scientific reports

Check for updates

OPEN

Machine learning identifies environmental drivers of pulmonary tuberculosis in Xinjiang a typical arid region of China

Feifei Li[1,4], Liping Zhang[1,4]✉, ChenChen Wang[2], Peiyao Zhou[1], Qin Xu[3] & Yanling Zheng[1]

The Xinjiang Uyghur Autonomous Region in northwest China experiences a disproportionately high burden of pulmonary tuberculosis (PTB) compared to global averages, yet the environmental determinants driving this epidemic in arid regions remain poorly understood. This study aims to quantify the combined effects of multiple environmental factors on PTB incidence, reveal their non-linear characteristics, and fill the research gap regarding the environmental driving mechanisms in the northwest region. This study integrated PTB incidence data from 14 regions in Xinjiang from 2010 to 2022, along with data on five air pollutants ($PM_{2.5}$, PM_{10}, NO_2, O_3, and CO) and four meteorological indicators (average temperature, average humidity, average wind speed, and average rainfall). Comparative modeling was conducted using the Gradient Boosting Decision Tree (GBDT) and the Extreme Gradient Boosting (XGBoost) models. The Shapley Additive Explanations (SHAP) values were employed to analyze variable contributions and exposure-response relationships. Model performance was evaluated using R^2, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The XGBoost model demonstrated superior performance in fitting complex non-linear relationships and handling high-dimensional data interactions, with a coefficient of determination of 0.91, significantly higher than the 0.49 achieved by the GBDT model. SHAP analysis revealed that PM_{10} was the most predominant risk factor (mean concentration of 142.10 μg/m³, exceeding the WHO guideline limit by 14 times; ranking first in SHAP contribution), followed by CO, average temperature, and $PM_{2.5}$. The exposure-response curves for PM_{10} and CO exhibited a monotonic increasing trend. There was a "protective threshold" for wind speed (4.0–5.5 m/s), beyond which aerosol dispersion mitigated PTB transmission. When precipitation exceeded 10 mm, the risk of PTB decreased, indicating either a protective or a promoting effect on disease transmission under specific conditions. Dust-related PM_{10} and coal combustion-derived CO are the primary environmental drivers of PTB in the arid Xinjiang ecosystem. The XGBoost-SHAP framework effectively elucidates complex environmental health effects. The findings support the formulation of regional prevention and control strategies targeting dust pollution and coal combustion emissions, providing a new pathway for environmental interventions to achieve the goal of ending tuberculosis.

Keywords Tuberculosis, Air pollution, Environmental exposure, Explainable machine learning, Arid region

Abbreviations

PTB	Pulmonary Tuberculosis
GBDT	Gradient Boosting Decision Tree
XGBoost	Extreme Gradient Boosting
SHAP	SHapley Additive exPanations
WHO	World Health Organization
AT	Average Temperature
WS	Average Wind Speed
AR	Average Rainfall

[1]Institute of Medical Engineering Interdisciplinary Research, College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830017, China. [2]Center for Disease Control and Prevention of Xinjiang Uygur Autonomous Region, Urumqi 830002, China. [3]The First Affiliated Hospital of Xinjiang Medical University, Urumqi 830013, China. [4]Feifei Li and Liping Zhang contributed equally to this work. ✉email: zhanglp1219@163.com

AH	Average Humidity
KNN	K-Nearest Neighbors
SD	Standard Deviations
IQR	Interquartile Ranges
VIF	Variance Inflation Factor
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
R^2	Coefficient of Determination
E- R	Exposure-Response
ORs	Odds Ratios

Tuberculosis remains a formidable public health challenge globally[1–3]. According to the latest *Global Tuberculosis Report 2024* released by the World Health Organization (WHO)[4], there were approximately 10.8 million new TB cases and 1.3 million TB-related deaths worldwide in 2023. China accounted for 7.1% (approximately 767,000 cases) of the global TB incidence, ranking third in terms of disease burden, following India and Indonesia. Notably, the prevalence of TB in China exhibits significant regional disparities. In 2021, the incidence rate in Xinjiang reached as high as 128 per 100,000 population, which was 2.4 times the national average (53.1 per 100,000 population), and the rate of decline over the past decade has lagged behind that of eastern provinces[5,6]. This abnormal distribution suggests that, in addition to biomedical factors, environmental driving mechanisms may be a crucial explanatory pathway[7,8].

A growing body of evidence indicates that environmental factors are closely associated with the transmission dynamics of TB[9]. Air pollutants, such as $PM_{2.5}$ and NO_2, can impair respiratory immune function[10–13], while meteorological conditions, such as temperature and humidity, indirectly modulate transmission risks by affecting the survival rate of *Mycobacterium tuberculosis* aerosols[14,15]. The Xinjiang Uygur Autonomous Region, situated in northwestern China and at the heart of the Eurasian continent, represents a quintessential arid and semi-arid zone. Its distinctive geographical structure—defined by the Tianshan Mountains partitioning the Junggar and Tarim Basins, and bordered by expansive deserts such as the Taklamakan-combines with an extreme continental climate featuring low annual precipitation, frequent dust storms, and a substantial dependence on coal for winter heating to form a unique environmental profile. Major emission sources comprise natural mineral dust derived from extensive arid terrains and desert areas, as well as anthropogenic pollutants originating from coal combustion used for centralized heating during the prolonged cold season, complemented by increasingly significant contributions from vehicular and industrial emissions in urban centers. These conditions result in consistently high levels of particulate matter (especially PM_{10}) and unique patterns of air pollutant exposure, which may significantly influence the transmission dynamics of respiratory infectious diseases like TB. However, the TB epidemic in Xinjiang has long been characterized by "three highs and one low" (high infection rate, high prevalence rate, high rural epidemic, and low annual decline rate), with a disease burden far exceeding the national average. There is an urgent need to delve into its environmental driving factors to formulate precise prevention and control strategies. Existing studies have predominantly focused on single environmental factors or employed traditional regression models[16–19], which are limited in capturing the combined effects of multiple factors and non-linear relationships. Moreover, these studies have predominantly been geographically concentrated in the developed eastern regions, with insufficient attention paid to the environmental specificities of the arid northwestern regions of China[20–22]. Additionally, traditional machine learning models have limitations in ranking variable importance and providing causal explanations, lacking in-depth exploration of model interpretability[23,24]. In response to the above-mentioned deficiencies, this study integrates TB incidence data, air pollutant concentrations, and meteorological indicators in Xinjiang from 2010 to 2022. It constructs and comparatively analyzes GBDT and XGBoost machine learning models, and uses SHAP values to elucidate the marginal contributions and action directions of various environmental factors.

This study is the first to apply the XGBoost-SHAP interpretable framework to quantify complex nonlinear interactions between multiple environmental factors and PTB in arid regions; identify dust-driven PM_{10} and coal-combustion derived CO as predominant drivers specific to arid ecosystems, differing from mechanisms in humid areas; propose targeted environmental intervention strategies for dust control and clean energy transition in northwestern China. These approaches provide new insights into the environmental driving mechanisms of TB and offer a scientific basis for the formulation of regional public health policies, contributing to the achievement of the global goal of ending the TB epidemic.

Methods

Data collection

Global and Chinese TB incidence data were obtained from the Global Tuberculosis Report 2023 published by the WHO and the Global Tuberculosis Database. The number of reported TB cases and incidence rates from 2010 to 2022 for 14 regions (prefectures) in Xinjiang were extracted from the "Infectious Disease Reporting Information Management System" of the China Information System for Disease Control and Prevention (excluding data from outside the province and the Xinjiang Production and Construction Corps). The data were aggregated according to the date of onset. Air pollutant monitoring data, including CO (mg/m^3), O_3 ($\mu g/m^3$), NO_2 ($\mu g/m^3$), $PM_{2.5}$ ($\mu g/m^3$), and PM_{10} ($\mu g/m^3$), were sourced from the China National Environmental Monitoring Centre's Real-time Air Quality Monitoring Platform (https://air.cnemc.cn/). Meteorological data, comprising average temperature (AT, ℃), average wind speed (WS, m/s), average rainfall (AR, mm), and average humidity (AH, %), were obtained from the National Oceanic and Atmospheric Administration of the United States.

