Journal of Hazardous Materials 500 (2025) 140539

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

Novel insights from multivariate analysis of metadata from a large PFAS remedial investigation dataset

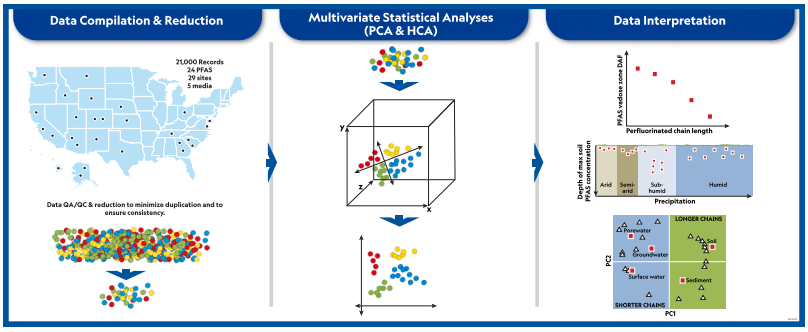
Dung Nguyen [a,*], Sonya Cadle [b], Teresa Verstraet [c], Lisa Kammer [d], Taire Van Scoy [b], Matt Anding [b], Richard Anderson [e]

[a] Weston Solutions, Inc., 615 2nd Avenue, Suite 350, Seattle, WA, USA
[b] Tepa, LLC, 9777 Pyramid Ct, Suite 265, Englewood, CO, USA
[c] Weston Solutions, Inc., 1536 Cole Blvd. Bldg.4, Suite 375, Lakewood, CO, USA
[d] Weston Solutions, Inc., 43 N Main Street, Concord, NH, USA
[e] Environmental Decontamination Limited, 161 Spanish Oak, Castroville, TX, USA

HIGHLIGHTS

- PCA and HCA simplify interpretation of large PFAS environmental datasets.
- PFAS cluster by chain length and media type across multiple site conditions.
- PFAS vertical transport varies with climate and hydraulic flux conditions.
- Results support refinement of site-specific knowledge frameworks and PFAS remediation strategies.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
PFAS
AFFF
Multivariate statistical analysis
Large datasets
Remedial investigations

ABSTRACT

The extensive use of per- and polyfluoroalkyl substances (PFAS) in aqueous film-forming foam (AFFF) has resulted in widespread environmental contamination. This study applies Principal Component Analysis (PCA) and Hierarchical Clustering Analysis (HCA) to approximately 21,000 PFAS records collected from 29 sites in the continental United States to identify compositional patterns, source signatures, and spatial trends across groundwater, porewater, surface water, soil, and sediment at AFFF source areas and downgradient plumes . A novel concept termed the estimated PFAS vadose zone dilution attenuation factor (the ratio between PFAS porewater and groundwater concentration) is defined herein and negatively correlated with perfluorinated chain length, reflecting sorption and mobility differences. PCA revealed clustering by chain length and media type: short-chain PFAS dominated aqueous samples, while long-chain PFAS were prevalent in soil and sediment. PFHxS was most abundant in groundwater and porewater; PFOS dominated surface water and solids. The maximum PFAS soil concentrations are generally limited to the upper 1 m in arid climates and deeper in sub-

* Corresponding author.
E-mail addresses: Zoom.Nguyen@WestonSolutions.com (D. Nguyen), Sonya.Cadle@tepa.com (S. Cadle), Teresa.Verstraet@WestonSolutions.com (T. Verstraet), Lisa.Kammer@WestonSolutions.com (L. Kammer), Taire.VanScoy@tepa.com (T. Van Scoy), Matt.Anding@tepa.com (M. Anding), hunter.anderson@edl-tech.com (R. Anderson).

https://doi.org/10.1016/j.jhazmat.2025.140539
Received 25 August 2025; Received in revised form 10 November 2025; Accepted 17 November 2025
Available online 19 November 2025

humid regions. Site-specific PFAS fingerprints were associated with hydroclimatic variables. PCA results suggest diffuse contamination patterns, likely due to historical AFFF application. These findings support refinement of the site-specific knowledge framework, inform risk assessments, and guide targeted sampling and remediation strategies at PFAS-impacted sites.

1. Introduction

Per- and polyfluoroalkyl substances (PFAS) are a diverse group of thousands of synthetic chemicals with unique physicochemical characteristics including exceptional chemical and thermal stability attributable to their strong carbon-fluorine bonds [1]. These properties have led to the widespread use of PFAS in various industrial processes and consumer products including nonstick cookware, stain-resistant fabrics, and, notably, aqueous film-forming foam (AFFF) used for Class B firefighting. The proliferation of PFAS in numerous industrial and commercial applications has resulted in their widespread occurrence in the environment.

As early as the mid-2000s, PFAS contamination began to emerge as a major environmental concern. Due to their high mobility in the environment and resistance to most biological and chemical degradation processes, PFAS including perfluorooctanoic acid (PFOA) and perfluorooctane sulfonate (PFOS) have been detected in various media including surface water, groundwater, soil, sediment, and porewater at hundreds of sites [2,3]. Exposure to PFAS has been associated with various adverse health impacts including developmental effects, immune system suppression, and increased risk of cancers [4,5], leading to increasing regulatory scrutiny at both the federal and state levels. In April 2024, the United States Environmental Protection Agency (EPA) finalized the first-ever National Primary Drinking Water Regulations that set Maximum Contaminant Levels (MCLs) for six PFAS, including MCLs of 4 parts per trillion (ppt) for PFOA and PFOS [6]. These regulations aim to reduce PFAS exposure to approximately 100 million people, prevent thousands of deaths, and reduce tens of thousands of serious illnesses in the US alone.

Since the mid-late 2000's, investigations conducted at the 29 sites included in this study have generated vast amounts of analytical data, often encompassing hundreds or thousands of samples at a single site across multiple environmental media (soil, sediment, surface water, groundwater, porewater), each with dozens of PFAS detections. This results in a complex, high-dimensional dataset that can be difficult to interpret using conventional univariate or bivariate analysis alone. Moreover, the fate and transport of PFAS in the environment are highly dependent on their perfluorinated chain length and functional group as well as a variety of site-specific factors including lithology, meteorology, hydrogeology, source composition, nature and extent of contaminant release, and physicochemical interactions with other contaminants. The relative contributions and interaction among these various factors on the overall magnitude of PFAS retention in unsaturated soils above the water table (i.e., vadose zone) and PFAS leaching to underlying groundwater remains somewhat uncertain and can only be fully interrogated by large empirical datasets from a diverse portfolio of contaminated sites. Therefore, there is a critical need to identify key compositional patterns, source signatures, and spatial trends across media and sites to improve the effectiveness and efficiency of current and future investigation and remediation efforts across different media at PFAS-impacted sites.

Multivariate statistical methods offer robust tools for analyzing complex datasets by uncovering patterns and relationships that may not be evident through traditional data analytical techniques. Principal component analysis (PCA), for instance, reduces data dimensionality by transforming original correlated variables into uncorrelated components that capture most of the data variance [7]. This can reveal groupings of PFAS that share similar spatial distributions, potentially indicating common sources, environmental fate and transport processes or treatment challenges. Hierarchical clustering analysis (HCA), another multivariate statistical technique, groups samples or variables based on similarity metrics, producing dendrograms that can help delineate site groupings, prioritize sampling areas or differentiate between background and site-related contamination [8]. The application of these techniques has shown promise in PFAS source identification, fate and transport characterization, and prioritization of sampling locations [9-12].

In the context of large industrial facilities having multiple sites on the same campus or corporations having multiple sites nationwide or internationally, both cross-media and cross-site comparisons are essential. Thus, multivariate approaches have the potential to streamline data interpretation, support refinement of the site-specific knowledge framework, and guide decision-making for future sampling and remediation strategies [13-17,12]. The primary objective of this study is to apply multivariate statistical techniques including PCA and HCA, in conjunction with traditional statistical techniques, to a large PFAS dataset collected in surface water, groundwater, porewater, soil, and sediment samples at 29 sites, of which several had multiple campuses that were physically separated from each other. By identifying key compositional patterns, source signatures, and spatial trends across media and sites, this work aims to demonstrate how multivariate statistical tools can simplify interpretation of large environmental datasets and improve overall site management practices.

