Exploratory analysis of excitation-emission matrix fluorescence spectra with self-organizing maps as a basis for determination of organic matter removal efficiency at water treatment works

Magdalena Bieroza,[1] Andy Baker,[2] and John Bridgeman[1]

Received 21 January 2009; revised 22 May 2009; accepted 3 September 2009; published 15 December 2009.

[1] In the paper, the self-organizing map (SOM) was employed for the exploratory analysis of fluorescence excitation-emission data characterizing organic matter removal efficiency at 16 water treatment works in the UK. Fluorescence spectroscopy was used to assess organic matter removal efficiency between raw and partially treated (clarified) water to provide an indication of the potential for disinfection by-products formation. Fluorescence spectroscopy was utilized to evaluate quantitative and qualitative properties of organic matter removal. However, the substantial amount of fluorescence data generated impeded the interpretation process. Therefore a robust SOM technique was used to examine the fluorescence data and to reveal patterns in data distribution and correlations between organic matter properties and fluorescence variables. It was found that the SOM provided a good discrimination between water treatment sites on the base of spectral properties of organic matter. The distances between the units of the SOM map were indicative of the similarity of the fluorescence samples and thus demonstrated the relative changes in organic matter content between raw and clarified water. The higher efficiency of organic matter removal was demonstrated for the larger distances between raw and clarified samples on the map. It was also shown that organic matter removal was highly dependent on the raw water fluorescence properties, with higher efficiencies for higher emission wavelengths in visible and UV humic-like fluorescence centers.

Citation: Bieroza, M., A. Baker, and J. Bridgeman (2009), Exploratory analysis of excitation-emission matrix fluorescence spectra with self-organizing maps as a basis for determination of organic matter removal efficiency at water treatment works, *J. Geophys. Res.*, *114*, G00F07, doi:10.1029/2009JG000940.

1. Introduction

[2] Organic matter is a complex heterogeneous mixture of chemical compounds ubiquitous in all natural and anthropogenically transformed environments, from marine deep-water, through estuarial to freshwater ecosystems. The inherent complexity of organic matter structure and function determines the substantial number of different physical and chemical methods used in organic matter characterization. Recent advances in spectrofluorometric techniques have concentrated on the development of more accurate, portable and faster instruments, enhanced optical analysis efficiency and have stimulated academic and industrial interest in utilization of intrinsic spectral properties of organic matter in characterization of its composition and role in a variety of ecosystems [*Mopper and Schultz*, 1993; *Coble*, 1996; *McKnight et al.*, 2001; *Stedmon et al.*, 2003; *Boehme et al.*, 2004; *Hudson et al.*, 2007].

[3] In the work presented here, fluorescence spectroscopy was used for the quantitative and qualitative characterization of organic matter during water treatment. During the water treatment process, raw water organic matter should be effectively removed prior to disinfection due to its propensity for forming toxic and carcinogenic disinfection by-products (DBPs) as a result of its chemical reaction with chlorine. As fluorescence measurements are rapid and noninvasive with the possibility for incorporation into online monitoring system, fluorescence spectroscopy can provide an accurate assessment of organic matter removal efficiency during treatment processes and facilitate online prediction of DBPs formation potential [*Bieroza et al.*, 2008]. Moreover, organic matter characterization can provide an insight into the dependence of organic matter removal efficiency on the character of organic matter, based on the presence, relative importance and spectral properties of particular fluorophores in raw water.

[4] Although the acquisition of fluorescence data with improved instrumentation has become easier and faster, the substantial amount of fluorescence data generated impedes the interpretation process and requires adequately robust statistical and computational analysis tools. A common output from fluorescence spectroscopy is the excitation-emission matrix (EEM), produced by scanning fluorescence intensity over a range of excitation and emission wavelengths,

[1]School of Civil Engineering, University of Birmingham, Birmingham, UK.
[2]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK.

0148-0227/09/2009JG000940$09.00

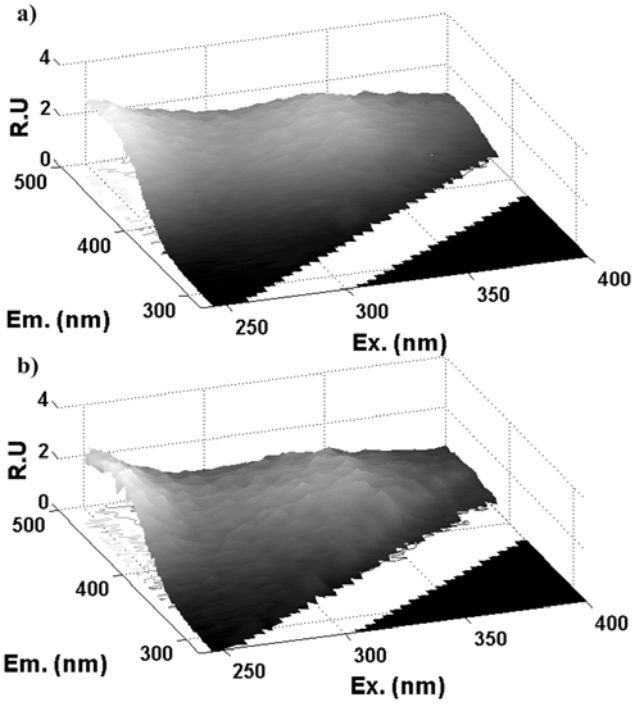

Figure 1. Excitation-emission spectra for (a) raw and (b) clarified water (site 1).

to produce a three-dimensional output comprising of more than 4000 fluorescence data points (Figure 1).

