters	Values
GCN	Hidden dimension	64
	K-neighbors	5
	Number of layers	2
TCN	Hidden channels	64
	Kernel size	3
	Dilation factors	[1, 2, 4, 8]
	Number of layers	4
iTransformer	Hidden dimension	128
	Number of attention heads	8
	Feed-forward dimension	512
	Number of encoder layers	2
Fully connected layer	Hidden layers	[64, 32, 1]
-	Dropout	0.2
-	Batch size	32
-	Learning rate	0.001
-	Epochs	100
-	Training method	Adam
-	Loss function	MSE

3. Results

3.1. Accuracy Comparison with Other Models

A sliding window approach was implemented for single-step prediction, wherein a 3 h lookback window provides input features to forecast PM_{2.5} concentrations at the subsequent time point [39]. To establish performance benchmarks, both conventional machine learning algorithms (RF and XGBoost) and classical deep learning architectures (CNN, TCN, CNN-TCN, and ResNet-TCN) served as baseline models. Additionally, two component variants were constructed to quantify the contribution of individual modules within GT-iFormer: iTransformer represents the global interaction modeling component in isolation, while GTCN-Block encompasses the integrated GCN-TCN fusion module without subsequent Transformer processing. These controlled comparisons provide for a systematic evaluation of the efficacy of each architectural element.

To ensure the reproducibility and stability of the reported results, all experiments were conducted using a fixed random seed and identical data splits across models. Since the training, validation, and testing sets were chronologically determined and remained unchanged for all experiments, the reported performance metrics correspond to a unique deterministic evaluation on the held-out test set. Given the large sample size (hourly observations over multiple years), the variance of evaluation metrics is minimal under repeated runs. Table 4 shows the results of single-step predictions on the Beijing and Shenzhen datasets. One interesting observation is that in all models, the RMSE for Beijing is higher, while the R^2 for Beijing is higher than that for Shenzhen. This seeming paradox is the characteristic of the $PM_{2.5}$ concentration in the two cities. Beijing has higher temporal variability of total $PM_{2.5}$ levels, implying larger absolute but also larger relative prediction errors (larger RMSE and R^2); hence, predicting Beijing is more challenging, as reflected by the R^2 values. In contrast, because of smaller fluctuation ranges of much lower $PM_{2.5}$ concentrations in Shenzhen, the relative errors in Shenzhen are higher due to a lower ratio of signal to noise, absolute errors are lower, and prediction is more difficult in relative terms, as evidenced by the relatively smaller R^2 values. This phenomenon highlights that there are inherent challenges in predicting low-concentration $PM_{2.5}$ even if the magnitude of the absolute error is small.

Table 4. Single-step prediction.

Model	Beijing			Shenzhen		
	RMSE	MAE	R^2	RMSE	MAE	R^2
RF	15.682	11.178	0.905	8.869	6.593	0.874
XGBoost	15.564	11.564	0.906	8.569	6.241	0.880
CNN	13.265	9.987	0.927	7.154	5.458	0.901
TCN	12.941	9.751	0.932	6.623	5.357	0.908
CNN-TCN	11.589	8.564	0.949	5.865	4.674	0.924
ResNet-TCN	10.423	7.571	0.958	4.716	3.787	0.932
iTransformer	10.314	7.054	0.961	4.242	3.374	0.945
GTCN-Block	9.647	5.725	0.972	4.089	2.941	0.950
GT-iFormer	8.781	5.193	0.978	3.871	2.552	0.957

Experimental results illustrate that deep learning models consistently perform better than traditional machine learning models in the two cities. Beijing's RMSE of RF and XGBoost are 15.682 and 15.564, with MAE values of 11.178 and 11.564, correspondingly, while deep learning models attained much smaller error measures. Note that the pure CNN alone reduces the RMSE to 13.265 (Beijing) and 7.154 (Shenzhen), with the MAE being 9.987 and 5.458. This gap in performance further demonstrates that neural network-based frameworks are more effective in modeling the complex nonlinear structures involved in daily $PM_{2.5}$ dynamics. It can be seen that there is a clear trend of enhancement among the deep learning baselines. The TCN performs better than the CNN, with RMSEs decreased from 13.265 to 12.941 (Beijing) and 7.154 to 6.623 (Shenzhen), thanks to dilated causal convolutions that increase the receptive field while not violating the temporal order. The hybrid CNN-TCN approach achieves additional improvements, with RMSEs of 11.589 and 5.865, and MAEs of 8.564 and 4.674 for Beijing and Shenzhen, correspondingly, showing that the integration of different convolutional schemes can better capture temporal features. ResNet-TCN obtains an even better result (RMSE of 10.423 and 4.716; MAE of 7.571 and 3.787) by adding residual connections that ease the flow of the gradients.

The iTransformer module is strong on its own, with RMSEs of 10.314 (Beijing) and 4.242 (Shenzhen), MAEs of 7.054 and 3.374, and R^2 values of 0.961 and 0.945, verifying attention along the variable dimension as an effective means to attend global inter-variable

relationships. GTCN-Block achieves the best results, with RMSEs (MAEs) of 9.647 (5.725) and 4.089 (2.941), and R^2 values of 0.972 and 0.950, indicating the benefits of fusing GCN-based inter-variable dependency modeling and TCN-based multi-scale temporal pattern extraction with residual-style fusion. The full version of the GT-iFormer model yields the best results for all the metrics in both cities: RMSE of 8.781, MAE of 5.193, and R^2 of 0.978 for Beijing; and RMSE of 3.871, MAE of 2.552, and R^2 of 0.957 for Shenzhen. Against GTCN-Block, to which it is superior, further integrating iTransformer leads to 8.98% (B) and 5.33% (S) reductions in RMSE, as well as MAE decreases of 9.29% and 13.23%, correspondingly, indicating that modeling global feature interactions has extra predictiveness beyond local spatiotemporal patterns. Against the strongest baseline (ResNet-TCN), GT-iFormer brings declines of 15.75% (Beijing) and 17.92% (Shenzhen) in RMSE, decreases of 31.40% and 32.61% in MAE, and increases in the R^2 from 0.958 to 0.978 and 0.932 to 0.957, respectively. Given the large number of testing samples, all reported R^2 values are statistically significant at $p < 0.001$ when tested against the null hypothesis of zero correlation. Therefore, statistical significance testing is not separately reported, as practical predictive performance rather than hypothesis testing is the primary focus of this study.

3.2. Medium-Term Prediction

To broadly assess the long-term predictive capacity of the models, medium-term prediction experiments were performed with forecast lengths of 2–10 h. We compared three variants of our model: iTransformer, GTCN-Block, and the full GT-iFormer model. The length of the historical input was set proportionally to the length of the prediction, with ratios of 2:1 (e.g., 4 h of historical data for 1–2 h of prediction). Table 5 shows the RSME results in both Beijing and Shenzhen's datasets.

Table 5. Medium-term prediction.

Task	Historical Time Period	iTransformer		GTCN-Block		GT-iFormer	
		Beijing	Shenzhen	Beijing	Shenzhen	Beijing	Shenzhen
1–2 h prediction	4 h	13.752	4.791	12.641	4.085	11.917	3.654
1–3 h prediction	6 h	16.845	5.012	15.626	5.738	14.542	4.517
1–6 h prediction	12 h	20.417	7.255	20.023	7.359	18.326	6.626
1–8 h prediction	16 h	25.126	9.692	24.245	9.254	22.641	8.347
1–12 h prediction	24 h	31.254	12.854	29.578	11.890	27.214	10.651

The same pattern is observed for all three models: as the forecast horizon increases, the prediction accuracy is gradually degraded. This effect is the result of the nature of medium-term AQ prediction, where the cumulative uncertainties in the evolving meteorological condition and pollutant dispersion procedures cause the errors of prediction to grow as time goes on. For GT-iFormer, the RMSE of Beijing increases from 11.917 (1–2 h) to 27.214 (1–10 h), with an increase of 128%, while it goes up from 3.654 to 10.651 for Shenzhen, with an increase of 191%. The more marked deterioration in Shenzhen can be explained by the fact that the underlying concentrations were much lower, and thus small absolute errors lead to larger relative uncertainties over long horizons.