Data pre-processing

To ensure data quality and consistency, a comprehensive pre-processing pipeline was implemented. Missing values in the dataset were addressed using the K-Nearest Neighbors (KNN) imputation method, which allowed for informed estimation of missing values based on the characteristics of the ten nearest data points. This method was selected for its ability to preserve the underlying data structure and reduce bias compared to mean or median imputation, particularly in time-series environmental data. In this study, Z-score standardization was used to process continuous variables and eliminate dimensional influence. The variance inflation factor (VIF) was utilized to detect multicollinearity, revealing potential collinearity among air pollutants and meteorological factors. The results indicated weak or no multicollinearity (Supplementary Material Table S1).

Statistical method

Statistical analysis

The preprocessed dataset was divided into a training subset containing 70% of the data and a test subset containing 30% of the data. Continuous variables were described using means and standard deviations (SD) or interquartile ranges (IQR). Pearson correlation analysis was used to investigate the correlations among all exposure variables. Additionally, restricted cubic splines were used to model potential non-linear relationships between each air pollutant exposure, meteorological factor, and TB incidence. All statistical analyses were primarily conducted using statistical packages in R 4.1.3, with the statistical significance level set at 0.05.

Gradient boosted decision tree

GBDT is a classical machine learning method proposed by Friedman et al.[25]. GBDT employs decision trees as weak learners and iteratively corrects model errors through a gradient descent strategy. In contrast to the parallel construction of random forests, the sequential training mechanism of GBDT enables it to focus more on correcting the residuals of the preceding models, often resulting in advantages in prediction accuracy[26]. In this study, an exhaustive grid search was conducted over a predefined parameter space to automate hyperparameter optimization. The performance of each unique hyperparameter combination was rigorously evaluated using a robust 10-fold cross-validation protocol. To mitigate the risk of overfitting and ascertain the optimal number of iterations for each specific hyperparameter configuration, early stopping was integrated into each training run during the cross-validation process. The optimal number of iterations is depicted in the supplementary material (Fig. S1). The final model reported in the results was trained utilizing the best-performing hyperparameter set identified through this optimization procedure. The implementation was based on the LightGBM framework[27], and a summary of the optimized hyperparameters is provided in Supplementary Table S2.

Extreme gradient boosting

XGBoost[28] is an efficient implementation of the GBDT proposed by Chen and Guestrin. Through technological innovations such as second-order derivative optimization, regularization control, and the Weighted Quantile Sketch, XGBoost significantly enhances the accuracy and training speed of traditional GBDT.

Key improvements include: Regularization in the loss function, the introduction of regularization terms into the loss, the function helps control model complexity (L1/L2 regularization), thereby mitigating overfitting. Second-order Taylor expansion, by utilizing both the first-order derivative (gradient) and the second-order derivative (Hessian) of the loss function, XGBoost optimizes the node splitting criterion. In a parallelized design, features such as pre-sorted feature blocks and cache optimization break through the computational bottlenecks of GBDT, enabling parallel processing. Automatic handling of missing values, XGBoost learns the default splitting direction for missing values, eliminating the need for preprocessing. XGBoost iteratively optimizes an objective function with regularization terms:

$$Obj^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T_t + \frac{1}{2} \lambda \|\omega_t\|^2 \quad (1)$$

Here, ${}^{c}Obj^{(t)}$ represents the total optimization objective at the t-th iteration, ^{c}n denotes the number of training samples, ${}^{c}g_i$ refers to the gradient, i.e., the first-order derivative of the loss function L with respect to the current predicted value, h_i represents the Hessian, i.e., the second-order derivative of the loss function ${}^{c}L$ with respect to the current predicted value, $f_t(x_i)$ is the predicted value of the t-th tree for the ${}^{c}x_i$ sample. The hyperparameter ${}^{c}\gamma$ is the penalty for leaf splitting to control the complexity of the tree, ${}^{c}T_t$ indicates the number of leaf nodes in the t-th tree, λ is the L2 regularization coefficient to prevent overfitting, and ${}^{c}\|\omega_t\|^2$ is the squared L2 norm of the leaf weight vector, which measures the model complexity. We set the tree depth to 4, the learning rate to 0.1, and the regularization coefficients γ = 0.3 and λ = 1. The optimal number of iterations is shown in the supplementary material (Fig. S2). The model was tuned through cross-validation, and a summary of the optimized hyperparameters is provided in Supplementary Table S2.

This study compares the predictive performance of the two models using the following quantitative metrics: R^2, RMSE, and MAE.

SHAP theory and interpretation

To enhance the interpretability of the machine learning model and quantify the marginal contribution of each environmental variable to the prediction of PTB incidence, we employed SHAP values. SHAP is a unified framework based on cooperative game theory that assigns an importance value to each feature for every individual prediction. Its core assumption is that the model's prediction result can be decomposed into the sum of contributions from each feature, namely, an additive explanatory model:

$$f(x) = \varphi_0 + \sum_{i=1}^{n} \varphi_i \tag{2}$$

Here, $f(x)$ represents the predicted value for a sample x, φ_0 is the baseline value (often the model's prediction with an empty feature set), $\varphi_i$and is the SHAP value of feature *i*. The SHAP value represents the contribution of that feature to the prediction result. Theoretically, SHAP values are precisely calculated by enumerating all possible feature combinations. For a sample containing features *n*, the formula for calculating the SHAP value φ_i of feature *i* is as follows:

$$\varphi_i = \sum\nolimits_{s \subseteq x\{x_i\}} \frac{|s|!\,(n - |s| - 1)!}{n!} \left[f(s \cup \{x_i\}) - f(s) \right] \tag{3}$$

In this formula, *s* is a feature subset that does not include feature *i*, is the size of the subset, and $|s|$ is the model's prediction under the feature subset *s*. This formula calculates the contribution of a feature *i* by computing the increment in the model's prediction when the feature *i* is added to different subsets *s*, and then taking a weighted sum based on the probability of each subset's occurrence.

SHAP typically graphically visualizes machine learning predictions to enhance presentation. For instance, the SHAP variable importance plot succinctly illustrates the contribution of each feature to the predictive performance: the larger the value, the greater the contribution[29,30]. This is of paramount importance for improving the interpretability and transparency of models, thereby aiding in enhancing the understanding and trust in disease screening.

Results

Descriptive analysis results

Figure 1 is a schematic diagram of the general situation of the study area, which illustrates the geographical, administrative, and environmental distribution patterns within the region, while also indicating the specific site locations of meteorological monitoring stations in Xinjiang. Xinjiang is situated in the northwest of China. Based on local geographical characteristics, this region can be divided into three parts: Southern Xinjiang, Northern Xinjiang, and Eastern Xinjiang. As illustrated in Fig. 2, which presents the epidemiological surveillance data of tuberculosis from 2010 to 2022, the global incidence rate of tuberculosis exhibited a gradual upward trend, increasing from 133 per 100,000 population to 138 per 100,000 population. China as a whole demonstrated a better performance compared to the global average, with its incidence rate dropping from 72 per 100,000 population to 58 per 100,000 population. However, the Xinjiang region displayed a distinct evolutionary pattern. From 2010 to 2017, the incidence rate in Xinjiang consistently remained higher than the national average, reaching its peak in 2018. Xinjiang had been in a state of hyper-high prevalence for an extended period before the rate began to decline.