2. Materials and methods

2.1. Data compilation and reduction

A total of approximately 9000 aqueous and 12,000 solid PFAS environmental sample results were examined as part of this study. The sites examined in this study are presented in Supplemental information (SI) SI-A. Environmental media sampled include soil and sediment as well as surface water, groundwater, and porewater collected at suspected AFFF source areas and plumes downstream of the source areas along the groundwater flow path (i.e., downgradient plumes). In addition to PFAS analytical data, site-specific geospatial and meteorological information including average annual temperature (temp), precipitation (precip), and evapotranspiration (ET) as well as depth to groundwater (DTW) were used in the multivariate analysis. A summary of the site-specific data is provided in SI-A.

Several data reduction steps were conducted prior to the multivariate statistical analysis. At locations where multiple rounds of groundwater sampling data were available, only the most recent data were used to minimize duplication. Field duplicates and other quality control/quality assurance (QA/QC) samples were also removed. Because early sampling efforts produced data based on a limited analyte list, only datasets with the full analytical suite of 24 PFAS (obtained via EPA Method 537 and presented in SI-B) that was later adopted widely were included in subsequent statistical analyses to ensure that all datasets were analyzed consistently. Additionally, only native (i.e., untreated) samples were used whereas samples collected as part of site remediation efforts were removed from the dataset. For general statistical evaluation of PFAS concentration and composition, this reduced PFAS dataset was used. Multivariate statistical analyses were performed on the median and the maximum PFAS concentrations observed in each environmental matrix to identify patterns in PFAS distribution and composition representative of a typical PFAS-impacted downgradient plume versus an AFFF source area, respectively.

For aqueous media including surface water, groundwater, and porewater, only the PFAS detected at an approximately 60 % frequency or higher were included in all statistical work described below. This resulted in 12 compounds (including 6:2 FTS, PFBA, PFPeA, PFHxA, PFHpA, PFOA, PFNA, PFBS, PFPeS, PFHxS, PFHpS, and PFOS) that were then subjected to the Robust Regression on Order Statistics (RROS via EPA ProUCL 5.2) for censored log-normally distributed environmental data (described by Helsel [18]) to assign values to PFAS data measured below the analytical limit of detection. Solid media were similarly subjected to RROS. Because data from many AFFF-impacted sites indicate that the highest PFAS concentrations are generally found in the organic-rich shallow soil, only data collected within the upper 0.3 m were included in the multivariate statistical analyses for soil samples [19-21]. Only the eight PFAS (including PFPeA, PFHxA, PFHpA, PFOA, PFNA, PFDA, PFHxS, and PFOS) detected in shallow soils at a frequency of approximately 60 % or higher were used in further statistical analyses. For sediment, the six PFAS meeting the detection frequency threshold of 60 % or higher were used for the statistical analyses and include FOSA, PFPeA, PFHxA, PFOA, PFHxS, and PFOS. Like the aqueous media, the censored solid data were treated similarly using the RROS approach. The summary statistics of the data used for the general and multivariate statistical analyses described below are presented in SI-C.

2.2. Statistical analyses

A statistical analysis of the exceptionally large dataset examined in this study using either univariate descriptive or explorative methods alone would be tedious, computationally tasking, and inefficient. Multivariate exploratory methods such as PCA, redundancy analysis, and factor analysis were considered; amongst these methods, PCA was found to be best suited due to its simplicity, interpretation quality, and usefulness in explaining the variation in our dataset. The main objective of a PCA is to transform a large number of correlated variables into an equal number of uncorrelated indices called principal components (PC). The first PC accounts for as much of the variance of the entire dataset as possible, and the second PC retains as much of the remaining variance as possible, and so on and so forth. The goal of PCA is to determine the minimum number of PCs accounting for the majority of the variance in the entire dataset, thereby reducing its overall data complexity and facilitating pattern identification.

PCA was conducted using XLSTAT [22] on standardized PFAS concentration data using the z-score normalization technique, with Pearson correlations used to mitigate the influence of differing measurement scales across sites. Each site (or physically separate campus, where applicable) was represented as a single point within the two-dimensional PC1/PC2 space. Two-dimensional plots of PC1 and PC2 scores were generated, and geospatial/meteorological covariates (average annual temperature, precipitation, evapotranspiration, and depth to groundwater) were overlaid to visually assess potential correlations with observed PFAS patterns. These overlays were used as exploratory visual aids only; no formal inference was made at this stage. Spatial contours were generated using RcGIS, a custom R library developed in-house, which implements thin-plate spline interpolation via the open-source fields package (CRAN). The batch_tps function within RcGIS allows automated contouring, interval specification, and cropping to a convex or concave hull. The function is openly documented and reproducible within standard R environments. A smoothing parameter ($\lambda = 5 \times 10^{-5}$) was chosen based on cross-validation to balance smoothness and overfitting.

To complement PCA, hierarchical cluster analysis (HCA) was conducted to identify clusters of sites exhibiting similar characteristics within the dataset. HCA was performed in XLSTAT using an agglomerative approach, Ward's linkage method, and Euclidean distance as the similarity measure. This combination minimizes within-cluster variance and facilitates interpretability of dendrograms. Clusters were inspected to evaluate grouping of sampling sites and PFAS compounds based on concentration patterns and co-occurrence.

3. Results and discussion

3.1. PFAS detection frequencies in different environmental media

The detection frequencies of PFAS (ordered in decreasing relative mobility in the x-axis) for different media examined and all samples collected as part of this study are shown below in Fig. 1. Shorter-chain PFAS (e.g., PFBS, PFBA, and PFHpA) were detected at high frequencies (often exceeding 80 %) in aqueous media including groundwater, porewater, and surface water. This is consistent with the high relative mobility of short-chain PFAS in the environment. In contrast, longer-chain PFAS were less frequently detected in aqueous media but are more often found in solid media including soil and sediment. PFHxS and PFOS were the most frequently detected analytes in all five matrices at a frequency exceeding 80 %. These trends underscore the influence of PFAS physicochemical properties on their partitioning and transport behavior.

3.2. PFAS concentrations in different environmental media

Typical concentration ranges of PFAS detected at a frequency of 60 % or higher across various media including groundwater, porewater, surface water, soil (all depths vs. the upper 0.3 m), and sediment at all sites examined in this study are presented in Fig. 2. The composition of PFAS varies considerably among these media of concern and is discussed in more detail below. Among the frequently detected compounds, PFOS generally exhibited the highest median concentrations in solid media [29 micrograms per kilogram (μg/kg) in the upper 0.3 m] and sediment (2.1 μg/kg). Notably, PFAS were observed at significantly higher concentrations in the upper 0.3 m of soil; PFAS distribution in soil as a function of depth is discussed in subsequent section. Similar to soil and sediment, PFOS was the most prevalent compound in surface water with a median concentration of approximately 130 ng/L. In contrast, the highest median concentrations in groundwater and porewater were observed for PFHxS at approximately 160 and 1200 ng/L, respectively. Porewater generally exhibited significantly higher median concentrations compared to groundwater; differences in median porewater and groundwater tend to decrease with perfluorinated chain length (discussed further below). In contrast to the median groundwater and porewater concentrations, where shorter-chain sulfonates and carboxylates tend to predominate, the maximum PFAS concentrations (representative of AFFF source areas) observed in groundwater and porewater are generally associated with PFOS and 6:2 FTS. Notable observations made with respect to PFAS concentration and composition in groundwater and porewater are further discussed below.

3.3. PFAS composition in different environmental media

Figs. 3**a** and 3**b** illustrate the differences in PFAS composition across the five media of concern (groundwater, porewater, surface water, soil, and sediment) based on maximum and median PFAS concentrations, respectively, and in decreasing relative mobility (see Attachment SI-D). In general, shorter-chain PFAS with higher relative mobility were detected with the highest relative abundance in surface water. Conversely, longer-chain PFAS that exhibit a lower relative mobility in the environment are more frequently detected in soil and sediment samples. Although the relative percent contributions of individual PFAS differ slightly between the maximum concentration dataset (representative of AFFF source areas – Fig. 3**a**) and the median dataset (more indicative of downgradient plume conditions – Fig. 3**b**), both datasets exhibit similar overall PFAS compositional patterns.