For the short-term prediction tasks (1–2 h and 1–3 h), GT-iFormer showed obvious advantages over both component versions. In the Beijing dataset, GT-iFormer achieved reductions in the RMSE of 13.35% and 13.68% when compared with iTransformer and 5.73% and 6.94% when compared with GTCN-Block for the predictions of 1–2 h and 1–3 h, respectively. In the Shenzhen dataset, the improvements are even more significant, with reductions in the RMSE of 23.72% and 9.88% over iTransformer, and 10.55% and 21.28%

over GTCN-Block. These findings suggest that the fusion of local spatiotemporal features and global inter-variable interactions benefits the prediction of near-future concentrations.

As we begin to look at more medium-term (1–6 h, 1–8 h, and 1–10 h) prediction horizons, the degradation of the results is evident for all the architectures. Yet, GT-iFormer still consistently outperforms in all of these cases. In the Beijing dataset, the RMSE of GT-iFormer is 10.24%, 9.89%, and 12.93% lower than that of iTransformer and 8.47%, 6.62%, and 7.99% lower than that of GTCN-Block for the predictions of 1–6 h, 1–8 h, and 1–10 h, respectively. For the Shenzhen dataset, the equivalent gains over iTransformer are 8.67%, 13.88%, and 17.14%, and over GTCN-Block they are 9.96%, 9.80%, and 10.42%. Curiously, GTCN-Block is (marginally) better than iTransformer for Beijing for nearly all prediction horizons considered but presents diverging outcomes for Shenzhen, implying that the local temporal patterns and the global features of interaction importance are data-dependent. The increasing performance difference as the forecast horizon lengthens demonstrates the necessity of considering multiple modeling views for capturing long-range dependencies.

A comparison of the two component options shows their complementary advantages. iTransformer is able to capture global inter-variable interactions via an attention mechanism, and GTCN-Block models local feature dependencies as well as multi-scale temporal dynamics by GCN-TCN integration. Neither individual module can compete with the full model of GT-iFormer; this indicates that the hierarchical fusion benefits from synergy. With residual-style fusion, GTCN-Block can also retain feature-level information to the global view from iTransformer to allow the PM_{2.5} dynamics to be more thoroughly expressed. It follows that architectural complexity is the source of slight improvements in one-step prediction but is the source of even more improvements in medium-term prediction. The consistently better accuracy of GT-iFormer over all prediction horizons also establishes its practical usefulness for real-time air quality management, as longer forecast windows facilitate more integrated early warning systems and proactive pollution mitigation schemes.

3.3. Interpretation Analysis of Influencing Factors

3.3.1. Relative Importance

Although the previous one-step and medium-term prediction experiments verify the forecasting ability of GT-iFormer, it is still essential to identify what input features lead to the predictions for scientific interpretation and policy making. In order to meet this demand of interpretability, SHAP analysis was introduced and used to measure the impact of features on the target as well as the fundamental mechanism of PM_{2.5} changes. SHAP analysis was performed on the testing data to provide an unbiased evaluation of interpretability [40]. Figure 6 shows the Beijing and Shenzhen ranked features in a SHAP-based importance sense, calculated by the mean absolute SHAP value of all predictions.

In Beijing, the air pollutant data account for 53.02% of total prediction importance, the influence of chemical composition and atmosphere condition appear balanced in this case, and meteorological factors account for 46.98%. Carbon monoxide (CO) is the most important predictor [41], with an importance of 28.231%, which is much higher than that of the other features. This preeminence mirrored both the status of CO as a direct combustion emission and a surrogate of incomplete combustion, which also produces fine particulates, especially in winter heating sources and vehicular emissions in the north of China. Relative humidity stands second at 20.435%, underlining its importance in enhancing particle hygroscopic growth and the speed of gas-to-particle conversion. When the humidity is high, PM_{2.5} particles absorb moisture, the particle deposits grow in size, and the formation of secondary aerosols is promoted [42]. Ozone (O₃) accounts for 13.228%, acknowledging its role in photochemical reactions leading to secondary organic aerosols in

the warmer months [43]. Atmospheric pressure (10.458%) and temperature (9.314%) also have substantial effects on atmospheric stability and boundary layer behaviors [44]. Sulfur dioxide (SO₂) has a proportion of 7.421%, which is consistent with its role as a precursor in secondary sulfate aerosols. Nitrogen dioxide (NO₂) is somewhat less important at 4.14%, then wind direction (3.339%), weather (1.995%), and wind speed (1.44%).

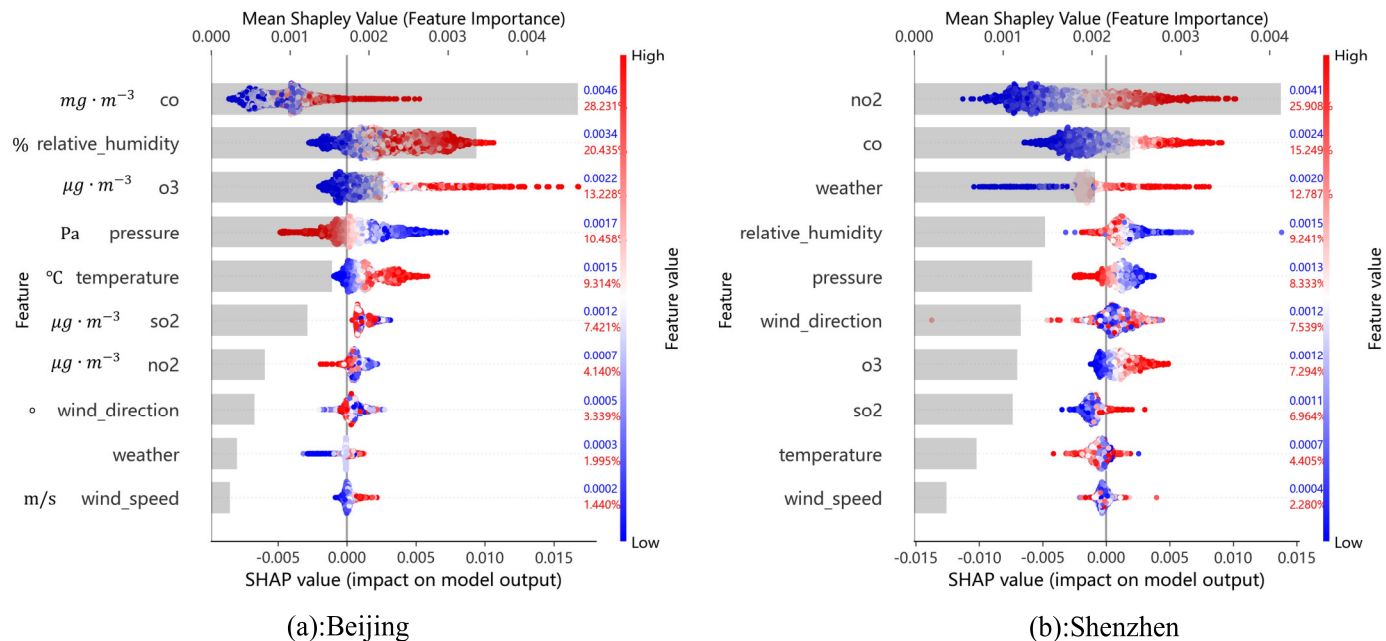


Figure 6. Feature importance analysis based on GT-iFormer. (a) Beijing; (b) Shenzhen.

