Interpretable Machine Learning Framework for Ozone Concentration Prediction Using TAP Data and SHAP Analysis in China

Baoli Jia*, Fayuan Zheng
Gansu Iron and Steel Vocational and Technical College, Jiayuguan 735100, China
Email: jbliww@126.com
*Corresponding author

Keywords: Online data, SHAP, O3, air pollution prevention and control, machine algorithm

Received: July 31, 2025

This study aims to enhance the monitoring and prediction capabilities of ozone pollution and support precise environmental governance and improvement of atmospheric quality. It constructs and compares multiple ozone concentration prediction models, including Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Random Forest (RF), based on machine learning methods. The study uses national online observation data from Tracking Air Pollution in China (TAP) spanning 2015–2025, covering 34 provinces. Before modeling, the data are subjected to missing value imputation, outlier removal, and normalization. Combined with Shapley Additive Explanations (SHAP) values, the study conducts model interpretability analysis to deeply reveal the key driving factors affecting ozone formation and their regional differences. The results show that the national annual average ozone concentration increases from 105.6 μg/m³ in 2015 to 130.2 μg/m³ in 2019, with an increase of 23.3%. The peak concentration in the Beijing-Tianjin-Hebei region in 2019 reached 182.8 μg/m³, exceeding the national limit by 114%. Ozone concentration decreased in 2020 due to the impact of the epidemic, but it was still 24.6 μg/m³ higher in 2024 than in 2015. In terms of model prediction performance, XGBoost performs the best nationwide and in all major regions. At the national level, its Mean Absolute Error (MAE) is 11.3 μg/m³, Root Mean Squared Error (RMSE) is 15.6 μg/m³, R^2 is 0.882, and Nash-Sutcliffe Efficiency coefficient (NSE) is 0.88. SHAP analysis indicates that day of year, temperature, and sunshine duration are the main driving factors for changes in national ozone concentration. NO_2 contributes significantly in the Beijing-Tianjin-Hebei region and the Fenwei Plain, while the effects of temperature and sunshine are the strongest in the Pearl River Delta region. This study enriches the understanding of the spatiotemporal dynamics and formation mechanisms of ozone pollution, and provides solid data support and theoretical basis for the scientific formulation of regional pollution prevention and control strategies.

Povzetek: Študija za napovedovanje ozona nad Kitajsko gradi več modelov na podatkih TAP 2015–2025 ter z razlago SHAP razkrije ključne značilke (letni dan, temperatura, osončenost, regionalni NO_2) za ciljno, regijsko prilagojeno upravljanje kakovosti zraka.

1 Introduction

With the rapid economic development and accelerated urbanization in China, air environmental quality issues have become increasingly prominent. Unlike traditional primary pollutants, ozone is a typical secondary pollutant formed by nitrogen oxides (NO_x) and volatile organic compounds (VOCs) under sunlight. It has strong oxidizing properties, causing severe harm to human health (respiratory and cardiovascular diseases) and ecosystems (crop yield reduction, forest chlorosis) [1]. In recent years, high-temperature and low-rainfall summers in many Chinese cities have driven ozone concentrations to record highs, exacerbating public health risks and imposing higher requirements on regional collaborative governance. Traditional ozone monitoring and prediction mainly rely on physical-chemical transport models (CTMs) and statistical regression models. Although CTMs have advantages in mechanistic research, they are sensitive to initial conditions, computationally intensive, and difficult to achieve real-time early warning [2]. However, statistical regression models are limited in accuracy and generalization ability, struggling to capture the comprehensive impacts of nonlinear, multi-factor coupling on ozone formation [3]. Furthermore, most current studies focus on macro-trend analysis, lacking interpretable insights into single high-concentration events and regional difference mechanisms, thus failing to provide quantitative bases for differentiated pollution control strategies [4]. In recent years, machine learning

methods have been widely applied to air pollution prediction due to their outstanding capabilities in nonlinear modeling, adaptive learning, and high-dimensional feature processing. Ensemble algorithms such as Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Random Forest (RF) have attracted attention for their excellent prediction accuracy and computational efficiency. However, the inherent interpretability limitation of black-box models hinders decision-makers from deeply trusting and applying model outputs [5]. To address this, the Shapley Additive Explanations (SHAP) method emerges, introducing Shapley values from game theory into machine learning interpretation. By using fair allocation theory, this method provides a unified and axiomatically satisfied measure of feature contributions for both single predictions and global model behavior, enabling model-agnostic and theoretically optimal interpretability analysis [6].

This study adopts the near-real-time updated Tracking Air Pollution in China (TAP) dataset and combines machine learning with the SHAP method. It can use large-scale, multi-source online data to improve the timeliness and accuracy of prediction. It deeply reveals the key driving factors affecting ozone formation and their regional differences from both local and global perspectives. The main innovations of this study are as follows: First, it integrates high spatiotemporal resolution TAP online observation data with multi-scale machine learning models. Second, it uses Bayesian optimization to improve the prediction performance of algorithms such as XGBoost. Third, it innovatively introduces SHAP interpretability analysis. This approach quantifies the local and global contributions of key driving factors such as day of year, temperature, sunshine duration, and NO_2, and reveals the differentiated characteristics across different regions and months. This study holds important theoretical value and application significance for realizing the integrated ozone governance closed loop of "prediction and early warning-precision prevention and control-evaluation and feedback". The study provides operable decision support for precise environmental governance.

2 Related works

In recent years, scholars in China and other countries have made remarkable progress in the research on ozone pollution prevention and control models. As early as 2007, Lasry et al. took Mediterranean cities in France as examples, used CTMs to simulate different emission scenarios, and evaluated the impact of the standard European short-term emergency action plan on ozone peak concentrations [7]. Li et al. argued that the greatest advantage of CTMs was in their ability to analyze the physical-chemical coupling relationship between precursor emissions and meteorological conditions [8]. With the improvement of big data and computing capabilities, statistical regression methods have been gradually replaced by machine learning models, and ensemble learning algorithms such as XGBoost and LightGBM have become the mainstream in recent years. Liu et al. found that after integrating lag features, the accuracy of XGBoost in predicting ozone concentrations was significantly improved, providing more robust data support for environmental monitoring and policy formulation [9]. Tang et al. integrated multi-source information such as reanalysis meteorological data based on the RF machine learning algorithm, and provided a new scheme for high-resolution ground concentration prediction of multiple pollutants [10]. To balance mechanism analysis and high-precision prediction, the academic community has begun to introduce interpretable machine learning methods into pollution control research to endow "black-box" models with transparency. Marvin et al. used feature selection methods to replace the traditional neighborhood feature constraints, and combined Shapley value to analyze the interpretability of the model, which improved the accuracy of high ozone concentration prediction [11]. Gagliardi and Andenna combined supervised and unsupervised machine learning algorithms with the Shapley value interpretation method, providing a scientific basis for formulating ozone prevention and control measures and health risk intervention measures [12].

In summary, CTMs offer clear advantages in mechanistic analysis, while machine learning models excel in predictive accuracy and computational efficiency. Interpretability methods further provide reliable support for decision-making. However, existing research lacks deep integration of high-resolution online observations with ensemble learning and interpretability analysis. Against this backdrop, the present study is the first to integrate TAP real-time monitoring data with ensemble algorithms such as XGBoost and LightGBM, along with SHAP-based interpretability analysis. This approach achieves accurate and efficient ozone prediction at national and regional scales, and uncovers the spatiotemporal heterogeneity of key features. The findings provide an innovative methodological framework for targeted pollution control. The summary of related work is shown in Table 1.

