IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 9, 2020, accepted July 20, 2020. Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2020.3013108

Signal Processing Methods to Interpret Polychlorinated Biphenyls in Airborne Samples

RYAN A. MCCARTHY[1], ANANYA SEN GUPTA[1,2], (Member, IEEE), BERNICE KUBICEK[1], ANDREW M. AWAD[3,4], ANDRES MARTINEZ[3,4], RACHEL F. MAREK[3,4], AND KERI C. HORNBUCKLE[3,4]

[1]Department of Electrical and Computer Engineering, The University of Iowa, Iowa City, IA 52242, USA
[2]Iowa Technology Institute (ITI), The University of Iowa, Iowa City, IA 52242, USA
[3]Department of Civil and Environmental Engineering, The University of Iowa, Iowa City, IA 52242, USA
[4]IIHR-Hydroscience and Engineering, The University of Iowa, Iowa City, IA 52242, USA

Corresponding author: Ananya Sen Gupta (ananya-sengupta@uiowa.edu)

This work was supported in part by the University of Iowa Center for Health Effects of Environmental Contamination (CHEEC), The University of Iowa through the Iowa Superfund Research Program under Grant NIEHS P42 ES 013661, and in part by the National Science Foundation under Grant 1808463.

ABSTRACT The main contribution of this interdisciplinary work is a robust computational framework to autonomously discover and quantify previously unknown associations between well-known (target) and potentially unknown (non-target) toxic industrial air pollutants. In this work, the variability of polychlorinated biphenyl (PCB) data is evaluated using a combination of statistical, signal processing, and graph-based informatics techniques to interpret the raw instrument signal from gas chromatography-mass spectrometry (GC/MS/MS) data sets. Specifically, minimum mean-squared techniques from the adaptive signal processing literature are extended to detect and separate coeluted (overlapped) peaks in the raw instrument signal. A graph-based visualization is provided which bridges two complementary approaches to quantitative pollution studies: (i) peak-cognizant target analysis (limits data analysis to few well-known compounds) and (ii) chemometric analysis (statistical large-scale data analysis) that is agnostic of specific compounds. Further, peak fitting techniques based on L2 error minimization are employed to autonomously calculate the amount of each PCB present with a normalized mean square error of -18.4851 dB. Graph-based visualization of associations between known and unknown compounds are developed through principal component analysis and both fuzzy c-means (FCM) and k-means clustering techniques are implemented and compared. The efficiency of these methods are compared using 150 air samples analyzed for individual PCBs with GC/MS/MS against traditional target-only techniques that perform analysis across only the known (target) PCBs. Parameter optimization techniques are employed to evaluate the relative contribution of PCB signals against ten potential source signals representing legacy signatures from historical manufacture of Aroclors and modern sources of PCBs produced as byproducts of pigment and polymer manufacturing. Aroclors 1232, 1254, 1016, and 1221 as well as non-Aroclor 3, 3', dichlorobiphenyl (PCB 11) were found in many of the samples as unique source signals that describe PCB mixtures in air samples collected from Chicago, IL.

INDEX TERMS Identifying sources, interpreting GC/MS/MS, PCBs, signal processing.

I. INTRODUCTION

Recent years have seen a surge in interdisciplinary research [1]–[6] combining signal processing and related analytical techniques for interpretation of environmental data sets [2]–[4], [6]–[14]. Statistical techniques [3], [6]–[13] have been employed successfully in interpretation of polychlorinated biphenyl (PCB) data. PCBs are a set of 209 bioaccumulating, persistent, and toxic compounds that are widely found in the environment worldwide. The 209 possible PCB compounds are referred to as congeners belonging to ten sets of PCB homolog isomers each with the same molecular mass. The relative concentrations of each PCB congener measured in active air samples varies as a function of proximity to sources, their specific physicochemical properties, meteorological conditions, and historical use [15], [16]. Although not intentionally produced today, PCBs are byproducts to certain manufacturing processes, still present

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

throughout the environment, and pose multiple health risks to those exposed to them [17]–[20].

PCBs have many different sources which are affected by microbial and environmental processes that change the relative mass of each congener. In a formative study from 1997, Frame analyzed commercial PCB mixtures called Aroclors for their specific PCB content [21]. Each Aroclor has a unique signal which can be used to determine the product or process producing these PCBs. This identification is useful for exposure management and remediation. This work hypothesizes that the evaluation of the raw chromatographic signal will uncover information about the environment that would not be detected from target analysis of the individual PCBs. Specifically, minimum mean-squared error techniques are extended in combination with adaptive signal processing which has been proposed in recent literature [22]–[28] to detect, separate, and analyze coeluted peaks within the raw signal to bridge the gap between peak-cognizant target analysis and statistical chemometric analysis.

This paper is divided into six sections. The remainder of this section describes the instruments used, current-state-of-the-art methods, and key contributions of this work in associating dominant peaks to hidden peaks within the signal as well as sampling locations. Section II describes the collection and extraction of sample data. Section III presents the peak fitting procedures of the signals and Section IV describes the association techniques implemented within this work. Finally, Sections V and VI discusses and compares the results produced from this work and presents conclusions.

A. INSTRUMENT DESCRIPTION AND PREPARATION OF THE RAW INSTRUMENT SIGNAL

The instrumentation used for generating the data used in this work identifies PCBs through gas chromatography-mass spectrometry (GC/MS/MS) in multiple reaction monitoring mode (MRM). This provides selective signal separations of all the known (target) PCBs by their homolog isomer for each sample [29], [30]. The selective signal separations in this work are exploited by fitting curves to the peaks to identify target PCBs within each sample, detailed in Section III. Direct peak interpretation is difficult due to retention time shifts of target PCBs on the chromatograms and non-linearity from sample to sample. To overcome this challenge, PCBs are typically measured in environmental samples by comparing the signal of a calibration standard solution run through GC/MS/MS in MRM mode of target PCB content against that of the prepared environmental sample [31], [32]. The raw instrument signal after this pre-processing is an information-rich signal representative of PCBs in the environment at the time the sample was collected.

B. CURRENT STATE-OF-THE-ART METHODS

Signal processing techniques employing constrained optimization techniques for peak extraction have been exhaustively researched in beamforming and sonar localization literature [33]–[37] as well as other applications e.g. real-time brain activity and heart rate monitoring [38], [39], to name a few. Signal processing techniques have also been employed for raw instrumental signal interpretation to provide better analysis in determining environmental pollutants [40], [41]. However, despite these recent computational advances, peak-cognizant raw signal interpretation beyond target compounds remains an open challenge, particularly for studying toxic air pollutants such as PCBs. Further discussion is provided below.

C. TARGET ANALYSIS VS. CHEMOMETRIC ANALYSIS

The raw instrument signal from gas-chromatographic and mass spectrometric instruments carries a wealth of information on the composition of complex mixtures [2], [42]–[46]. However, most chemical analysis and expert interpretation of the raw signal is target-based (e.g. [42] and references in [44]) i.e., focused only on the contribution of target compounds whose chemical properties are well-known, and which occupy specific positions in the retention time of the instrument signal. Target analysis, while extremely important and relevant to interpret the dominant or known part of the instrument signal, provide limited opportunity to exploit the full informational power that sophisticated analytical hardware can offer. For example, hundreds if not thousands, of non-target compounds that manifest as unknown peaks within the raw instrument signal can provide hitherto unforeseen knowledge of environmental pollutants within passive air samples.

Raw signal analysis itself is not new. The rich and growing field of chemometrics [47]–[49] already provide many statistical techniques to analyze the peaks within the raw instrument signal on a large scale. However, purely statistical methods are compound-agnostic and as such, provide insight into the aggregate behavior of the raw signal, e.g. dominant trends in a principal component analysis (PCA) [47]. Aggregate studies are useful to understand broader trends but, as yet, are not designed to detect compound-specific information, particularly from unknown toxic contaminants that can be buried in the larger statistical behavior of the raw signal (e.g. hidden against more dominant targets, or aligned along less dominant PCA components). This is an important distinction against the currently available techniques in both environmental chemistry and statistical methods; as currently, no technique exists that can discover and disentangle the signature of highly toxic yet unknown contaminants which chemists do not look for and which peak-agnostic multivariate chemometric analysis fail to detect. There is, therefore, a compelling need to bridge the gap between purely target-driven methods, as pursued by chemists that provide in-depth knowledge of a few well-known compounds, and purely statistical methods, which are compound-agnostic.

The GC/MS/MS MRM raw signal interpretation routinely excludes many of the non-targeted analytes found in the samples thus eliminating key connections between non-target analytes and target PCBs in an environmental sample. In this work, non-target analytes are defined as chemicals that have

gone through the GC/MS/MS and appear as peaks in the chromatograms in more than 50% of the samples but are not in the calibration solutions. A more comprehensive data interpretation can be achieved without such filtering; however, these non-target analytes are traditionally ignored to improve detection and identification of target PCBs within each sample.

Current-state-of-the-art analysis of PCBs using GC/MS/MS focus solely on target PCBs and ignore other potential co-indicators of PCB sources. Presently, modeling GC/MS/MS and calculating the contribution of sources have produced many algorithms to aid in the process. Listed next are approaches from the current art that are well known and have offered important advancements in the calculation of PCBs and sources. Discussion of how the proposed computational techniques complement and potentially enhance the current art in raw signal interpretation can be found throughout the manuscript.

1) MODELING GC/MS/MS - Recent computational techniques proposed include algorithms that model raw gas chromotagrophic signals, e.g. PARAFAC and PARAFAC2 [7], [8], [14], [50]. PARAFAC uses a N-way PCA decomposition method assuming low-rank N-linearity which breaks down the array into sets of scores and loadings that are mainly unique estimates of the underlying peaks in the data. Further, using an alternating least squares algorithm, the model fits the curve of the analyte in the total ion chromatogram (TIC) data set [14], [50]. While PARAFAC can locate peaks in a sample, it struggles to detect the retention time shifts of the peaks from sample to sample and can overlook analytes that are small peaks in the data [7], [50]. In PARAFAC2 however, the data does not require low-rank linearity and can allow deviation in the data. This decomposition similarly breaks down the array into a set of scores and loading which are unique estimates of the underlying data and uses similar procedures as PARAFAC [7], [8], [50]. Although PARAFAC2 fixes the retention time shift problem by using a time loading matrix for each sample and a one-component model, it is still challenged in determining the number of components required for the peak, identifying PCBs, and finding other reoccurring chemicals within the chromatographic data sets [8], [50]. The motivation in this work is to complement such modeling approaches and employ robust signal processing techniques that glean as much peak information as possible against the ambient noise in the raw signal. This enables robust joint and compound-cognizant interpretation of target and non-target peaks from the raw signal.
2) IDENTIFYING SOURCES - Analysis to identify sources can be done using linear regression models [9]–[13] or even positive matrix factorization (PMF) [10]–[13], [51]–[54]. These algorithms attempt to solve for the various percentages of sources through linear equations. Although these algorithms pose unique solutions to the problem, they are limited in their estimations. Linear regression models are limited to linear relationships and is sensitive to outlier data when calculating mixture percentages, and PMF requires source weights that can influence the outcomes of the percentages.

D. BACKGROUND MOTIVATION

PCBs are frequently detected in different environmental compartments such as air, water and sediment, and even in human serum, and provide a well-established basis to compare different samples. The interpretation of target PCBs and sources can be greatly enhanced by incorporating non-target analytes found within the GC/MS/MS topography. While recent chromatography interpretations of GC/MS/MS data sets have proposed various methods to locate target PCBs, approaches to finding non-target analytes are rare. Moreover, contributions of various analytical techniques to target PCBs have partial limitations and would benefit from deeper analysis.

E. KEY CONTRIBUTIONS

The scope of this work is to automate and enhance target-centric raw signal processing such that the end result is a compound-cognizant graph-based peak profile. The value of the work lies in automating the process to avoid human confirmation bias in peak selection, while also allowing human interpretation using the peak-cognizant graph visualization as well as statistical clustering analysis presented in this work. This approach connects peak-specific interpretation, as is commonly done in traditional target analysis and related peak-mapping efforts [2], [42]–[46], to purely chemometric interpretation [47]–[49]. Furthermore, as noted in Section V-G and related discussion for Table 1, the technique is capable of isolating non-target peaks that may coelute or elute in close proximity to target peaks. Through these techniques, as when applied to target-centric raw signals, such as presented here, the contribution of non-targets can be isolated and quantified. Such non-target identification can detect chemical threats from toxic contaminants, e.g. introduced into the environment by hostile agents or as byproducts of unknown sources, which would otherwise remain undiagnosed in routine target and regulatory analysis. This is particularly applicable to toxins that are similar in chemical composition and retention time to known targets, and hence will be captured in the raw GC-MS signal, but only as non-targets which may coelute or elute in close proximity.

