Hindawi
Journal of Spectroscopy
Volume 2023, Article ID 7105348, 13 pages
https://doi.org/10.1155/2023/7105348

Research Article

Multiple PAHs' Detection under CDOM Interference Based on Excitation-Emission Matrix and Interval Selection

Ruizhuo Li,[1,2] **Limin Gao**,[1,2] **Guojun Wu**,[1,3] **and Jing Dong**[1,2]

[1]*Xian Institute of Optics and Precision Mechanics, Chinese Academy of Science, Xi'an 710119, China*
[2]*College of Photoelectricity, University of Chinese Academy of Science, Beijing 100049, China*
[3]*Qingdao Marine Science and Technology Center, Qingdao 266237, Shandong, China*

Correspondence should be addressed to Guojun Wu; wuguojun@opt.ac.cn

Received 13 September 2023; Revised 31 October 2023; Accepted 15 November 2023; Published 8 December 2023

Academic Editor: Mohd Sajid Ali

Fluorescence technology is an effective tool for detecting polycyclic aromatic hydrocarbons (PAHs) in water. However, the accuracy of fluorescence detection is reduced by the spectral overlap of different PAHs and coexisting colored dissolved organic matter (CDOM). In this study, a single-excitation interval selection method based on an excitation-emission matrix is proposed to quantify four PAHs: fluorene, pyrene, phenanthrene, and benzo(a)pyrene under CDOM interference. The optimal excitation wavelength for each PAH was obtained by stability analysis, based on which the optimal emission interval was obtained by chaotic particle swarm optimization. The partial least squares (PLS) prediction models of four PAHs under interference were established. On comparing with other modeling methods, the results show that the models with interval selection have better prediction accuracy (mean relative error < 10%) under CDOM interference. The recovery rate and limit of detection of the method were also evaluated. This study provides a new and helpful strategy for fluorescence detection of interfering PAHs in water.

1. Introduction

Polycyclic aromatic hydrocarbons (PAHs) are a class of organic compounds consisting of multiple aromatic rings, which are strongly carcinogenic, teratogenic, and mutagenic [1, 2]. PAHs can enter water bodies through a variety of routes, including wet and dry depositions, road runoff, industrial wastewater, fossil fuel combustion, and atmospheric sedimentation. Because of their severe risks to the environment and human health, PAHs are considered to be a priority contaminant for monitoring by various countries and organizations [3, 4]. Therefore, the development of rapid and efficient PAH detection methods is crucial to evaluate water quality and establish effective pollution control measures [5].

High-performance liquid chromatography (HPLC) and gas chromatography combined with mass spectrometry (GC-MS) are common methods to measure PAH in water because of their high accuracy [6]. However, the measurement usually requires analyte extraction and intensive sample preparation, which is time-consuming and laborious [7]. Some other techniques, such as spectrometric analysis [8, 9], capillary electrophoresis [10], and immunological detection [11], have been proposed for the convenient detection of PAHs. Among these methods, molecular fluorescence analysis is widely used due to its inherent sensitivity, selectivity, and versatility [12]. With the development of data acquisition systems and data analysis techniques, fluorescent measurement has shown exciting prospects in trace PAH identification and quantification [13, 14].

Three-dimensional (3D) fluorescence spectroscopy, also known as excitation-emission matrix (EEM), has attracted wide attention due to its abundant spectral information and adaptability in complex situations. EEM provides a comprehensive view of the fluorescent properties, allowing for the identification and analysis of complex mixtures of fluorophores. However, the presence of substances such as metal ions, colored dissolved organic matter (CDOM), and surfactant molecules in samples can interfere with the

fluorescence spectra of target PAHs, thus affecting the detection performance [15]. CDOM, a mixture of various organic compounds, is a ubiquitous natural source of fluorescence, and humic acid (HA) plays a major role in its absorbance and fluorescence properties [16]. Chemometric methods such as PARAFAC provide the second-order advantage of spectral separation from the EEM [17, 18]. However, PARAFAC requires the test spectra to have the same size as the modeled spectra, necessitating the measurement and preprocessing of the full EEM of the sample, which complicates the measurement and analysis processes.

Another strategy to mitigate interference from other fluorophores is variable selection, where the most important variables on the spectrum are used for modeling. For two-dimensional spectra such as near-infrared and ultraviolet spectra, numerous studies have shown that wavelength selection removes uninformative or interfering variables and improves the model interpretability [19, 20]. The variable selection for EEM is more complex, as it involves both excitation and emission dimensions. Previous studies have shown the potential of EEM combined with wavelength selection for the quantitative detection of fluorescent organics [21–23]. However, most studies have been devoted to the detection of target substances, with less focus on wavelength selection and modeling under interference. Therefore, we would like to explore PAH's wavelength selection method and modeling capabilities under CDOM interference.

This study intends to (1) investigate the optimal excitation wavelengths and emission intervals in multiple PAH detection and the interference of CDOM on the interval selection, (2) develop prediction models for PAHs based on the selected variables, and (3) evaluate the validity of the selected wavelengths by comparing them with other methods. Twenty-four samples were configured for calibration, each sample consisting of four PAHs (fluorene, pyrene, phenanthrene, and benzo(a) pyrene) and HA. A wavelength selection algorithm based on the reliability analysis and chaotic particle swarm algorithm was developed to optimize the spectral variables in EEM. The optimal excitation wavelengths and emission wavelength intervals in the presence/absence of CDOM interference were compared by preprocessing the spectra and further analysis, and the optimal models were established. This study aims to provide theoretical and experimental guidance for the detection of PAHs under CDOM interference.

2. Materials and Methods

2.1. Samples and Measurements. Four different PAHs used in this study were fluorene (FLU), pyrene (PYR), phenanthrene (PHE), and benzo(a)pyrene (BaP). These PAHs are commonly found in different water sources and signify the impact of human activities. BaP, in particular, is recognized as one of the most hazardous and cancer-causing PAHs. Most PAHs are insoluble in water, and the difference in fluorescence intensity makes their detection more complex [24]. PAH reagents were obtained from Aladdin Reagent Co., Ltd. and were of analytical grade, requiring no further purification. PAHs were dissolved in Milli-Q water and configured as a 100 μg/L primary stock solution. In particular, a small amount of ethanol is added during BaP dissolution to ensure that the concentration of BaP is consistent with that of other PAHs. HA was also obtained from Aladdin Reagent Co., and was dissolved in NaOH solution to prevent any impact on the pH. The resulting HA solution was then diluted to 100 mg/L.

Two sample sets were prepared using stock solutions: Mix-PAH and HA-PAH. The samples in the Mix-PAH set comprised four distinct PAHs, with concentrations ranging from 1 to 20 μg/L. In the HA-PAH set, the concentration of PAH mirrored that of the Mix-PAH set, and the differentiating factor lies in the introduction of three distinct concentrations of HA (2 mg/L, 5 mg/L, and 10 mg/L). The detailed concentration settings are shown in Table 1.

All the fluorescence measurements were implemented on the F-7000 spectrometer (Hitachi, Japan) at a constant temperature of 25°C. The excitation wavelength range of the instrument was set to 200 nm–400 nm with a step size of 10 nm, and the emission wavelength range was set to 250 nm–450 nm with a step size of 1 nm. The resulting Mix-PAH and HA-PAH datasets comprised 24 sample measurements, with 21 excitation and 201 emission wavelength points. Three blank samples were also measured, and blank subtraction was made to mitigate spectral background.