[5] Fluorescence intensity occurs primarily in the regions corresponding with spectral location of particular fluorophores (humic-, fulvic- and protein-like fluorescence). The traditional approach to EEM spectra analysis is for an expert to read the maximum fluorescence intensity from defined spectral regions (peaks). This "peak-picking" method significantly reduces the amount of fluorescence data to a subset of excitation-emission pairs with corresponding fluorescence intensity values. The high dimensionality of the initial fluorescence data set is therefore reduced in a supervised manner through expert analysis of specific regions of known spectral composition. Thus, to retrieve fluorescence information on the variability of organic matter accurately and effectively, fluorescence EEMs should be analyzed as a whole with the use of appropriate statistical techniques. The most common techniques of EEM data analysis include various multiway methods, such as Principal Components Analysis (PCA, unsupervised algorithm) and Parallel Factor Analysis (PARAFAC, supervised algorithm). Exploratory analysis of fluorescence spectra has been primarily conducted with the use of traditional statistical methods of unsupervised data analysis, such as Principal Components Analysis or Principal Filter Analysis [*Brunsdon and Baker*, 2002; *Persson and Wedborg*, 2001; *Boehme et al.*, 2004; *Spencer et al.*, 2007]. Those methods of fluorescence data decomposition (PCA, PARAFAC) have a range of limitations (e.g., poor noise tolerance) and require a time-consuming interpretation and validation of resultant components [*Bro*, 1998; *Stedmon et al.*, 2003]. Therefore, the development and use of more robust tools for fluorescence data decomposition and pattern recognition is warranted. In this paper the application of an unsupervised method of fluorescence data decomposition, the self-organizing map (SOM), was employed to discern patterns within a fluorescence data set and provide implications for removal of organic matter in drinking water treatment.

[6] The SOM is an example of an unsupervised classification algorithm in which a pattern (if it exists) is assigned to a category, not specified or not known a priori by the domain-expert analyzing the fluorescence data. This approach is often used in data clustering, where the input feature space (here related to fluorescence EEM data) is explored to discern any reasonable relationships among the data, often without prior knowledge or assumptions on the data set given. In a SOM, the feature extraction from the input domain is performed with a nonlinear (SOM) transformation of the input data onto a *k*-dimensional map (grid). A feature describes an elementary pattern of information that represents partial aspects or properties of an item [*Kohonen*, 1998]. In fluorescence data analysis with SOM, the extracted features can be referred to presence of particular fluorophore (or group of fluorophores) or its specific spectral properties.

[7] The SOM is an example of two-layered Artificial Neural Network (ANN), consisting of a number of interconnected single processing units called neurons or nodes. ANNs can be considered as parallel interconnected networks of single computational elements (neurons) organized in hierarchical way, with structure and functions that imitate the biological nervous system. ANNs have the ability to learn the pattern from the input features or the model input-output relationship based on the training algorithms where weights vectors are stored in connections between neurons which are adjusted to minimize the overall error of network prediction. In a SOM network the connection weights of the size of the input data *m* are stored in input neurons and during training are projected onto *k*-dimensional output space (Figure 2) [*Kohonen*, 1998; *Rhee et al.*, 2005; *Garcia et al.*, 2007].

[8] The aim of this paper is to demonstrate the use of the SOM technique for exploratory analysis of fluorescence EEMs characterization of water treatment works (WTW) performance in terms of organic matter removal. The basic concepts, application and interpretation of a SOM are presented to familiarize fluorescence EEM users with this robust and efficient feature extraction technique.

2. Materials and Methods

2.1. Fluorescence Data

[9] Fluorescence spectroscopy measurements and total organic carbon (TOC) analyses were carried out on samples of raw and clarified water from 16 surface WTWs, collected

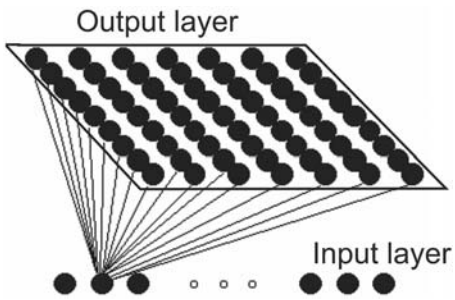

Figure 2. Self-organizing map. Black dots denote nodes of the network.

Table 1. Summary of Catchments and Organic Matter Characteristics[a]

Site	Source	Typical Catchment Land Use	Mean TOC			Hydrophobicity			Microbial Fraction		
			Mean	SD	Class	Mean Peak C Emission	SD	Class	Mean Peak T Intensity	SD	Class
1	river[b]	A 26%, U 21%	3.0	1.2	Lv	426	6.4	I	48.2	20.7	Hv
2	river[b]	P 30%, A 26%	3.0	1.1	Lv	425	6.6	I	22.6	6.1	L
3	river[b]	A 63%, P 24%	5.0	1.9	I	426	6.0	I	44.8	12.3	Hv
4	river[b]	A 38%, P 31%	4.1	1.3	I	430	10.7	I	24.9	4.5	I
5	river[b]	A 63%, P 24%	5.1	2.0	I	425	7	I	44.6	11.5	H
6	river	A 65%, P 25%	6.0	1.3	H	423	3.6	I	44.3	5.6	H
7	river	P 44%, C 19%	3.3	0.7	I	420	5.3	H_{phil}	49.5	17	H
8	river	A 48%, P 24%	7.0	1.1	H	420	5.1	H_{phil}	53.8	9	H
9	river	A 65%, P 25%	6.8	0.6	H	420	4.4	H_{phil}	46.9	5.4	H
10	river	A 43%, P 30%	4.2	0.7	I	422	7.7	I	34.9	3.5	I
11	river	P 55%, C 15%,	4.6	0.9	I	428	8.0	I	34.5	6.7	I
12	river	A 38%, P 31%,	5.2	0.9	I	426	7.0	I	39.3	7.6	I
13	river	A 55%, I 32%,	6.0	1.4	H	424	8.4	I	56.1	13.7	Hv
14	reservoir	P 48%, O 39%	5.6	1.6	Hv	449	7.4	H_{phob}	15.8	3.1	L
15	reservoir	P 45%, A 33%	2.7	0.8	L	441	13.1	H_{phob}	14.6	2.5	L
16	reservoir	P 76%, F 10%	6.7	1.2	H	427	7.5	I	41.7	2.9	H