More specifically, the aim of this work is to propose and test novel combinations of peak fitting PCB chromatograms, applying principal component analysis (PCA) and both k-means and c-means clustering, L2 minimization calculations to analyze sources and mixtures of contaminants, and provide a further examination of signals more topologically to develop deeper analysis of the data. An important distinction between the technique presented here and other methods is the potential to discover hidden peaks within the samples through automating (peak-cognizant) detection and interpretation while preserving the identity

of target peaks within the signal. In this case, the peaks within the MRM data sets are fitted and target peaks are autonomously identified based on the calibrations and their retention times. Once performed, hidden peaks are identified from the remaining peaks within the MRM or TIC signals. The objective is to extend the scope of target PCB analysis to include detection of non-target peaks within the various samples and better identify sources of PCBs in the sample locations. While target peaks dominate the GC/MS/MS signal, the unutilized contribution of non-target peaks can also be employed to distinguish related samples. Employing these techniques and methods will significantly enhance the already well-established knowledge of PCBs and sources through deeper analysis of GC/MS/MS data. Further, these techniques can relate compound-cognizant target analysis with compound-agnostic and purely statistical approaches to create an in-depth dictionary of underlying information hidden within the signals.

II. DESCRIPTION OF DATA

A. DATA

The data set originated from 150 air samples collected with active high-volume air samplers (Hi-Vols) deployed across the Chicago metropolitan area from 2007 to 2009 [29]. The data for this paper were generated by instrument analysis as follows: sample extracts were analyzed using a GC-MS/MS (Agilent 7000 Triple Quad with Agilent 7890A GC and Agilent 7693 autosampler equipped with a Supelco SPB-Octyl capillary column) in multiple-reaction monitoring (MRM) mode [18], [20]. Analytical quality control included surrogate standards recoveries, replicates, laboratory and field blanks, and standard reference material. The MRMs produce twelve chromatograms for each sample, representing the chromatographic signal for different mass transition ions (10 transitions for unlabeled PCBs and 2 for mass-labeled PCB standards). Further, one total ion chromatogram (TIC) is obtained for each sample, representing the combined MRM signals. The same MRM and TIC were obtained for the calibration solutions containing 209 PCB congeners. Theoretically, each peak within the TIC and MRM signal corresponds to a PCB congener or, in a lesser extent, to a non-target chemical found in the sample. Heights or relative intensity were used as the total amount of the PCB congener or chemical found. The heights of the peaks in the calibration solutions were used for calculating the mass of each congener through computation of the relative response factor (RRF) [29].

B. SOFTWARE

The algorithms and analysis done in this work were developed with the MATLAB R2018a software (The Mathworks, Inc. USA). The following toolboxes were installed with the MATLAB 2018a software to implement the algorithms: Fuzzy Logic toolbox, Optimization toolbox, and Statistics and Machine Learning toolbox. The MRM and TIC chromatographic signals were first manually adjusted for appropriate and consistent baseline using Agilent's software MassHunter (Version B.06.00, ©Agilent Technologies, Inc).

III. FITTING SIGNALS PROCEDURE

The total ion chromatogram (TIC) signal obtained for each sample represents the linear superposition of the B combined multiple reaction monitoring mode (MRM) signals. This is expressed as:

$$T[x] = \sum_{i=1}^{B} H_i[x] \tag{1}$$

where $T[x]$ and $H[x]$ are the TIC at the x^{th} time instance and the i^{th} MRM signal, respectively. The MRM raw signal specifically isolates individual groups, and therefore has higher-precision compound-specific information, though it does not convey the total contribution of all the chemicals captured in the TIC signal. Therefore, to autonomously capture the individual contributions of different contaminants within a Hi-Vol sample, it is imperative to analyze the available MRM raw signals for different compound groups. This enables derivation of the individual peak heights, corresponding to individual PCB congeners (target peaks), as well as significant non-target peaks that contribute towards the aggregated TIC signal. This section highlights, describes, and presents the procedure in the order used to analyze the raw MRM signals.

A. SHIFTING MRM SAMPLE SIGNALS

Changes in the GC/MS/MS column temperature, dimension, or carrier gas linear velocity through the GC/MS/MS columns cause retention time shifts which can impact the analysis of identifying PCB congeners in the sample signals [41]. To alleviate this issue and aid in identifying PCB congeners, the signals are shifted to align the peaks in the MRM sample signals. The peaks are aligned by identifying the standards, the largest peaks in the signal, within the samples and adjusting the retention times to the standards in the calibration samples. The adjusted TIC signal, $T[x]$, is expressed as:

$$T[x] = \sum_{i=1}^{B} H_i[x - f] \tag{2}$$

where f is the difference in retention time from each MRM's respective standard to the calibration's standard.

B. DETERMINING PEAK MAXIMA AND LOCAL MINIMA

To remove the noise from each MRM sample signal, the maximum signal value, s_{max}, is determined in each of the twelve MRM chromatograms. The background noise floor is selected as τs_{max} where any part of the raw signal with amplitude $s \geq \tau s_{max}$ is used for peak detection (detailed in Fig. 1). The background noise threshold is selected empirically as $\tau = 5 \times 10^{-5}$. The threshold is chosen by examining the previously analyzed MRM data and determining an average baseline value; this corresponds to a maximum signal-to-noise ratio (SNR) of $\sim$ 43dB for the maximum value of

Pseudo-code showing steps for peak detection

Step 1: Determine if there is a local peak maximum at retention time X_1. X_{other} corresponds to other retention times around X_1 that we are using to check if X_1 is a local maximum.

$$If\ X_1 > \forall X_{other}\ where\ X_{other} = (X_{Max} - 20 \ldots X_{Max} + 20) \rightarrow X_1\ is\ maximum\ or\ X_{max},$$

Step 2: Determine if there is a local peak minimum at retention time X_1. X_{other} corresponds to other retention times around X_1 that we are using to check if X_1 is a local minimum.

$$If\ X_1 < \forall X_{other}\ where\ X_{other} = (X_{Min} - 70 \ldots X_{Min} + 70) \rightarrow X_1\ is\ minimum\ or\ X_{min},$$

FIGURE 1. Pseudo-code for peak detection with the raw signal.

the raw signal, i.e., treating s_{max} as the signal amplitude. The choice of maximum SNR allowed robust detection of any peaks, target or non-target, that fall within 43 dB of the highest peak within the raw signal. After the noise removal step, the pseudo-code given in Fig. 1 Step 1 is used to determine PCB peaks, where x is the index corresponding to the retention time on the MRM. A peak spread threshold of 20 indices is used to ensure determination of a distinct peak.

C. APPLYING A CURVE FIT FOR PEAKS

To obtain the minima, local minimum heights were identified by using the original MRM sample signals for twelve different compound groups [29] and finding the smallest relative intensity in the signal relative to nearby points seen in Fig. 1 Step 2. A threshold of 70 indices is implemented to minimize the capture of local minimums within larger peaks.

To eliminate further noise in the peaks and evaluate the contribution of the peak to the signal, H[x], a best fit cosine curve is applied to the signal peaks (example seen in Fig. 2).

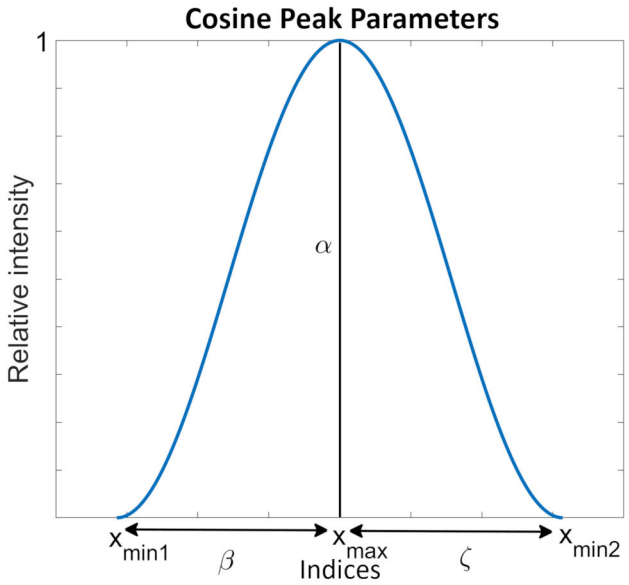

FIGURE 2. Example of cosine peak used for modeling peaks. α adjusts the height of the cosine peak, β varies the width of the left side of the peak and ζ varies the width of the right side of the peak.

A cosine model is used since it provided a superior goodness of fit to other popular peak shapes such as Gaussian models. To computationally calculate the portion of each peak the maxima and minima are first determined using the pseudo-code and specific steps detailed in Fig. 1. There are two cases when modeling peaks within the raw signal, described below in detail.

1) CASE 1: SINGLE PEAK, NON-COELUTED PEAKS

In this scenario, a peak is identified at x_{max} between two minima at x_{min1} and x_{min2}, where $x_{min1} < x_{max} < x_{min2}$, using the pseudo-code depicted in Fig. 1. Because peaks within the signals can be asymmetrical, each half of the peak is fitted separately. The signal between x_{min1} and x_{min2} is defined as $s[x]$. Half of the peak, $r_\beta[x]$, is fitted as:

$$r_\beta[x] = \alpha\left(\frac{\cos(\beta\theta) + 1}{2}\right) \tag{3}$$

where $\theta = -\pi, \ldots, 0$, α varies the height of the peak and is initialized to $s[x_{max}]$, and β varies the width of the half peak and is initialized to the distance between $s[x_{min\ 1}]$ and $s[x_{max}]$. The other half of the peak, $r_\zeta[x]$, is fitted using:

$$r_\zeta[x] = \alpha\left(\frac{\cos(\zeta\theta) + 1}{2}\right) \tag{4}$$

where $\theta = 0, \ldots, \pi$, α is the same as (3), and ζ determines the width of the half peak and is initialized to the distance between $s[x_{max}]$ and $s[x_{min2}]$. The modeled peak, $R[x]$, is written as:

$$R[x] = r_\beta[x] + r_\zeta[x] \tag{5}$$

Equation (5) is optimized by minimizing the objective function for the best fit curve, $P[x]$, as:

$$P[x] = \sum_{i=1}^{n} ||s[x] - R_i[x]||_2^2 \tag{6}$$

2) CASE 2: MULTIPLE PEAKS: COELUTED PEAKS

When there are multiple peaks around the same retention time identified using the pseudo-code depicted in Fig. 1, it is possible to separate them using a similar technique as before. Identifying ϱ peaks between two minima indicates a coelution of peaks within the signal. To determine the contributions of each peak to $s[x]$, each peak is fitted using (3)-(5). There are three cases in which the peak fitting is initialized:

(i) Peak lies between a local minimum and another peak, i.e. $x_{min1} < x_{max1} < x_{max2}$, the β in (3) is initialized as described in case 1 and ζ in (4) as the distance between x_{max1} and x_{max2}.

(ii) Peak lies between a peak and local minima, i.e. $x_{max2} < x_{max3} < x_{min2}$, the ζ in (4) is initialized as described in case 1 and β in (3) as the distance between x_{max1} and x_{max2}.

(iii) Peak between two peaks, i.e. $x_{max1} < x_{max2} < x_{max3}$, the β is initialized as the distance between x_{max1} and x_{max2} and ζ as the distance between x_{max2} and x_{max3} in (3) and (4) respectively.

