Biogeosciences, 10, 5875–5887, 2013
www.biogeosciences.net/10/5875/2013/
doi:10.5194/bg-10-5875-2013

Biogeosciences Open Access

Technical Note: Dissolved organic matter fluorescence – a finite mixture approach to deconvolve excitation-emission matrices

A. Butturini and E. Ejarque

Departament d'Ecologia, Facultat de Biologia, Universitat de Barcelona (UB), Avinguda Diagonal 643, 08028 Barcelona, Spain

Correspondence to: A. Butturini (abutturini@ub.edu)

Received: 28 February 2013 – Published in Biogeosciences Discuss.: 8 March 2013
Revised: 17 July 2013 – Accepted: 29 July 2013 – Published: 6 September 2013

Abstract. The analysis of the shape of excitation-emission matrices (EEMs) is a relevant tool for exploring the origin, transport and fate of dissolved organic matter (DOM) in aquatic ecosystems. Within this context, the decomposition of EEMs is acquiring a notable relevance. A simple mathematical algorithm that automatically deconvolves individual EEMs is described, creating new possibilities for the comparison of DOM fluorescence properties and EEMs that are very different from each other. A mixture model approach is adopted to decompose complex surfaces into sub-peaks. The laplacian operator and the Nelder-Mead optimisation algorithm are implemented to individuate and automatically locate potential peaks in the EEM landscape. The EEMs of a simple artificial mixture of fluorophores and DOM samples collected in a Mediterranean river are used to describe the model application and to illustrate a strategy that optimises the search for the optimal output.

1 Introduction

Since the pioneering works of Traganza (1969) and Coble et al. (1990) the analysis of fluorescence properties of dissolved organic matter (DOM) in aquatic ecosystems has become an essential technique in the exploration of qualitative changes and fate of dissolved organic carbon in aquatic ecosystems (Hudson et al., 2007; Fellman et al., 2010; Ishii and Boyer, 2012).

This analytical technique has benefited considerably from instrumentation advances that facilitate the rapid analysis of large amount of samples in a short period of time and allow the storage of large datasets. This has motivated the generation of excitation-emission matrices (EEMs). EEMs are contour plots in which fluorescence intensities are plotted as a function of excitation (typically 230–450 nm) and emission (typically 300–600 nm) wavelengths. Excitation wavelengths represent the wavelength delivered to the aqueous sample, thus inducing fluorescence, while emission wavelengths represent the wavelength of the resulting fluorescence. Form of a EEM responds to a complex mixture of fluorescent compounds (fluorophores). The main challenge consists in individuating the location and relevance of fluorescence events that compose the fluorescence spectra. To date, decomposition of fluorescence spectra is performed with advanced supervised (PCA, N-PLS, PARAFAC) or unsupervised (self-organising map) statistical multivariate techniques (Bieroza et al., 2009). Those algorithms strongly enhanced the study of DOM. However, their use is a matter of debate (Fellmann et al., 2010) and deep analysis (Bieroza et al., 2009). Multivariate tools are usually executed with datasets that include a large number of EEMs (typically more than 100) and robust results are more easily obtained when a dataset integrates samples that follow gradual gradients (Stedmon and Bro, 2008). Conversely, to our knowledge, an algorithm that decomposes the signal of individual EEMs is currently unavailable. Such an algorithm would allow researchers to evaluate DOM quality using a reduced number of EEMs, as well as the freedom to compare and evaluate EEMs that do not necessarily follow any gradient.

In this note, we introduce an alternative approach that integrates a simple surface analysis with the finite distribution mixture (FDM) modelling. Similarly to the tools mentioned previously, FDM is widely used for data mining and pattern recognition. It assumes that a single complex surface

(an EEM for example) can be deconvolved into n subjacent peaks (Frjguhwirth-Schnatter, 2006). In consonance with multiway techniques, FDM assumes that peaks behave independently, without interference between them. The basic difference with respect to the multiway techniques lies in the assumption that in FDM, peaks fit a predefined probabilistic density function. A peak is simply a mathematical unit that isolates a fluorescence event. This unit is not necessarily a synonym of "fluorophore". EEMs from pure fluorescent substances frequently show single or multiple peaks that can be roughly approximated to a Gaussian bell (for example, see Boehme and Coble, 2000; Yamashita, and Tanoue, 2003; Hudson et al., 2007). On the other hand, due to their chemical intrinsic complexity, fluorescence events from natural DOM samples, can not be attributed solely to specific fluorophores (Del Vecchio and Blough, 2004; Chen and Kenny, 2007). However, the Gaussian shapes frequently emerge when we observe fluorescence signal in natural samples. Boehme and Coble (2000), reported the detection of a small pool of fluorophores with dual peaks with "circular contours". EEM from algae extract show two–three clear peaks that call in mind a Gaussian bell (Her et al., 2003). Therefore, it is not surprising that researchers attempt to fit fluorescence signals with one-dimensional Gaussian distribution (Korshin et al., 1999; Westerhoff et al., 2001)

These preliminary considerations are at the heart of the idea to adopt the FDM to decompose the fluorescence events in individual EEM. In this note, besides the model description, a strategy to optimise the selection of an optimal model for an EEM is reported. The model is initially applied to a simple EEM generated with two well-known fluorescent substances (quinine sulphate and tryptophan) that lead the reader through the different methodological steps. Successively, it is applied to a heterogeneous dataset that includes 21 EEMs collected along the main stem of an impacted Mediterranean river. Implementation of FDM is executed with Wolfran Mathematica® program (version 8 was used in this study). A didactical example of the model and its implementation with this software is available at the following link http://hdl.handle.net/2445/33820. However, FDM can be computed with any other mathematical software.

2 Model description

Within the FDM context, an EEM is a bivariate matrix ($f_{(xy)}$) that can be described as the sum of n distributions ($c_{(x,y)}$, Eq. 1). In FDM research, Gaussian distributions are the most used probability models (Frjguhwirth-Schnatter, 2006). However, in our FDM the Gaussian distribution incorporate an asymmetric parameter to capture peaks with eventual asymmetries and/or long tails (Kato et al., 2002).

The parameters that describe each distribution are: their mean $\mu_i(\mu_{ix}, \mu_{iy})$, deviation $\sigma_i, (\sigma_{ix}, \sigma_{iy})$, height a_i# (a_{ix}, a_{iy}) and skewness $r_i(r_{ix}, r_{iy})$:

$$z_{(x,y)} = \sum_{i=1}^{n} c_{(x,y)_i} \quad (1)$$

$$c_{(x,y)_i} = a_i e^{-\left\{ \begin{array}{l} \frac{(\mu-\mu_i)^2}{2\sigma_i^2} \text{ if } \mu > \mu_i \\ \frac{(\mu-\mu_i)^2}{2r_i^2\sigma_i^2} \text{ otherwise} \end{array} \right.} \quad (2)$$

where $z_{(x,y)}$ is the sum of n peaks $c_{(x,y)}$ that fit a bivariate asymmetric Gaussian distribution model. If the asymmetric parameter, r, is equal to the unity, Eq. (2) is equivalent to a classic Gaussian distribution.

