
Electronic Supplementary Materials

Text S1. Calculation of the distribution coefficient

The distribution coefficient (Kd) is a critical parameter governing the transport of solutes such as heavy metals in porous media. Here, Kd values for Fe, Cu, and Zn were calculated using the analytical solution of the one-dimensional convection–diffusion equation for instantaneous pulse input, as follows:

$$c(x,t) = \frac{M}{\sqrt{4\pi D_e(t_i - t_1)}} \exp\left(-\frac{(x - u_e(t_i - t_1))^2}{4D_e(t_i - t_1)}\right), \quad (\text{S1})$$

$$R = 1 + \frac{\rho_b}{\theta} K_d, \quad (\text{S2})$$

where $c(x,t)$ is the solute concentration at position x and time t . M is the total mass released per unit cross-sectional area, D_e is the effective dispersion coefficient, u_e is the effective solute velocity, t_i and t_1 are specific times, and R is the retardation factor.

Text S2. SHAP analysis

A Kernel SHAP explainer was employed in this study, which approximates model behavior through sampling. This method is applicable to a wide range of machine learning models, including tree-based algorithms (e.g., Random Forest and XGBoost) as well as non-linear models such as SVM. Kernel SHAP was selected due to its model-agnostic nature, which does not depend on the internal structure of the model, thereby enabling consistent interpretation across heterogeneous model architectures. To ensure the explainer fully captured the model's learning behavior, the entire training dataset (103 samples) was used as the background dataset when constructing the SHAP explainer. In SHAP, the background dataset is used to simulate the model's response when features are missing, and its composition critically influences the stability and credibility of the explanations. Using all training samples as the background dataset enhances behavioral coverage and produces SHAP values that more accurately represent each feature's contribution to the model's decision-making process, thereby improving both interpretability and explanatory value.

Table S1. Optimal Model Hyperparameters and Hyperparameter Optimization Results

Heavy metal	Input parameters	Optimal model	Model parameters	RMSE	MAE	R ²	WI
Fe	GEO	RF	{ n_estimators: 420, max_depth: 11, min_samples_split: 4, min_samples_leaf: 5, max_features: None, bootstrap: True }	0.086	0.059	0.622	0.837
Fe	GEO	XGBoost	{ n_estimators: 359, max_depth: 6, learning_rate: 0.020, subsample: 0.742, colsample_bytree: 0.650, min_child_weight: 9, gamma: 0.028, reg_alpha: 0.273, reg_lambda: 1.168 }	0.088	0.063	0.611	0.834
Fe	GEO	SVM	{ kernel: poly, C: 8.405, gamma: scale, epsilon: 0.014, degree: 4 }	0.080	0.056	0.641	0.844
Fe	GEO & CH	RF	{ n_estimators: 301, max_depth: 11, min_samples_split: 2, min_samples_leaf: 1, max_features: log2, bootstrap: False }	0.070	0.054	0.729	0.873
Fe	GEO & CH	XGBoost	{ n_estimators: 490, max_depth: 4, learning_rate: 0.145, subsample: 0.981, colsample_bytree: 0.764, min_child_weight: 10, gamma: 0.001, reg_alpha: 0.019, reg_lambda: 4.790 }	0.051	0.066	0.710	0.880
Fe	GEO & CH	SVM	{ kernel: rbf, C: 3.381, gamma: auto, epsilon: 0.013, degree: 5 }	0.080	0.061	0.595	0.842
Fe	GEO & CH & AWQ	RF	{ n_estimators: 400, max_depth: 17, min_samples_split: 4, min_samples_leaf: 1, max_features: sqrt, bootstrap: False }	0.061	0.045	0.796	0.896

Fe	GEO & CH & AWQ	XGBoost	{ n_estimators: 160, max_depth: 6, learning_rate: 0.020, subsample: 0.903, colsample_bytree: 0.769, min_child_weight: 6, gamma: 0.026, reg_alpha: 0.254, reg_lambda: 1.300 }	0.048	0.062	0.805	0.899
Fe	GEO & CH & AWQ	SVM	{ kernel: rbf, C: 10.000, gamma: auto, epsilon: 0.020, degree: 1 }	0.064	0.048	0.707	0.879
Zn	GEO	RF	{ n_estimators: 150, max_depth: 10, min_samples_split: 8, min_samples_leaf: 2, max_features: log2, bootstrap: True }	0.095	0.143	0.848	0.913
Zn	GEO	XGBoost	{ n_estimators: 170, max_depth: 4, learning_rate: 0.161, subsample: 0.689, colsample_bytree: 0.992, min_child_weight: 7, gamma: 0.006, reg_alpha: 0.022, reg_lambda: 3.639 }	0.143	0.092	0.848	0.911
Zn	GEO	SVM	{ kernel: poly, C: 4.011, gamma: scale, epsilon: 0.048, degree: 3 }	0.159	0.107	0.831	0.900
Zn	GEO & CH	RF	{ n_estimators: 417, max_depth: 8, min_samples_split: 3, min_samples_leaf: 2, max_features: None, bootstrap: True }	0.140	0.092	0.863	0.921
Zn	GEO & CH	XGBoost	{ n_estimators: 300, max_depth: 6, learning_rate: 0.298, subsample: 0.795, colsample_bytree: 0.964, min_child_weight: 7, gamma: 0.001, reg_alpha: 0.282, reg_lambda: 4.080 }	0.079	0.109	0.928	0.950
Zn	GEO & CH	SVM	{ kernel: rbf, C: 4.876, gamma: scale, epsilon:	0.129	0.095	0.903	0.940