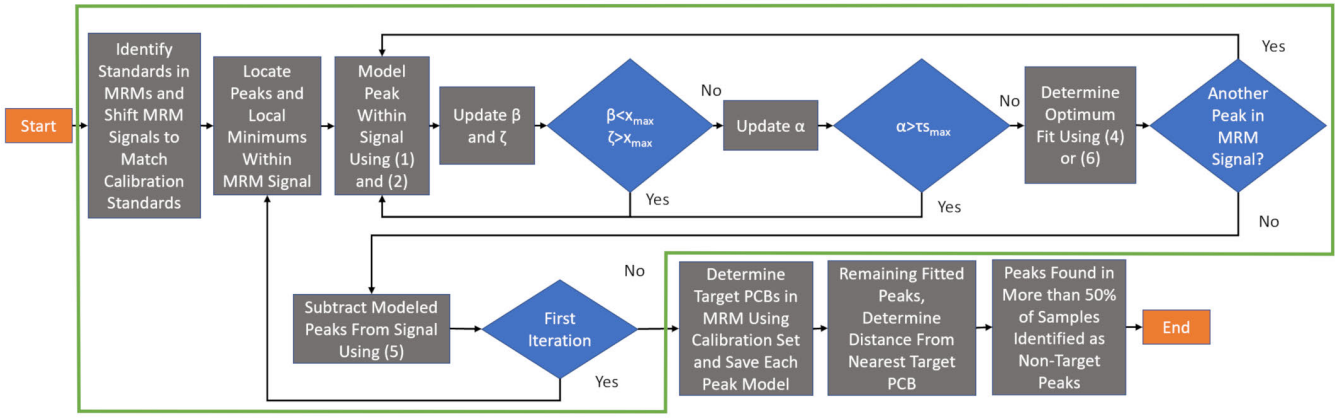

FIGURE 3. Procedure of fitting peaks within the signal. Once all the peaks have been fitted, the fitted curves are subtracted from the raw signal and fit the resulting peaks in the new signal (denoted in the green box). Once all the peaks and coeluted peaks have been determined, target peaks are determined from the calibration signals. Once the target peaks have been identified, non-target peaks that occur greater than 50% of the signals are found over the sample signals.

To optimize the fitted curve, $P[x]$, and contribution of each peak to $s[x]$, use the following:

$$P[x] = \sum_{i=1}^{n} ||s_i[x] - \sum_{j=1}^{\varrho} R_i^j[x]||_2^2 \tag{7}$$

where j denotes the j^{th} fitted peak in the raw signal. Equation (7) thus optimizes the overall fit of each of the peaks to the coeluted signal, $s[x]$, and determines each peak's contribution.

D. DISCOVERING COELUTED PEAKS

Because of the properties of the chromatographic column used in the GC/MS/MS, some PCBs that go through the GC/MS/MS will reach the detector at the same retention time and create further coelution of peaks. To better identify these coeluted peaks, the N fitted curves detailed in Section III-C., $P[x]$, are subtracted from the original raw signal, $H[x]$, to get a new signal, $A[x]$, containing hidden peaks. This is done by using:

$$A[x] = H[x] - \sum_{j=1}^{N} P_j[x] \tag{8}$$

where j denotes the j^{th} fitted curve in the raw signal. If there are hidden non-target peaks, they can be identified in the new signal $A[x]$. Section III A-D is repeated using $A[x]$ as the new raw signal to determine if any peaks were unseparated (coeluted) within the signal. The procedure of this technique can be seen in Fig. 3. To demonstrate how well fitted this model is, one sample's twelve modeled MRM signals were added to produce a TIC sample signal seen in Fig. 4.

IV. AUTOMATIC DETECTION AND INTERPRETATION OF RAW INSTRUMENT SIGNAL

The peak model presented in Section III is used to autonomously detect hundreds of target and non-target peaks

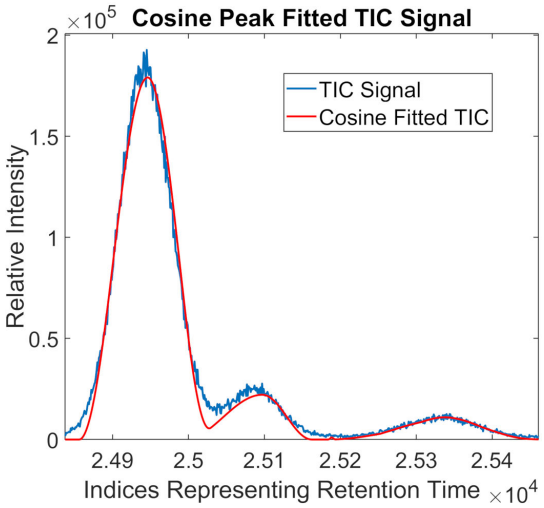

FIGURE 4. Fit of the TIC data set. The blue line is the TIC signal $T[x]$ and the red line is the linear summation of fitted cosine peaks in each MRM samples.

from the raw instrument signal and interpreted their associations using a combination of statistical and geometric clustering techniques.

A. DETERMINING PCB CONGENER

A PCB was identified in the MRM sample if the peak's retention time was closest to a PCB's peak retention time in the same MRM calibration solution's chromatogram (within a 0.07-minute range). After a target PCB was determined, the PCB's modeled peak was stored in the constructed TIC signal. For each target PCB found in each of the MRM sample signals, the fitted peak was shifted back to its original position to match the original MRM sample signal.

B. DETERMINING NON-TARGET ANALYTES

Implementing the peak model from Section III A-D, non-target analytes in the MRM and TIC data sets are then

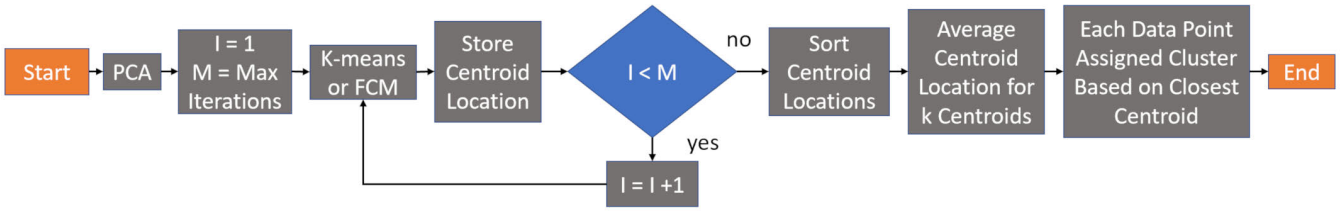

FIGURE 5. Procedure of PCA, k-means and c-means clustering.

analyzed. After the target PCBs are determined, a new signal is reconstructed for the MRM data sets using only these target PCBs. The signal is subtracted from the original data sets, theoretically uncovering peaks that are not identified as PCBs. After applying this approach to all the data sets, the newly found peaks are compared to each other to determine non-target peaks. The non-target peaks are determined by first finding the difference in retention times of the uncovered peaks to the peak of the closest target PCB. Then, the difference is used to search all 150 samples to find similarities. If a peak is found present more than 50% of the time, it is labeled as a potential non-target analyte within the data. This method is applied to both MRM and TIC data sets.

C. PRINCIPAL COMPONENT ANALYSIS (PCA)

After determining the PCB peak height values found in each sample and normalizing using the internal standard heights of both the calibration solutions and samples, principal component analysis (PCA) is implemented to find relationships within the peak heights of the data. In this case, PCA is particularly useful to determine correlations from similar instances of peaks within the samples. The data matrix V, comprising of all the identified target peak heights for each sample, can be described by the product between the scores matrix G and the transpose, T, of the loading matrix L with an added residual matrix E.

$$V = GL^T + E \tag{9}$$

Equation (9) produces the principal components which are a linear combination of the original variables. The principal components are ordered according to the amount of variance explained in V, i.e. principal component 1 represents the dominant variation while principal component 2 represents the second most. For this work, the scores matrix G is used for clustering and is plotted in the first two or three principal components to visualize the associations within the data.

D. K-MEANS CLUSTERING

The k-means algorithm implements an unsupervised machine learning algorithm to associate a data set within an N dimensional space into k clusters. These clusters are created by using k centroids to assign the data to each cluster. The choice of k was determined based on empirical observations and the elbow method (see Appendix B). The k-means objective function minimized is shown in (10).

$$J_M = \sum_{j=1}^{k} \sum_{i=1}^{n} ||g_i - c_j||^2 \tag{10}$$

where k is the number of clusters, n is the number of data-points, g is the i^{th} data-point in the scores matrix G, and c is the centroid for cluster j. Each value is assigned a cluster using:

$$c^{(i)} = \arg\min_j ||g_i - c_j||^2 \tag{11}$$

The j^{th} centroid location, c_j, is updated using:

$$c_j := \frac{\sum_{i=1}^{m} \delta(c_j = i) g_i}{\sum_{i=1}^{m} \delta(c_j = i)} \tag{12}$$

where $\delta = 1$ if g_i belongs to the j^{th} cluster. Equation (12) is iterated until Equation (10) converges on a local or global minima. Because of the high variability between runs of the k-means clustering, this technique is implemented multiple times to find an average centroid location to consistently cluster the data and determine associations (Fig. 5). From the resulting clusters, PCB congeners or sample locations are associated with each other in multiple principal components to find connections between the clustering of PCBs or locations to potential sources.

E. FUZZY C-MEANS CLUSTERING (FCM)

The fuzzy c-means (FCM) algorithm is used in conjunction with the k-means algorithm to provide validation to the clustering parameters chosen. The FCM objective function minimized is shown in (13).

$$J_M = \sum_{i=1}^{n} \sum_{j=1}^{k} \mu_{i,j}^{m} \|g_i - c_j\|^2 \tag{13}$$

where n is the total number of data-points, k is the total number of clusters, $\mu_{i,j}$ is the calculated degree of membership of the i^{th} data-point to the j^{th} cluster, m is the fuzzy partition matrix exponent, g_i is the i^{th} data-point, and c_j is the j^{th} centroid location. The fuzzy partition matrix exponent is chosen and explained further in Appendix A. Similar to k-means, this is an iterative algorithm, and starts by assigning random degrees of memberships to the data-points. The j^{th}

centroid, c_j is calculated and updated with Equation (14).

$$c_j := \frac{\sum_{i=1}^{n} \mu_{i,j}^m g_i}{\sum_{i=1}^{n} \mu_{i,j}^m} \tag{14}$$

where the same variable definitions hold. After centroid location calculation the degree of membership matrix, $\mu_{i,j}$, is calculated and updated using Equation (15).

$$\mu_{i,j} = \frac{1}{\sum_{l=1}^{k} \left(\frac{\|g_i - c_j\|}{\|g_i - c_l\|} \right)^{\frac{2}{m-1}}} \tag{15}$$

The FCM is iteratively run until a maximum number of iterations has been updated or the objective function improves less than a specified threshold, whichever occurs first. Due to the variability in resultant centroid locations, the FCM algorithm is performed multiple times and an overall average of the centroid locations are used for association determination (Fig. 5). After calculating the average centroid locations, the final degree of membership is calculated from Equation (15), where the maximum degree of membership is used to assign a cluster number to the data-point.

F. ACCURATELY DETERMINING SOURCES

Geographical locations can have different sources of PCBs thus it is imperative to identify and localize these sources using different techniques. Previously, sources have been identified by either using linear regression models [9]–[13] or even PMF [10]–[13], [51]–[54] on other data sets. In this work, sources are identified by correcting the mass of each PCB congener using the calibration solutions, RRF, and surrogate standard recoveries for each sample, then finding the mass fraction of the PCB congeners in each sample. The percentages of Aroclor sources, γ, of PCBs are identified using an L_2 minimization optimization technique that considers PCB mass fractions and minimizes the error, ϵ. The percentages of Aroclor sources is calculated using:

$$\epsilon = \min_{\gamma}((D\gamma - b)^2) \tag{16}$$

where D is the mass fraction contribution of PCBs in each Aroclor mixture (found in [21]), γ are the factors (sources) represented as columns of Aroclor data that are being calculated (restricted to $\gamma \geq 0$), and b is the mass fraction of the PCBs found in a sample. Equation (16) can also be written as:

$$\gamma = (D^T D)^{-1} D^T b \tag{17}$$

This technique is implemented with MATLAB's `fmincon` function to parameterize the results given. Although reference [21] provides feasible sources, non-Aroclor mixtures present and other potential mixtures not considered that are not present are of interest to this study. Referring to references [17]–[20], new factors are added to D to calculate the percentages of each mixture. To ensure random mixtures are also being considered, a Monte Carlo approach is employed to calculate an average percentage of each mixture used.