In Eqs. (1) and (2), estimates of the unknown parameters ($\mu_i, \sigma_{i,}, a_i$ and r_i) are performed following two main steps:

Step A: Surface analysis to detect and locate the potential peaks in $f_{(xy)}$, $L_n = \{\mu_{1,} \mu_{2,} \mu_{3,\ldots,} \mu_n\}$;

Step B: Optimal model selection criteria and estimate of the parameters a_i ,σ_i and r_i.

To avoid chemically meaningless results, the only requirement is that all selected peaks must have a positive height ($a_i > 0$). Steps A and B are detailed below.

2.1 Step A: Detection and location of candidate peaks.

It consists of an analysis of the surface of $f_{(xy)}$ to detect the position of potential peaks in a EEM. This step combines two search strategies:

a. Detection of global and local maxima in the $f_{(xy)}$:

$$\text{Max} f_{(xy)} = \{\mu_a, \mu_b, \mu_c, \ldots, \mu_n\} \quad (3)$$

b. Detection of local minima (μ'_i) of the differential Laplacian operator of $f_{(xy)}$ ($\nabla^2 f$):

$$\text{Min}\nabla^2 f = \{\mu'_a, \mu'_b, \mu'_c \ldots \mu'_n\} \quad (4)$$

$\nabla^2 f$ describes the sum of the second derivative of $f_{(xy)}$ with respect to x and y (Ganza and Vorozhtsov, 1996). It is used to detect shoulders and edges in complex surfaces. In chemometrics, the local minimum of second derivative is used to identify the position of non-evident peaks in complex chromatograms (Stevenson et al., 2010). Here, we extend this idea to two dimensions.

The search for maxima in $f_{(xy)}$ and minima in $\nabla^2 f$ is performed with the Nelder-Mead optimisation algorithm under constrained conditions (Horst and Pardalos, 1995). The sensitivity of this algorithm can be increased or reduced by modifying selected parameters (namely: the contraction ratio, the expansion ratio, the reflection ratio and the shrink ratio). In our application we used the standard values for these parameters (0.5, 2, 1, and 0.5 respectively, Nelder and Mead, 1965) as they guaranteed an exhaustive search of main local minima in $\nabla^2 f$ into a relatively short computational time. The $\nabla^2 f$ operator is sensible to edges. Therefore, minimum

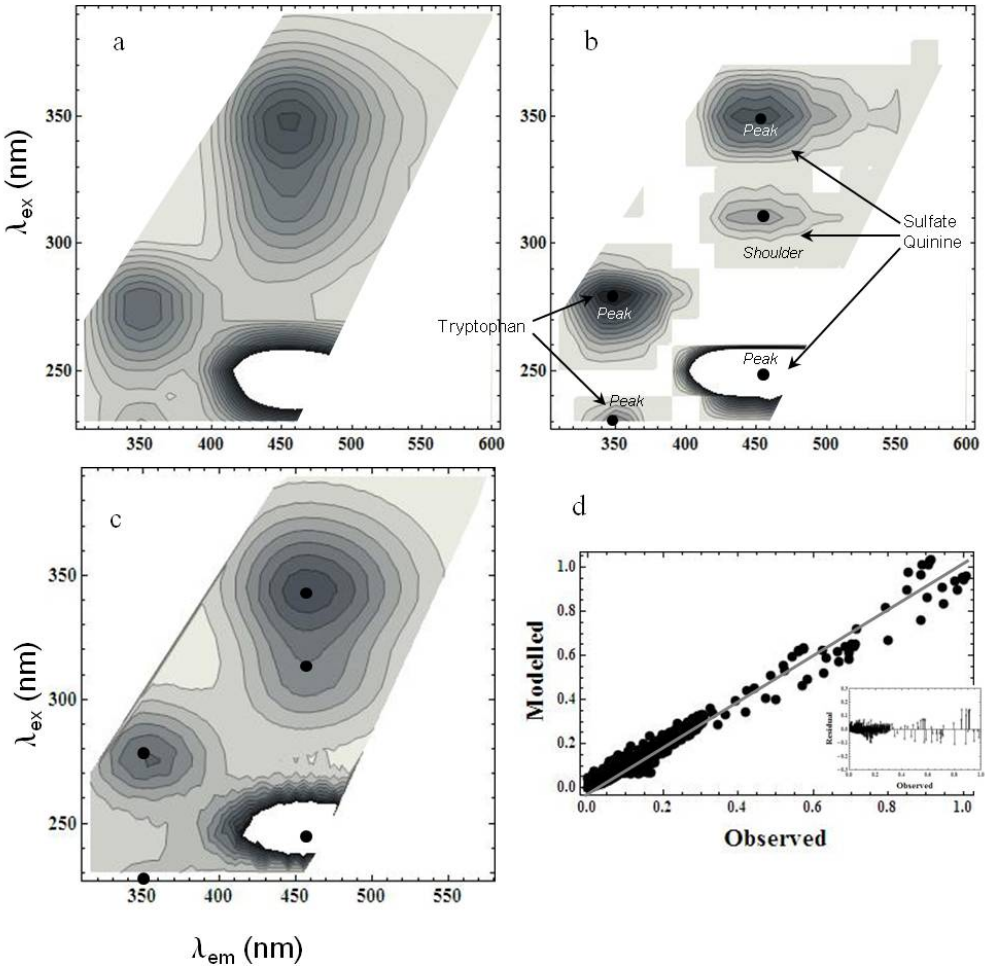

Fig. 1. (a) Contour plot for tryptophan quinine sulfate (TQS) EEM; **(b)** its laplacian $\nabla^2 f$; **(c)** the modelled EEM; **(d)** Comparison of the modeled vs. observed TQS EEM (values in Raman units; $r^2 = 0.985$, slope of the fit is 0.97 ± 0.0026). Dots represent each individual fluorescence. The gray line shows the 1 : 1 line. The inset shows the residuals with respect to to the magnitude of the fluorescence signal.

in $\nabla^2 f$ surface found in the proximity of the Raman and Rayleigh-Tyndall scattering are omitted.