			0.010, degree: 4 }				
Zn	GEO & CH & AWQ	RF	{ n_estimators: 330, max_depth: 14, min_samples_split: 2, min_samples_leaf: 5, max_features: None, boot- strap: False }	0.062	0.099	0.896	0.936
Zn	GEO & CH & AWQ	XGBoost	{ n_estimators: 55, max_depth: 7, learn- ing_rate: 0.238, subsample: 0.783, colsample_bytree: 0.897, min_child_weight: 9, gamma: 0.000, reg_alpha: 0.479, reg_lambda: 4.359 }	0.117	0.088	0.914	0.943
Zn	GEO & CH & AWQ	SVM	{ kernel: rbf, C: 2.824, gamma: scale, epsilon: 0.022, degree: 5 }	0.096	0.075	0.949	0.964
Cu	GEO	RF	{ n_estimators: 90, max_depth: 14, min_samples_split: 9, min_samples_leaf: 1, max_features: log2, boot- strap: True }	0.309	0.191	0.689	0.858
Cu	GEO	XGBoost	{ n_estimators: 330, max_depth: 10, learn- ing_rate: 0.151, subsample: 0.637, colsample_bytree: 0.612, min_child_weight: 2, gamma: 0.001, reg_alpha: 0.230, reg_lambda: 3.990 }	0.189	0.308	0.692	0.860
Cu	GEO	SVM	{ kernel: poly, C: 8.240, gamma: scale, epsilon: 0.021, degree: 5 }	0.311	0.184	0.716	0.865
Cu	GEO & CH	RF	{ n_estimators: 263, max_depth: 12, min_samples_split: 8, min_samples_leaf: 2, max_features: None, boot- strap: False }	0.308	0.197	0.702	0.867

Cu	GEO & CH	XGBoost	{ n_estimators: 220, max_depth: 4, learning_rate: 0.025, subsample: 0.669, colsample_bytree: 0.660, min_child_weight: 4, gamma: 0.000, reg_alpha: 0.348, reg_lambda: 4.630 }	0.207	0.336	0.659	0.854
Cu	GEO & CH	SVM	{ kernel: linear, C: 2.806, gamma: auto, epsilon: 0.032, degree: 4 }	0.368	0.259	0.572	0.836
Cu	GEO & CH & AWQ	RF	{ n_estimators: 402, max_depth: 18, min_samples_split: 9, min_samples_leaf: 4, max_features: None, bootstrap: True }	0.099	0.062	0.896	0.936
Cu	GEO & CH & AWQ	XGBoost	{ n_estimators: 195, max_depth: 4, learning_rate: 0.068, subsample: 0.691, colsample_bytree: 0.941, min_child_weight: 4, gamma: 0.027, reg_alpha: 0.015, reg_lambda: 8.631 }	0.263	0.176	0.773	0.891
Cu	GEO & CH & AWQ	SVM	{ kernel: linear, C: 4.956, gamma: auto, epsilon: 0.037, degree: 5 }	0.282	0.189	0.734	0.878