G. RELATIVE RETENTION TIME PROXIMITY

Relative retention time proximity is a useful criterion to associate peaks that closely elute and therefore, may be useful to associate within and beyond PCA clusters. A node edge graph visualizes the relationship between PCBs with respect to their retention time that are harder to determine by just examining the TIC or MRM signal. This relation can be useful in identifying closely eluted PCBs and identifying peaks as PCB targets. Fig. 10 is created using the average retention time location of each PCB congener across all samples. The relative proximity of each target PCB within the TIC signal is found by implementing a nearest neighbor technique. This step allows for a better understanding of the location of PCB targets within the TIC signal. These graphs are shown to demonstrate the co-occurrence associations between PCBs and not as knowledge graph informatics as [55].

V. RESULTS AND DISCUSSION

The resulting fitted peaks described in detail throughout Section III was used for the clustering analysis performed throughout the results and discussion, Section V. The peak fitting procedure had an average normalized mean square error of -18.4851 dB to the signal.

A. CROSS-COMPARISON BETWEEN AUTONMOUSLY DETECTED TARGET PEAKS AND EXPERT-ANNOTATED TARGET PEAKS

This work aims to automate the time-consuming process of identifying target PCB peaks and mapping them to known target compounds. Therefore, it is imperative to provide quantitative comparisons between what our algorithm finds against expert-annotated ground truths validated over the same data. The ground truth for retention times for each of the 209 target PCBs considered in this work are based on existing calibration standards documented in [29]. Therefore, based on the documented retention times for each target peak in the standards, any identified peaks can be mapped to specific target PCB. Additionally, for each sample, ground truths are further established based on manually executed expert validation of whether or not a peak was visually identified at the stipulated retention time. Depending on the PCB composition of the sample, a particular peak may (or may not) be present in the raw signal, as all PCBs may not be present in detectable concentrations in every sample.

Table 1 compares the target PCBs identified with this algorithmic technique to what was manually identified in [29]. A raw signal peak identified manually or using our autonomous technique based on Equations (3)-(7) is labeled "positive" if a PCB is also identified within ($\pm$)0.07 minutes of the listed retention time in the calibration standard, where 0.07 minutes is the sampling time period of the raw signal. Otherwise, we assign the label "negative" to a detected peak, which cannot be mapped to a target PCB in the standard; either due to the peak falling outside the 0.07 minutes range or for being an undocumented non-target peak. Similarly,

the labels "manual" and "autonomous" are respectively assigned for peaks detected by an expert, as detailed in [29], or autonomously identified using the signal processing techniques detailed in this work. Specifically, the four possible labels are described as below:

(i) Manual negative: A PCB peak is identified in the standard but the manual inspection cannot find the peak in the sample;
(ii) Manual positive: A PCB peak is identified in the standard and the manual inspection can find the peak in the sample;
(iii) Autonomous negative: A target PCB peak is identified in the standard but the algorithm does not detect the peak in the raw signal for the sample;
(iv) Autonomous positive: A target PCB peak is identified in the standard and the algorithm also detects the peak in the raw signal for the sample.

Each entry of Table 1 shows the number of raw signal peaks that meet the criteria for the corresponding row and column labels. For example, the entry "Manual positive/Autonomous positive" indicates that 23411 raw signal peaks were autonomously identified across the full dataset which also matched with the ground truth of manually validated target PCB peaks.

TABLE 1. Cross-Comparison between target PCBs retention times found autonomously through the described technique and expert-validated target PCBs retention times provided from previous work [29] as ground truths. The full definition of terms are provided in Section V-A.

results	Manual positive	Manual negative
Autonomous positive	23411	2519
Autonomous negative	454	16

The results of Table 1 can be summarized as: 98.1% of manually determined target peaks could also be detected as target PCB peaks by our autonomous method. On the other hand, 90.3% of autonomously detected target peaks, i.e., peaks that occurred within ± 0.07 minutes of a listed target PCB in the calibration standard, could be matched with target peaks that were manually determined over the raw signal sample. Therefore, the proposed algorithm discovered 2519 peaks over the whole data that corresponded to a target PCB based on the standard but were missed in manual inspection. We also observe that 16 raw signal peaks, listed as target PCBs in the standard sample, were not found both by manual inspection or the proposed autonomous method. Therefore, Table 1 provides validation that our method autonomously identifies target peaks from the raw signal with 98.1% accuracy, while identifying extra target peaks missed by manual detection. Any peak-allocation error in the technique is attributed to any residual baseline noise in the MRM data sets or larger retention time shifts that can be accounted for by the raw signal sampling interval. These autonomously identified peaks, which are mapped into specific target PCBs, provide the basis for automated peak-cognizant interpretation that most autonomous chemometric studies cannot offer.

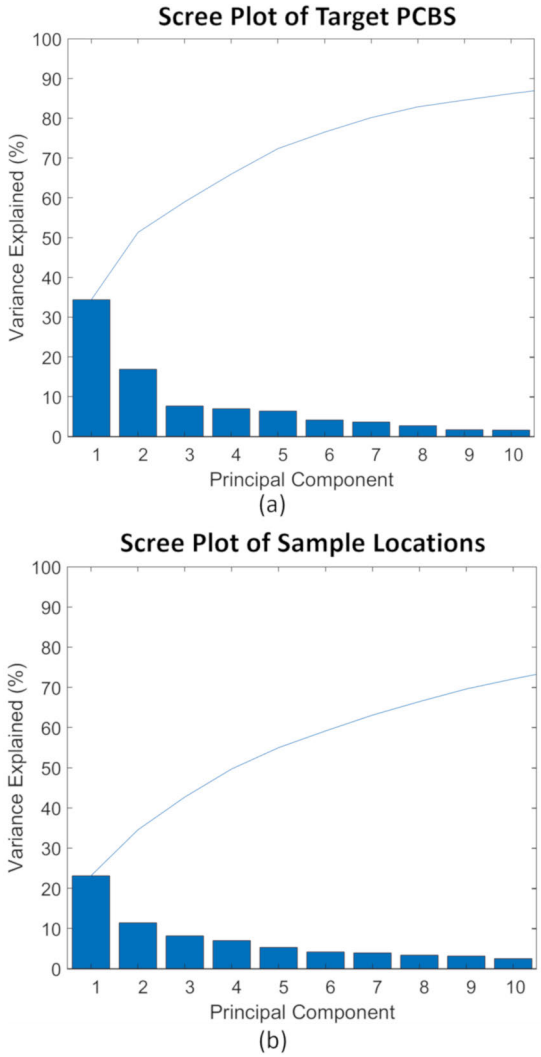

FIGURE 6. Scree plots of target PCB (a) and sample location (b) data. (a) The majority of the variance explained is found within the first 3 principal components. (b) The majority of the variance explained if found within the first 5 principal components. For this work, only the principal components that contribute to ∼ 95% of the variance explained are considered in clustering.

Reproducing this rich peak-cognizant information manually, annotated as specific target PCBs, and visualized based on their relative proximity as in Fig. 10, will be overwhelmingly expensive in expert personnel time and subject to human bias.

For the PCA visualization, the peak heights were determined for each PCB and these values were normalized based on their surrogate standards and were standardized sample to sample. The scree plots are shown in Fig. 6 and plots of the data in the principal components are seen in Fig. 7 and Fig. 8. To better visualize the clusters and plotted data in Fig. 7 and Fig. 8, the first two principal components were plotted in Fig. 7 and first three principal components were plotted in Fig. 8.

B. COMPARISON OF K-MEANS AND FCM CLUSTERING

To associate the data presented in this work and cluster either PCBs or sampling locations, both k-means and FCM

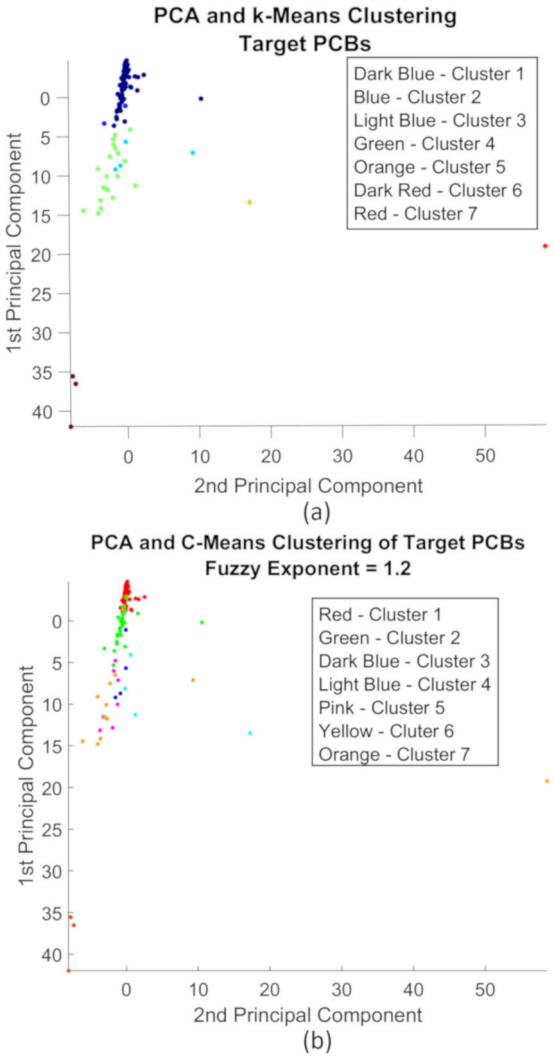

FIGURE 7. First 2 principal components of the target PCB data are plotted. (a) k-means clustering of target PCB data with 7 clusters. (b) FCM clustering of target PCB data with 7 clusters and fuzzy exponent, *m*, as 1.2. Both clustering techniques offer unique solutions to the data, however, FCM clusters the larger group of data into separate clusters, i.e. cluster 1 and cluster 2, and clusters two outlier points that are far apart together, i.e. cluster 7.

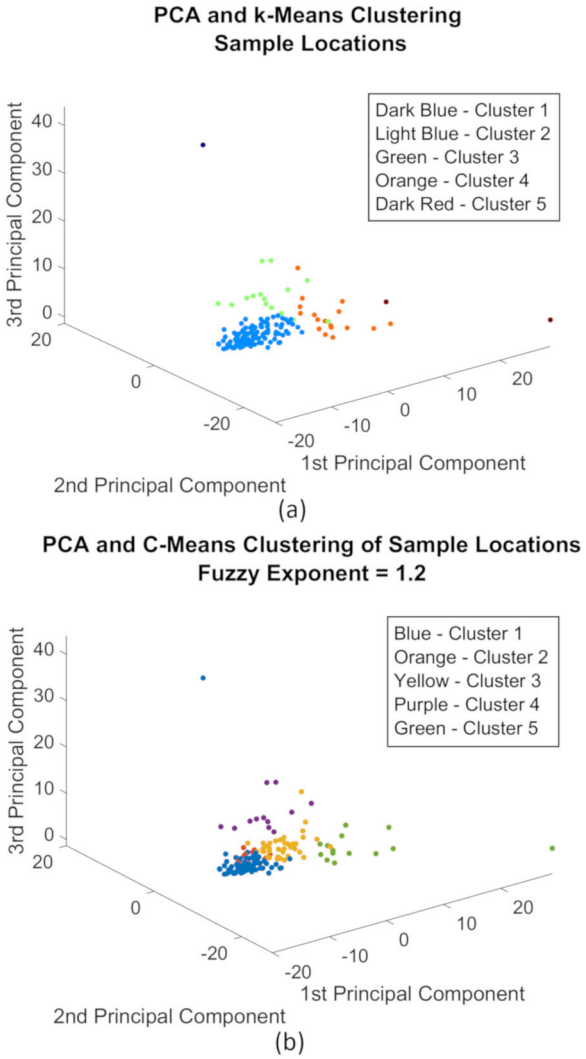

FIGURE 8. First 3 principal components of the sample location data are plotted. (a) k-means clustering of sample location data with 5 clusters. (b) FCM clustering of sample location data with 5 clusters and fuzzy exponent, *m*, as 1.2. Both clustering techniques offer unique solutions to the data, however, FCM clusters two outlier points into the larger groups, i.e. cluster 5, and cluster 2 separates closely related data.

clustering are examined. FCM is implemented with a choice of the fuzzy partition exponent, m, in Equation (13) as 1.2 (discussed in Appendix A). The optimal number of centroids for PCB data and sampling locations is 7 and 5 respectively (discussed in Appendix B). Although FCM considers all points in the data for each centroid to make a soft decision, the computational time of the technique is slower. Further, the clusters created using FCM does not partition the points into distinct clusters and can have difficulty finding optimal associations. Seen in Fig. 7 and Fig. 8, the larger group of points is clustered differently and includes outlier points within the relatively closer groups of points as compared to k-means clustering. In Fig. 7 (a), the k-means clustering technique clusters the larger group of data together while in Fig. 7 (b), FCM splits the larger group into separate clusters. Although this provides more separation of data, the c-means technique clusters outliers together with the larger group and can provide misleading associations. In Fig. 8 (a), the k-means technique clusters outlier data as their own clusters and clusters the larger group of data into 3 separate clusters. In Fig. 8 (b), the FCM technique clusters the larger group of data into separate clusters and associates outlier data within the larger groups providing incorrect associations of the data. From the above analysis, k-means clustering is chosen and plotted in the results for the rest of this work.