2.2. PAH Concentration Prediction Model. The prediction model of PAH concentration was established based on partial least squares (PLS) regression. PLS decomposes the independent and dependent variables into latent variables to determine the most important components and establishes the least squares equation that maximizes the correlation between the latent variables [25]. The number of latent variables (nLVs) in PLS is usually determined by k-fold cross-validation. In the presence of redundancy in spectra data, variable selection has been proven to be an effective way to improve model performance [26]. This study uses selected spectral data as independent variables and PAH concentrations as dependent variables.

The PLS prediction model is evaluated using root mean square error (RMSE), coefficient of determination (R^2), and mean relative error (MRE).

$$\mathrm{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n}\left(y_o - y_p\right)^2\right]^{1/2},$$
$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_o - y_p\right)^2}{\sum_{i=1}^{n}\left(y_o - \overline{y}\right)^2}, \quad (1)$$
$$\mathrm{MRE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_o - y_p\right|}{y_o},$$

where y_o and y_p are observed and predicted values, $\overline{y}$ is the average of y_i, and n is the number of samples. In general, a smaller RMSE and an R^2 closer to 1 indicate that the model has a better predictive ability. RMSECV, RMSEC, and RMSEP are the RMSE of the cross-validation, calibration set, and prediction set, respectively.

Table 1: PAH concentration values of the Mix-PAH and HA-PAH datasets (μg/L).

No.	FLU	PYR	PHE	BaP
1	20.0	20.0	20.0	20.0
2	13.3	13.3	13.3	13.3
3	5.0	5.0	5.0	5.0
4	3.3	3.3	3.3	3.3
5	1.0	1.0	1.0	1.0
6	2.0	1.0	5.0	5.0
7	2.0	10.0	10.0	10.0
8	2.0	20.0	20.0	20.0
9	5.0	2.0	5.0	10.0
10	5.0	5.0	2.0	20.0
11	5.0	10.0	20.0	2.0
12	5.0	20.0	10.0	5.0
13	10.0	2.0	10.0	20.0
14	10.0	5.0	20.0	10.0
15	10.0	10.0	2.0	5.0
16	10.0	20.0	5.0	2.0
17	20.0	2.0	20.0	5.0
18	20.0	5.0	10.0	2.0
19	20.0	10.0	5.0	20.0
20	20.0	20.0	2.0	10.0
21	15.0	10.0	7.5	12.5
22	7.5	13.0	20.0	7.0
23	5.0	8.7	13.3	4.7
24	10.0	6.7	5.0	8.3
$T1$	3.0	6.0	9.0	12.0
$T2$	7.5	17.5	16.0	9.0
$T3$	12.5	12.5	5.0	3.0
$T4$	17.5	2.5	12.5	6.0
$T5$	15.0	15.0	20.0	16.0

2.3. Particle Swarm Optimization (PSO) and Chaotic PSO (CPSO). In the study related to two-dimensional spectroscopy, the wavelength selection method is mainly divided into two categories: wavelength point selection (WPS) and wavelength interval selection (WIS) [27]. Monte Carlo uninformative variables elimination (MC-UVE) and changeable size moving window partial least squares (CSMWPLS) are two representative methods of WPS and WIS [28, 29]. Due to the apparent continuity of molecular spectra, WIS has an advantage over WPS in the interpretation of the model. However, WIS usually divides intervals according to specific rules, which are less flexible than WPS. Particle swarm optimization (PSO) is an intelligent optimization method based on the principle of bionomics [30]. Inspired by the social behavior of fish schooling or bird flocking observed in nature, PSO optimizes by simulating the information interaction between population members. PSO randomly generates I particles in the solution space and each particle has D dimensions. Each particle represents an individual solution, and these particles have position characteristics and velocity characteristics to guide the flight of particles. The objective function evaluated the fitness of the particle to determine the individual best position p_i and the global best position p_g. p_{best} is the best solution that one particle can achieve so far, and g_{best} is the global best solution of all particles within the swarm. At the t-th iteration, the position of each particle would be updated as follows:

$$v_{i,d}^{t+1} = \omega * v_{i,d}^{t} + c_1 * r_1 * \left(p_{i,d}^{t} - x_{i,d}^{t}\right) + c_1 * r_2 * \left(p_{g,d}^{t} - x_{i,d}^{t}\right), \tag{2}$$

$$x_{i,d}^{t+1} = x_{i,d}^{t} + v_{i,d}^{t+1}, \tag{3}$$

where ω is the inertia weight, $v_{i,d}$ and $x_{i,d}$ are the particle velocity and particle position in d-th dimension of i-th particle, respectively, r_1 and r_2 are random numbers between (0, 1), and c_1 and c_2 are learning factors to control step length of each iteration. The position and velocity of particles are limited to a specific range to ensure the movement of particles is reasonable.

Benefiting from the mutual cooperation between particles, standard PSO tends to converge quickly [31]. However, the performance of PSO depends mainly on the initial parameters, and it is easy to converge early and fall into local optima. Also, as the number of iterations increases, each particle is more and more similar to the optimal particle, reducing the particle population's diversity [32].

In order to enrich the search behavior and avoid falling into the local optimum, chaotic perturbation was applied to the search of PSO [33, 34]. The chaotic variable has ergodicity, pseudorandomness, and irregularity, which are determined by a deterministic equation. As a typical chaos, the logistic mapping equation is defined as follows:

$$\alpha^{n+1} = \mu\alpha^{n}\left(1 - \alpha^{n}\right), \quad n = 0, 1, 2, 3 \ldots, \tag{4}$$

where μ is the control parameter; the track of chaotic variable travels ergodically over the whole search space when $\mu = 4$ and $\alpha^0 \notin (0, 0.25, 0.5, 0.75, 1)$. The chaotic value of a logistic map for 100 iterations where $\alpha^1 = 0.3$ is shown in Figure 1.

The CPSO procedure using logistic functions is described as follows:

Step 1: Mapping the original variable $X = [x^1 \cdots x^d \cdots x^D]$ into a chaotic variable $A_0 = [\alpha_0^1 \cdots \alpha_0^d \cdots \alpha_0^D]$ according to the following rules:

$$\alpha^d = \frac{x^d - x_{\min}^d}{x_{\max}^d - x_{\min}^d}, \tag{5}$$

where term $x_{\max}^d$ and $x_{\min}^d$ represent the maximum and minimum bound in the d-th dimension of particles, respectively.

Step 2: Calculate the new chaotic variables A_n for the next iteration according to the logistic equation (4) until the maximum iteration N is reached.

Step 3: The chaotic sequence $A_1, \ldots, A_N$ is reverse transformed into decision variables $X_1, \ldots, X_N$. The variable of the d-th dimension in X_n is transformed into the following way:

$$x_n^d = x_{\min}^d + \alpha_n^d \times \left(x_{\max}^d - x_{\min}^d\right). \tag{6}$$

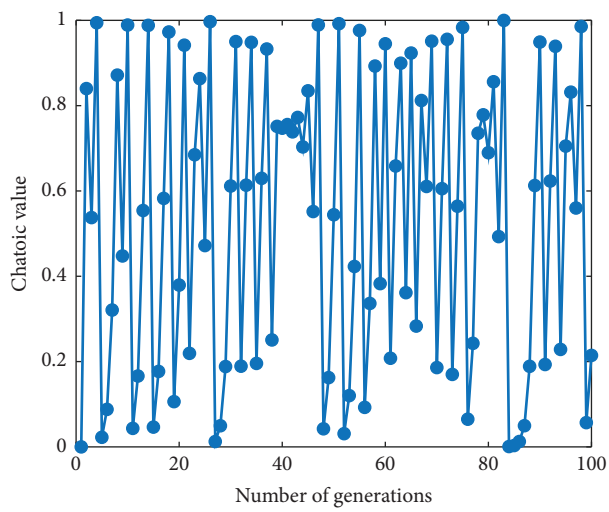

Figure 1: Variation of logistic function for 100 iterations.

