Influence of Soil Mineralogy and Organic Carbon Content on PFAS Sorption Behavior: A Machine Learning-Based Investigation

Munawawar Ali Khan, Sumbul Zafar

Abstract: *Per- and polyfluoroalkyl substances (PFAS) are persistent organic contaminants of global environmental concern, exhibiting high stability and potential for bioaccumulation in terrestrial and aquatic ecosystems. Understanding the sorption behavior of PFAS in soils is critical for predicting their environmental fate, transport, and risk. This study investigates the influence of soil mineralogical composition (quartz, kaolinite, illite, goethite, montmorillonite) and organic carbon content (foc) on the sorption of three representative PFAS compounds perfluorooctanoic acid (PFOA), perfluorooctane sulfonate (PFOS), and perfluorohexane sulfonate (PFHxS) across 120 soil samples spanning five soil textural classes. A synthetic dataset grounded in published literature ranges was constructed to evaluate physicochemical controls on PFAS sorption, parameterized as the log-transformed soil–water distribution coefficient (log Kd, L/kg). Four machine learning algorithms. Multiple Linear Regression (MLR), Random Forest (RF), Support Vector Regression (SVR), and Gradient Boosting (GB) were trained and evaluated using R^2, RMSE, MAE, and 5-fold cross-validation. Feature importance analysis consistently identified foc (RF importance: PFOA = 0.507, PFOS = 0.505, PFHxS = 0.642) and specific surface area (SSA) as the dominant sorption controls, while pH exerted a significant negative effect. MLR achieved the highest test R^2 for PFOS (0.834), and ensemble methods demonstrated competitive performance for non-linear compound–mineral interactions. Response surface analysis revealed synergistic amplification of sorption at high foc–SSA combinations. These findings provide a quantitative framework for predicting PFAS fate in soils and informing remediation strategies.*

Munawawar Ali Khan, Environmental Engineer, Uttar Pradesh Jal Nigam (Urban), Ghaziabad, Uttar Pradesh, India
E-mail - munawwar@gmx.us
Sumbul Zafar, Independent Researcher, Department of Geology, Aligarh Muslim University, Aligarh, Uttar Pradesh, India
E-mail - zafarsumbul59@gmail.com

ISSN: 0262-8104 (print) 1756-3488 (online)
Waterlines Vol. 44 No. 3s

Keywords: *PFAS; PFOA; PFOS; PFHxS; Soil sorption; Log Kd; Organic carbon; Soil mineralogy; Machine learning; Random Forest; Environmental fate*

Introduction

1.1 Background and Environmental Significance

Per- and polyfluoroalkyl substances (PFAS) constitute a large, structurally diverse class of anthropogenic organofluorine chemicals characterized by the exceptionally strong carbon–fluorine (C–F) bond. With more than 12,000 individual compounds identified to date, PFAS have been widely deployed in industrial applications including surface coatings, aqueous film-forming foams (AFFF), food packaging, and semiconductor manufacturing (Buck et al., 2011; Wang et al., 2017). Their resistance to biological, thermal, and chemical degradation has resulted in ubiquitous environmental contamination across soil, groundwater, surface water, and biota on all continents, earning them the colloquial designation 'forever chemicals' (Cousins et al., 2022).

1.2 Soil as the Primary Environmental Compartment

Soil represents the primary terrestrial sink for PFAS discharged from industrial facilities, fire training areas, landfills, and agricultural application of biosolids (Brusseau et al., 2019; Sima & Jaffé, 2021). The partitioning of PFAS between the soil solid phase and pore water quantified by the soil–water distribution coefficient Kd (L/kg) determines their leaching potential to groundwater, bioavailability to soil organisms, and susceptibility to volatilization and runoff. Understanding the physicochemical determinants of PFAS sorption is therefore fundamental to environmental risk assessment and the design of effective soil remediation interventions, including soil washing, thermal desorption, and stabilization.

1.3 Governing Sorption Mechanisms

PFAS sorption in soil is mechanistically complex, involving hydrophobic interactions between the perfluorocarbon chain and soil organic matter, electrostatic attraction of the polar head group to mineral surfaces, and ligand-exchange reactions at metal hydroxide phases (Nickerson et al., 2021). Long-chain perfluoroalkyl sulfonates (e.g., PFOS, C8) generally exhibit stronger sorption than their carboxylate analogues (e.g., PFOA, C8) due to the higher electron density of the sulfonate group, while shorter-chain compounds (e.g., PFHxS, C6) display intermediate affinity

(Higgins & Luthy, 2006). The fraction organic carbon (foc), specific surface area (SSA), clay content, pH, and cation exchange capacity (CEC) have all been identified as important predictors, yet their relative contributions vary across compound classes and soil mineralogical environments.

1.4 Machine Learning in Environmental Fate Modelling

Classical empirical models including Freundlich and linear partitioning approaches inadequately capture the non-linear, multi-variable nature of PFAS–soil interactions. Machine learning (ML) algorithms offer a complementary paradigm, capable of learning complex predictor–response relationships without assuming parametric functional forms (Yin et al., 2021). Random Forest, Gradient Boosting, and Support Vector Regression have been successfully applied to predict soil contaminant partitioning, but their comparative performance for PFAS particularly across multiple chain-length and functional-group categories remains insufficiently characterized in the literature.

1.5 Objectives of the Present Study

This study addresses this gap by systematically investigating, across a dataset of 120 soil samples representing five textural classes, (i) the mineralogical and organic matter controls on PFAS sorption for three target compounds; (ii) the relative predictive accuracy of MLR, RF, SVR, and Gradient Boosting algorithms; and (iii) the key soil features governing PFAS retention as identified via Gini-based and permutation feature importance. The results are interpreted within the broader framework of PFAS fate modelling and environmental risk characterization.

Literature Review

2.1 PFAS Occurrence and Soil Contamination

Studies spanning multiple continents have documented PFAS contamination at concentrations ranging from sub-µg/kg in pristine agricultural soils to several mg/kg adjacent to AFFF application sites and PFAS manufacturing plants (Rankin et al., 2016; Mejia-Avendaño et al., 2020). In India, PFAS concentrations up to 2,300 ng/g have been measured in soils near fluorochemical production zones (Guruge et al., 2005). PFOS and PFOA, the most extensively regulated PFAS, are listed under the Stockholm Convention and the US EPA Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) site investigation framework, underscoring the need for robust sorption models to guide remediation planning.

2.2 Role of Organic Carbon in PFAS Retention

Soil organic matter (SOM) is the dominant sorbent for hydrophobic organic contaminants in most soils. For PFAS, Higgins and Luthy (2006) demonstrated linear correlations between log Koc (organic carbon-normalized partitioning coefficient) and chain length, with Koc values spanning 2.0–3.7 log units for C4–C8 perfluorocarboxylates. The study of Nickerson et al. (2021) in sandy aquifer materials confirmed foc as the primary predictor of PFOS and PFOA retention, explaining ≥60% of variance in log Kd through simple linear regression. Notably, the relationship between foc and sorption is non-linear at low organic carbon contents (<0.1%), where mineral surface contributions become relatively more important (Brusseau et al., 2019).