C. PCA OF TARGET PCBs

This section is focused on the target PCBs and their contribution throughout each sample clusters. PCA and clustering

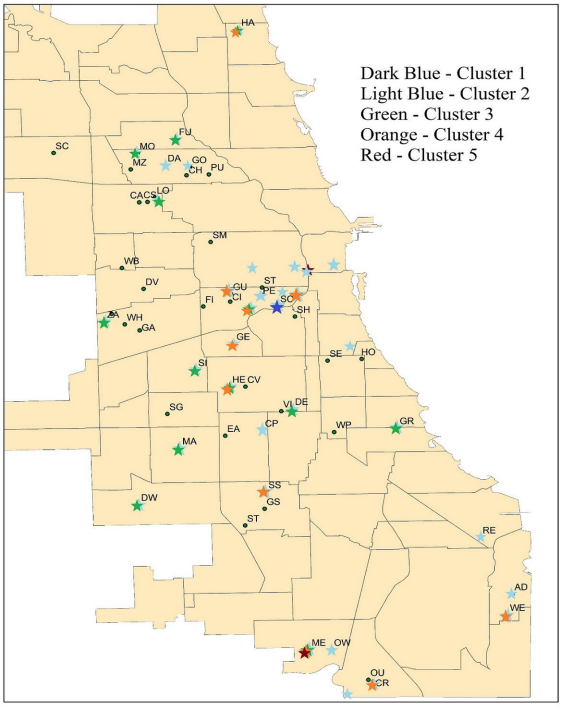

FIGURE 9. **The Chicago land area map labeled with sample clusters. The cluster coloring comes from the PCA plot in Fig. 8 (a).**

of the PCBs across all the samples are plotted in Fig. 7 (a) where the first two principal components make up roughly 50% of the variance explained. This plot shows clusters of potential significance across the samples that could be related to sources. Applying this concept, major PCBs are identified that are contributing as outliers to overall PCB profiles such as PCB 3 and PCB 5 which form their own separate respective clusters. Further, when non-target analytes are identified this method is used to make associations with Aroclors aiding in fingerprinting the pollutants.

D. PCA OF GEOGRAPHICAL SAMPLING LOCATIONS

Clustering geographical sampling locations is based on target PCBs and plotted in Fig. 8 (a) where the first three principal components consist of 45% of the variance explained. Further, the color of the cluster is plotted for each of the samples on a map to get a better topological view seen in Fig. 9. Although seasonal clustering was not a particular focus of this work, it was noticed that all the winter samples clustered together independent of the location (Cluster 2) while all other seasons were scattered throughout the various clusters. With samples of the same season that were collected around the same time, this method can produce associations between various locations to identify sources based on proximity of their geographical sampling location.

E. RETENTION TIME PROXIMITY

Cluster analysis identifies two distinct groups of target PCB retention time proximity shown in Fig. 10. Fig. 10 (b) shows a subgraph that only contains PCBs that cluster in the same group and Fig. 10 (c) shows PCBs that clustered in completely different groups. This provides two different ways of interpreting the data. In the former, it means that the general proximity of the PCBs in retention time may not be coeluting as much or that these PCBs together show up with the same relative intensity throughout all the samples. The latter case demonstrates that either the PCBs are coeluting and that some of the actual concentration may be in another PCB or that these PCBs appear close to one another and further analysis of peaks that show up around the larger PCB can be based on the other PCB. Although these are two different ways of analyzing this, it provides a better understanding of the locations of the PCBs in the chromatogram and identification of PCBs in the TIC data more topologically.

F. DETERMINING SOURCES

The percentage of Aroclor only contribution to each location was determined and displayed in Fig. 11 using Equation (16). Focusing on the two largest percentages of Aroclors in each location, Aroclors 1232, 1254, 1016, and 1221 were highly present (Fig. 11 (a)). Further, with added uncertainty to Equation (16), there is a large contribution of 3, 3', dichlorobiphenyl (PCB 11) and Aroclors 1232, 1254, 1016, and 1221 as shown in Fig. 11 (b). Aroclors 1254, 1016, and 1221 were produced and sold by Monsanto for use in products such as capacitors, adhesives, and rubbers [15]. These Aroclors can be primarily seen as products used in construction throughout Chicago, however, Aroclor 1232 was not expected to be an important source. Aroclor 1232 was sold in small quantities compared to other Aroclors and is only found in a few products such as hydraulic fluids or adhesives that could have similar contributions as Aroclors 1221 and 1254 [15], [16]. Environmental weathering by microbial dechlorination, environmental distillation, and atmospheric reaction may have changed the PCB mixtures present in Chicago air to resemble Aroclor 1232. In addition to detection of Aroclors, PCB 11 is present in each sample. PCB 11 is a current byproduct of pigment manufacturing and was not present in the Aroclor mixtures sold, and now banned, more than 40 years ago [17], [30]. It has a lower molecular weight and is more volatile than most of the PCBs present in Aroclors which could explain its large relative abundance. Because paint is applied to surfaces in thin coats, PCB 11 likely volatilizes into the air more efficiently than heavier PCB congeners. This is also seen in [19].

G. NON-TARGET PCB PEAKS

Although no non-target peaks of significant size were identified within the MRM data, there were hidden non-target peaks that coeluted with target PCBs in the TIC data set. Fig. 12 shows representative non-target peaks that appeared in the TIC signal after deleting the modeled TIC signal with color coding corresponding to cluster colors in Fig. 13. The non-target peaks thus isolated are significant from a data analysis perspective. This is because the TIC signal is generated

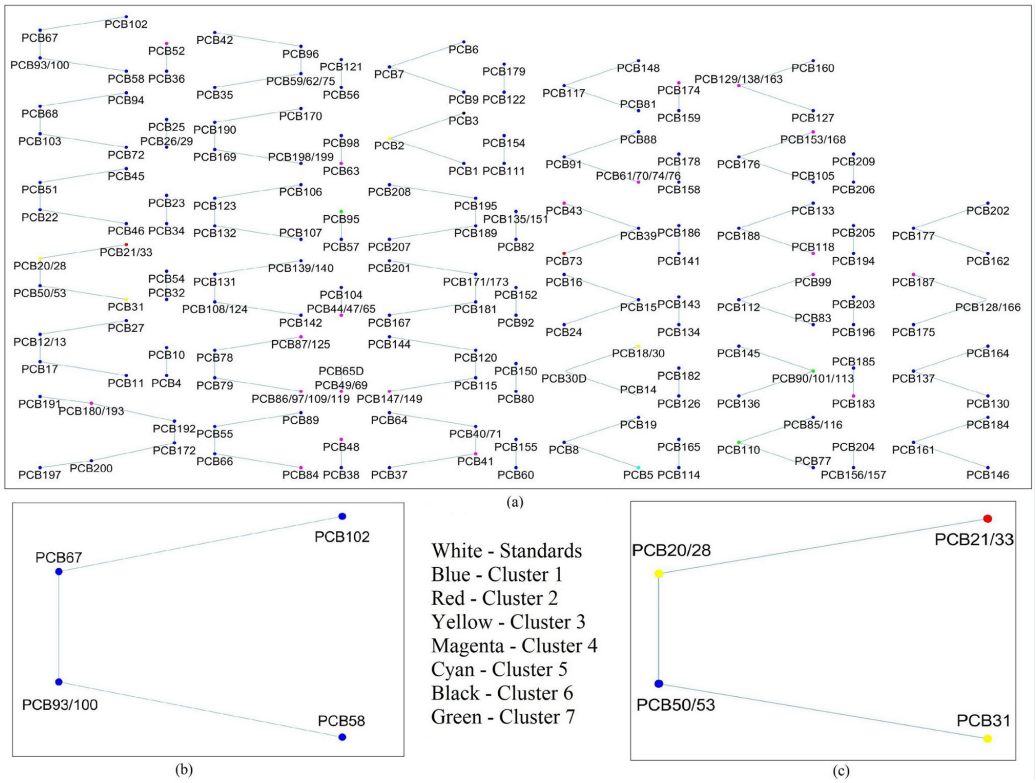

FIGURE 10. Graph of detected target PCBs within the GC/MS/MS data sets. The nodes represent PCBs with colors corresponding to their cluster color, and the edges represent a connection to another PCB that has retention time closest to it (a). The retention times used are the average retention time found throughout each of the data sets. Found within (a) are two cases: 1) where the PCBs all have the same cluster color (b) and 2) where the PCBs are all closest to each other but are within different clusters (c).

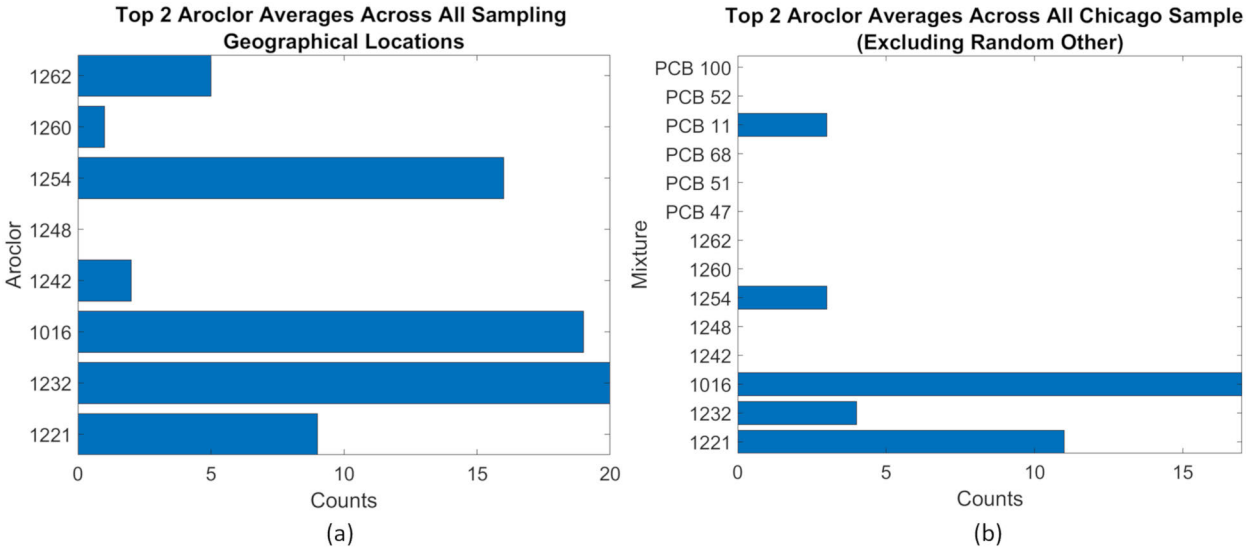

FIGURE 11. Two highest percentages of Aroclors in the geographical locations are counted excluding random other mixture. (a) Top 2 Aroclor only percentages are considered from each sample using target PCBs peak heights and Equation (9) to determine percentages (Appendix C). (b) Top 2 Aroclor only percentages are considered from each sample using target PCBs peak heights and Equation (9) with contributions of each PCB to each Aroclor considered from [8] to determine percentages (Appendix D). Further, different mixtures of PCB only and random mixtures of PCBs are considered. In both (a) and (b), Aroclor 1016 is one of the most present Aroclors across all sampling locations.

from archived samples that are traditionally filtered to screen out chemicals that are not PCBs. Such target-selective chemical filtering is a standard laboratory protocol in a majority of environmental studies and most public-domain data archives assume such target-selective filtering has been successfully performed. However, based on the raw TIC signal analysis

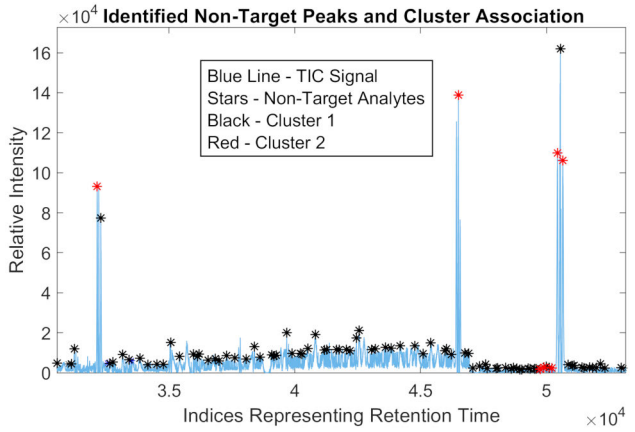

FIGURE 12. Plot of one of the sample TIC data sets after the subtraction of target PCB peaks. A few of the potentially found non-target peaks within the TIC data set are pictured and are indicated by the stars at the location they were found. The coloring is based on PCA and k-means clustering.