Step 4: The fitness of the decision variables $X_1, \ldots, X_N$ is calculated, and if optimal fitness is better than g_{best}, retain the new solution.

Step 5: Shrink the chaotic search space according to the following equation. The subsequent chaos search searches within the new search space.

$$\begin{aligned} x_{\min}^{d,n+1} &= \max\left(x_{\min}^{d,n}, x^{d,n} - \text{rand} * \left(x_{\max}^{d,n} - x_{\min}^{d,n}\right)\right), \\ x_{\max}^{d,n+1} &= \min\left(x_{\max}^{d,n}, x^{d,n} + \text{rand} * \left(x_{\max}^{d,n} - x_{\min}^{d,n}\right)\right). \end{aligned} \tag{7}$$

2.4. Wavelength Selection Method. Unlike infrared or ultraviolet spectra, accurate measurement must optimize excitation and emission wavelengths in EEM. Suppose the EEM spectral data include M excitation wavelengths, J emission wavelengths, and K samples, the concentration matrix Y consists of K samples and the concentration of P PAHs.

The schematic procedure of the single-excitation interval selection is shown in Figure 2 briefly, and the method is described below in detail:

The optimal excitation wavelength of each PAH is selected according to Steps 1 to 3.

Step 1: The emission spectral data corresponding to each excitation wavelength are used as the independent variable, and the 1st column of Y (concentration of the 1st PAH) is the dependent variable. M prediction models are established, and RMSECV values corresponding to each excited wavelength are obtained by random k-fold cross-validation.

Step 2: Calculate RMSECV values for 2-P substances as in Step 2. The RMSECV matrix has P rows and M columns.

$$\text{RMSECV} = \begin{vmatrix} \text{RMSECV}_{1,1} & \cdots & \text{RMSECV}_{1,M} \\ \vdots & \text{RMSECV}_{p,m} & \vdots \\ \text{RMSECV}_{P,1} & \vdots & \text{RMSECV}_{P,M} \end{vmatrix}. \tag{8}$$

Step 3: Execute Step 1 to Step 2 N times and calculate the stability $S_{p,\text{m}}$ (m-th excitation wavelength for p-th PAH) according to the following equation:

$$S_{p,m} = \frac{F_{p,m}}{\sum_{i=1}^{N}\left(\text{RMSECV}_{p,m}/N\right)} = \begin{vmatrix} S_{1,1} & \cdots & S_{1,M} \\ \vdots & \ddots & \vdots \\ S_{P,1} & \cdots & S_{P,M} \end{vmatrix}, \tag{9}$$

where $F_{p,m}$ is the frequency at which $\text{RMSECV}_{p,m}$ becomes the minimum value in the p-th row of the RMSECV matrix. A higher S indicates a higher stability of the excitation wavelength. For the p-th PAH, the wavelength with the highest stability in the p-th row is chosen as the optimal excitation wavelength Ex_p. In order to reduce the limitation of excessive excitation wavelength, only the optimal one rather than combinations of multiple excitation wavelengths is picked.

After obtaining each PAH's optimal excitation wavelength Ex_p, the corresponding emission spectral data (K^*J) are extracted to select the optimal emission interval. The CPSO algorithm was utilized to regulate the process of interval selection according to the following steps.

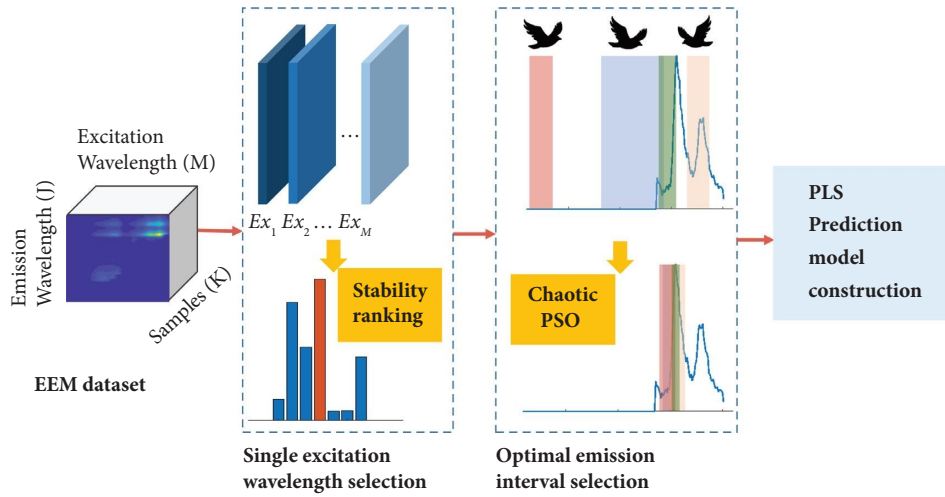

Figure 2: Flowchart of the single-excitation interval selection of EEM.

Step 4: Initialize the chaos control parameter Flag = 0 and an appropriate integer Q. Generate I particles, each containing two dimensions x_start and x_end. The variables between x_start and x_end are used as independent variables in the PLS model. The positions and velocities of the particles are randomized in [0, 1] and [1, J], and the fitness of each particle is calculated with fitness = RMSECV (X_c).

Step 5: Update the position and velocity of the particle according to the updated formulas (2)-(3). If g_{best} remains the same as the previous generation, Flag = Flag + 1;

Step 6: If Flag > Q, chaotic perturbation is performed on the particles. The particles are arranged in ascending order of fitness, and the best 20% particles are chaotically perturbed in the search space according to the logistic optimization method.

Step 7: If a better g_{best} is obtained after chaotic perturbation, update the p_g and g_{best}, then reset Flag = 0. Generate 80% I particles randomly and evaluate the fitness of these particles. The new swarm is formed from newly generated and retained particles.

Step 8: Go back to Step 5 until termination conditions are met, then output p_g (x_start, x_end) as the best emission spectral interval. Build the prediction model based on the selected variables.

2.5. Software Tools. All EEM data analysis and model development were conducted in MATLAB 2018a (The MathWorks Inc., USA). The drEEM toolbox (https://www.models.life.ku.dk) is used for data integration and scattering elimination.

3. Results and Discussion

3.1. Fluorescence Spectra of Four PAHs. The EEMs of FLU, PYR, PHE, and BaP are shown in Figure 3. It can be seen that the fluorescence spectrum of FLU is in the range of excitation (Ex) 230–310 nm/emission (Em) 280–360 nm, with a fluorescence peak at Ex 270 nm/Em 310 nm. The fluorescence spectrum of PYR is in the range of Ex 260 nm–350 nm/Em 340 nm–400 nm, with two fluorescence peaks at Ex 330 nm/Em 380 nm and Ex 270 nm/Em 380 nm. The fluorescence spectrum of PHE is in the range of Ex 230–310 nm/Em 340–400 nm, and the fluorescence peaks are located at Ex 250 nm/Em 360 nm and Ex 280 nm/Em 360 nm. The fluorescence spectrum of BaP is in the range of Ex 250–400 nm/Em 380–450 nm, and the prominent fluorescence peaks are located at Ex 380 nm/Em 400 nm and Ex 380 nm/Em 430 nm; there are two weaker peaks at Ex 290 nm. The ratio of peak fluorescence intensity of FLU, PYR, PHE, and BaP was 7 : 4 : 1 : 45 at the concentration of 20 μg/L.

Figure 4 shows the effect of HA fluorescence on PAH detection. Figures 4(a) and 4(b) are PAH's fluorescence spectrum without/with HA being added. The spectrum of HA displays a continuous absorption and emission band, which is widely present at Ex > 250 nm and Em > 350 nm. Moreover, the fluorescence intensity of HA increases as the excitation wavelength increases. HA and PAH fluorescence have different degrees of overlap, so the fluorescence characteristics of PAH are blurred, which degrades the modeling accuracy significantly by simply using peak intensity.