Table 1: Summary of related works

Study	Model	Data Type	Main Contribution/Performance	Limitation
Lasry et al. (2007)	CTM	Urban emission inventory + meteorological data	Simulated ozone response under emergency measures	High computational cost, poor real-time performance
Li et al. (2023)	CTM	Atmospheric chemistry + meteorological data	Revealed coupling between precursors and meteorology	Sensitive to initial conditions, complex parameterization
Liu et al. (2025)	XGBoost and other ML	Historical O_3 and NO_2 concentrations	Improved accuracy by incorporating lagged features	Lack of interpretability methods
Tang et al. (2024)	RF	Reanalysis data + satellite AOD + CTM output	Achieved 1 km resolution multi-pollutant estimation	Limited interpretability of features
Marvin et al. (2022)	ML + SHAP	Meteorological + topographic data	Enhanced ozone prediction in complex terrain, identified key drivers	Limited data, regional applicability
Gagliardi & Andenna (2025)	Supervised + unsupervised ML + SHAP	Multi-source monitoring data	Unveiled meteorology-dominated drivers and nonlinear relations	Did not consider high-frequency online observations
This study	XGBoost, LightGBM + SHAP	TAP online monitoring data + meteorological data	Nationwide and regional O_3 high-accuracy prediction; revealed spatiotemporal heterogeneity of key drivers	——

3 Research methodology

3.1 Prevention and control of ozone pollution

Ground-level ozone (O_3) is a typical secondary pollutant. It is primarily formed through a series of photochemical reactions involving precursor gases, nitrogen oxides (NO_x) and VOCs, under ultraviolet (UV) radiation. The chemical reactions are as follows [13-15]:

$$NO_2 + h\nu(\lambda < 420nm) \rightarrow NO + O \quad (1)$$

$$O + O_2 + M \rightarrow O_3 + M \quad (2)$$

$$O_3 + NO \rightarrow NO_2 + O_2 \quad (3)$$

$h\nu$ is UV light. λ is the wavelength.

In addition, VOCs generate peroxy radicals (RO_2) under photochemical activation, which accelerate the NO to NO_2 conversion cycle [16]:

$$k(T) = Aexp\left(-\frac{E_a}{RT}\right) \quad (4)$$

$k(T)$ is the rate constant at temperature T. $Aexp$ is the frequency factor. E_a is the activation energy, and R is the gas constant. The principle of ozone formation is shown in Figure 1 [17].

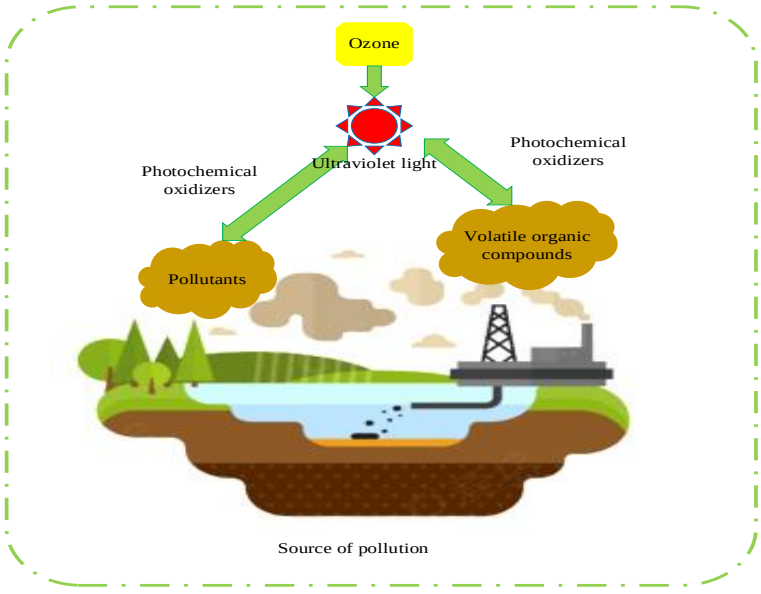


Figure 1: Formation principle of ozone

The precursors NO_x and VOCs have complex and diverse sources, including coal-fired power plants, motor vehicle exhaust, and biological emissions. Their formation mechanisms exhibit nonlinear characteristics and are affected by the interaction between regional pollution transport and meteorological conditions. Therefore, a single emission reduction measure is difficult to significantly reduce ozone concentration. The key characteristics of ozone pollution control are illustrated in Figure 2 [18].

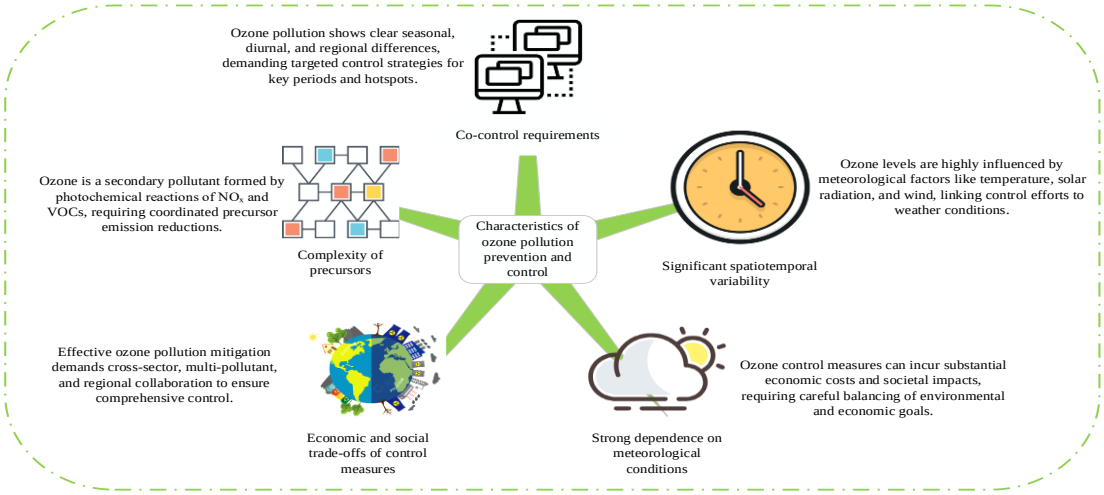


Figure 2: Characteristics of ozone pollution prevention and control

3.2 National ozone concentration prediction model based on online data

This study aims to explore the capability of predicting ozone concentrations nationwide and in key regions based on multi-source online observation data and machine learning methods, with a focus on evaluating the advantages of machine learning models integrated with SHAP interpretation in capturing regional heterogeneity and short-term sudden pollution events. To clarify the research direction, the following research questions are proposed in this study:

Research Question 1: Can machine learning models based on SHAP interpretation outperform traditional learning methods in predicting ozone concentrations nationwide and in key regions?

Research Question 2: Do the contributions of various environmental and meteorological factors to ozone concentrations exhibit significant heterogeneity across different regions?

This study uses online observation data from the TAP dataset [19]. Developed by Tsinghua University in collaboration with multiple universities, this dataset integrates multi-source information including ground observations, satellite remote sensing, emission inventories, and model simulations, and provides multi-scale, near-real-time concentration data of atmospheric pollutants in China. This study selects ozone concentration data and corresponding meteorological variables from 34 representative cities during the period 2015–2025. The data preprocessing process is as follows: First, data from stations with a missing rate exceeding 10% are excluded, leaving a total of approximately 45,620 valid observation points. For locally missing data, Lagrange interpolation is used for imputation. Considering its smoothness and fidelity for continuous observation data in time series, this method can better maintain the continuity of local trends compared with K-Nearest Neighbors (KNN) or Multiple Imputation by Chained Equations (MICE) methods. Subsequently, the input features are normalized to the range of [0,1] using the min-max method, and Box-Cox transformation is performed to reduce the impact of skewed distribution on model training. In the Box-Cox transformation, the λ parameter selected for each feature is automatically estimated by the maximum likelihood method, and the specific values are shown in Table 2.