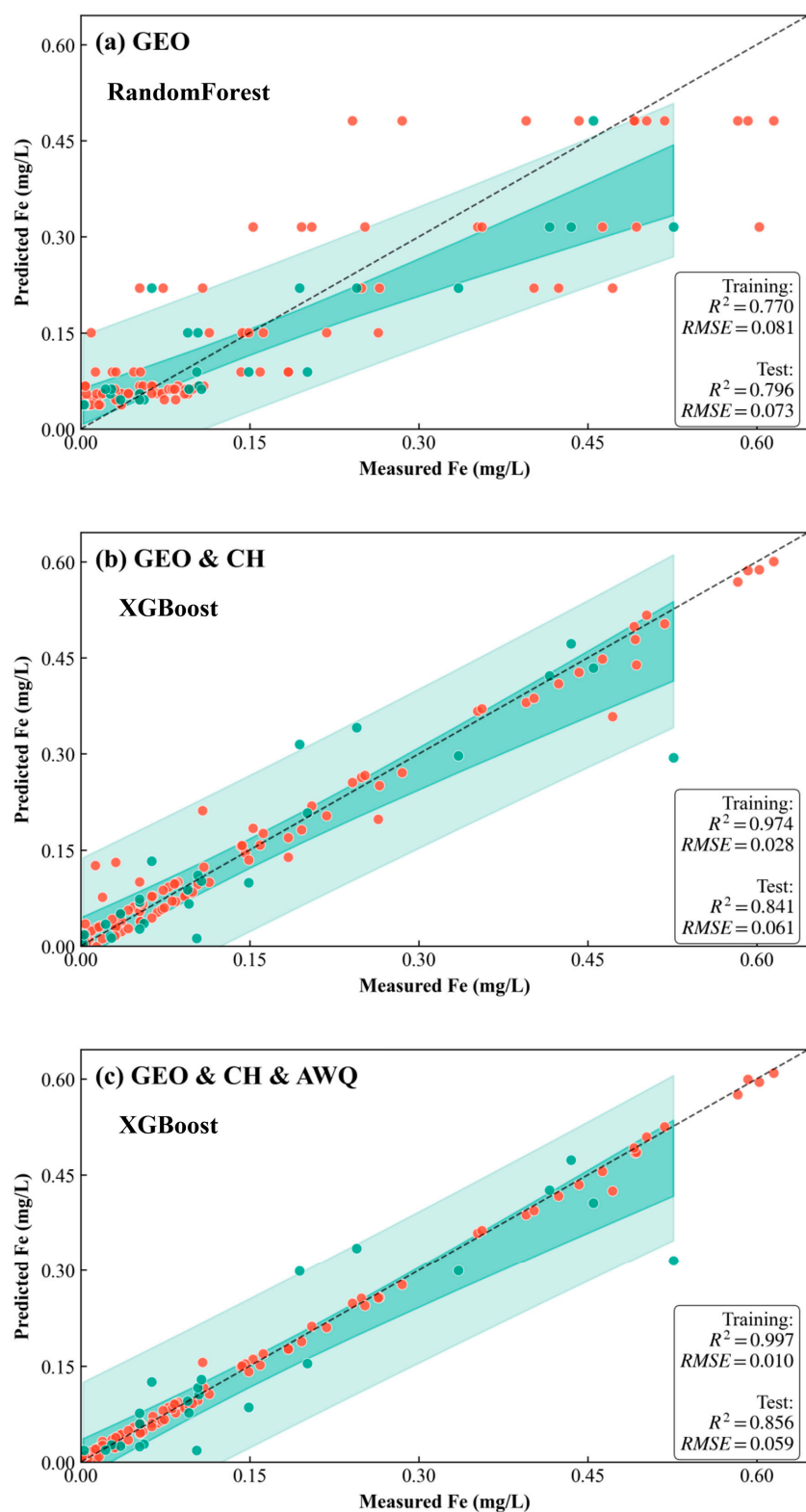


Figure S1. Predicted versus measured concentrations for Fe (a, b, c) with prediction models exhibiting the highest R^2 values. Red points indicate training datasets, green points indicate test datasets, the black line is the 1:1 prediction line, light green shading shows the 95% prediction confidence interval, and dark green shading shows the 95% prediction confidence interval.