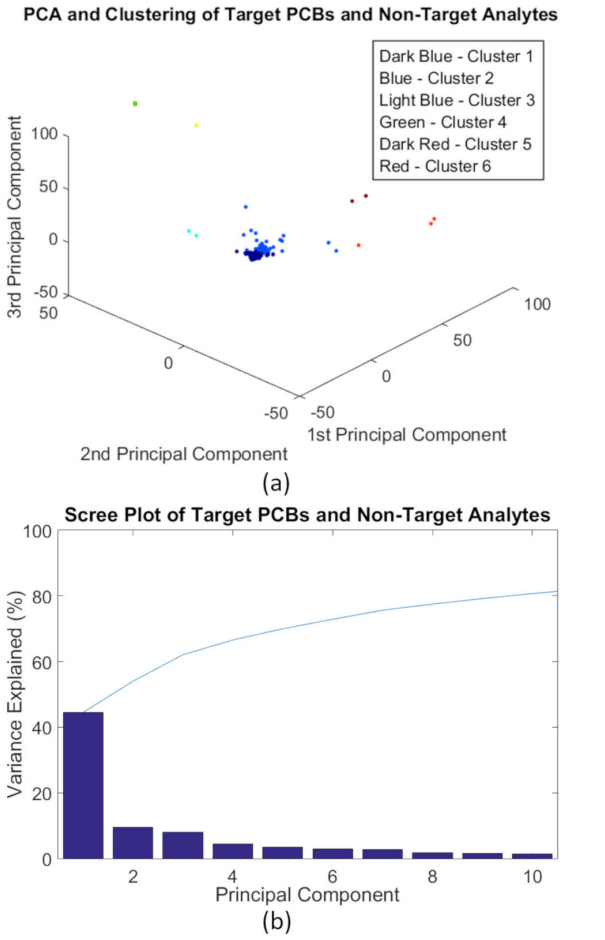

FIGURE 13. (a) First 3 principal components plotted for target PCBs and non-target analytes found within the TIC. Clustering analysis performed with k-means clustering with 6 clusters. (b) Scree plot of target PCBs and non-target analytes found in TIC signal data. Over 50% of the variance explained is explained within the first 3 principal components.

from such a representative data archive it appears that some non-target chemicals still pass through as undetected contaminants. These contaminants, while potentially closely associated with target analytes, coelute with the target compounds as hidden TIC peaks, and normally would not be detected or accounted for in any target-based or statistical analysis. Therefore, while the motivation for raw signal analysis is to find target and non-target compounds, from a purely traditional target-oriented perspective, such hidden non-target contaminants discovered in the TIC signal are significant for two reasons: (i) to test using GC/MS/MS raw signal analysis whether the laboratory protocols can indeed screen out most non-target analytes in the samples as typically desired in target-oriented studies, and (ii) to validate whether computational techniques, such as those proposed here, can indeed discover hidden non-target analytes that coelute with target PCBs.

H. PCA OF TARGET PCBs AND NON-TARGET ANALYTES

For this section, both the target PCBs identified before and the non-target analytes identified in the TIC data set were used. Examples of a few non-target analytes identified can be seen in Fig. 12. PCA and clustering of PCBs and non-target analytes are plotted in Fig. 13. The first three principal components make up roughly 60% of the variance explained. This plot shows clustering of target PCBs and non-target analytes across the samples that could impact source discoveries. Although a large amount of the non-target analytes clustered as outliers, some were clustered with target PCBs like PCB 5 and PCB 1. This is a significant finding as these non-targets, which closely associate with target PCBs, would otherwise never be included in traditional contaminant studies. Although this technique was implemented for PCBs and non-target analytes, this same approach can be used for specific sample locations. The idea behind this technique is to find non-target analytes and target PCBs that associate together to better identify sources of PCBs in the sample locations. This approach was only implemented for non-target analytes in the TIC data set because there were more significant peaks in the TIC than the MRM data sets.

VI. CONCLUSION

This work proposes a novel combination of various computational techniques to automate peak-cognizant detection and interpretation of PCBs found within GC/MS/MS data sets. Specifically, peak modeling and L2 error minimization techniques are employed to autonomously detect target and previously undetected non-target peaks from the raw instrument signal. Then, a combination of PCA and k-means clustering techniques are employed to isolate groups of PCB congeners that are potentially associated with each other to demonstrate how they manifest in the environment. Individual contributions of Aroclors across a diverse portfolio of Chicago air samples are isolated. Utilizing these techniques and concepts can aid in discovering and interpreting all the information inherent within the GC/MS/MS signal. This type

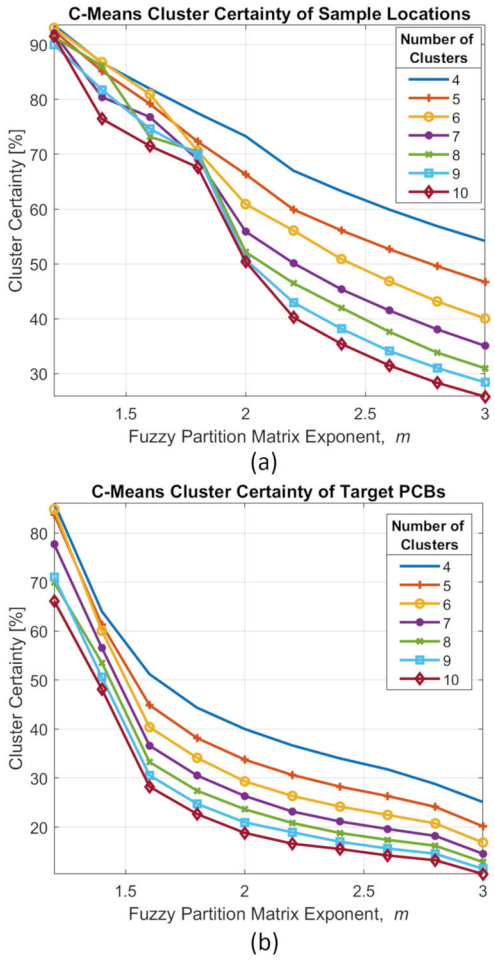

FIGURE 14. Average degree of membership of points to a cluster as a function of cluster number and fuzzy partition matrix exponent, *m*. (a) shows c-means clustering for the sample locations while (b) shows the clustering for target PCBs. As the fuzzy partition matrix exponent, *m*, increases the average cluster certainty decreases (average cluster uncertainty increases) since there are more options for a data-point to belong too.

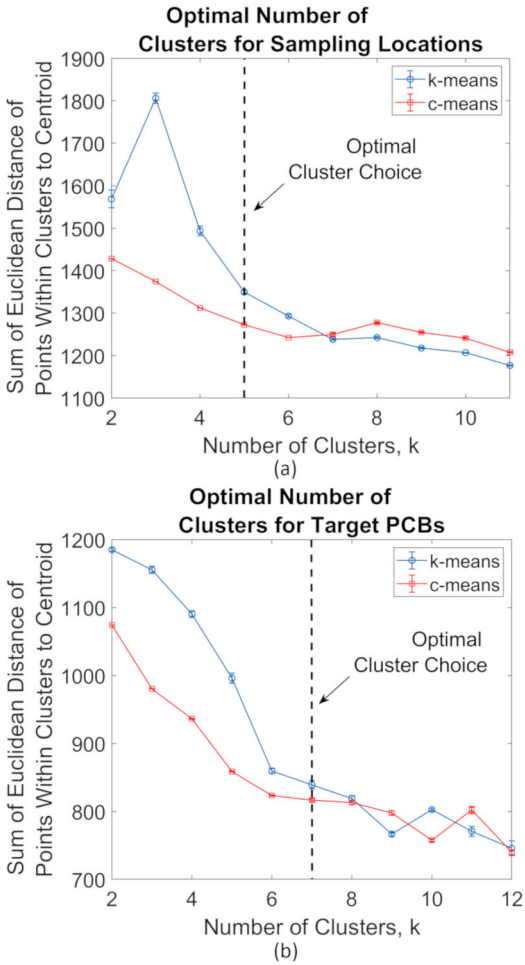

FIGURE 15. Plot of sum of euclidean distance of points within a cluster to their respective centroid vs number of clusters, k, using the iterative process depicted in Fig. 5. The sum of euclidean distances were averaged by implementing k-means and fuzzy c-means clustering 20 unique times for each k.

of comprehensive and quantitative analysis is valuable to environmental science in two significant ways:

(i) By design, these developed techniques are not biased towards target contaminants, which are typically employed in traditional GC/MS/MS interpretation, and therefore, can be used to discover unknown contaminants that might prove critical to air pollution studies.

(ii) These novel techniques are peak-based, and therefore compound-cognizant, unlike purely statistical chemometric methods [7], [8]. This approach allows interpretation of large-scale statistical results based on PCA and k-means clustering at the level of individual compounds. Therefore, the methods can bridge the gap between compound-agnostic statistical interpretation and compound-specific (target-based) studies [7]–[13], [17]–[21], [29], [51]–[54].

In summary, the major science return of these techniques is connecting target compounds (known PCB congeners) with potentially significant but previously unknown non-target compounds and allow comprehensive automated (peak-cognizant) analysis of raw GC-MS (and combinations thereof) signals across large data repositories. While this work reports the findings across 150 active air samples, this technique has the potential to be applied across much larger scales of data and repositories.

APPENDIX A
FCM CLUSTERING CERTAINTY AND FUZZY PARTITION EXPONENT

Equations (13)-(15) depend primarily on the fuzzy partition exponent, m, to update the cost function and centroid location within each cluster. The fuzzy partition exponent dictates how fuzzy the results will be and often can skew the results determined by the relative inter-distance of the data-points. While choosing a random m may yield results, further observations into the exponent m is plotted in Fig. 14. An important observation to note is the increasing m value causes the uncertainty of points within a cluster to increase. Further, the number of clusters impacts the cluster certainty because the clusters will be close together making it difficult to distinguish the optimal cluster for the point to belong to.