PCA was also employed to assess the relative abundance of different target PFAS in the maximum concentration samples observed in each

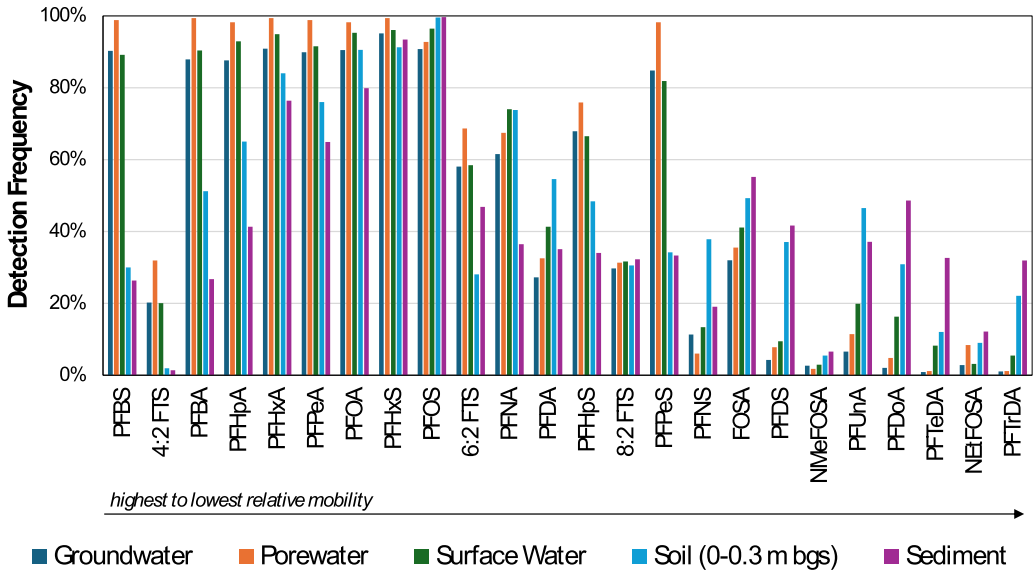


Fig. 1. PFAS detection frequencies in different media.

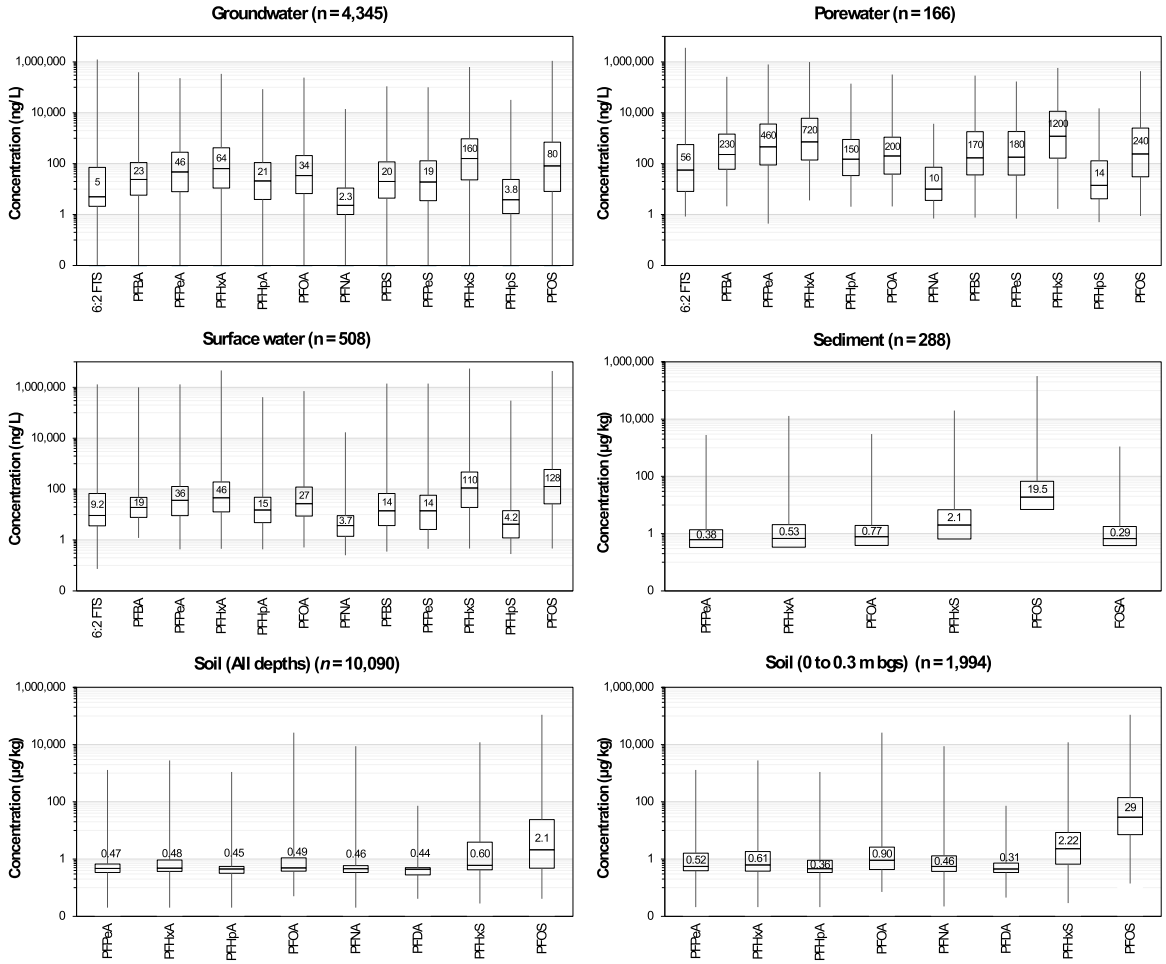


Fig. 2. PFAS concentration ranges in various media. *Reporting limits were used for non-detects. The lower and upper vertical bars represent the minimum and maximum concentrations. The lower, middle, and upper horizontal line represent the 25th, 50th (median), and 75th percentile concentrations. The PFAS concentration shown at the center of each box represents the median concentration observed in each dataset.*

media of concern as shown in Fig. 4. In general, PFAS with similar chain-lengths and relative mobilities tend to cluster together in the two-dimensional PCA plot consisting of PC1 and PC2 that collectively capture nearly 77 % of the variability in the dataset; the Pearson correlation coefficients among the PFAS frequently observed in each media are presented in Attachment SI-E. Consistent with the PFAS

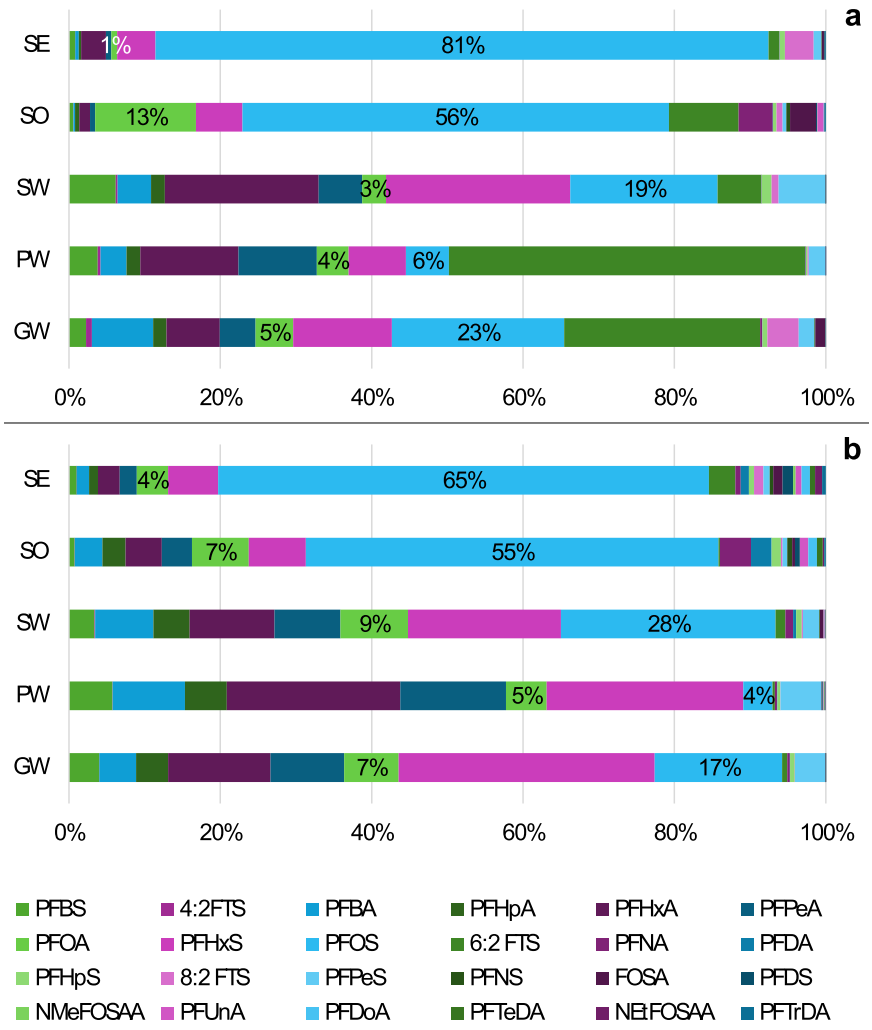


Fig. 3. PFAS composition across various media using maximum (a) and median concentrations (b) in decreasing relative mobility from left to right. *Reporting limits were used for non-detects. PFOA and PFOS percent contributions are shown. GW = groundwater. PW = porewater. SW = surface water. SO = soil. SE = sediment.*