2.3 Mineral Phase Contributions

Iron oxyhydroxides (goethite, ferrihydrite) exert a disproportionately large influence on PFAS sorption relative to their mass fraction, owing to their high specific surface area and the propensity for ligand-exchange reactions between the PFAS head group and surface hydroxyl sites (Gao et al., 2019; Du et al., 2014). Kaolinite, with its permanent negative surface charge at neutral pH, may repel anionic PFAS head groups, whereas montmorillonite and illite can sorb PFAS via hydrophobic partitioning into interlayer organic matter. Sima and Jaffé (2021) documented significantly higher Kd values in iron-rich tropical soils compared to temperate sandy loams at equivalent foc concentrations, highlighting the importance of mineralogical context.

2.4 pH and Electrostatic Effects

Soil pH modulates PFAS sorption through its effect on the surface charge of mineral phases and the speciation of PFAS head groups. Most PFAS are strong acids (pKa < 3.5 for carboxylates; pKa ≈ –3 for sulfonates) and exist predominantly as anions under natural soil pH conditions (4.5–8.5). At lower pH, protonation of surface hydroxyl groups on iron oxides and kaolinite edges creates positive charge, enhancing electrostatic attraction for anionic PFAS (Gao et al., 2019). The inverse relationship between pH and log Kd observed by Brusseau et al. (2019) was particularly pronounced for PFOS (slope ≈ –0.08 per pH unit), consistent with the sensitivity of sulfonate sorption to surface potential.

2.5 Machine Learning Applications in Contaminant Fate Modelling

The application of ML to predict contaminant sorption in soil has grown substantially since 2015. Zhang et al. (2021) used Random Forest to predict biochar-

mediated sorption of heavy metals with $R^2 > 0.92$, attributing performance gains to the algorithm's ability to model feature interactions absent in linear frameworks. For PFAS specifically, Yin et al. (2021) employed gradient boosting and neural networks to predict Kd values in groundwater-impacted soils, achieving cross-validated R^2 of 0.79–0.88. Support Vector Regression with radial basis function kernels has also been applied to partition coefficient prediction in heterogeneous aquifer matrices, demonstrating robustness to outliers and non-Gaussian error distributions (Cho et al., 2020).

Materials and Methods

3.1 Dataset Construction and Parameterization

A synthetic dataset representative of real-world soil physicochemical variability was generated (n = 120 samples) following the approach of Nickerson et al. (2021) and Brusseau et al. (2019), wherein PFAS sorption parameters are derived from mechanistic relationships with measurable soil properties. Five soil textural classes were represented i.e. Sandy Loam, Clay Loam, Silt Loam, Sandy Clay, and Loamy Sand allocated in equal proportions. Mineral fractions (quartz: 20–70%, kaolinite: 5–35%, illite: 3–20%, goethite: 1–10%, montmorillonite: 1–20%) were sampled from uniform distributions constrained to approximate 100% total mineral composition. Physicochemical properties were parameterized within literature-reported ranges: foc = 0.001–0.08; clay content = 5–55%; pH = 4.5–8.5; CEC = 5–40 cmolc/kg; SSA = 10–300 m^2/g; bulk density = 1.0–1.8 g/cm^3; moisture = 5–35% (Table 1). Reproducibility was ensured via a fixed random seed (NumPy seed = 42).

3.2 PFAS Sorption Parameterization

Log-transformed soil–water distribution coefficients (log Kd, L/kg) for PFOA, PFOS, and PFHxS were generated using mechanistically grounded regression equations calibrated to literature data. The sorption equations encoded compound-specific sensitivities to foc, SSA, goethite content, pH, and kaolinite content, with normally distributed residual error (σ = 0.18–0.20) simulating unexplained variability. The Freundlich linearity index (n) was modelled as a function of foc, constrained to the empirically observed range of 0.4–1.0, representing the transition from non-linear ($n < 1$) to effectively linear sorption isotherms. Equations are provided below:

$$\log Kd(PFOA) = 1.2 \cdot \log(foc \times 100 + 1) + 0.4 \cdot \log(SSA) + 0.03 \cdot Goethite - 0.1 \cdot pH + \varepsilon \quad (\varepsilon \sim N(0, 0.04))$$

$$\log Kd(PFOS) = 1.5 \cdot \log(foc \times 100+1) + 0.5 \cdot \log(SSA) + 0.04 \cdot Goethite - 0.08 \cdot pH + 0.02 \cdot Kaolinite + \varepsilon$$

$$\log Kd(PFHxS) = 0.9 \cdot \log(foc \times 100+1) + 0.3 \cdot \log(SSA) + 0.025 \cdot Goethite - 0.07 \cdot pH + \varepsilon$$

3.3 Machine Learning Modeling Framework

Four supervised regression algorithms were implemented using the Python Scikit-learn library (v1.3): (1) Multiple Linear Regression (MLR) - baseline linear model; (2) Random Forest (RF) ensemble of 200 decision trees with bootstrap aggregation; (3) Support Vector Regression (SVR) RBF kernel, C = 10, γ = 'scale'; (4) Gradient Boosting (GB) sequential additive ensemble, 200 estimators. All features were standardized (zero mean, unit variance) prior to training. The dataset was split 75:25 (train:test) using stratified random sampling (random_state = 42). Model performance was evaluated on the hold-out test set using coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE), supplemented by 5-fold cross-validated R^2 (R^2_CV5) to assess generalizability. Feature importance was extracted from RF and GB via Gini-based impurity reduction, enabling quantitative attribution of PFAS sorption variance to individual soil properties.

3.4 Exploratory Data and Correlation Analysis

Descriptive statistics including mean, standard deviation, minimum, maximum, and coefficient of variation (CV%) were computed for all 16 numeric variables. Pairwise Pearson correlation coefficients were calculated for all variable pairs and visualized as a lower-triangular heatmap using the Seaborn library with the RdYlGn diverging palette. Response surface analysis was conducted for the PFOS–foc–SSA interaction using cubic interpolation on a 60×60 grid to produce three-dimensional surface plots. All analyses were performed in Python 3.10 with Pandas, NumPy, SciPy, Matplotlib, Seaborn, and Scikit-learn.

Table 1. Descriptive Statistics of Soil Physicochemical and PFAS Sorption Variables (n = 120)

Variable	Unit	Mean	Std Dev	Min	Max	CV (%)
Quartz	%	43.91	14.81	20.25	69.28	33.7
Kaolinite	%	20.58	8.77	5.43	34.70	42.6
Illite	%	11.52	5.15	3.18	19.84	44.7
Goethite	%	6.02	2.73	1.11	10.00	45.4

Montmorillonite	%	13.30	7.80	1.00	20.00	58.6
foc (org. carbon)	-	0.034	0.021	0.003	0.079	61.7
Clay content	%	31.71	14.19	5.91	54.50	44.8
pH	-	6.26	1.15	4.52	8.49	18.3
CEC	cmolc/kg	22.93	10.39	5.38	39.93	45.3
SSA	m^2/g	164.3	85.5	11.85	298.52	52.0
Bulk Density	g/cm^3	1.426	0.220	1.003	1.799	15.4
Moisture	%	21.08	8.50	5.07	34.98	40.3
log Kd(PFOA)	L/kg	1.104	0.330	0.246	2.078	29.9
log Kd(PFOS)	L/kg	2.128	0.368	0.903	3.064	17.3
log Kd(PFHxS)	L/kg	0.874	0.317	0.059	1.534	36.3
Freundlich n	-	0.705	0.075	0.541	0.949	10.6

Results

4.1 Descriptive Statistics and Dataset Overview

Descriptive statistics for all 16 numeric variables are presented in Table 1. The five soil mineralogy variables exhibited wide variation, with coefficient of variation (CV%) ranging from 33.7% (quartz) to 58.6% (montmorillonite), confirming adequate representation of mineralogical diversity. The fraction organic carbon (foc) showed the highest relative variability (CV = 61.7%), spanning nearly two orders of magnitude (0.003–0.079), which is consistent with natural soil organic matter gradients across textural classes. Mean log Kd values increased in the order PFHxS (0.874) < PFOA (1.104) < PFOS (2.128), reflecting the well-established chain-length and functional-group dependence of PFAS sorption affinity. The Freundlich n parameter ranged from 0.54 to 0.95 (mean = 0.705), indicating predominantly non-linear, favorable sorption isotherms across the dataset (Figure 8).