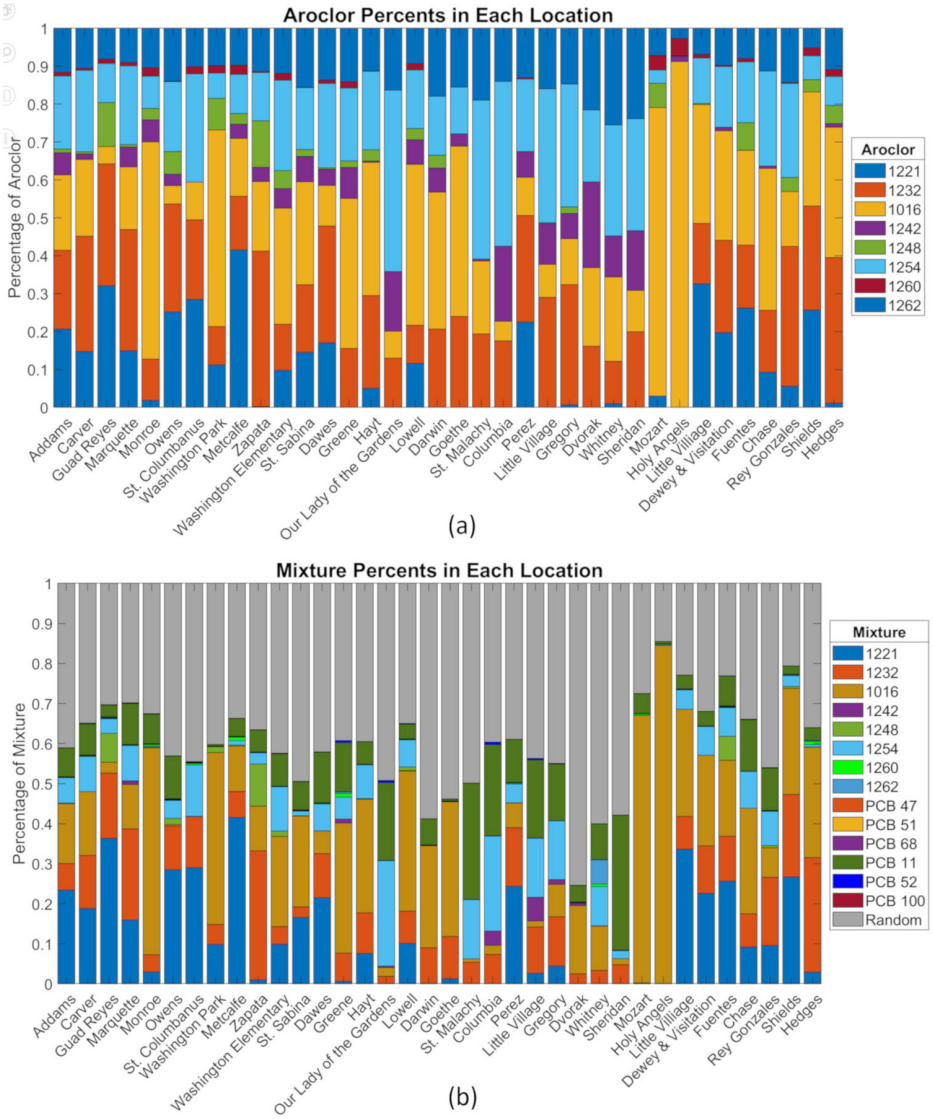

FIGURE 16. Percentage of each Aroclor contributing to the PCB values found for each location. Only Aroclors are considered in [8].

For this work, a smaller m value is implemented to ensure clustering of neighboring data.

APPENDIX B
DETERMINING K CLUSTERS

For practical purposes, the sum of euclidean distances from each point within each cluster is considered and plotted to determine the optimal choice of k clusters. Mathematically, the sum of euclidean distances from each point within each cluster is expressed as:

$$d = \sum_{i=1}^{k} \sum_{g_j \epsilon K_i} ||c_i - g_j||_2^2 \tag{18}$$

where g is the set of points in the scores matrix G, K is the set of k clusters, and c_i is the center of cluster i. Summing across multiple replications of similar number of clusters measures the variability of the points within cluster and describes how compact the clusters are within the data. To confirm the correct number of clusters used, d is plotted for different number of clusters implemented. Seen in Fig. 15 is the optimal choice of clusters for the scores matrix, G, of both PCB and and sample location data. The choice of clusters is based on empirical observations and the slope of the plot in Fig. 15.

APPENDIX C
AROCLOR PERCENTAGE PRESENCE

Contributions of each Aroclor are determined using Equation (17) for each sampling location. The mass fraction contribution matrix, D, is determined by normalizing the PCBs found in the samples using internal standards to the calibration data. The relative response factor (RRF) of each PCB is determined using the normalized peak heights to the calibration data. The percents of only Aroclor contributions for each sampling location are determined using [5]. The two

highly present Aroclors found across all sampling locations are plotted in Fig. 11 and discussed further in Section V-E.

APPENDIX D
MIXTURE PERCENTAGE PRESENCE

Contributions of each Aroclor are determined using Equation (17) for each sampling location. The mass fraction contribution matrix, D, is determined by normalizing the PCBs found in the samples using internal standards to the calibration data. The relative response factor (RRF) of each PCB is determined using the normalized peak heights to the calibration data. The percents of Aroclor and other mixture contributions for each sampling location are determined. While Aroclor only contributions values can be found in [5], other PCB only contributions and random mixtures of varying PCBs are considered to determine mixture percentages. The two highly present mixutres found across all sampling locations are plotted in Fig. 11 and discussed further in Section V-E.

APPENDIX E
NOMENCLATURES

Acronym	Definition
PCB	Polychlorinated Biphenyl
Congener	One of the 209 well defined chemical PCB compounds
Aroclor	Mixture of well known PCB congeners with unique signals
RRF	Relative Response Factor
GC/MS/MS	Gas Chromatography-Mass Spectrometry
PCA	Principal Component Analysis
TIC	Total Ion Chromatogram
MRM	Multiple Reaction Monitoring
Hi-Vols	High-Volume Air Samplers
SNR	Signal-to-Noise Ratio
FCM	Fuzzy c-means Clustering
PMF	Positive Matrix Factorization

APPENDIX F
MRM MASS TRANSITIONS (M/Z)

Cl Homolog	Precursor Ion	Product Ion
1	188.0	152.0
2	222.0	152.0
3	258.0	186.0
4	291.9	222.0
5	325.0	255.9
6	359.8	289.9
7	393.8	323.9
8	429.7	259.8
9	463.7	393.8
10	497.7	427.9

ACKNOWLEDGMENT

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

[1] J. Lee, T. H. Park, H. S. Kang, and S.-H. Lim, "Miniaturized gas chromatography module with micro posts embedded MEMS column for the separation of exhaled breath gas mixtures," in *Proc. IEEE SENSORS*, Oct. 2016, vol. 37, no. 2, pp. 1–3.

[2] A. S. Gupta, B. Meyer, and E. Overton, "Quantifying weathering profiles of environmental contaminants from marine and coastal oil spills using signal processing techniques," in *Proc. OCEANS MTS/IEEE Charleston*, Oct. 2018, pp. 1–4.

[3] A. Skarysz, Y. Alkhalifah, K. Darnley, M. Eddleston, Y. Hu, D. B. McLaren, W. H. Nailon, D. Salman, M. Sykora, C. L. P. Thomas, and A. Soltoggio, "Using capillary gas chromatography to determine polychlorinated biphenyls (PCBs) in electrical insulating liquids," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[4] N. S. A. Zubir, M. A. Abas, N. Ismail, N. A. M. Ali, M. H. F. Rahiman, N. K. Mun, M. N. Taib, and N. T. Saiful, "Pattern classifier of chemical compounds in different qualities of agarwood oil parameter using scale conjugate gradient algorithm in MLP," in *Proc. IEEE 13th Int. Colloq. Signal Process. Appl. (CSPA)*, Mar. 2017, pp. 18–22.

[5] Y. Fu, X. Wan, X. Zhang, G. Fang, and J. Yi, "Side peak interference mitigation in FM-based passive radar via detection identification," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 53, no. 2, pp. 778–788, Apr. 2017.

[6] G. W. Johnson, R. Ehrlich, W. Full, and S. Ramos, "Principal components analysis and receptor models in environmental forensics," in *Introduction to Environmental Forensics*, 3rd ed. New York, NY, USA: Academic, 2015, ch. 18, pp. 609–653.

[7] J. M. Amigo, T. Skov, R. Bro, J. Coello, and S. Maspoch, "Solving GC-MS problems with PARAFAC2," *TrAC Trends Anal. Chem.*, vol. 27, no. 8, pp. 714–725, Sep. 2008.

[8] J. M. Amigo, M. J. Popielarz, R. M. Callejón, M. L. Morales, A. M. Troncoso, M. A. Petersen, and T. B. Toldam-Andersen, "Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis," *J. Chromatography A*, vol. 1217, no. 26, pp. 4422–4429, Jun. 2010.

[9] R. Corbella, M. A. Rodriguez-Delgado, and F. J. Garcia Montelongo, "Contribution to the identification and quantitation of aroclor mixtures by least-squares analysis of gas chromatographic data," *J. Chromatograph. Sci.*, vol. 36, no. 7, pp. 372–378, Jul. 1998.

[10] F. Risso, A. Magherini, M. Ottonelli, E. Magi, S. Lottici, S. Maggiolo, M. Garbarino, and R. Narizzano, "A comprehensive approach to actual polychlorinated biphenyls environmental contamination," *Environ. Sci. Pollut. Res.*, vol. 23, no. 9, pp. 8770–8780, May 2016.

[11] M. Zhang and P. B. Harrington, "Automated pipeline for classifying aroclors in soil by gas chromatography/mass spectrometry using modulo compressed two-way data objects," *Talanta*, vol. 117, pp. 438–491, Dec. 2013.

[12] C. Y. Ma and C. K. Bayne, "Differentiation of aroclors using linear discrimination for environmental samples analyzed by electron capture negative ion chemical ionization mass spectrometry," *Anal. Chem.*, vol. 65, no. 6, pp. 772–777, Mar. 1993.

[13] S. C. Karcher, M. J. Small, and J. M. Vanbriesen, "Statistical method to evaluate the occurence of PCB transformations in river sediments with application to hudson river data," *Environ. Sci. Technol.*, vol. 38, no. 24, pp. 6760–6766, Dec. 2004.

[14] R. Bro, "PARAFAC. Tutorial and applications," *Chemometric Intell. Lab. Syst.*, vol. 38, no. 2, pp. 149–171, Oct. 1997.

[15] O. Faroon, S. R. Corporation, and J. Olson, "Toxicology profile for polychlorinated biphenyls (PCBs)," U.S. Dept. Health Human Services, Washington, DC, USA, Tech. Rep. CDC 6480, Nov. 2000. [Online]. Available: https://stacks.cdc.gov/view/cdc/6480

[16] T. Kopp, *PCBs in the United States Industrial Use and Environmental Distribution*. Washington, DC, USA, Feb. 1976. [Online]. Available: https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=20001275.TXT

[17] D. Hu and K. C. Hornbuckle, "Inadvertent polychlorinated biphenyls in commercial paint Pigments," *Environ. Sci. Technol.*, vol. 44, no. 8, pp. 2822–2827, Apr. 2010.

[18] N. J. Herkert, J. C. Jahnke, and K. C. Hornbuckle, "Emissions of tetrachlorbiphenyls (PCBs 47,51, and 68) form polymer resin on kitchen cabinets as a non-aroclor source to residential air," *Environ. Sci. Technol.*, vol. 52, no. 9, pp. 5154–5160, Apr. 2018.

[19] R. A. Hites, "Atmospheric concentrations of PCB-11 near the great lakes have not decreased since 2004," *Environ. Sci. Technol. Lett.*, vol. 5, no. 3, pp. 131–135, Feb. 2018.

[20] R. F. Marek, P. S. Thorne, N. J. Herkert, A. M. Awad, and K. C. Hornbuckle, "Airborne PCBs and OH-PCBs inside and outside urban and rural U.S. Schools," *Environ. Sci. Technol.*, vol. 51, no. 14, pp. 7853–7860, Jun. 2017.

[21] G. M. Frame, "A collaborative study of 209 PCB congeners and 6 aroclors on 20 different HRGC columns 2. Semi-quantitative aroclor congener distributions," *Fresenius J. Anal. Chem.*, vol. 357, no. 6, pp. 714–722, Mar. 1997.

[22] M. A. Morgan and B. W. McDaniel, "Transient electromagnetic scattering: Data acquisition and signal processing," *IEEE Trans. Instrum. Meas.*, vol. 37, no. 2, pp. 263–267, Jun. 1988.

[23] K. Peabody, A. Husain, M. H. Tang, and Z. Macek, "Digital signal processing aids cholesterol plaque detection," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, May 1995, pp. 1173–1176. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=480446

[24] U. Madhow, "MMSE interference suppression for timing acquisition and demodulation in direct-sequence CDMA systems," *IEEE Trans. Commun.*, vol. 46, no. 8, pp. 1065–1075, Aug. 1998.

[25] A. Mirbagheri, "Linear MMSE receivers for interference suppression & multipath diversity combining in long-code DS-CDMA systems," UWSpace, 2003. [Online]. Available: http://hdl.handle.net/10012/780

[26] S. Uhlich and B. Yang, "MMSE estimation in a linear signal model with ellipsoidal constraints," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2009, pp. 3249–3252.

