Migration of Total Petroleum Hydrocarbons and heavy metals Contaminants in the Soil-Groundwater Interface of Petrochemical Site using machine learning: Impacts of Convection and Diffusion

Yingdong Wu[1,2], Jiang Yu[1,2,3,*], Zhi Huang[1,2], Yinying Jiang[1,2], Zixin Zeng[1,2], Lei Han[1,2], Siwei Deng[1,4], Jie Yu[1,2]

[1]*Department of Environmental Science and Engineering, College of Architecture and Environment, Sichuan University, Chengdu, 610065, PR China*

[2] *Institute of New Energy and Low Carbon Technology, Sichuan University, Chengdu, 610065, PR China*

[3] *Yibin Institute of Industrial Technology, Sichuan University, Yibin 644000, PR China*

[4]*Soil and Groundwater Pollution Prevention Research Institute, Sichuan Academy of Eco-Environmental Sciences, 610046, Chengdu, P.R. China*

* Corresponding author

E-mail address: yuj@scu.edu.cn(J.Yu).

The supplementary material (total of 27 pages including coversheet) contains 3 texts (Text S1-S3), 5 tables (Table S1-S5), and 14 figures (Fig. S1-S14).

TEXT S1 Description of Bayesian Optimization

1. What is Bayesian Optimization?

Bayesian Optimization is a sequential optimization strategy used to find the best parameters of a model by building a probabilistic model, typically a Gaussian Process, of the objective function. Instead of evaluating all possible combinations of parameters, Bayesian optimization selects parameter values based on a balance between exploring unknown regions of the parameter space and exploiting regions known to have high performance. By iteratively refining the model based on prior evaluations, it efficiently finds the optimal configuration of hyperparameters, reducing computational cost compared to grid or random search methods.

2. Parameter Search Space in Bayesian Optimization

In my models, we used different parameter ranges for Bayesian optimization across various machine learning algorithms to fine-tune their performance:

RF: The number of estimators (n_estimators) was searched between 100 and 500, maximum tree depth (max_depth) between 10 and 50, and minimum samples required to split a node (min_samples_split) from 2 to 10.

XGB: The search space included n_estimators from 50 to 300, learning rate (learning_rate) from 0.01 to 0.3, and maximum depth (max_depth) between 3 and 9.

SVM: The regularization parameter C was searched between 0.1 and 100 on a logarithmic scale, with gamma tested for both "scale" and "auto" values, and kernel functions tested between 'linear' and 'rbf'.

KNN: The number of neighbors (n_neighbors) was varied from 3 to 9, with weighting schemes (weights) between 'uniform' and 'distance'.

DTree: The max_depth was explored between 10 and 90, and min_samples_split from 2 to 10.

ANN: The hidden layer sizes (hidden_layer_sizes) were tested with combinations of (50,), (100,), (50, 50), and (100, 50), activation functions (activation) between 'relu' and 'tanh', solvers (solver) between 'adam' and 'lbfgs', and learning rates (learning_rate) between 'constant' and 'adaptive'.

EBM: The max_bins was set between 64 and 256, max_leaves between 3 and 7, and learning_rate from 0.01 to 0.1.

These ranges were designed to explore diverse settings and ensure the optimal configuration of each

algorithm is efficiently found through Bayesian optimization.

TEXT S2 Description of SHAP

SHAP (Shapley Additive Explanations) is grounded in game theory and leverages Shapley values to explain the contribution of each feature in a model's predictions. The core idea is to distribute the prediction among the input features fairly, based on their contribution to the final output. Below are some key equations and concepts that underlie SHAP:

1. Shapley Value Formula

For a feature i, the Shapley value $\emptyset_i$is calculated as:

$$\emptyset_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

Where: N is the set of all features;S is a subset of N that does not contain feature i; f(S) represents the prediction of the model using only the features in subset S; $f(S \cup \{i\}) - f(S)$ calculates the marginal contribution of feature i to the subset S.

This equation ensures that each feature's contribution is fairly attributed by considering all possible combinations of feature subsets.

2. Additive Feature Attribution

SHAP falls under additive feature attribution methods, where the explanation model is assumed to be linear with respect to the contributions of the features. The additive nature can be expressed as:

$$f(x) = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i$$

Where: f(x) is the output of the machine learning model for input x; $\emptyset_0$ is the base value (i.e., the mean prediction across all samples); $\emptyset$ represents the Shapley value of feature iii, which indicates the contribution of feature i to the model prediction for the specific input x.

3. Efficiency, Symmetry, and Additivity

The Shapley values satisfy several important properties, which make them suitable for explaining machine

learning models:

Efficiency: The sum of all feature contributions ϕi\phi_iϕi equals the difference between the prediction and the base value.

Symmetry: If two features contribute equally to a model, their Shapley values will be identical.

Additivity: The Shapley values for multiple models can be combined to represent the ensemble.

4. Approximation for Large Models

For complex machine learning models, exact computation of Shapley values using the above formula can be computationally expensive. Therefore, approximate methods like KernelSHAP and TreeSHAP have been developed to compute Shapley values more efficiently.

KernelSHAP: A model-agnostic approach that approximates the Shapley values using weighted linear regression.

TreeSHAP: A specialized algorithm for decision trees, which allows for the efficient computation of exact Shapley values in tree-based models like random forests and gradient-boosting machines.

In summary, SHAP provides a robust, mathematically grounded approach for interpreting model predictions, ensuring transparency in machine learning through a fair distribution of feature contributions based on Shapley values. These formulas and principles make SHAP a powerful tool for model interpretability.

TEXT S3 Derivations and Assumptions of Convective and Diffusive Factors

1. Derivation and Assumptions for Convective Factor (Equation 3)

Equation (3) represents the change in pollutant concentration due to the convective process. Convection describes the transport of pollutants driven by fluid movement, and the change in concentration over time can be approximated by the time derivative of concentration:

$$\mathrm{C_S} = \frac{\partial C}{\partial t}$$

Where: C_s is the convective factor, representing the rate of change in pollutant concentration over time; C is the pollutant concentration; T is time.

Assumptions:

Assumption 1: The convective process is steady, with pollutant transport dominated by the macroscopic movement of groundwater, and microscopic diffusion effects are neglected.

Assumption 2: The pollutant concentration changes over time can be approximated as a steady process, ignoring short-term fluctuations.

This assumption simplifies the description of the convective process, allowing it to capture the time-related changes in pollutant transport with fewer parameters.

2. Derivation and Assumptions for Diffusive Factor (Equation 4)

Equation (4) describes how the vertical concentration gradient of pollutants is calculated during the diffusion process. Based on Fick's law of diffusion, the rate of diffusion can be expressed as the gradient of concentration with respect to depth:

$$\mathrm{D_S} = \frac{\partial C_n}{\partial z} = \frac{C_n - C_{n-1}}{Z_n - Z_{n-1}}$$

Where: D_s is the diffusive factor, representing the vertical change in concentration; C_n and C_{n-1} represent the pollutant concentrations in the n-th and n−1-th soil layers, respectively; Z_n represent the depths of the n-th and n−1-th layers.

Assumptions:

Assumption 1: The diffusion process occurs predominantly in the vertical direction, with lateral diffusion effects neglected.

Assumption 2: The change in pollutant concentration can be approximated by discrete soil layers, with uniform concentration assumed within each layer.

This method simplifies the complex three-dimensional diffusion process by focusing solely on the vertical concentration changes, enabling effective modeling of diffusion under limited data conditions.

Table S1 Comparison of the advantages and disadvantages of machine learning algorithms used in this study.