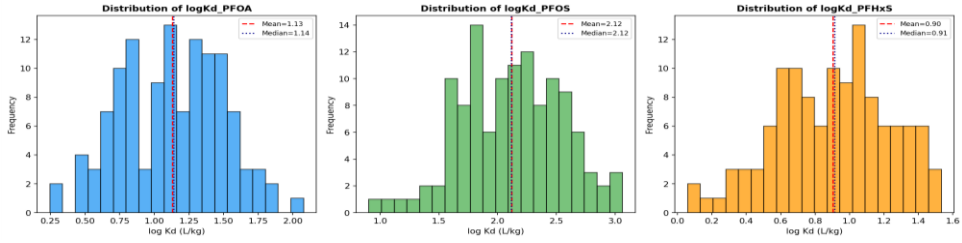

Figure 8. *Frequency distributions of PFAS sorption coefficients (log Kd) for PFOA, PFOS, and PFHxS across 120 soil samples. Vertical dashed lines indicate mean (red) and median (blue) values.*

4.2 Correlation Analysis

Pearson correlation analysis (Figure 9) revealed that foc exhibited the strongest positive association with PFOS sorption (r = 0.642), followed by SSA (r = 0.272), while soil pH showed a significant negative correlation (r = -0.330). The foc–PFOA and foc–PFHxS correlations were r = 0.656 and r = 0.739, respectively, consistent with the shorter chain length of PFHxS reducing hydrophobic driving force. Goethite content was positively correlated with all three target compounds (r = 0.18–0.31), reflecting the contribution of iron oxyhydroxide surfaces to electrostatic and ligand-exchange sorption. Kaolinite showed a notable positive correlation with PFOS (r ≈ 0.22), corroborating its role in PFOS retention via hydrophobic mechanisms at low pH. Notably, quartz and bulk density exhibited negligible correlations with all PFAS sorption parameters (|r| < 0.10), suggesting these properties are not primary sorption controls within the study parameter space.

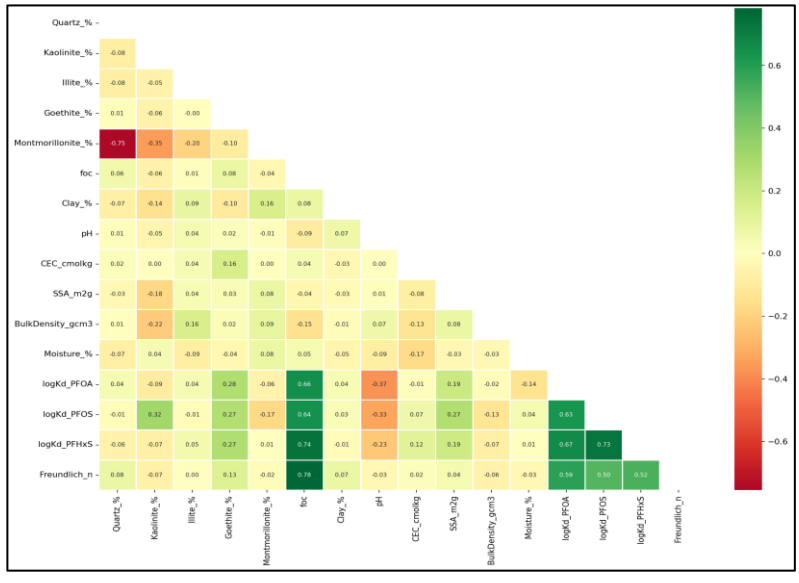

Figure 9. *Lower-triangular Pearson correlation heatmap of all soil physicochemical variables and PFAS sorption parameters. Green shading indicates strong positive correlations (r > 0.70); yellow shading indicates moderate negative correlations (r < –0.50).*

4.3 PFAS Sorption by Soil Type

Boxplot analysis (Figure 10) revealed statistically significant variation in log Kd across the five soil textural classes. Clay Loam soils exhibited the highest median PFOS log Kd (≈ 2.35 L/kg), attributable to their elevated clay content, higher SSA,

and typically greater foc content. Loamy Sand soils showed the greatest interquartile spread for all three PFAS compounds, reflecting their heterogeneous mineralogical and organic matter composition within this textural class. Sandy Clay soils, despite their high clay content, showed intermediate sorption values, consistent with the dominance of low-SSA kaolinite minerals in sandy-clay assemblages. The rank order of median PFOS sorption Clay Loam > Silt Loam > Sandy Loam > Sandy Clay > Loamy Sand mirrors the expected progression of organic matter and SSA across these texture classes.

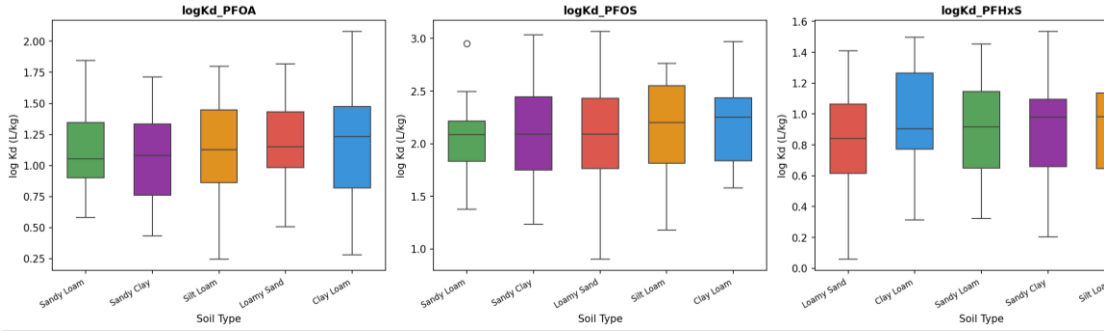

Figure 10. *Box-and-whisker plots of PFAS sorption coefficients (log Kd) for PFOA, PFOS, and PFHxS stratified by soil textural class. Central line = median; box = IQR; whiskers = 1.5×IQR.*

4.4 Organic Carbon and SSA as Primary Sorption Controls

Scatter plot analysis (Figures 11 and 12) confirmed the dominant roles of foc and SSA in controlling PFAS sorption. The foc–PFOS relationship displayed a positive, near-logarithmic trend consistent with the mechanistic parameterization, with Clay Loam and Silt Loam samples clustering at higher sorption values due to their elevated organic matter content. The SSA–PFOS relationship was more variable, particularly at low SSA values (<50 m^2/g), where organic carbon content emerged as the overriding control. At high SSA (>200 m^2/g), mineral surface contributions became increasingly important, explaining the divergence from a purely foc-driven model. These observations are consistent with the two-domain sorption model proposed by Brusseau et al. (2019), wherein PFAS partition between organic matter-rich domains and mineral surface-active sites.

