ACCSCIENCE PUBLISHING

Asian Journal of Water, Environment and Pollution

ORIGINAL RESEARCH ARTICLE

Driving factors of chlorinated OPEs pollution in industrial soils: an empirical study based on machine learning and a novel petrochemical source

Chi Zhu[1†], Lu Cao[1†], Enze Pei[2], Xin Huang[3], Shixiang Dai[4], Liang Ding[1], Bo Wang[3], and Changsheng Qu[1*]

[1]Key Laboratory of Environmental Remediation and Ecological Health, Ministry of Industry and Information Technology, Jiangsu Environmental Engineering Technology Co., Ltd., Nanjing, Jiangsu, China

[2]Department of Environmental Engineering, School of Environmental Science and Engineering, Suzhou University of Science and Technology, Suzhou, Jiangsu, China

[3]Department of Environmental Science, School of Environmental and Safety Engineering, Liaoning Petrochemical University, Fushun, Liaoning, China

[4]State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing, Jiangsu, China

[†]*These authors contributed equally to this work.*

***Corresponding author:**
Changsheng Qu
(changshengqu_jsep@163.com)

Citation: Zhu C, Cao L, Pei E, Huang X, Dai S, Ding L, Wang B, Qu C. Driving factors of chlorinated OPEs pollution in industrial soils: an empirical study based on machine learning and a novel petrochemical source. *Asian J Water Environ Pollut.* 2026;23(4):026200139. doi: 10.36922/AJWEP026200139

Received: May 11, 2026

Revised: June 8, 2026

Accepted: June 10, 2026

Published online: July 1, 2026

Abstract

Chlorinated organophosphate esters (Cl-OPEs), a major class of organophosphate ester pollutants, are ubiquitously detected in industrial soils, yet quantitatively apportioning these contaminants remains a critical challenge. Traditional machine learning-based source apportionment approaches are often undermined by the spatial clustering of monitoring data, a phenomenon known as geographical bias. To address this limitation, this study first identified geographical bias through exploratory data diagnostics and subsequently constructed a robust random forest-based machine learning prediction framework that excludes geographical features to eliminate the confounding effect of spatial bias. The optimized model exhibited excellent reliability in predicting soil Cl-OPEs concentrations (test set R^2 = 0.90). Model interpretability analysis via feature importance assessment confirmed E-waste processing (site function category_E-waste processing) as the primary driver of Cl-OPEs pollution. To validate the model and explore overlooked sources, we also conducted an empirical case study of the petrochemical industry. Comparative analysis not only confirmed E-waste processing areas as the primary source but also identified the petrochemical industry as a previously underestimated yet significant source of emissions, with its coil Cl-OPEs levels markedly exceeding those of other traditional industrial areas. Overall, this study proposes a novel, geographical-bias-free machine learning source apportionment method and provides new preliminary empirical evidence supporting pollution control of Cl-OPEs.

Keywords: Organophosphate esters; Machine learning; Soil pollution; Industrial areas

1. Introduction

Organophosphate esters (OPEs), serving as both flame retardants and plasticizers, are widely incorporated into materials such as plastics, rubber, coatings, and electronic

products.[1] Because OPEs are not chemically bonded to the polymer matrix but are merely physically mixed, they are readily released into the environment during product use and disposal.[2] In recent years, numerous toxicological studies have indicated that some OPEs possess neurotoxicity, endocrine-disrupting effects,[3] and potential carcinogenic risks, posing a latent threat to ecosystems and human health.[4] Among various environmental media, industrial soil serves as a critical sink and secondary source for OPEs.[5] On one hand, as a long-term depositional medium for industrial facilities and waste disposal sites, soil can accumulate high concentrations of OPEs[6]; on the other hand, OPEs in soil can persistently impact the surrounding environment through pathways such as atmospheric emissions, runoff transport, and resuspension.[7] Therefore, systematically identifying the spatial patterns and driving mechanisms of OPE pollution in industrial soils is of great scientific significance for conducting precise risk assessments and formulating differentiated control strategies.[8]

Although existing research has pointed out that different industrial types can lead to variations in soil OPE concentrations, related studies still face several challenges.[9] Firstly, the distribution of OPEs is co-influenced by multi-scale factors, and traditional statistical methods have limitations in capturing the non-linear associations among these multiple factors.[10] Machine learning (ML) technology offers a powerful tool for this, but its application is often constrained by the data's structural limitations.[11] In particular, environmental monitoring data frequently exhibit spatial clustering (i.e., excessive sampling in a few hot-spot areas), which leads to a critical methodological challenge: Does the model learn the universal causes of pollution, or does it merely learn to recognize specific geographical regions?[12,13] This issue of "geographical bias" can cause the model to fail completely when predicting in new regions and is a core obstacle that must be overcome in current ML-based source apportionment.[14]

To circumvent the aforementioned geographical bias, this study proposes a novel ML source apportionment method. The method first identifies and quantifies spatial clustering in the data through data diagnostics (e.g., violin plots). Based on this, the study deliberately removes geographical coordinate features to construct a more robust and generalizable random forest (RF) model. This forces the model to focus on more universal physicochemical driving factors, such as industrial type and operational mode. The resulting model demonstrated high reliability in predicting the concentrations of chlorinated organophosphate esters (Cl-OPEs; test set $R^2 = 0.90$).

The core scientific questions of this study are: After excluding geographical bias, what are the true driving factors of Cl-OPEs pollution in industrial soils? And, is the currently known list of pollution sources complete? To answer these questions, this study aims to: (i) Utilize the highly reliable ML model (RF) and interpretability analysis (feature importance plot) to identify the primary driving factors of Cl-OPEs pollution; (ii) by introducing a new empirical case study from the petrochemical industry, validate the model's conclusions while simultaneously exploring previously underestimated sources of Cl-OPEs pollution. The selected study areas encompass a broad geographical gradient spanning tropical to temperate zones across East and West Asia, and incorporate highly diverse industrial typologies, including historic E-waste dismantling hubs, comprehensive manufacturing zones, and a major petrochemical complex. This multi-regional and multi-typological diversity ensures that the integrated dataset captures a comprehensive spectrum of industrial emission profiles, thereby establishing a highly representative framework for identifying universal, de-geographicalized driving mechanisms of soil Cl-OPE contamination. The findings of this study can provide a key scientific basis for differentiated pollution control of Cl-OPEs and for updating their pollution-source inventory.

2. Materials and methods

2.1. Study area and data sources

The data for this study were derived from a systematic review and integration of published academic literature concerning multiple representative industrial regions. These regions cover a multi-waste comprehensive disposal area in North China (Jinghai District, Tianjin), an industrial hub in Southwest China (Chengdu City), the renowned E-waste dismantling center in South China (Guiyu Town, Guangdong), a comprehensive industrial city in East China (Jinan City, Shandong), and a typical highly industrialized representative in West Asia (Bursa, Turkey) (Table 1). To empirically evaluate the applicability of the model's source apportionment insights and explore potential missing emission sources, this study additionally introduced a new empirical case study: Soil samples collected in 2024 from a petrochemical industrial park in Yancheng, Jiangsu, China. This data will be used for empirical comparison in subsequent analyses. The concentration data utilized for this petrochemical empirical snapshot were acquired from an unpublished internal collaborative monitoring campaign that screened emerging contaminants across prominent industrial sectors in Jiangsu Province, China. Specifically, a cluster of five high-precision sampling points was established within a representative petrochemical functional zone in the Yancheng region of Northern

Table 1. Key characteristics of the five study areas

Study area	Geographic location	Area characteristics	Main industrial/pollution source types	Reference
Jinghai, Tianjin	North China, Ziya Environmental Industrial Park	A comprehensive disposal site for various wastes, including E-waste	Recycling and dismantling of E-waste, waste plastics, old home appliances, and automobiles	15
Chengdu, Southwest China	Southwest China	The most important central city in Western China, a high-tech base and transportation hub	Covers various industrial parks such as automotive, pharmaceutical, furniture, shoemaking, and logistics	16
Guiyu, South China	South China	Historically famous E-waste dismantling center, later integrated into an industrial park	E-waste dismantling and recycling	17
Jinan City	East China, capital of Shandong Province	Typical second-tier city in China, population over eight million	Comprehensive industry; sampling sites near textile, E-waste dismantling, and plastic waste recycling plants	18
Bursa, Turkey	Turkey	Turkey's fourth-largest city, a highly industrialized city	Focus on textile, automotive, spare parts, machinery, and metal industries	19

Jiangsu (latitude range: 34.35354602°N–34.35440556°N, longitude range: 119.7767976°E–119.7782297°E). Core soil samples were systematically extracted from a vertical profile spanning 0–3 m during the winter season (January to February 2024). Through selective instrumental analysis, a total of eight specific OPE congeners were cross-identified and quantified, which perfectly bifurcated into two Cl-OPEs (tris(3-chloropropyl) phosphate and tris(2-chloroethyl) phosphate), two alkyl OPEs (dominated by tributyl phosphate), and four aryl OPEs. The compiled mean concentration of total Cl-OPEs reached 200 $ng \cdot g^{-1}$ (ppb), embedded within a total baseline OPE burden of 798 $ng \cdot g^{-1}$ and an analytical relative error of 41%. The six core methodological design attributes and key pollution parameters governing this empirical case study are systematically summarized in Table 2.