The mean exposure concentrations of all air pollutants and meteorological factors in Xinjiang are summarized in Table 1. Notably, the annual mean concentrations of both $PM_{2.5}$ and PM_{10} substantially exceeded the limits set by the WHO (Air Quality Guidelines 2021: $PM_{2.5}$ = 5 μg/m^3, PM_{10} = 10 μg/m^3) and China's National Ambient Air Quality Standards (GB 3095 – 2012: $PM_{2.5}$ = 35 μg/m^3, PM_{10} = 70 μg/m^3). All meteorological factors exhibited varying degrees of fluctuation. The IQR was used to quantify the variability of each factor around the median. A larger IQR indicates greater variability in the data. AR showed the most significant fluctuation (IQR = 12.77 mm), indicating substantial variation in precipitation levels across the observation period, which is a characteristic feature of the arid continental climate in Xinjiang. In contrast, WS showed the least variability

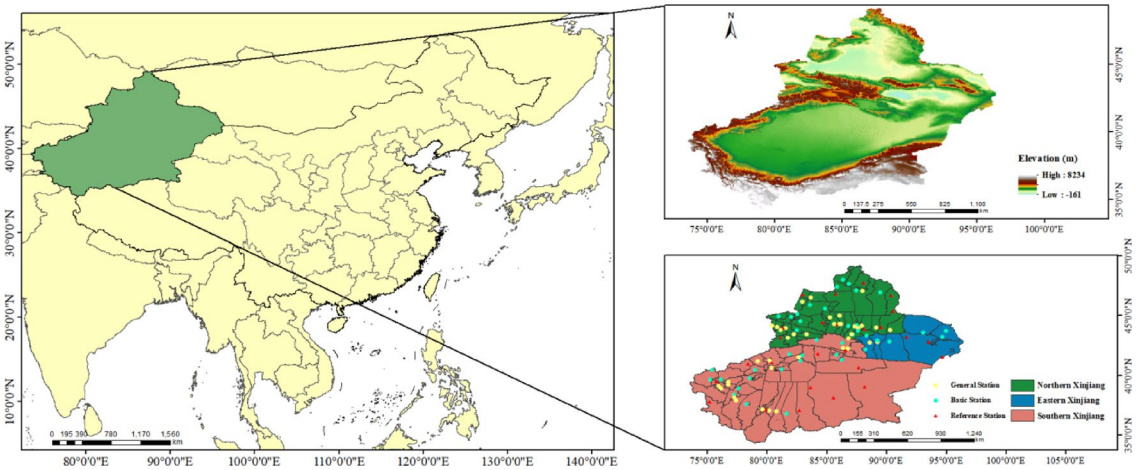

Fig. 1. Overview map of the study area.

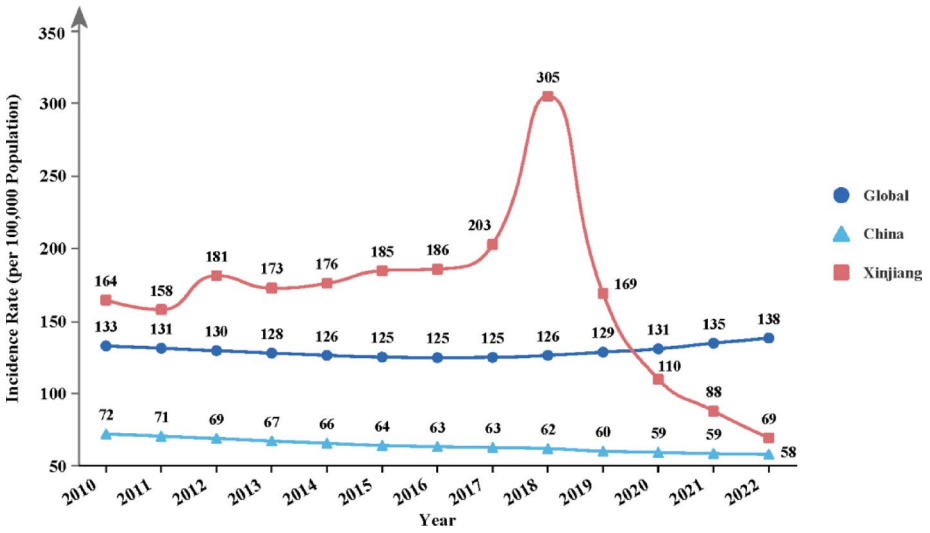

Fig. 2. Trend of tuberculosis incidence rates globally, in China and in Xinjiang from 2010 to 2022.

Air pollutants/ Meteorological factors	Mean	SD	IQR	(Min-Max)
$PM_{2.5}$ (μg/m^3)	40.92	15.93	22.64	12.65–96.53
PM_{10} (μg/m^3)	142.10	104.89	117.58	12.80-456.03
NO_2 (μg/m^3)	27.80	11.44	15.68	9.40-76.71
O_3 (μg/m^3)	62.41	13.87	14.78	11.33-104.74
CO (mg/m^3)	1.28	0.52	0.69	0.37–2.89
AT (℃)	6.21	2.43	3.86	2.09–10.98
WS (m/s)	4.68	0.96	0.93	3.26–7.94
AR (mm)	13.19	8.56	12.77	1.36–40.52
AH (%)	50.74	6.74	12.62	39.36–61.73

Table 1. Descriptive statistics of exposure concentrations of air pollutants and meteorological factors in Xinjiang. SD, standard deviations. IQR, interquartile range.

(IQR = 0.93 m/s), suggesting relatively stable wind conditions. Figure 3 presents the results of the Pearson correlation analysis between air pollutants and meteorological factors in Xinjiang, revealing the interactions among various environmental elements. In terms of internal correlations among air pollutants, PM_{10} and $PM_{2.5}$ exhibit an extremely close relationship, with a correlation coefficient as high as 0.86 ($P < 0.001$). Additionally, a positive correlation is observed between $PM_{2.5}$ and NO_2, with $r = 0.45$ ($P < 0.001$). The correlation coefficient between NO_2 and CO is $r = 0.35$ ($P < 0.001$), indicating a certain positive correlation in their concentration changes. In contrast, NO_2 and O_3 show a negative correlation, with $r = -0.42$ ($P < 0.001$), meaning that when the concentration of NO_2 increases, the concentration of O_3 tends to decrease. CO demonstrates a positive correlation with both PM_{10} and $PM_{2.5}$, with correlation coefficients of 0.23 ($P < 0.005$) for both, suggesting that changes in CO concentration have similar impacts on the concentrations of these two particulate matters. From the perspective of the associations between air pollutants and meteorological factors, there is a significant correlation between particulate matter and AH. Specifically, PM_{10} exhibits a strong negative correlation with AH, with $r = -0.65$ ($P < 0.001$); $PM_{2.5}$ also shows a strong negative correlation with AH, with $r = -0.56$ ($P < 0.001$). The effects of AR on different particulate matters vary. The correlation coefficient between AR and PM_{10} is–0.57 ($P < 0.001$), while that between AR and $PM_{2.5}$ is–0.48 ($P < 0.001$), which is relatively weaker. WS also has a non-negligible impact on particulate matter concentrations. WS shows a significant negative correlation with $PM_{2.5}$, with $r = -0.33$ ($P < 0.001$), and a similarly significant negative correlation with PM_{10}, with $r = -0.31$ ($P < 0.001$). Furthermore, there are numerous significant correlation relationships among meteorological factors. AH and WS exhibit a significant negative correlation, with $r = -0.35$ ($P < 0.001$), whereas AH and AR show an extremely strong positive correlation, with $r = 0.83$ ($P < 0.001$). Meanwhile, WS and AT are positively correlated, with $r = 0.35$ ($P < 0.001$); however, WS and AR are negatively associated, with $r = -0.34$ ($P < 0.001$).

Nonlinear exposure-response (E-R) relationship patterns between atmospheric pollutants, meteorological factors, and tuberculosis risk

The E-R curves in Fig. 4 depict the nonlinear associations between environmental factors and PTB incidence, expressed as Odds Ratios (ORs). An OR represents the ratio of the odds of PTB incidence occurring at a specific

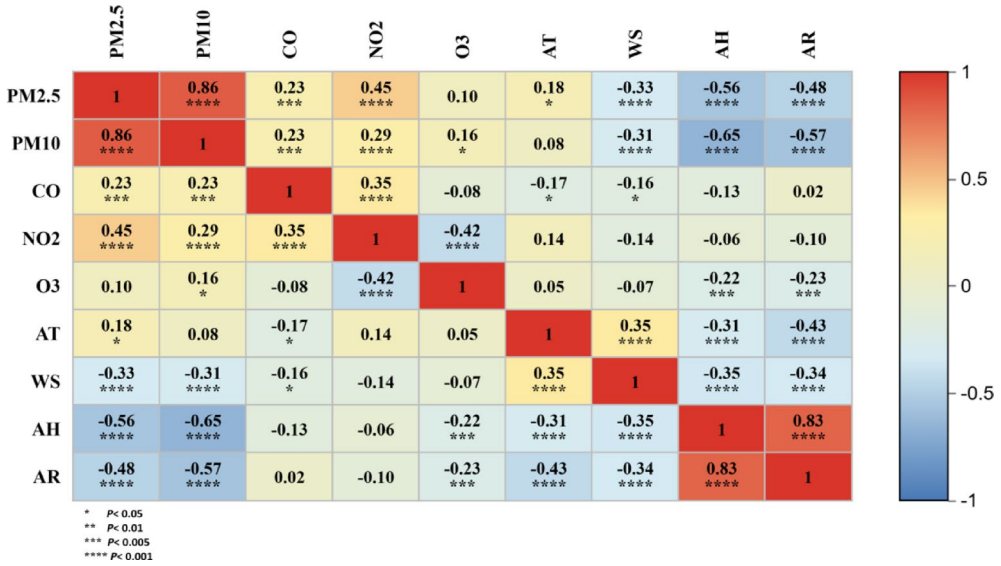

Fig. 3. The relationship between air pollutants and meteorological factors.