Once Max $f_{(xy)}$ and Min $\nabla^2 f$ are obtained, results are joined to sort all distinct coordinates that appear in the two lists:

$$L_n = \mathrm{Max} f_{(xy)} U \mathrm{Min} \nabla^2 f \quad (5)$$

where L_n is the list of the potential n peaks in $f_{(x,y)}$. In complex surfaces the Nelder-Mead algorithm can be easily trapped in local minima (or maxima) that are very close to each other and, presumably, are identifying the same peak. From a statistical perspective it is assumed that these neighbour peaks fall into the same cluster. In this case, it is necessary to merge them into a single coordinate. The search for clusters is performed according to the *fixed radius near neighbour* approach (Bentley et al., 1977): at each detected coordinate (μ_i), a circular influence area (IA_i) of radius R is associated ($\mathrm{IA}_i = \pi R^2$), centred at the point μ_i. The value of the radius R is the same for all detected μ_i and is fixed to set the IA value to 10 % of the planar area of the surface matrix. Those coordinates (different from μ_i) that fall within the area IA_i of μ_i are automatically grouped into a same cluster. Two criteria are established to assign a coordinate to each cluster:

Criteria # 1 (applicable for Eqs. 3 and 5): the coordinate with the highest maxima is selected, the rest are discarded.

Criteria # 2 (applicable for Eq. 4): the coordinate with the lowest $\nabla^2 f$ is selected, the rest are discarded.

Figure 1 provides a visual example of methodological steps explained previously. In this example, the EEM of a mixture of two fluorescence substances is used (referred to as the TQS sample). The substances are: tryptophan (dissolved in deionised water) and quinine sulfate (dissolved initially in 50 mM H_2SO_4). Tryptophan is an amino acid with two fluorescence peaks at emission of $\sim$ 350 nm. Quinine sulfate shows two clear maxima at emission of $\sim$ 450 nm and a characteristic shoulder between these two peaks (Fig. 1a).

In this EEM the identification of the local maxima of quinine sulphate and tryptophan is straightforward (Fig. 1a).

In parallel, the contour plot of the laplacian operator ($\nabla^2 f$, Fig. 1b) evidences the presence of five local minima: two of them ($\mu_1 = \{277, 350\}$ and $\mu_2 = \{230, 350\}$) are attributable to the two local maxima of tryptophan; other two coincide with the two local maxima of sulphate quinine ($\mu_3 = \{245, 450\}$ and $\mu_4 = \{345, 450\}$); a fifth local minima is at $\mu'_5 = \{305, 450\}$ and it locates the position of an additional subtle peak between the two quinine sulphate maxima. Therefore, according to Eq. (5), L_n is a list of five coordinates, μ_i, ($L_3 = \{\mu_1, \mu_2, \mu_{3,}\mu_{4,}\mu'_5\} = \{\{230, 350\}, \{277, 350\}, \{245, 450\}, \{345, 450\}, \{305, 450\}\}$).

Figure 1c shows the TQS modelled spectra obtained assuming that peaks fit the asymmetric Gaussian distribution. In this example $r^2 = 0.985$. Figure 1d shows the relationship between the observed and modelled fluorescence measured at each $\lambda_{ex}/\lambda_{em}$ pairs. Points are located around the 1 : 1 line (the slope of the fit is 0.97 ± 0.0026) and magnitude of residuals do not show a trend with respect to the magnitude of the fluorescence signal (see the inset in the bottom right corner).

To test if the introduction of the asymmetric parameter, r_i (Eq. 2), into the Gaussian distribution helps to improve the model fit we modelled the TQS EEM assuming that peaks fit the classic Gaussian distribution as suggested by Westerhoff et al. (2001). Therefore, in Eq. (2) we forced the condition r_i ={1,1}. The fit is still reasonably good, however, the r^2 decreased to 0.96 (the slope of the fit is 0.945 ± 0.0076) and residuals are larger than those estimated previously (Supplement 1). Figure 2 compares in more detail the outputs obtained with the two distribution approaches along the emission fluorescence spectra ($\lambda_{ex} = 350$ nm). The plot reveals the failure of the classical Gaussian distribution to fit reasonably well the spectra at $\lambda_{em} < 450$ nm and highlights that the introduction of an asymmetry factor in the Gaussian distribution improve notably the model goodness.

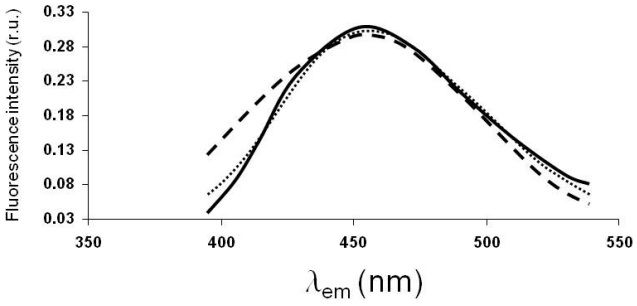

Fig. 2. Emission fluorescence spectra ($\lambda_{ex} = 350$ nm) for the TQS sample (solid line) and model output with the asymmetric Gaussian distribution (dotted line) and the classic Gaussian distribution (dashed line).

2.2 Step B: Optimal model selection criteria

The mixture of tryptophan and quinine sulfate produces a simple EEM with clear and unambiguous peaks. Therefore, the identification of the number and coordinates of peaks to implement into the model is a simple task. However, this task can be much more difficult when EEMs from natural samples are analysed.

In a complex EEM, in which we ignore the number peaks, Step A identifies a set of n potential peaks and their coordinates (L_n). However, the possibility to overestimate the number peaks exists. To individuate the optimal number of peaks in more complex EEMs and reduce the risk of overestimating the model parameters (i.e., the number of peaks) we adopt the Bayesian Information Criterion (BIC) descriptor:

$$BIC_i = -2\ln(ML_i) + k_i \ln(O) \quad (6)$$

where ML_i is the maximized likelihood of the model associated to the subset i; k_i number of input parameters (i.e., number of element in the subset i); O is the sample size. The model with the smallest BIC value is selected as the optimal model (Schwarz, 1978).

The search of the model with the lowest BIC value is performed according the following the procedure. First all, we extract all i distinct proper subsets of L_n:

$$P(L_n) = \{\{\mu_1\}_1, \{\mu_2\}_2, \{\mu_3\}_3, \{\mu_1, \mu_2\}_4 \ldots\ldots \{\mu_1, \mu_2, \mu_3, \ldots, \mu_n\}_i\} \quad (7)$$

where $i = 2^n - 1$.

Successively, FDM (Eqs. 1 and 2) is run for each i subset. Finally, among all possible i subsets, the optimal model is that with the lowest BIC. The optimal model is considered valid if it explains more than the 99 % of the measured variance (i.e., $r^2 > 0.99$). This threshold is similar to that reported by Stedmon and Bro (2008) to individuate "a reasonable fit for EEM data" with parallel factor analysis (PARAFAC), probably the most commonly used in EEMs deconvolution.