[a]Typical catchment land use, selected, types, of the largest percentage in total catchment area: A, nonirrigated arable land; P, pastures; C, other cultivated areas; U, urban fabric; I, industrial, transport or commercial units; G, green urban areas; F, forests; O, other areas. Mean TOC: >6.0 mg/l indicates high TOC. Mean TOC < 3.0 mg/l indicates low TOC. Hydrophobicity measured as peak C emission wavelength: >440 nm, hydrophobic (H_{phob}); <420 nm, hydrophilic (H_{phil}). Microbial fraction measured as peak T intensity: >40 au, high microbial; <20 au, low microbial; coefficient of variation measured as a mean coefficient of organic matter properties: >30%, high variability; <10%, low variability. Classification: L, low; I, intermediate; H, high; v, variable.

[b]Direct abstraction from river to WTW.

monthly between August 2006 and February 2008 [*Bieroza et al.*, 2008]. The WTWs are located in the Midlands region, central UK and are owned and operated by Severn Trent Water Ltd. The WTWs treat a range of raw waters, from upland sources exhibiting natural organic matter with high TOC concentrations, to lowland sources reflecting anthropogenically impacted microbial organic matter character (Table 1). Fluorescence results from conventional peak picking approaches, and comparison with organic matter removal in drinking water treatment works, can be found in the work of *Bieroza et al.* [2008].

[10] Laboratory analytical methods have been presented in detail previously [*Bieroza et al.*, 2008]. In summary, organic matter fluorescence of unfiltered samples was measured using a Cary Eclipse Fluorescence Spectrophotometer (Varian, Surrey, UK), by scanning excitation wavelengths from 200 to 400 nm in 5 nm steps, and detecting the emitted fluorescence in 2 nm steps between 280 and 500 nm. Excitation and emission slit widths were set to 5 nm and photomultiplier tube voltage to 725 V. In order to maintain the consistency of measurement conditions, blank scans with a sealed cell containing deionized water and the measurement of the intensity of Raman line of water at 348 nm excitation wavelength, were run systematically following the procedure presented elsewhere [*Baker*, 2001]. The mean Raman value during the study period was 22.3 intensity units, (1 SD = 0.5). All the fluorescence intensities were corrected and calibrated to a Raman peak intensity of 20 units at 396 (392–400) nm emission wavelength.

[11] TOC was measured using a Shimadzu TOC-V-CSH analyzer with auto-sampler TOC-ASI-V. The nonpurgable organic carbon (NPOC) determination method was employed and the result NPOC was calculated as a mean of the three valid measurements. The typical error of the analyses was less than 10% indicating sufficient precision of the TOC measurements.

[12] Examples of typical EEMs for raw and clarified water are presented in Figure 1. From the fluorescence measurements, EEMs of each sample display the intensity of fluorescence against wavelengths at which excited organic matter fluorophores emitted the light. Fluorescence regions can be attributed to both natural fluorescence (humic- and fulvic-like), defined as peaks A and C [*Coble*, 1996] and microbial derived organic matter (tryptophan- and tyrosine-like fluorescence, defined as peaks T and B) at shorter emission wavelengths [*Coble*, 1996; *Stedmon et al.*, 2003]. Peak C (fulvic-like) fluorescence intensity has been shown to exhibit a general correlation with TOC [*Hudson et al.*, 2007; *Cumberland and Baker*, 2007]. In a related study, a strong linear correlation was observed between fulvic-like fluorescence intensity reduction between raw and clarified water and TOC removal measured independently ($R^2 = 0.90$) [*Bieroza et al.*, 2008].

[13] Prior to SOM analyses, the fluorescence data were preprocessed with scripts written in Matlab® 7.7 with the Statistics Toolbox 7.0 and Neural Network Toolbox 6.0.1. All data analyses were carried out on a 512 MB Dual Pentium III PC computer. First, fluorescence spectra were normalized to the Raman scatter peak (at 348 nm excitation wavelength) of deionized water by subtracting the Raman signal from the raw data [*Determann et al.*, 1998; *Stedmon et al.*, 2003]. The Rayleigh and Raman scatter were removed, assuming that the position of the Rayleigh scatter occurs at the excitation equal to the emission wavelength (first-order) or double excitation (second-order) and the position of Raman line is at constant energy shift with respect to the first-order Rayleigh scatter [*Bahram et al.*, 2006]. The fluorescence regions containing redundant information in areas where excitation wavelengths are larger than emission wavelengths and of low signal-to-noise ratio for excitation wavelengths less than 240 nm were removed from further analysis by replacement with NaN (Not A Number) [*Bro*, 1998; *Stedmon et al.*, 2003]. The resultant EEMs ranged from 240 to 400 nm excitation and from 300 to 500 nm emission wavelengths respectively. Therefore, the final data set used in the SOM analysis comprised 625 samples of raw and clarified water and 2515 fluorescence excitation-emission wavelengths.

[14] Finally, fluorescence data scaling (data variance was normalized to one) and mean centering (by subtracting off

variable means) was performed to reduce the concentration effects exhibited by intensity [*Boehme et al.*, 2004].

2.2. Kohonen's Self-Organizing Map

[15] To understand the mechanism of SOM network training and its ability for data reduction, some fundamental concepts of ANNs are summarized below.