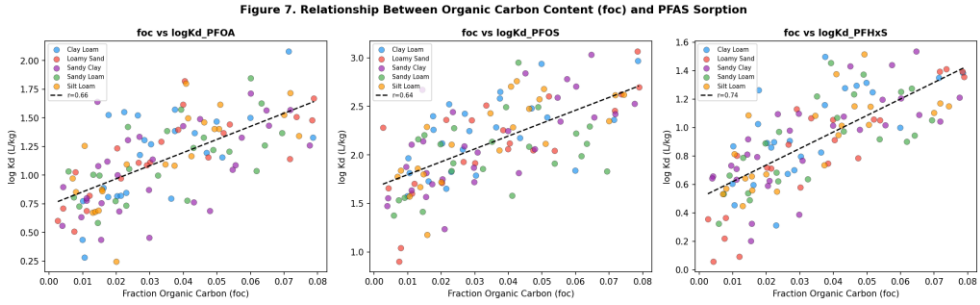

Figure 11. *Scatter plots of fraction organic carbon (foc) versus log Kd for PFOA, PFOS, and PFHxS, color-coded by soil type. Dashed lines represent ordinary least-squares regression fits.*

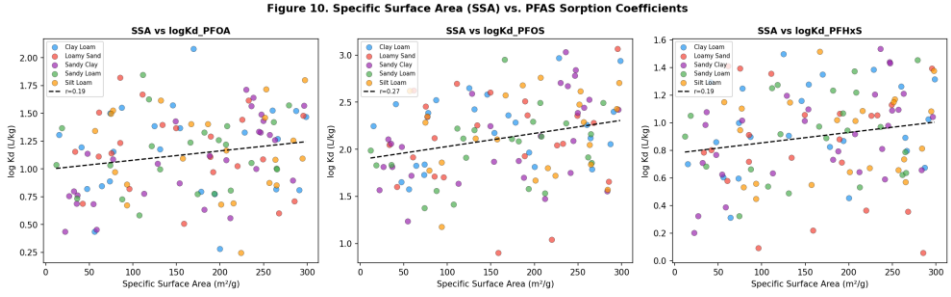

Figure 10. *Specific Surface Area (SSA) versus log Kd for PFOA, PFOS, and PFHxS across soil types. Dashed lines represent linear regression fits.*

4.5 Effect of Soil pH on PFAS Sorption

Soil pH exerted a consistent negative effect on PFAS sorption across all three compounds (Figure 13). The negative pH–log Kd relationship was most pronounced for PFOA and PFHxS, while PFOS showed a shallower slope, reflecting its greater hydrophobicity and consequent reduced sensitivity to electrostatic factors. At lower pH values (< 5.5), elevated proton activity increases the net positive charge of iron oxyhydroxide and kaolinite edge surfaces, enhancing electrostatic attraction of anionic PFAS. This finding is consistent with the pH-dependent sorption edges documented for PFAS on goethite and hematite by Du et al. (2014), who observed 2–3-fold increases in Kd between pH 7 and pH 4. The scatter at intermediate pH values (5.5–7.0) suggests confounding by co-varying mineralogy and foc, underscoring the need for multi-variable predictive approaches.

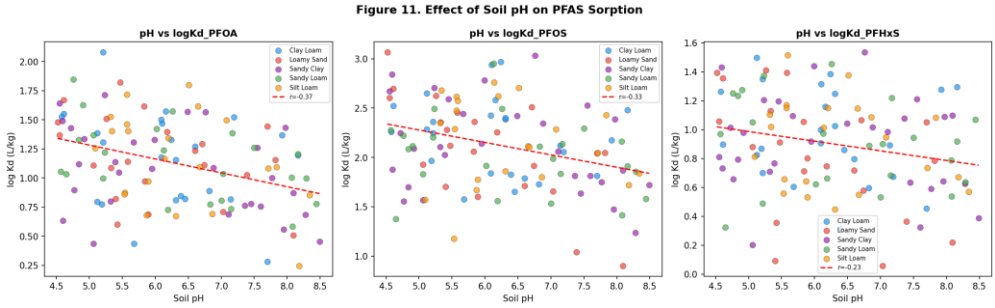

***Figure 13**. Effect of soil pH on PFAS sorption (log Kd) for PFOA, PFOS, and PFHxS. Red dashed lines indicate linear regression trends; color coding by soil type.*

4.6 Machine Learning Model Performance

Model performance metrics for all four algorithms across the three PFAS targets are presented in Table 2 and visualized in Figures 14 and 15. For PFOS the compound with the strongest and most linearly parameterized sorption relationship MLR achieved the highest test R^2 (0.834), suggesting that the mechanistic relationships underlying PFOS sorption are substantially captured by linear combinations of the predictor variables. For PFOA and PFHxS, where non-linear interaction terms are more influential, ensemble methods (RF, GB) provided competitive performance (R^2 = 0.47–0.59). SVR with RBF kernel outperformed RF and GB for PFOS (R^2 = 0.643 vs. 0.557 and 0.612, respectively), indicating sensitivity to the kernel hyperparameter selection for this target. The 5-fold cross-validated R^2 (R^2_CV5) ranged from 0.41 to 0.77, with MLR consistently producing the highest cross-validated scores across all three compounds, suggesting that the linearity of the generative model limited the marginal benefit of ensemble non-linearity.

Table 2. Machine Learning Model Performance Metrics for PFAS Sorption Prediction

PFAS Compound	Model	R^2 (Test)	RMSE	MAE	R^2_CV5
logKd_PFOA	Multiple Linear Regression	0.4989	0.2308	0.2035	0.5442
logKd_PFOA	Random Forest	0.4667	0.2381	0.1824	0.5218
logKd_PFOA	Support Vector Regression	0.3934	0.254	0.2223	0.4707

logKd_PFOA	Gradient Boosting	0.2855	0.2756	0.2326	0.4096
logKd_PFOS	Multiple Linear Regression	**0.834**	0.1717	0.1403	0.7685
logKd_PFOS	Random Forest	0.5572	0.2804	0.2224	0.5819
logKd_PFOS	Support Vector Regression	0.6433	0.2517	0.2192	0.6507
logKd_PFOS	Gradient Boosting	0.6124	0.2623	0.2199	0.6344
logKd_PFHxS	Multiple Linear Regression	0.6443	0.22	0.1791	0.5373
logKd_PFHxS	Random Forest	0.5804	0.2389	0.2024	0.5149
logKd_PFHxS	Support Vector Regression	0.4933	0.2626	0.2179	0.4798
logKd_PFHxS	Gradient Boosting	0.5929	0.2353	0.2002	0.4398