3.2. Results of Wavelength Selection

3.2.1. Optimal Excitation Wavelength Selection. The excitation wavelengths of four PAHs were selected, as described in Section 2.4. Emission spectral data corresponding to the 1st–21st excitation wavelength were extracted as independent variables, and then, 21*4 = 84 PLS models were established with concentrations of FLU, PYR, PHE, and BaP as dependent variables. Figure 5 shows the variation of explained variance and RMSECV with nLVs (FLU, Ex = 280 nm). In this case, the nLVs were decided to be four because the first four latent variables contain 99% of the total variation, and the RMSECV was stabilized at a lower level. The prediction models corresponding to each excitation

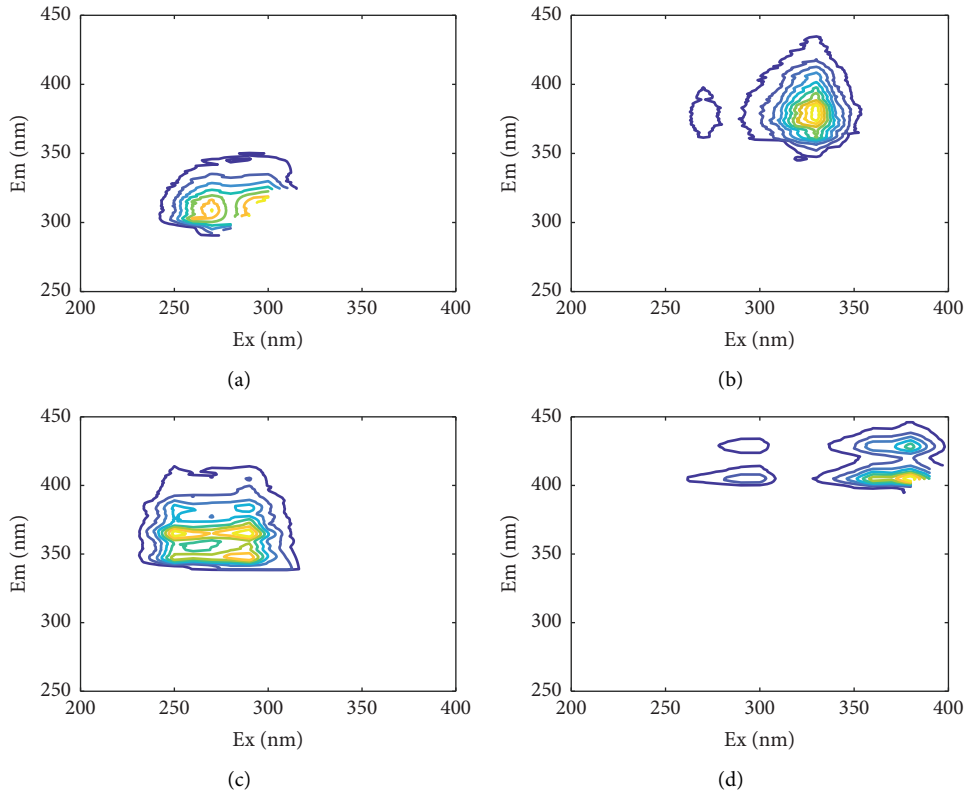

Figure 3: EEM contours of (a) FLU, (b) PYR, (c) PHE, and (d) BaP.

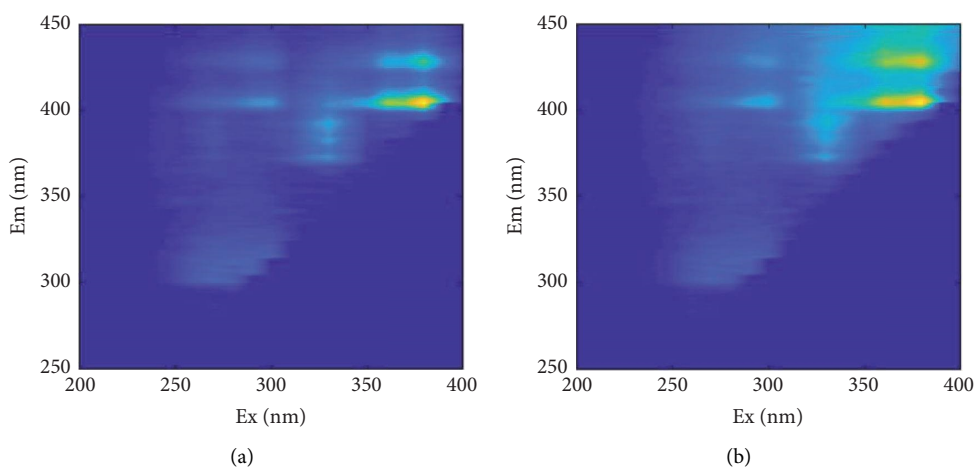

Figure 4: Spectra of a sample in (a) Mix-PAH dataset and (b) HA-PAH dataset. The concentration of HA is 10 mg/L.

wavelength were built 50 times in the Mix-PAH dataset and HA-PAH dataset, and the stability of each excitation wavelength for four PAHs was obtained and is shown in Figure 6.

In the presence of HA fluorescence, the excitation wavelength distribution of HA-PAH changes obviously. As Figure 6 shows, since the fluorescence of HA is stronger at longer wavelengths, there is a significant blue shift in the stability of the excitation wavelength. For example, the stability of PYR decreases at 330 nm and increases at 310 nm and that of BaP decreases at 390 nm and increases at 370 nm. The blue shift phenomenon is particularly evident for FLU. The optimal excitation wavelength for the Mix-PAH dataset is 300 nm, but the optimal excitation wavelength of HA-PAH is moved to 270 nm under the influence of HA, and the stabilty of between 250 nm and 270 nm is significantly improved. The stabilty of PHE varies insignificantly, and the optimal excitation wavelength remains at 250 nm. This suggests that the effectiveness of each excitation wavelength varies in different environmental conditions,

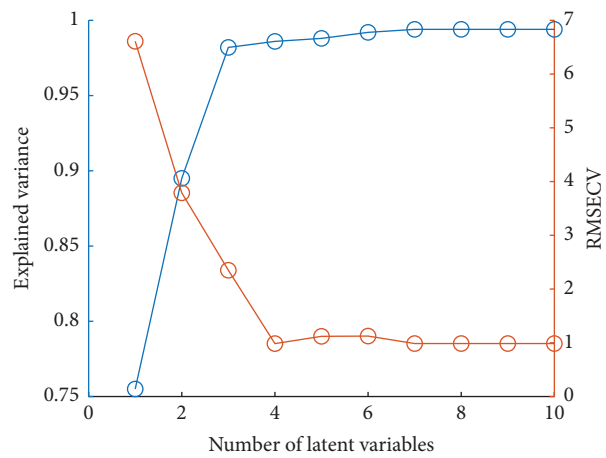

Figure 5: Explained variance and RMSECV variation with nLVs.