Shenzhen displays quite a different pattern of feature importance. The importance of air pollutants is 55.42%, and the importance of meteorological factors is 44.58%. In defiance to its minimal contribution in Beijing (4.14%), nitrogen dioxide (NO₂) is the most important predictor, with 25.908% of importance. This contrast is due to the characteristics of Shenzhen's pollution sources: it is a new city with high vehicle density and very limited heavy industry, and so the primary pollution source is traffic-related emissions rather than combustion heating [45]. Carbon monoxide (CO) is second at 15.249%, but significantly lower than Beijing's 28.231%, which verifies that there are fewer smoking sources related to combustion in Shenzhen's subtropical climate, as residents do not have to use heating in winters. Weather (12.787%), with its significantly greater value than that in Beijing (1.995%), implies that Shenzhen's maritime climate with a high frequency of weather changes might have a greater effect on PM_{2.5}. Relative humidity (9.241%) and atmospheric pressure (8.333%) also have a moderate impact, yet wind direction (7.539%) appears to be more dominant, which may be associated with the significance of sea-land breeze circulation in this coastal city. Similar shares are added by ozone (7.294%) and sulfur dioxide (6.964%), while temperature (4.405%) and wind speed (2.28%) come up as the least important.

Some major findings from the cross-city comparison are as follows. First, the dominant pollutant predictors are quite different: CO dominates in Beijing, as related to combustion sources, while NO₂ dominates in Shenzhen, as related to traffic emissions. Second, the sensitivity to meteorology is different between the two climate types: Beijing's PM_{2.5} is more sensitive to humidity and pressure-related atmospheric stability, while Shenzhen's concentrations are more susceptible to weather and wind. Third, although the influential factors between the two cities are different, in both cities, more than 53% of the predictive importance is contributed by air pollutants, which further highlights the dominant position of emission sources.

3.3.2. Nonlinear SHAP Dependence

Relative importance tells us which features are the most informative for prediction, but it does not provide insights into how changing feature values affect $\text{PM}_{2.5}$ concentrations. To overcome this limitation, and to explore nonlinear relations between each single feature and model output within their corresponding domains, SHAP dependence plots are used. Figure 7 shows SHAP dependence plots for all ten input features for Beijing and Shenzhen, demonstrating the changes in SHAP values as feature magnitudes change. These plots show intricate nonlinear relationships, important threshold effects, and surprisingly large cross-city differences in pollutant behavior.

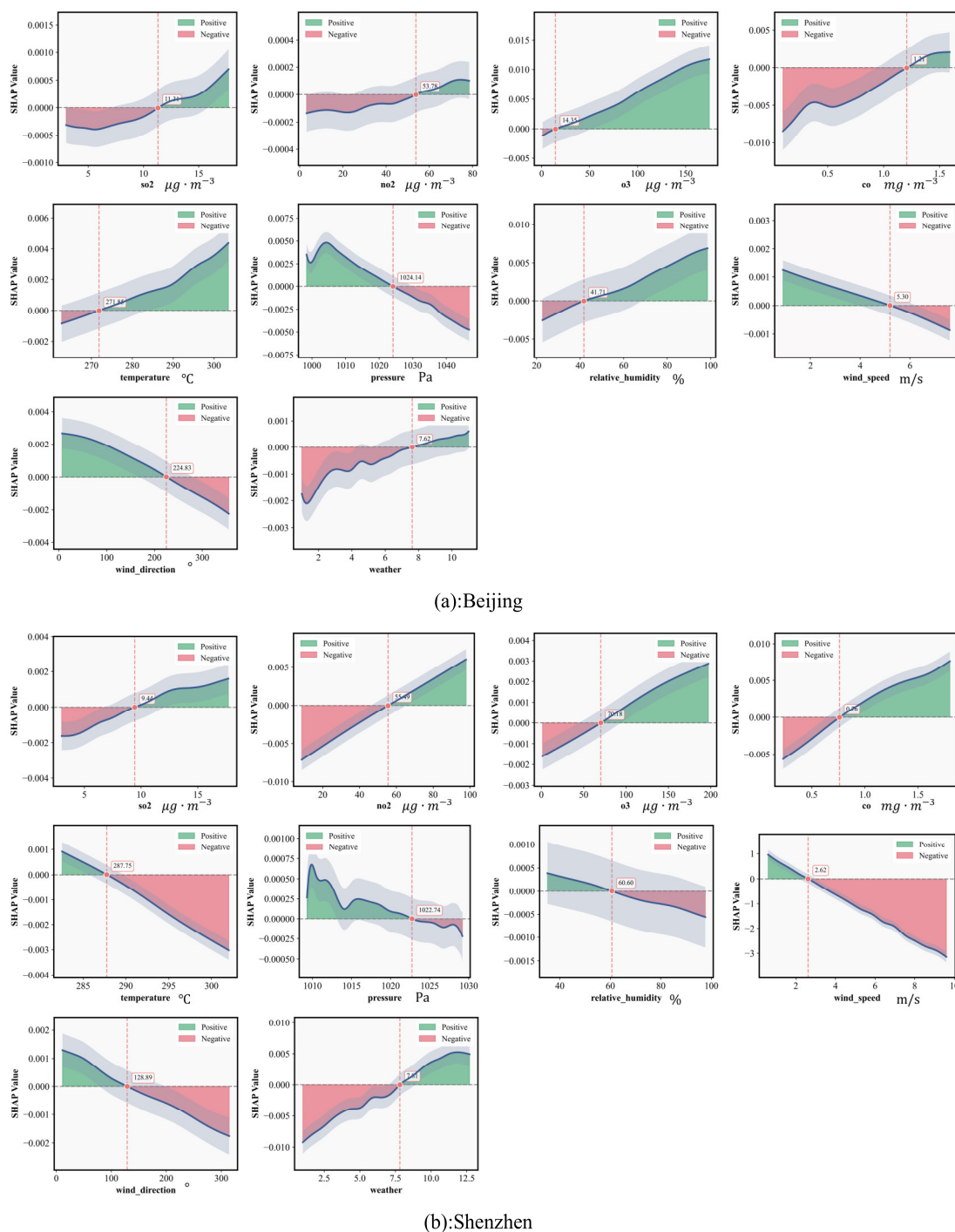


Figure 7. SHAP partial dependence plots for variables. (a) Beijing; (b) Shenzhen.

CO shows different threshold phenomena in both cities and SHAP contributions change from negative to positive at different change points. The threshold for Beijing is 1.214 mg/m^3 and Shenzhen has a smaller threshold of 0.76 mg/m^3 . At low levels of these thresholds, the suppressive effect of CO on predicted $\text{PM}_{2.5}$ probably represents a clean atmosphere. Low ambient carbon monoxide (CO) concentrations generally indicate reduced combustion-related emissions, since CO is primarily emitted from incomplete combustion processes such as fossil fuel or biomass burning [46]. SHAP values monotonically increase with CO concentration when it is above the threshold values, suggesting that it is a key tracer for incomplete combustion processes. The lower threshold of Shenzhen demonstrates that relatively small increases in CO cause substantial increases in $\text{PM}_{2.5}$ in this clean area, whereas the higher threshold of Beijing represents an adjustment for high background combustion emissions. Sulfur dioxide has nearly the same change trend in both cities, with turning points at $11.31 \text{ }\mu\text{g/m}^3$ (BJ) and $9.44 \text{ }\mu\text{g/m}^3$ (SZ). After those thresholds, SO_2 's contribution increases with a larger slope because of concentration-driven efficiency in forming sulfate aerosols through atmospheric oxidation. The higher threshold in Shenzhen indicates increased sensitivity to changes in SO_2 in the subtropical environment [47]. Nitrogen dioxide flips red and becomes green at $53.78 \text{ }\mu\text{g/m}^3$ in Beijing and $55.49 \text{ }\mu\text{g/m}^3$ in Shenzhen. (Firstly, it is remarkable how similar these two thresholds are given the very different pollution in the two cities.) Such uniformity in behavior may be because the process by which NO_2 contributes to secondary nitrate particle formation via photochemical reactions follows similar chemical kinetics globally [48]. Ozone has city-specific threshold effects. The turning point in Beijing is $14.35 \text{ }\mu\text{g/m}^3$, much lower than that in Shenzhen ($70.18 \text{ }\mu\text{g/m}^3$). This dramatic disparity is attributed to the contrasting photochemical regimes; Beijing's continental climate, with severe winter pollution episodes, means that even relatively low O_3 levels were associated with active photochemistry implicated in secondary aerosol formation. In fact, positive contributions to $\text{PM}_{2.5}$ due to the oxidation effect of O_3 on particulate matter formation can only be observed with highly enhanced O_3 concentrations in Shenzhen's subtropical climate with higher background O_3 levels, implying that oxidizing processes induced by O_3 on particulate matter formation might be overwhelmed by other dominant processes under low O_3 concentrations [49].