compositional analysis depicted in Fig. 3, the PCA analysis indicates that certain PFAS tend to be associated with particular media. For example, short-chain PFAS, including PFPeS and PFHxS, are found in abundance in surface water samples, while long-chain PFAS and PFAA precursors with relatively low mobility such as PFNS, NetFOSA, NMeFOSA, PFDoA, and PFTrDA are generally observed in soil samples. Similarly, certain long-chain PFAAs including PFDS, PFTeDA, and PFOS are more likely to be found in sediment samples. Short-chain PFAS with relatively high mobility such as 4:2 FTS, PFBA, PFBS, PFPeA, and PFHpA are generally associated with porewater and groundwater samples. Fig. 4 further illustrates the utility of PCA in visually distilling complex data patterns that are less readily discernible using conventional data visualization techniques as shown in Fig. 3. In addition to differentiating differences in PFAS composition, PCA results also indicate that a higher percentage of these short-chain PFAS are generally observed in porewater and at higher concentrations compared to groundwater. The observed differences in PFAS composition between porewater and groundwater are likely attributable to variable physiochemical properties among various PFAS (discussed further below).

3.4. PFAS vadose zone dilution attenuation factors

Fig. 5 graphically depicts the ratios of PFAS concentrations in porewater to groundwater at the 25th percentile, median (50th percentile), 75th percentile, and maximum concentrations as a function of perfluorinated carbon chain length and functional group. These ratios, referred hereafter as the estimated PFAS vadose zone dilution attenuation factors (DAFs), represent the degree of dilution and attenuation that occurs as PFAS migrate from the vadose zone porewater to underlying groundwater at PFAS-impacted sites examined in this study. A DAF is a dimensionless ratio representing the reduction in contaminant concentration between a source and a downgradient receptor in groundwater driven by processes such as contaminant mixing, adsorption, absorption, and/or dispersion. While traditionally used in the context of groundwater transport, we herein extend this concept to describe PFAS attenuation mechanisms from vadose zone to underlying groundwater including sorption to organic-rich soil fractions and partitioning to air-water interfaces.

Across all concentration percentiles, an apparent correlation between the estimated PFAS vadose zone DAFs and perfluorinated chain length was observed for both perfluorinated carboxylic acids (PFCAs) and perfluorinated sulfonic acids (PFSAs). Elevated DAFs were observed for short-chain PFAS including PFBA and PFBS, where such DAFs commonly exceeded 10:1 particularly within the 25th and 75th percentile datasets. In contrast, substantially lower corresponding estimated DAFs, ranging between 6:1 and 1:1, were seen with longer-chain

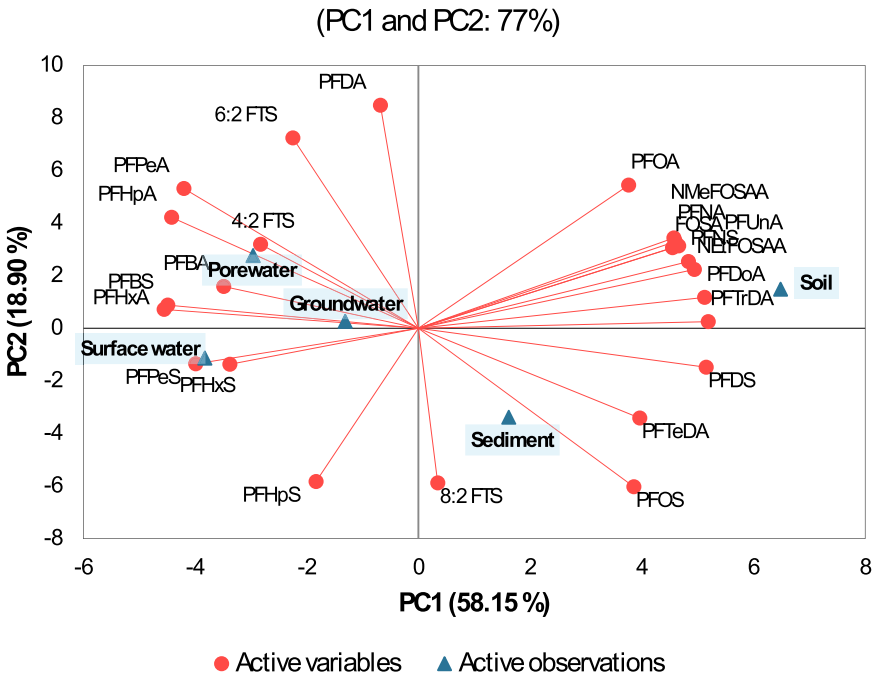


Fig. 4. Relative abundance of target PFAS in various environmental media using PCA. *Maximum PFAS concentrations observed in each media of interest at each site were used.*

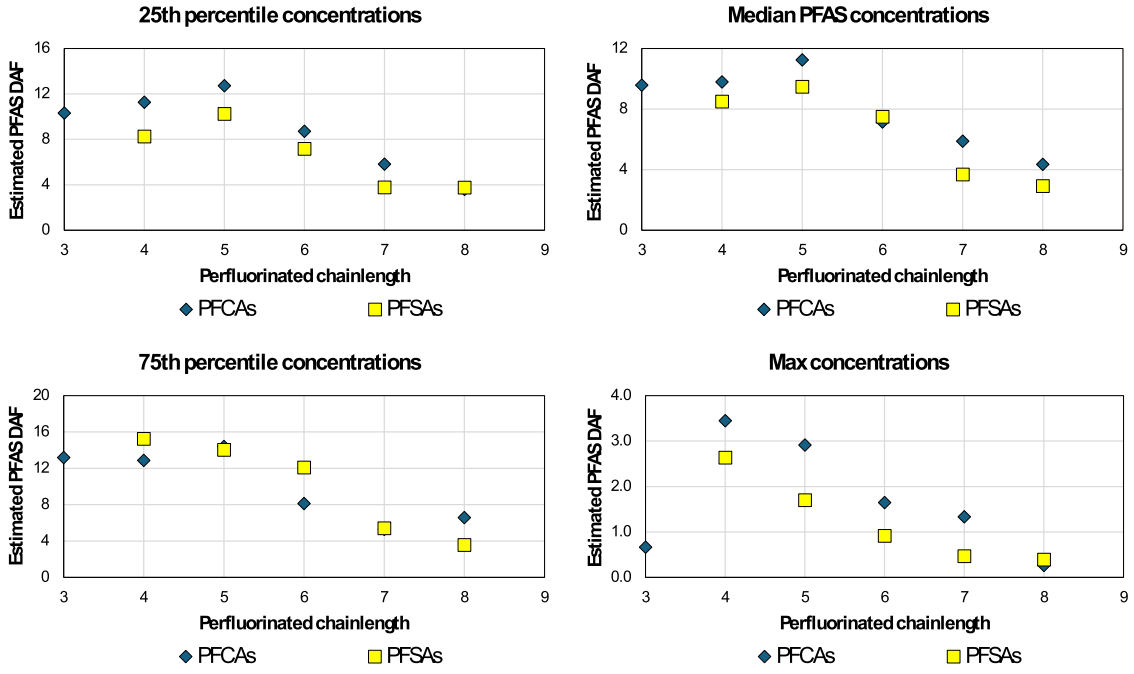


Fig. 5. Estimated PFAS vadose zone DAFs as a function of perfluorinated chain length and functional group. *PFAS vadose zone DAFs were calculated as PFAS porewater concentration (measured via lysimetry in the vadose zone) divided by the corresponding concentrations in the underlying groundwater. Reporting limits were used for non-detects.*

PFAS such as PFOA and PFOS. These apparent correlations are consistent with the physicochemical behavior of PFAS in unsaturated subsurface environments. Shorter-chain PFAS generally exhibit lower hydrophobicity and surface activity, resulting in reduced sorption affinity to soil. Consequently, they are more prone to aqueous dispersion leading to relatively lower groundwater concentrations as a result of groundwater mixing. In contrast, longer-chain PFAS, due to their higher hydrophobicity and stronger affinity sorption to soil, are more heavily retarded limiting their migration into groundwater contributing to lower vadose zone DAFs, particularly for groundwater samples collected at the water table.