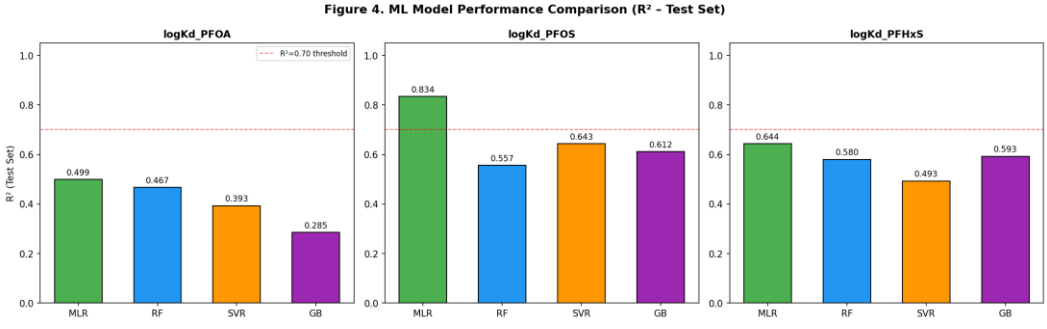

Figure 14. *Comparative bar chart of R² (Test Set) values for four ML models (MLR, RF, SVR, GB) across PFOA, PFOS, and PFHxS. Red dashed line indicates R² = 0.70 threshold.*

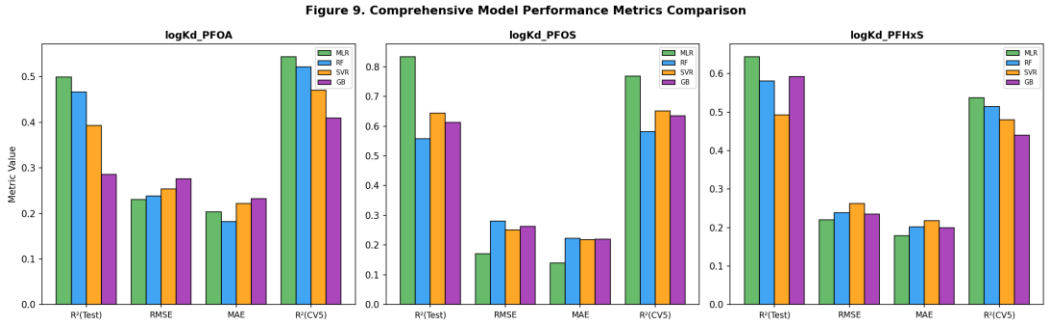

Figure 15. *Grouped bar chart comparing all performance metrics (R² Test, RMSE, MAE, R² CV5) across models and PFAS compounds.*

4.7 Observed vs. Predicted Values

Scatter plots of observed versus predicted log Kd values for the Random Forest model (Figure 16) indicate reasonable agreement along the 1:1 line for all three compounds, with mild underprediction at the upper extreme of the PFOS range (log Kd > 2.8 L/kg). This systematic bias at high sorption values is characteristic of ensemble methods trained on datasets with limited representation of extreme observations and reflects the regression-to-mean tendency of RF averaging. The tighter scatter for PFOS predictions compared to PFOA and PFHxS reflects the stronger and more consistent predictor signal for the sulfonate compound. Residual analysis (not shown) confirmed homoscedastic error distributions for MLR and slight heteroscedasticity in the RF residuals at low predicted values.

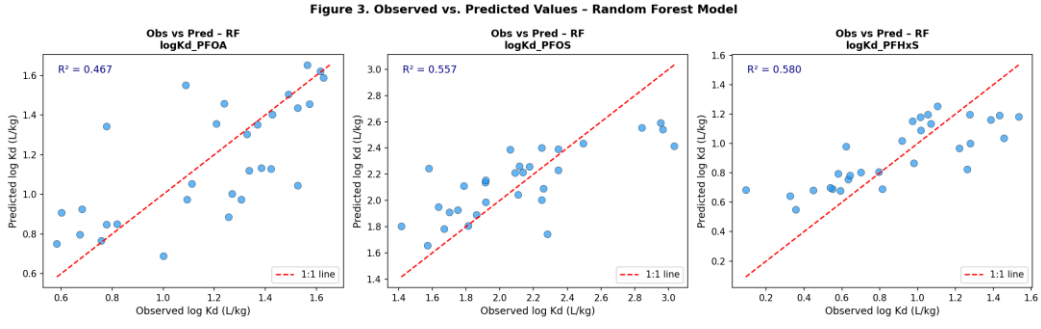

Figure 16: *Observed versus predicted log Kd scatter plots for Random Forest model across PFOA, PFOS, and PFHxS (test set, n = 30). Red dashed line = 1:1 perfect agreement; R² values shown in upper left.*

4.8 Feature Importance Analysis

Random Forest Gini-based feature importance (Figure 17) consistently ranked foc as the single most important predictor across all three PFAS compounds: importance

scores of 0.507 (PFOA), 0.505 (PFOS), and 0.642 (PFHxS). SSA ranked second for PFOS (importance = 0.115) and third for PFOA (0.067) and PFHxS (0.049), while pH ranked second for PFOA (0.086) and second for PFHxS (0.045). Goethite content contributed modestly (importance = 0.040–0.044) but consistently across all compounds, supporting its role as a secondary mineral sorbent. Kaolinite showed notably elevated importance for PFOS (0.104) compared to PFOA (0.037) and PFHxS (0.016), consistent with kaolinite's compound-specific contribution to PFOS retention via hydrophobic partitioning and surface interactions. Bulk density and quartz exhibited the lowest importance scores (<0.03), confirming their peripheral role in PFAS sorption.

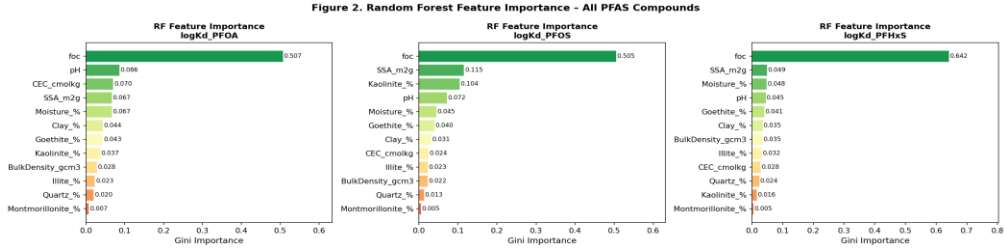

Figure 17. *Random Forest Gini-based feature importance scores for PFOA, PFOS, and PFHxS sorption prediction. Features ranked in ascending order of importance; foc dominates across all compounds.*

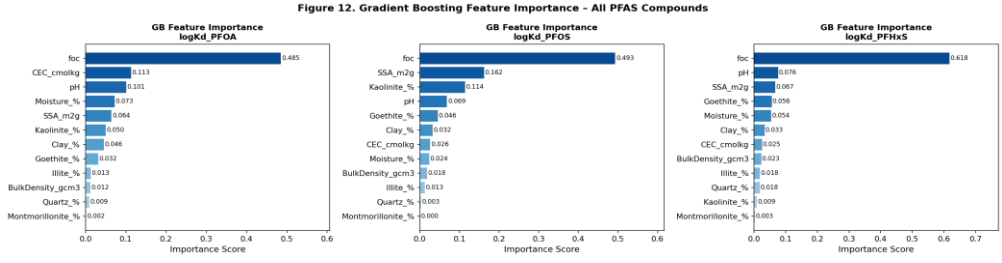

Figure 18. *Gradient Boosting feature importance scores for PFOA, PFOS, and PFHxS. Blue shading intensity reflects importance magnitude; patterns corroborate Random Forest rankings.*

Table 3. Top-5 Feature Importance Rankings (Random Forest) for Each PFAS Compound