Temperature contraries between the two cities, form a strong cross-city disparity. In Beijing, the temperature effect shifts from negative to positive at 271.85 K ($-1.30 \text{ }^\circ\text{C}$), as warmer temperatures promote photochemical activity and the formation of secondary aerosols. In Shenzhen, the trend is reversed, with a turn of temperature from a positive to a negative influence at 287.75 K ($14.60 \text{ }^\circ\text{C}$). This contrast is indicative of fundamentally different temperature–pollution dynamics. In cold winter in Beijing's continental climate, low temperatures inhibit boundary layer mixing and chemical reactions, while warming facilitates pollution generation [36]. In the subtropical climate of Shenzhen, warming could potentially promote convective mixing and aid in pollutant dispersal, while cold waves may trap them [50]. Atmospheric pressure reveals the same negative effects in both cities, with the turning points at 1024.14 hPa for Beijing and 1022.74 hPa for Shenzhen. High-pressure systems are associated with atmospheric stability and diminished vertical mixing, which encourage pollutants build-up. The lower bound for Shenzhen is in line with Shenzhen's lower average pressure in the subtropical coastal region. Relative humidity shows the opposite trend. In Beijing, humidity changes from a negative to positive influence at 41.71% , indicating that humidity promotes hygroscopic growth as well as aqueous-phase chemistry that enhances particle mass. In Shenzhen, the trend flips, changing from a positive to negative influence at 60.60% . This distinction may well represent the fact that Shenzhen starts out with higher baseline humidity and so too much moisture may lead to wet deposition and particle loss as opposed to accumulation. Wind speed consistently

disperses in both cities, with positive and negative SHAP value contributions for values below and above the thresholds [51,52]. The threshold of Beijing, 5.30 m/s, is much higher than that of Shenzhen, 2.62 m/s; this is because more intense winds are needed to clean the air in more polluted air environments with complicated terrains, such as in Beijing [53], while the coastal geographic feature of Shenzhen promotes air cleaning with weaker winds via sea–land breeze circulation. Directions of wind reveal different threshold orientations. Beijing becomes negatively influenced at 224.83° , with its south–southwesterly flows [54]. Southerly winds ($90\text{--}225^\circ$) carry pollution from industry areas, while northwesterly winds are associated with much cleaner air. Shenzhen’s threshold at 128.89° corresponds well with its coastal location, where southeasterly winds from the South China Sea carry cleaner maritime air and northerly winds bring potential pollutants from land-based industrial regions [54]. The weather is similarly patterned in both cities, with break-even points equal to 7.62 (Beijing) and 7.81 (Shenzhen). Based on the coding scheme in Table 2, categories 1–7 (rain, snow and clear sky states) produce negative SHAP contributions via wet deposition or increased vertical mixing under a clear state. Categories 8–13 (cloudy to hazy conditions) have positive values, reflecting suppressed vertical mixing and the enhanced near-surface concentration due to increased atmospheric stratification stability trapping pollutants near the surface [55].

Some important findings emerge from the cross-city comparison of threshold behavior. First, Shenzhen has generally lower pollutant thresholds (except for NO_2) than Beijing, which suggests that people are more sensitive to changes in pollutant concentrations under the low-concentration environment. Second, more contrasting patterns are displayed by meteorological factors, and temperature and humidity exhibit reverse relationships between the two cities, which is the result of the completely different climate–pollution interaction mechanisms in the two cities. Thirdly, wind-related thresholds reveal that Beijing needs stronger ventilation for pollutants to be dispersed compared to the coastal city of Shenzhen.

4. Discussion

4.1. Model Performance Analysis

The experimental results reveal a consistent performance hierarchy, with deep learning architectures substantially outperforming conventional machine learning methods across both single-step and medium-term forecasting tasks. This superiority stems from deep learning’s capacity for automatic hierarchical feature extraction and nonlinear dependency modeling from high-dimensional input sequences. Similar performance advantages of deep learning over traditional machine learning models in $\text{PM}_{2.5}$ prediction have been widely reported in recent studies, particularly when long temporal dependencies are involved [56]. These studies emphasize the importance of deep architectures in capturing nonlinear atmospheric dynamics. Conventional approaches such as RF and XGBoost, although robust for tabular data, are not deep enough in architecture to capture temporal complexity and interactions between variables in the process of generating $\text{PM}_{2.5}$ concentration. A note on comparing GT-iFormer to its building block variants across the prediction settings: (1) architectural complexity only brings relatively small gains for one-step prediction, but (2) these gains are aggregated remarkably for longer-term prediction. This finding is consistent with the observation that transformer-based architectures demonstrate more significant advantages in multi-step forecasting scenarios, where long-range dependencies dominate predictive performance [57].

For one-step prediction, GT-iFormer achieves remarkable reductions in RMSE of 8.98% (Beijing) and 5.33% (Shenzhen) over GTCN-Block, suggesting that the performance improvements that can be expected for immediate prediction are incremental rather than

revolutionary. But then, as the horizon moves toward medium-term ranges, the gap in performance systematically increases. For Beijing, the reductions in RMSE compared to GTCN-Block increase from 5.73% (1–2 h) to 7.99% (1–10 h), while the improvements over iTransformer are slightly worse, from 13.35% to 12.93%. This echoes the result that hierarchical feature integration is more important when the temporal dependencies are beyond one step.

The strengths and weaknesses of iTransformer and GTCN-Block are further realized in comparison. iTransformer is superior in modeling global inter-variable interactions by attending along the variable dimension; both iTransformer and GTCN-Block capture temporal dependencies locally through windowed embeddings, but GTCN-Block models local spatial feature dependencies through the GCN and local temporal feature dependencies through the TCN. The mixed trends across cities—that GTCN-Block performs better than iTransformer in Beijing but not consistently in Shenzhen—imply that the importance of local modeling relative to global modeling depends on data characteristics. The performance of neither component alone is comparable to that of GT-iFormer, which further demonstrates the mutually reinforcing effect of their integration.

Cross-city comparison provides fascinating insight into the difficulty of prediction. It is important to interpret the evaluation metrics in light of the different concentration baselines of the two cities. In this study, the average $\text{PM}_{2.5}$ concentration in Beijing is $34.285 \mu\text{g}/\text{m}^3$, whereas in Shenzhen it is $18.521 \mu\text{g}/\text{m}^3$. Since RMSE and MAE are scale-dependent absolute error metrics, higher background concentrations and larger fluctuation amplitudes naturally lead to larger absolute error values. Therefore, the larger RMSE (8.781) and MAE (5.193) observed in Beijing primarily reflect its higher pollution level and wider variability range rather than inferior predictive capability. In contrast, Shenzhen exhibits a lower RMSE (3.871) and MAE (2.552) due to its substantially lower mean concentration and narrower fluctuation range. However, the R^2 measures the proportion of variance explained by the model relative to the total variance. The larger variability in Beijing increases the explainable variance, allowing the model to achieve a higher R^2 (0.978). Conversely, the smaller variance and weaker signal-to-noise ratio in Shenzhen constrain the achievable R^2 (0.957), even though the absolute errors are smaller. This phenomenon is statistically consistent and does not indicate model instability. Instead, it reflects the inherent differences in data distribution characteristics between high-pollution and low-pollution regions.