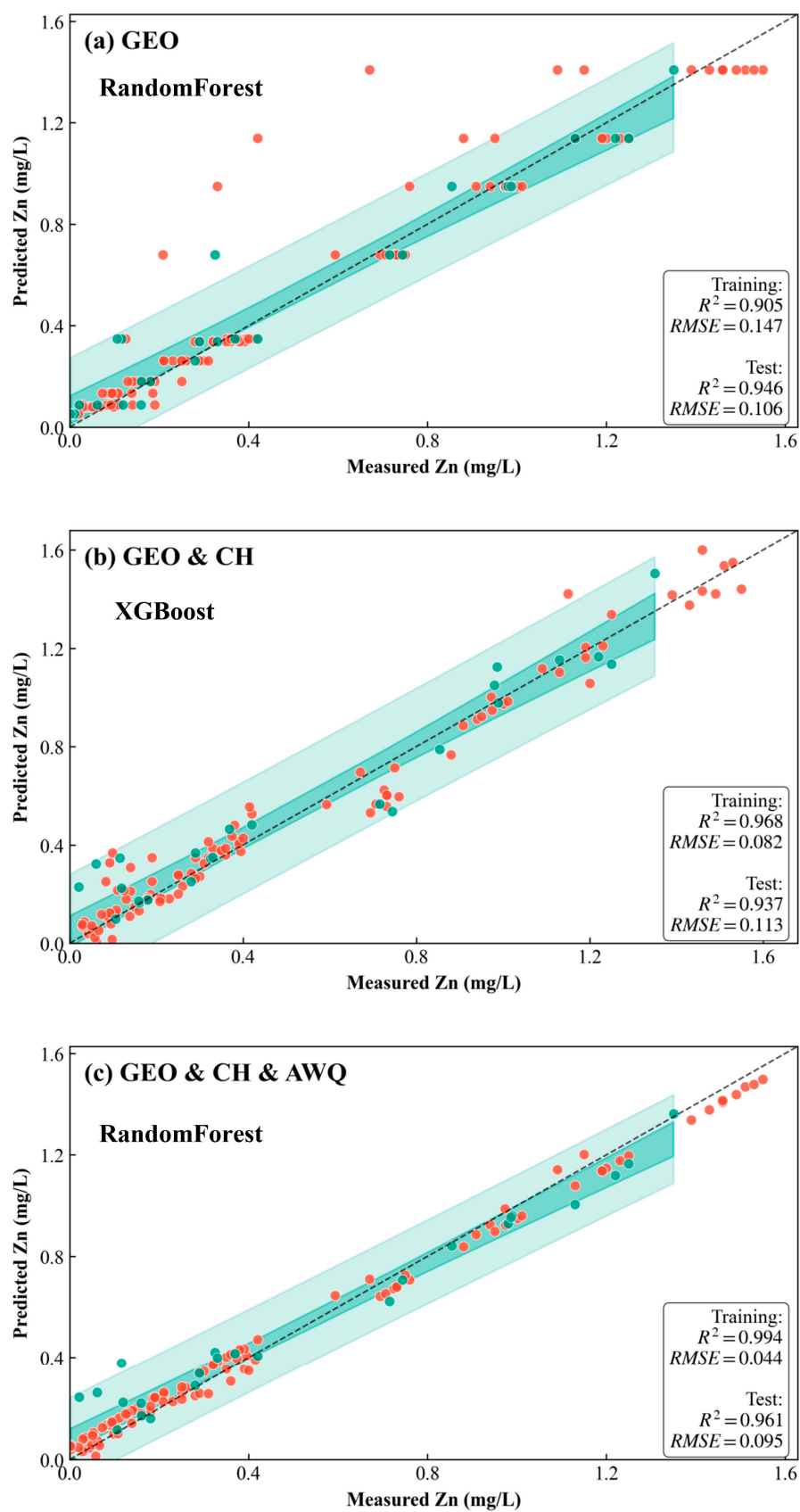


Figure S2. Predicted versus measured concentrations for Zn (a, b, c) with prediction models exhibiting the highest R^2 values.