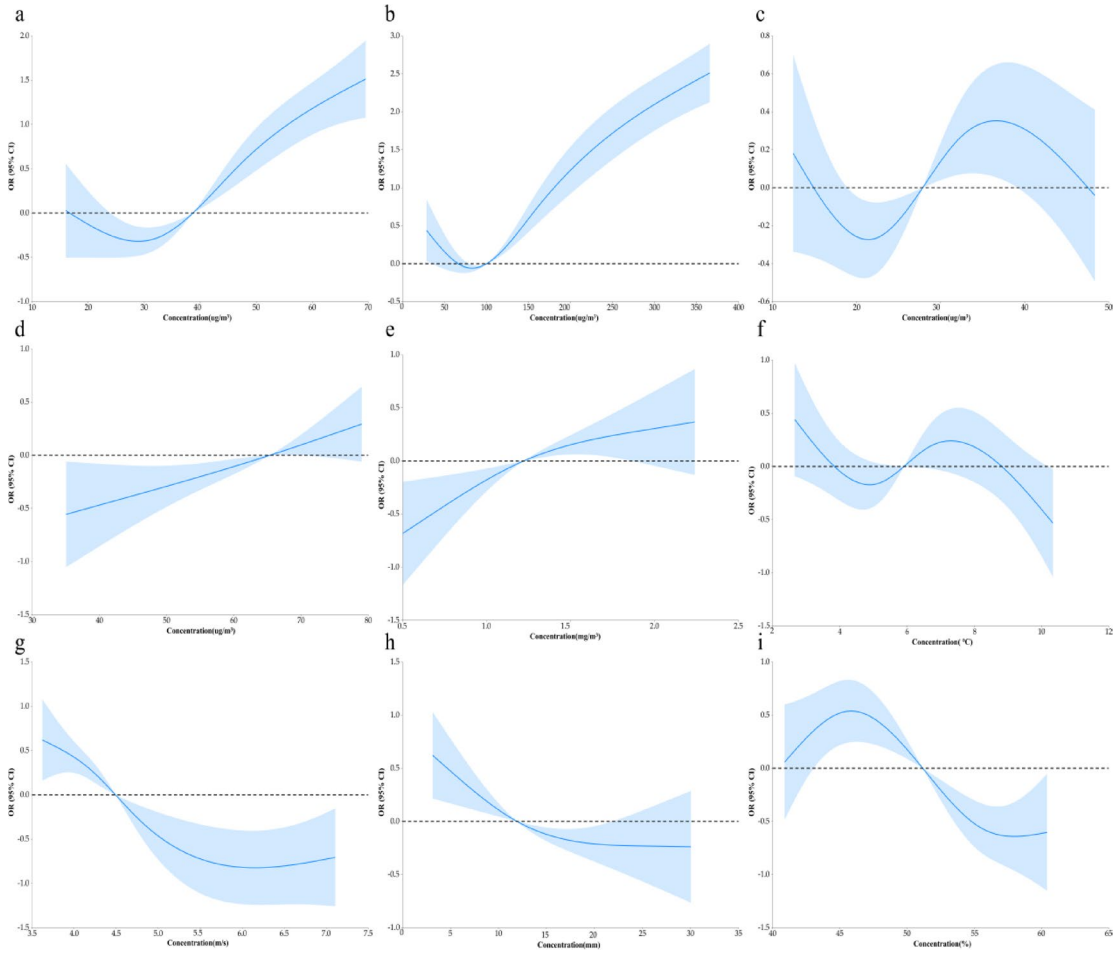

Fig. 4. The exposure-response relationship between air pollutants and meteorological factors and the incidence of tuberculosis. (**a**) $PM_{2.5}$. (**b**) PM_{10}. (**c**) NO_2. (**d**) O_3. (**e**) CO. (**f**) AT. (**g**) WS. (**h**) AR. (**i**) AH.

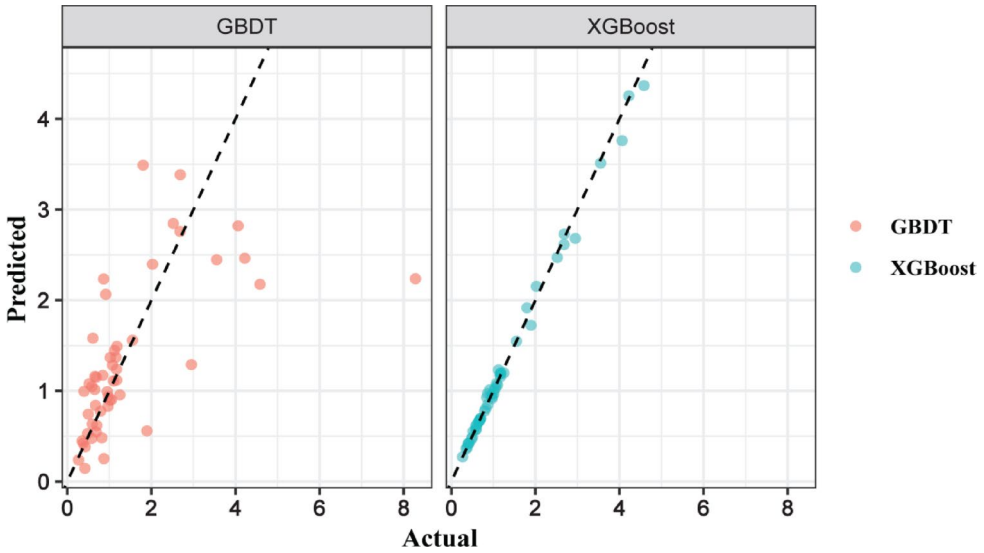

Fig. 5. Actual vs. Predicted Comparison.

Model	R^2	RMSE	MAE
XGBoost	0.906	0.438	0.084
GBDT	0.492	0.996	0.580

Table 2. Evaluation of model fitting effect. Evaluation of model prediction effect.

exposure level compared to a reference level (typically the median exposure value in this study). An OR greater than 1 indicates an increased risk of PTB, while an OR less than 1 suggests a protective effect against the disease. A detailed analysis of Fig. 4 reveals several key patterns. The concentrations of $PM_{2.5}$ and PM_{10} exhibit a pattern characterized as "gentle at low exposure, steep increase at high exposure" (Fig. 4a, b). This indicates that the risk of PTB increases marginally at lower concentrations but rises sharply when pollutant levels exceed a certain threshold. The E-R curves for NO_2 and AT display a wavy, non-monotonic shape (Fig. 4c, f). The risk fluctuates at lower concentrations and gradually stabilizes as concentrations increase, suggesting a complex, nonlinear influence on PTB transmission. In contrast, O_3 and CO show a predominantly monotonic increasing trend (Fig. 4d, e), implying a consistent rise in PTB risk with rising concentrations of these pollutants. The curve for WS presents a distinct inverted J-shape (Fig. 4g). A "protective threshold" is observed within the range of 4.0–5.5 m/s, where the OR is at its lowest. Deviating from this optimal range, either with lower or higher wind speeds, results in a significant increase in PTB risk. This is likely because moderate winds promote aerosol dispersion, while calm conditions lead to pollutant accumulation, and very strong winds may resuspend dust particles. AR demonstrates a significant protective effect when it exceeds approximately 10 mm (Fig. 4h), as precipitation likely removes airborne particles through wet deposition. The E-R curve for AH is relatively flat (Fig. 4i), though a slight increase in risk is observed in low-humidity intervals (< 45%).