Although Shenzhen has relatively smaller RMSE values due to a lower baseline of $\text{PM}_{2.5}$ concentration, the R^2 values of Shenzhen are lower than Beijing under all models. Similar phenomena have been reported in low-pollution regions, where reduced variability weakens the signal-to-noise ratio and thus limits predictive skill [58]. This means that prediction of low-concentration $\text{PM}_{2.5}$ is an intrinsically ill-posed problem due to weakened signal-to-noise ratios. The more severe deterioration in performance in Shenzhen as the prediction horizons expand (191% RMSE increase vs. 128% for Beijing) reconfirms that it becomes even more challenging to keep predictive accuracy when baseline variability is small. These observations have practical implications when choosing the appropriate model: for short-term nowcasting purposes, simpler architectures may be adequate, but for medium-term prediction, more complete architectures such as GT-iFormer are necessary, as strong long-range dependency modeling is essential.

4.2. Cross-City Comparison of Influencing Factor Patterns

The importance of SHAP analysis also indicates significant discrepancies across regions regarding influencing factors on $\text{PM}_{2.5}$ concentration, which can be attributed to the different sources of pollution and climate types between Beijing and Shenzhen. One of the salient results is the opposing dominant pollutant predictors. Carbon monoxide is by far the

dominant one in Beijing (28.231%), while in Shenzhen, the dominant one is nitrogen dioxide (25.908%). This combustion-dominated pattern in northern Chinese cities has been widely documented, particularly during heating seasons when coal combustion and industrial emissions remain significant contributors [59,60]. Shenzhen's NO₂ dominance represents a traffic-dominated pollution signature typical of contemporary coastal megacities with no winter heating demand but with high vehicular density. Previous studies in the Pearl River Delta region have similarly identified vehicular emissions as a major contributor to NO₂ and secondary PM formation [61,62]. Differing dominant features notwithstanding, both cities demonstrate air pollutants as a collective overriding predictor, contributing over 53% to predictive importance, and meteorological factors explain all others. Such convergence highlights that emission sources are the main factors shaping PM_{2.5} levels, and meteorological factors are modulatory influences, driving forces. The contrasting dominance of meteorological parameters illustrates two distinct climate-dependent pollution processes. This difference can be further explained by the interaction between emission structure and atmospheric background conditions. In Beijing, stable synoptic conditions and frequent winter heating enhance the accumulation of combustion-related pollutants, making CO a strong proxy for primary emissions, while relative humidity facilitates secondary aerosol formation under stagnant conditions. In contrast, Shenzhen's subtropical coastal climate is characterized by stronger atmospheric mixing and the absence of centralized winter heating, so traffic-related NO₂ becomes a more sensitive indicator of local emission intensity, and the influence of meteorology is more closely linked to rapid weather transitions rather than sustained accumulation processes. Relative humidity is the second highest ranked parameter in Beijing (20.435%), but the fourth in Shenzhen (9.241%), underlining its pertinence in the drier continental climate of Beijing. The enhancing role of humidity in secondary aerosol formation has been extensively discussed, especially under high-moisture stagnant conditions that promote heterogeneous reactions [63]. Weather conditions have the reverse distribution: they have a very marginal impact in Beijing (1.995%), but they rank as the third most important factor in Shenzhen (12.787%), which is not surprising, since Shenzhen is located in the maritime climate and abrupt changes in weather are major factors driving pollution.

These regional differences have immediate consequences for pollution control and forecasting strategies. Reducing combustion-related emissions (CO and SO₂) in heating and industrial activities should be a priority in Beijing and humidity-based early warning systems should be used in high-moisture situations since secondary aerosol formation is enhanced during such periods. Shenzhen needs to concentrate on traffic emissions reduction (NO₂) and combine meteorological forecasting with pollution prediction models, such as strong weather–PM_{2.5} relationships. The fact that sources of emissions have a greater impact than meteorological conditions in both cities highlights the critical importance of implementing source control policies in both cities rather than focusing on their meteorological sensitivity.

4.3. Threshold Effects and Their Implications for Air Quality Management

The SHAP dependence plot captures the most basic property of PM_{2.5} formation: for nearly all the input dimensions, the PM_{2.5} response exhibits strong threshold behaviors instead of simple linearity. Nonlinear and threshold-type relationships between gaseous precursors and PM_{2.5} have been widely reported in the context of secondary aerosol formation during haze episodes [62]. This prevalence of inflection points indicates that variations in PM_{2.5} are controlled by regime-dependent dynamics, such that minor modifications in predictor values at regime boundaries can induce outsized responses. The detection of concentration-dependent transitions in gaseous pollutions sheds mechanistic light on

the secondary route of aerosol formation. CO transitions at 1.214 mg/m^3 in Beijing, but in Shenzhen at a significantly lower value of 0.76 mg/m^3 , indicating increased sensitivity for the low-concentration situation. Beneath these levels, the effects of gaseous pollutants on the formation of $\text{PM}_{2.5}$ are non-significant, implying that the gas-to-particle transformation needs to meet a certain threshold of the precursor concentration. Such concentration-dependent transitions are consistent with regime-shift behaviors reported in emission–concentration response analyses [64]. SHAP contributions grow super-linearly beyond critical values. This means that emission cuts that result in concentrations lower than the specified thresholds may give disproportionate benefits for air quality. The meteorological thresholds show significant cross-city diversity. Relative humidity transitions at 41.71% in Beijing, while its direction reverses at 60.60% in Shenzhen, implying different moisture–pollution interactions between continental and oceanic climates. Wind speed thresholds are quite different (5.30 m/s for Beijing vs. 2.62 m/s in Shenzhen), suggesting that coastal regions have good dispersion even at lower wind speeds. The dispersion effect of wind speed and its regional heterogeneity have also been confirmed in multiple observational studies across northern and southern China [65]. Opposite trends of temperature were observed for both cities, suggesting that region-specific response models are necessary.

Such threshold behaviors provide useful information for air quality management. In addition, control of emissions should focus on reducing the concentration of precursors below the identified thresholds rather than on applying equal percent reductions. For the city of Beijing, if the concentrations of CO and SO_2 are kept at 1.214 mg/m^3 and $11.31 \text{ }\mu\text{g/m}^3$, respectively, the formation pathways for secondary aerosols could be significantly interrupted. The lower limit (CO: 0.76 mg/m^3 , SO_2 : $9.44 \text{ }\mu\text{g/m}^3$) for Shenzhen requires stricter emission standards to keep the current air quality status. Second, early warning systems should be based on threshold-related triggers that are adapted to the specific local conditions, such as high humidity ($>42\%$), low wind speed ($<5.3 \text{ m/s}$), and southwesterly wind (Beijing: pollution alert should be issued). Another early warning system could be when humidity is lower than 60% and when the wind speed is down to 2.6 m/s with northerly wind patterns in Shenzhen. Thirdly, the thresholds for favorable (categories 1–7) and unfavorable (categories 8–13) weather conditions are the same for both cities, allowing for the use of routine weather forecasts in the pollution prediction system. Environmental authorities may use imminent overcast conditions as signals to enforce temporary emission restrictions or public advisories on health. These regionally defined, threshold-based management strategies allow for more effective use of resources as compared to blanket control measures that do not account for meteorology. Similar threshold-informed air quality management frameworks have been proposed in recent policy-oriented atmospheric research [66].