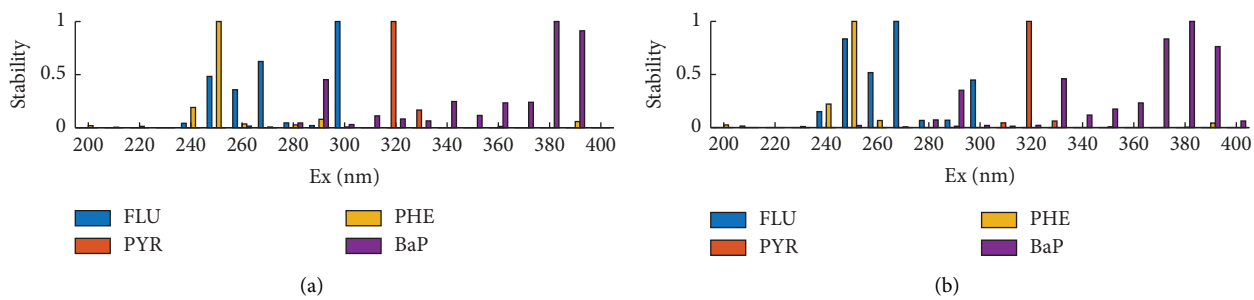

Figure 6: Stability distribution of excitation wavelengths in (a) Mix-PAH and (b) HA-PAH datasets.

resulting in a variation in reliability. It is helpful to build a better prediction model by reweighing the effective and interference information.

3.2.2. Optimal Emission Interval Selection. After determining the excitation wavelengths of the four PAHs, the spectral variables corresponding to the optimal excitation wavelengths were extracted, and CPSO was used to optimize the wavelength interval. A swarm containing 30 particles is created, and x_start and x_end are randomly distributed between 1 and 201, with the restriction x_start < x_end-nLVs to ensure that the number of interval variables is larger than the nLVs. ω decreases linearly from 0.9 to 0.4, accelerating constant $c_1 = c_2 = 2$, and particle velocity was limited in [−2, 2]. The maximum number of iterations of the particle swarm algorithm is 100; while g_{best} is not updated for 10 (corresponding Q) consecutive iterations, chaotic perturbations are applied to update related parameters.

The fitness variation with iterations during the operation of the CPSO is shown in Figure 7. It can be seen that with the increase in the number of iterations, the fitness value gradually decreases until the maximum number of iterations is reached. The result of standard PSO with the same random seed is also shown for comparison. PSO and CPSO have the same trend in the initial iteration. However, in the 26th iteration, chaotic perturbation is applied for the first time, reducing the fitness and showing the difference between CPSO and PSO. Chaotic perturbation was also applied in the 41st, 42nd, 43rd, and 61st iterations, but chaotic perturbation in the 41st and 42nd iterations did not produce better results. None of the chaotic perturbations output better results after the 76th iteration, which indicates that a stable solution has been obtained in the search space. Obviously, CPSO gets a better fitness value and a higher rate of convergence.

Because of the different random numbers, the optimal interval chosen for each run is not exactly the same. The RMSECVs were calculated 50 times, and the frequency distribution was obtained to evaluate the importance of each variable. Figure 8 shows different frequency distributions of four PAHs of the Mix-PAH dataset. CPSO mainly selects the regions near 330 nm, 370 nm, 350 nm, and 410 nm for FLU, PYR, PHE, and BaP, respectively, which is consistent with the spectra distribution. The frequency of selected variables decreases sequentially from the center to both sides, indicating a gradual decrease in the importance of these variables.

The different frequency distributions of HA-PAH are shown in Figure 9. Because HA has a stronger fluorescence at longer wavelengths, the interval center of FLU was transformed to 320 nm from 330 nm, indicating a tendency towards shorter wavelengths. Compared with Mix-PAH, the wavelength interval of PYR showed a significant wavelength broadening to capture more informative data for modeling. The distribution of PHE's wavelength did not change

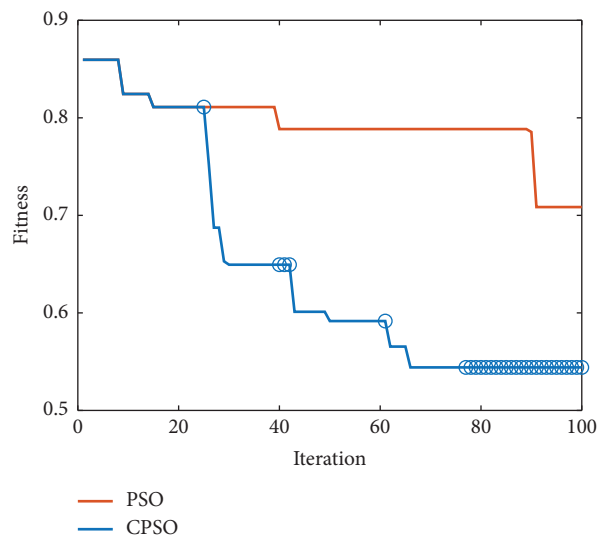

Figure 7: Trend graph of the fitness for PSO and CPSO. The red and blue lines represent the optimization of PSO and CPSO under the same random seed, respectively.

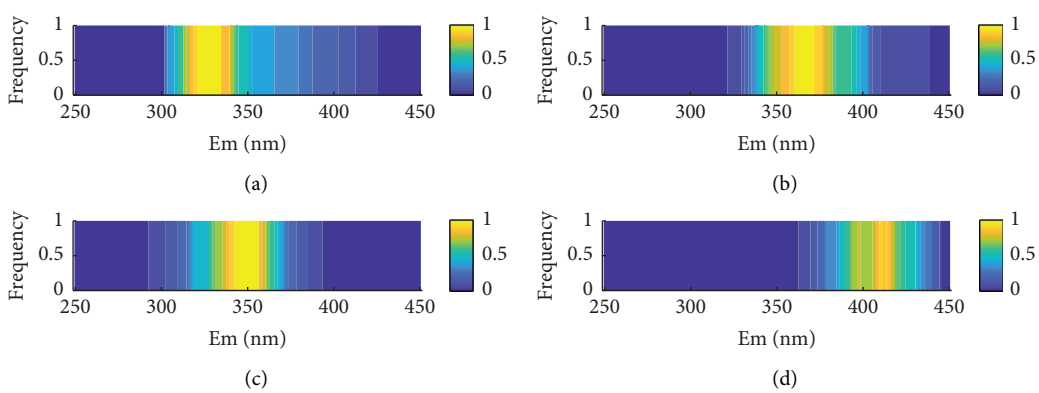

Figure 8: Frequency distribution of variables selected on (a) FLU, (b) PYR, (c) PHE, and (d) BaP by CPSO in the Mix-PAH dataset. Yellow indicates that the corresponding wavelength is selected at a higher frequency.

significantly, with a slight contraction and shift towards shorter wavelengths observed. The fluorescence distribution of BaP showed a more concentrated trend, which is concentrated in the peak BaP region (400–430 nm). The results show that compared with the Mix-PAH dataset, the emission distribution of HA-PAH produces different degrees of deviation to obtain more effective information.

The obtained variable distributions decide the optimal emission interval; the RMSECV corresponding to each frequency interval is calculated starting from the interval with the highest frequency to eliminate the interference information as much as possible, and the interval with the smallest RMSECV is selected as the modeling interval.

3.2.3. Model Establishment and Results of Concentration Prediction. The spectra of the optimal intervals of five test samples were measured to verify the validity of the selected intervals. A single measurement takes less than 30 seconds, which is less time than a full EEM (2-3 minutes) and chromatography. Based on the selected spectral intervals, single-excitation interval PLS prediction models for HA-PAH were established. The relation between predicted and actual values is shown in Figure 10. For the results of FLU, PYR, and BaP, each data point is close to the projected regression line, concluding that the predicted model fits the data well. By comparison, the predicted results for PHE showed more significant deviations, partly because of the severe spectral overlap between PHE and other fluorophores and partly because of the relatively lower fluorescence efficiency. The prediction values of the test set are shown in Table 2. The MRE values of Mix-PAH were 2.05%, 4.87%, 6.90%, and 3.20%, respectively, which achieved satisfactory results. The MRE values of HA-PAH were 4.30%, 4.54%, 7.39%, and 3.36%, respectively. Despite the fluorescence interference from HA, the MRE values of HA-PAH were still lower than 10%, indicating that the wavelength selection method screened out the most informative wavelength variables for modeling under HA interference.