Model fitting and performance analysis

As shown in Fig. 5, a comparison of the scatter plots of predicted values versus actual values reveals significant performance disparities between the XGBoost model and the GBDT model in tuberculosis data modeling. The predicted points of the XGBoost model are more closely clustered along the ideal fitting line (dashed line), and the range of residual distribution is notably narrower. This indicates that the XGBoost model has a significantly superior ability to capture complex nonlinear relationships in the data compared to the GBDT model, thereby reducing systematic biases in the prediction results.

Table 2 presents the quantitative evaluation results of the fitting performance of the two models on the tuberculosis dataset. The coefficient of determination of the XGBoost model ($R^2 = 0.906$) is significantly higher than that of the GBDT model ($R^2 = 0.492$), suggesting that the XGBoost model accounts for a substantially larger proportion of the data variance. Moreover, the XGBoost model demonstrates a marked improvement in prediction stability, as evidenced by its significantly lower RMSE and MAE compared to the GBDT model, indicating better error control capabilities.

The residual distribution plots of the XGBoost model presented in Fig. 6 offer intuitive diagnostic evidence. The model residuals for both the training set (Fig. 6a) and the test set (Fig. 6b) exhibit a random distribution pattern centered around the zero axis, with no discernible directional trends or regular fluctuations. This is consistent with the expected form of random errors.

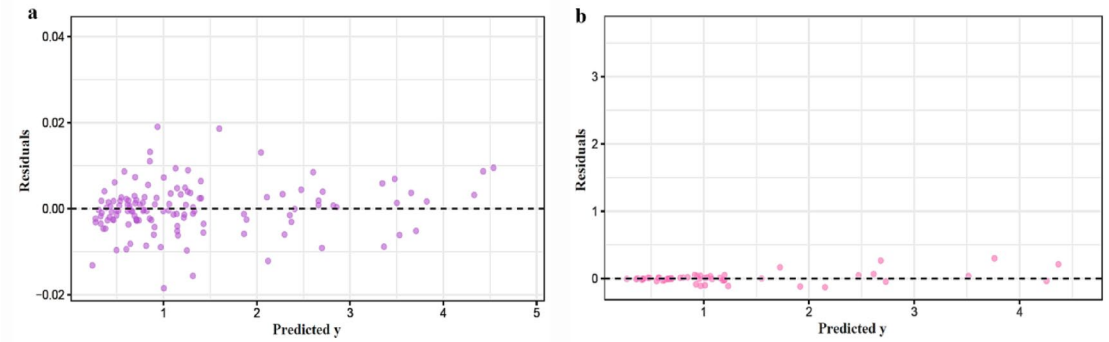

Fig. 6. Residuals vs. Predictions. (**a**) Train. (**b**) Test.

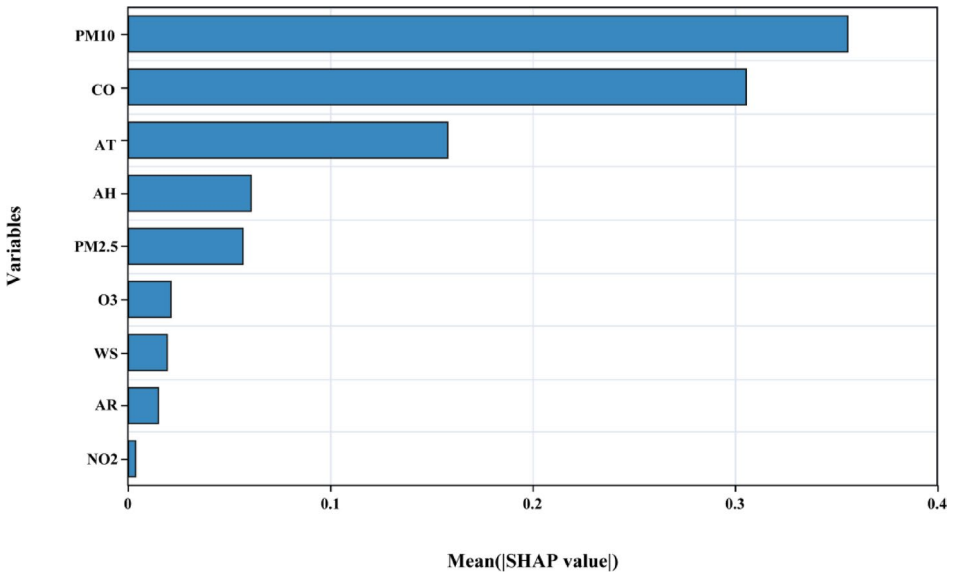

Fig. 7. Feature importance of the SHAP-based XGBoost explainer.

Analysis of feature contributions in the model

The global feature importance analysis (Fig. 7) conducted on the XGBoost model based on the SHAP method has unveiled a complex association pattern between the incidence rate of tuberculosis and environmental drivers in the Xinjiang region. The quantitative evaluation results indicate significant hierarchical differences in the contributions of various environmental variables to the disease burden. Among the multiple environmental indicators included in the analysis, PM_{10} exhibits a dominant influence, with its mean absolute SHAP value ranking first. This suggests that this variable has the strongest explanatory power in the model prediction process.

CO concentration, as a secondary influential factor, follows closely behind, and its feature importance score also reaches a statistically significant level, implying that gaseous pollutants may play an important role in the transmission dynamics of tuberculosis. Although AT, AH, and $PM_{2.5}$ have relatively lower contributions, they still maintain a moderate level of feature importance, indicating that these meteorological quality parameters constitute key components in the environmental risk spectrum for tuberculosis incidence. The feature importance scores for variables such as AR and NO_2 are at a lower magnitude. Among them, NO_2 has the weakest influence, suggesting that it may only have a marginal effect on the occurrence and development of the disease under specific exposure scenarios.

Discussion

This study focused on the association between the incidence data of PTB and environmental factors in the Xinjiang region from 2010 to 2022. By constructing GBDT and XGBoost machine learning models and combining them with SHAP interpretability analysis techniques, the study systematically dissected the nonlinear association

mechanisms and feature importance hierarchies between environmental exposure factors and tuberculosis incidence. The research findings provide new scientific evidence and decision-making support for tuberculosis prevention and control in arid regions.

This study indicates that the incidence rate of PTB in Xinjiang reached its peak in 2018 and then experienced a sharp decline in 2019. The main reasons for this phenomenon may be as follows: Firstly, the large historical base of PTB patients in Xinjiang has led to a slow decline in the epidemic, with the PTB incidence rate consistently ranking first in the country over the years[31]. Secondly, the implementation of a policy combining passive surveillance with active case-finding for tuberculosis, along with the comprehensive roll-out of initiatives such as screening for suspected PTB patients during the national health check-up in 2018 and PTB screening among high-risk populations, significantly contributed to the marked increase in the number of reported cases in Xinjiang in 2018. In addition, under the guidelines and instructions of the national investigation into under-reporting and under-registration of PTB cases and diagnostic re-verification work in China in 2017[32], Xinjiang launched a tuberculosis technical assistance program, which promoted the improvement of local PTB detection capabilities. This serves as another crucial reason for the sharp surge in the number of reported PTB cases in 2018. From 2018 to 2022, the PTB incidence rate in Xinjiang has shown a year-on-year decrease. This may be attributed to the continuous implementation of the tuberculosis technical assistance program, which has effectively enhanced the comprehensive technical skills of PTB prevention and treatment personnel in the region[33]. Moreover, patients identified through active screening have all been included in management, resulting in a certain degree of decline in the PTB incidence rate in Xinjiang after 2019.

In the evaluation of model performance, the XGBoost framework significantly outperformed GBDT in capturing complex environmental exposure-disease relationships ($R^2 = 0.91$ vs. 0.49). This is mainly attributed to its use of second-order Taylor expansion optimization, L1/L2 regularization constraints, and weighted quantile sketch techniques, which endow it with stronger feature capture capabilities when dealing with high-dimensional interactive features. This finding aligns with the theoretical advantages of the XGBoost algorithm proposed by Chen and Guestrin[28] and validates its applicability through empirical research in the field of public health.