5. Conclusions

This work presents GT-iFormer, an end-to-end interpretable deep learning framework that combines a GCN, TCN, and an inverse Transformer (iTransformer) for the prediction of $\text{PM}_{2.5}$ concentration. The framework employs a GTCN-Block, which nests the GCN and TCN by means of a residual-like fusion, along with iTransformer for modeling global inter-variable interactions. Analysis with SHAP is embedded to achieve transparent model interpretability. Comprehensive experiments on the hourly data of Beijing (2020–2021) and Shenzhen (2021) lead to the following observations:

- (1) GT-iFormer consistently outperforms all competing models in predictive accuracy with different forecast horizons in both cities. In single-step prediction, the model attains an RMSE of $8.781 \text{ }\mu\text{g/m}^3$ and R^2 of 0.978 for Beijing and an RMSE of $3.871 \text{ }\mu\text{g/m}^3$

and R^2 of 0.957 for Shenzhen, surpassing all baselines. Against the strongest baseline, ResNet-TCN, GT-iFormer brings RMSE reductions of 15.75% (Beijing) and 17.92% (Shenzhen). The performance advantages become larger for medium-term prediction, i.e., the RMSE improvements become larger for the longer prediction horizons, which demonstrates that hierarchical feature integration is effective in capturing long-range temporal dependencies.

- (2) The ablations of GT-iFormer with its atomic component variants (iTransformer and GTCN-Block) differentiate that each building block construct synergistically in terms of forecasting accuracy. GTCN-Block captures local feature dependencies and multi-scale temporal patterns via GCN-TCN fusion, and iTransformer captures global inter-variable relations. Neither by itself is competitive with the full model, illustrating the complementary advantage of integrating them in a hierarchical fashion.
- (3) Relative importance analysis using SHAP reveals different regional patterns for the driving factors of PM_{2.5}. Predictive capacity is dominated by air pollutants in both cities (53.02% in Beijing and 55.42% in Shenzhen), while the rest is explained by meteorological factors. Yet, there is a great contrast between dominant predictors: CO is the first in Beijing (28.231%), indicating combustion-related sources, while NO₂ is the first in Shenzhen (25.908%), suggesting the predominance of traffic emissions. These results indicate that region-specific emission control measures targeting different sources of pollution are justified.
- (4) All input features have prominent nonlinear and threshold effects on PM_{2.5} concentration with significant variations across cities. Different critical inflection points occur in different cities: for CO, it is 1.214 mg/m³ (Beijing) vs. 0.76 mg/m³ (Shenzhen), for relative humidity it is 41.71% (Beijing) vs. 60.60% (Shenzhen), and for wind speed it is 5.30 m/s (Beijing) vs. 2.62 m/s (Shenzhen).

Although these are valid contributions, there are some limitations that need to be addressed. First, the transferability of the model to other cities with similar or dissimilar pollution profiles needs to be further tested. In this study, experiments were conducted only in Beijing and Shenzhen, and external validation using datasets from other climatic zones or regions with different topographical and emission characteristics was not performed. Therefore, the general applicability of the proposed framework should be interpreted with caution. Secondly, the present framework is based solely on past observations, without taking into account physical constraints or chemistry mechanisms. In addition, the SHAP-based feature importance analysis may vary under different historical window settings, and thus the interpretability of the results is conditional on the adopted training configuration. Moreover, although a fixed training–testing split was adopted to evaluate model performance, a systematic cross-validation strategy across multiple meteorological cycles was not implemented. The stability and robustness of the model under different seasonal or interannual conditions therefore require further investigation in future research. Future work should further validate in more cities with various climatic zones, incorporate physics-informed constraints for improving mechanistic interpretability, and investigate online operational implementation for real-time air quality management applications.

Author Contributions: Conceptualization, D.L.; Methodology, D.L.; Software, D.L.; Validation, D.L.; Formal analysis, D.L.; Investigation, J.W.; Resources, M.L.; Data curation, H.H.; Writing—original draft, D.L.; Writing—review & editing, D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (42374024 and 42274029), the Beijing Nova Program (20230484270), and the Doctor Graduate Scientific Research Ability Improvement Project of Beijing University of Civil Engineering and Architecture (DG2025034).

Institutional Review Board Statement: No applicable.

Informed Consent Statement: No applicable.

Data Availability Statement: All data generated or analyzed during this study are included in this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sangkham, S.; Phairuang, W.; Sherchan, S.P.; Pansakun, N.; Munkong, N.; Sarndhong, K.; Islam, M.A.; Sakunkoo, P. An update on adverse health effects from exposure to PM_{2.5}. *Environ. Adv.* **2024**, *18*, 100603. [\[CrossRef\]](#)
2. Wang, X.; Tian, G.; Yang, D.; Zhang, W.; Lu, D.; Liu, Z. Responses of PM_{2.5} pollution to urbanization in China. *Energy Policy* **2018**, *123*, 602–610. [\[CrossRef\]](#)
3. Yan, Y.; Ren, P.; Meng, Q. Quantitative evaluation of the synergistic effects of multiple meteorological parameters on air pollutants based on generalized additive models. *Urban Clim.* **2024**, *55*, 101965. [\[CrossRef\]](#)
4. Ju, T.; Wang, L.; Li, B.; Lv, Z.; Gu, Z. Study on correlation between atmospheric pollutants and meteorological factors during extreme weather in Mexico based on GTWR model. *Air Qual. Atmos. Health* **2025**, *18*, 1489–1505. [\[CrossRef\]](#)
5. Zhou, Y.; Han, Y.; Hu, Z.; Zhou, Q.; Liu, Y.; Dong, L.; Xiao, P. Characteristics of Planetary Boundary Layer Height (PBLH) over Shenzhen, China: Retrieval Methods and Air Pollution Conditions. *Remote Sens.* **2025**, *17*, 3937. [\[CrossRef\]](#)
6. Lu, S.; Li, X.; Zhao, T.; Ma, Y.; Wang, Y.; Zhang, Y.; Luo, Y.; Xin, Y. Impact of thermal structure of planetary boundary layer on aerosol pollution over urban regions in Northeast China. *Atmos. Pollut. Res.* **2023**, *14*, 101665. [\[CrossRef\]](#)
7. Su, T.; Li, Z.; Kahn, R. Relationships between the planetary boundary layer height and surface pollutants derived from lidar observations over China: Regional pattern and influencing factors. *Atmos. Chem. Phys.* **2018**, *18*, 15921–15935. [\[CrossRef\]](#)
8. Gao, Z.; Do, K.; Li, Z.; Jiang, X.; Maji, K.J.; Ivey, C.E.; Russell, A.G. Predicting PM_{2.5} levels and exceedance days using machine learning methods. *Atmos. Environ.* **2024**, *323*, 120396. [\[CrossRef\]](#)
9. Zhou, S.; Wang, W.; Zhu, L.; Qiao, Q.; Kang, Y. Deep-learning architecture for PM_{2.5} concentration prediction: A review. *Environ. Sci. Ecotechnol.* **2024**, *21*, 100400. [\[CrossRef\]](#)
10. Li, D.; Wang, J.; Tian, D.; Chen, C.; Xiao, X.; Wang, L.; Wen, Z.; Yang, M.; Zou, G. Residual neural network with spatiotemporal attention integrated with temporal self-attention based on long short-term memory network for air pollutant concentration prediction. *Atmos. Environ.* **2024**, *329*, 120531. [\[CrossRef\]](#)
11. Parmezan, A.R.S.; Souza, V.M.A.; Batista, G.E.A.P.A. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. *Inf. Sci.* **2019**, *484*, 302–337. [\[CrossRef\]](#)
12. Khan, D.R.; Patankar, A.B.; Khan, A. An experimental comparison of classic statistical techniques on univariate time series forecasting. *Procedia Comput. Sci.* **2024**, *235*, 2730–2740. [\[CrossRef\]](#)
13. Zheng, H.; Shen, L.; Tang, A.; Luo, Y.; Hu, H.; Du, B.; Wen, Y.; Tao, D. Learning from models beyond fine-tuning. *Nat. Mach. Intell.* **2025**, *7*, 6–17. [\[CrossRef\]](#)
14. Herm, L.-V.; Heinrich, K.; Wanner, J.; Janiesch, C. Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *Int. J. Inf. Manag.* **2023**, *69*, 102538. [\[CrossRef\]](#)
15. Wang, Y.-Z.; He, H.-D.; Huang, H.-C.; Yang, J.-M.; Peng, Z.-R. High-resolution spatiotemporal prediction of PM_{2.5} concentration based on mobile monitoring and deep learning. *Environ. Pollut.* **2025**, *364*, 125342. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Han, S.; Xun, Y.; Cai, J.; Yang, H.; Li, Y. DyGraphformer: Transformer combining dynamic spatio-temporal graph network for multivariate time series forecasting. *Neural Netw.* **2025**, *181*, 106776. [\[CrossRef\]](#)
17. Lu, H.; Wang, L.; Ma, X.; Cheng, J.; Zhou, M. A survey of graph neural networks and their industrial applications. *Neurocomputing* **2025**, *614*, 128761. [\[CrossRef\]](#)
18. Li, D.; Han, H.; Yu, H.; Wang, J.; Liu, M.; Zou, G.; Wang, L. Spatiotemporal Graph Convolutional Attention Network for Air Quality Index Prediction of Beijing, Shanghai and Shenzhen. *Atmosphere* **2025**, *16*, 1314. [\[CrossRef\]](#)
19. Hu, W.; Zhang, Z.; Zhang, S.; Chen, C.; Yuan, J.; Yao, J.; Zhao, S.; Guo, L. Learning spatiotemporal dependencies using adaptive hierarchical graph convolutional neural network for air quality prediction. *J. Clean. Prod.* **2024**, *459*, 142541. [\[CrossRef\]](#)
20. Zou, G.; Lai, Z.; Wang, T.; Liu, Z.; Li, Y. MT-STNet: A Novel Multi-Task Spatiotemporal Network for Highway Traffic Flow Prediction. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 8221–8236. [\[CrossRef\]](#)
21. Ye, Y.; Cao, Y.; Dong, Y.; Yan, H. A graph neural network and Transformer-based model for PM_{2.5} prediction through spatiotemporal correlation. *Environ. Model. Softw.* **2025**, *191*, 106501. [\[CrossRef\]](#)
22. Kim, D.-Y.; Jin, D.-Y.; Suk, H.-I. Spatiotemporal graph neural networks for predicting mid-to-long-term PM_{2.5} concentrations. *J. Clean. Prod.* **2023**, *425*, 138880. [\[CrossRef\]](#)