Table 2: The λ parameter values selected for each feature

Features	Box-Cox λ Parameter	Description
Day Order	0.15	Approximates linear transformation
Temperature	0.23	Mitigates skewed distribution
Sunshine Duration	0.1	Mitigates skewed distribution
NO_2	0.05	Mitigates skewed distribution
Wind Speed	0	Logarithmic transformation
Relative Humidity	0.18	Mitigates skewed distribution
Air Pressure	1	No transformation needed

This study adopts a rigorous machine learning modeling framework to construct and validate five predictive models: XGBoost [20], LightGBM [21], RF [22], SVM [23], and Backpropagation Neural Network (BPNN) [24]. The complete features after feature selection are shown in Table 3.

Table 3: Final selected feature list

Features	Description
Day Order	Sequential day within the year, capturing seasonality
Temperature	Daily mean temperature (°C)
Sunshine Duration	Actual daily sunshine hours (h)
NO_2	Daily mean NO_2 concentration ($\mu g/m^3$)
Wind Speed	Daily mean wind speed (m/s)
Relative Humidity	Daily mean relative humidity (%)
Air Pressure	Daily mean atmospheric pressure (hPa)

Subsequently, Bayesian optimization algorithm is used to iteratively optimize the core hyperparameters of each model. The specific values of the hyperparameters of this research model are shown in Table 4.

Table 4: Final model hyperparameters

Model	Hyperparameter	Final Value
XGBoost	Learning rate	0.05
	Maximum tree depth	6
	Subsample ratio of training instances	0.8
	Subsample ratio of features per tree	0.7
	Number of boosting trees	500
LightGBM	Learning rate	0.03
	Maximum number of leaves	31
	Maximum tree depth	7
	Number of boosting trees	400
RF	Number of trees	300
	Maximum tree depth	10
	Minimum samples required to split	2
	Minimum samples required at leaf node	1
SVM	Penalty parameter	10
	Kernel coefficient	0.01
BP Neural Network	Neurons in hidden layers	(50,30)
	Initial learning rate	0.001
	Maximum iterations	500

In the model training phase, time-series stratified sampling is adopted for dataset division, with 70% allocated to the training set and 30% to the test set. Subsequently, dual-track validation involving time-sliding window prediction and independent test set is conducted to comprehensively evaluate the generalization ability of the model and its performance differences across the whole country and key regions. The workflow of machine learning-based ozone concentration prediction is shown in Figure 3 [25].

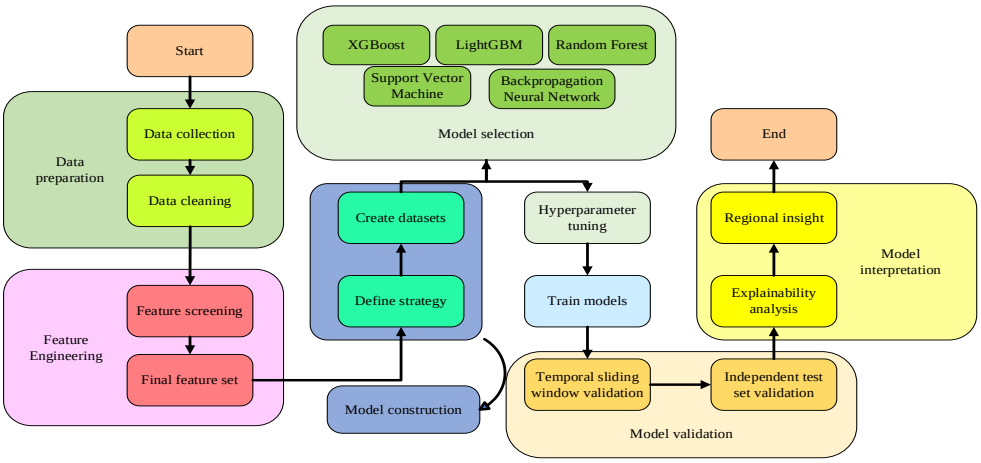


Figure 3: Ozone concentration prediction process based on machine learning

To comprehensively evaluate the predictive performance of each model, this study adopts the following six evaluation metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), and Nash-Sutcliffe Efficiency Coefficient (NSE). The equations for each metric are defined as follows [26]:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

$$R^2 = 1 - \frac{i=1}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (8)$$

n is the total number of samples. y_i is the i-th observed value (true value). $\hat{y}_i$ is the i-th predicted value, and $\bar{y}$ is the arithmetic average of true values.

3.3 Interpretability analysis of prediction models based on SHAP value

To quantify the contribution of each feature to ozone prediction from a game-theoretic perspective, this study introduces the SHAP method [27]. In the specific implementation process, considering that the prediction models in this study are mainly based on tree models, the TreeSHAP variant is adopted to improve computational efficiency and interpretation accuracy. The TreeSHAP algorithm is optimized for tree structures, enabling fast and accurate calculation of feature contribution while ensuring the fair distribution principle of Shapley values, which is suitable for interpretive analysis of large-scale environmental monitoring data. The interpretability analysis workflow based on SHAP values is illustrated in Figure 4 [28].

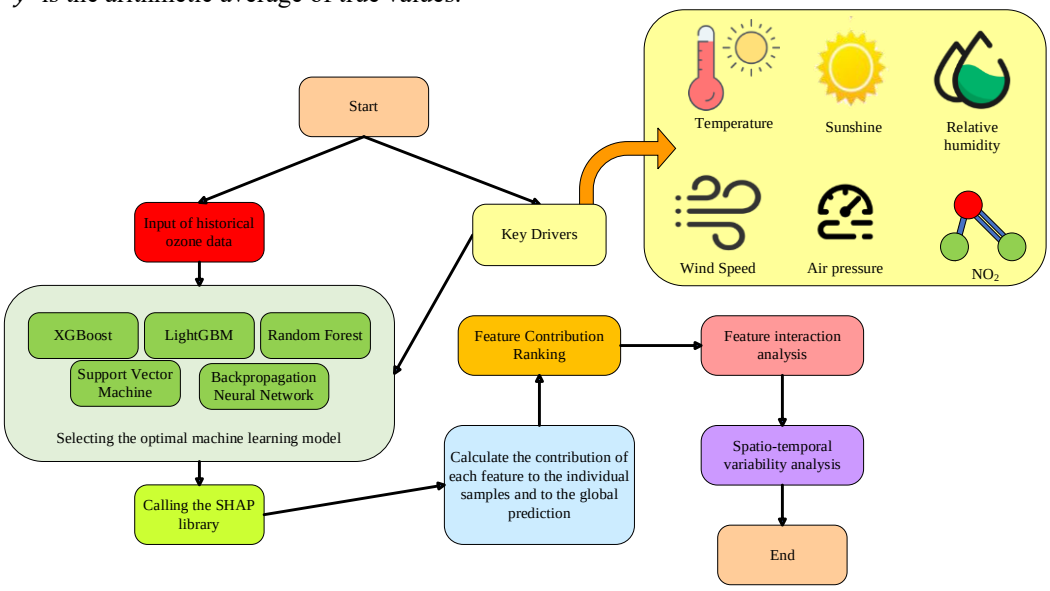


Figure 4: Explanatory analysis process of prediction model based on SHAP value

Let the feature set $F = [1,2,\ldots,D]$ (D features in total), and define the calculation formula of the value function $v(S)$ as follows [29]:

$$v(S) = \mathbb{E}[f(X)|X_S = x_S] \quad (9)$$

$v(S)$ is the expected output of the model prediction when the current sample value x_S is taken for any subset $S \sqsubseteq F$ while preserving S features. Shapley value ϕ_k represents the marginal average incremental contribution

of the kth feature in all possible cooperative subsets, and its calculation equation is as follows [30]:

$$\phi_k = \sum_{S \subseteq F \setminus (k)} \frac{|S|!(D-|S|-1)!}{D!} [v(S \cup \{k\}) - v(S)] \quad (10)$$