The apparent correlations between the estimated PFAS vadose zone DAFs and perfluorinated chain length observed herein also suggest that PFAS concentrations measured in groundwater can be used to estimate the rough order-of-magnitude porewater concentrations that are the primary driver for vertical PFAS migration from vadose zone to

underlying groundwater at AFFF-impacted areas. While lysimetry has been used to directly measure PFAS porewater concentrations at impacted sites [23,19,24-26], it has not been used extensively in all PFAS investigation efforts. Therefore, the aforementioned estimation of PFAS porewater concentrations may have some utility and thus should be further investigated.

3.5. PFAS soil concentration as a function of depth and climate type

Total target PFAS (the log-transformed sum of the 24 PFAS listed in SI-B) distribution in soil as a function of depth and by climate type (arid, semi-arid, sub-humid, and humid) is illustrated in Fig. 6. Across all sites examined, the highest sums of the 24 target PFAS were consistently observed within the vadose zone (i.e., above the water table). In arid and semi-arid climates, these concentrations were predominantly confined to the upper 1 m of soil. In contrast, at sub-humid and humid sites, peak PFAS concentrations were generally detected at slightly greater depths, typically between 1 and 1.5 m below ground surface (bgs). Notably, sites where the highest target PFAS concentrations were observed at deeper depths were generally associated with a sub-humid rather than a humid climate. The observed variability in depth profiles is clearly attributable to differences in evapotranspiration rates and the overall magnitude of flushing within the vadose zone similar to the results of [27]. In sub-humid and humid environments, increased precipitation and sustained soil moisture extend typical wetting fronts and the zero flux plane facilitating the downward transport of PFAS, resulting in peak concentrations observed at greater depths. Additionally, soils in more humid regions often exhibit higher organic matter content and finer textures [28-30] which can influence PFAS sorption dynamics and retardation in the vadose zone. The data presented herein suggest that conditions typically encountered in a sub-humid region consisting of relatively higher precipitation, coarser-grained soils with lower organic matter result in deeper PFAS contamination all else equal. These findings underscore the importance of climate-driven hydrological processes governing the vertical distribution of PFAS in the vadose zone and warrant further investigation.

3.6. Influence of site-specific factors on PFAS maximum concentrations

Fig. 7 graphically illustrates in a 2-dimensional PCA space the relationships among the frequently detected PFAS using the maximum dataset for each media at each site (coded with states and numbers to maintain confidentiality) and site-specific parameters including temperature, precipitation, evaporation, latitude, longitude, and depth to water. The site-specific meteorological information and DTW information are provided in SI-A. Pearson correlation coefficients among co-occurring PFAS are presented in SI-E. The corresponding PCA and HCA results generated using the median dataset are shown in SI-F. Heat maps of site-specific meteorological and DTW information overlaid onto the two-dimensional PC1/PC2 space were generated for better visualization of site characteristics responsible for observed PFAS patterns across different sites and sampling media (SI-G).

In groundwater (Fig. 7**a**), PC1 and PC2 captured approximately 75 % of the data variance. Specifically, PC1 explained 65.18 % of the variance with the greatest contribution from PFPeA, PFHxS, and PFOS whereas PC2 only captured less than 10 % of the data variance and was primarily attributable to PFNA. The highest PFAS concentrations were generally observed at sites characterized by shallow groundwater tables, elevated temperatures, and high evaporation rates. Sites such as CO-02, NV-01, and UT-01 exhibited some of the highest PFAS concentrations. PFNA was rarely detected at high concentrations across most sites with the exception of CO-01. Its distinct separation from other PFAS clusters in the PCA space suggests its potential utility as a site-specific signature compound for forensics purposes. The short-chain and mobile PFBA was not commonly found at high concentrations in groundwater at these sites. In porewater (Fig. 7**b**), PC1 and PC2 explained 52.97 % and

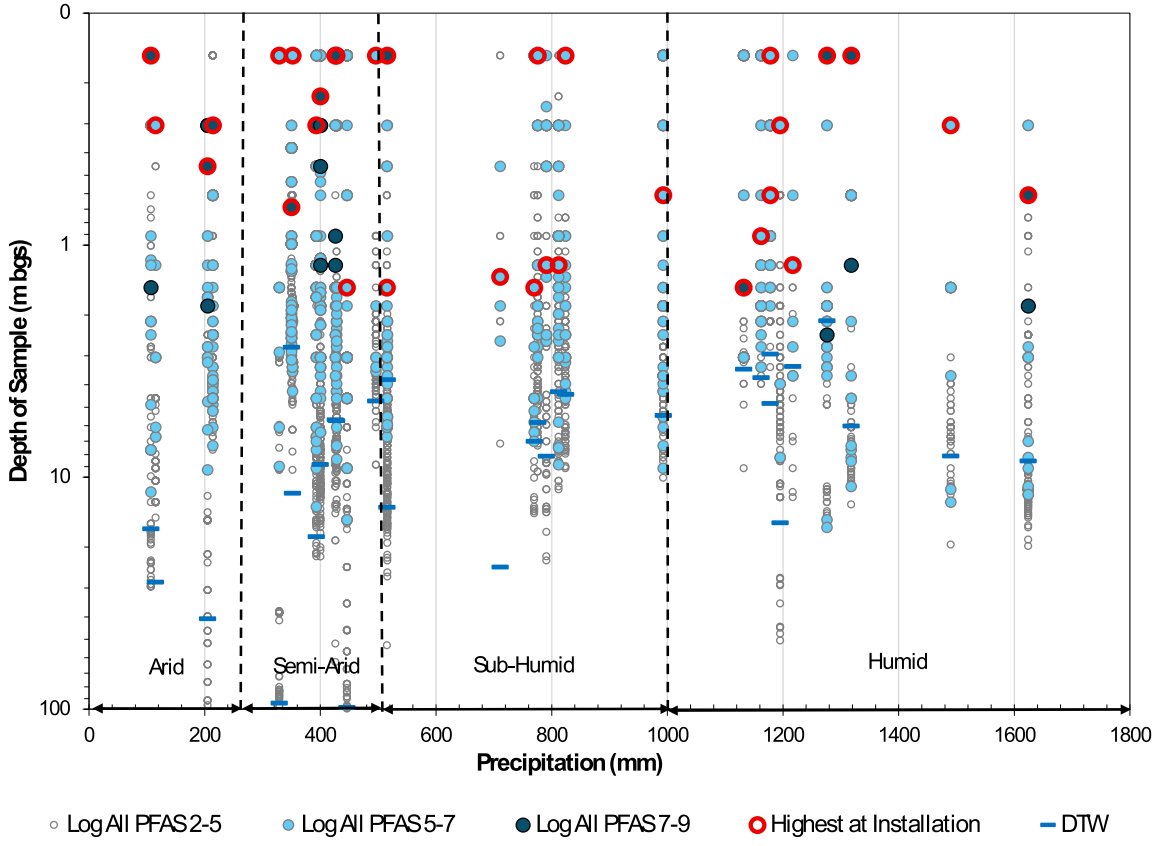


Fig. 6. Total target PFAS distribution in soil as a function of depth and climate type. *Reporting limits were used for non-detects.*