Name	Description	Advantages	Disadvantages
Artificial Neural Network (ANN)	ANNs consist of interconnected "neurons" or nodes in multiple layers, including input, hidden, and output layers. Deep learning models, a subset of ANNs, involve stacking many layers to model highly complex tasks.	①Can model highly non-linear and complex relationships. ②Effective for large datasets and tasks like image recognition,speech processing.	①Requires large amounts of data and computational resources for training. ②Difficult to interpret due to its black-box nature.
Decision Tree (DTree)	Decision trees split the dataset into smaller subsets by creating decision nodes based on feature values, leading to predictions at the tree's leaves.	①Easy to interpret and understand, even for non-experts. ②Handles both numerical and categorical data.	①Prone to overfitting, especially with deep trees. ②Sensitive to small variations in data, leading to instability.
Explainable Boosting Machine (EBM)	EBM is a type of Generalized Additive Model (GAM) that remains interpretable while capturing both linear and non-linear relationships. It's particularly designed for transparency.	①Highly interpretable, suitable for scenarios requiring explainability. ②Handles complex relationships while providing a clear understanding of each feature's impact.	①Training can be slower compared to other ensemble models ②May not be as powerful as less interpretable methods like XGBoost in terms of predictive accuracy
K-nearest Neighbors (KNN)	KNN is a non-parametric algorithm that predicts the value of a target variable by averaging the values of the K-nearest data points in the feature space.	①Simple and intuitive to understand and implement. ②No need for a training phase—predictions are made directly from the dataset.	①Computationally expensive during prediction, especially with large datasets. ②Sensitive to the choice of K and scaling of features.
Linear Regression (LR)	Linear regression models the relationship between the dependent variable and one or more independent variables using a straight line (linear function).	①Simple to implement and interpret. ②Computationally efficient and well-suited for small datasets.	①Assumes linearity, which may not always be accurate. ②Vulnerable to multicollinearity (correlated features) and outliers.
Random Forest (RF)	Random Forest builds multiple decision trees (each using a random subset of features) and averages their predictions to improve accuracy and reduce overfitting.	①High accuracy and robustness to overfitting. ②Can handle missing data and feature importance can be derived.	①Difficult to interpret individual trees within the forest. ②Can be computationally intensive and slow on very large datasets.
Support Vector Machine (SVM)	SVMs find the hyperplane that best separates data points from different classes. For regression tasks, SVM can also fit data with a margin of tolerance.	①Effective in high-dimensional spaces and non-linear problems. ②Versatile with the use of different kernel functions (linear, polynomial, radial basis function).	①Slow for large datasets and challenging to tune the hyperparameters (like the choice of kernel). ②Memory-intensive, as it requires storing the entire training set for predictions.
XGBoost (XGB)	XGBoost is an efficient implementation of gradient-boosting algorithms, focusing on improving prediction accuracy by iteratively adding decision trees.	①Highly efficient and scalable, suitable for large datasets. ②Handles missing data automatically and includes regularization techniques to reduce overfitting.	①Complex to tune and requires careful optimization of hyperparameters. ②Can be slow to train for very large datasets if not optimized properly.

Table S2 Basic statistical parameters of TPH and heavy metal concentrations in soil and groundwater at the study site.

				TPH	As	Co	Ni	Pb
Soil (mg/kg)	Statistics based on depth	0 - 0.5 m (n=459)	Min	ND	5.110	2.582	5.337	2.880
			Max	30000.000	51.227	280.953	714.864	1114.107
			Mean	502.397	14.352	15.289	31.839	38.383
			SD	1997.717	8.316	13.538	46.308	55.208
			DR%	97.17	100	100	100	100
			CV	3.976	0.579	0.885	1.454	1.438
		0.5 - 1.5 m (n=527)	Min	ND	1.165	3.866	12.146	9.849
			Max	26412.000	95.890	65.525	207.718	836.996
			Mean	654.204	14.218	14.273	27.957	34.464
			SD	1634.338	7.765	7.760	14.864	40.483
			DR%	93.36	100	100	100	100
			CV	2.498	0.546	0.544	0.532	1.175
		1.5 - 2.5 m (n=478)	Min	6.000	2.049	4.164	11.087	11.705
			Max	39715.027	52.037	51.979	125.782	190.574
			Mean	821.069	13.650	14.089	27.547	30.798
			SD	2427.848	7.371	9.423	20.130	18.624
			DR%	100	100	100	100	100
			CV	2.957	0.540	0.669	0.731	0.605
		2.5 - 4.0 m (n=366)	Min	ND	1.615	4.964	10.272	13.102
			Max	8421.575	52.319	54.347	107.521	1204.567
			Mean	244.352	14.361	14.503	28.639	34.197
			SD	851.084	7.541	6.412	11.682	70.566
			DR%	90.16	100	100	100	100
			CV	3.483	0.525	0.442	0.408	2.064
	Soil reference value*[1]			826	40	40	150	400
Groundwater (μg/L)	Perched water (n=46)		Min	40.000	0.300	0.780	0.710	1.450
			Max	44500.000	63.400	2778.000	2367.000	249.000
			Mean	4887.347	5.555	63.167	54.980	28.535
			SD	9020.447	12.247	391.899	333.803	42.375
			CV	1.846	2.205	6.204	6.071	1.485
			OSR%	95.65	4.348	2.174	2.174	6.522
	Pore water (n=69)		Min	30.000	0.300	0.153	0.924	0.621
			Max	4140.000	18.500	93.225	282	358.593
			Mean	556.429	3.572	11.730	16.110	26.484
			SD	962.932	3.561	18.427	35.471	50.682
			CV	1.731	0.997	1.571	2.202	1.914
			OSR%	17.391	5.797	0	2.899	4.348
	Groundwater reference value*[2]			500	10	100	100	100

[1]: Risk intervention values (GB36600-2018)[1];[2]: Risk intervention values (GB-T14848-2018[2]；GB3838-2002[3])；ND: Not detected; SD: Standard Deviation; DR: Detection rate; CV: Coefficient of Variation; OSR: Over-standard rate.

Table S3 The model fitting results

		Pre-training		Validation		Test	
		R^2	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
RF	TPH	0.951	0.354	0.883	0.568	0.830	0.614
	As	0.929	0.103	0.728	0.197	0.800	0.170
	Co	0.948	0.082	0.695	0.174	0.697	0.173
	Ni	0.975	0.050	0.861	0.102	0.883	0.100
	Pb	0.994	0.030	0.978	0.067	0.964	0.072
XGB	TPH	0.916	0.463	0.875	0.588	0.825	0.622
	As	0.883	0.133	0.717	0.201	0.783	0.177
	Co	0.870	0.130	0.704	0.171	0.693	0.174
	Ni	0.984	0.041	0.880	0.095	0.870	0.105
	Pb	1.000	0.001	0.918	0.130	0.962	0.074
SVM	TPH	0.818	0.682	0.730	0.863	0.693	0.825
	As	0.792	0.177	0.739	0.193	0.780	0.179
	Co	0.792	0.164	0.698	0.173	0.682	0.177
	Ni	0.934	0.082	0.860	0.103	0.810	0.127
	Pb	0.970	0.065	0.741	0.231	0.796	0.170
KNN	TPH	1.000	0.000	0.548	1.116	0.458	1.096
	As	1.000	0.000	0.671	0.217	0.739	0.195
	Co	1.000	0.000	0.511	0.220	0.553	0.210
	Ni	1.000	0.000	0.605	0.172	0.670	0.168
	Pb	1.000	0.000	0.634	0.275	0.561	0.249
ANN	TPH	0.842	0.635	0.868	0.602	0.810	0.649
	As	0.781	0.182	0.684	0.212	0.812	0.165
	Co	0.852	0.138	0.599	0.199	0.672	0.180
	Ni	0.937	0.080	0.857	0.104	0.834	0.119
	Pb	0.995	0.027	0.991	0.042	0.983	0.049
DTree	TPH	0.942	0.385	0.779	0.781	0.690	0.829
	As	0.911	0.116	0.622	0.232	0.600	0.241
	Co	0.928	0.096	0.497	0.223	0.525	0.217
	Ni	0.962	0.062	0.804	0.121	0.817	0.125
	Pb	0.998	0.015	0.988	0.049	0.960	0.075
EBM	TPH	0.911	0.476	0.897	0.512	0.812	0.645
	As	0.840	0.156	0.724	0.198	0.791	0.174
	Co	0.858	0.136	0.708	0.170	0.692	0.175
	Ni	0.933	0.083	0.884	0.093	0.884	0.099
	Pb	0.987	0.043	0.956	0.095	0.981	0.051
LR	TPH	0.376	1.263	0.375	1.312	0.387	1.166
	As	0.670	0.224	0.706	0.205	0.709	0.206
	Co	0.529	0.247	0.516	0.219	0.495	0.224
	Ni	0.712	0.171	0.639	0.165	0.544	0.197
	Pb	0.522	0.260	0.626	0.278	0.558	0.250