$D!$ is the number of all features, $|S|!(D-|S|-1)!$ is the normalization factor of subset weight, and $[v(S \cup \{k\}) - v(S)]$ is the incremental income of feature k after adding subset S. Taking the primary model prediction as a "cooperative game" to distribute income, the value function of sample x can be obtained, and the calculation equation is as follows:

$$v_x(S) = \mathbb{E}_{X_{\setminus S}}[f(x_S, X_{\setminus S})] - \mathbb{E}[f(X)] \quad (11)$$

x_S represents the observed value of the sample on the feature of subset S. $X_{\setminus S}$ represents the random distribution of other features. Therefore, the equation for calculating the SHAP value of the k-th feature under sample x is as follows:

$$\phi_k = \sum_{S \subseteq F \setminus (k)} \frac{|S|!(D-|S|-1)!}{D!} [v_x(S \cup \{k\}) - v_x(S)] \quad (12)$$

$$f(x) = \phi_0 + \sum_{k=1}^{D} \phi_k \quad (13)$$

$\phi_0 = \mathbb{E}[f(X)]$ is the baseline prediction value of the model, and $f(x)$ is additivity.

4 Experimental performance assessment results

4.1 Performance assessment results

The comparative results of the annual average ozone concentration in major regions of China are shown in Figure 5.

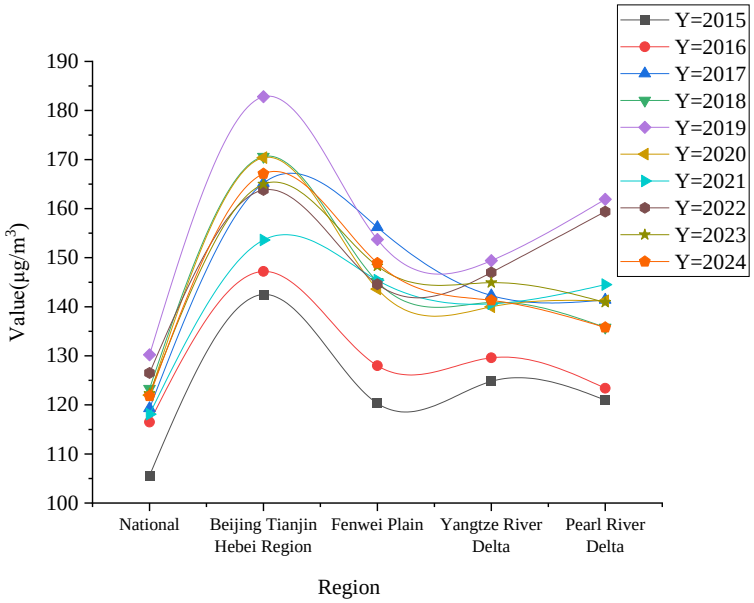


Figure 5: Comparative results of time variation of average annual ozone concentration in major regions of China

In Figure 5, the national ozone concentration presents a three-stage trend of "rapid rise-short-term decline-stabilization at a high level": the baseline value was 105.6 μg/m³ in 2015, reached a peak of 130.2 μg/m³ in 2019 (with an increase of 23.3%), dropped to 122 μg/m³ in 2020 due to emission reduction during the epidemic (a decrease of 6.3%), rebounded to 126.5 μg/m³ in 2022, and then stabilized at around 122 μg/m³ from 2023 to 2024, which was still 15.3% higher than that in 2015. At the regional level, the Beijing-Tianjin-Hebei region suffered from persistent severe pollution, with a concentration of 167.1 μg/m³ in 2024 (an increase of 17.2% over the decade). The Fenwei Plain showed a fluctuating downward trend after an abnormal surge in 2017, and its concentration in 2024 was 23.8% higher than that in 2015. The Yangtze River Delta remained relatively stable, with a concentration in 2024 5.4% lower than the peak in 2019, making it the only region with a continuous downward trend. The Pearl River Delta experienced severe fluctuations, with a concentration dropping to 135.8 μg/m³ in 2024 (a decrease of 16.1% from the peak), but the stability of emission reduction measures was insufficient.

The spatial distribution results and time distribution results of the average annual ozone concentration in various provinces in China are shown in Figs. 6 and 7. The gradient color in the figure indicates the ozone concentration in different regions (low in blue and high in red).

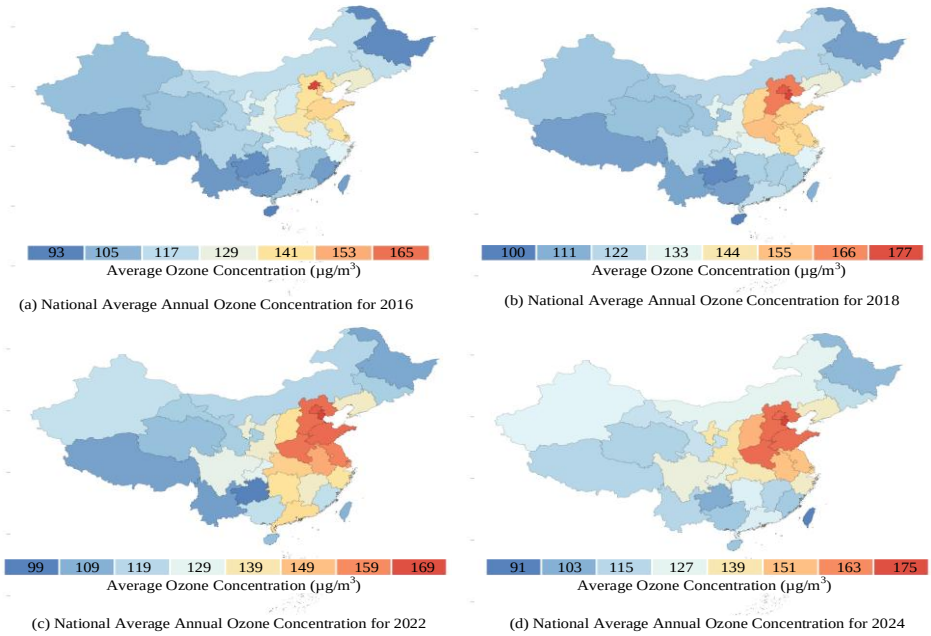


Figure 6: Spatial distribution results of average annual ozone concentration in provinces of China