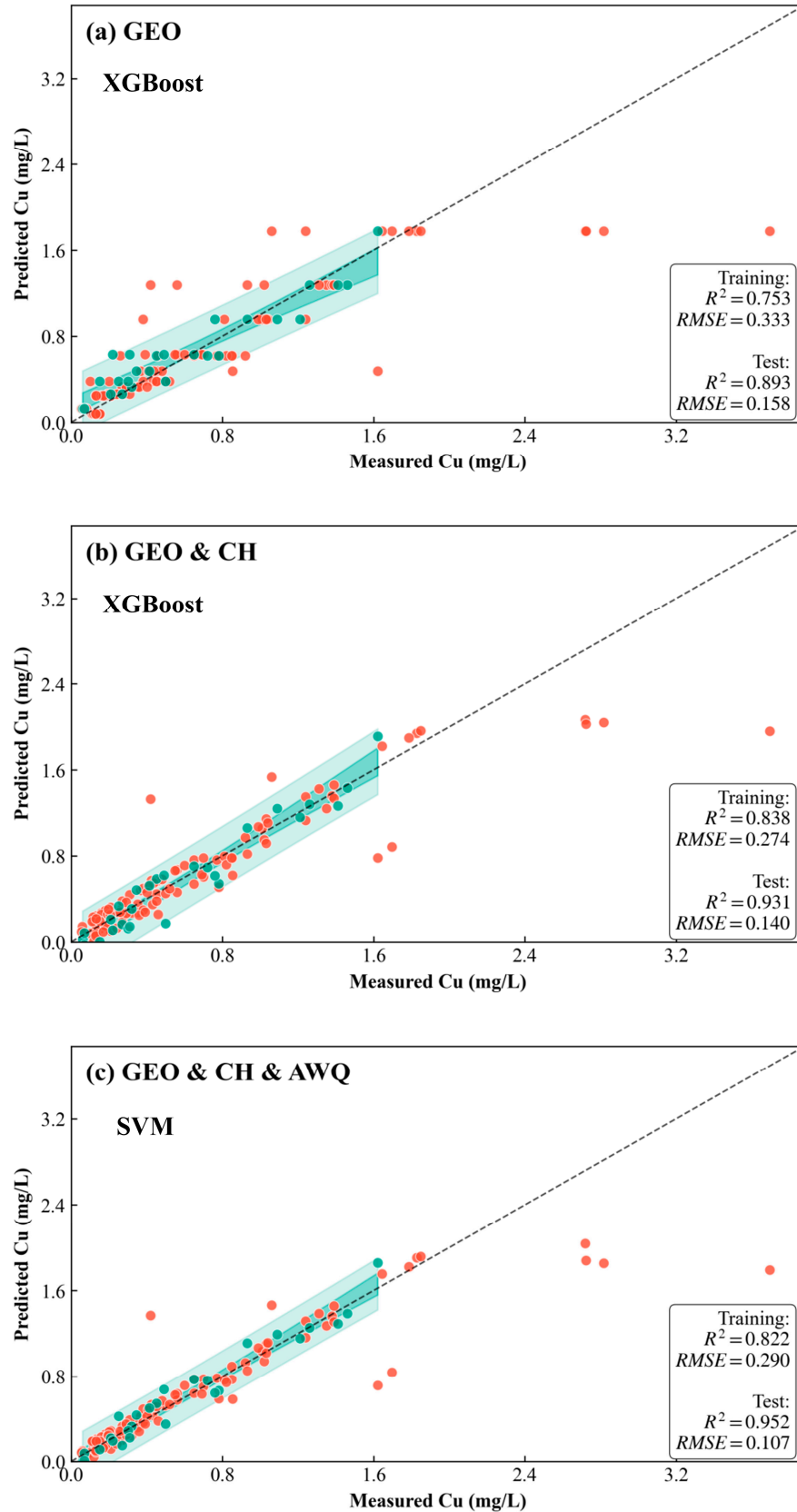


Figure S3. Predicted versus measured concentrations for Cu (a, b, c) with prediction models exhibiting the highest R^2 values.

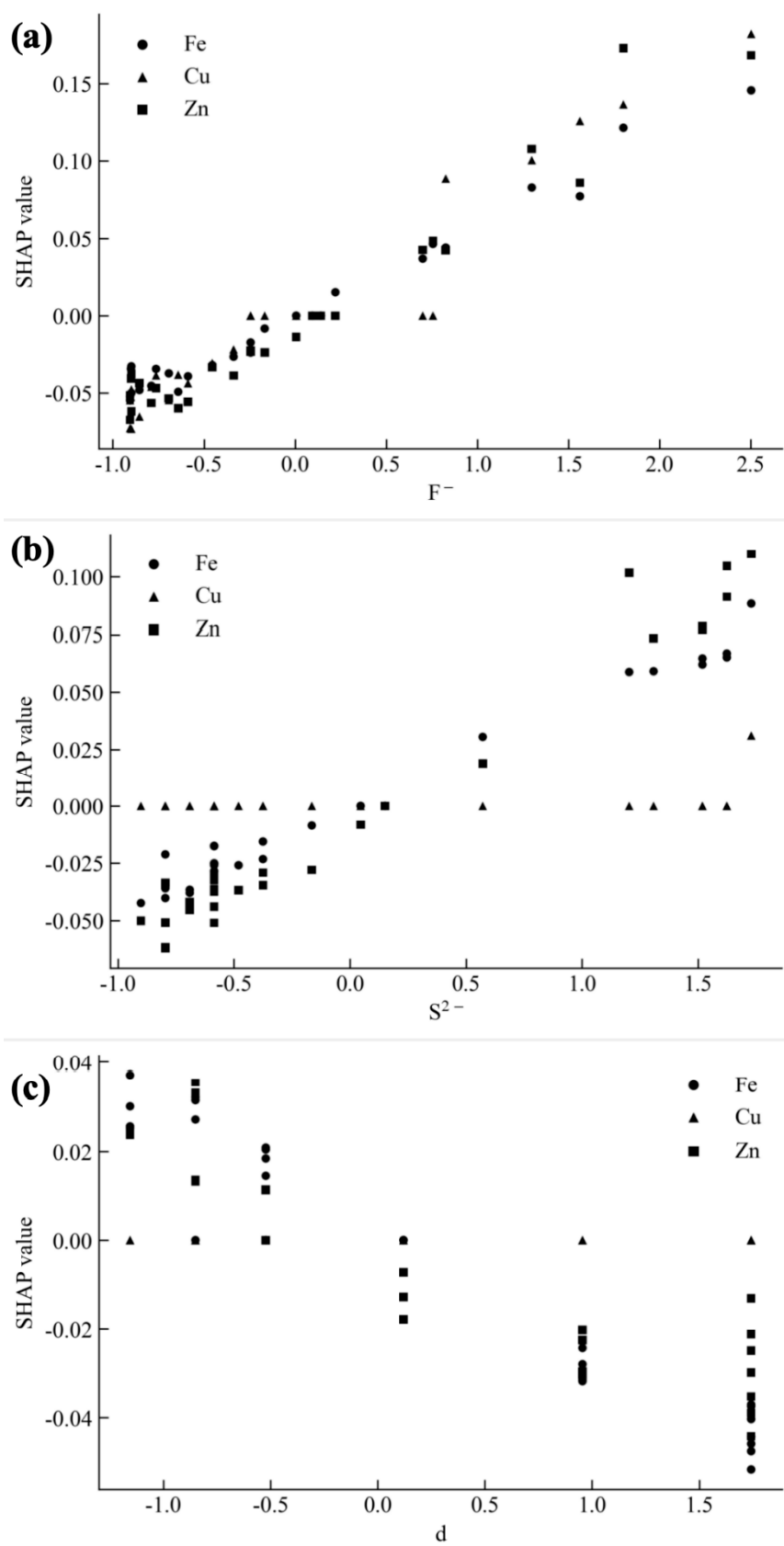


Figure S4. SHAP single analysis between (a) F^- and Fe, Cu, Zn; (b) S^{2-} and Fe, Cu, Zn; and (c) distance and Fe, Cu, Zn