Table S4 The Hyperparameter results of different models

	Hyperparameter	TPH	As	Co	Ni	Pb
RF	Max_depth	13	10	10	10	10
	Min_samples_split	10	6	2	2	4
	n_estimators	132	237	322	127	500
XGB	Learning_rate	0.059	0.048	0.169	0.126	0.3
	Max_depth	4	3	3	4	8
	n_estimators	104	217	50	150	104
SVM	C	8.030	1.023	1.251	3.341	9.493
	gamma	scale	scale	scale	scale	scale
	kernel	rbf	rbf	rbf	rbf	rbf
KNN	n_neighbors	9	9	8	6	7
	weigths	distance	distance	distance	distance	distance
DTree	Max_depth	10	10	10	10	10
	Min_samples_split	9	9	10	9	6
ANN	activation	tanh	tanh	tanh	tanh	tanh
	Hidden_layer_sizes	(100.50)	(100.50)	(50,)	(50,)	(50,50)
	Learning_rate	constant	constant	constant	constant	constant
	solver	adam	adam	lbfgs	lbfgs	lbfgs
EBM	Learning_rate	0.066	0.010	0.010	0.010	0.010
	Max_bins	223	177	254	222	202
	Max_leaves	4	3	3	3	3

Table S5 Linear Regression Results

			TPH	As	Co	Ni	Pb
Perched water	Convection	Beta	0.365	0.056	0.142	0.796	0.188
		p	0.049*	0.836	0.531	0.071	0.528
	Diffusion	Beta	0.888	0.363	0.193	0.193	0.114
		p	0.000***	0.190	0.399	0.649	0.701
Pore water	Convection	Beta	0.239	0.229	0.110	0.006	0.333
		p	0.298	0.464	0.715	0.989	0.202
	Diffusion	Beta	-0.012	0.414	0.359	-0.240	-0.192
		p	0.959	0.191	0.242	0.587	0.457