The core of the integrated dataset includes four categories of response variables: Total OPEs concentration and the concentrations of its three major chemical subclasses (chlorinated, alkyl, and aryl OPEs), with all units standardized to $ng \cdot g^{-1}$ dry weight. To investigate their driving forces, we also extracted multiple predictive factors, including the longitude and latitude of the sampling points, regional type, main industrial activities, sampling time and season, and specific operational modes. To ensure the accuracy and comparability of data integration across literature, this study implemented a rigorous quality assurance/quality control (QA/QC) process. This process first ensured indicator consistency by unifying target compounds, meaning only OPE species reported in all five source papers were retained for subsequent analysis. Subsequently, we converted all concentration data to a uniform unit ($ng \cdot g^{-1}$ dry weight) and confirmed that each study employed mature, reliable quantitative analysis techniques (e.g., gas chromatography [GC], mass spectrometry [MS]) and primarily used external standard methods for quantification, providing a methodological basis for data comparability. To address potential data heterogeneity and rigorously validate inter-study comparability, we systematically cross-examined the extraction protocols, clean-up methods, and QA/QC parameters across the five literature sources, as synthesized in Table 3. Methodologically, the extraction techniques encompassed vortex-ultrasonic extraction, solvent extraction, ultrasonic extraction, orbital shaking, and Soxhlet extraction, while sample clean-up was executed via Florisil solid-phase extraction (SPE), alumina/silica column chromatography, or Oasis HLB SPE. Analytical detection relied on robust instrument configurations including GC–MS, GC–tandem mass spectrometry (MS/MS), GC–electron ionization–MS, and high-performance liquid chromatography–MS/MS. Regarding QC, the reported recovery ranges spanned 73%–120%, 70.5%–135%, 84%–121%, and 65%–108% for surrogates (>80% for targets), with limits of detection established at 0.010–0.080 ng/g where explicitly specified in the source texts. This granular technical profile confirms that, despite inherent operational variations, all independent datasets strictly conformed to acceptable standards for environmental

Table 2. Core methodological attributes and quantified pollution metrics of the newly introduced empirical case study in the Yancheng petrochemical industrial zone

No.	Core methodological attribute	Quantified empirical specifications and value
1	Sampling season and time	Winter (January 2024 – February 2024)
2	Detected organophosphate ester (OPE) species	8 congeners total (chlorinated: 2; alkyl: 2; aryl: 4)
3	Total chlorinated OPEs concentration	200 $ng \cdot g^{-1}$ (ppb; analytical relative error: 41%)
4	Site function category	Petrochemical industrial area
5	Geographic coordinates	Lat: 34.35354602°N–34.35440556°N, Lon: 119.7767976°E–119.7782297°E
6	Sampling soil depth	0–3 m (vertical soil core profile)
7	Total alkyl OPEs concentration	505 $ng \cdot g^{-1}$ (ppb)
8	Total aryl OPEs concentration	93 $ng \cdot g^{-1}$ (ppb)
9	Total OPEs burden	798 $ng \cdot g^{-1}$ (ppb)

Table 3. Methodological and quality assurance/quality control parameter synthesis of the five integrated literature sources

Reference	Extraction method	Clean-up	Instrumental analysis	Recovery range	Limits of detection range
15	Vortex-ultrasonic extraction	Florisil SPE	GC–MS	Not reported	Not reported (<MDL)
16	Solvent extraction	Alumina/silica column chromatography	GC–MS	73%–120%	Not reported
17	Ultrasonic extraction	Florisil SPE	GC–MS/MS	70.5%–135%	Not reported
18	Orbital shaking	Oasis HLB SPE	HPLC–MS/MS	84%–121%	0.010–0.080 ng/g
19	Soxhlet extraction	Florisil SPE	GC–EI–MS	65%–108% (surrogates) >80% (targets)	Not reported

Abbreviations: EI: Electron ionization; GC: Gas chromatography; HPLC: High-performance liquid chromatography; MDL: Method detection limit; MS: Mass spectrometry; MS/MS: Tandem mass spectrometry; SPE: Solid-phase extraction.

trace analysis. Finally, after checking the data for outliers, we eliminated two samples with obvious entry errors to ensure the reliability of the final dataset.

Following this processing, the final dataset comprised 141 sampling points and 13 variables. Although this represents a typical small-sample scenario compared with large-scale ML training sets in other fields, it accurately reflects the dual challenges of collection cost and geographical distribution in cross-regional environmental monitoring. Therefore, the dataset covers a broad geographical gradient from tropical to temperate zones, spanning West Asia to East Asia (approximately 23°N–40°N latitude, approximately 29°E–117°EB longitude). Its rigorous QC and good spatial representativeness jointly provide a high-quality data foundation for the subsequent exploration of robust ML modeling under small-sample conditions.[20]

2.2. Data preprocessing and feature engineering

Before formal modeling, we executed a systematic preprocessing and feature engineering workflow on the integrated dataset to ensure data quality and optimize model performance.[21] In this study, we decided to focus the analysis on Cl-OPEs. This decision was based on their status as one of the main categories of OPEs, possessing significant environmental persistence and potential biotoxicity. Consequently, the total concentration of Cl-OPEs was selected as the sole output target for all subsequent

models. The core methodological challenge of this study was to circumvent potential geographical bias. Traditional ML source apportionment is susceptible to "sampling clustering" of monitoring data.[22] If the model learns geographical shortcuts (e.g., identifying specific latitudes) rather than universal driving factors, its generalization ability will be severely compromised.[23,24] Therefore, to build a more robust model focused on physicochemical drivers, this study made a critical decision during preprocessing: The longitude and latitude features were explicitly removed from the input feature set. The input features we ultimately used for modeling were restricted to the following three categories of more universal variables: Site function category (categorical), season (categorical), and OPE types (numerical, representing the number of OPE species detected). Although the exclusion of explicit geographical coordinates introduces a methodological trade-off that may abstract away macro-scale environmental variations—such as regional climate gradients or long-range atmospheric transport pathways—this strategy is critical to mitigate spatial confounding. Given the severe spatial clustering demonstrated in our exploratory data diagnostics (Section 3.1), incorporating raw latitude and longitude data would inevitably cause tree-based ensemble algorithms to learn localized spatial autocorrelation shortcuts rather than universal physicochemical relationships. To preserve essential environmental complexity without introducing geographical bias, the model relies on functional proxy variables: "Site function category" implicitly captures the density of localized industrial layouts and point-source emission footprints, while "season" serves as a temporal-climatic indicator accounting for temperature-dependent atmospheric partitioning and precipitation-driven deposition patterns. This framework successfully isolates the predictive architecture from spatial artifacts while preserving the core environmental mechanisms governing the distribution of soil Cl-OPEs. Subsequently, we performed a systematic preprocessing workflow. We first removed samples with more than 10% missing data, then imputed the remaining missing values using the median (for numerical types) and mode (for categorical types). For discrete features such as site function category and season, one-hot encoding was applied to convert them into binary dummy variables that the model can interpret. OPE types, as the only numerical input feature, were used directly. Next, this cleaned and feature-engineered dataset (n = 141) was split into a training set (113 samples) and a test set (28 samples) at an 8:2 ratio using stratified sampling. To enhance the robustness of model evaluation, this study employed a five-fold cross-validation strategy during model development.

2.3. Machine learning model development and performance assessment

Machine learning, as a powerful data-driven technology, can automatically learn patterns from complex data and make predictions.[25] In this study, to accurately predict the sole target variable—total Cl-OPEs concentration in industrial soils—we employed and evaluated four representative ML algorithms: Regression chain (RC),[26,27] support vector regression (SVR),[28] K-nearest neighbors (KNN),[29] and RF.[30] To systematically benchmark these algorithms while safeguarding against data leakage during validation, an integrated computational architecture was established. The comprehensive ML workflow—incorporating stratified data partitioning, sequential pipeline preprocessing, and grid-search hyperparameter optimization under a five-fold cross-validation framework—is schematically illustrated in Figure 1. The fundamental mathematical formulations, operational principles, and theoretical limits of each of the four individual regression algorithms are detailed in Appendix A1 to maintain focus on the holistic modeling implementation. This study will compare the performance of these four models in predicting Cl-OPEs concentration to select the optimal model.