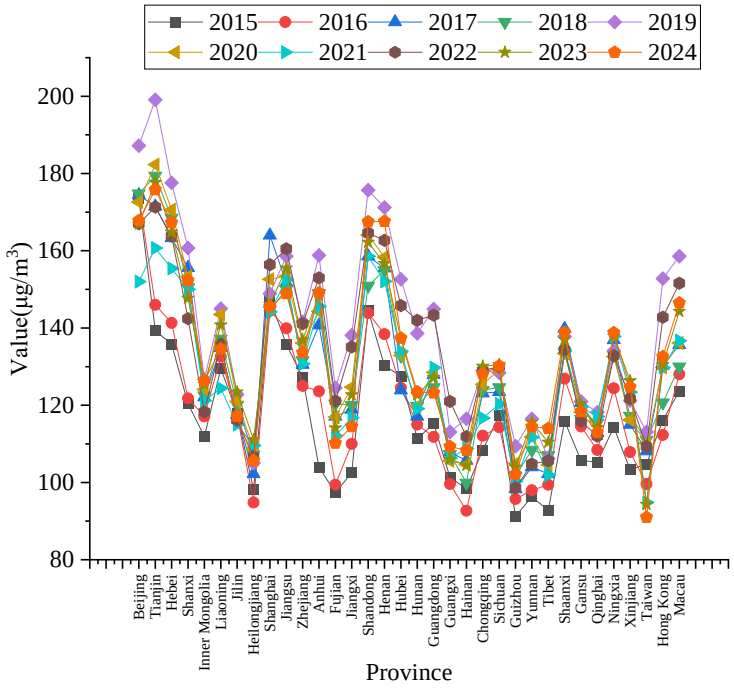
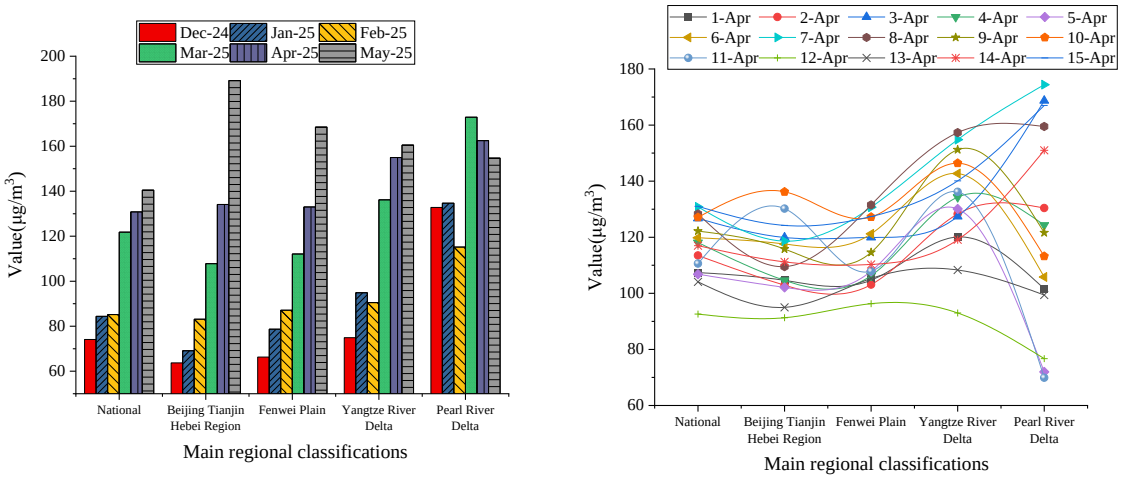


Figure 7: Comparative results of time variation of average annual ozone concentration in provinces of China

In Figure 6 and Figure 7, the annual average ozone concentration across provinces in China presents a spatial pattern of "higher in the east than in the west, and higher in the north than in the south": Provinces with high levels of industrialization and urbanization in the central and eastern regions (such as those in the Beijing-Tianjin-Hebei region, Shandong, Henan, as well as Jiangsu, Shanghai, Zhejiang in the Yangtze River Delta, and Guangdong in the Pearl River Delta) have relatively high ozone concentrations, mostly exceeding 150 μg/m³. Provinces in the western region such as Tibet, Guizhou, and Yunnan have lower concentrations, mostly around 100 μg/m³. This pollution pattern is related to economic development, industrial emissions, and meteorological conditions. In terms of time, the annual average ozone concentration across all provinces in China generally shows a fluctuating upward trend: it continued to rise to a peak from 2015 to 2019, decreased in 2020 due to the epidemic, rebounded in some provinces starting from 2021 and remained at a high level from 2022 to 2023. In 2024, some provinces saw a slight decrease, but the overall concentration still fluctuated at a high level, indicating significant pressure on pollution control.

Figure 8 presents a comparison of the monthly and

daily average ozone concentrations over time for major regions nationwide.



(a) Average monthly ozone concentration

(b) Average daily ozone concentration in April 2025

Figure 8: Comparative results of time variation of average monthly and daily ozone concentration in main areas of China

In Figure 8a, the monthly average concentration in all regions shows an upward trend from winter to spring. The national average increases from 74.1 μg/m³ to 140.5 μg/m³, with the Beijing-Tianjin-Hebei region recording the largest increase (from 63.7 μg/m³ to 189.2 μg/m³). The Pearl River Delta has the highest initial value (132.8 μg/m³), but it decreases slightly after reaching a peak of 172.9 μg/m³ in March. In Figure 8b, the national average of daily average concentration rises from 107.4 μg/m³ to 131.0 μg/m³ and has significant low points, while the Pearl River Delta exhibits the most severe fluctuations (ranging from 69.9 μg/m³ to 174.4 μg/m³). Overall, the monthly and daily average concentrations in the Beijing-Tianjin-Hebei region and the Yangtze River Delta rise steadily, while the Pearl River Delta and the Fenwei Plain show stronger spatiotemporal heterogeneity. This highlights the seasonal accumulation and short-term suddenness of ozone pollution.

In this study, rolling time-series K-fold cross-validation (K=5) is adopted for model training and testing to ensure the robustness of model evaluation. The predictive capabilities of different models for ozone concentrations nationwide and in key regions are presented in Table 5.

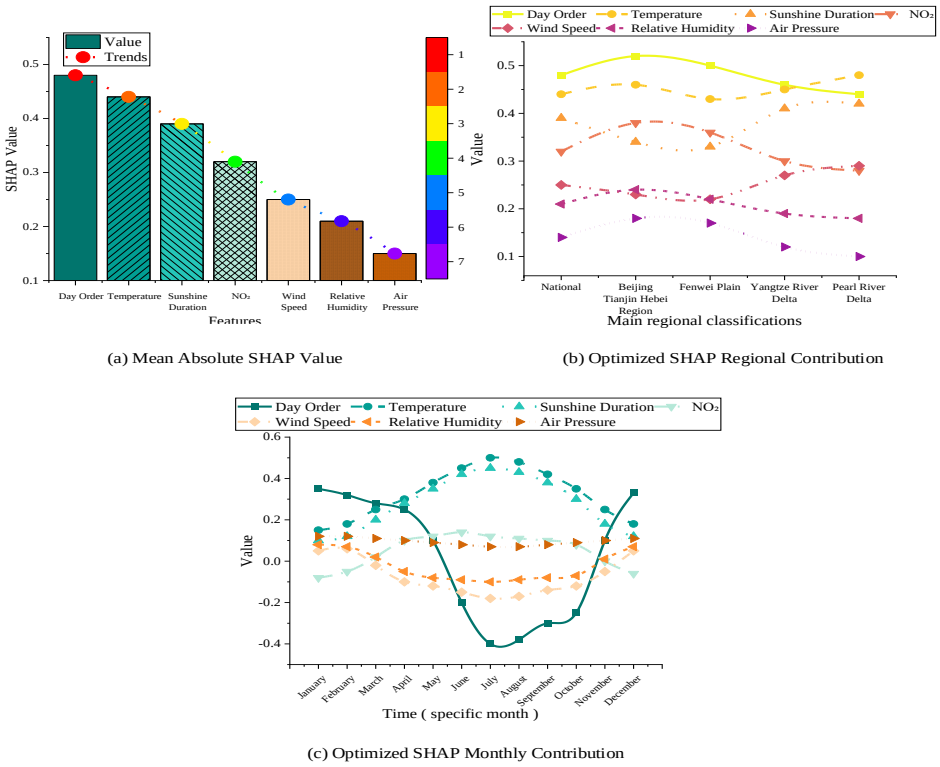
Table 5: The prediction ability of different models for ozone concentration in the whole country and key areas