The SHAP value analysis further revealed that PM_{10} holds a dominant position among environmental drivers in Xinjiang, followed by CO, AT, and $PM_{2.5}$. This conclusion is consistent with the findings of a multi-center study in East Asia[34], which identified particulate matter pollution as an important risk factor for respiratory infectious diseases. However, in Xinjiang, the contribution of PM_{10} is significantly higher than that of $PM_{2.5}$. This discrepancy may be attributable to the unique geographical and climatic characteristics of Xinjiang. Specifically, the arid climate, prevalent dust storms, and basin topography promote the generation and persistence of coarse particulate matter (PM_{10}), leading to its dominance over $PM_{2.5}$ as the primary particulate pollutant. The arid climate and frequent sandstorms in Xinjiang result in an annual average PM_{10} concentration of 142.1 μg/m^3 (14 times the WHO limit), far exceeding the levels in eastern Chinese cities. The coarse particulate matter in Xinjiang mainly originates from natural dust processes, such as the frequent dust storms originating from the surrounding deserts (e.g., the Taklamakan Desert), which is a dominant source throughout the year. Sand dust particles may exacerbate infection risks by damaging the respiratory mucosal barrier and promoting the colonization of Mycobacterium tuberculosis, a hypothesis that resonates with the pathological studies conducted by Rivas-Santiago et al.[35] and Glass[36].

As a secondary risk factor, CO exhibits a monotonically increasing exposure-response curve, which is closely related to widespread anthropogenic emissions from residential coal combustion for heating during the extended winter, a common practice in both rural and urban areas of Xinjiang. Raqib et al. confirmed in their study on household air pollution that CO can enhance tuberculosis susceptibility by inhibiting macrophage immune function[37]. This mechanism has also received support in a tuberculosis study in Lima, Peru[38], but it has not been highlighted in previous studies in the humid eastern regions of China and other research[39,40]. The annual average temperature in Xinjiang is only 6.2℃, with a significant seasonal temperature difference of 8.9℃. Extreme low-temperature conditions (< 4℃) may prolong the survival time of Mycobacterium tuberculosis in aerosols[41] and simultaneously increase indoor congregation behaviors among the population, thereby elevating the risk of droplet transmission.

The effects of meteorological factors exhibit significant nonlinear characteristics. There exists a "protective threshold" of wind speed WS ranging from 4.0 to 5.5 m/s, which is consistent with the dynamics of aerosol dispersion[42]. When WS is below this threshold, air stagnation leads to the accumulation of pollutants, while excessively high wind speeds may lift pathogen-bearing particulate matter from the surface[43]. When AR exceeds 10 mm, it can reduce the risk of transmission through the wet deposition effect. However, this effect is weaker in arid regions compared to humid areas, reflecting regional differences in dust deposition mechanisms. This study found that the risk of PTB increases when AH is below 45%, which is consistent with research conclusions that dry environments in temperate regions promote droplet transmission[44]. However, high humidity (>55%) does not show a significant protective effect, differing from conclusions in tropical studies[45]. This discrepancy may be related to the arid baseline characteristics of Xinjiang.

It is noteworthy that O_3 shows a positive association with PTB incidence, but its SHAP importance ranking is moderately low. This is mainly because the annual average concentration of O_3 in Xinjiang (62.4 μg/m^3) is lower than that in severely ozone-polluted areas in inland China, and only 14.3% of samples exceed the threshold concentration of 80 μg/m^3. Experimental studies have shown that O_3 may only significantly inhibit the phagocytic function of alveolar macrophages through oxidative stress when its concentration is above 80 μg/m[3][46]. The chemical mechanism of O_3 generation being suppressed under high PM_{10} conditions[47] further explains why its effect is masked by dominant factors. This is consistent with the conclusions of an Indian study[48] but differs from a report in Los Angeles, USA[49], where low particulate matter levels make O_3 the primary risk factor. The exposure-response curve of NO_2 exhibits a wavy pattern, and its SHAP importance ranking is the lowest. This nonlinear pattern suggests a possible duality in its role: the risk increases in the low-concentration zone (< 30

μg/m^3), which may reflect the spread of pathogens carried by primary pollutants from traffic emissions[50]; the effect weakens in the high-concentration zone (>50 μg/m^3), possibly due to the consumption of NO_2 through photochemical reactions to generate O_3, forming a "NO_2 - O_3" antagonistic effect[51,52]. In addition, the overall NO_2 concentration in Xinjiang is relatively low (mean value of 27.8 μg/m^3), with only 7.2% of samples exceeding China's secondary standard (40 μg/m^3), further limiting its overall influence.

Through multi-regional comparisons, this study found that the environmental exposure characteristics and disease-association patterns in Xinjiang are significantly different from those in inland China and other international regions[53–55]. Compared with other developed regions in China[56–58], Xinjiang's unique arid climate, frequent sandstorms, and energy structure characteristics lead to significant regional specificities in the intensity and pathways of environmental factors' effects on PTB incidence. These differences may result in varying intensities and patterns of environmental factors' impacts on PTB incidence in different regions. At the methodological level, this study conducted a comparative analysis of model selection, verifying the advantages of the XGBoost-SHAP framework in analyzing complex environmental health effects, providing a methodological reference for similar studies.

Limitations and policy implications

Despite its contributions, this study is subject to several limitations. Firstly, at the data level, relying on aggregated incidence data and environmental monitoring data from Xinjiang may introduce ecological bias. Moreover, individual-level environmental exposure and health data were not included, making it impossible to directly establish individual-level causal associations. Secondly, socioeconomic indicators (per capita GDP, medical accessibility), behavioral factors (smoking rate, population aggregation), and biological factors (HIV co-infection, drug resistance) were not incorporated into the model, which may overestimate the contribution of environmental factors, especially in impoverished and medically under-resourced areas such as rural Xinjiang. Furthermore, while SHAP values quantify the importance of environmental predictors, the findings lack direct mechanistic support from experimental studies, such as the survival of M. tuberculosis in aerosols under varying temperature/humidity conditions or the immunosuppressive pathways of air pollutants. Finally, the exposure-response relationships identified in the arid environment of Xinjiang may not be generalizable to humid regions due to fundamental differences in pollutant composition and climate patterns.

Notwithstanding these limitations, our findings carry important implications for public health policy. The identification of PM_{10} (largely dust-derived) and CO (primarily from coal combustion) as the dominant environmental drivers of PTB in Xinjiang suggests that regional control strategies should prioritize integrated dust suppression measures (e.g., afforestation, soil stabilization) and accelerated clean energy transition (e.g., replacing coal-based heating with electricity or natural gas in rural households). Moreover, the non-linear effects of meteorological factors suggest that public health advisories and intervention plans could be optimized based on seasonal weather patterns—for instance, issuing health warnings during low-wind periods (< 4.0 m/s) or high-dust seasons to reduce outdoor exposure. These targeted environmental interventions, combined with ongoing biomedical strategies, could significantly reduce the TB burden in arid regions and contribute to achieving the WHO's End TB goals.

Conclusions

This study systematically evaluated the composite effects of multiple environmental factors on the incidence of PTB within the specific environmental context of Xinjiang by integrating PTB incidence data, concentrations of various air pollutants, and meteorological indicators from 2010 to 2022. By comparing two machine learning models, GBDT and XGBoost, and employing the SHAP value interpretation framework, this study quantified the marginal contributions of individual environmental factors and revealed the direction and magnitude of the effects exerted by the primary environmental drivers.

The key findings are as follows: First, the XGBoost model demonstrated superior performance in fitting complex nonlinear relationships and handling high-dimensional data interactions ($R^2 = 0.91$), significantly outperforming the GBDT model ($R^2 = 0.49$). Second, SHAP analysis indicated that PM_{10} (mean concentration: 142.1 μg/m^3, exceeding the WHO guideline limit by 14-fold) was the most prominent environmental risk factor for PTB incidence in Xinjiang, followed by CO, mean temperature, and $PM_{2.5}$. Both PM_{10} and CO exhibited monotonic exposure-response relationships, with PTB risk increasing as their concentrations rose. Additionally, meteorological factors demonstrated significant nonlinear characteristics: a "protective threshold" for wind speed was identified within the 4.0–5.5 m/s range, while precipitation exceeding 10 mm showed a trend of reducing PTB risk.