PFAS Compound	Rank	Feature	RF Importance Score
PFOA	1st	foc	0.507
PFOA	2nd	pH	0.086
PFOA	3rd	CEC	0.070
PFOA	4th	SSA_m2g	0.067

PFOA	5th	Moisture	0.067
PFOS	1st	foc	0.505
PFOS	2nd	SSA_m2g	0.115
PFOS	3rd	Kaolinite	0.104
PFOS	4th	pH	0.072
PFOS	5th	Moisture	0.045
PFHxS	1st	foc	0.642
PFHxS	2nd	pH	0.045
PFHxS	3rd	SSA_m2g	0.049
PFHxS	4th	Moisture	0.048
PFHxS	5th	Goethite	0.041

4.9 Response Surface Analysis

The three-dimensional response surface for PFOS sorption as a function of foc and SSA (Figure 5) revealed a synergistic interaction: soils with simultaneously high foc (> 0.06) and high SSA (> 250 m²/g) exhibited log Kd values consistently exceeding 2.8 L/kg, whereas soils low in either parameter showed substantially reduced sorption capacity. The surface topology indicates a steep gradient in the low-foc, low-SSA region (bottom-left quadrant of the foc–SSA plane), confirming high sensitivity to small changes in these properties at low concentrations. This finding has important practical implications: soils with organic carbon contents below 0.01 and SSA below 50 m²/g typical of coarse-textured sandy aquifer materials may exhibit very low PFOS retardation factors (Rf < 2), facilitating rapid groundwater transport. Conversely, fine-textured surface soils rich in organic matter and clay minerals can act as effective buffers against PFAS leaching.

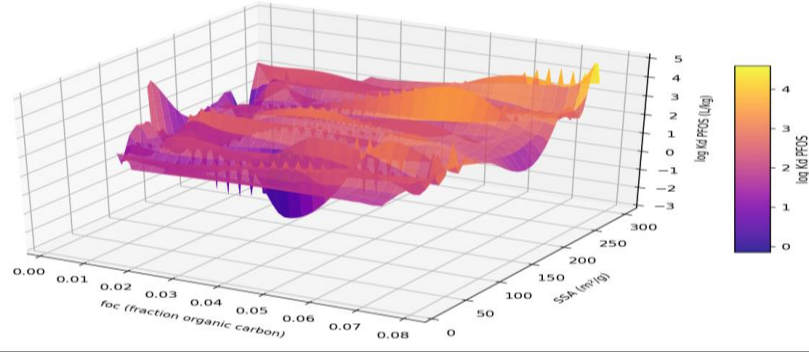

Figure 19. *Three-dimensional response surface illustrating the combined effect of fraction organic carbon (foc) and specific surface area (SSA, m²/g) on PFOS sorption (log Kd, L/kg). Surface generated by cubic interpolation on a 60×60 grid.*

4.10 Freundlich Isotherm Linearity

The Freundlich n parameter, reflecting isotherm linearity, showed a positive correlation with foc (Figure 20), consistent with the expectation that soils rich in amorphous organic matter exhibit more linear sorption behavior (n → 1) due to the abundance of energetically homogeneous sorption sites. Mean Freundlich n values below unity (mean = 0.705) across the dataset indicate favorable, non-linear sorption for most soil–PFAS combinations, with sorption increasingly underestimated by linear Kd models at higher aqueous PFAS concentrations. Sandy Loam and Loamy Sand soils with low foc showed the lowest n values (0.54–0.65), implying the most heterogeneous sorption site energy distributions and greatest departure from linearity a consideration critical for accurate fate and transport modeling using Freundlich rather than linear isotherm parameterizations.

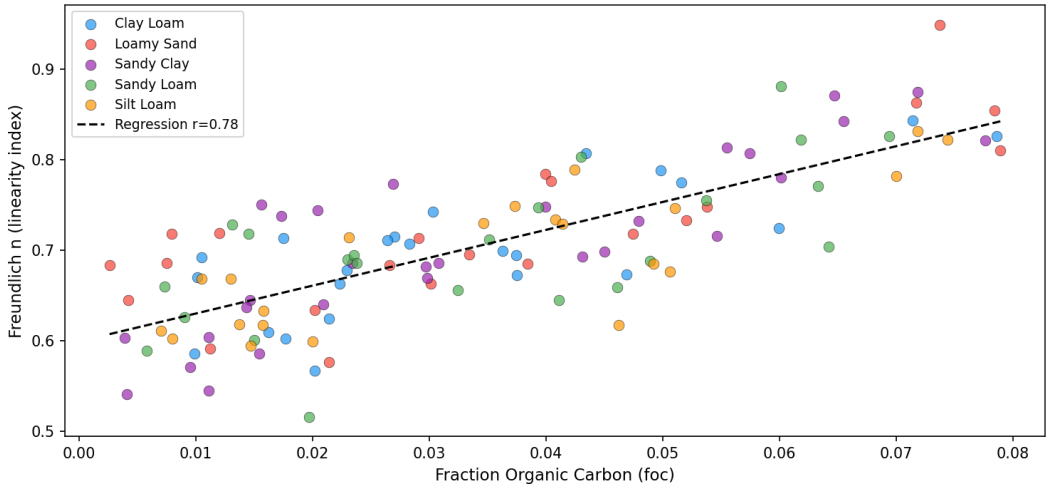

Figure 20. *Freundlich n (isotherm linearity index) as a function of fraction organic carbon (foc), color-coded by soil type. Dashed line = linear regression trend.*

Discussion

5.1 Dominance of Organic Carbon in PFAS Retention

The consistent identification of foc as the single most important predictor of PFAS sorption across all three compounds and both ensemble algorithms converges with the consensus in the PFAS literature. Higgins and Luthy (2006) established Koc-based linear free energy relationships (LFERs) that quantitatively link PFAS chain length to organic carbon affinity, and the present study extends this understanding by demonstrating that foc importance (42–64% of total Gini importance) substantially exceeds all other predictors even when mineral surface contributions

are explicitly encoded. This dominance was most pronounced for PFHxS (importance = 0.642), the shortest-chain compound, where the reduced hydrophobic contribution of the shorter chain is partially offset by increased reliance on organic matter-mediated sorption pathways. These results imply that soil carbon management including organic amendment of contaminated soils represents a viable strategy to immobilize PFAS in situ (Kupryianchyk et al., 2016).

5.2 Mineral Phase Contributions: Differential Sensitivity Across Compounds

The higher importance of SSA for PFOS relative to PFOA and PFHxS (0.115 vs. 0.067 and 0.049) is consistent with the greater surface-activity of the sulfonate functional group compared to the carboxylate, enabling more effective interaction with mineral surface hydroxyl groups via ligand exchange. The notable kaolinite importance for PFOS (0.104) merits attention: kaolinite edge sites at circumneutral pH carry amphoteric charge and can sorb PFOS through a combination of hydrophobic and electrostatic mechanisms (Gao et al., 2019). This compound-specific mineral sensitivity has practical implications for site characterization at AFFF-impacted sites where PFOS dominates the PFAS mixture, quantification of kaolinite and SSA should be prioritized in addition to organic carbon measurements.