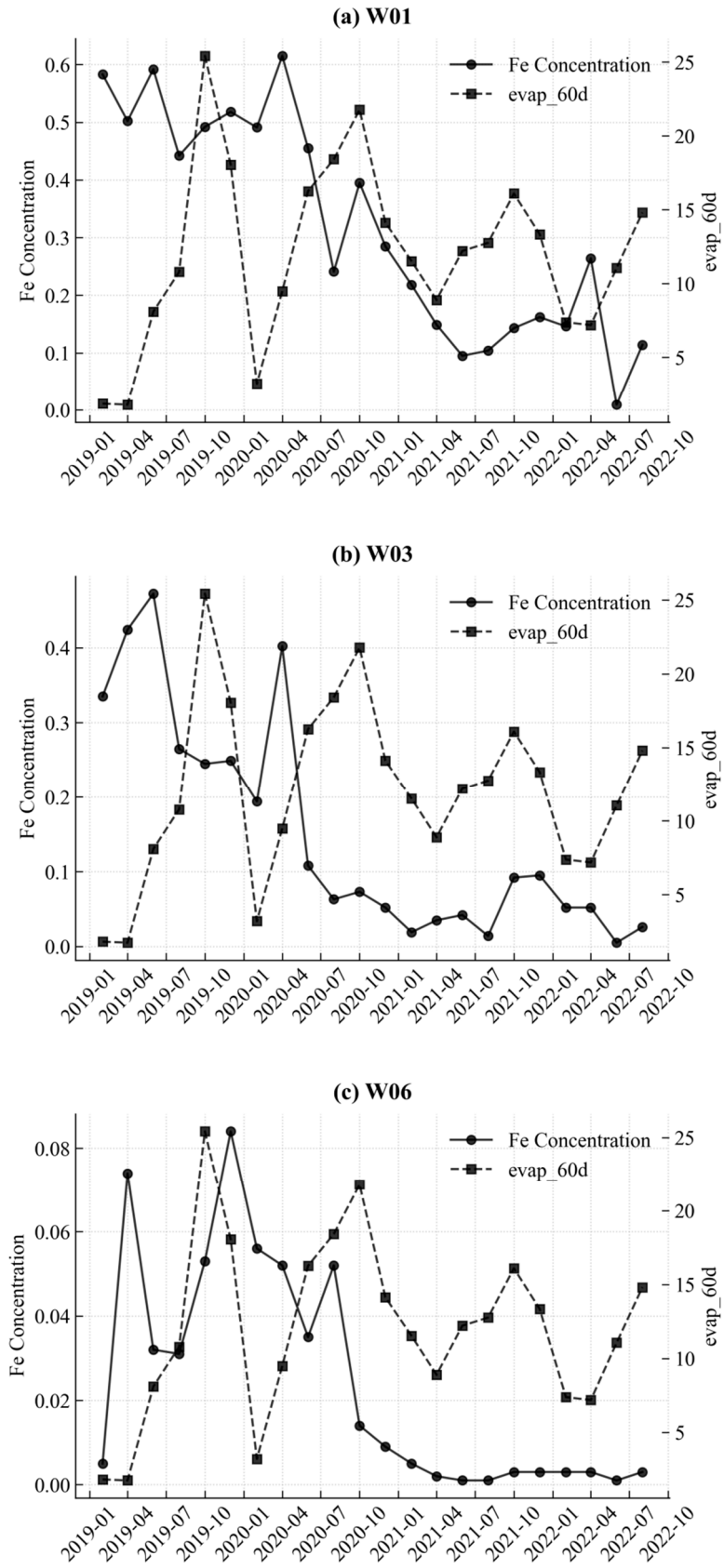


Figure S5. Temporal trends in Fe concentrations and cumulative evaporation over 60-day windows.