The results of this study underscore that, within the arid ecosystem of Xinjiang, dust-related PM_{10} and coal combustion-derived CO are key environmental drivers of PTB transmission. This study not only provides a novel perspective for understanding the environmental mechanisms underlying PTB in arid regions but also demonstrates the effectiveness of the XGBoost-SHAP framework in dissecting complex environmental health effects. From a public health practice standpoint, the findings support the development of targeted regional policies for dust pollution control and clean energy promotion, offering a scientific basis for accelerating the achievement of the goal of ending the tuberculosis epidemic through environmental interventions.

Future research should aim to incorporate individual-level data, socioeconomic variables, and biomarker information, while validating the generalizability of the exposure-response relationships established in this study across different climatic zones and pollution contexts through multicenter collaborations. Such efforts will further advance the field of environmental tuberculosis toward precision and interpretability.

Data availability

The global and Chinese tuberculosis incidence data utilized in this study were obtained from the "Global Tuberculosis Report 2023" published by the WHO and the global tuberculosis database. The dataset for the Xinjiang region can be made available upon reasonable request to the corresponding author.

Received: 15 July 2025; Accepted: 24 September 2025
Published online: 30 October 2025

References

1. Achkar, J. M. Prospects and challenges of a new live tuberculosis vaccine. *Lancet Respir. Med.* **7**, 723–725. https://doi.org/10.1016/S2213-2600(19)30277-2 (2019).
2. Fogel, N. & Tuberculosis A disease without boundaries. *Tuberculosis* **95**, 527–531. https://doi.org/10.1016/j.tube.2015.05.017 (2015).
3. Godoy, S. et al. Risk of tuberculosis among pulmonary tuberculosis contacts: the importance of time of exposure to index cases. *Ann. Epidemiol.* **91**, 12–17. https://doi.org/10.1016/j.annepidem.2024.01.004 (2024).
4. Goletti, D., Meintjes, G., Andrade, B. B., Zumla, A. & Lee, S. S. Insights from the 2024 WHO global tuberculosis Report - More comprehensive Action, Innovation, and investments required for achieving WHO end TB goals. *Int. J. Infect. Dis.* **150** https://doi.org/10.1016/j.ijid.2024.107325 (2025).
5. Feng, Q. et al. Roadmap for ending TB in China by 2035: the challenges and strategies. *Biosci. Trends*. https://doi.org/10.5582/bst.2023.01325 (2024).
6. Wang, Y. S., Zhu, W. L., Li, T., Chen, W. & Wang, W. B. Changes in newly notified cases and control of tuberculosis in china: Time-series analysis of surveillance data. *Infect. Dis. POVERTY*. **10** https://doi.org/10.1186/s40249-021-00806-7 (2021).
7. Wang, Q. et al. Association of air pollutants and meteorological factors with tuberculosis: A National multicenter ecological study in China. *Int. J. Biometeorol.* https://doi.org/10.1007/s00484-023-02524-1 (2023).
8. de Glanville, W. A. et al. General contextual effects on neglected tropical disease risk in rural Kenya. *PLoS Negl. Trop. Dis.* **12** https://doi.org/10.1371/journal.pntd.0007016 (2018).
9. Martinez, L. et al. Detection, survival and infectious potential of Mycobacterium tuberculosis in the environment: A review of the evidence and epidemiological implications. *Eur. Respir. J.* **53** https://doi.org/10.1183/13993003.02302-2018 (2019).
10. Wang, X. Q. et al. Short-term effect of particulate air pollutant on the risk of tuberculosis outpatient visits: A multicity ecological study in Anhui, China. *Atmos. Environ.* **280** https://doi.org/10.1016/j.atmosenv.2022.119129 (2022).
11. Popovic, I. et al. A systematic literature review and critical appraisal of epidemiological studies on outdoor air pollution and tuberculosis outcomes. *Environ. Res.* **170**, 33–45. https://doi.org/10.1016/j.envres.2018.12.011 (2019).
12. Ibironke, O. et al. Urban air pollution particulates suppress human T-Cell responses to Mycobacterium tuberculosis. *Int. J. Environ. Res. Public Health*. **16** https://doi.org/10.3390/ijerph16214112 (2019).
13. Sarkar, S. et al. Season and size of urban particulate matter differentially affect cytotoxicity and human immune responses to Mycobacterium tuberculosis. *PLOS ONE*. **14** https://doi.org/10.1371/journal.pone.0219122 (2019).
14. Turner, R. D. et al. Tuberculosis infectiousness and host susceptibility. *J. Infect. Dis.* **216**, 636–S643. https://doi.org/10.1093/infdis/jix361 (2017).
15. Peters, J. S. et al. Advances in the understanding of Mycobacterium tuberculosis transmission in HIV-endemic settings. *LANCET Infect. Dis.* **19**, E65–E76. https://doi.org/10.1016/S1473-3099(18)30477-8 (2019).
16. Helmy, H., Kamaluddin, M. T. & Iskandar, I. Suheryanto. Investigating Spatial patterns of pulmonary tuberculosis and main related factors in Bandar Lampung, Indonesia using geographically weighted Poisson regression. *Trop. Med. Infect. DISEASE*. **7** https://doi.org/10.3390/tropicalmed7090212 (2022).
17. Nie, Y. et al. Effects and interaction of meteorological factors on pulmonary tuberculosis in Urumqi, China, 2013–2019. *Front. PUBLIC. HEALTH*. **10** https://doi.org/10.3389/fpubh.2022.951578 (2022).
18. Khaliq, A., Batool, S. A. & Chaudhry, M. N. Seasonality and trend analysis of tuberculosis in Lahore, Pakistan from 2006 to 2013. *J. Epidemiol. Global Health*. **5**, 397–403. https://doi.org/10.1016/j.jegh.2015.07.007 (2015).
19. Li, Z., Zhang, L. & Liu, Y. Analysis of the epidemiological trends of tuberculosis in China from 2000 to 2021 based on the joinpoint regression model. *BMC Infect. Dis.* **24** https://doi.org/10.1186/s12879-024-10126-4 (2024).
20. Luo, D. et al. Spatial spillover effect of environmental factors on the tuberculosis occurrence among the elderly: A surveillance analysis for nearly a dozen years in Eastern China. *BMC Public. Health*. **24** https://doi.org/10.1186/s12889-024-17644-5 (2024).
21. Li, X. et al. Tuberculosis infection in rural labor migrants in Shenzhen, China: Emerging challenge to tuberculosis control during urbanization. *Sci. Rep.* **7** https://doi.org/10.1038/s41598-017-04788-1 (2017).
22. Lin, D. et al. A genome epidemiological study of Mycobacterium tuberculosis in subpopulations with high and low incidence rate in Guangxi, South China. *BMC Infect. Dis.* **21** https://doi.org/10.1186/s12879-021-06385-0 (2021).
23. Vowels, M. J. Trying to outrun causality with machine learning: Limitations of model explainability techniques for exploratory research. *Psychol. Methods*. https://doi.org/10.1037/met0000699 (2024).
24. Li, Z., Zhou, H., Xu, Z. & Ma, Q. Machine learning and public health policy evaluation: Research dynamics and prospects for challenges. *Front. Public Health*. **13** https://doi.org/10.3389/fpubh.2025.1502599 (2025).
25. Jerome, H. F. Greedy function approximation: A gradient boosting machine. *Annals Stat.* **29**, 1189–1232. https://doi.org/10.1214/aos/1013203451 (2001).
26. Natekin, A. & Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobotics.* **7** https://doi.org/10.3389/fnbot.2013.00021 (2013).
27. Hajihosseinlou, M., Maghsoudi, A. & Ghezelbash, R. A novel scheme for mapping of MVT-Type Pb-Zn prospectivity: LightGBM, a highly efficient gradient boosting decision tree machine learning algorithm. *Nat. Resour. Res.* **32**, 2417–2438. https://doi.org/10.1007/s11053-023-10249-6 (2023).
28. Tianqi, C., Carlos, G. & XGBoost: A scalable tree boosting system. *arXiv - CS - Machine Learning*. arxiv-1603.02754 (2016).
29. Lundberg, S. M. et al. From local explanations to global Understanding with explainable AI for trees. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-019-0138-9 (2020).
30. Yi, F. et al. XGBoost-SHAP-based interpretable diagnostic framework for alzheimer's disease. *BMC Med. Inf. Decis. Mak.* https://doi.org/10.1186/s12911-023-02238-9 (2023).
31. Deng, W. et al. Genotypic diversity of Mycobacterium tuberculosis isolates and its association with drug-resistance status in Xinjiang, China. *Tuberculosis* **128** https://doi.org/10.1016/j.tube.2021.102063 (2021).
32. Long, Q., Guo, L., Jiang, W., Huan, S. & Tang, S. Ending tuberculosis in China: Health system challenges. *Lancet Public Health*. **6**, E948–E953 (2021).
33. Zhang, H. et al. Guiding tuberculosis control through the healthy China initiative 2019–2030. *CHINA CDC Wkly.* **2**, 948–952. https://doi.org/10.46234/ccdcw2020.236 (2020).
34. Liu, C. et al. Associations between ambient fine particulate air pollution and hypertension: A nationwide cross-sectional study in China. *Sci. Total Environ.* **584**, 869–874. https://doi.org/10.1016/j.scitotenv.2017.01.133 (2017).