*:$p<0.05$; ***:$p<0.001$

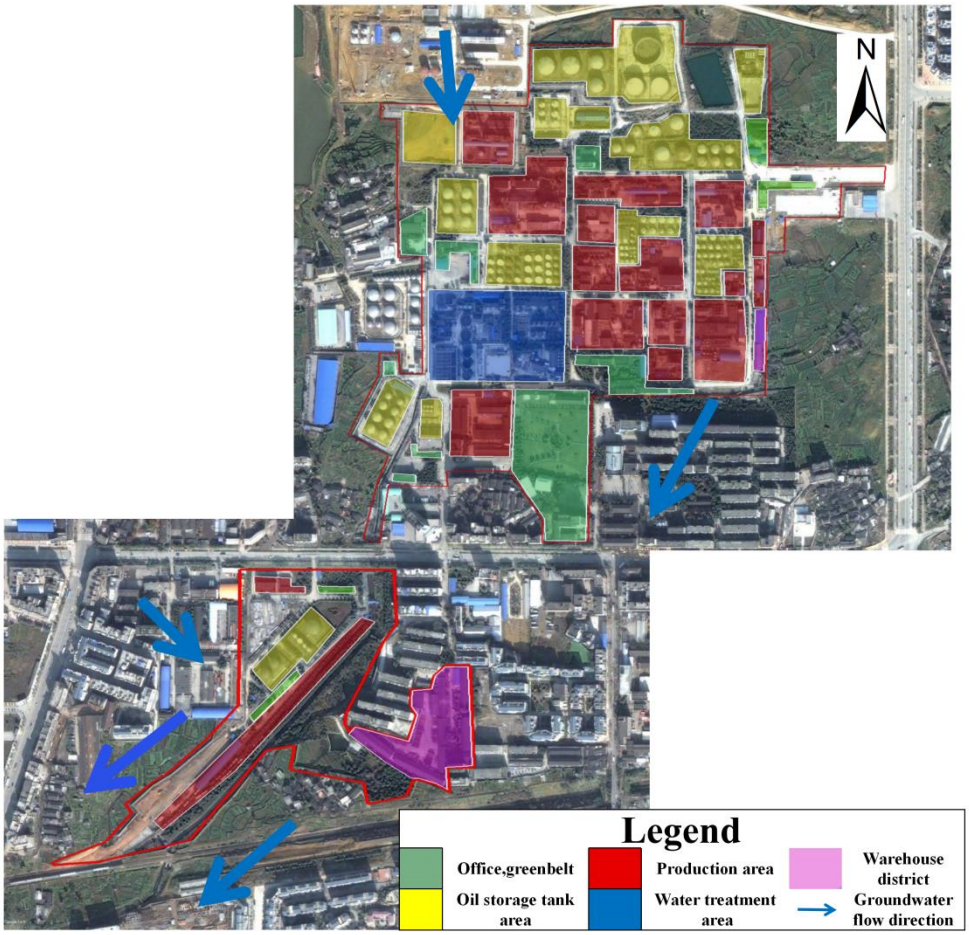

Fig. S1. Functional Zoning of the Study Site

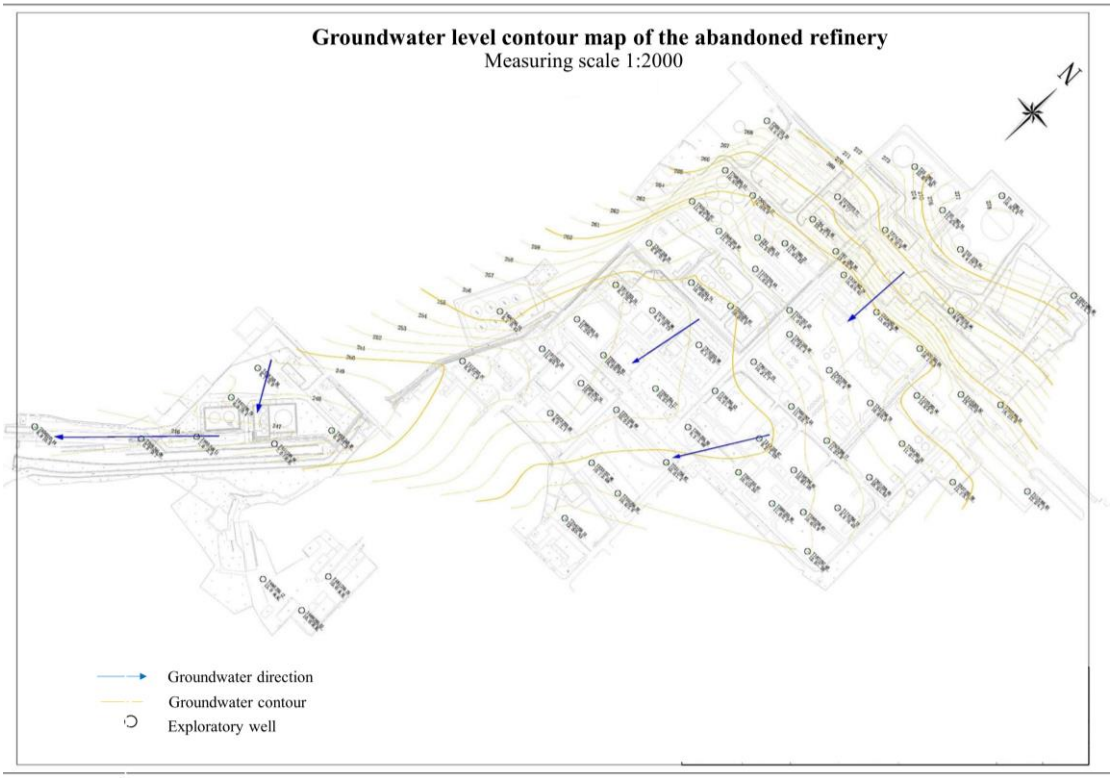

Fig. S2 Groundwater level contour map of the abandoned refinery

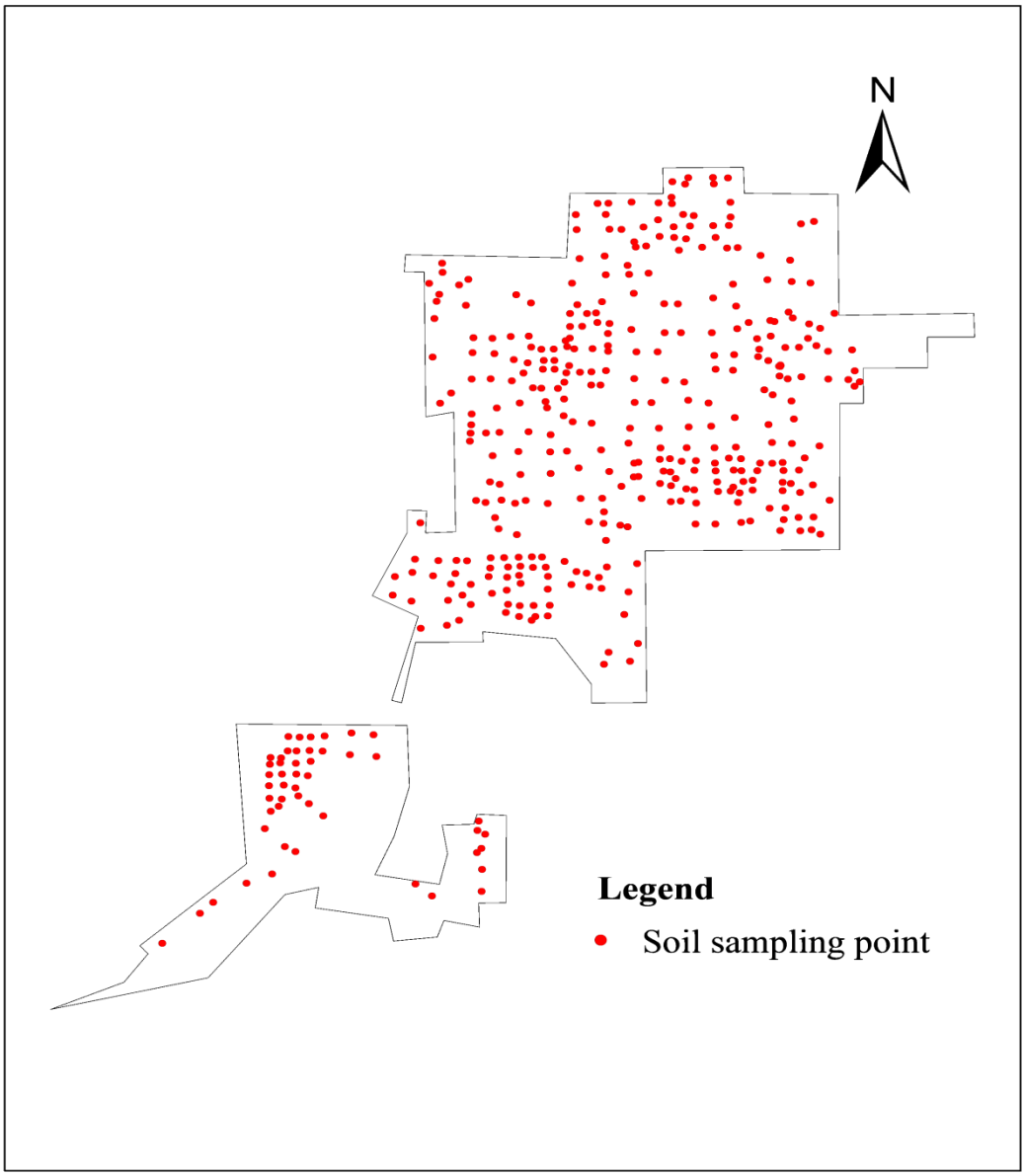

Fig. S3. Sampling Locations in the Abandoned Refinery's Soil

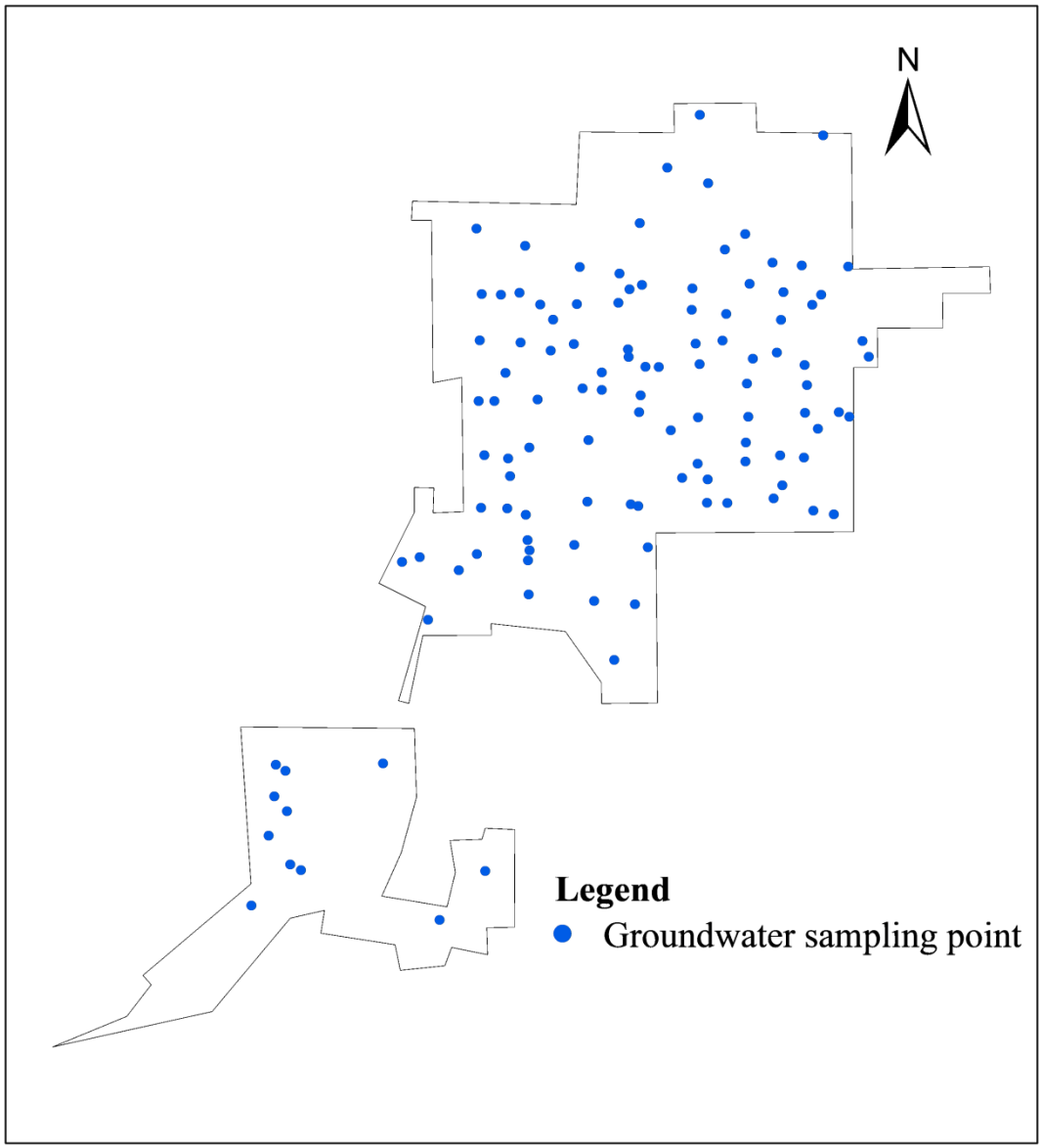

Fig. S4. Groundwater Sampling Locations at the Abandoned Refiner

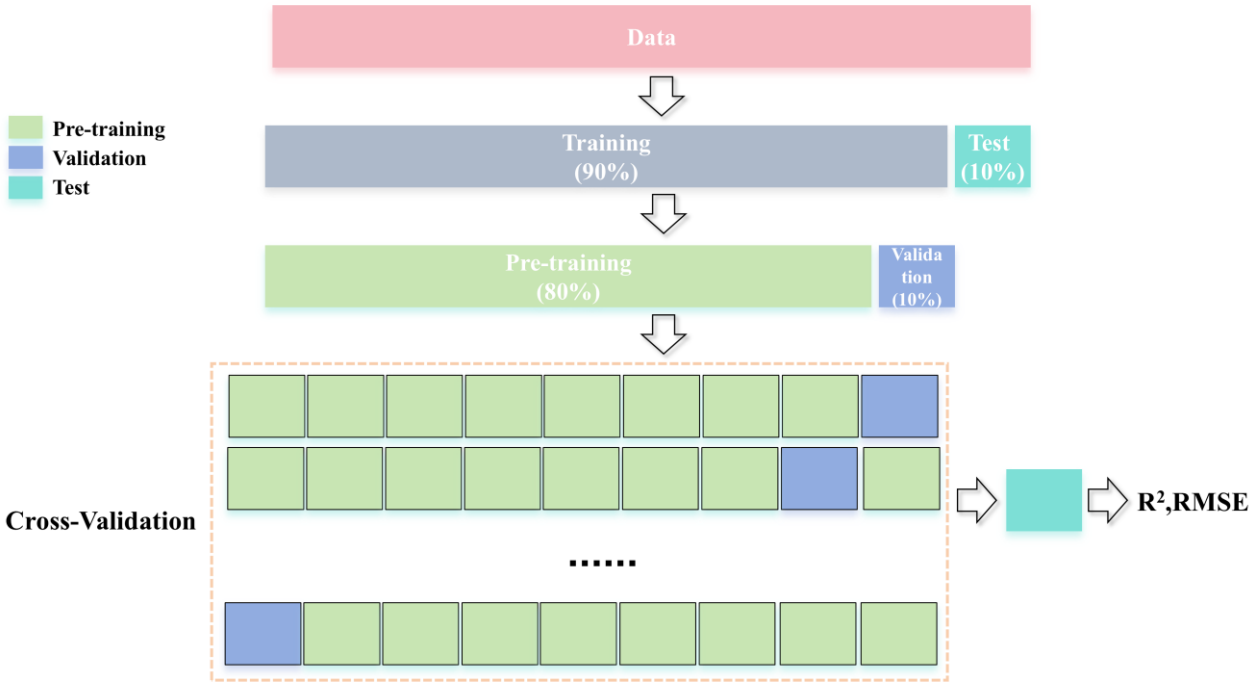

Fig. S5. Model data segmentation diagram

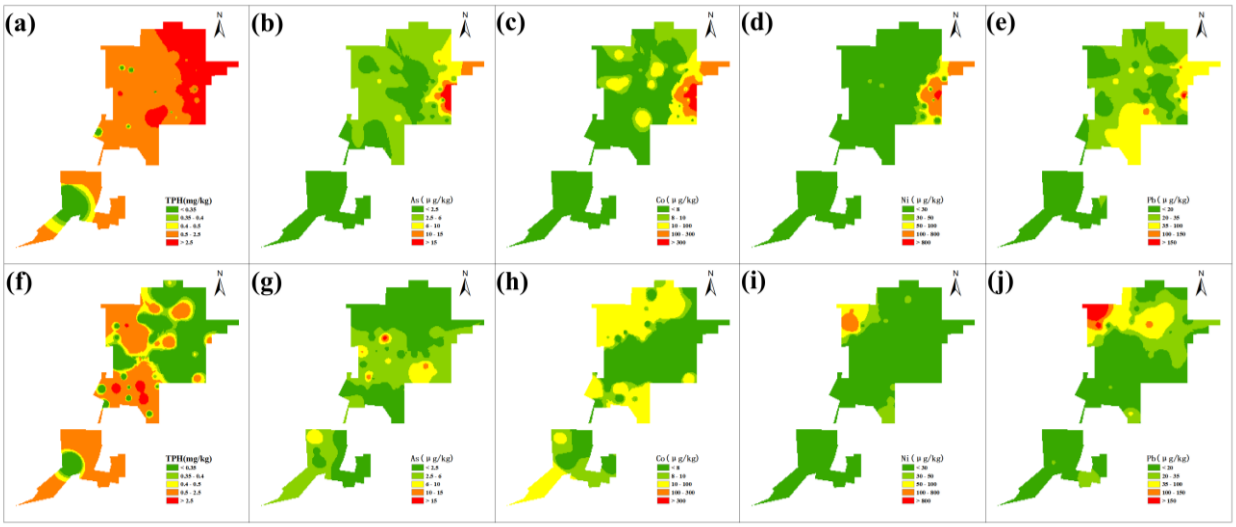

Fig. S6. Spatial Distribution of TPH and HMs in Groundwater. Panels (a) to (e) display the distribution of TPH, As, Co, Ni, and Pb in the perched water. Panels (f) to (j) illustrate the distribution of TPH, As, Co, Ni, and Pb in pore water. The Risk intervention values (GB-T14848-2018[2]; GB3838-2002[3]) for pollutants are as follows: TPH is set at 0.5 mg/kg, As at 10 μg/kg, Co at 100 μg/kg, Ni at 100 μg/kg, and Pb at 100 μg/kg. The orange or red areas indicate that the pollutant levels in these regions exceed the standard values.