[16] ANNs are powerful, nonparametric, parallel computational tools frequently employed for data classification and calibration [*Bos et al.*, 1993; *Despagne and Massart*, 1998; *Basheer and Hajmeer*, 2000] due to their ability to model nonlinearity, and properties such as fault and noise tolerance (ability of processing noisy, uncertain data), self-modeling, self-learning (by example) and generalization capabilities [*Basheer and Hajmeer*, 2000]. ANN functions are facilitated by special structure, consisting of the number of processing units (neurons) arranged in interconnected layers. In a typical ANN, the input layer neurons are responsible for presenting the data to the network, whereas the output layer neurons generate the overall network response to the input data. Neurons in the hidden and output layers are active and perform computational transformations by summing up their input connection weights multiplied by the output of corresponding neurons from the preceding layer and generating the output that is passed to the successive layer. The iterative process of adjusting connection weights between neurons of different layers is called network training.

[17] Kohonen's self-organizing map comprises two fully connected layers (Figure 2).

[18] For the purpose of the SOM analysis, fluorescence three-dimensional EEMs of raw and clarified water for 16 WTWs were deconvoluted to two-dimensional vectors, where each column corresponds with one emission-excitation wavelength pair. The neurons in the input layer of the SOM are connected with each input sample (EEM converted to vector) and have an associated reference vector that contains SOM weights. The reference vector (also called the codebook or weight vector) can be defined as $d_i = [d_{i1}\, d_{i2} \ldots d_{im}]$, where m is equal to the dimension of the input vectors (2515 fluorescence excitation-emission pairs).

[19] The reference vector and the location on the map are the positions of the neuron in input and output space and the initial high-dimensional matrix of input data can be projected with the SOM algorithm on a two-dimensional map comprising output neurons (Figure 2) [*Kohonen*, 1998, 2001].

[20] The SOM training is an iterative process in which for each input sample comprising unfolded EEM, the neuron with reference vector weights most similar to the input vector is first identified (winner or best matching unit, BMU). Thus, for each input EEM presented to the SOM network, the output neuron with the reference vector most similar to the vector representation of EEM is selected. Once the best matching reference vector for each input EEM vector is found, its weights and the weights of its neighboring neurons are modified and moved toward the input vector (self-organization feature of the algorithm) (equation (1)):

$$w_i(k+1) = w_i(k) + \varepsilon(k) h_p(i,k)\{x_j(k) - w_i(k)\} \qquad (1)$$

where $w_i(k)$ is the previous weight of neuron, $w_i(k+1)$ is the new weight of neuron, $\varepsilon(k)$ is the learning rate, $h_p(i, k)$ describes the neighborhood of the winning neuron, k is the number of epochs (a finite set of input patterns presented sequentially) and p is the index of the winning neuron. The learning rate describes the speed of the training process ($0 < \varepsilon(k) < 1$) and decreases monotonically during the training phase. The topological neighborhood can be described as a neighborhood set of array points *Nc* around the given node *c*. During the training of the map the radius of the *Nc* (a size of the neighborhood set) decreases monotonically to enable the global ordering of the map. Thus, the projection of the input data is done in two phases: the main training (large *Nc* radius) and fine adjustment of the map (small *Nc* radius) [*Kohonen*, 2001].

[21] The trained network activates the appropriate output neurons of the network according to the input samples without prior knowledge of the process that produced the data and its distribution. Consequently, the analysis of the networks output, provides the basis for extraction of relationships and regularities from the original data. A complete description of the SOM algorithm can be found in [*Kohonen*, 2001].

[22] Fluorescence data decomposition with SOM was carried out in Matlab® with the use of the SOM toolbox version 2 [*Kohonen*, 1998] providing scripts for algorithm implementation and validation (testing) and various tools for visualization and analysis of the obtained results.

[23] The input matrix, comprising 625 samples and 2515 excitation-emission pairs, was presented to the nodes of the SOM input layer simultaneously (batch mode) and neuron weights were initialized using linear initialization along the two greatest eigenvectors of the input matrix [*Kohonen*, 2001]. The size of the output layer was determined by finding the ratio of the two greatest eigenvalues of the input matrix. The final map contained 120 nodes (size 15 × 8).

3. Results

[24] The SOM training is an unsupervised process of adjusting the connection weights (reference vectors) between nodes in the input and output (map) layers (Figure 2). The training is completed once for each input sample its representation in the form of the reference vector with the weights most similar to the input data (best matching unit) is found. The analysis of the SOM output requires the use of various visualization and clustering tools, providing substantial information on the input data distribution and relationships with the measured variables.

[25] Figure 3 presents some basic SOM visualization methods, including unified distance matrix algorithm (U matrix, Figure 3a), samples distribution on the map with cluster borders determined (Figure 3b) and single hit histograms (Figure 3c). The U matrix [*Ultsch*, 1993] is the most common graphical representation of the SOM structure, in which distances between neighboring map units are calculated and visualized using gray or color scale on the trained map [*Park et al.*, 2003]. Compared with the original map size (15 by 8 neurons), the U matrix comprises additional map units to visualize the distances between neurons. High values on the U matrix (light areas) indicate large distances between neighboring units and hence can be helpful in determining the cluster borders as clusters typically form uniform areas of low values (dark areas). From Figure 3a it can be observed that the cluster structure of the fluorescence data is not well defined as