Region	Model	MAE (μg/m³)	RMSE (μg/m³)	R^2	NSE	MAE vs XGBoost t-test (p-value)	RMSE vs XGBoost t-test (p-value)
National	DT	16.3±1.2	23.2±2.1	0.738±0.03	0.73±0.03	<0.001	<0.001
	RF	12.4±0.9	17.3±1.4	0.854±0.02	0.85±0.02	0.02	0.03
	SVM	19.3±1.4	26.1±2.2	0.669±0.03	0.66±0.03	<0.001	<0.001
	BP	19.5±1.5	26.1±2.3	0.669±0.03	0.66±0.03	<0.001	<0.001
	XGBoost	11.3±0.8	15.6±1.1	0.882±0.02	0.88±0.02	-	-
	LGBM	11.5±0.9	16.0±1.2	0.878±0.02	0.87±0.02	0.21	0.25
Beijing–Tianjin–Hebei Region	DT	17.5±1.3	24.5±2.2	0.720±0.03	0.71±0.03	<0.001	<0.001
	RF	13.0±1.0	18.0±1.5	0.840±0.02	0.83±0.02	0.03	0.04
	SVM	20.0±1.5	27.0±2.3	0.650±0.03	0.64±0.03	<0.001	<0.001
	BP	20.2±1.5	27.2±2.4	0.650±0.03	0.64±0.03	<0.001	<0.001
	XGBoost	12.0±0.9	16.5±1.2	0.870±0.02	0.86±0.02	-	-

	LGBM	12.1±0.9	16.7±1.3	0.865±0.02	0.86±0.02	0.28	0.3
	DT	17.0±1.2	24.0±2.1	0.730±0.03	0.72±0.03	<0.001	<0.001
	RF	12.8±0.9	17.8±1.4	0.850±0.02	0.84±0.02	0.04	0.05
	SVM	19.8±1.4	26.8±2.2	0.660±0.03	0.65±0.03	<0.001	<0.001
Fenwei Plain	BP	20.0±1.5	27.0±2.3	0.660±0.03	0.65±0.03	<0.001	<0.001
	XGBoost	11.8±0.8	16.2±1.1	0.880±0.02	0.87±0.02	-	-
	LGBM	12.0±0.9	16.5±1.2	0.875±0.02	0.87±0.02	0.23	0.27
	DT	15.0±1.1	22.0±2.0	0.750±0.03	0.74±0.03	<0.001	<0.001
	RF	11.2±0.8	16.0±1.3	0.860±0.02	0.85±0.02	0.02	0.03
	SVM	18.0±1.3	25.0±2.1	0.680±0.03	0.67±0.03	<0.001	<0.001
Yangtze River Delta	BP	18.5±1.4	25.5±2.2	0.680±0.03	0.67±0.03	<0.001	<0.001
	XGBoost	10.5±0.7	15.0±1.1	0.890±0.02	0.88±0.02	-	-
	LGBM	10.7±0.8	15.2±1.1	0.885±0.02	0.88±0.02	0.24	0.28
	DT	14.0±1.0	21.0±1.9	0.760±0.03	0.75±0.03	<0.001	<0.001
	RF	10.5±0.7	15.0±1.2	0.870±0.02	0.86±0.02	0.02	0.03
	SVM	17.5±1.3	24.0±2.1	0.690±0.03	0.68±0.03	<0.001	<0.001
Pearl River Delta	BP	18.0±1.4	24.5±2.2	0.690±0.03	0.68±0.03	<0.001	<0.001
	XGBoost	10.0±0.7	14.0±1.0	0.900±0.02	0.89±0.02	-	-
	LGBM	10.2±0.7	14.3±1.0	0.888±0.02	0.89±0.02	0.2	0.22

In Table 5, the XGBoost model performs the best with the highest prediction accuracy and good stability nationwide and in all major regions. On a national scale, the MAE of XGBoost is 11.3 μg/m³, the RMSE is 15.6 μg/m³, the R^2 is 0.882, and the NSE is 0.88, which is significantly better than traditional regression and neural network models. At the regional level, the XGBoost model achieves the best performance in the Pearl River Delta, with an MAE of only 10.0 μg/m³, an RMSE of 14.0 μg/m³, and an R^2 of 0.900. In the Beijing-Tianjin-Hebei region and the Fenwei Plain, the MAE values are 12.0 μg/m³ and 11.8 μg/m³ respectively, the RMSE values are 16.5 μg/m³ and 16.2 μg/m³ respectively, and the R^2 values both exceed 0.86. This indicates that the ensemble learning method has significant advantages in capturing the spatiotemporal variation characteristics of ozone concentration. Moreover, the results of the paired t-test show that the differences in MAE and RMSE between XGBoost and other models are statistically significant ($p<0.05$), which further verifies the reliability and scientific nature of its prediction performance.

The interpretable analysis results based on SHAP value are shown in Figure 9.



(a) Mean Absolute SHAP Value

(b) Optimized SHAP Regional Contribution

(c) Optimized SHAP Monthly Contribution

Figure 9: Interpretable analysis results based on SHAP value

Analysis of Figure 9 indicates that, at the national scale, calendar day, temperature, and sunshine duration are the dominant factors driving ozone concentration variations, consistently showing significant influence. The average contribution value of NO_2 in the Beijing-Tianjin-Hebei region is 0.38, and that in the Fenwei Plain is 0.36. Both values are higher than the national average of 0.32 and significantly higher than those in the Yangtze River Delta and the Pearl River Delta. This indicates that in the Beijing-Tianjin-Hebei region and the Fenwei Plain, the contribution ratio of NO_2 to the model prediction results is significantly higher than that in other regions, which reflects the key role of precursor emissions in photochemical reactions in these regions. In contrast, the Yangtze River Delta and Pearl River Delta regions exhibit greater contributions from temperature and sunshine duration, with the strongest promoting effects observed in the Pearl River Delta. Wind speed and relative humidity have comparatively smaller overall contributions. However, wind speed plays a relatively larger role in the Pearl River Delta and Yangtze River Delta, indicating that atmospheric dispersion conditions in these regions significantly modulate ozone concentration changes. These findings provide targeted guidance for region-specific ozone pollution control strategies across China.

The SHAP summary results of the model and the dependence results of the top three important features are shown in Figure 10.

In Figure 10, there are obvious differences in the importance and action direction of each feature on the model output. From the perspective of the SHAP value of day of year, it generally makes a positive contribution to the prediction results. With the increase of day of year, the SHAP value shows an upward trend, indicating that the model is sensitive to the cumulative effect of the time series. For temperature, the SHAP value fluctuates greatly: samples with high temperature correspond to positive contributions, while samples with low temperature correspond to negative contributions. This reflects the nonlinear dependence of temperature on the prediction results. Regarding sunshine duration, its SHAP value contributes more to the model output when the sunshine value is relatively low or extremely high, while the variation range in the middle interval is relatively small. This indicates that sunshine duration has an obvious threshold effect on the prediction results. On the whole, the dependence plots of these three features show that day of year has a steady positive dependence, temperature has a nonlinear fluctuating dependence, and sunshine duration has a threshold-type dependence. These plots provide an intuitive reference for the interpretation of model features.