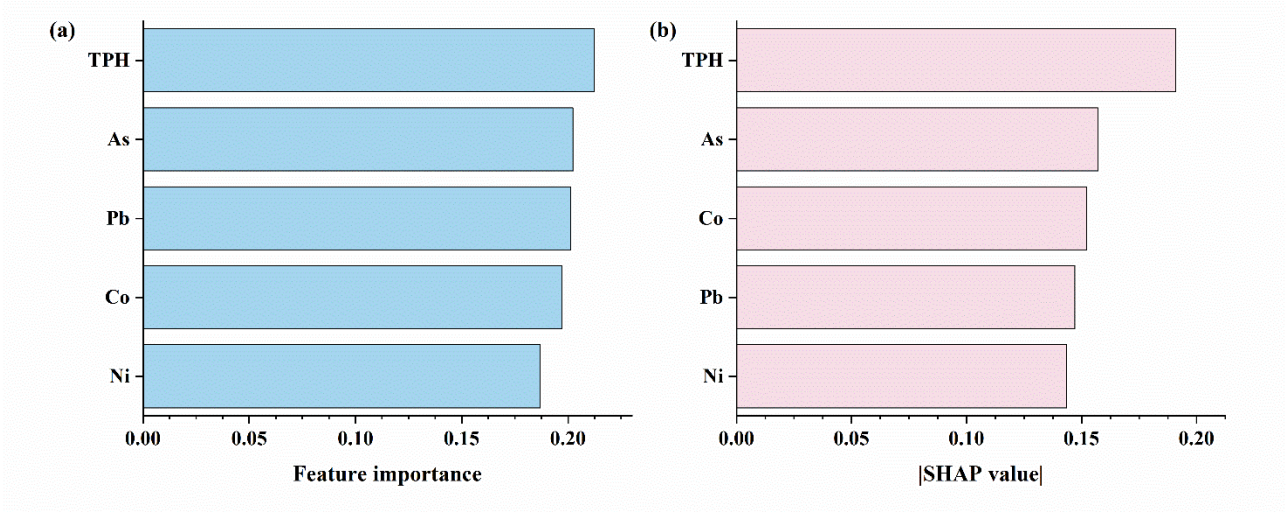

Fig. S7. Relationship between soil depth and concentrations of TPH and HMs. (a) Based on importance indices from Random Forest; (b) based on SHAP value.

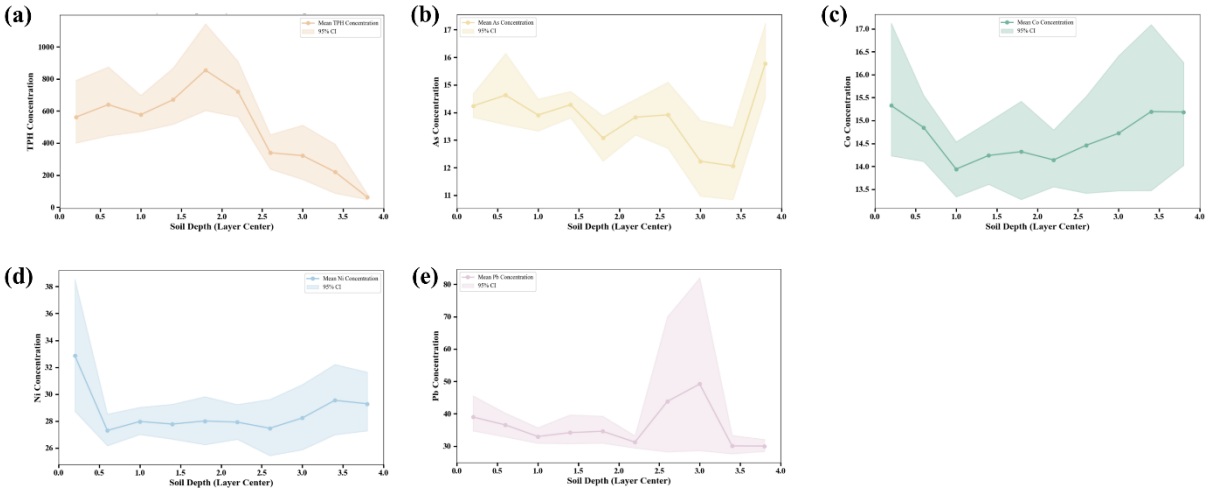

Fig. S8. Bootstrapped confidence intervals. (a) TPH, (b) As, (c) Co, (d) Ni, (e)Pb.

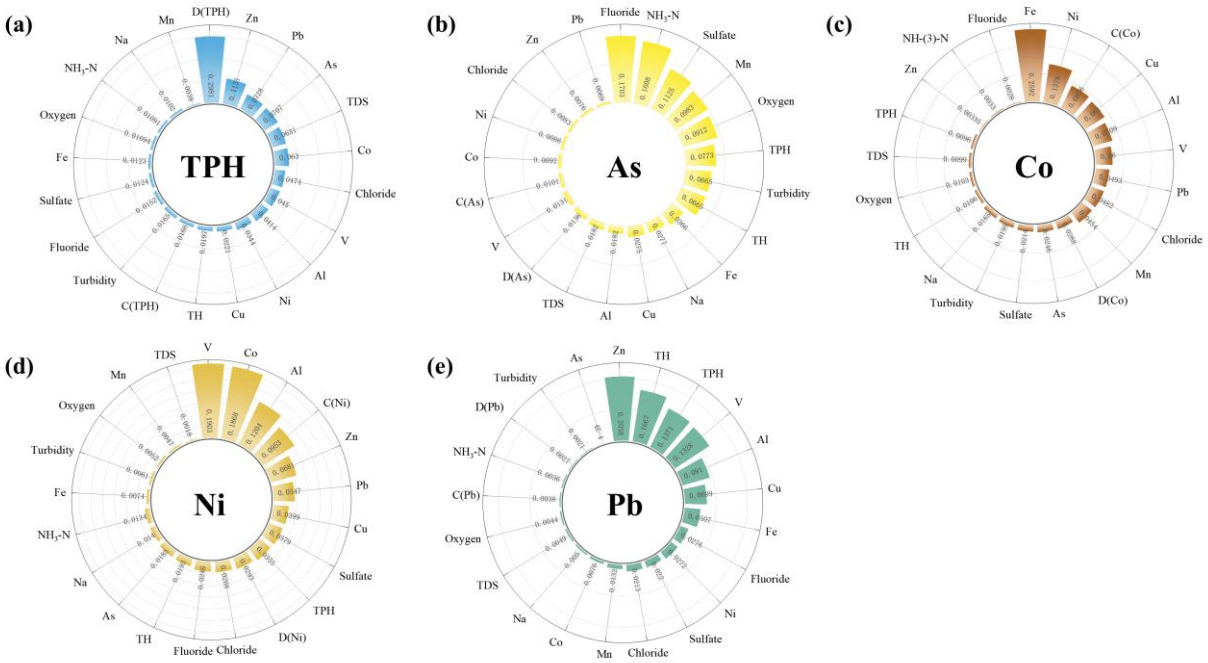

Fig. S9. Importance Indices of the Random Forest Model for TPH and HMs in perched water. TDS: Total dissolved solids; TH: Total hardness; D(.): Diffusion of TPH, As, Co, Ni, Pb; C(.): Convection of TPH, As, Co, Ni, Pb.