Figure 3 describes this process for a real EEM from the dataset. In this example, the step A identifies nine potential peaks (n =9) that generate $2^9 - 1 = 511$ subsets. Each subset has one model output. Out of them 53 (~ 10 %) generated "good" model outputs ($r^2 > 0.99$, the black dots in Fig. 3a). The number of peaks of these candidate models ranged between five and nine. Within this reduced pool of models, the model with the lowest BIC values is that one with eight peaks (BIC $= -5086$, $r^2 = 0.996$, Fig. 3b). This is the optimal model. Conversely, the model with nine peaks and higher r^2 (0.9962) is discarded because of the higher BIC value (-5078).

The output of the optimal model consists of a table that includes the selected statistical descriptors of each single peak (Eq. 2): position into the excitation-emission plane (μ_i), deviation (σ_i), height (a_i) and asymmetry (r_i). These values allow us to calculate the volume of a specific peak and thus to estimate its contribution with respect to the total fluorescence of a specific EEM. The eight peaks of the optimal model of the previous example are shown in Fig. 4.

The search for the optimal model can be accelerated by removing all the subsets that do not have any chance to

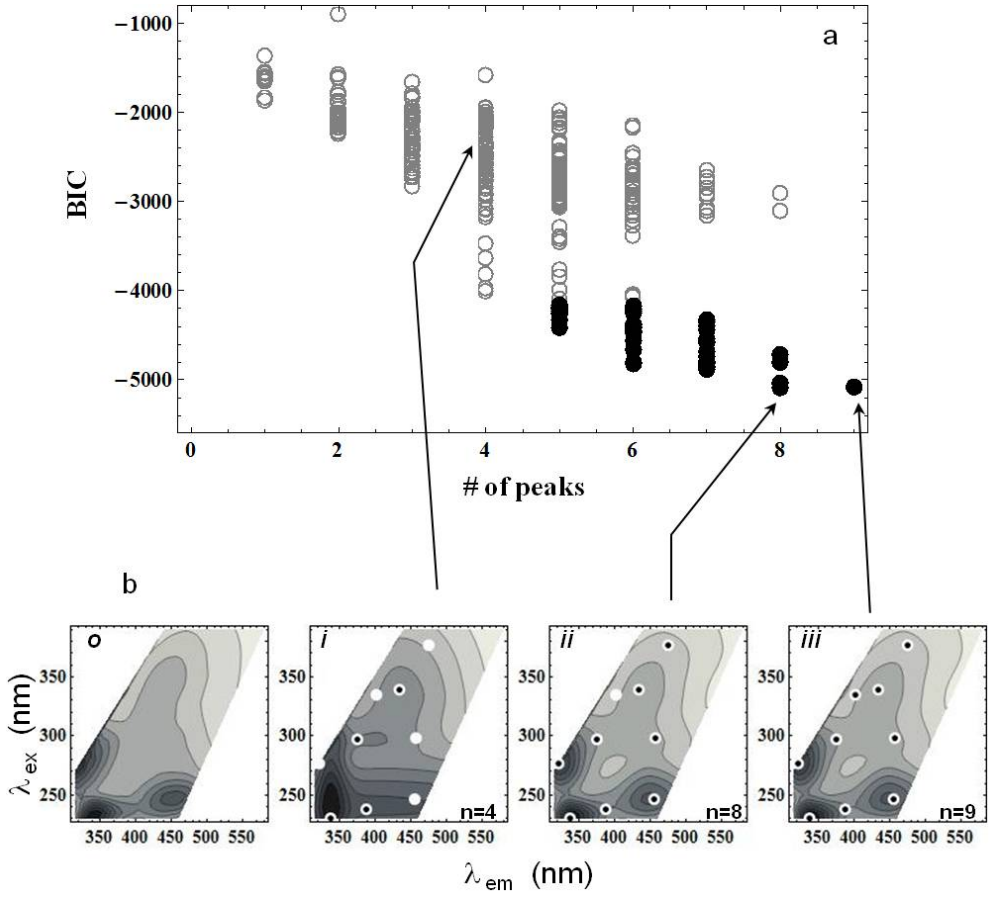

Fig. 3. Visual example of the optimal model selection process. This example refers to the sample S17 with nine potential peaks ($n = 9$). **(a)** shows the relationship between BIC values and number of peaks obtained executing the FDM $z_{(x,y)}$ (Eqs. 1 and 2) for all possible subsets i of the nine potential peaks, where $i = 2^9 - 1 = 511$. Gray disks and black dots discern modelled EEM adjust with r^2 lower and higher than 0.99, respectively. **(b)** shows the contour plots of original EEM sample (o), and three model outputs with a "poor" adjust (i), "optimal" adjust (i.e., lower BIC values, (ii) and overfitted adjust (i.e., larger number of peaks, (iii). Large white and small black dots in contours plots show location of potential and selected peaks, respectively. Each contours represents 10 % of total intensity.

generate a reasonably satisfactory output. A criteria might be to remove all subsets shorter than the length of the list obtained with Max $f_{(xy)}$ (Eq. 3). In fact, it is highly improbable to fit reasonably well $f_{(xy)}$ with less peaks that those detected with Eq. (3). In the preceding example, the numbers of subsets decreased to 64. These 64 subsets include the 53 "good" model outputs that were previously identified in the group of 511 subsets.

3 Model application to a dataset

3.1 The dataset and fluorescence measurements

The dataset integrates 21 EEMs (labelled from S1 to S21) obtained along the main stem of la Tordera river, a 60 km-long human-impacted Mediterranean river, which drains a catchment of 870 km^2 located 70 km to the northeast of Barcelona (Catalonia, Spain). Samples were collected in April 2012, under basal discharge conditions (2 m^3 s^{-1} at the outlet). DOC concentration ranged from 0.5 ppm in headwaters, to 2.9 ppm. Fluorescence analyses were performed with a Shimadzu RF-5301 PC spectrofluorometer. Raw EEM data were corrected and normalised following the steps described in Goletz et al. (2011). Data were normalised by the area under the Raman peak of a deionised water sample at $\lambda_{ex} = 350$ nm and $\lambda_{em} = \{371 \div 428\}$ nm (Lawaetz and Stedmon, 2009). Inner filter effects were corrected by comparing absorbance measurements according to Lackowicz (2006). Absorption spectra were measured with a UV-Visible spectrophotometer UV1700 Pharma Spec (Shimadzu). Each EEM consists of a $\{x, y, z\}$ list of 1050 elements.