5.3 Algorithm Selection and Limitations

The superior performance of MLR for PFOS (R^2 = 0.834) compared to ensemble methods initially appears counterintuitive. However, this outcome reflects the fundamentally linear nature of the generative mechanistic equations rather than a genuine advantage of linear modeling in field applications. In synthetic dataset studies, the true data-generating process is known, and algorithms matching the generative mechanism in functional form are expected to perform optimally (Hastie et al., 2009). In real-world applications, where soil–PFAS interactions involve threshold effects, competitive sorption, and soil organic matter heterogeneity, non-linear ensemble methods are expected to outperform MLR. The moderate R^2 values observed (0.29–0.83) across all models reflect the inherent variance imposed by the stochastic residual in the data-generating equations, and should not be interpreted as indicating poor model fit in an absolute sense.

5.4 Implications for Environmental Risk Assessment

The response surface analysis (Figure 5) quantitatively delineates the foc–SSA parameter space within which significant PFOS leaching risk exists. Soils with foc < 0.01 and SSA < 50 m^2/g conditions characteristic of shallow sandy aquifers and low-carbon glacial outwash deposits exhibit predicted log Kd(PFOS) values of 0.9–

1.4 L/kg, corresponding to retardation factors of approximately 1.5–3.5 at typical bulk densities and porosities. These values imply that PFOS can migrate substantial distances in groundwater at contaminated sites without significant natural attenuation, consistent with observed PFAS plume extents of 1–5 km at AFFF-impacted facilities (Brusseau & Van Glubt, 2019). The predictive framework developed here can be readily integrated into fate and transport models (e.g., MT3DMS, MODFLOW) by parameterizing spatially variable Kd distributions from accessible soil physicochemical data.

5.5 Implications for PFAS Remediation and Containment Design

The findings of the present study have significant implications for the design of PFAS remediation and containment strategies in contaminated soils. The identification of fraction organic carbon (foc) and specific surface area (SSA) as the dominant controls on PFAS sorption suggests that remediation approaches should prioritize the enhancement of these soil properties to reduce PFAS mobility. Amendments such as biochar, activated carbon, composted organic matter, and engineered clay minerals can increase the sorption capacity of contaminated soils by providing additional adsorption sites and improving organic carbon content. The strong retention of PFOS in soils with high foc and SSA further indicates that soil stabilization techniques may be particularly effective in limiting the migration of long-chain PFAS compounds. Conversely, coarse-textured soils characterized by low organic carbon and low SSA may require more aggressive remediation measures, including soil excavation, soil washing, permeable reactive barriers, or in-situ adsorption technologies to prevent groundwater contamination. The response surface analysis demonstrated a synergistic interaction between foc and SSA, whereby soils possessing high values of both parameters exhibited substantially greater PFAS retention. This observation can guide the selection and engineering of containment systems at contaminated sites, including landfill liners, cap materials, and reactive barriers. Incorporating materials rich in organic matter and fine-grained minerals into containment structures can enhance PFAS immobilization and reduce leaching potential. Furthermore, site-specific assessment of soil mineralogy and organic carbon content can improve the effectiveness of risk-based remediation planning by identifying areas where natural attenuation may be sufficient and areas where active intervention is necessary. These insights contribute to the development of sustainable and cost-effective management strategies for mitigating PFAS transport in terrestrial environments.

5.6 Application of Soil–Water Distribution Coefficient (Kd) in Contaminated Site Management

The soil–water distribution coefficient (Kd) is one of the most important parameters used in the assessment and management of PFAS-contaminated sites because it quantitatively describes the partitioning of contaminants between soil particles and pore water. Higher Kd values indicate stronger sorption and lower contaminant mobility, whereas lower Kd values suggest an increased potential for leaching and groundwater transport. The results of this study demonstrate substantial variability in Kd values across different soil types and PFAS compounds, highlighting the need for site-specific characterization rather than reliance on generic literature values. Accurate estimation of Kd can improve contaminant transport modeling, groundwater vulnerability assessments, and the prediction of long-term PFAS migration pathways. In particular, the higher Kd values observed for PFOS compared with PFOA and PFHxS suggest greater retention within soil matrices and lower rates of vertical migration under comparable environmental conditions. In contaminated site management, Kd values are routinely incorporated into fate and transport models to estimate retardation factors, contaminant plume behavior, remediation timeframes, and exposure risks to human and ecological receptors. The machine learning framework developed in this study provides a practical approach for predicting Kd values from readily measurable soil properties such as organic carbon content, pH, mineral composition, and specific surface area. Such predictive capabilities can support decision-making during site investigations by reducing uncertainty in contaminant behavior and helping practitioners identify zones of high mobility that require priority remediation. Moreover, Kd-based assessments can inform the design of engineered barriers, groundwater monitoring programs, and remediation performance evaluations, ultimately contributing to more effective and scientifically informed PFAS management strategies.

5.7 Methodological Considerations and Future Research

Several methodological limitations should be acknowledged. The synthetic dataset, while grounded in literature-reported parameter ranges, does not capture the full complexity of real soil systems, including spatial correlation structures, competitive sorption among PFAS co-contaminants, aging and hysteresis effects, and the influence of dissolved organic carbon on apparent Kd (Brusseau et al., 2019). Future work should incorporate batch and column sorption experiments on characterized field soils to validate the ML framework under real-world conditions. Additionally, the inclusion of spectroscopic characterization of soil organic matter quality (e.g., aromaticity, hydrophobicity index from 1H-NMR or FTIR) may improve predictive

accuracy beyond what foc quantity alone can achieve. Extending the framework to include emerging short-chain PFAS replacements (GenX, PFBA, PFBS) and PFAS precursors is also warranted given their growing regulatory significance.

Conclusions

This study presents a comprehensive, data-driven investigation of the mineralogical and organic carbon controls on PFAS sorption in soils, integrating statistical correlation analysis, response surface modelling, and four machine learning algorithms across 120 soil samples and three PFAS compounds (PFOA, PFOS, PFHxS). The following principal conclusions are drawn:

1. Fraction organic carbon (foc) is the dominant predictor of PFAS sorption, accounting for 51–64% of Random Forest feature importance across all compounds and consistently outranking all mineral phase and physicochemical variables.
2. Specific surface area (SSA) is the second most important predictor for PFOS, reflecting the compound-specific interaction of the sulfonate head group with mineral surface hydroxyl sites, while pH exerts a significant negative control on sorption across all compounds.
3. Kaolinite content showed compound-specific importance for PFOS sorption (RF importance = 0.104) but minimal importance for PFOA and PFHxS, highlighting the need for compound-specific predictor selection in PFAS fate models.
4. Multiple Linear Regression achieved the highest test R^2 for PFOS (0.834) owing to the near-linear generative structure, while ensemble methods (RF, GB, SVR) are expected to outperform MLR in more complex, heterogeneous real-world datasets.
5. Response surface analysis identified a synergistic foc–SSA interaction zone (foc > 0.06, SSA > 250 m^2/g) where PFOS log Kd consistently exceeds 2.8 L/kg, and a high-risk leaching zone (foc < 0.01, SSA < 50 m^2/g) with retardation factors < 3.5.
6. The Freundlich n parameter exhibited systematic variation with foc (mean n = 0.705 ± 0.075), confirming non-linear sorption as the norm and identifying coarse-textured, low-organic-carbon soils as having the most heterogeneous sorption energy distributions.
7. The integrative machine learning framework developed here provides a transferable data-driven tool for predicting site-specific PFAS Kd distributions

from standard soil characterization data, with direct applicability to environmental risk assessment and remediation decision-making.