Two common modeling methods without variable selection, peak-picking and full-spectrum (FS-PLS), and two

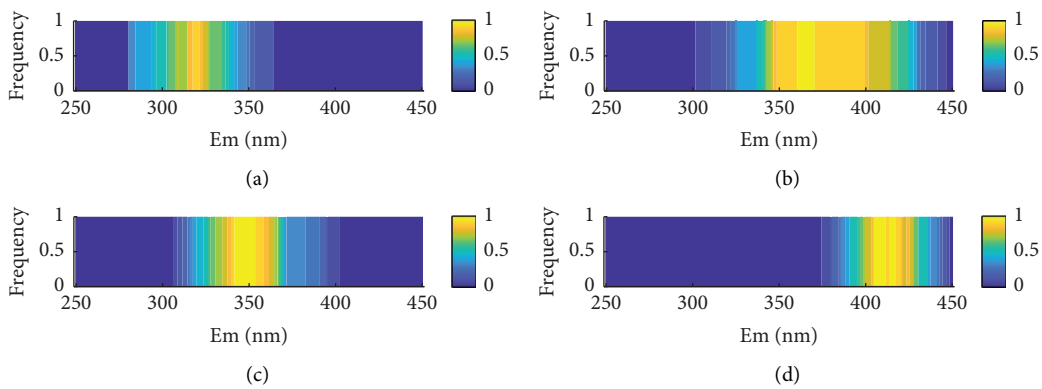

Figure 9: Frequency distribution of variables selected in (a) FLU, (b) PYR, (c) PHE, and (d) BaP by CPSO in the HA-PAH dataset.

Table 2: Prediction values of single-excitation interval PLS in the test samples (μg/L).

	Actual concentration	3.0	7.5	12.5	17.5	15.0	MRE (%)
FLU	Prediction concentration (Mix-PAH)	3.90	7.35	12.88	17.77	14.86	2.05
	Prediction concentration (HA-PAH)	3.31	7.36	12.19	18.31	15.84	4.30
	Actual concentration	6.0	17.5	12.5	2.5	15.0	
PYR	Prediction concentration (Mix-PAH)	6.41	16.31	13.24	2.07	15.13	4.87
	Prediction concentration (HA-PAH)	4.98	16.95	12.90	2.52	14.18	4.54
	Actual concentration	9.0	16.0	5.0	12.5	20.0	
PHE	Prediction concentration (Mix-PAH)	8.43	13.75	6.59	12.78	19.46	6.90
	Prediction concentration (HA-PAH)	7.37	15.37	5.18	11.47	17.97	7.39
	Actual concentration	12.0	9.0	3.0	6.0	16.0	
BaP	Prediction concentration (Mix-PAH)	12.51	9.44	3.57	6.04	15.70	3.20
	Prediction concentration (HA-PAH)	12.47	9.52	3.70	6.13	15.72	3.36

MRE values are listed in the last column.

representative wavelength selection methods, CSMWPLS and MC-UVE, were applied to construct prediction models to evaluate the performance of the proposed method. The excitation wavelength of FS-PLS, CSMWPLS, and MC-UVE was determined by the same method in Section 2.3. In determining the optimal modeling interval with CSMWPLS, the window size was expanded from 10 to 100 with an interval of 10, and the window was slid on the corresponding emission wavelengths to select the best interval by RMSECV. Similarly, MC-UVE selected the first $10^{*}n$ ($n = 1, 2, 3, \ldots, 10$) wavelengths for cross-validation and selected the variable set with the smallest RMSECV for modeling.

The results of different modeling approaches on the HA-PAH dataset are listed in Table 3. For HA-PAH, the peak-picking method has the highest RMSEP values, while the R^2 is also the smallest, indicating that the predictive performance of the univariate prediction model is severely weakened in HA-PAH samples. FS-PLS outperformed the peak-picking method, indicating the advantage of multiple linear regressions. However, some variables containing irrelevant information are still involved in the modeling, which limits the accuracy of the model. In the presence of HA, variable selection still exhibited its contribution to modeling (lower RMSEP and higher R^2), and the selected wavelengths reduced the interference of CDOM to some extent and had more valid information in the model construction. The model accuracy and linearity of both MC-UVE and CSMWPLS exceeded FS-PLS, confirming the presence of a large number of uninformative and interfering variables in the emission spectra, and the wavelength selection method extracted the critical information. The RMSEPs of single-excitation interval PLS are 0.5588, 0.6598, 1.3502, and 0.4654, respectively, with more satisfactory results compared to CSMWPLS and MC-UVE.

In order to validate the performance of the PLS models for PAH measurement, recovery experiments were conducted, and the results are displayed in Table 4. Four PAHs with concentrations of 2.38, 1.19, and 0.71 μg/L were added to the original solution, and each sample was measured three times. The spiked concentrations were calculated to obtain average recoveries of 103.3%, 103.2%, 107.4%, and 101.4%, respectively. The results showed that the spiked samples with different concentrations had similar recoveries, which were all close to 100%, indicating that the developed model has good precision and reliability.

3.2.4. LOD and LOQ for the Simultaneous Determination of Four PAHs with HA. The limit of detection (LOD) and limit of quantitation (LOQ) are critical metrics that combine the sensitivity and precision of the analytical determination, playing a pivotal role in the assessment of the model's