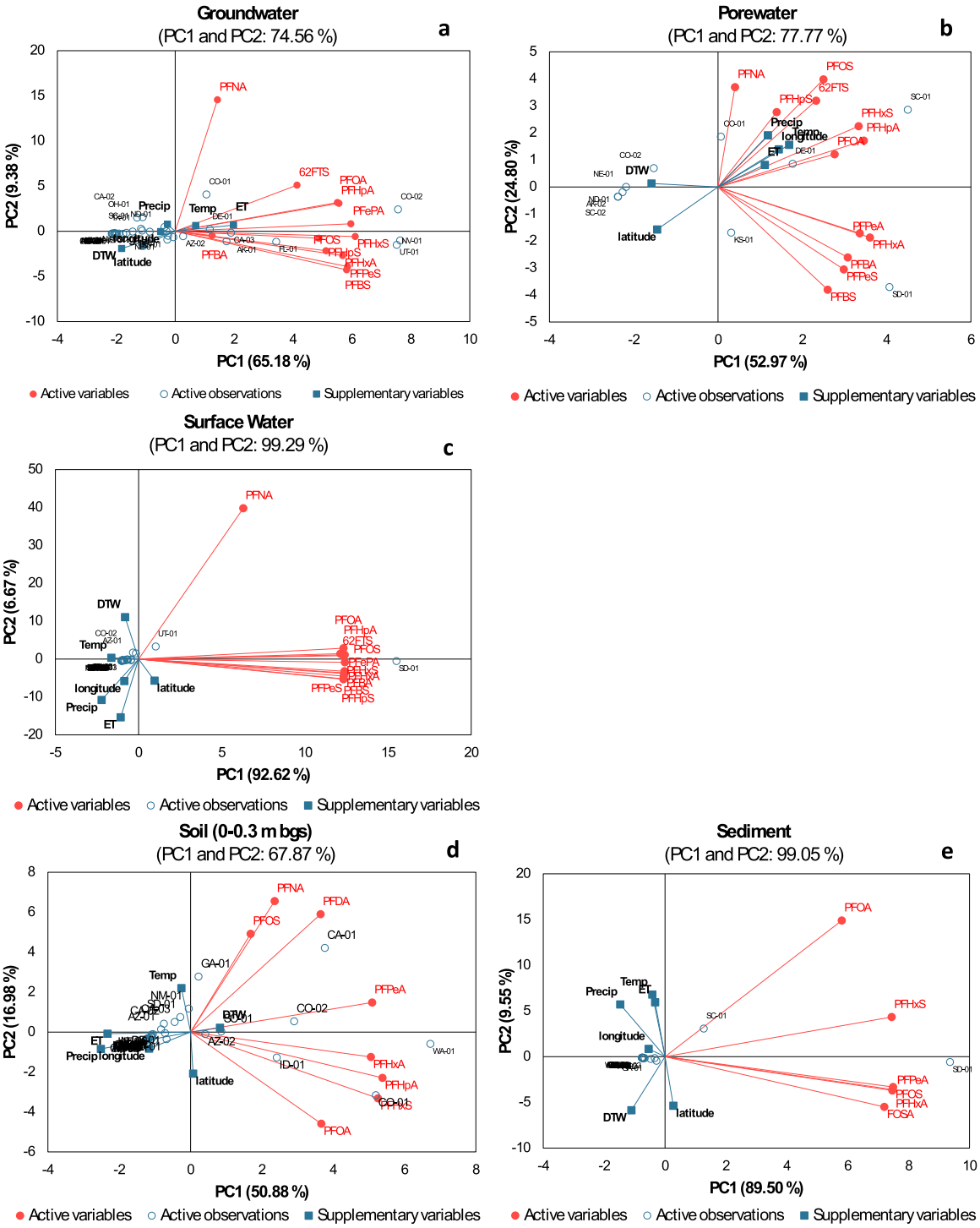


Fig. 7. PCA results for groundwater (a), porewater (b), surface water (c), surficial soil (0–0.3 m bgs) (d), and sediment (e). *Maximum PFAS concentrations observed at each site used. RROS used to assign values to non-detects.*

24.80 % of the data variance, respectively. PFAS concentration patterns in porewater were consistent with those observed in groundwater, where the highest levels were typically found at sites with shallow groundwater, warm climates, and high precipitation and evaporation. In contrast to groundwater, short-chain PFAS including PFBA was more consistently observed at high concentrations in porewater. Additionally, a more distinct clustering of short- versus long-chain PFAS was observed in porewater. The highest concentrations of long-chain PFAS including PFOS, 6:2 FTS, PFHxS, and PFHpA were seen at SC-01 whereas SD-01 exhibited the highest concentrations of short-chain PFAS including PFBA, PFBS, and PFPeS. Similar to groundwater, the highest PFNA concentrations detected in porewater were found at CO-01. In surface water (Fig. 7**c**), nearly all data variance (99.09 %) was captured by PC1 with no apparent clustering of different PFAS. All PFAS in the surface water maximum data clustered together with the exception of PFNA, suggesting that many PFAS are commonly observed in surface water. Such a finding would fit the typical release profile in that AFFF contains a high percentage of water, which would be expected to run off into surface water bodies during a firefighting event. The highest PFAS concentrations, comprising both short- and long-chain compounds, were

observed at SD-01. In surficial soil collected in the upper 1 foot (Fig. 7**d**), PC1 and PC2 explained 50.88 % and 16.98 % of the data variance, respectively. Elevated PFAS levels were generally associated with sites experiencing low precipitation, low evapotranspiration, and deeper groundwater tables. Distinct clustering of short- versus long-chain PFAS was observed in soil. Long-chain PFAS such as PFNA and PFDA were predominantly found at CA-01, while the highest PFHxS concentrations were detected in soil at CO-01. In sediment (Fig. 7**e**), PC1 accounted for 89.50 % of the variance with strong contribution from PFHxS and PFOS. Sediment samples from SC-01 were primarily composed of PFOA whereas the highest PFHxS and PFOS concentrations were observed in sediment at SD-01.

It is important to note that limited surface water and sediment samples were collected in the data set used for this study, which may have influenced the findings presented herein. Additionally, similar PFAS clustering and influence of site-specific characteristics on PFAS distribution were generally observed when the median PFAS dataset was used in the analysis described (SI-E). The consistency in key findings across the maximum and median datasets (representative of source versus plume conditions, respectively) may suggest that AFFF releases were more likely spatially widespread than typical point-source releases associated with legacy contaminants such as chlorinated solvents.