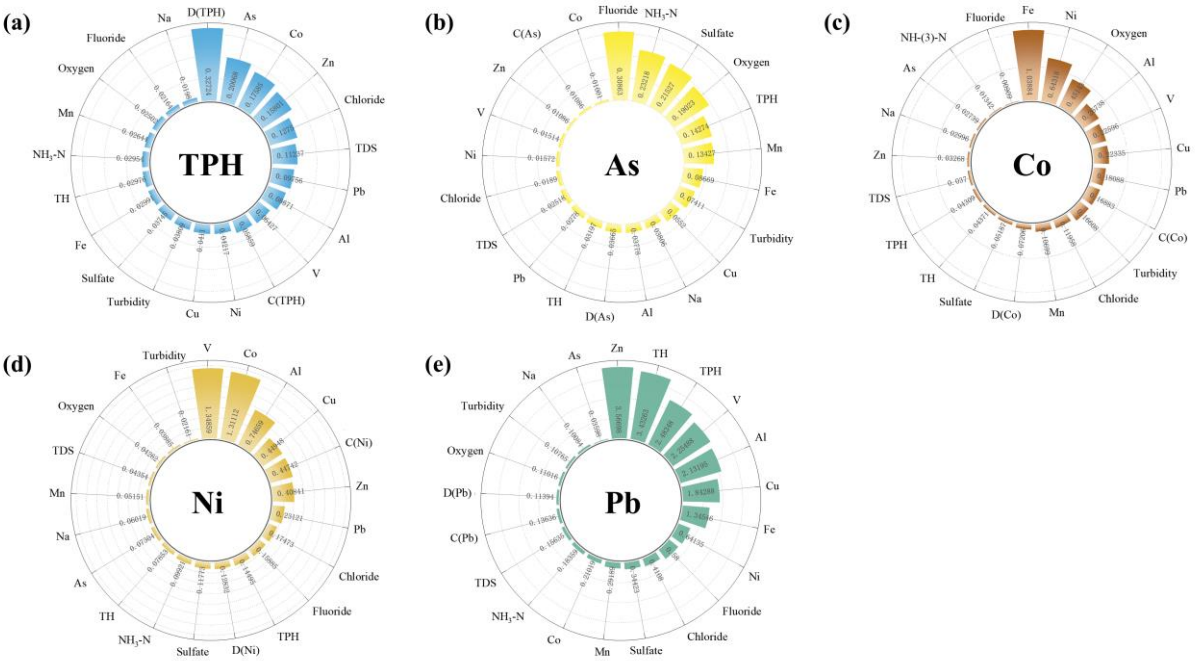

Fig. S10. SHAP value for TPH and HMs in perched water. TDS: Total dissolved solids; TH: Total hardness; D(.): Diffusion of TPH, As, Co, Ni, Pb; C(.): Convection of TPH, As, Co, Ni, Pb.

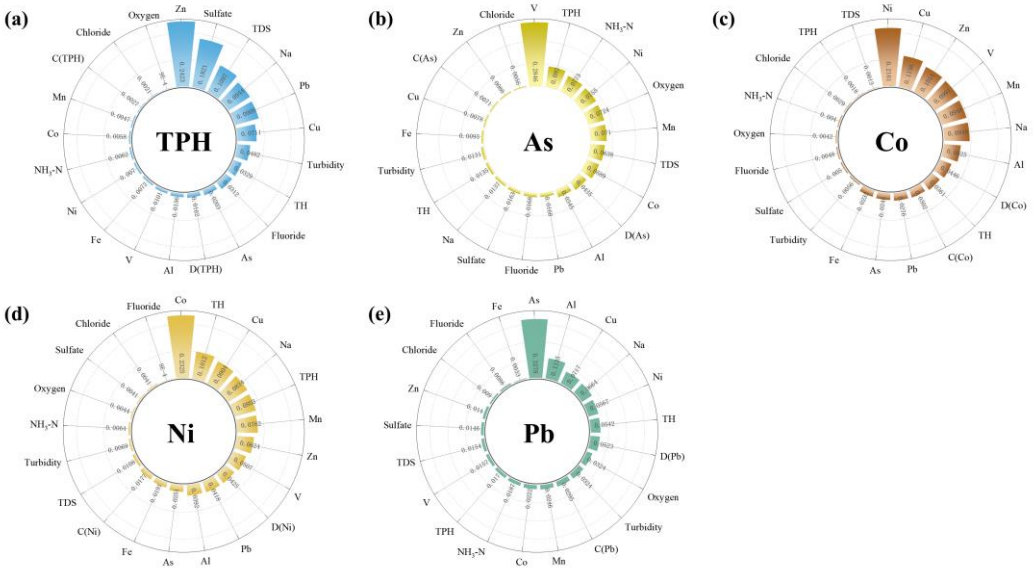

Fig. S11. Importance Indices of the Random Forest Model for TPH and HMs in pore water. TDS: Total dissolved solids; TH: Total hardness. D(.): Diffusion of TPH, As, Co, Ni, Pb; C(.): Convection of TPH, As, Co, Ni, Pb.

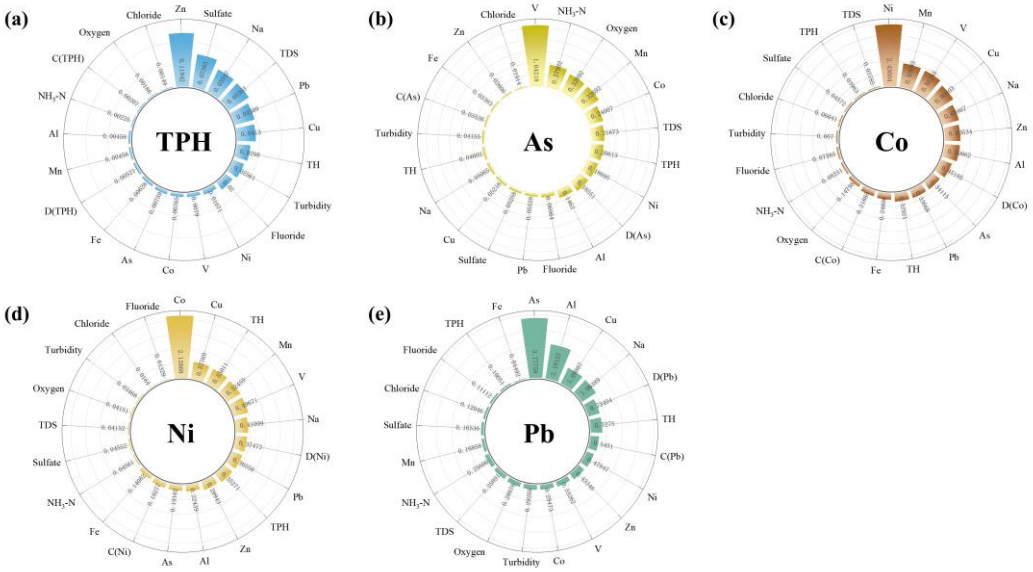

Fig. S12. SHAP value for TPH and HMs in pore water. TDS: Total dissolved solids; TH: Total hardness. D(.): Diffusion of TPH, As, Co, Ni, Pb; C(.): Convection of TPH, As, Co, Ni, Pb.