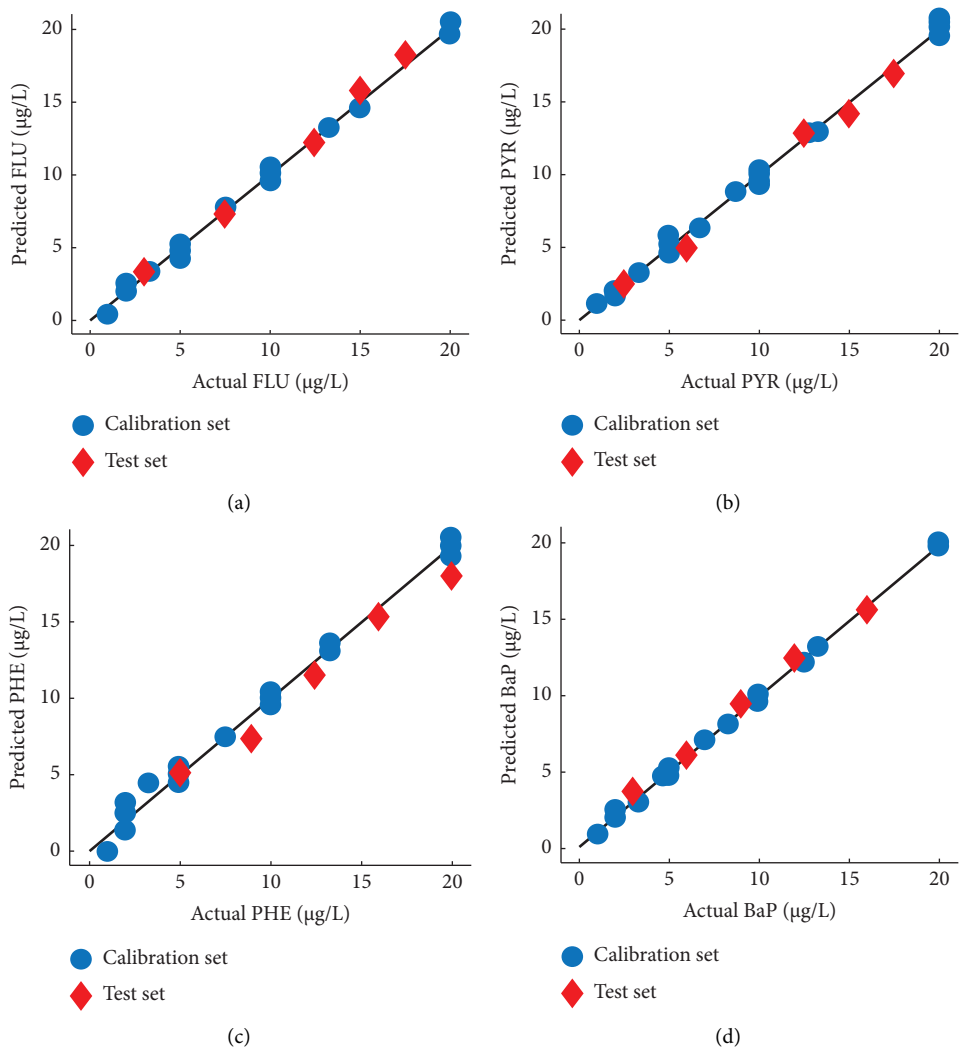

Figure 10: Relation between the predicted concentration and the observed concentration of (a) FLU, (b) PYR, (c) PHE, and (d) BaP in the HA-PAH dataset.

Table 3: Prediction results of different methods in the HA-PAH dataset (μg/L).

	Method	RMSEP	R_c^2	R_t^2
FLU	Peak-picking	0.6436	0.9810	0.9865
	FS-PLS	0.5915	0.9924	0.9939
	CSMWPLS	0.5884	0.9980	0.9965
	MC-UVE	0.5739	0.9975	0.9916
	Single-excitation interval PLS	0.5588	0.9974	0.9943
PYR	Peak-picking	2.9173	0.7608	0.8553
	FS-PLS	0.8908	0.9974	0.9887
	CSMWPLS	0.8074	0.9928	0.9917
	MC-UVE	0.7536	0.9969	0.9990
	Single-excitation interval PLS	0.6598	0.9969	0.9911
PHE	Peak-picking	2.8100	0.3122	0.4313
	FS-PLS	1.6104	0.8950	0.9562
	CSMWPLS	1.5048	0.9090	0.9770
	MC-UVE	1.4578	0.8725	0.9350
	Single-excitation interval PLS	1.3502	0.9925	0.9828
BaP	Peak-picking	0.5109	0.9959	0.9921
	FS-PLS	0.4950	0.9992	0.9953
	CSMWPLS	0.4975	0.9992	0.9956
	MC-UVE	0.4765	0.9986	0.9957
	Single-excitation interval PLS	0.4654	0.9991	0.9966

R_c^2 and R_t^2 are the coefficients of determination of the calibration set and test set.

Table 4: Results of the recovery tests for four PAH measurements.

Spiked PAHs	Original concentration	Added concentration	Measured concentration	Recovery (%)	Average recovery (%)
FLU	9.88 ± 0.19	2.38	12.39 ± 0.31	105.4	103.3
	9.88 ± 0.19	1.19	11.17 ± 0.22	108.4	
	9.88 ± 0.19	0.71	10.57 ± 0.17	97.2	
PYR	10.16 ± 0.23	2.38	12.75 ± 0.20	109.0	103.2
	10.16 ± 0.23	1.19	11.32 ± 0.19	97.7	
	10.16 ± 0.23	0.71	10.89 ± 0.32	102.8	
PHE	9.84 ± 0.39	2.38	12.39 ± 0.14	107.2	107.4
	9.84 ± 0.39	1.19	11.14 ± 0.21	109.5	
	9.84 ± 0.39	0.71	10.59 ± 0.23	105.6	
BaP	9.96 ± 0.14	2.38	12.30 ± 0.18	98.2	101.4
	9.96 ± 0.14	1.19	11.19 ± 0.17	103.5	
	9.96 ± 0.14	0.71	10.69 ± 0.12	102.4	

performance. For the multiple linear regressions such as PLS, the LOD and LOQ can be calculated following the method pioneered by Olivieri et al.:

$$\mathrm{LOD} = 3.3\left(\mathrm{SEN}_n^{-2}\,\mathrm{var}\,(x) + h\mathrm{SEN}_n^{-2}\,\mathrm{var}\,(x) + h\,\mathrm{var}\left(y_{\mathrm{cal}}\right)\right)^{1/2},$$
$$\mathrm{LOQ} = 10\left(\mathrm{SEN}_n^{-2}\,\mathrm{var}\,(x) + h\,\mathrm{SEN}_n^{-2}\,\mathrm{var}\,(x) + h\,\mathrm{var}\left(y_{\mathrm{cal}}\right)\right)^{1/2}, \tag{10}$$

where SEN represents the sensitivity associated with the specific detection substance, var (x) is the variance in instrumental signals, h stands for the leverage of each sample, and var (y_{cal}) is the variance in the calibration concentrations [35]. Each sample in the multivariate model has a specific leverage value h due to differences in spectral characteristics. Therefore, the LOD of the multivariate model is distributed within a leverage value modulated interval, which is different from the univariate model. Here, average LOD and average LOQ were used as metrics to assess model performance.

Under HA interference, the LOD values calculated for the four PAHs were 0.4412, 0.4475, 0.9417, and 0.2088, respectively. Accordingly, the LOQ values of the four PAHs were consistent with the trend of LOD, which were 1.3369, 1.2652, 2.8536, and 0.6326, respectively. The different LOD values among PAHs are mainly due to the difference in fluorescence signal intensity. PAHs with stronger fluorescence tend to have smaller regression coefficients, resulting in larger SEN. In addition, due to the interference of different concentrations of HA, PAHs with weaker fluorescence intensity usually have higher var (y_{cal}). Consequently, FLU, PYR, and BaP have better LOD and LOQ, especially BaP, with the strongest fluorescence signal. In contrast, FLU with the weakest fluorescence intensity had the highest LOD and LOQ.

4. Conclusions

A method integrating EEM and single-excitation interval selection was proposed for detecting four PAHs under CDOM interference. For each PAH, the optimal excited wavelength and emission interval are optimized for modeling, and multiple PLS prediction models are built to predict the different PAHs. Under CDOM interference, the MRE of this method is less than 10%, and it also has smaller RMSEPs and higher R^2 compared to the other modeling methods, which show satisfactory results. Sound detection limits and recovery rates further validate the reliability of the model. Compared with chromatography and full EEM, this study provides a new method for rapid and accurate analysis of multiple PAHs in water. Further research will explore methods to compensate for the effects of temperature and turbidity to improve the reliability of predictive models in real environments.

Data Availability

The data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant no. 2022YFC3103900).

References

[1] A. Patel, S. Shaikh, K. R. Jain, C. Desai, and D. Madamwar, "Polycyclic aromatic hydrocarbons: sources, toxicity, and remediation approaches," *Frontiers in Microbiology*, vol. 11, Article ID 562813, 2020.

[2] C. A. Menzie, B. B. Potocki, and J. Santodonato, "Exposure to carcinogenic PAHs in the environment," *Environmental Science and Technology*, vol. 26, no. 7, pp. 1278–1284, 1992.

[3] Iarc Working Group on the Evaluation of Carcinogenic Risks to Humans, "Some non-heterocyclic polycyclic aromatic hydrocarbons and some related exposures," *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, vol. 92, pp. 1–853, 2010.