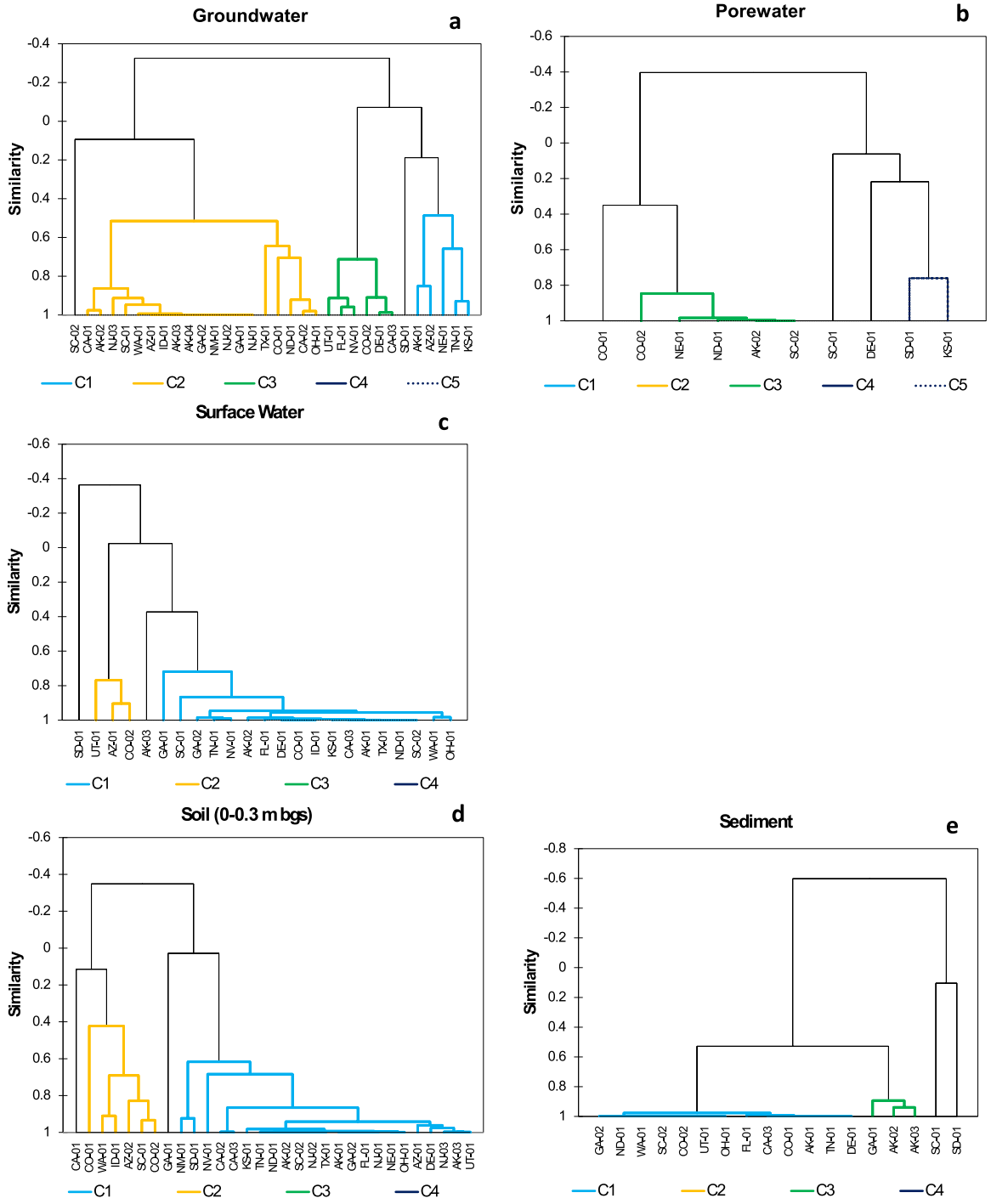


Fig. 8. HCA results for groundwater (a), porewater (b), surface water (c), surficial soil (0–0.3 m bgs) (d), and sediment (e). *Maximum PFAS concentrations observed at each site used. RROS used to assign values to non-detects.*

Unlike traditional plumes where contaminants are transported downgradient from discrete release points, the similar patterns in PFAS soil distribution observed in source and downgradient areas suggest broader dispersal mechanisms at these sites, likely due to the nature of AFFF application and site use history.

3.7. Hierarchical clustering analysis

HCA was performed to assess the similarity among environmental samples collected from five distinct media: groundwater, porewater, surface water, soil, and sediment. The resulting dendrograms (Fig. 8) illustrate the degree of similarity among sites within each media, with clustering patterns providing insight into the heterogeneity and potential environmental processes influencing each system. The groundwater dendrogram revealed the highest degree of structural complexity, with five distinct clusters (C1–C5) identified. This pronounced clustering suggests substantial heterogeneity among groundwater samples, potentially reflecting temporal, spatial, meteorological, and hydrogeological variability as well as localized differences in AFFF release mechanisms and AFFF formulations at the sites examined in this study.

In contrast, the surface water samples exhibited a relatively homogeneous clustering pattern, with most samples grouped into a single dominant cluster (C1). This uniformity likely reflects the common process in which PFAS migration occurs from surficial soil to surface water runoff due to the high volume of water released with the AFFF, during high precipitation events or via groundwater-surface water interactions. The porewater and soil dendrograms displayed intermediate levels of clustering complexity, each forming two to three distinct clusters. These patterns indicate moderate heterogeneity, potentially driven by localized variations in soil lithology, organic content, depth to groundwater, and other factors. The sediment dendrogram was characterized by a lack of well-defined clusters, with most samples remaining ungrouped and only a single small cluster (light blue) identified. This lack of a distinct clustering pattern suggests high variability in sediment composition, which may be attributed to heterogeneous depositional environments, varying grain sizes, or differential diagenetic processes. It should be noted that a limited number of sediment samples were available in the dataset and such limitations may have impacted the results observed herein. However, the dendrograms highlight distinct patterns of environmental heterogeneity across media, with groundwater and potentially sediment exhibiting the greatest complexity, and surface water the least. Across all media, the grouping of sites via hierarchical clustering was consistent with results generated by PCA.

4. Conclusions

Results from this statistical analysis of a large PFAS dataset examined in this study suggest that PFAS composition varies considerably within different environmental media at AFFF-impacted sites. In our dataset, PFOS was the most prevalent PFAS detected in soil, sediment, and surface water whereas PFHxS was predominantly detected in groundwater and porewater. Shorter-chain PFAS with higher relative mobility are generally detected with the highest relative abundance in surface water. On the other hand, long-chain PFAS and PFAA precursors with relatively low mobility such as PFNS, NetFOSA, NMeFOSA, PFDoA, and PFTrDA are generally observed in soil samples. Consistent with observations of the soil data, select long-chain PFAAs including PFDS, PFTeDA, and PFOS are more likely found in sediment samples. Short-chain PFAS with relatively high mobility such as 4:2 FTS, PFBA, PFBS, PFPeA, and PFHpA are commonly found in both porewater and groundwater, albeit with different degrees of contribution.

An apparent correlation between the estimated PFAS vadose zone DAFs (defined as the ratio between PFAS porewater concentration measured in the vadose zone and that measured in the corresponding underlying groundwater) and perfluorinated chain length was observed for both PFCAs and PFSAs, confirming that the degree to which PFAS migrate from the vadose zone to underlying groundwater is highly dependent on their physicochemical properties and that PFAS concentrations measured in groundwater may be used to estimate rough order-of-magnitude porewater concentrations. Although lysimetry has been used to directly measure PFAS porewater concentrations at impacted sites, it has not been used extensively in PFAS remedial investigations. In some cases, a gross estimation of PFAS porewater concentrations based on groundwater monitoring data (which is more commonly collected) may have some utility, particularly in risk assessments, and thus should be further investigated.

Across all sites examined, the highest sums of the 24 target PFAS were consistently observed within the vadose zone. In arid and semi-arid climates, these concentrations were predominantly confined to the upper 1 m of soil. In contrast, at sub-humid and humid sites, peak PFAS concentrations were generally detected at slightly greater depths, typically between 1 and 1.5 m bgs. Notably, sites where the highest target PFAS concentrations were observed in soil at deeper depths were generally associated with the sub-humid instead of the humid climate. The observed variability in PFAS soil distribution across different climate regimes may be attributable to differences in soil properties, hydrogeological conditions, and PFAS' physicochemical properties. Distinct PFAS and site clustering patterns were identified using PCA and HCA, demonstrating their ability to reduce the multidimensional complexity typically observed with large PFAS datasets. PCA was particularly beneficial in discerning PFAS compositional patterns that are difficult to visualize using conventional analytical techniques.

The consistency in key statistical findings across the maximum and median datasets (representative of source versus plume conditions, respectively) may suggest that AFFF releases were more likely spatially widespread than typical point-source releases associated with legacy contaminants such as chlorinated solvents. The similar patterns in PFAS soil distribution observed in source and downgradient areas suggest broader dispersal mechanisms at sites examined in this study, likely due to the nature of AFFF application and site use history. Additional work should be performed to determine whether these findings are specific to the 29 sites examined in this study or broadly applicable across other PFAS-impacted sites. This distinction could provide critical insights for refining the site-specific knowledge framework, improving the accuracy of risk assessments and guiding the development of targeted sampling and remediation strategies. Such efforts would help ensure that management approaches account for the unique distribution patterns and transport behaviors associated with AFFF-derived PFAS, ultimately supporting more effective and defensible decision-making.

It should be noted that all PFAS analyses for this study were conducted using EPA Method 537. While appropriate at the time, such method is now limited in scope relative to more recent analytical advancements. As newer PFAS analytical methods such as EPA Method 1633 become more widely adopted and ultrashort-chain PFAS (e.g., trifluoroacetic acid) receive increased regulatory attention, future investigations would benefit from incorporating these advanced PFAS analytical methods and expanded PFAS analyte lists to develop more unique chemical signatures and to further augment the utility of PCA. Additionally, due to the confidential nature of the sites examined in our study, the site clustering patterns identified by HCA were not further complemented by geospatial analyses, which could otherwise provide additional insights into PFAS source allocation and attribution. These study-specific limitations should be considered when interpreting the robustness of multivariate findings, adapting statistical models to site-specific conditions, and identifying opportunities where additional analytical and spatial data can strengthen forensic interpretations and support informed site management decisions.

Environmental Implication

Results from this study demonstrate that traditional and multivariate statistical techniques including PCA and HCA can serve as

complementary tools for simplifying and interpreting large PFAS datasets. Insights on PFAS distribution across various media examined can support better understanding of site-specific conditions, inform risk assessments, and guide future sampling and remediation strategies at PFAS-impacted sites.