The code for this study was written in Python 3.9, and the four ML models were implemented in a Jupyter Notebook (version 6.4.5) environment. Data processing and visualization relied mainly on the Numpy (1.20.3), Pandas (1.3.4), and Matplotlib (3.4.3) libraries, while model construction was based on the Scikit-learn (1.0.2) framework. Model performance was evaluated primarily on the test set using three metrics: The coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE). Higher R^2 and lower MAE and RMSE values indicate better model performance, and the optimal model was determined by a comprehensive assessment of these three indicators. R^2, MAE, and RMSE are shown in **Equations 1–3**:

$$R^2 = 1 - \sum_{i=1}^{n} (y_i - x_i)^2 / \sum_{i=1}^{n} (y_i - y)^2 \quad (1)$$

$$RMSE = \sqrt{1/n \times \sum_{i=1}^{n} \left(y_i - x_i\right)^2} \quad (2)$$

$$MAE = 1/n \times \sum_{i=1}^{n} \left|y_i - x_i\right| \quad (3)$$

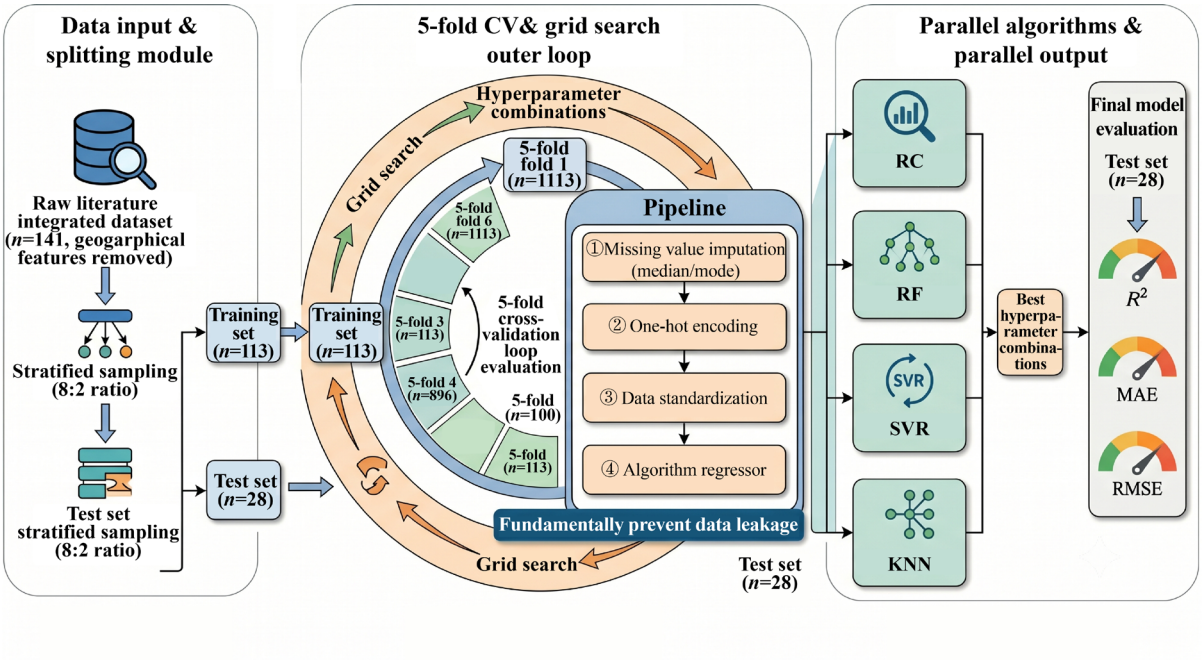

Figure 1. Conceptual topology of the integrated machine learning pipeline, demonstrating the parallel execution of the four candidate regressors coupled with sequential preprocessing, grid-search hyperparameter tuning, and five-fold cross-validation
Abbreviations: CV: Cross-validation; KNN: K-nearest neighbors; MAE: Mean absolute error; R2: Coefficient of determination; RC: Regression chain; RF: Random forest; RMSE: Root mean square error; SVR: Support vector regression.

where y_i represents the actual observed value, x_i indicates the predicted values, y signifies the mean of the actual values, and n is the total number of data points in the test set.

Furthermore, to mitigate the statistical limitations of small-sample training frameworks (n = 141) and to safeguard against the absence of completely independent external validation datasets, an ensemble-based uncertainty quantification framework was implemented for the optimal model. Because the RF architecture operates as an ensemble of independent decision trees (B = 100 in this study), the model's structural confidence and prediction stability can be mathematically characterized by assessing the variance of outputs across individual learners. For any given input feature vector x, the prediction uncertainty $\sigma^2(x)$ is quantified according to **Equation 4**:

$$\sigma^2(x) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\left(f_b(x) - \hat{f}(x)\right)\right)^2 \quad (4)$$

where B represents the total number of decision trees constructed within the forest, $f_b(x)$ denotes the independent prediction generated by the b-th individual decision tree, and $\hat{f}(x)$ signifies the final ensemble mean prediction reported as the model's ultimate output. This variance metric serves as an internal robustness diagnostic, capturing the epistemic uncertainty of the ML framework across different pollutant concentration brackets, thereby conceptually reinforcing the homoscedasticity evaluations conducted within our residual diagnostics.

Hyperparameter selection is a critical step that determines the upper limit of an ML model's performance, directly impacting its prediction accuracy and generalization ability.[31] To achieve systematic, efficient, and rigorous parameter optimization, ensure the standardization of data preprocessing, and fundamentally prevent data leakage during cross-validation, this study encapsulated the data preprocessing steps and the model itself into a unified ML pipeline. Specifically, this pipeline sequentially connected four steps: Missing value imputation, one-hot encoding, data standardization, and the final model (e.g., RF, SVR). This not only simplified the workflow but, more importantly, prevented data leakage during cross-validation (i.e., preventing information from the validation set from influencing the preprocessing of the training set, such as the calculation of standardization parameters).[32] Subsequently, we employed the classic grid search method, combined with five-fold cross-validation, to perform integrated hyperparameter tuning on this

pipeline.[32] Grid search systematically traverses all possible combinations within the parameter grid predefined for the model.[33] Supported by cross-validation, the performance of each parameter combination is measured by its average validation score across the five data subsets. This strategy of encapsulating preprocessing and model training within a pipeline for tuning ensures that the optimal solution within the specified parameter space is found. It fundamentally guarantees that every step of data preprocessing is tailored only to the current training data, thereby maximizing fairness in model evaluation and the generalization ability of the selected hyperparameters on unseen data.[34]

2.4. Model interpretability analysis

To uncover the key drivers of the ML model's predictions, this study employed the permutation importance (PI) technique for model interpretation. The core principle of PI is to measure a feature's importance by observing the decrease in model performance after randomly shuffling the values of that single feature variable in the dataset.[35] If shuffling a variable leads to a significant drop in model performance, it indicates that its contribution is high. The main advantages of this method are its non-parametric nature, its applicability to any type of ML model, and its ability to intuitively quantify the global average contribution of variables.

Mathematically, the PI $I(X_j)$ for a specific feature X_j is defined according to **Equation 5**:

$$I\left(X_j\right) = L\left(y, M\left(X^{perm\_j}\right)\right) - L\left(y, M\left(X\right)\right) \tag{5}$$

where M represents the trained ML model, L denotes the evaluation loss function, y signifies the true vector of soil Cl-OPE concentrations, X is the baseline validation matrix, and X_j represents the matrix where the values of the target feature X_j have been randomly shuffled. While **Equation 5** rigorously quantifies the predictive driving capacity of individual features based on model error inflation, it inherently captures exploratory statistical associations rather than empirical mechanistic causality. This mathematical boundary is particularly vital when categorical features dominate the input space and are one-hot encoded; in such configurations, the permutation of individual dummy variables evaluates the localized split vulnerability within the tree architecture rather than defining immutable, structured causal pathways.

2.5. Empirical case study comparative analysis method

To achieve the second core objective of this study—validating the model's findings and exploring potential new pollution sources—this study employed qualitative and quantitative comparative analyses. This method aims to compare pollution levels between this study's empirical case study data (i.e., the Yancheng, Jiangsu petrochemical site mentioned in Section 2.1) and the literature dataset (n = 141). Specifically, we will use a box plot to visualize the distribution of total Cl-OPE concentrations (including the median and quartiles) across different industrial site types (e.g., E-waste processing, manufacturing) within the literature dataset. Subsequently, we will mark the measured concentration value from the petrochemical site as a single data point on the same chart, thereby allowing for a direct visual assessment of the petrochemical industry's pollution level relative to traditional industrial sources. To further upgrade this comparative analysis from a qualitative visual assessment to a rigorous quantitative metric, the relative enrichment factor (REF) is introduced to mathematically define the deviation degree of the novel petrochemical source against the model-established traditional industrial baseline, as expressed in **Equation 6**:

$$REF = \frac{C_{petrochemical}}{\tilde{C}_{manufacturing}} \tag{6}$$

where $C_{petrochemical}$ represents the measured concentration of soil Cl-OPEs in the newly sampled petrochemical industrial park (ng·g^{-1}), and $\tilde{C}_{manufacturing}$ signifies the median concentration of the "manufacturing" site category calculated from the literature-derived baseline dataset. A REF value significantly greater than 1 indicates that the newly introduced industry's pollution magnitude systematically exceeds that of the traditional industrial baseline identified by the ML model, thereby providing a numerical indicator for identifying emission profiles that are underestimated.