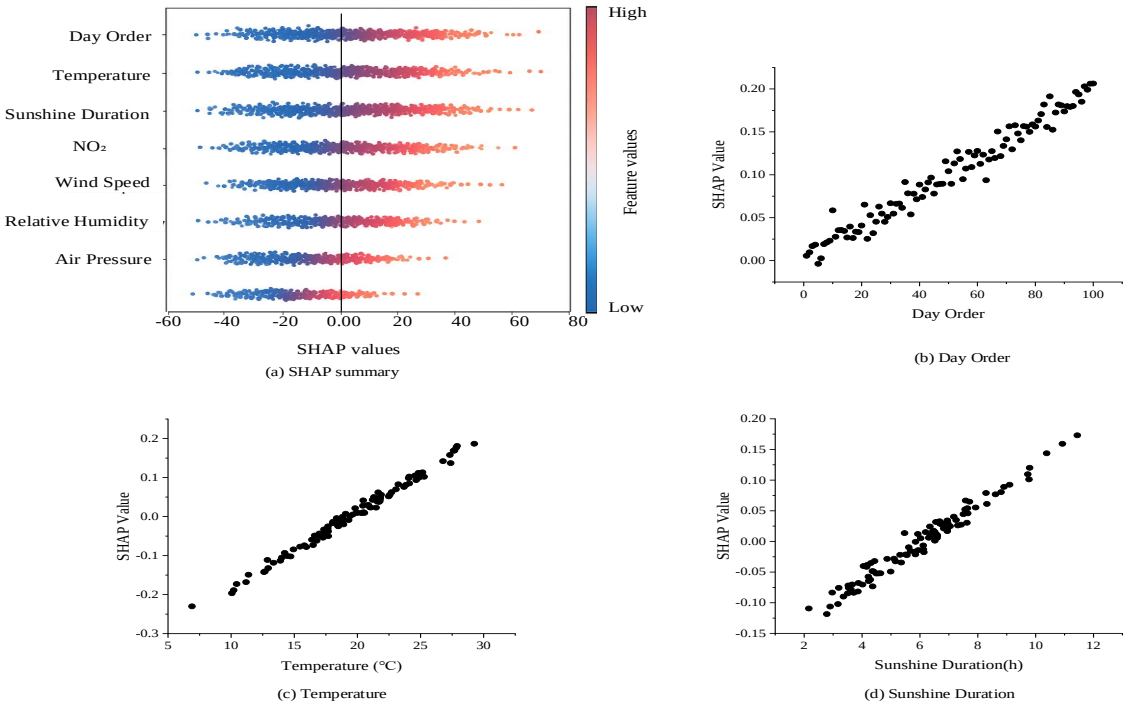


Figure 10: Interpretable analysis results based on SHAP value

4.2 Discussions

Compared with existing CTMs, statistical regression models, and traditional machine learning models, the model in this study performs excellently in both prediction accuracy and robustness. On a national scale, the XGBoost model has an MAE of 11.3 μg/m³, an RMSE of 15.6 μg/m³, and an R^2 of 0.882, which is significantly better than decision tree, support vector machine, and BPNN models ($p<0.05$). In the Pearl River Delta region, the MAE of the XGBoost model is only 10.0 μg/m³, the RMSE is 14.0 μg/m³, and the R^2 reaches 0.900. The main reasons for the performance advantages of this study are as follows: High spatiotemporal resolution online observation data are adopted, which enhances the comprehensiveness and representativeness of the model input information. Through feature selection and lag feature construction, the nonlinear impacts of precursor emissions and meteorological conditions on ozone formation are effectively captured. Based on the ensemble learning algorithm, the model exhibits excellent fitting ability and generalization performance in handling the nonlinear coupling relationships of multiple variables, thereby significantly improving the prediction accuracy and stability.

Based on the results of SHAP analysis, this study reveals the regional heterogeneity of driving factors for ozone formation. In regions dominated by precursor emissions, such as the Beijing-Tianjin-Hebei region and the Fenwei Plain, NO_2 and VOCs make significant contributions to changes in ozone concentration. In contrast, regions dominated by meteorological conditions, such as the Yangtze River Delta and the Pearl River Delta, are mainly affected by factors like temperature and sunshine. This indicates that ozone pollution control in different regions should adopt differentiated strategies targeting the main driving factors to improve the efficiency and accuracy of prevention and control.

On the basis of scientific analysis, this study can further provide a quantitative basis for policy-making. On the one hand, for regions dominated by precursor emissions, priority should be given to several key strategies. These include: Strengthening the coordinated emission reduction of key emission sources such as NO_2 and VOCs. Optimizing industrial, transportation, and energy structures. Realizing precise source control in combination with high-frequency emission monitoring. On the other hand, for regions dominated by meteorological conditions, it is necessary to strengthen dynamic regulation and emergency response. This is particularly important under extreme meteorological conditions such as high temperature and low wind speed. Specific measures should include temporary traffic restrictions and staggered production for high-emission enterprises. In addition, nationwide, the ozone online monitoring network and integrated intelligent forecasting system should be continuously improved. Through data sharing and cross-departmental collaborative governance, the prevention and control of ozone pollution should be promoted towards digitalization, intelligence, and refinement to provide reliable support for public health and ecological environment security.

5 Conclusion

By integrating multi-scale ozone observation data with an interpretable machine learning framework, this study systematically reveals the spatiotemporal evolution characteristics and formation mechanisms of ozone pollution in China. The results show that the national

ozone concentration from 2015 to 2024 presents a three-stage trend of "rise-decline-stabilization", and the concentration in 2024 is still 15.3% higher than the baseline value in 2015. At the regional level, the Beijing-Tianjin-Hebei region and the Fenwei Plain maintain a continuously high concentration level, highlighting the pressure of pollution control. Meanwhile, the economically dense areas in eastern China have exceeded the standard for a long time, showing a differentiation pattern of "higher in the east and lower in the west", which confirms that industrial emissions and urbanization are the core features. Based on the prediction of the XGBoost model and SHAP analysis, the ozone formation mechanism shows obvious regional heterogeneity: The Beijing-Tianjin-Hebei region and the Fenwei Plain are mainly dominated by the emission of precursors such as NO_2 (with SHAP contribution ratio > 0.35), while the Yangtze River Delta and the Pearl River Delta are sensitive to temperature and sunshine. The model achieves the best prediction performance in the Pearl River Delta, providing a scientific reference for dynamic and precise management and control.

The limitations of this study are mainly reflected in factor coverage and model input: limited by existing monitoring data and available variables, some short-term sudden pollution events and complex chemical reaction mechanisms have not been fully captured. In addition, the quantitative analysis of external factors such as social and economic development and policy intervention is relatively limited. Future research can improve the model's ability to predict short-term pollution events. This can be achieved by introducing socio-economic datasets such as traffic flow and industrial activities. At the same time, it is necessary to strengthen the integration of multi-source and multi-dimensional data. This will enhance the model's adaptability to extreme meteorological conditions and nonlinear changes. In terms of model deployment and application, this study also provides practical implications: although XGBoost has high prediction accuracy, there are certain costs in training time and SHAP feature contribution calculation. In the future, more efficient computing strategies or model compression methods can be explored to balance accuracy and computing efficiency, thereby providing operable guidance for model developers and environmental management practitioners.