35. Rivas-Santiago, C. E. et al. Air pollution particulate matter alters antimycobacterial respiratory epithelium innate immunity. *Infect. Immun.* **83**, 2507–2517. https://doi.org/10.1128/IAI.03018-14 (2015).
36. Glass, R. I. & Rosenthal, J. P. International approach to environmental and lung health A perspective from the Fogarty international center. *Ann. Am. Thorac. Soc.* **15**, S109–S113. https://doi.org/10.1513/AnnalsATS.201708-685MG (2018).
37. Raqib, R. et al. Association of household air pollution with cellular and humoral immune responses among women in rural Bangladesh. *Environ. Pollut.* **299** https://doi.org/10.1016/j.envpol.2022.118892 (2022).
38. Carrasco-Escobar, G., Schwalb, A., Tello-Lizarraga, K., Vega-Guerovich, P. & Ugarte-Gil, C. Spatio-temporal co-occurrence of hotspots of tuberculosis, poverty and air pollution in Lima, Peru. *Infect. Dis. Poverty*. https://doi.org/10.1186/s40249-020-00647-w (2020).
39. Wu, Z. et al. Trends of outdoor air pollution and the impact on premature mortality in the Pearl river delta region of Southern China during 2006–2015. *Sci. Total Environ.* **690**, 248–260. https://doi.org/10.1016/j.scitotenv.2019.06.401 (2019).
40. Xiang, K. et al. Association between ambient air pollution and tuberculosis risk: A systematic review and meta-analysis. *Chemosphere* https://doi.org/10.1016/j.chemosphere.2021.130342 (2021).
41. Barbier, E., Rochelet, M., Gal, L., Boschiroli, M. L. & Hartmann, A. Impact of temperature and soil type on Mycobacterium Bovis survival in the environment. *PLoS ONE.* **12** https://doi.org/10.1371/journal.pone.0176315 (2017).
42. von Schoenberg, P. et al. Aerosol dynamics and dispersion of radioactive particles. *Atmos. Chem. Phys.* **21**, 5173–5193. https://doi.org/10.5194/acp-21-5173-2021 (2021).
43. Chaloupecka, H. et al. Investigating the formation of microplastic aerosols and their dispersion in urban environments: A comparative physical modelling study of aerosol and gas dispersion. *Atmosp. Pollut. Res.* **16** https://doi.org/10.1016/j.apr.2025.102481 (2025).
44. Fan, X. et al. Numerical investigation of the effects of environmental conditions, droplet size, and social distancing on droplet transmission in a street Canyon. *Build. Environ.* **221** https://doi.org/10.1016/j.buildenv.2022.109261 (2022).
45. de Castro Fernandes, F. M., Martins, E. S., Sampaio Pedrosa, A., Nantua Evangelista, M. & D. M. & D. S. Relationship between Climatic factors and air quality with tuberculosis in the federal District, Brazil, 2003–2012. *Brazil. J. Infect. Dis.* **21**, 369–375. https://doi.org/10.1016/j.bjid.2017.03.017 (2017).
46. Marimon, O. et al. An oxygen-sensitive toxin-antitoxin system. *Nat. Commun.* **7** https://doi.org/10.1038/ncomms13634 (2016).
47. He, H. et al. Mineral dust and nox promote the conversion of SO_2 to sulfate in heavy pollution days. *Sci. Rep.* **4** https://doi.org/10.1038/srep06092 (2014).
48. Kota, H. et al. Year-long simulation of gaseous and particulate air pollutants in India. *Atmos. Environ.* **180**, 244–255. https://doi.org/10.1016/j.atmosenv.2018.03.003 (2018).
49. Jassal, M. S., Bakman, I. & Jones, B. Correlation of ambient pollution levels and heavily-trafficked roadway proximity on the prevalence of smear-positive tuberculosis. *Public Health.* **127**, 268–274. https://doi.org/10.1016/j.puhe.2012.12.030 (2013).
50. Sarkar, S. et al. Suppression of the NF-κB pathway by diesel exhaust particles impairs human antimycobacterial immunity. *J. Immunol.* **188**, 2778–2793. https://doi.org/10.4049/jimmunol.1101380 (2012).
51. Wu, Z. et al. Trends of outdoor air pollution and the impact on premature mortality in the Pearl river delta region of Southern China during 2006–2015. *Sci. Total Environ.* https://doi.org/10.1016/j.scitotenv.2019.06.401 (2019).
52. Zhang, Y. et al. Numerical investigations of reactive pollutant dispersion and personal exposure in 3D urban-like models. *Build. Environ.* https://doi.org/10.1016/j.buildenv.2019.106569 (2020).
53. Kaspersen, K. A. et al. Exposure to air pollution and risk of respiratory tract infections in the adult Danish population-a nationwide study. *Clin. Microbiol. Infect.* **30**, 122–129. https://doi.org/10.1016/j.cmi.2023.10.013 (2024).
54. Liu, Y. et al. Ambient air pollution exposures and newly diagnosed pulmonary tuberculosis in Jinan, China: A time series study. *Sci. Rep.* **8** https://doi.org/10.1038/s41598-018-35411-6 (2018).
55. Wang, X. Q. et al. Short-term effect of ambient air pollutant change on the risk of tuberculosis outpatient visits: A time-series study in Fuyang, China. *Environ. Sci. Pollut. Res.* **29**, 30656–30672. https://doi.org/10.1007/s11356-021-17323-7 (2022).
56. Chen, C. et al. High latent TB infection rate and associated risk factors in the Eastern China of low TB incidence. *PLoS ONE.* https://doi.org/10.1371/journal.pone.0141511 (2015).
57. You, S., Tong, Y. W., Neoh, K. G., Dai, Y. & Wang, C. H. On the association between outdoor PM2.5 concentration and the seasonality of tuberculosis for Beijing and Hong Kong. *Environ. Pollut.* **218**, 1170–1179. https://doi.org/10.1016/j.envpol.2016.08.071 (2016).
58. Xiong, Y. et al. Association of daily exposure to air pollutants with the risk of tuberculosis in Xuhui district of Shanghai, China. *Int. J. Environ. Res. Public Health*. **19** https://doi.org/10.3390/ijerph19106085 (2022).

Acknowledgements

This study would like to express our sincere gratitude to the WHO, the global tuberculosis database, as well as the staff of the Urumqi Center for Disease Control and Prevention for their invaluable assistance in field investigations, management, and data collection. We are also grateful to the National Natural Science Foundation of China and the Xinjiang Outstanding Young Talents Program for their financial support of this study.

Author contributions

Conception and design of study: Feifei Li, Liping Zhang; Acquisition of data: ChenChen Wang; Analysis and interpretation of data: Feifei Li, Peiyao Zhou; Drafting the manuscript: Feifei Li; Revising the manuscript critically for important intellectual content: Qin Xu, Yanling Zheng and ChenChen Wang. All authors read and approved the final manuscript.

Funding

This work was supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (Grant No. 2022D01C473); National Natural Science Foundation of China (Grant Nos. 72174175, 72163033); and the Xinjiang Outstanding Young Talent Program-Young Innovative Science and Technology Talents (Project Number: 2024TSYCCX0080).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/1

0.1038/s41598-025-21930-6.

Correspondence and requests for materials should be addressed to L.Z.