3. Results and discussion

3.1. Data preliminary analysis and geographical bias diagnosis

Building a robust ML model relies on a deep understanding of the underlying data features. We first conducted a correlation analysis on all potential variables in the dataset (Figure 2). The Pearson correlation coefficient matrix comprehensively illustrates the relationships among all variables.[36] Consistent with the research objectives defined in Section 2.2, our analysis focused on the total Cl-OPEs concentration row. We observed a strong positive correlation (r = 0.74) between this target variable and the total concentration of OPEs, indicating that Cl-OPEs are a critical component of overall pollution. Furthermore, the

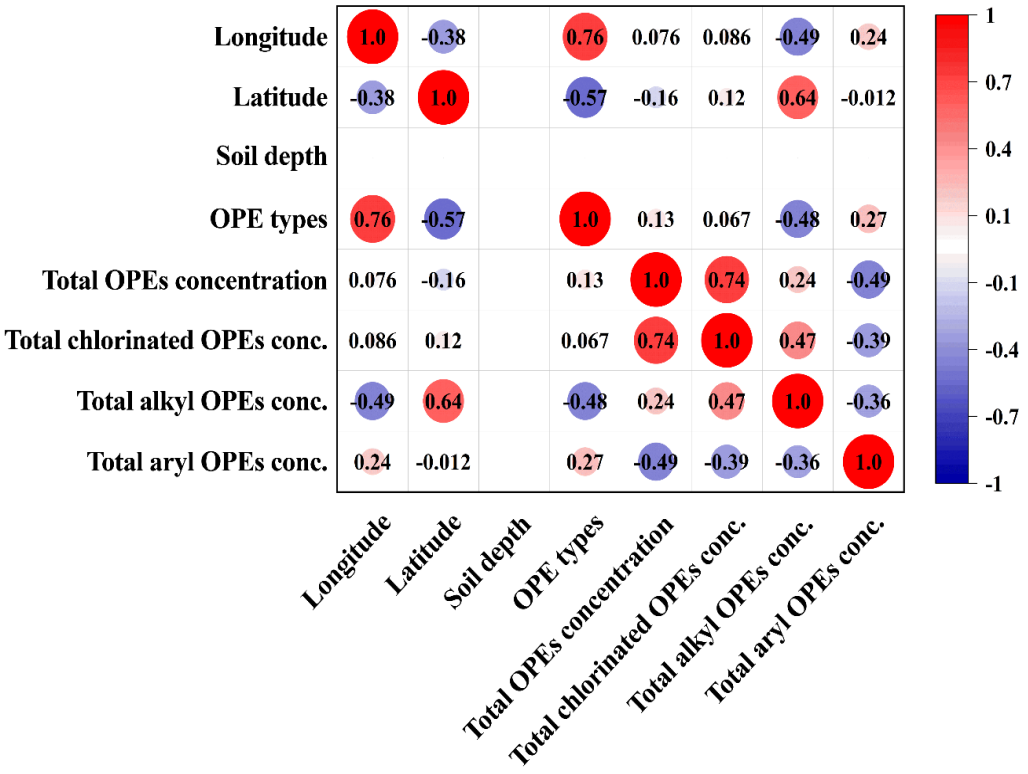

Figure 2. Heatmap of Pearson correlation coefficients between any two features
Abbreviation: conc.: Concentration.

heatmap provided crucial diagnostic information: A high positive correlation ($r = 0.76$) exists between longitude and OPE types, suggesting preliminary spatial clustering.To further diagnose the spatial distribution characteristics of the data, we plotted violin plots for key variables (Figure 3). This figure focuses on the selected target variable (total concentration of Cl-OPEs) alongside geographical and key input features used for diagnosis.[37] As illustrated, the total concentration of Cl-OPEs exhibits a typical right-skewed (positive skew) distribution, where the majority of samples have low concentrations while a few exhibit extremely high values. Most critically, the distribution of the latitude feature reveals a significant bimodal pattern. This finding strongly confirms the presence of severe sampling clustering in our data's geographic space, indicating that samples are primarily concentrated in two distinct regions.

3.2. Model optimization and performance evaluation

To identify the key drivers of the total concentration of Cl-OPEs, this study adopted the de-geographical modeling strategy formulated in Section 2.2. This strategy retained only the site function category, season, and OPE types as input features and focused on the total concentration of Cl-OPEs as the sole target variable. The variables used for modeling and their statistical descriptions are presented in Table 4.

This study aimed to evaluate the predictive capability of four mainstream ML models—RC, RF, SVR, and KNN—for the total concentration of Cl-OPEs in industrial soils. To determine the optimal performance of each model, a grid search was employed to tune key hyperparameters, and five-fold cross-validation was utilized to ensure the robustness and reliability of the evaluation results, thereby identifying the optimal hyperparameter combinations (Table 5).

The performance comparison of the four optimized regression models on the training and test sets is presented in Table 6 and Figure 4. Comprehensive comparative analysis indicates that the RF model demonstrates superior performance, being the only architecture to successfully combine high prediction accuracy with

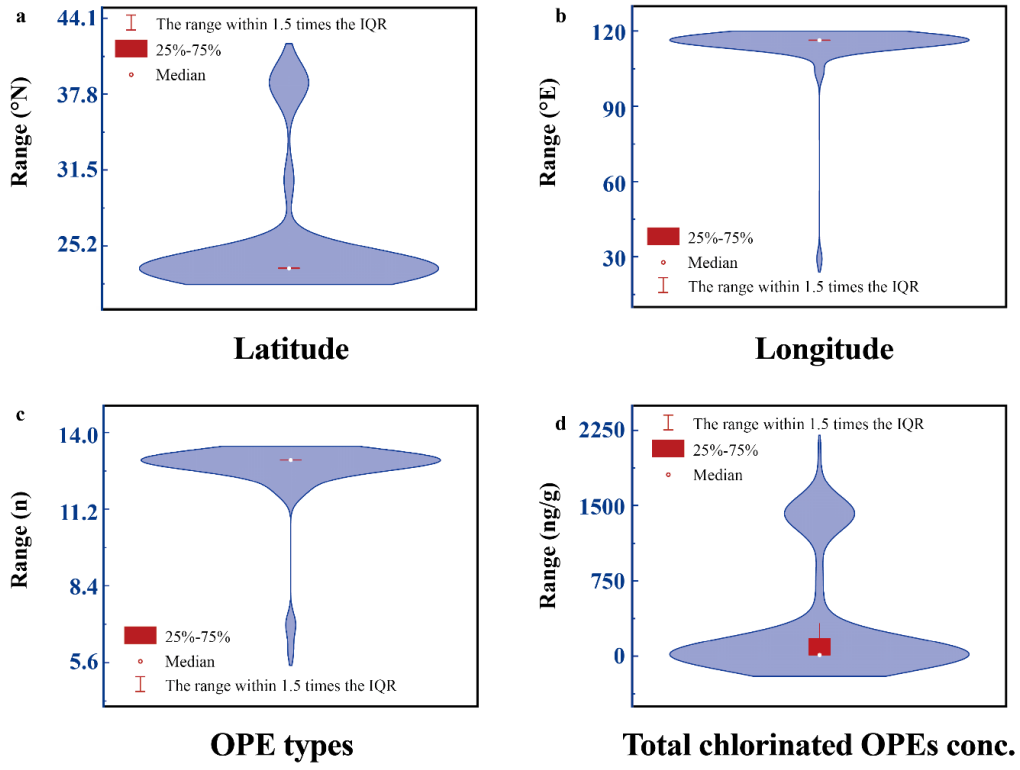

Figure 3. Violin plots of data distribution for four variables
Abbreviations: conc.: Concentration; IQR: Interquartile range.

Table 4. Dataset description analysis of input and output features

Variable	Number	Unit	Data range	Mean	Standard deviation
Input features	–	–	–	–	–
Sampling site	1	–	–	–	–
Sampling area type	2	–	–	–	–
Operational mode	3	–	–	–	–
Sampling time	4	–	–	–	–
Season	5	–	–	–	–
Longitude	6	–	29.10–117.11	114.64	10.64
Latitude	7	–	23.29–40.27	26.64	6.12
Soil depth	8	m	0–0.10	0.10	0
Organophosphate ester (OPE) types	9	–	6–13	12.43	1.53
Output target	–	–	–	–	–
Total OPEs concentration	10	ng·g^{-1}	60.22–19,952.62	1,848.98	3,529.00
Total chlorinated OPEs concentration	11	ng·g^{-1}	10.80–2,107.66	359.60	586.75
Total alkyl OPEs concentration	12	ng·g^{-1}	5.90–305.50	26.50	36.58
Total aryl OPEs concentration	13	ng·g^{-1}	0–442.64	91.07	67.92

Table 5. Hyperparameter optimization ranges and optimal settings for four machine learning models

Model	Hyperparameter	Tuning algorithms	Tuning range	Optimal parameter
Regression chain	Alpha	Grid search	0.01, 0.1, 0.5, 1, 5, 10	10
Random forest	Max_depth	Grid search	3, 5, 7	7
	Min_samples_leaf		3, 5, 7	3
	N_estimators		100, 150	100
Support vector regression	C	Grid search	1, 10, 100	100
	Kernel		Linear, rbf	Linear
K-nearest neighbors	N_neighbors	Grid search	3, 5, 7, 9, 11	3

Table 6. Performance metrics of the four machine learning models on the training and test sets for predicting total chlorinated organophosphate esters concentration

Performance indicators	RC	RF	SVR	KNN
Training set R^2	1.00	0.87	1.00	1.00
Training set RMSE	21.23	211.79	0.10	0.00
Training set MAE	14.33	65.08	<0.10	0.00
Test set R^2	0.81	0.90	0.82	0.84
Test set RMSE	249.31	181.03	243.67	226.09
Test set MAE	169.13	56.11	162.07	71.59

Abbreviations: KNN: K-nearest neighbors; MAE: Mean absolute error; R^2: Coefficient of determination; RC: Regression chain; RF: Random forest; RMSE: Root mean square error; SVR: Support vector regression.

robust generalization capability. As synthesized in the harmonized Table 6, the RF model achieved a high R^2 score of 0.90 on the test set, noticeably outperforming RC (0.81), SVR (0.82), and KNN (0.84). More critically, the RF model exhibited exceptional structural stability; its training set R^2 (0.87) is highly consistent with its test performance, thereby avoiding the overfitting trap. In sharp contrast, the RC, SVR, and KNN models all exhibited a dual failure characterized by perfect or near-perfect training indicators (R^2 = 1.00) paired with substantially degraded testing metrics. This symptomatic gap strongly indicates severe overfitting, which frequently arises when high-capacity non-parametric algorithms are trained on relatively small environmental datasets (n = 141). In these configurations, RC, SVR, and KNN effectively memorized the structural noise and local spatial clusters present within the training partition rather than distilling generalized physicochemical rules. Conversely, the ensemble mechanism of RF—which aggregates predictions from numerous decorrelated decision trees—effectively mitigates single-tree variance and training-set-specific noise, thereby securing superior predictive reliability on unseen test samples.