DOC concentration and total DOM fluorescence covary significantly through the river main stem ($r^2 = 0.695$, d.f. $= 19$, $p < 0.001$, Fig. 5). This relationship suggests that fluorescent DOM might be a relevant component of dissolved organic matter. Total DOM fluorescence signal increased

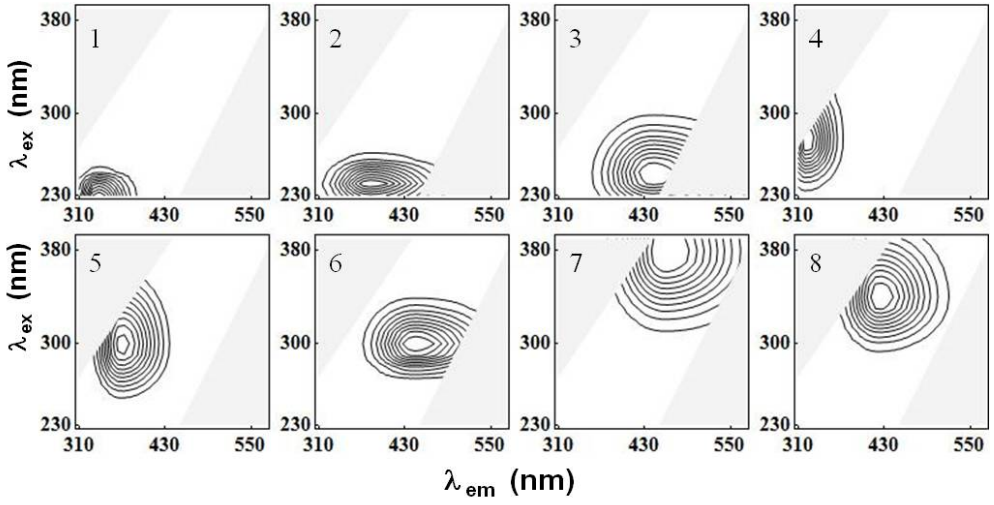

Fig. 4. Shape of the eight peaks individuated in sample S17 after the optimal selection model process illustrated in Fig. 3b (contour plot (ii)). Each contours represents 10 % of total intensity.

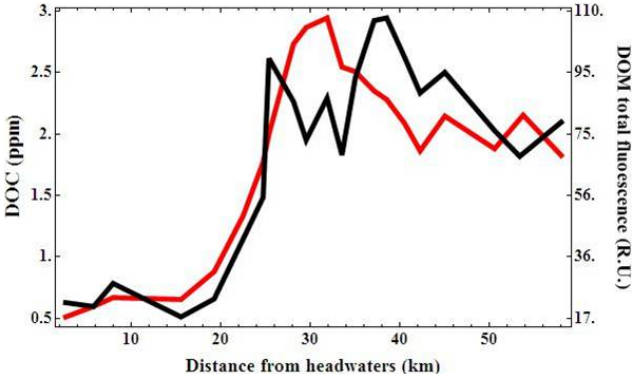

Fig. 5. DOC (red line) and DOM total fluorescence (black line) along the Tordera river main stem. DOC and total fluorescence significantly co-vary ($r^2 = 0.695$, d.f. $= 20$, $p < 0.001$).

abruptly from 25 to 28 km and from 35 to 40 km from the river source as a consequence of point source anthropogenic inputs (mainly waste water treatment plants and industrial effluents). In more detail, shapes of EEMs are extremely variable also and differences among them do not follow a clear gradient along the river main stem (Fig. 6).

3.2 Deconvolution output

Table 1 summarizes deconvolution results for each EEM. Optimal models presented r^2 values ranking between 0.993 (S3) and 0.999 (S6). Figure 6 allows comparing visually the original EEMs and their respective modelled versions for five samples. The fit between modelled and observed EEMs is shown as well. All fits are close to the 1 : 1 line and residuals do not show any clear trend with respect to the magnitude of the fluorescence signal (see the inset of the scatter plots).

At step A, the number n of potential peaks, in each EEM, ranges between seven and eleven and it decreases to five (S13) and ten (S19) after step B. The number of selected peaks tends to increase down river, however, the trend is not significant ($r = 0.53$, d.f $= 19$, $p > 0.01$).

3.3 Data analysis of deconvolved EEMs

Once all EEMs have been deconvolved the implementation of the model is finished. Hereafter the information obtained from deconvolution can be managed and analysed in consonance with the objectives of each study.

If we focus on our study case, it might be of interest to trace in detail the fate and dynamics (in term of intensity or volume) of a specific peak along the river stem, or to explore if some peaks co-vary among the dataset. To address these aspects it is necessary to execute an exploratory analysis to assign those peaks whose coordinates, μ_i, are reasonably close to each other, to a unique "characteristics peak".

The first step consists of exploring how positions of all selected peaks, from all samples, are distributed in the excitation-emission plane (Fig. 7a). Points are visually clustered into ten groups. These clusters become more intelligible if we convert the scatter plot into an array to generate a contour plot (Fig. 7b). Contour lines also help to assign a centre for each cluster. These centres describe the coordinates of the potential "characteristic" peaks. If it is considered relevant, it is possible to be more meticulous by adding new clusters ad hoc. For instance, we might be interested in exploring if and when a peak shifts in a small portion of the excitation-emission plane. Figure 7b shows that in area $\lambda_{ex} < 250$ mn and $420 < \lambda_{em} < 500$ nm (a fulvic-humic like region named "A", Ishii and Boyer, 2012) the contour plot might suggest the presence of two groups of points very close to each other. Therefore, we split the points located into this small zone