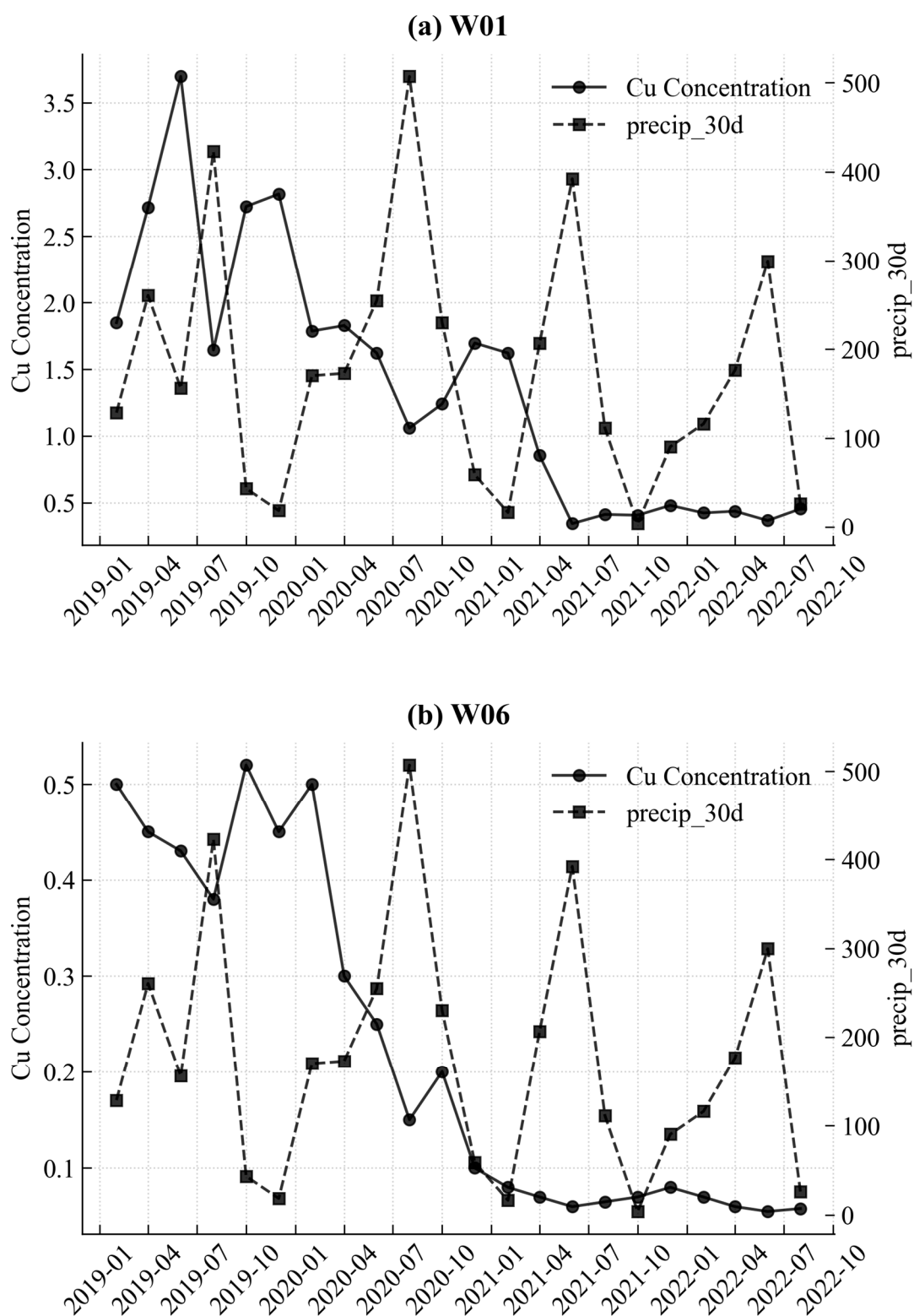


Figure S6. Temporal trends in Cu concentrations and cumulative precipitation over 30-day windows.