The predicted-versus-actual scatter plots (Figure 4) provide the most intuitive visual evidence for comparing the goodness of fit and generalization capabilities of the four models. The analysis was based on two core criteria: (i) The deviation of data points (especially the test set) from the 1:1 diagonal (i.e., prediction accuracy); and (ii) the consistency between the training set (blue dots) and the test set (red triangles) (i.e., the degree of overfitting). Both RC and SVR exhibited a dual failure of severe overfitting and low prediction accuracy. Their training sets (blue dots) were perfectly aligned with the 1:1 diagonal (corresponding to R^2 =1.00), whereas their test sets (red triangles) performed poorly, forming high-variance clouds loosely distributed on either side of the diagonal, visually confirming their low test set R^2 values (both 0.51). The KNN model represented a more typical case of overfitting. It also performed perfectly on the training set (blue dots, R^2 = 1.00), but significant deviations were observed on the test

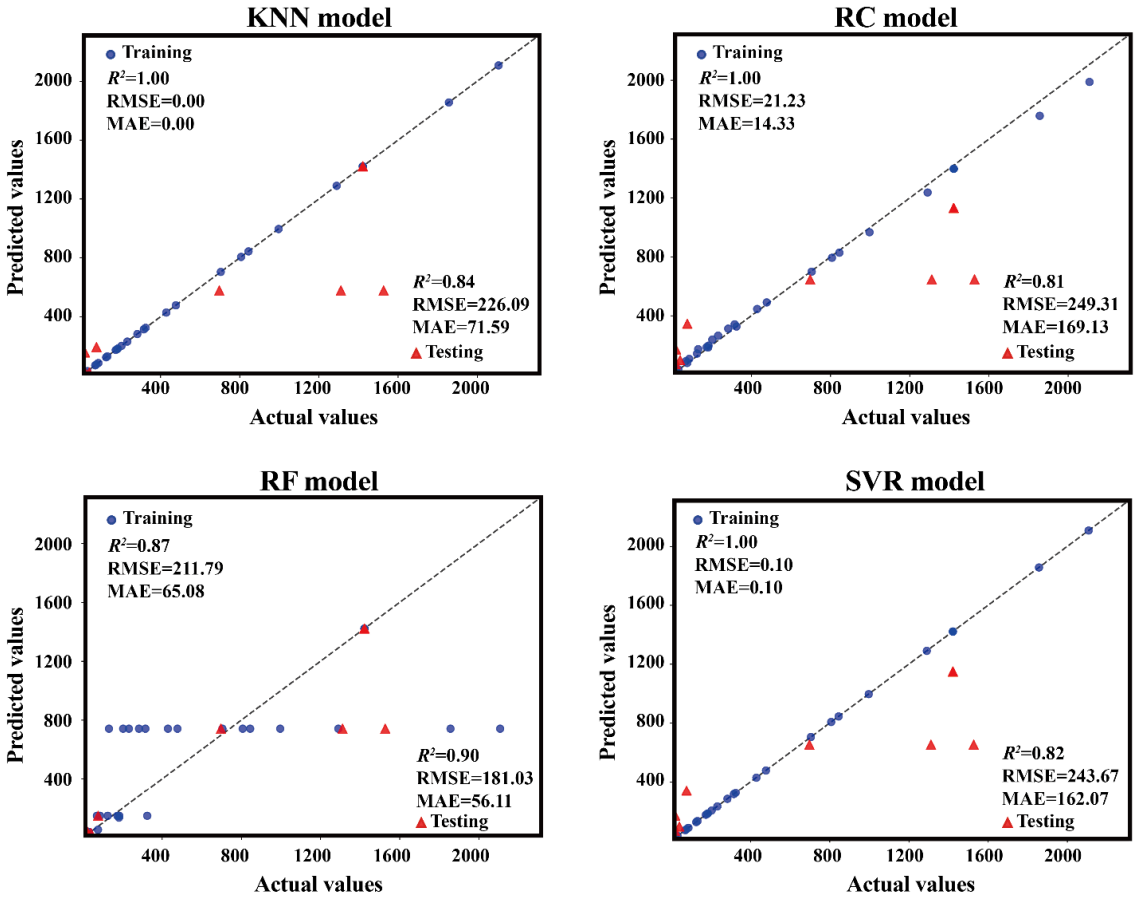

Figure 4. Scatter plots of predicted versus actual values for RC, RF, SVR, and KNN models after optimization
Abbreviations: KNN: K-nearest neighbors; MAE: Mean absolute error; R^2: Coefficient of determination; RC: Regression chain; RF: Random forest; RMSE: Root mean square error; SVR: Support vector regression.

set (red triangles). This large visual discrepancy between training and testing indicated that the KNN model was merely memorizing the training data, limiting its ability to generalize (R^2 = 0.79) and making it less reliable than the RF model.

In sharp contrast, RF was the only model to demonstrate healthy and efficient performance. Its data points (both training and testing) were highly concentrated near the y = x diagonal, indicating high prediction accuracy across all concentration ranges. Crucially, the distribution patterns of the training set (blue dots, R^2 = 0.87) and the test set (red triangles, R^2 = 0.90) were highly consistent. The training set was not perfectly aligned with the diagonal, proving that the model did not memorize the data but successfully learned universal laws. In summary, the comparative analysis in Figure 4 provides compelling evidence that the RC, SVR, and KNN models all suffer from severe overfitting issues. Only the RF model visually satisfied both high accuracy (concentrated points) and strong generalization (training/testing consistency), qualifying it as the optimal model for subsequent analysis.

To further evaluate model performance from a diagnostic perspective, this study introduced residual analysis (Figure 5). Residual plots provide a rigorous diagnosis of the validity of the four models from the perspective of error structure. An ideal regression model should have residuals randomly distributed around the zero line (y = 0) without displaying any specific pattern as predicted values change, i.e., satisfying homoscedasticity. The RC, SVR, and KNN models all exhibited fatal structural defects. First, their residuals on the training set (blue dots) were perfectly zero, again exposing extreme overfitting to training data. More seriously, the residuals from their test sets (red triangles) all exhibited clear trumpet or funnel-shape patterns; that

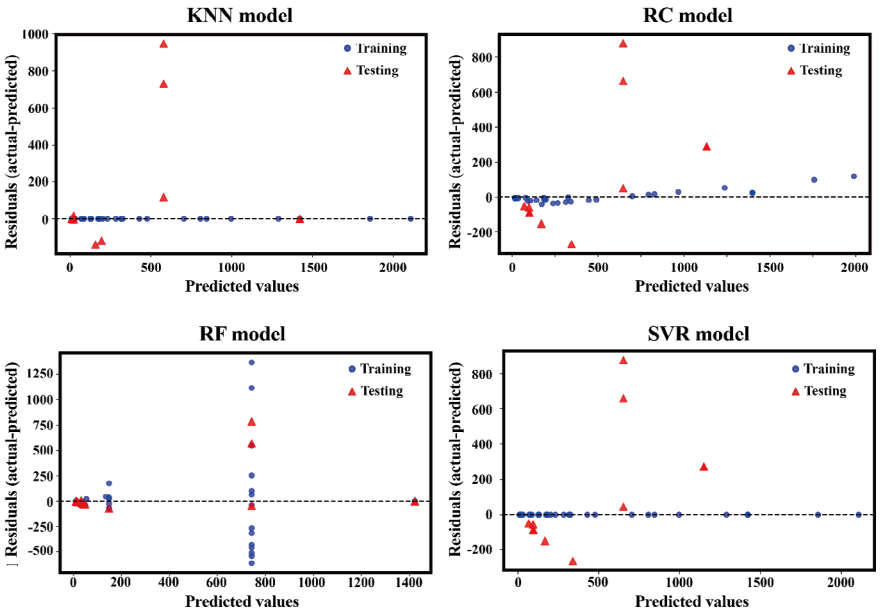

Figure 5. Residual plots of RC, RF, SVR, and KNN models after optimization
Abbreviations: KNN: K-nearest neighbors; RC: Regression chain; RF: Random forest; SVR: Support vector regression.

is, as the predicted concentration increased, the fluctuation range (variance) of the residuals systematically increased. This phenomenon, known as heteroscedasticity, indicates extremely low reliability in predicting high-concentration samples, violating the basic assumptions of regression models.