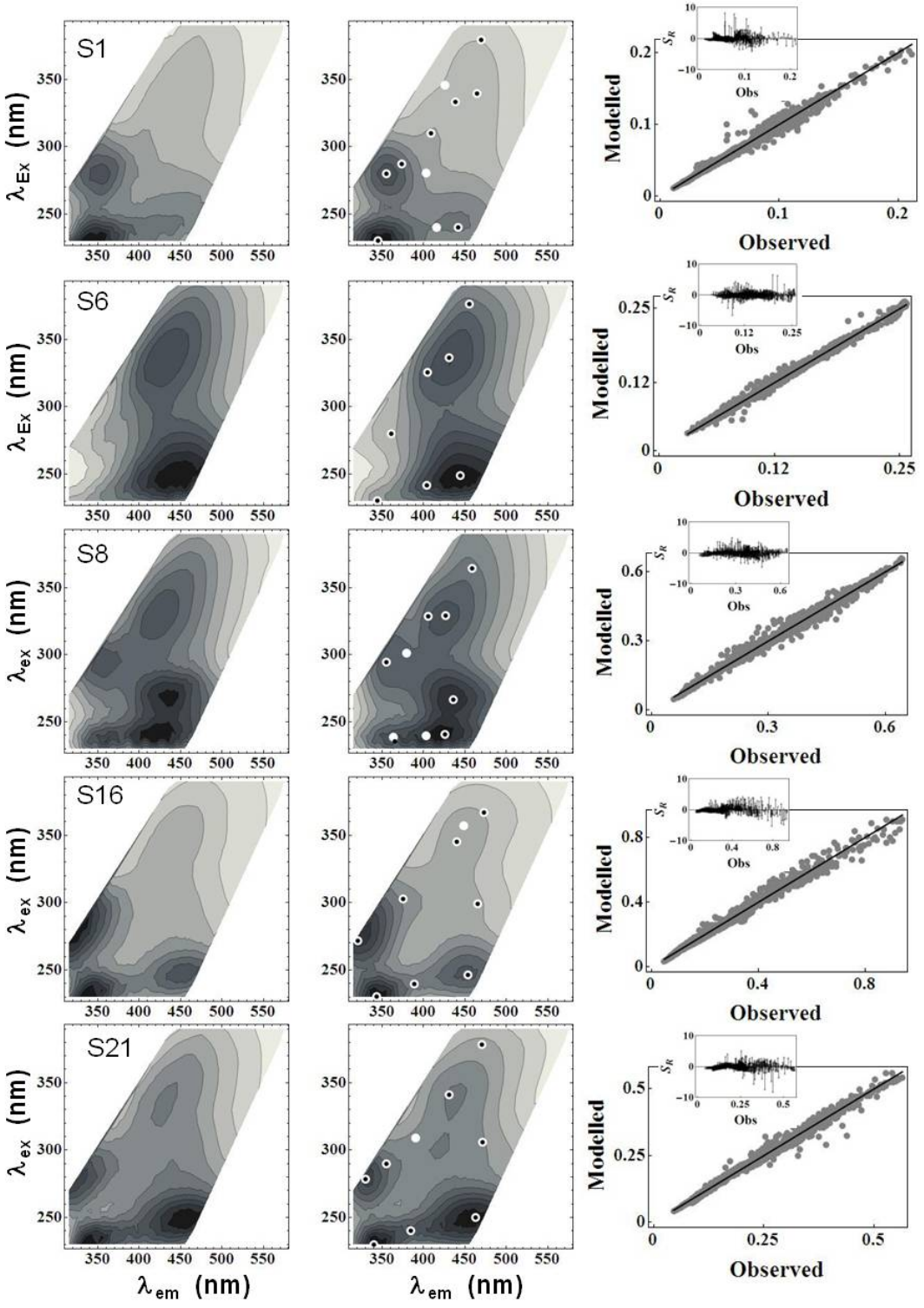

Fig. 6. Examples of observed (left) and modelled (center) EEMs from la Tordera river from five sampling sites. Each contour represents 10 % of the total intensity. Large white and small black dots in contours plots show the location of potential and selected peaks respectively. Scatter plots on the right show the fit between modelled and observed EEMs (values in Raman units). Each dot represents a single fluorescence pair that compose the observed and modelled EEM. Solid line shows the 1 : 1 line. The inset shows the plot of standardized residuals (SR) respect to the observed (Obs) data.

Table 1. DOC concentration and deconvolution results for each DOM sample. The number of initial potential peaks (step A) and selected peaks after step B, goodness-of-fit parameters (r^2 and BIC) of each optimal FDM model are provided. Numbers in parenthesis in the first column show the distance (km) of each sampling point from the headwaters.

		Surface Analysis: Step A		Deconvolution: Step B		
Sample	DOC (ppm)	Max $f_{(x,y)}$ (#)[a]	L_n (#)[b]	Selected Peaks (#)	r^2	BIC
S_1 (3)	0.5	3	11	8	0.994	−7606
S_2 (6)	0.6	2	10	6	0.997	−8753
S3 (8)	0.7	4	10	6	0.993	−47077
S4 (16)	0.7	2	9	7	0.998	−9595
S5 (19)	0.9	2	9	6	0.997	−8319
S6 (22)	1.3	2	7	7	0.999	−7965
S7 (24)	1.8	2	8	6	0.994	−5687
S8 (26)	2	5	9	7	0.998	−5362
S9 (28)	2.7	3	7	7	0.998	−5844
S10 (30)	2.9	3	8	8	0.997	−5830
S11 (32)	2.9	3	9	8	0.998	−5743
S12 (34)	2.5	4	7	6	0.997	−5994
S13 (35)	2.5	3	8	5	0.995	−4816
S14 (37)	2.4	4	7	7	0.995	−4321
S15 (39)	2.3	5	9	8	0.996	−4621
S16 (41)	2.1	4	9	8	0.995	−4649
S17 (42)	1.9	4	9	8	0.996	−5086
S18 (45)	2.1	4	9	9	0.998	−5578
S19 (51)	1.9	4	11	10	0.998	−6023
S20 (53)	2.2	4	11	8	0.998	−6173
S21 (58)	1.8	4	9	8	0.998	−5908

[a] Length of list Max $f_{(x,y)}$ (Eq. 3);
[b] number of potential peaks (length of list L_n, Eq. 5)

into two clusters. In consequence, in our dataset we discern eleven clusters, each one with its own centre (the characteristic peaks, labelled P1 to P11). To identify the boundary of each cluster the Voronoi diagram tessellation approach is implemented (Aurenhammer and Klein, 2000). All points that lie within a region are assigned to the centres of that region (Fig. 7c).

The Voronoi diagram is a method used to divide a surface into "*n*" polygons. The boundary of each polygon goes through the middle of a segment that joins two adjacent centres. These segments are obtained with the Delaunay triangulation algorithm. All points that lie within a polygon are assigned to the centres of that polygon. Both the Voronoi diagram and the Delaunay triangulation are usually implemented in the mathematical software.

The Voronoi diagram, as any other clustering approach, has some inherent limits. In this case, it is the user experience that determines the number and position of centres. To evaluate our visual approach, a hierarchical clustering analysis (HCA) is executed to individuate automatically the clusters in the scatter plot of Fig. 7a. In the HCA the Euclidean distance and the agglomerative function are implemented. To identify the optimal number of clusters, the "silhouette" test is adopted (Rousseeuw, 1987). The HCA detected nine clusters (Supplement 2). Eight clusters out of eleven coincided with those identified with the Voronoi diagram suggesting that the visual approach is quite accurate. The difference among the two approaches, lies in the area $\lambda_{ex} < 275$ nm and $420 < \lambda_{em} < 470$ nm. In this portion, HCA individuated one cluster, while three clusters (P4, P5 and P7, Fig. 7c) were identified visually.

Peaks P4 and P5 are both fulvic-humic like peaks, very close to each other, described previously and their identification responds to our interest to trace how a peak shifts in position in this small portion of the excitation-emission plane (see below for results). With respect to P7, a peak with similar $\lambda_{ex}/\lambda_{em}$ coordinates has been detected in outlets of wastewater treatment plants in another study (Saadi et al., 2006). It appears to have a different origin with respect to P4 and P5. Therefore, it seems reasonably, in a biogeochemical context, to preserve its individuality.