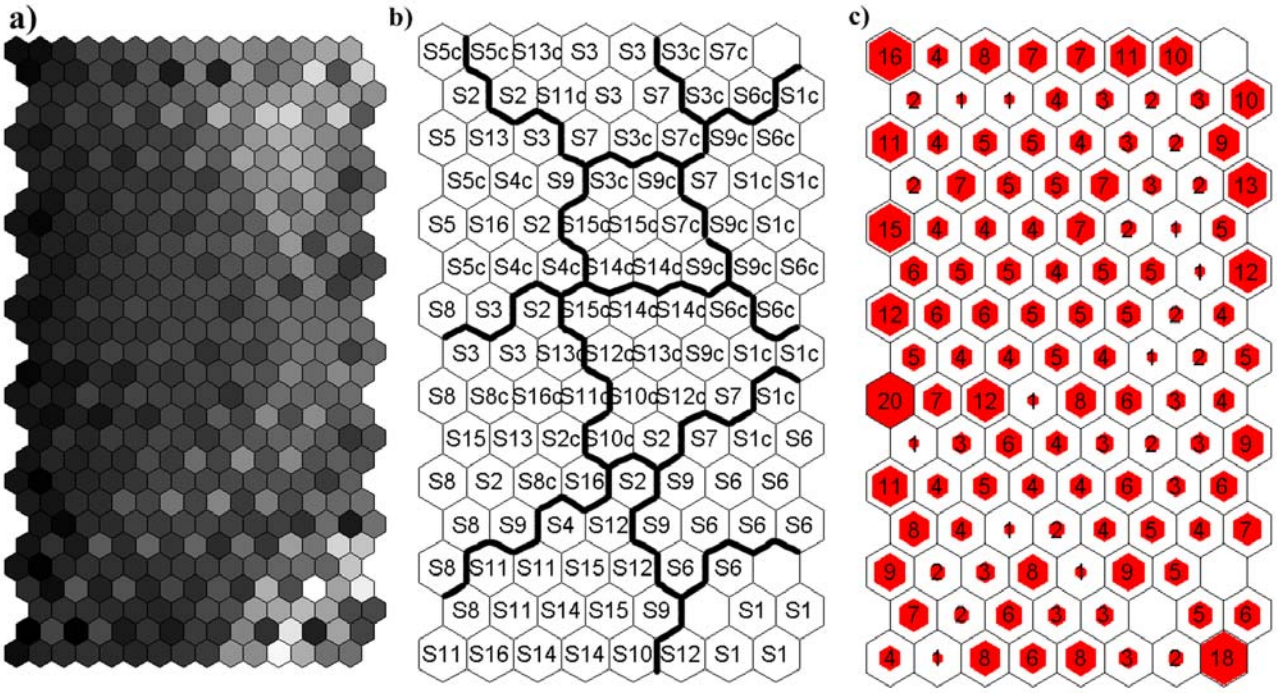

Figure 3. Visualization of the SOM map for fluorescence data: (a) U matrix, (b) sample distribution and clusters, and (c) hit histogram. Notation used, e.g., S1 and S1c, denotes raw and clarified water of site 1.

the light and dark areas on the U matrix are not easily partitioned. However, the presence of a few clusters can be discerned (i.e., in the lower right-hand side corner of the map). As the U matrix provides the information on the cluster structure, the number of valid clusters can be derived from the k-means algorithm [*Jain and Dubes*, 1988]. The k-means clustering algorithm is used for minimizing the sum of squared Euclidean distances between the input (unfolded EEMs) and the SOM reference vectors. The best clustering minimizes the sum of the squared distances (and also the Davies-Bouldin index) between each input data vector and its nearest cluster center [*Davies and Bouldin*, 1979]. Here the optimum number of ten clusters was determined by running the k-means algorithm multiple times for different number of clusters and selecting the solution with the lowest sum of the squared distances (Figure 3b). To correlate the cluster pattern on the map with samples distribution, for each map node the most frequent best matching unit with assigned site number was found. Each cluster contains sites with similar organic matter properties measured with excitation-emission spectra (Figure 3b). Distinctively unique spectral properties of raw water can be discerned for sites 1 and 6, which are both reservoir abstractions (raw water is stored in reservoirs prior to uptake for treatment), whereas the distribution of other sites is more complex. A good discrimination between raw and clarified water fluorescence properties occurs for sites located at the bottom of the map, sites 1, 6, 10, 11, 14. The opposite is observed for sites clustered at the top of the SOM map (2, 3, 5), with the raw and clarified water of similar properties as indicated by the short distances on the map. Finally, the distribution of samples on the SOM map can be portrayed with hit histograms (Figure 3c). For each neuron the hit characteristic is calculated on the basis of the map response to the input data. The size of the marker indicates how many times each map unit was the BMU for the data set. It can be seen that data is uniformly distributed over the map, with a number of neurons located at the edges of the map being the most frequent BMUs (neuron 1–16 hits, neuron 5–15 hits, neuron 9–20 hits, neuron 120–18 hits).

[26] While the U matrix, cluster structure and hit histograms reveal a pattern of samples distribution on the map, the reference vectors of selected neurons and component planes exhibit the significance of particular fluorescence variables. From the hit histogram in Figure 3c it was inferred that some neurons represent the greatest number of fluorescence samples. Thus the spectral properties derived from the reference vectors of those neurons can provide the important information on the dominant fluorescence features of the data set. The EEMs of two neurons located at the left-hand side of the SOM map (Figures 4a and 4c), indicate the predominance of the humic-like fluorescence at the lower, UV excitation wavelengths (250–300 nm). Additionally, a shift toward higher emission wavelengths (400–450 nm) can be observed for neuron 9 (Figure 4c). Samples projected onto the upper SOM neurons demonstrate a substantial contribution of protein-like fluorescence as the tryptophan-like center located at excitation-emission wavelengths of 280/350 nm can be discerned for both neuron 1 and 109 (Figures 4a and 4b). The distinctively different fluorescence properties can be observed in the excitation-emission spectra of neuron 120 with the humic-like fluorescence peak shifted toward higher excitation and emission wavelengths (Figure 4d). These results are in accordance with the analysis of the response of each fluorescence variable (component) in the form of component planes (Figure 5). The component planes depict the values of the reference vectors for different fluorescence variables and allow the correlation between the samples distribution and particular excitation-emission wavelengths and