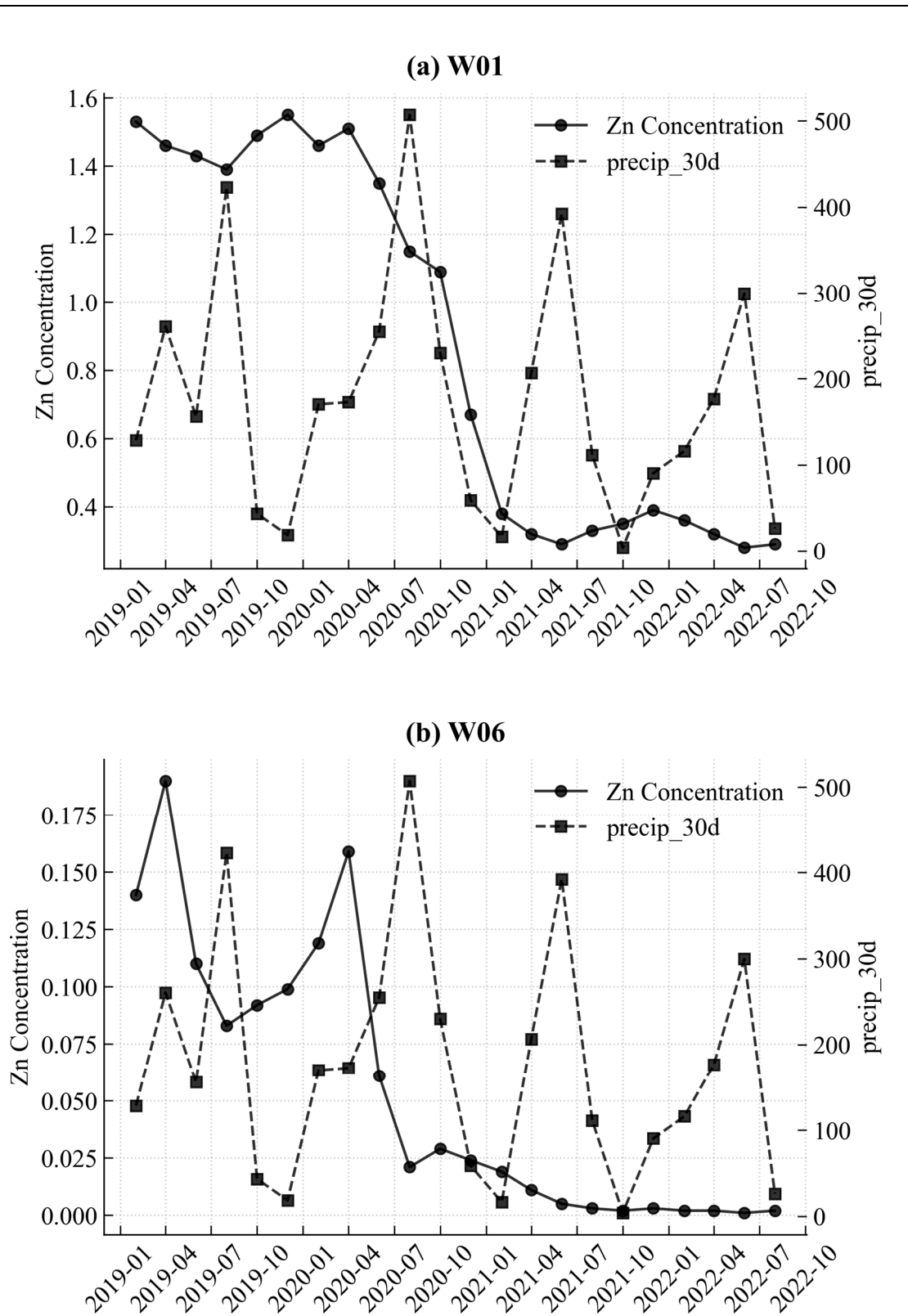


Figure S7. Temporal trends in Zn concentrations and cumulative precipitation over 30-day windows.

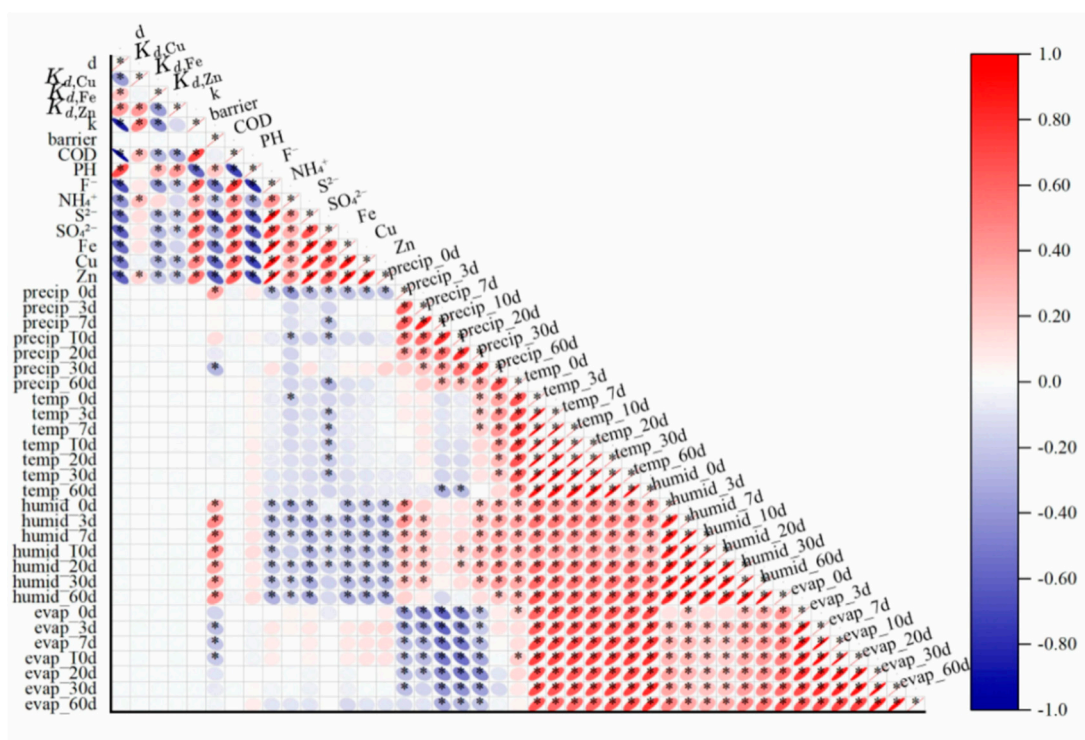


Figure S8. Pearson correlation coefficients for 37 parameters within the GEO, CH, AWQ datasets and the 3 target metal concentrations (*: p value < 0.05).