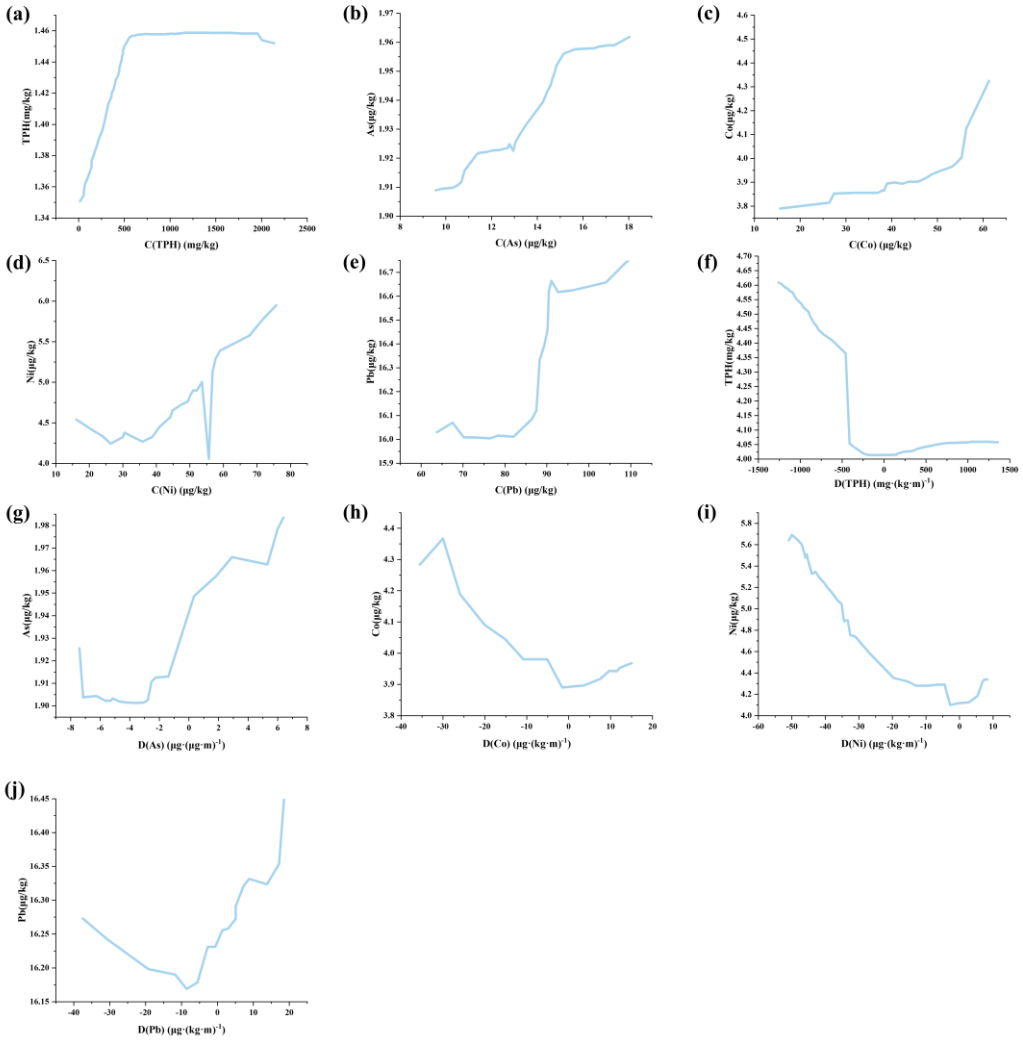

Fig. S13. The dependency plots between contaminant concentrations in perched water and convection/diffusion. (a) – (g) represent the dependency plots of TPH, As, Co, Ni and Pb on convection; (h) – (m) represent the dependency plots of TPH, As, Co, Ni and Pb on diffusion.

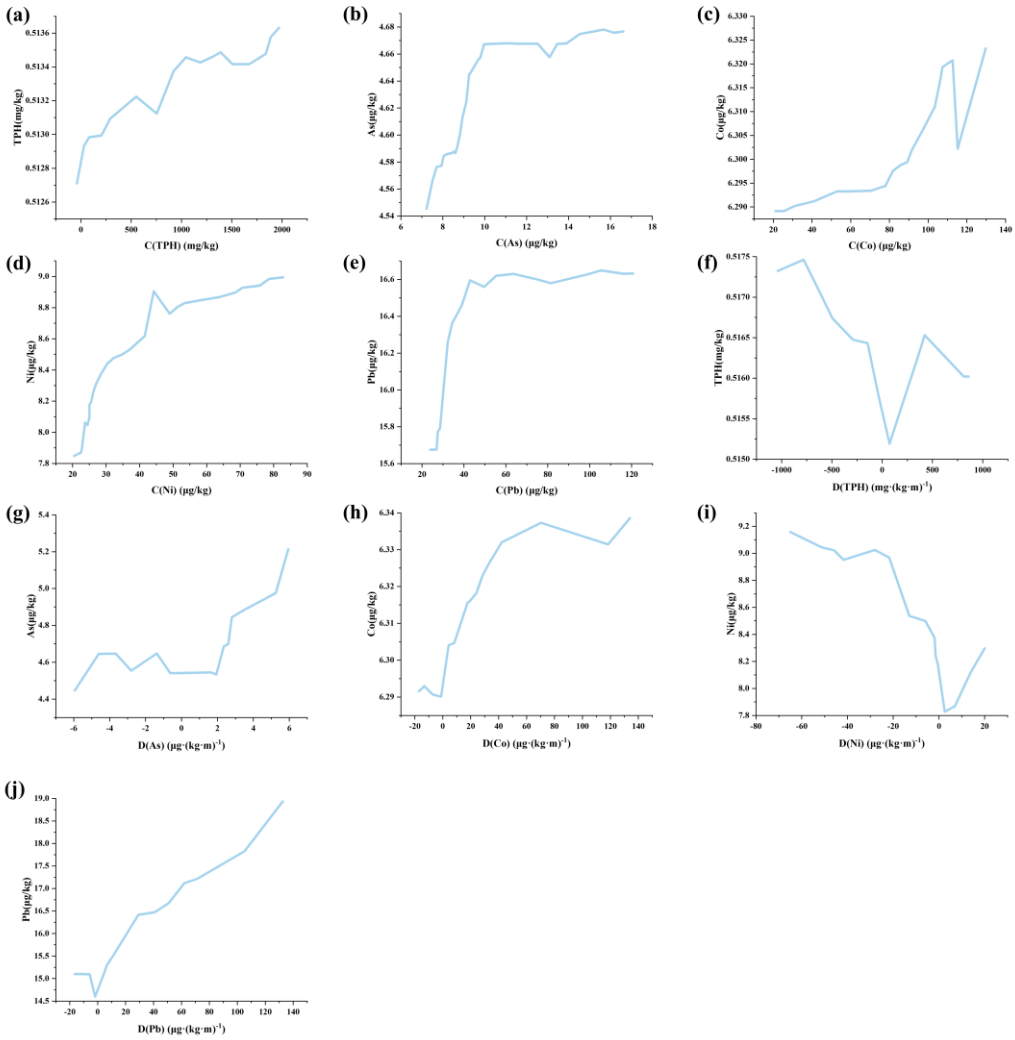

Fig. S14. The dependency plots between contaminant concentrations in pore water and convection/diffusion. (a) – (f) represent the dependency plots of TPH, As, Co, Ni and Pb on convection; (g) – (k) represent the dependency plots of TPH, As, Co, Ni and Pb on diffusion.

References

[1] P.R.C. Ministry of Ecology and Environment Soil Environmental Quality Risk Control Standard for Soil Contamination of Development Land (2018). China Beijing (Ed.)

[2] P.R.C. Ministry of Ecology and Environment. Quality Standard for Groundwater (2018). China Beijing (Ed.)

[3] P.R.C. Ministry of Ecology and Environment. Environmental Quality Standards for Surface Water (2002). China Beijing (Ed.)