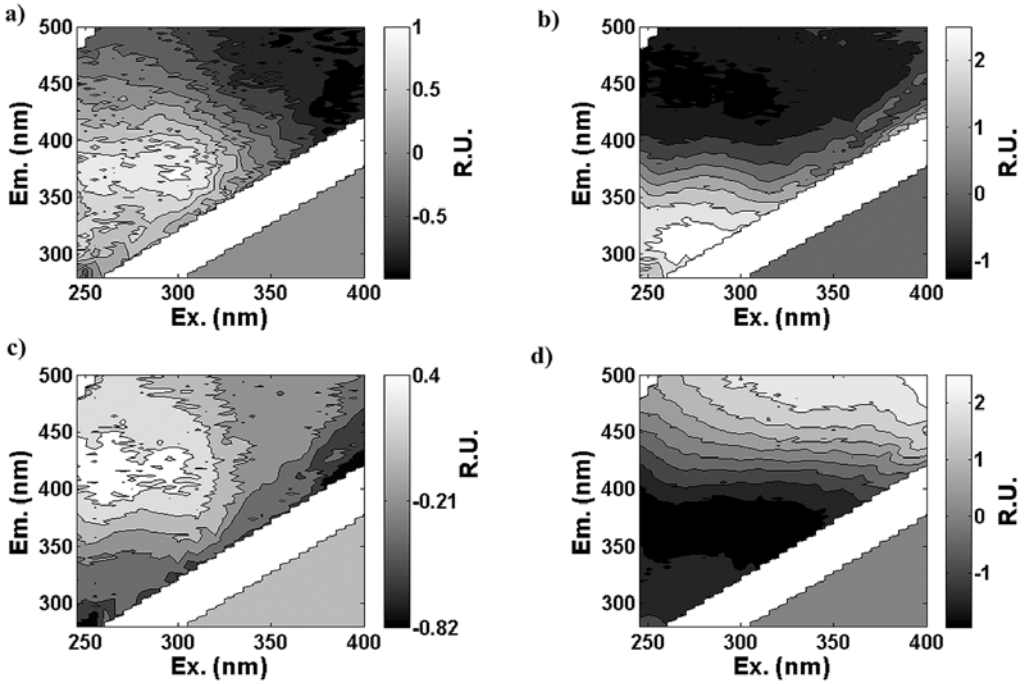

Figure 4. Reference vector plots for selected SOM neurons with the highest number of hits. (a) Neuron 1, (b) neuron 109, (c) neuron 9, and (d) neuron 120.

hence main fluorophores. Thus, for each excitation-emission wavelength pair, a corresponding component plane can be obtained that enables correlation between sample location on the map and fluorescence properties. In Figure 5 the component planes at excitation wavelength of 280 nm were shown for different emission wavelengths (300, 350, 400, 450 and 500 nm). The high values in the component plane denote a higher fluorescence intensity. As the emission wavelength increases, the center of the highest values moves from the top to the bottom of the map with the relative increase in fluorescence intensities (compare maximum values for each component plane). For example, in Figure 5b the component plane for excitation wavelength of 280 nm and emission wavelength of 350 nm is presented, the spectral area related to tryptophan-like fluorescence. The higher values on this component plane indicate the predominance of highly microbial organic matter for sites located in the upper part of the map.

[27] In Figure 6 the hit histograms for sites of different organic matter properties and efficiency of organic matter removal are shown. The geometric distance between raw and clarified water samples correlates with the organic matter removal, with the higher removal for more distant raw and clarified samples. Thus sites 1 and 8 tend to have better organic matter removal than sites 5 and 3. The greater the spread of water samples of particular type on the map, the greater the variation in spectral properties can be discerned. Site 1 represents uniform raw water properties, whereas a greater variation is typical for sites 5, 3 and 8. It can be concluded that site 5 (relatively hydrophilic raw water organic matter with lower emission wavelength) has poorer removal compared to site 1, where the hydrophobic character of the organic matter enhances the efficiency of the treatment process. As stated above, the location in the upper part of the SOM (lower emission wavelengths) corresponds to the increased inputs of the microbial fraction. Thus, distinctive spectral properties can be attributed to sites 3, 5 and 7, where microbial fraction related to tryptophan-like fluorescence has a significant contribution in the raw water fluorescence

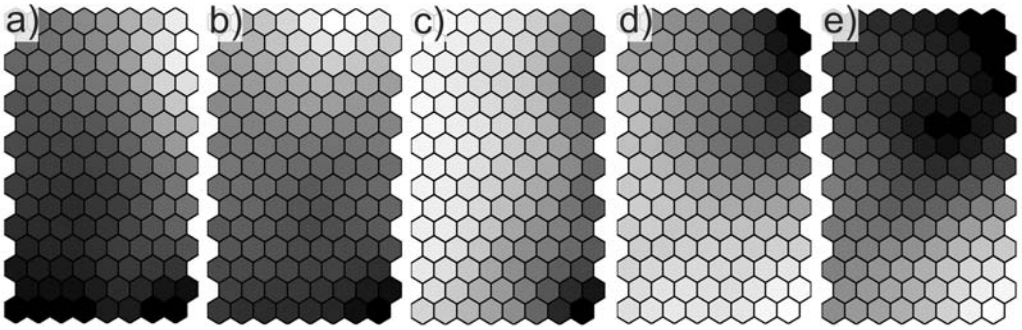

Figure 5. Component planes for the excitation-emission wavelengths with fixed excitation at 280 nm. Emission wavelength of (a) 300, (b) 350, (c) 400, (d) 450, and (e) 500 nm.