[27] J. Simsa, "Linear adaptive blind MMSE detectors for DS-CDMA signals," in *Proc. 17th Int. Conf. Radioelektronika*, Jun. 2007, pp. 1–6.

[28] H.-T. Li and P. M. Djurie, "MMSE estimation of nonlinear parameters of multiple linear/quadratic chirps," *IEEE Trans. Signal Process.*, vol. 46, no. 3, pp. 796–800, Mar. 1998.

[29] D. Hu, H.-J. Lehmler, A. Martinez, K. Wang, and K. C. Hornbuckle, "Atmospheric PCB congeners across chicago," *Atmos. Environ.*, vol. 44, no. 12, pp. 1550–1557, Apr. 2010.

[30] A. M. Awad, A. Martinez, R. F. Marek, and K. C. Hornbuckle, "Occurence and distribution of two hydroxylated polychlorinated biphenyl congeners in chicago air," *Environ. Sci. Technol. Lett.*, vol. 3, no. 2, pp. 47–51, Jan. 2016.

[31] *Polychlorincated Biphenyls (PCBs) by Gas Chormatography, Revision 1*, document Method 8082A (SW-846), U.S. EPA, 2007. [Online]. Available: https://www.epa.gov/sites/production/files/2015-12/documents/8082a.pdf

[32] *Chlorincated Congeners in Water, Soil, Sediment, Biosolids, and Tissue by HRGC/HRMS*, document Method 1668b, U.S. EPA, 2008. [Online]. Available: https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P1005EUE.TXT

[33] A. Baldacci and G. Haralabus, "Signal processing for an active sonar system suitable for advanced sensor technology applications and environmental adaptation schemes," in *Proc. 14th Eur. Signal Process. Conf.*, Sep. 2006, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7071237

[34] W. C. Knight, R. G. Pridham, and S. M. Kay, "Digital signal processing for sonar," *Proc. IEEE*, vol. 69, no. 11, pp. 1451–1506, Nov. 1981.

[35] S. Haykin and B. Kosko, *Intelligent Signal Processing*. Piscataway, NJ, USA: IEEE Press, 2001.

[36] P. R. Atkins, T. Collins, and K. G. Foote, "Transmit-signal design and processing strategies for sonar target phase measurement," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 1, pp. 91–104, Jun. 2007.

[37] Z. Wei, J. Huang, and Y. Hui, "Adaptive-beamforming-based multiple targets signal separation," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Sep. 2011, pp. 1–4.

[38] Z. Ni, L. Wang, J. Meng, F. Qiu, and J. Huang, "EEG signal processing in anesthesia feature extraction of time and frequency parameters," *Procedia Environ. Sci.*, vol. 8, pp. 215–220, Jan. 2011.

[39] A. K. Ahmadi, P. Moradi, M. Malihi, S. Karimi, and M. B. Shamsollahi, "Heart rate monitoring during physical exercise using wrist-type photoplethysmographic (PPG) signals," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7319800

[40] R. K. Jain, J. M. F. Moura, and C. E. Kontokosta, "Big data + big cities: Graph signals of urban air pollution [exploratory SP]," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 130–136, Sep. 2014.

[41] D. Rood, "Gas chromatography problem solving and troubleshooting," *J. Chromatograph. Sci.*, vol. 35, no. 136, pp. 239–240, May 1997.

[42] W. V. Ligon and R. J. May, "Target compound analysis by two-dimensional gas chromatography—Mass spectrometry," *J. Chromatography A*, vol. 294, pp. 77–86, Jan. 1984.

[43] A. Sen Gupta, C. M. Reddy, and R. Nelson, "Systems and methods for topographic analysis," U.S. Patent 8 838 393, Sep. 16, 2014.

[44] H. Ghasemi Damavandi, A. Sen Gupta, R. K. Nelson, and C. M. Reddy, "Interpreting comprehensive two-dimensional gas chromatography using peak topography maps with application to petroleum forensics," *Chem. Central J.*, vol. 10, no. 1, pp. 1–14, Nov. 2016.

[45] R. Bruflodt, R. K. Nelson, E. C. Arrington, D. Valentine, A. S. Gupta, K. Lemkau, V. Kivenson, and C. M. Reddy, "Fingerprinting the refugio oil spill using topographic signal processing of two-dimensional gas chromatographic images," in *Proc. OCEANS Anchorage*, Sep. 2017, pp. 1–4.

[46] H. Ghasemi Damavandi, A. Sen Gupta, G. Canahuate, C. M. Reddy, and R. Nelson, "Robust oil-spill forensics and petroleum source differentiation using quantized peak topography maps," 2018, *arXiv:1807.07484*. [Online]. Available: http://arxiv.org/abs/1807.07484

[47] I. T. Jolliffe, *Principal Component Analysis*. Hoboken, NJ, USA: Wiley, 2002.

[48] A. Demiriz, K. P. Bennett, C. M. Breneman, and M. J. Embrechts, "Support vector machine regression in chemometrics," in *Proc. 33rd Symp. Interface Comput. Sci. Statist.*, 2001, pp. 1–9.

[49] T. Howley, M. G. Madden, M.-L. O'Connell, and A. G. Ryder, "The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data," *Knowl.-Based Syst.*, vol. 19, no. 5, pp. 363–370, Sep. 2006.

[50] M. Kamstrup-Nielsen, L. Johnsen, and R. Bro, "Core consistency diagnostic in PARAFAC2," *J. Chemometrics*, vol. 27, no. 5, pp. 149–171, May 2013.

[51] L. A. Rodenburg and D. K. Ralston, "Historical sources of polychlorinated biphenyls to the sediment of the new York/New Jersey harbor," *Chemosphere*, vol. 169, pp. 450–459, Feb. 2017.

[52] S. Du, T. J. Belton, and L. A. Rodenburg, "Source apportionment of polychlorinated biphenyls in the tidal delaware river," *Environ. Sci. Technol.*, vol. 42, no. 11, pp. 4044–4051, Jun. 2008.

[53] L. A. Rodenburg, S. Du, D. E. Fennell, and G. J. Cavallo, "Evidence for widespread dechlorination of polychlorinated biphenyls in groundwater, landfills, and wastewater collection systems," *Environ. Sci. Technol.*, vol. 44, no. 19, pp. 7534–7540, Oct. 2010.

[54] L. A. Rodenburg, S. Du, B. Xiao, and D. E. Fennell, "Source apportionment of polychlorinated biphenyls in the New York/New Jersey harbor," *Chemosphere*, vol. 83, no. 6, pp. 792–798, Apr. 2011.

[55] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proc. IEEE*, vol. 104, no. 1, pp. 11–33, Jan. 2016.

RYAN A. MCCARTHY received the dual B.A. degree in engineering physics and scandinavian Studies from the Augustana College, Rock Island, IL, USA, in 2017. He is currently pursuing the Ph.D. degree in electrical and computer science engineering with The University of Iowa, Iowa City, IA, USA. From 2014 to 2015, he worked at the Research and Development Department, Miner Enterprises Inc. In 2016, he was an Engineer at Crawford Company. Since 2017, he has been a Research Assistant with the Electrical and Computer Science Engineering Department, The University of Iowa. His research interests include applications of signal processing, underwater acoustics, and data analysis. He became a member of Sigma Pi sigma, in 2016, and a member of Phi Beta Kappa, in 2017. He was named to the Hampshire Society of the National Football Foundation, in 2017.

ANANYA SEN GUPTA (Member, IEEE) received the M.S. and Ph.D. from the University of Illinois at Urbana-Champaign, in 2001 and 2006, respectively.

From 2008 to 2012, she was a Postdoctoral Scholar and Researcher at Woods Hole Oceanographic Institution, working in undersea signal processing and petroleum forensics. Since 2013, she has been an Assistant Professor with the Electrical and Computer Engineering Department, The University of Iowa, Iowa City. Her research interests include the nexus of signal processing, pattern recognition, and knowledge discovery, with emphasis on applications to environmental chemistry, underwater acoustics, and space plasma physics. She seeks to develop geometric computational techniques that enable sophisticated representation, localization, tracking, and classification of raw instrument signals generated by diverse environmental contaminants, laboratory conditions, and natural phenomena. Her algorithms have been applied to shallow water acoustic communications, fingerprinting oil spills, sonar target recognition in high-clutter coastal environments, and tracking high-energy plasmospheric events on Earth and Mars. She currently leads an interdisciplinary research team of graduate and undergraduate students, several of whom have received multiple student research awards under her mentorship. Her research of Iowa EPSCoR project was recently featured in the ISGC 2015-2016 STIMULI EPsCOR report distributed to Congress.

Dr. Sen Gupta has received the Teaching Award, in 2015, and three mentor awards from the Iowa Space Grant Consortium, in 2016 and 2017. She is currently an Associate Editor of IEEE Access, a Guest Editor of the IEEE Journal of Oceanic Engineering Special Issue in "Underwater Acoustic Propagation Physics and Signal Processing Techniques for Shallow Water Acoustic Communications" and a Technical Committee Member of the IEEE Oceanic Engineering Society.

BERNICE KUBICEK received the dual B.S. degree in mechanical engineering and electrical engineering from the Milwaukee School of Engineering, Milwaukee, WI, USA, in May 2019. She is currently pursuing the Ph.D. degree in electrical and computer engineering with The University of Iowa, Iowa City, IA, USA. In 2017, she worked as a Process Engineer Intern at Hentzen Coatings Inc., Milwaukee. From 2018 to 2019, she worked as an Electrical Design Engineering Intern at Astronautics Corporation of America, Milwaukee. She has been a Research Assistant at the Electrical and Computer Engineering Department, The University of Iowa, since June 2019. Her research interests include active sonar, various feature extraction techniques, and applications of signal processing.

ANDREW M. AWAD received the B.S. degree in physics and astronomy, the B.A. degree in music, and the M.S. degree in environmental science from The University of Iowa. He is a Project Engineer with cGMP Consulting. He is currently working with a large bio-pharmaceutical manufacturer leading projects to track production processes and streamline data collection. He has five years of scientific research experience, with an emphasis in the analysis of PCBs, and their chemical byproducts in environmental air, soil, and sediment samples.

ANDRES MARTINEZ received the B.S. degree in biochemical engineering from the Pontificia Universidad Católica de Valparaiso, Chile, the M.S. degree in environmental technology from Imperial College London, and the Ph.D. degree in environmental engineering from The University of Iowa. He is currently an Assistant Research Engineer at IIHR-Hydroscience and Engineering, and an Adjunct Assistant Professor at the Department of Civil and Environmental Engineering, The University of Iowa. He has more than ten years of scientific research experience, during which he has developed expertise in the areas of field sampling, development of analytical method and passive sampling devices, and analysis of organic compounds, such as PCB in complex environmental matrices, environmental modeling, and data analysis. He has more than 20 peer reviewed articles in high impact scientific journals.

RACHEL F. MAREK received the B.A. degree in chemistry from the Grinnell College and the Ph.D. degree in environmental engineering from The University of Iowa. She is currently an Assistant Research Scientist at the IIHR-Hydroscience and Engineering, The University of Iowa, and a Researcher with the Iowa Superfund Research Program. She was a GAANN Fellow of the U.S. Department of Education, The University of Iowa. Her research interests include sources and fate of environmental contaminants, such as PCBs, siloxanes, and pesticides and their breakdown products in abiotic and biotic environmental matrices and whether people, especially children, are exposed to these harmful chemicals. She is a member of the American Chemical Society, the Society for Environmental Toxicology and Chemistry, and the International Association for Great Lakes Research. As a Graduate Student, she won the C. Ellen Gonter Graduate Student Paper Award from the American Chemical Society, in 2013.

KERI C. HORNBUCKLE received the B.A. degree in chemistry from the Grinnell College, and the Ph.D. degree in environmental engineering and science from the University of Minnesota. She is currently the Donald E. Bently Professor of engineering at the Department of Civil and Environmental Engineering and a Research Engineer at the IIHR-Hydroscience and Engineering, The University of Iowa. She is also the Director of the Iowa Superfund Research Program, a research center funded by the National Institute for Environmental Health Sciences. She is an expert on the sources and transport of polychlorinated biphenyls (PCBs), synthetic fragrances, perfluorinated compounds, siloxanes, current use and legacy pesticides, and other persistent organic pollutants. She is an Associate Editor of the *Journal of the American Chemical Society* and *Environmental Science & Technology*.