CRediT authorship contribution statement

Lisa Kammer: Writing – review & editing, Validation, Supervision. **Teresa Verstraet:** Writing – original draft, Formal analysis, Data curation. **Matt Anding:** Writing – review & editing, Supervision. **Taire Van Scoy:** Writing – review & editing, Validation, Supervision. **Sonya Cadle:** Writing – original draft, Validation, Project administration, Methodology, Investigation, Formal analysis. **Dung Nguyen:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Richard Anderson:** Writing – review & editing, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jhazmat.2025.140539.

Data availability

Data will be made available on request.

References

[1] Glüge, J., Scheringer, M., Cousins, I.T., DeWitt, J.C., Goldenman, G., Herzke, D., et al., 2020. An overview of the uses of per- and polyfluoroalkyl substances (PFAS). Environ Sci: Process Impacts 22 (12). https://doi.org/10.1039/d0em00291g.

[2] Anderson, R.H., Long, G.C., Porter, R.C., Anderson, J.K., 2016. Occurrence of select perfluoroalkyl substances at U.S. air force aqueous film-forming foam release sites other than fire-training areas: field-validation of critical fate and transport properties. Chemosphere 150, 678–685. https://doi.org/10.1016/j.chemosphere.2015.11.036.

[3] Hu, X.C., Andrews, D.Q., Lindstrom, A.B., et al., 2016. Detection of poly- and perfluoroalkyl substances (PFASs) in U.S. drinking water linked to industrial sites, military fire training areas, and wastewater treatment plants. Environ Sci Technol Lett 3 (10), 344–350. https://doi.org/10.1021/acs.estlett.6b00260.

[4] Agency for Toxic Substances and Disease Registry (ATSDR). Toxicological profile for perfluoroalkyls; 2021. ⟨https://www.atsdr.cdc.gov/toxprofiles/tp200.html⟩.

[5] National Academies of Sciences, Engineering, and Medicine (NASEM). Guidance on PFAS exposure, testing, and clinical follow-up; 2022. https://doi.org/10.17226/26156.

[6] Environmental Protection Agency (EPA). PFAS strategic roadmap: EPA's commitments to action 2021–2024; 2024. ⟨https://www.epa.gov/pfas⟩.

[7] Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philos Trans R Soc A Math Phys Eng Sci 374 (2065), 20150202. https://doi.org/10.1098/rsta.2015.0202.

[8] Wang, Z., Cousins, I.T., Scheringer, M., Hungerbuehler, K., 2021. Multivariate statistical analysis of PFAS profiles in environmental samples: a review. TrAC Trends Anal Chem 136, 116171. https://doi.org/10.1016/j.trac.2021.116171.

[9] Andrews, D.Q., DeWitt, J.C., Grandjean, P., 2022. Applying analytical chemistry and epidemiology to inform risk management of PFAS. Environ Health 21, 70. https://doi.org/10.1186/s12940-022-00884-z.

[10] Guelfo, J.L., Adamson, D.T., 2018. Evaluation of a national data set for insights into sources, composition, and concentrations of per- and polyfluoroalkyl substances (PFASs) in U.S. drinking water. Environ Pollut 236, 505–513. https://doi.org/10.1016/j.envpol.2018.01.066.

[11] Yu, N., Guo, H., Yang, J., et al., 2020. Application of cluster analysis and principal component analysis in identifying potential sources of per- and polyfluoroalkyl substances (PFASs) in contaminated sites. Sci Total Environ 729, 138839. https://doi.org/10.1016/j.scitotenv.2020.138839.

[12] Zhang, X., Lohmann, R., Dassuncao, C., Hu, X.C., Weber, A.K., Vecitis, C.D., et al., 2016. Source attribution of poly- and perfluoroalkyl substances (PFASs) in surface waters from Rhode Island and the New York metropolitan area. Environ Sci Technol Lett 3, 316–321.

[13] Evans, P.J., Nguyen, D.D., Chappell, R.W., Whiting, K., Gillette, J., Bodour, A., et al., 2014. Factors controlling in situ biogeochemical transformation of trichloroethene: column study. Groundw Monit Remediat 34 (3), 65–78.

[14] Kulkarni, P.R., Andrzejczyk, N.E., Gavaskar, A., Cartwright, A., Adamson, D.T., Cook, J., et al., 2025. Characteristics of aqueous film forming foam (AFFF) sites impacted with per- and polyfluoroalkyl substances (PFAS): a 37-site study. Water Res 285, 124124. https://doi.org/10.1016/j.watres.2025.124124.

[15] Muniz, D.H.F., Oliveira-Filho, E.C., 2023. Multivariate statistical analysis for water quality assessment: a review of research published between 2001 and 2020. Hydrology 10 (10), 196. https://doi.org/10.3390/hydrology10100196.

[16] Patel, P.S., Pandya, D.M., Shah, M., 2023. A holistic review on the assessment of groundwater quality using multivariate statistical techniques. Environ Sci Pollut Res 30, 85046–85070. https://doi.org/10.1007/s11356-023-27605-x.

[17] Saab, C., Zéhil, G.P., 2025. Statistical analysis techniques in water quality monitoring: a review. Stoch Environ Res Risk Assess 39, 3723–3760. https://doi.org/10.1007/s00477-025-03035-8.

[18] Helsel, D.R., 2005. Nondetects and data analysis: statistics for censored environmental data. Wiley-Interscience, Hoboken, NJ, p. 250 (p xv).

[19] Anderson, R.H., Field, J.B., Dieffenbach-Carle, H., Elsharnouby, O., Krebs, R.K., 2022. Assessment of PFAS in collocated soil and porewater samples at an AFFF-impacted source zone: field-scale validation of suction lysimeters. Chemosphere 308 (1), 136247.

[20] Brusseau, M.L., Anderson, R.H., Guo, B., 2020. PFAS concentrations in soils: background levels versus contaminated sites. Sci Total Environ 740, 140017.

[21] McGarr, J.T., Mbonimpa, E.G., McAvoy, D.C., Soltanian, M.R., 2023. Fate and transport of per- and polyfluoroalkyl substances (PFAS) at aqueous film forming foam (AFFF) discharge sites: a review. Soil Syst 7 (2), 53. https://doi.org/10.3390/soilsystems7020053.

[22] Lumivero. XLSTAT – advanced analytics in excel; 2025. ⟨https://lumivero.com/products/xlstat/⟩.

[23] Anderson, R.H., 2021. The case for direct measures of soil-to-groundwater contaminant mass discharge at AFFF-impacted sites. Environ Sci Technol 55 (10), 6580–6583. https://doi.org/10.1021/acs.est.1c01543.

[24] Costanza, J., Clabaugh, C.D., Leibli, C., Ferreira, J., Wilkin, R.T., 2025. Using suction lysimeters for determining the potential of per- and polyfluoroalkyl substances to leach from soil to groundwater: a review. Environ Sci Technol 59 (9), 4215–4229. https://doi.org/10.1021/acs.est.4c10246.

[25] Rayner, J.L., Lee, A., Corish, S., Leake, S., Bekele, E., Davis, G.B., 2024. Advancing the use of suction lysimeters to inform soil leaching and remediation of PFAS source zones. Groundw Monit R 44, 49–60. https://doi.org/10.1111/gwmr.12670.

[26] Schaefer, C.E., Lavorgna, G.M., Lippincott, D.R., Nguyen, D.D., Christie, M., Shea, S., et al., 2022. A field study to assess the role of air-water interfacial sorption on PFAS leaching in an AFFF source area. J Contam Hydrol 248, 104001.

[27] Anderson, R.H., Adamson, D.T., Stroo, H.F., 2019. Partitioning of poly-and perfluoroalkyl substances from soil to groundwater within aqueous film-forming foam source zones. J Contam Hydrol 220, 59–65.

[28] Amooh, M.K., Bonsu, M., 2015. Effects of soil texture and organic matter on evaporative loss of soil moisture. J Glob Agric Ecol 3 (3), 152–161.

[29] Franzluebbers, A.J., 2022. Soil organic matter, texture, and drying temperature effects on water content. Soil Water Manag Conserv 86 (4), 1086–1095.

[30] Li, H., Bulcke, J.V., Mendoza, O., Deroo, H., Haesaert, G., Dewitte, K., et al., 2022. Soil texture controls added organic matter mineralization by regulating soil moisture – evidence from a field experiment in a maritime climate. Geoderma 410, 115690.