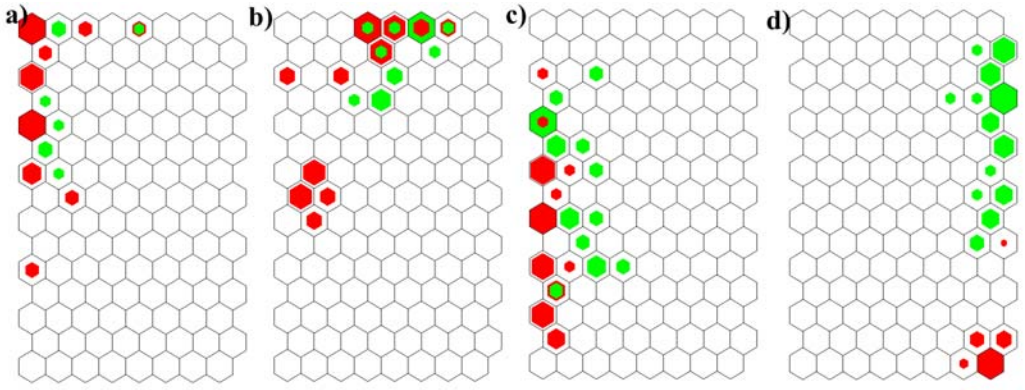

Figure 6. Hit histograms for raw (red) and clarified (green) water. (a) Site 5, (b) site 3, (c) site 8, and (d) site 1.

signature. The presence of microbial organic matter indicates that those sites could be more prone to algal outbreaks which can further deteriorate drinking water quality.

4. Discussion

[28] Fluorescence excitation-emission spectra contain substantial amount of information on organic matter characteristics, however special computational and statistical techniques are required to preprocess the data, remove noise and redundant features like Raman and Rayleigh scatter and to distinguish patterns of interest from a matrix background. To identify significant features or patterns from multivariate, high-dimensional fluorescence space, a dimensionality reduction, data projection and feature extraction methods should be employed.

[29] In the study the unsupervised, nonparametric algorithm of cluster analysis, visualizing and projecting multivariate data was applied, the self-organizing map (SOM). The approach is commonly used in exploratory data analysis. However, only few examples of application of SOM to fluorescence EEM spectra deconvolution could be found in fluorescence-related literature [*Lee et al.*, 2005; *Rhee et al.*, 2005]. Here, the SOM results provide a good discrimination between water treatment sites on the basis of the spectral properties of raw and partially treated water. The evaluation of the SOM properties regarding sample distribution and the importance of fluorescence variables provides significant information on the relationships between raw and clarified water organic matter composition. The distribution of the samples on the map corresponds with the excitation-emission properties (Figure 5). Horizontal and vertical axes correspond to fluorescence emission and excitation wavelengths, with increasing values from the top to the bottom and from the left to the right respectively. Moreover, the diagonal that joins the upper left with lower right corner of the map is the line of the greatest changes in variance within the data set and discriminates the sites of radically different organic matter spectral properties. The reservoir sites 1 and 6, demonstrate the most uniform fluorescence properties of raw water, whereas sites with abstraction from the rivers exhibit more profound changes in organic matter inputs and thus fluorescence signals (site 3, 5, 8, 9, 10, 13) (Figure 3b). The organic matter removal efficiency derived from the fluorescence intensity decrease between raw and clarified water stages is of the primary significance for the formation of disinfection by-products (DBPs). The distance between nodes of SOM map indicates the similarity of the fluorescence samples and thus can demonstrate the relative changes in organic matter character and quantity between raw and clarified water. It was found that the distances on the SOM map between raw and corresponding clarified water samples correlate with the efficiency of organic matter removal measured as a decrease in fulvic-like fluorescence intensity. The larger distance between raw and clarified water samples on the map can be correlated with higher organic matter removal efficiency, e.g., sites 1, 6, 12, 14 and 15, whereas the higher degree of samples clustering can be explained with poorer decrease in organic matter quantity (sites 5, 8, 10, 16) (Figure 6). It was shown that the organic matter removal is highly dependent on the raw water fluorescence properties, with higher efficiencies for higher emission wavelengths in visible and UV humic-like fluorescence centers. The shift toward higher emission wavelengths is indicative of the increased content of more hydrophobic organic matter fraction which is easier to remove during the treatment process. Hence, the exploratory analysis of the fluorescence data with SOM provides a substantial amount of information pertinent to drinking water organic matter properties and removal.

5. Conclusions

[30] This paper has introduced a robust unsupervised algorithm of the SOM applied to fluorescence data analysis. The technique was employed for the characterization of fluorescence excitation-emission spectra of organic matter in surface waters abstracted at 16 surface WTW in the Midlands region of the UK. Although fluorescence data contain a substantial amount of information on the organic matter properties, the high dimensionality of the data generates difficulties in recognition of meaningful relationships between sample distribution and spectral properties of organic matter. In the paper a novel approach to the analysis and interpretation of fluorescence data with SOM was presented. The SOM facilitated pattern recognition of the fluorescence data and revealed linkages between samples distribution and the importance of the particular spectral properties. With reference to the fluorescence differences between raw and partially treated water, the SOM enabled correlation of the organic matter removal efficiency with the organic matter

properties derived from the fluorescence excitation-emission spectra. These results demonstrate that SOM can be a powerful decomposition tool for fluorescence data analysis and, with the use of available toolboxes the implementation and interpretation process can be as straightforward as more common statistical methods.

[31] **Acknowledgments.** The authors are grateful for the financial and logistical support provided by Severn Trent Water Ltd and the University of Birmingham. The authors also acknowledge the laboratory support provided by Ian Boomer and Andy Moss.

References

Bahram, M., R. Bro, C. Stedmon, and S. Afkhami (2006), Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation, *J. Chemometr.*, *20*(3–4), 99–105, doi:10.1002/cem.978.

Baker, A. (2001), Fluorescence excitation-emission matrix characterization of some sewage-impacted rivers, *Environ. Sci. Technol.*, *35*(5), 948–953, doi:10.1021/es000177t.

Basheer, I. A., and M. Hajmeer (2000), Artificial neural networks: Fundamentals, computing, design, and application, *J. Microbiol. Methods*, *43*(1), 3–31, doi:10.1016/S0167-7012(00)00201-3.