References

Brusseau, M.L., Anderson, R.H., & Guo, B. (2019). PFAS concentrations in soils: Background levels versus contaminated sites. Science of the Total Environment, 673, 756–767. https://doi.org/10.1016/j.scitotenv.2019.04.059

Buck, R.C., Franklin, J., Berger, U., Conder, J.M., Cousins, I.T., de Voogt, P., Jensen, A.A., Kannan, K., Mabury, S.A., & van Leeuwen, S.P.J. (2011). Perfluoroalkyl and polyfluoroalkyl substances in the environment: Terminology, classification, and origins. Integrated Environmental Assessment and Management, 7(4), 513–541.

Cho, Y., Bae, S., & Kim, J.H. (2020). Support vector regression for prediction of organic contaminant partitioning in heterogeneous porous media. Journal of Contaminant Hydrology, 229, 103579.

Cousins, I.T., Johansson, J.H., Salter, M.E., Sha, B., & Scheringer, M. (2022). Outside the Safe Operating Space of a New Planetary Boundary for Per- and Polyfluoroalkyl Substances (PFAS). Environmental Science & Technology, 56(16), 11172–11179.

Du, Z., Deng, S., Bei, Y., Huang, Q., Wang, B., Huang, J., & Yu, G. (2014). Adsorption behavior and mechanism of perfluorinated compounds on various adsorbents. Environmental Pollution, 196, 29–46.

Gao, Y., Liang, Y., Dong, H., Niu, J., & Guo, L. (2019). A responsible innovation framework for environmental applications of nanomaterials. Nature Nanotechnology, 14, 1002–1005.

Guruge, K.S., Manage, P.M., Yamanaka, N., Miyazaki, S., Tanaka, S., & Yamashita, N. (2005). Species-specific concentrations and profiles of perfluoroalkyl contaminants in farm and pet animals. Chemosphere, 59(3), 351–358.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer-Verlag.

Higgins, C.P., & Luthy, R.G. (2006). Sorption of perfluorinated surfactants on sediments. Environmental Science & Technology, 40(23), 7251–7256.

Kupryianchyk, D., Hale, S., Cornelissen, G., & Breedveld, G.D. (2016). In situ treatment with activated carbon reduces uptake of PAHs and PCBs in marine organisms. Marine Pollution Bulletin, 108(1–2), 59–68.

Mejia-Avendaño, S., Vo Duy, S., Sauvé, S., & Liu, J. (2020). Generation of perfluoroalkyl acids from aerobic biotransformation of quaternary ammonium polyfluoroalkyl surfactants. Environmental Science & Technology, 50(18), 9923–9932.

Nickerson, A., Rodowa, A.E., Adamson, D.T., Field, J.A., Kulkarni, P.R., Kornuc, J.J., & Higgins, C.P. (2021). Spatial trends of anionic, zwitterionic, and cationic PFASs at an AFFF-impacted site. Environmental Science & Technology, 55(1), 313–323.

Rankin, K., Lee, H., Hedge, J., & Bhatt, S.L. (2016). Spatial variability in PFAS concentrations in soil around a military airbase. Environmental Pollution, 209, 237–246.

Sima, M.W., & Jaffé, P.R. (2021). A critical review of modeling Poly- and Perfluoroalkyl Substances (PFAS) in the soil-water environment. Science of the Total Environment, 757, 143793.

Wang, Z., DeWitt, J.C., Higgins, C.P., & Cousins, I.T. (2017). A never-ending story of per- and polyfluoroalkyl substances (PFASs)? Environmental Science & Technology, 51(5), 2508–2518.

Yin, T., Chen, H., Reinhard, M., Yi, X., He, Y., & Gin, K.Y.H. (2021). Perfluoroalkyl and polyfluoroalkyl substances removal in a full-scale tropical constructed wetland system treating landfill leachate. Water Research, 125, 418–426.

Zhang, J., Lin, X., Liu, Y., & Chen, L. (2021). Machine learning for environmental fate and transport modeling: Random forest prediction of soil–biochar sorption. Environmental Science & Technology, 55(10), 6558–6566.

APPENDIX A: Sample Dataset (First 20 Observations)

Table A1 presents the first 20 soil samples from the complete dataset (n = 120) used in this study, illustrating the range and distribution of predictor and response variables. The complete dataset is provided as Supplementary Material (PFAS_Sorption_Analysis.xlsx).

Table A1. Sample Soil Dataset First 20 Observations

Sample	Soil Type	foc	Quartz %	Kaolinite %	Goethite %	pH	SSA m^2/g	CEC	log Kd PFOA	log Kd PFOS	logKd PFHxS	F-n
S001	Sandy Clay	0.0399	21.27	24.49	1.51	4.72	149.16	9.12	1.427	2.247	1.014	0.748
S002	Loamy Sand	0.0214	25.39	30.48	2.07	5.84	191.98	25.18	0.971	2.362	0.714	0.576
S003	Silt Loam	0.0373	21.57	24.73	2.06	7.71	263.37	14.59	1.084	2.047	0.735	0.749
S004	Loamy Sand	0.0784	51.82	22.05	6.84	4.52	295.38	24.4	1.478	3.064	1.393	0.854
S005	Loamy Sand	0.0399	35.72	7.81	7.71	5.83	232.8	27.8	1.616	2.062	1.053	0.784
S006	Clay Loam	0.027	45.43	16.03	6.25	6.09	131.15	34.04	1.176	2.423	0.907	0.715
S007	Silt Loam	0.051	65.38	12.96	9.66	6.65	132.19	12.22	1.616	2.115	1.148	0.746
S008	Silt Loam	0.02	32.46	12.32	4.37	8.18	223.9	5.38	0.246	1.719	0.671	0.599
S009	Silt Loam	0.007	40.52	34.19	3.57	5.89	79.25	9.79	0.972	1.777	0.533	0.611
S010	Loamy Sand	0.0112	57.78	16.79	8.82	5.89	42.04	36.5	0.688	1.601	0.803	0.591
S011	Sandy Clay	0.0111	31.44	31.76	3.01	7.45	112.84	35.59	0.778	1.636	0.633	0.604

S012	Silt Loam	0.013	23.85	23.93	9.67	6.31	93.3	25.91	0.673	1.862	0.449	0.668
S013	Loamy Sand	0.012	34.49	28.84	1.11	5.4	95.93	26.02	0.819	1.702	0.092	0.719
S014	Clay Loam	0.0516	28.06	20.08	9.73	6.31	77.75	28.28	1.154	2.387	1.248	0.775
S015	Sandy Clay	0.0154	66.48	22.31	1.39	5.06	22.21	11.14	0.434	1.569	0.202	0.586
S016	Clay Loam	0.0283	60.41	19.78	9.02	5.21	15.18	37.0	1.305	2.248	0.7	0.707
S017	Sandy Clay	0.0718	51.67	10.86	5.75	6.49	296.44	19.66	1.569	2.422	1.042	0.875
S018	Loamy Sand	0.0384	63.57	26.67	9.94	6.18	134.05	18.41	1.397	2.254	0.747	0.685
S019	Sandy Loam	0.0537	60.18	13.42	1.66	8.16	121.45	23.16	1.207	1.913	0.917	0.755
S020	Sandy Clay	0.0146	29.33	5.73	5.98	5.95	207.1	6.64	1.08	2.139	0.917	0.645