[4] M. M. Mumtaz, J. D. George, K. W. Gold, W. Cibulas, and C. T. Derosa, "ATSDR evaluation of health effects of

chemicals. IV. Polycyclic aromatic hydrocarbons (PAHs): understanding a complex problem," *Toxicology and Industrial Health*, vol. 12, no. 6, pp. 742–971, 1996.

[5] N. R. Ekere, N. M. Yakubu, T. Oparanozie, and J. Ihedioha, "Levels and risk assessment of polycyclic aromatic hydrocarbons in water and fish of Rivers Niger and Benue confluence Lokoja, Nigeria," *Journal of Environmental Health Science and Engineering*, vol. 17, no. 1, pp. 383–392, 2019.

[6] United States Environmental Protection Agency, *EPA Method 8270E(SW-846): Semivolatile Organic Compounds by Gas Chromatography/Mass Spectrometry (GC/MS)*, United States Environmental Protection Agency, Washington, DC, USA, 2014.

[7] N. D. Forsberg, G. R. Wilson, and K. A. Anderson, "Determination of parent and substituted polycyclic aromatic hydrocarbons in high-fat salmon using a modified QuEChERS extraction, dispersive SPE and GC-MS," *Journal of Agricultural and Food Chemistry*, vol. 59, no. 15, pp. 8108–8116, 2011.

[8] E. Mansouri, V. Yousefi, V. Ebrahimi et al., "Overview of ultraviolet-based methods used in polycyclic aromatic hydrocarbons analysis and measurement," *Separation Science and Technology*, vol. 3, no. 4, pp. 112–120, 2020.

[9] D. E. Zacharioudaki, I. Fitilis, and M. Kotti, "Review of fluorescence spectroscopy in environmental quality applications," *Molecules*, vol. 27, no. 15, p. 4801, 2022.

[10] L. Ferey and N. Delaunay, "Capillary and microchip electrophoretic analysis of polycyclic aromatic hydrocarbons," *Analytical and Bioanalytical Chemistry*, vol. 407, no. 10, pp. 2727–2747, 2015.

[11] X. Meng, Y. Li, Y. Zhou et al., "Real-time immuno-PCR for ultrasensitive detection of pyrene and other homologous PAHs," *Biosensors and Bioelectronics*, vol. 70, pp. 42–47, 2015.

[12] L. Greene, B. Elzey, M. Franklin, and S. O. Fakayode, "Analyses of polycyclic aromatic hydrocarbon (PAH) and chiral-PAH analogues-methyl-β-cyclodextrin guest-host inclusion complexes by fluorescence spectrophotometry and multivariate regression analysis," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 174, pp. 316–325, 2017.

[13] R. Yang, G. Dong, X. Sun et al., "Feasibility of the simultaneous determination of polycyclic aromatic hydrocarbons based on two-dimensional fluorescence correlation spectroscopy," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 190, pp. 342–346, 2018.

[14] F. Cyr, M. Tedetti, F. Besson, N. Bhairy, and M. Goutx, "A glider-compatible optical sensor for the detection of polycyclic aromatic hydrocarbons in the marine environment," *Frontiers in Marine Science*, vol. 6, p. 110, 2019.

[15] S. Li, Y. Chen, J. Zhang et al., "The relationship of chromophoric dissolved organic matter parallel factor analysis fluorescence and polycyclic aromatic hydrocarbons in natural surface waters," *Environmental Science and Pollution Research*, vol. 25, no. 2, pp. 1428–1438, 2018.

[16] E. S. Boyle, N. Guerriero, A. Thiallet, R. Del Vecchio, and N. V. Blough, "Optical properties of humic substances and CDOM: relation to structure," *Environmental Science and Technology*, vol. 43, no. 12, pp. 4612–2268, 2009.

[17] R. Yang, N. Zhao, X. Xiao, S. Yu, J. Liu, and W. Liu, "Determination of polycyclic aromatic hydrocarbons in the presence of humic acid in water," *Applied Spectroscopy*, vol. 70, no. 9, pp. 1520–1528, 2016.

[18] N. Ferretto, M. Tedetti, C. Guigue, S. Mounier, R. Redon, and M. Goutx, "Identification and quantification of known polycyclic aromatic hydrocarbons and pesticides in complex mixtures using fluorescence excitation–emission matrices and parallel factor analysis," *Chemosphere*, vol. 107, pp. 344–353, 2014.

[19] Y. H. Yun, H. Li, L. R. E Wood et al., "An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 111, pp. 31–36, 2013.

[20] H. Wang, P. Chen, J. Dai et al., "Recent advances of chemometric calibration methods in modern spectroscopy: algorithms, strategy, and related issues," *TRAC, Trends in Analytical Chemistry*, vol. 153, Article ID 116648, 2022.

[21] C. Hao, Y. Wang, G. Wang, and Z. Zhu, "An assemble based on clustering and Monte Carlo for the wavelengths selection of excitation emission fluorescence spectra," *Applied Sciences*, vol. 10, no. 3, p. 1134, 2020.

[22] X. Wu, J. Li, J. Jie, and H. Zheng, "Three-dimensional fluorescence spectra model optimisation for water quality analysis based on particle swarm optimisation," *International Journal of Wireless and Mobile Computing*, vol. 7, no. 2, pp. 200–206, 2014.

[23] Y. Zhang, D. Zhu, Y. Chen, Z. Liu, W. Duan, and S. Li, "Wavelength selection method of algal fluorescence spectrum based on convex point extraction from feature region," *Spectroscopy and Spectral Analysis*, vol. 42, pp. 3031–3038, 2022.

[24] L. Wang, X. Ren, X. Wang et al., "Polycyclic aromatic hydrocarbons (PAHs) in the upstream rivers of Taihu Lake Basin, China: spatial distribution, sources and environmental risk," *Environmental Science and Pollution Research*, vol. 1-10, 2022.

[25] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001.

[26] Y. Yun, Y. Liang, G. Xie, H. Li, D. Cao, and Q. Xu, "A perspective demonstration on the importance of variable selection in inverse calibration for complex analytical systems," *Analyst*, vol. 138, no. 21, pp. 6412–6421, 2013.

[27] Y. Yun, H. Li, B. Deng, and D. Cao, "An overview of variable selection methods in multivariate analysis of near-infrared spectra," *TRAC, Trends in Analytical Chemistry*, vol. 113, pp. 102–115, 2019.

[28] W. Cai, Y. Li, and X. Shao, "A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 90, no. 2, pp. 188–194, 2008.

[29] Y. Du, Y. Liang, J. Jiang, R. J. Berry, and Y. Ozaki, "Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares," *Analytica Chimica Acta*, vol. 501, no. 2, pp. 183–191, 2004.

[30] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the ICNN'95- International Conference on Neural Networks*, vol. 4, pp. 1942–1948, Perth, Australia, November, 1995.

[31] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," *IEEE Transactions on Evolutionary Computation*, vol. 69-73, 1998.

[32] H. Wang, H. Sun, C. Li, S. Rahnamayan, and J. Pan, "Diversity enhanced particle swarm optimization with neighborhood search," *Information Sciences*, vol. 223, pp. 119–135, 2013.

[33] B. Liu, L. Wang, Y. Jin, F. Tang, and D. Huang, "Improved particle swarm optimization combined with chaos," *Chaos, Solitons and Fractals*, vol. 25, no. 5, pp. 1261–1271, 2005.

[34] X. Xu, H. Rong, M. Trovati, M. Liptrott, and N. Bessis, "CS-PSO: chaotic particle swarm optimization algorithm for solving combinatorial optimization problems," *Soft Computing*, vol. 22, no. 3, pp. 783–795, 2018.

[35] A. C. Olivieri, "Analytical figures of merit: from univariate to multiway calibration," *Chemical Reviews*, vol. 114, no. 10, pp. 5358–5378, 2014.