Bieroza, M., A. Baker, and J. Bridgeman (2008), Relating freshwater organic matter fluorescence to organic carbon removal efficiency in drinking water treatment, *Sci. Total Environ.*, *407*, 1765–1774, doi:10.1016/j.scitotenv.2008.11.013.

Boehme, J., P. Coble, R. Conmy, and A. Stovall-Leonard (2004), Examining CDOM fluorescence variability using principal component analysis: Seasonal and regional modelling of three-dimensional fluorescence in the Gulf of Mexico, *Mar. Chem.*, *89*(1–4), 3–14, doi:10.1016/j.marchem.2004.03.019.

Bos, M., A. Bos, and W. E. van der Linden (1993), Data processing by neural networks in quantitative chemical analysis, *Analyst*, *118*(4), 323–328, doi:10.1039/an9931800323.

Bro, R. (1998), Multi-way analysis in the food industry: Models, algorithms, and applications, Ph.D. dissertation, Dep. of Dairy and Food Sci., R. Veterinary and Agric. Univ., Copenhagen, Denmark.

Brunsdon, C., and A. Baker (2002), Principal filter analysis for luminescence excitation-emission data, *Geophys. Res. Lett.*, *29*(24), 2156, doi:10.1029/2002GL015977.

Coble, P. G. (1996), Characterization of marine and terrestrial DOM in seawater using excitation-emission spectroscopy, *Mar. Chem.*, *51*(4), 325–346, doi:10.1016/0304-4203(95)00062-3.

Cumberland, S. A., and A. Baker (2007), The freshwater dissolved organic matter fluorescence-total organic carbon relationship, *Hydrol. Process.*, *21*(16), 2093–2099, doi:10.1002/hyp.6371.

Davies, D. L., and D. W. Bouldin (1979), Cluster separation measure, *IEEE Trans. Pattern Anal. Mach. Intell.*, *1*(2), 224–227, doi:10.1109/TPAMI.1979.4766909.

Despagne, F., and D. L. Massart (1998), Neural networks in multivariate calibration, *Analyst*, *123*(11), 157–178, doi:10.1039/a805562i.

Determann, S., J. M. Lobbes, R. Reuter, and J. Rullkötter (1998), Ultraviolet fluorescence excitation and emission spectroscopy of marine algae and bacteria, *Mar. Chem.*, *62*(1–2), 137–156, doi:10.1016/S0304-4203(98)00026-7.

Garcia, J. S., G. A. da Silva, M. A. Arruda, and R. J. Poppi (2007), Application of Kohonen neural network to exploratory analyses of synchroton radiation x-ray fluorescence measurements of sunflower metalloproteins, *XRay Spectrom.*, *36*(2), 122–129, doi:10.1002/xrs.950.

Hudson, N. J., A. Baker, and D. Reynolds (2007), Fluorescence analysis of dissolved organic matter in natural, waste and polluted waters—A review, *River Res. Appl.*, *23*(6), 631–649, doi:10.1002/rra.1005.

Jain, A., and R. Dubes (1988), *Algorithms for Clustering Data*, Prentice-Hall, Upper Saddle River, N. J.

Kohonen, T. (1998), The self-organizing map, *Neurocomputing*, *21*(1), 1–6, doi:10.1016/S0925-2312(98)00030-7.

Kohonen, T. (2001), *Self-Organizing Maps*, 3rd ed., Springer, Berlin.

Lee, K. I., Y. S. Yim, S. W. Chung, J. Wei, and J. I. Rhee (2005), Application of artificial neural networks to the analysis of two-dimensional fluorescence spectra in recombinant E coli fermentation processes, *J. Chem. Technol. Biotechnol.*, *80*(9), 1036–1045, doi:10.1002/jctb.1281.

McKnight, D. M., et al. (2001), Spectrofluorometric characterization of dissolved organic matter for indication of precursor organic material and aromaticity, *Limnol. Oceanogr.*, *46*(1), 38–48.

Mopper, K., and C. A. Schultz (1993), Fluorescence as a possible tool for studying the nature and water column distribution of DOC components, *Mar. Chem.*, *41*(1–3), 229–238, doi:10.1016/0304-4203(93)90124-7.

Park, Y. S., R. Céréghino, A. Compin, and S. Lek (2003), Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters, *Ecol. Modell.*, *160*(3), 265–280, doi:10.1016/S0304-3800(02)00258-2.

Persson, T., and M. Wedborg (2001), Multivariate evaluation of the fluorescence of aquatic organic matter, *Anal. Chim. Acta*, *434*, 179–192, doi:10.1016/S0003-2670(01)00812-1.

Rhee, J. I., K. I. Lee, C. K. Kim, Y. S. Yim, S. W. Chung, J. Wei, and K. H. Bellgardt (2005), Classification of two-dimensional fluorescence spectra using self-organizing maps, *Biochem. Eng. J.*, *22*(2), 135–144, doi:10.1016/j.bej.2004.09.008.

Spencer, R. G. M., et al. (2007), Discriminatory classification of natural and anthropogenic waters in two UK estuaries, *Sci. Total Environ.*, *373*, 305–323, doi:10.1016/j.scitotenv.2006.10.052.

Stedmon, C. S., S. Markager, and R. Bro (2003), Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy, *Mar. Chem.*, *82*(3–4), 239–254, doi:10.1016/S0304-4203(03)00072-0.

Ultsch, A. (1993), Self-organizing neural networks for visualization and classification, in *Information and Classification*, edited by O. Opitz, B. Lausen, and R. Klar, pp. 307–313, Springer, Berlin.

A. Baker, School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham B15 2TT, UK. (a.baker.2@bham.ac.uk)

M. Bieroza and J. Bridgeman, School of Civil Engineering, University of Birmingham, Birmingham B15 2TT, UK. (mzb605@bham.ac.uk; j.bridgeman@bham.ac.uk)