In contrast, RF was the only model to exhibit a healthy residual distribution. First, its training set residuals did not zero out but were randomly scattered around the zero line, indicating strong generalization (no overfitting). Second, for both training and test sets, the residuals maintained a relatively constant variance across the entire range of predicted values, indicating an unstructured, random distribution. This ideal homoscedasticity confirms that the errors of the RF model are accidental (white noise) rather than due to systematic defects or bias. Consequently, the diagnostic results from the residual plots confirm that RC, SVR, and KNN models are unreliable due to severe overfitting and heteroscedasticity. Only the RF model simultaneously satisfied the ideal characteristics of high prediction accuracy and homoscedastic, unbiased residuals. Therefore, this study ultimately selected RF as the optimal model for subsequent interpretability analysis and apportionment of driving factors.

3.3. Driving factor importance ranking

To further investigate the intrinsic drivers of the model's decision-making, this study employed the PI method to quantitatively evaluate the contribution of each input feature (Figure 6). The analysis reveals the key drivers of the model's decisions and reflects its learning behavior, shaped by the specific dataset used in this study.

As shown in the analysis, the site function category_E-waste processing emerges as the dominant primary driving factor. This implies that during decision-making, more than half of the model's weight relies on the specific information of "whether the site is an E-waste processing area." Closely following is the number of detected OPE species (OPE types), serving as the sole secondary driving factor. This finding holds significant scientific value, indicating that the chemical complexity of the pollution source (i.e., the diversity of detected OPEs) is also a critical predictive indicator. This provides a logical complementary validation with the E-waste processing feature, as E-waste sites typically contain a mixture of plastics and components from diverse sources, resulting in a necessarily higher number of OPE species (e.g., 12–13 types) compared to other single-source industrial sites (e.g., 6–7 types).

The most diagnostically valuable finding from this plot is the immense "importance gap" between the top two features and all remaining features. The importance of the site function category_informal recycling (mixed) is merely 0.0009—more than two orders of magnitude lower than the second-ranked feature. This gap implies that the contributions of all other factors, including other industrial types (e.g., manufacturing, logistics) and

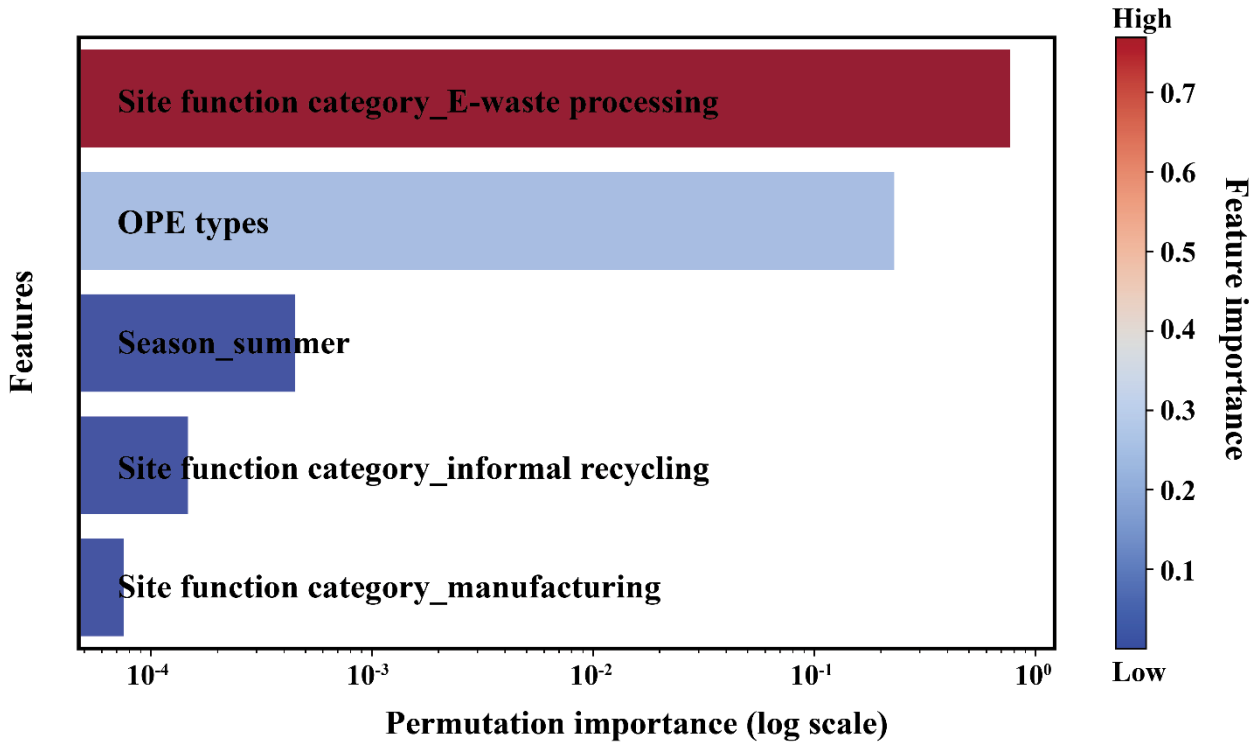

Figure 6. Feature contribution analysis based on permutation importance
Abbreviation: OPE: Organophosphate ester.

sampling seasons (e.g., season_summer, season_winter), are statistically negligible (all < 0.001) for predicting Cl-OPE concentrations.

In summary, the feature importance analysis provides robust quantitative evidence confirming that, after excluding geographical bias, the model successfully learned a highly concentrated relationship: Cl-OPEs pollution in industrial soils is not a generalized industrial issue but is almost exclusively determined by two factors—the type of industrial activity (E-waste processing) and the complexity of the pollution source (OPE types).

Crucially, in alignment with the interpretive limitations detailed in Section 2.4, these established rankings must be interpreted as dominant predictive driving capacities and exploratory statistical associations, rather than definitive mechanistic or causal source apportionments. While "site function category_E-waste processing" strongly governs the RF's decision pathways under small-sample configurations (n = 141), tree-based predictive dependencies do not automatically imply strict physical or chemical causality. Consequently, these findings serve as highly reliable data-driven diagnostic indicators to guide targeted regulatory monitoring, rather than an absolute, deterministic allocation of environmental pollution kinetics.

3.4. Empirical validation and identification of potential pollution sources

Based on the literature dataset, the ML model (R^2 =0.90) successfully identified "E-waste processing" as the primary driving factor for Cl-OPEs pollution. However, a model's capability is limited by the breadth of its training data. Therefore, key scientific questions arise: (i) Can this model finding be validated in real-world data? (ii) Do existing literature data cover all major pollution sources?

To answer these questions, we employed an empirical comparative case study method. Figure 7 visually compares the total concentrations of Cl-OPEs across different industrial site types in the literature dataset (box plots) with our empirical case study (red data point). This figure reveals two key findings. First, it robustly validates the model's conclusion. As shown in Figure 7, the pollution concentration (box median and distribution range) at "E-waste processing" sites is significantly higher than that of "manufacturing" and "informal recycling

(mixed)" categories, which is entirely consistent with the driving factor importance ranking identified by the model. Second, Figure 7 reveals a significant new phenomenon: The petrochemical industry is a previously underestimated important source of pollution. As indicated by the red data point in the figure, the Cl-OPEs concentration at the petrochemical site in our empirical case study is significantly higher than that in traditional non-E-waste industrial areas, such as manufacturing and informal recycling (whose medians are both close to 10 $ng \cdot g^{-1}$).

This empirical finding (high concentration in the petrochemical sector) is of great significance. It suggests that while our model (based on literature data) correctly pinpointed E-waste as the largest known culprit, the existing literature dataset likely failed to fully cover all pollution source types. The empirical data from this study provide preliminary evidence that the petrochemical industry may also be a significant source of Cl-OPEs emission, offering a new perspective for future pollution source apportionment and risk control of Cl-OPEs.