Most of the identified peak coordinates matched well (or were close) to those found in the literature with visual peak-picking (Baker, 2002; Hudson et al., 2007) or with multivariate methods (Cory and McKnight, 2005; Fellman et al., 2010). Peaks from P1 to P3 encompass the protein-like

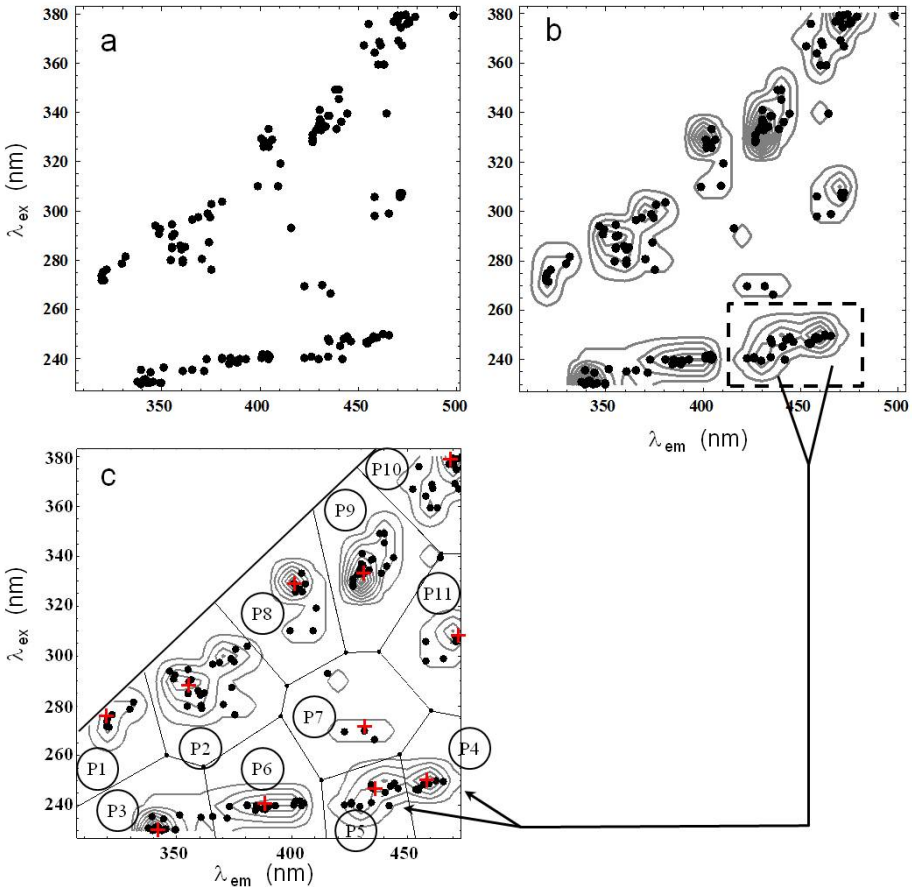

Fig. 7. (a) Scatter plot showing the position of all the peaks identified in all samples after the deconvolution (black dots) within the excitation vs. emission plane; **(b)** Contour plot of the scatter plot generated by converting the two-dimensional data into an array of counts of width of 10. This data representation allows to discern visually ten groups. Dashed rectangle highlights points located at $\lambda_{ex} < 250$ nm and $420 < \lambda_{em} < 500$ nm. These points are further split into two clusters to obtain a total number of eleven clusters; **(c)** Voronoi diagram of the eleven peaks showing the limits of each cluster. Red "+" symbols show the centre of each cluster. Numbers into white circles label the eleven peaks in the dataset.

region (frequently named peaks B, T1, and T2, Coble et al., 1998). The remaining peaks (from P4 to P11) are located in the humic-fulvic like region (peaks A, C and M, according the nomenclature proposed by Coble et al., 1998; Table 2).

Significant correlations between peak intensities (a_i) are observed. Some of them relate a relationship between peaks within the humic/fluvic like region (P4 with P11, $r = 0.74$, $p < 0.001$) or within the protein-like region (P1 with P3, $r = 0.89$, $p < 0.001$). Although the reduced number of samples obligated to be cautious, these results suggest that P1 and P3 (or P4 and P11) might describe a fluorophore with two maxima. Other correlations related protein-like peaks (P3 and P2) with that of humic/fluvic like (P4, P11 and P10) suggesting that these peaks might have a common origin, probably linked to the anthropogenic inputs along the river channel (Table 3).

The strategy to discern "characteristics peaks" very close to each other, allows to analyse in detail how their position and relevance (in terms of peak intensity, a_i) change within a small region of the excitation-emission plane. For instance, P4 ($\lambda_{ex}/\lambda_{em} \sim 250/460$ nm) and P5 ($\lambda_{ex}/\lambda_{em} \sim 246/441$ nm) are both located within the region traditionally named peak "A" ($\lambda_{ex} \leq 260$, $400 \leq \lambda_{em} \leq 500$, Ishii and Boyer, 2012). P5 appears between the headwaters and 35 km. Downriver, it disappears and is replaced by P4. In fact, these two peaks never coincided in the same sample, with the exception of S15 (Fig. 8a). This shift in peak positions toward larger emission wavelength might indicate changes in DOM molecular weight along the river main stem: from relatively small compounds in head waters to larger ones downriver (Ishii and Boyer, 2012).

In the protein-like region, the P2 ($\lambda_{ex}/\lambda_{em} \sim 290/356$ nm) appears at head waters with a maximum between 24 and 40 km. It coexists with P1 ($\lambda_{ex}/\lambda_{em} \sim 272/319$ nm; Fig. 8b).

Table 2. Coordinates and brief description of each one of the eleven main peaks detected in the dataset after the deconvolution.

Peak id	λ_{ex} (nm)	λ_{em} (nm)	# cases[c]	Conventional Class[a]					
				Protein like			Humic/Fulvic like		
				B	T1	T2	A	M	C
P1	272	319	7	X					
P2	290	356	21		X				
P3	231	339	17			X			
P4	250	460	9				X		
P5	246	441	14				X		X
P6	239	385	17				X		X
P7	269	433	5				X[b]		
P8	326	402	11					X	
P9	332	431	21					X	X
P10	380	471	21						X
P11	307	471	7			unknown			

[a] Coble et al. (1998).
[b] Detected in a wastewater treatment plant (Saadi et al., 2006)
[c] Number of EEMs, in which the peak has been detected.