References

[1] Kong L, Song M, Li X, Liu Y, Lu S, Zeng L, et al. Analysis of China's PM2. 5 and ozone coordinated control strategy based on the observation data from 2015 to 2020. Journal of Environmental Sciences, 2024, 138(1): 385-394. https://doi.org/10.1016/j.jes.2023.03.030

[2] Li Z, Bi J, Liu Y, Hu X. Forecasting O_3 and NO_2 concentrations with spatiotemporally continuous coverage in southeastern China using a Machine learning approach. Environment International, 2025, 195(1): 109249. https://doi.org/10.1016/j.envint.2024.109249

[3] Carbo-Bustinza N, Iftikhar H, Belmonte M, Cabello-Torres RJ, Cruz ARHDL, López-Gonzales JL. Short-term forecasting of Ozone concentration in metropolitan Lima using hybrid combinations of time series models. Applied Sciences, 2023, 13(18): 10514. https://doi.org/10.3390/app131810514

[4] Rahman A, Nasher N M R. Forecasting hourly ozone concentration using functional time series model-A case study in the coastal area of bangladesh. Environmental Modeling & Assessment, 2024, 29(1): 125-134. https://doi.org/10.1007/s10666-023-09928-8

[5] Xie J, Tang X, Zheng F, Wang X, Ding N, Song Y, et al. Improvement of the ozone forecast over Beijing through combining the chemical transport model with multiple machine learning methods. Atmospheric Pollution Research, 2024, 15(8): 102184. https://doi.org/10.1016/j.apr.2024.102184

[6] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017, 30. https://doi.org/10.48550/arXiv.1705.07874

[7] Lasry F, Coll I, Fayet S, Havre M, Vautard R. Short-term measures for the control of ozone peaks: expertise from CTM simulations. Journal of atmospheric chemistry, 2007, 57(1): 107-134. https://doi.org/10.1007/s10874-007-9062-1

[8] Li J, Jang J, Zhu Y, Wang C, Xing S, Dong J, et al. Development of a recurrent spatiotemporal deep-learning method coupled with data fusion for correction of hourly ozone forecasts. Environmental Pollution, 2023, 335(1): 122291. https://doi.org/10.1016/j.envpol.2023.122291

[9] Liu Z, Lu Z, Zhu W, Yuan J, Cao Z, Cao T, et al. Comparison of machine learning methods for predicting ground-level ozone pollution in Beijing. Frontiers in Environmental Science, 2025, 13(1): 1561794. https://doi.org/10.3389/fenvs.2025.1561794

[10] Tang B, Stanier CO, Carmichael GR, Gao M. Ozone, nitrogen dioxide, and PM2. 5 estimation from observation-model machine learning fusion over S. Korea: Influence of observation density, chemical transport model resolution, and geostationary remotely sensed AOD. Atmospheric Environment, 2024, 331(1): 120603. https://doi.org/10.1016/j.atmosenv.2024.120603

[11] Marvin D, Nespoli L, Strepparava D, Medici V. A data-driven approach to forecasting ground-level ozone concentration. International Journal of Forecasting, 2022, 38(3): 970-987. 10.1016/j.ijforecast.2021.07.008

[12] Gagliardi R V, Andenna C. Exploring the Influencing Factors of Surface Ozone Variability by Explainable Machine Learning: A Case Study in the Basilicata Region (Southern Italy). Atmosphere, 2025, 16(5): 491. https://doi.org/10.3390/atmos16050491

[13] Haagen-Smit A J. Chemistry and physiology of Los Angeles smog. Industrial & Engineering Chemistry, 1952, 44(6): 1342-1346. https://doi.org/10.1021/ie50510a045

[14] Chapman S. XXXV. On ozone and atomic oxygen in the upper atmosphere. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1930, 10(64): 369-383. https://doi.org/10.1080/14786443009461588

[15] Chapman S. Bakerian Lecture-Some phenomena of the upper atmosphere. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 1931, 132(820): 353-374. https://doi.org/10.1098/rspa.1931.0105

[16] Chu W, Li H, Ji Y, Zhang X, Xue L, Gao J, et al. Research on ozone formation sensitivity based on observational methods: Development history, methodology, and application and prospects in China. Journal of Environmental Sciences, 2024, 138(1): 543-560. https://doi.org/10.1016/j.jes.2023.02.052

[17] Wang X, Shao T, Qin J, Li Y, Long X, Jiang D, et al. Promotion effect of micro-hole in dielectric on ozone generation of dielectric barrier discharge. Ozone: Science & Engineering, 2024, 46(4): 345-354. https://doi.org/10.1080/01919512.2023.2301548

[18] Li Y, Wu Z, Ji Y, Chen T, Li H, Gao R, et al. Comparison of the ozone formation mechanisms and VOCs apportionment in different ozone pollution episodes in urban Beijing in 2019 and 2020: Insights for ozone pollution control strategies. Science of The Total Environment, 2024, 908(1): 168332. https://doi.org/10.1016/j.scitotenv.2023.168332

[19] Geng G, Xiao Q, Liu S, Liu X, Cheng J, et al. Tracking air pollution in China: near real-time PM2. 5 retrievals from multisource data fusion. Environmental Science & Technology, 2021, 55(17): 12106-12115. https://doi.org/10.1021/acs.est.1c01863

[20] Guo Q, He Z, Wang Z. The characteristics of air quality changes in Hohhot City in China and their relationship with meteorological and socio-economic factors. Aerosol and Air Quality Research, 2024, 24(5): 230274. https://doi.org/10.4209/aaqr.230274

[21] Kumar G D, Tyagi S, Pradhan K C, Shah A. District-level rainfall and cloudburst prediction using XGBoost: A machine learning approach for early warning systems. Informatica, 2025, 49(2). https://doi.org/10.31449/inf.v49i2.7612

[22] Chen C. Football match analysis and prediction based on light GBM decision algorithm. Informatica, 2024, 48(16). https://doi.org/10.31449/inf.v48i16.6263

[23] Oukhouya M H, Angour N, Aboutabit N, Hafidi I. Comparative analysis of ARDL, LSTM, and XGBoost models for forecasting the moroccan stock market during the COVID-19 pandemic. Informatica, 2025, 49(14). https://doi.org/10.31449/inf.v49i14.5751

[24] Rawat R, Raj ASA, Chakrawarti RK, Sankaran KS, Sarangi SK, Rawat H, et al. Enhanced cybercrime detection on twitter using aho-corasick algorithm and machine learning techniques. Informatica, 2024, 48(18). https://doi.org/10.31449/inf.v48i18.6272

[25] Lai H. Predicting the growth value of technology enterprises with an optimized back-propagation neural network. Informatica, 2024, 48(16). https://doi.org/10.31449/inf.v48i16.6437

[26] Cheng M, Fang F, Navon IM, Zheng J, Zhu J, Pain C. Assessing uncertainty and heterogeneity in machine learning-based spatiotemporal ozone prediction in Beijing-Tianjin-Hebei region in China. Science of the Total Environment, 2023, 881(1): 163146. https://doi.org/10.1016/j.scitotenv.2023.163146

[27] Yao T, Lu S, Wang Y, Li X, Ye H, Duan Y, et al. Revealing the drivers of surface ozone pollution by explainable machine learning and satellite observations in Hangzhou Bay, China. Journal of Cleaner Production, 2024, 440(2): 140938. https://doi.org/10.1016/j.jclepro.2024.140938

[28] Zhao B, Wang S, Hao J. Challenges and perspectives of air pollution control in China. Frontiers of Environmental Science & Engineering, 2024, 18(6): 68. https://doi.org/10.1007/s11783-024-1828-z

[29] Nath S J, Girach I A, Harithasree S, Bhuyan K, Ojha N, Kumar M. Urban ozone variability using automated machine learning: inference from different feature importance schemes. Environmental Monitoring and Assessment, 2024, 196(4): 393. https://doi.org/10.1007/s10661-024-12549-7

[30] Han L, Zhao J, Gao Y, et al. Prediction and evaluation of spatial distributions of ozone and urban heat island using a machine learning modified land use regression method. Sustainable Cities and Society, 2022, 78(1): 103643. https://doi.org/10.1016/j.scs.2021.103643