To shift the scientific focus deeper into the chemical nature of the pollution itself and unveil the structural fingerprint of this novel petrochemical source, the chlorinated contribution fraction (F_{Cl}) within the total OPE mass burden is quantitatively formulated using **Equation 7**:

$$F_{Cl} = \frac{C_{Cl-OPEs}}{C_{Total\ OPEs}} \times 100\% \quad (7)$$

where $C_{Cl\text{-}OPEs}$ signifies the mean concentration of soil Cl-OPEs (200 $ng \cdot g^{-1}$) and $C_{Total\ OPEs}$ represents the total compiled OPE burden (798 $ng \cdot g^{-1}$), yielding an F_{Cl} value of 25.06%. To provide a comprehensive environmental screening of this pollution matrix, the congener-specific distribution across the three OPE subclasses and their corresponding environmental behaviors within the deep vertical soil cores (0–3 m) are systematically summarized in Table 7. Geochemically, alkyl OPEs accounted for the absolute dominant share (505 $ng \cdot g^{-1}$, 63.28%), which is closely tied to the extensive use of tributyl phosphate as an industrial solvent and defoamer in petrochemical processing lines. However, the prominent accumulation of Cl-OPEs (200 $ng \cdot g^{-1}$) signals a critical environmental threat. Compared to alkyl and aryl counterparts, these chlorinated species exhibit significantly lower soil organic carbon-water partitioning coefficients (K_{oc}) and exceptional resistance to microbial cleavage, allowing them

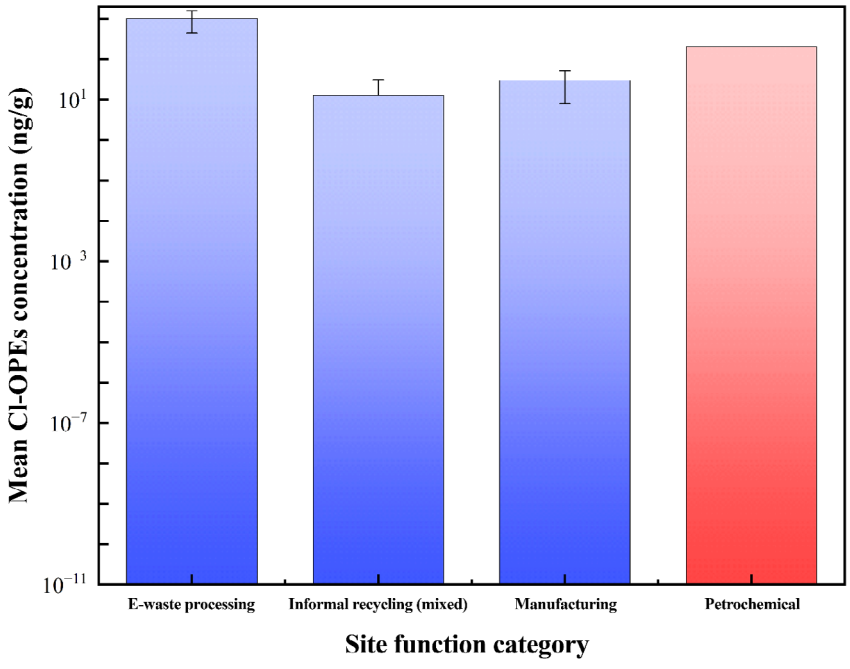

Figure 7. Comparison of total concentrations of chlorinated organophosphate esters (Cl-OPEs) in soils across different industrial function zones and an empirical case study of the petrochemical industry

Table 7. Congener-specific concentration profiles and environmental pollution characteristics of OPEs in the petrochemical study area

Organophosphate ester (OPE) subclass	Quantified congeners identified in this study	Concentration ($ng \cdot g^{-1}$)	Primary environmental behavior and pollution implication
Chlorinated OPEs	Tris(3-chloropropyl) phosphate, tris(2-chloroethyl) phosphate	200	Exceptional persistence; low K_{oc} values driving high vertical mobility into the 0–3 m deep soil profile
Alkyl OPEs	Tributyl phosphate (dominant category), tris(2-butoxyethyl) phosphate	505	High industrial usage as solvents/defoamers; acts as the primary mass burden contributor
Aryl OPEs	Diethyl-3-hydroxyphenyl phosphate, diethyl-2-hydroxyphenyl phosphate, diphenyl phosphate, triphenyl phosphate	93	High affinity for soil organic matter; limited migration potential, but possesses latent endocrine-disrupting risks

to persistently bypass surface attenuation and migrate into deeper soil profiles, thus amplifying the long-term groundwater contamination risks.

3.5. Policy implications and research prospects

By employing an ML method that mitigates geographical bias, this study provides two distinct, highly feasible policy implications for controlling Cl-OPE pollution in industrial soils. First, regarding known risks, regulatory resources should be precisely concentrated on the E-waste processing industry. Our model ($R^2 = 0.90$) provides strong quantitative evidence that E-waste processing is the primary driver of Cl-OPEs pollution, with a contribution far exceeding that of other industrial sectors. This indicates that Cl-OPEs pollution is not a universal industrial problem but a highly industry-clustered one. Therefore, environmental management authorities should formulate and enforce the strictest source-control standards and emission-monitoring protocols specifically for process links such as E-waste dismantling, shredding, and recycling. Second, regarding potential risks, the petrochemical industry should be immediately included in the routine monitoring list for Cl-OPEs. Our empirical research provides preliminary evidence that the petrochemical industrial area is also an important, previously underestimated source of Cl-OPEs pollution, with pollution levels significantly higher than those of traditional manufacturing. This finding sounds an alarm for environmental regulation, suggesting that the existing pollution source inventory may contain omissions. Regulatory agencies should immediately conduct targeted investigations and monitoring of petrochemical parks (especially factories involved in plastic additives, flame-retardant synthesis, or their use). Concurrently, when translating these modeling outcomes into regional soil management frameworks, such policy implications should be interpreted as adaptive diagnostic recommendations rather than definitive, prescriptive criteria for rigid regulatory enforcement. Given that the established feature dependencies under small-sample configurations (n = 141) primarily reflect exploratory statistical associations rather than absolute empirical kinetics or exhaustive source apportionments, environmental frameworks should prioritize flexible, tiered monitoring strategies. This approach balances immediate data-driven predictive screening with long-term mechanistic validation, safeguarding against premature or overly restrictive multi-scale regulatory interventions based on localized modeling abstractions.

Simultaneously, this study opens up two clear directions for future scientific research. First, regarding methodology, future environmental ML research involving cross-regional data should treat the diagnosis of geographical bias as a necessary prerequisite to ensure the robustness and universality of model conclusions. Second, regarding source apportionment, in-depth follow-up research must be conducted on the petrochemical source identified in this study. Future work should focus on: (i) Sampling in a broader range of petrochemical parks to verify the universality of this issue; and (ii) identifying, through mechanistic research, which specific processes in petrochemical production (e.g., catalysis, synthesis, additive mixing) lead to the release of Cl-OPEs.

4. Conclusion

Focusing on the source apportionment of driving factors for Cl-OPEs pollution in industrial soils, this study

successfully constructed and validated an ML method that circumvents geographical bias. Methodologically, this study confirmed that geographical bias is a key pitfall in ML-based environmental source apportionment and proposed an effective diagnose-and-remove solution. The constructed de-geographicalized RF model demonstrated extremely high prediction accuracy (test set R^2 = 0.90) and strong generalization ability (homoscedastic, unbiased residuals), proving the robustness of the method. Regarding driving factor identification, the model provided strong quantitative evidence that E-waste processing is the primary driver of Cl-OPEs pollution. Its importance, combined with the complexity of pollution sources (OPE types), determines the Cl-OPE pollution level, while the contributions of other industrial types and seasonal factors are statistically negligible. Regarding pollution source exploration, this study provided preliminary evidence through an empirical case study indicating that the petrochemical industry (petrochemical) is a previously underestimated, important potential pollution source. The Cl-OPEs concentration at the petrochemical site was significantly higher than at traditional manufacturing, providing an important new perspective for updating the Cl-OPEs pollution source inventory and risk control. In summary, this study not only provides a rigorous method for avoiding bias in environmental ML but also, through the dual pathways of model validation and empirical findings, provides key scientific bases for the precise source apportionment of Cl-OPEs (which should focus on E-waste) and future research (which should include the petrochemical industry).

Acknowledgments

The authors appreciate the experimental support and valuable suggestions from the relevant research institutions and colleagues.

Funding

This work was financially supported by the National Key Research and Development Program of China (Grant No. 2022YFC3703200).

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Chi Zhu, Changsheng Qu
Data curation: Lu Cao, Liang Ding
Formal analysis: Chi Zhu, Lu Cao, Enze Pei
Investigation: Lu Cao, Enze Pei, Xin Huang, Shixiang Dai
Methodology: Chi Zhu, Xin Huang
Resources: Liang Ding, Changsheng Qu
Validation: Xin Huang, Bo Wang
Writing–original draft: Chi Zhu, Lu Cao
Writing–review & editing: Bo Wang, Shixiang Dai, Changsheng Qu

Availability of data

All data generated or analyzed during this study are included in this published article. No external datasets were generated or analyzed.