23. Shen, X.; Shen, J.; Zhang, Y.; Wu, H.; Wang, Y. A multi-source feature engineering-enhanced framework for mid-to-long-term EV charging load forecasting: Integrating self-adaptive optimization and BiLSTM-iTransformer predictor. *Energy* **2025**, *339*, 139020. [\[CrossRef\]](#)
24. Jia, W.; Sun, J.; Guan, S. FPQ-iTransformer: A novel iTransformer-based model for accurate seawater salinity prediction. *Environ. Model. Softw.* **2025**, *192*, 106529. [\[CrossRef\]](#)
25. Zhang, R.; Awang, N. A novel ST-iTransformer model for spatio-temporal ambient air pollution forecasting. *J. Big Data* **2025**, *12*, 101. [\[CrossRef\]](#)
26. Xin, N.; Su, J.; Hasan, M.M. MMformer with adaptive attention: Advancing multivariate time series forecasting for environmental applications. *Appl. Soft Comput.* **2026**, *186*, 114090. [\[CrossRef\]](#)
27. Li, Y.; Zhu, Y.; Wu, J. Critical paths and sectors for upstream and downstream carbon emissions of digital industry: A case study of Beijing, China. *J. Clean. Prod.* **2025**, *528*, 146745. [\[CrossRef\]](#)
28. Ding, J.; Wang, K. Spatial pattern and morphological characteristics of industrial production space and influential factors in the Pearl River Delta urban agglomeration. *Prog. Geogr.* **2016**, *35*, 610–621. [\[CrossRef\]](#)
29. Li, D.; Wang, L.; Wang, J.; Chen, C.; Xiao, X.; Zou, G. Spatiotemporal networks for multi-city and multi-task air pollutant prediction—Beijing, Shanghai and Shenzhen as examples. *Urban Clim.* **2025**, *63*, 102584. [\[CrossRef\]](#)
30. Zhang, K.; Yang, X.; Cao, H.; Thé, J.; Tan, Z.; Yu, H. Multi-step forecast of PM_{2.5} and PM₁₀ concentrations using convolutional neural network integrated with spatial-temporal attention and residual learning. *Environ. Int.* **2023**, *171*, 107691. [\[CrossRef\]](#)
31. Zhang, B.; Zou, G.; Qin, D.; Ni, Q.; Mao, H.; Li, M. RCL-Learning: ResNet and convolutional long short-term memory-based spatiotemporal air pollutant concentration prediction model. *Expert Syst. Appl.* **2022**, *207*, 118017. [\[CrossRef\]](#)
32. Zhang, J.; Li, M.; Xu, Y.; He, H.; Li, Q.; Wang, T. StrucGCN: Structural enhanced graph convolutional networks for graph embedding. *Inf. Fusion* **2025**, *117*, 102893. [\[CrossRef\]](#)
33. Wang, Y.; Deng, L.; Zheng, L.; Gao, R.X. Temporal convolutional network with soft thresholding and attention mechanism for machinery prognostics. *J. Manuf. Syst.* **2021**, *60*, 512–526. [\[CrossRef\]](#)
34. Li, T.; Si, H.; Qiu, J.; Li, J.; Gong, Y. A hybrid algorithm of TCN-iTransformer for aircraft aerodynamic parameter estimation based on dual attention mechanism. *Aerosp. Sci. Technol.* **2025**, *164*, 110350. [\[CrossRef\]](#)
35. Lei, K.; Wang, M.; Wang, M.; Liu, Q.; Zhang, F.; Xing, M.; Wu, W.; Jiang, F.; Guo, X.; Han, Q.; et al. SHAP explainable PSO-CNN-BiLSTM for 6-hour prediction analysis of urban PM_{2.5} and O₃ concentrations. *Atmos. Pollut. Res.* **2025**, *16*, 102705. [\[CrossRef\]](#)
36. Wei, Q.; Chen, Y.; Zhang, H.; Jia, Z.; Yang, J.; Niu, B. Simulation and prediction of PM_{2.5} concentrations and analysis of driving factors using interpretable tree-based models in Shanghai, China. *Environ. Res.* **2025**, *270*, 121003. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Zeng, Q.; Cao, Y.; Fan, M.; Chen, L.; Zhu, H.; Wang, L.; Li, Y.; Liu, S. Fine particulate matter concentration prediction based on hybrid convolutional network with aggregated local and global spatiotemporal information: A case study in Beijing and Chongqing. *Atmos. Environ.* **2024**, *333*, 120647. [\[CrossRef\]](#)
38. Tariq, S.; Tariq, S.; Kim, S.; Woo, S.S.; Yoo, C. Distance adaptive graph convolutional gated network-based smart air quality monitoring and health risk prediction in sensor-devoid urban areas. *Sustain. Cities Soc.* **2023**, *91*, 104445. [\[CrossRef\]](#)
39. Zou, G.; Zhang, B.; Yong, R.; Qin, D.; Zhao, Q. FDN-learning: Urban PM_{2.5}-concentration Spatial Correlation Prediction Model Based on Fusion Deep Neural Network. *Big Data Res.* **2021**, *26*, 100269. [\[CrossRef\]](#)
40. Liu, S.; Wang, G.; Kong, F.; Zhao, N.; Gao, W.; Zhang, H. PM_{2.5} pollution characteristics, drivers, and regional transport during different pollution levels in Linyi, China: An integrated PMF-ML-SHAP framework and transport models. *J. Hazard. Mater.* **2025**, *494*, 138534. [\[CrossRef\]](#)
41. Chen, Z.; Chen, D.; Zhao, C.; Kwan, M.-p.; Cai, J.; Zhuang, Y.; Zhao, B.; Wang, X.; Chen, B.; Yang, J.; et al. Influence of meteorological conditions on PM_{2.5} concentrations across China: A review of methodology and mechanism. *Environ. Int.* **2020**, *139*, 105558. [\[CrossRef\]](#)
42. Cheng, Y.; He, K.-b.; Du, Z.-y.; Zheng, M.; Duan, F.-k.; Ma, Y.-l. Humidity plays an important role in the PM_{2.5} pollution in Beijing. *Environ. Pollut.* **2015**, *197*, 68–75. [\[CrossRef\]](#)
43. Wang, F.; An, X.; Sheng, L. Extreme ozone pollution over North China linked to Rossby waves induced by North Atlantic sea surface temperature. *Atmos. Res.* **2026**, *331*, 108668. [\[CrossRef\]](#)
44. Wei, Q.; Zhang, H.; Yang, J.; Niu, B.; Xu, Z. PM_{2.5} concentration prediction using a whale optimization algorithm based hybrid deep learning model in Beijing, China. *Environ. Pollut.* **2025**, *371*, 125953. [\[CrossRef\]](#)
45. Geng, G.; Liu, Y.; Liu, Y.; Liu, S.; Cheng, J.; Yan, L.; Wu, N.; Hu, H.; Tong, D.; Zheng, B.; et al. Efficacy of China's clean air actions to tackle PM_{2.5} pollution between 2013 and 2020. *Nat. Geosci.* **2024**, *17*, 987–994. [\[CrossRef\]](#)
46. Li, Y.; Ma, Z.; Han, T.; Quan, W.; Wang, J.; Zhou, H.; He, D.; Dong, F. Long-term declining in carbon monoxide (CO) at a rural site of Beijing during 2006–2018 implies the improved combustion efficiency and effective emission control. *J. Environ. Sci.* **2022**, *115*, 432–442. [\[CrossRef\]](#) [\[PubMed\]](#)