References

1. Ye L, Li J, Gong S, *et al.* Established and emerging organophosphate esters (OPEs) and the expansion of an environmental contamination issue: A review and future directions. *J Hazard Mater.* 2023;459:132095.

   doi: 10.1016/j.jhazmat.2023.132095

2. Zhao F, Tian S, Liu J, *et al.* Traditional and novel organophosphate esters in atmosphere of greenhouse covered with mulch films: Seasonal variations, partitioning and exposure risks. *J Hazard Mater.* 2025;494:138633.

   doi: 10.1016/j.jhazmat.2025.138633

3. Hu N, Liu X, Zeshan M, *et al.* Unveiling sources of organophosphate esters in marine environments utilizing multi-factor multi-modal high-dimensional clustering algorithm. *Water Res.* 2025;283:123886.

   doi: 10.1016/j.watres.2025.123886

4. Li B, Yang H, Zhang Y, *et al.* Pollution characteristics and health risk assessment of organophosphate esters in agricultural products from different regions of China. *Environ Res.* 2025;278:121675.

   doi: 10.1016/j.envres.2025.121675

5. Mo WQ, Huang ZS, Li QQ, *et al.* Spatial variation, emissions, transport, and risk assessment of organophosphate esters in two large petrochemical complexes in southern China. *J Environ Manag.* 2024;367:122106.

   doi: 10.1016/j.jenvman.2024.122106

6. Gong S, Ren K, Ye L, Deng Y, Su G. Suspect and nontarget screening of known and unknown organophosphate esters (OPEs) in soil samples. *J Hazard Mater.* 2022;436:129273.

   doi: 10.1016/j.jhazmat.2022.129273

7. Zhong M, Zhou M, Tang J, Ren J, Ma R, Li X. Occurrence, spatial distributions, sources, and potential risks of organophosphate esters in Yarlung Tsangpo River and its main tributaries on the Tibetan Plateau. *Environ Pollut.* 2025;376:126385.

   doi: 10.1016/j.envpol.2025.126385

8. Hoang AQ, Karyu R, Tue NM, *et al.* Comprehensive characterization of halogenated flame retardants and

organophosphate esters in settled dust from informal e-waste and end-of-life vehicle processing sites in Vietnam: Occurrence, source estimation, and risk assessment. *Environ Pollut.* 2022;310:119809.

doi: 10.1016/j.envpol.2022.119809

9. Zhu C, Yu Z, Chen Y, *et al.* Distribution patterns and origins of organophosphate esters in soils from different climate systems on the Tibetan Plateau. *Environ Pollut.* 2024;351:124085.

doi: 10.1016/j.envpol.2024.124085

10. Zhou L, Wang X, Zhang X, *et al.* Besides traditional organophosphate esters: The ecological risks of emerging organophosphate esters in the Yangtze River basin cannot be ignored. *Environ Pollut.* 2025;367:125585.

doi: 10.1016/j.envpol.2024.125585

11. Teshome FT, Bayabil HK, Schaffer B, Ampatzidis Y, Hoogenboom G. Improving soil moisture prediction with deep learning and machine learning models. *Comput Electron Agric.* 2024;226:109414.

doi: 10.1016/j.compag.2024.109414

12. Duan D, Wang P, Rao X, *et al.* Identifying interactive effects of spatial drivers in soil heavy metal pollutants using interpretable machine learning models. *Sci Total Environ.* 2024;934:173284.

doi: 10.1016/j.scitotenv.2024.173284

13. Siqueira RG, Moquedace CM, Fernandes-Filho EI, *et al.* Modelling and prediction of major soil chemical properties with Random Forest: Machine learning as tool to understand soil-environment relationships in Antarctica. *CATENA*. 2024;235:107677.

doi: 10.1016/j.catena.2023.107677

14. Withana PA, Li J, Senadheera SS, Fan C, Wang Y, Ok YS. Machine learning prediction and interpretation of the impact of microplastics on soil properties. *Environ Pollut.* 2024;341:122833.

doi: 10.1016/j.envpol.2023.122833

15. Wang Y, Zhu H, Yao Y, *et al.* Distribution of Organophosphate Flame Retardants (OPFRs) in Soil and Outdoor Settled Dust from A Multi-waste Recycling Area. In: Proceedings of the POPs Forum 2017 and the 12th Symposium on Persistent Organic Pollutants. Wuhan, China; 2017:2.

16. YYin H, Liu L, Xiong Y, Qiao Y. Pollution characteristics and risk assessment of organophosphate esters (OPEs) in typical industrial parks in Southwest China. *Environ Sci Pollut Res.* 2024;31(24):35206-35218.

doi: 10.1007/s11356-024-33160-w

17. Ge X, Ma S, Zhang X, Yang Y, Li G, Yu Y. Halogenated and organophosphorous flame retardants in surface soils from an e-waste dismantling park and its surrounding area: Distributions, sources, and human health risks. *Environ Int.* 2020;139:105741.

doi: 10.1016/j.envint.2020.105741

18. Sun Y, Zhu H. A pilot study of organophosphate esters in surface soils collected from Jinan City, China: implications for risk assessments. *Environ Sci Pollut Res.* 2021;28(3):3344-3353.

doi: 10.1007/s11356-020-10730-2

19. Kurt-Karakus P, Alegria H, Birgul A, Gungormus E, Jantunen L. Gungormus, and L. Jantunen. Organophosphate ester (OPEs) flame retardants and plasticizers in air and soil from a highly industrialized city in Turkey. *Sci Total Environ.* 2018;625:555-565.

doi: 10.1016/j.scitotenv.2017.12.307

20. Tang J, Sun J, Ke Z, *et al.* Organophosphate esters in surface soils from a heavily urbanized region of Eastern China: Occurrence, distribution, and ecological risk assessment. *Environ Pollut.* 2021;291:118200.

doi: 10.1016/j.envpol.2021.118200

21. Sun Y, Chen S, Jiang H, *et al.* Towards interpretable machine learning for observational quantification of soil heavy metal concentrations under environmental constraints. *Sci Total Environ.* 2024;926:171931.

doi: 10.1016/j.scitotenv.2024.171931

22. Wang J, Huang R, Liang Y, *et al.* Prediction of antibiotic sorption in soil with machine learning and analysis of global antibiotic resistance risk. *J Hazard Mater.* 2024;466:133563.

doi: 10.1016/j.jhazmat.2024.133563

23. Zhang L, Heuvelink GB, Mulder VL, Chen S, Deng X, Yang L. Using process-oriented model output to enhance machine learning-based soil organic carbon prediction in space and time. *Sci Total Environ*. 2024;922:170778.

doi: 10.1016/j.scitotenv.2024.170778

24. Yan Y, Yang Y. Revealing the synergistic spatial effects in soil heavy metal pollution with explainable machine learning models. *J Hazard Mater.* 2025;482:136578.

doi: 10.1016/j.jhazmat.2024.136578

25. Feng L, Khalil U, Aslam B, *et al.* Evaluation of soil texture classification from orthodox interpolation and machine learning techniques. *Environ Res.* 2024;246:118075.

doi: 10.1016/j.envres.2023.118075

26. Badounas I, Bozikas A, Pitselis G. A robust random coefficient regression representation of the chain-ladder method. *Ann Actuar Sci* 2022;16(1):151-182.

doi: 10.1017/S1748499521000154

27. Mohammed S, Arshad S, Bashir B, *et al.* Evaluating machine learning performance in predicting sodium adsorption ratio for sustainable soil-water management in the eastern

Mediterranean. *J Environ Manag*. 2024;370:122640.

doi: 10.1016/j.jenvman.2024.122640

28. Wang YG, Wu J, Hu ZH, McLachlan GJ. A new algorithm for support vector regression with automatic selection of hyperparameters. *Pattern Recognit*. 2023;133:108989.

doi: 10.1016/j.patcog.2022.108989

29. Zhang Z. Introduction to machine learning: k-nearest neighbors. *Ann Transl Med*. 2016;4(11):218.

doi: 10.21037/atm.2016.03.37

30. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. *WIREs Data Min Knowl Discov.* 2019;9(3):e1301.

doi: 10.1002/widm.1301

31. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*. 2020;415:295-316.

doi: 10.1016/j.neucom.2020.07.061

32. Occhipinti A, Rogers L, Angione C. A pipeline and comparative study of 12 machine learning models for text classification. *Expert Syst Appl.* 2022;201:117193.

doi: 10.1016/j.eswa.2022.117193

33. Belete DM, Huchaiah MD. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *Int J Comput Appl*. 2022;44(9):875-886.

doi: 10.1080/1206212X.2021.1974663

34. Karl F, Pielok T, Moosbauer J, *et al.* Multi-Objective Hyperparameter Optimization in Machine Learning—An Overview. *ACM Trans. Evol Learn Optim*. 2023;3(4):1-50.

doi: 10.1145/3610536

35. Rengasamy D, Mase JM, Kumar A, *et al.* Feature importance in machine learning models: A fuzzy information fusion approach. *Neurocomputing*. 2022;511:163-174.

doi: 10.1016/j.neucom.2022.09.053

36. JJebarathinam C, Home D, Sinha U. Pearson correlation coefficient as a measure for certifying and quantifying high-dimensional entanglement. *Phys Rev A* 2020;101(2):022112.

doi: 10.1103/PhysRevA.101.022112

37. Kenny M, Schoen I. Violin SuperPlots: visualizing replicate heterogeneity in large data sets. *Mol Biol Cell* 2021;32(15):1333-1334.

doi: 10.1091/mbc.E21-03-0130

Appendix

A1. Theoretical background and operational principles of candidate machine learning regressors

The working principles of these models are distinct. Support vector regression is a regression method based on support vector machine principles, whose core idea is to construct an optimal hyperplane in a high-dimensional feature space that minimizes error across all sample points. In contrast, K-nearest neighbors is an instance-based, non-parametric method that does not build an explicit functional model; instead, it makes predictions by averaging the K nearest neighbors of the test sample in the training dataset. Random forest is a powerful ensemble learning method that builds and combines multiple independent decision trees for prediction, effectively reducing the risk of overfitting associated with a single decision tree and enhancing model stability. This study will compare the performance of these four models in predicting Cl-OPEs concentration to select the optimal model.