47. Cao, J.; Qiu, X.; Gao, J.; Wang, F.; Wang, J.; Wu, J.; Peng, L. Significant decrease in SO₂ emission and enhanced atmospheric oxidation trigger changes in sulfate formation pathways in China during 2008–2016. *J. Clean. Prod.* **2021**, *326*, 129396. [\[CrossRef\]](#)
48. Goldstein, S.; Samuni, A. Chemistry of nitrogen dioxide and its biological implications. *Redox Biochem. Chem.* **2024**, *7*, 100020. [\[CrossRef\]](#)
49. Qu, Y.; Wang, T.; Yuan, C.; Wu, H.; Gao, L.; Huang, C.; Li, Y.; Li, M.; Xie, M. The underlying mechanisms of PM_{2.5} and O₃ synergistic pollution in East China: Photochemical and heterogeneous interactions. *Sci. Total Environ.* **2023**, *873*, 162434. [\[CrossRef\]](#)
50. Sardoei, A.S.; Fazeli-Nasab, B. Physiological and Molecular Mechanisms of Freezing in Plants. *Phyton-Int. J. Exp. Bot.* **2025**, *94*, 1601–1630. [\[CrossRef\]](#)
51. Evans, C.; Zhu, L.; Baker, K.; Meng, L. Unveiling multiscale drivers of wind speed in Michigan using machine learning. *npj Clim. Atmos. Sci.* **2025**, *8*, 286. [\[CrossRef\]](#)
52. Hong, T.; Yim, S.H.L.; Heo, Y. Interpreting complex relationships between urban and meteorological factors and street-level urban heat islands: Application of random forest and SHAP method. *Sustain. Cities Soc.* **2025**, *126*, 106353. [\[CrossRef\]](#)
53. Alomari, Y.; Andó, M. SHAP-based insights for aerospace PHM: Temporal feature importance, dependencies, robustness, and interaction analysis. *Results Eng.* **2024**, *21*, 101834. [\[CrossRef\]](#)
54. Liu, Y.; Xu, Y.; Zhang, F.; Shu, W. A preliminary study on the influence of Beijing urban spatial morphology on near-surface wind speed. *Urban Clim.* **2020**, *34*, 100703. [\[CrossRef\]](#)
55. Xie, Y.; Sengupta, M.; Yang, J.; Habte, A.; Buster, G.; Benton, B.; Foster, M.; Heidinger, A.; Liu, Y. The influence of cloud cover on the reliability of satellite-based solar resource data. *Renew. Sustain. Energy Rev.* **2025**, *208*, 115070. [\[CrossRef\]](#)
56. Bai, S.; Kolter, J.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**. [\[CrossRef\]](#)
57. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115. [\[CrossRef\]](#)
58. Di, Q.; Wang, Y.; Zanobetti, A.; Wang, Y.; Koutrakis, P.; Choirat, C.; Dominici, F.; Schwartz Joel, D. Air Pollution and Mortality in the Medicare Population. *New Engl. J. Med.* **2017**, *376*, 2513–2522. [\[CrossRef\]](#)
59. Shang, J.; Khuzestani, R.B.; Tian, J.; Schauer, J.J.; Hua, J.; Zhang, Y.; Cai, T.; Fang, D.; An, J.; Zhang, Y. Chemical characterization and source apportionment of PM_{2.5} personal exposure of two cohorts living in urban and suburban Beijing. *Environ. Pollut.* **2019**, *246*, 225–236. [\[CrossRef\]](#)
60. Wang, Y.; Yao, L.; Wang, L.; Liu, Z.; Ji, D.; Tang, G.; Zhang, J.; Sun, Y.; Hu, B.; Xin, J. Mechanism for the formation of the January 2013 heavy haze pollution episode over central and eastern China. *Sci. China Earth Sci.* **2014**, *57*, 14–25. [\[CrossRef\]](#)
61. Chen, Y.; Zhu, L.; Wang, S.; Henze, D.K.; Fu, T.-M.; Zhang, L.; Wang, X. Unraveling the complexities of ozone and PM_{2.5} pollution in the Pearl River Delta region: Impacts of precursors emissions and meteorological factors, and effective mitigation strategies. *Atmos. Pollut. Res.* **2025**, *16*, 102368. [\[CrossRef\]](#)
62. Huang, R.-J.; Zhang, Y.; Bozzetti, C.; Ho, K.-F.; Cao, J.-J.; Han, Y.; Daellenbach, K.R.; Slowik, J.G.; Platt, S.M.; Canonaco, F.; et al. High secondary aerosol contribution to particulate pollution during haze events in China. *Nature* **2014**, *514*, 218–222. [\[CrossRef\]](#)
63. Cheng, Y.; Zheng, G.; Wei, C.; Mu, Q.; Zheng, B.; Wang, Z.; Gao, M.; Zhang, Q.; He, K.; Carmichael, G.; et al. Reactive nitrogen chemistry in aerosol water as a source of sulfate during haze events in China. *Sci. Adv.* **2016**, *2*, e1601530. [\[CrossRef\]](#)
64. Li, H.; Zheng, B.; Lei, Y.; Hauglustaine, D.; Chen, C.; Lin, X.; Zhang, Y.; Zhang, Q.; He, K. Trends and drivers of anthropogenic NO_x emissions in China since 2020. *Environ. Sci. Ecotechnol.* **2024**, *21*, 100425. [\[CrossRef\]](#)
65. Tai, A.P.K.; Mickley, L.J.; Jacob, D.J. Correlations between fine particulate matter (PM_{2.5}) and meteorological variables in the United States: Implications for the sensitivity of PM_{2.5} to climate change. *Atmos. Environ.* **2010**, *44*, 3976–3984. [\[CrossRef\]](#)
66. Zhao, N.; Zhang, H.; Wang, G. Revealing the nonlinear responses of PM_{2.5} and O₃ to VOC and NO_x emissions from various sources in Shandong, China. *J. Hazard. Mater.* **2025**, *489*, 137655. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.