Table 3. Pairwise Pearson correlations values between the peak intensities identified in the 21 EEMs after the deconvolution. Values in bold represent the significant correlations ($P < 0.001$, d.f. $= 19$).

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
P1	1	0	0	0	0	0	0	0	0	0	0
P2	0.19	1	0	0	0	0	0	0	0	0	0
P3	**0.89**	0.45	1	0	0	0	0	0	0	0	0
P4	**0.73**	0.29	**0.83**	1	0	0	0	0	0	0	0
P5	−0.10	0.29	−0.06	−0.34	1	0	0	0	0	0	0
P6	0.52	0.26	0.39	0.43	−0.12	1	0	0	0	0	0
P7	−0.27	0.19	−0.33	−0.33	0.28	0.20	1	0	0	0	0
P8	−0.16	0.01	−0.15	−0.32	0.29	0.18	0.29	1	0	0	0
P9	0.25	0.50	0.43	0.35	0.27	0.20	0.07	−0.02	1	0	0
P10	0.17	**0.75**	0.38	0.35	0.30	0.34	0.35	0.30	0.61	1	0
P11	**0.70**	0.22	**0.74**	**0.74**	−0.44	0.56	−0.23	0.03	0.18	0.17	1

They are considered labile substrates related to different biological processes such as bacterial (Hudson et al., 2007), dead organisms or primary producers leachates (Fellman et al., 2010). Additionally, these signals are frequently associated to wastewater treatment plant effluents (Saadi et al., 2006; Baker, 2002). The elevated intensities of these protein-like peaks in the dataset, presumably indicates the inputs of anthropogenic origins (Hudson at al., 2007; Saadi et al., 2006). The abrupt detection of P1 might indicate the increase of contribution of degraded proteins/peptides downriver (Fellman et al., 2010).

Finally, peaks P7 and P11 (within the humic-fulvic like region) are detected abruptly in two different points of the river continuum: P7 appears at 24 km, meanwhile P11 at 37 km. Both peaks disappeared down waters (Fig. 8c) suggesting high bioavailability or high photo-degradation rates.

The clear relationship between DOC and total DOM fluorescent signal mentioned previously (see Fig. 5), represents a starting point to explore in more detail which regions of the EEMs co-vary more significantly with DOC. To address to this question we firstly estimated the volume of each peak detected in EEMs dataset (Eq. 2 is used to calculate the volume of each peak). Successively, we executed a step-wise linear multiple regression to extract those peaks that more significantly co-vary with DOC. This analysis reveals that DOC strongly co-varies with six peaks: P2, P5, P8, P9, P10 and P11 ($r^2 = 0.96$, d.f. $= 15$, $p < 0.001$, Fig. 9). Peaks P8, P9 (humic-like, $p < 0.00057$ and $p < 0.0006$, respectively) and P2 (protein-like, $p < 0.006$) are the fluorescent events more significantly related to DOC concentration. Additionally, the analysis shows that the emergence of P8 is associated with the beginning of the increase of DOC from 20 km. Successively, P2 and P10 acquired more relevance in coincidence with the higher DOC concentrations (from 25 to 36 km, Fig. 9).

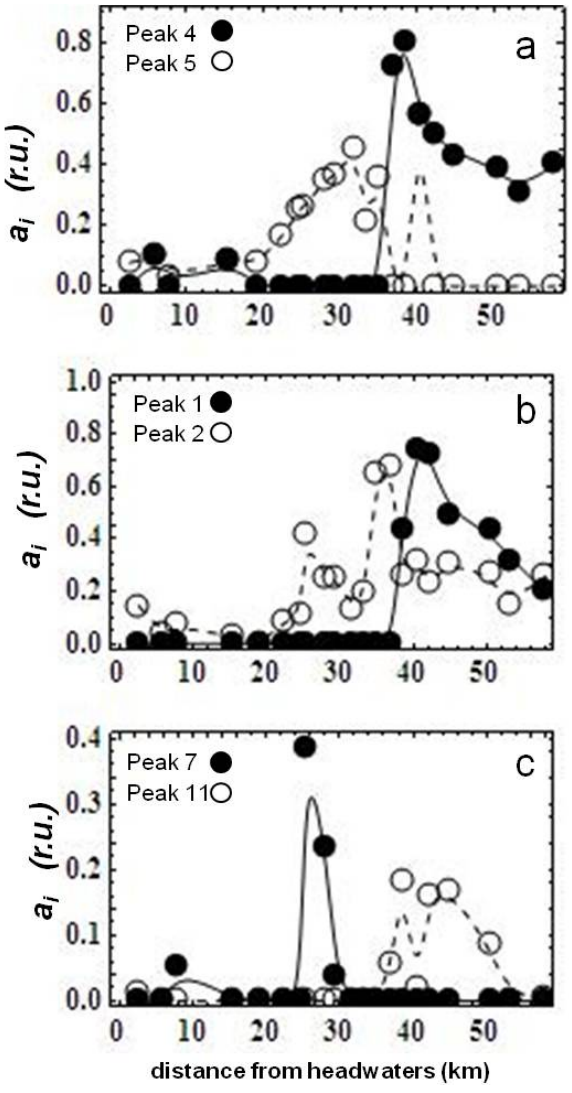

Fig. 8. Intensity of several peaks (in Raman units) detected along the la Tordera river. Solid and dotted lines show the smoothed trend of each peak. Lines are obtained with a cubic B-spline function and have only a descriptive purpose.

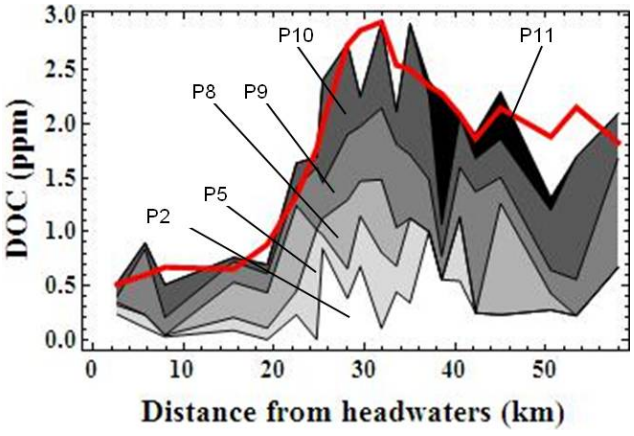

Fig. 9. Comparison of the observed DOC spatial dynamic (red line) with that modelled with six peaks (P2, P5, P8, P9, P10, P11) according the step-wise linear multiple linear regression ($r^2 = 0.969$, d.f. $= 14$, $p < 0.001$). Different gray tonality depicts the contribution of each peak.

4 Conclusion

Advances in our knowledge of DOM fluorescence properties in aquatic ecosystems strongly benefit from technological advances in data acquisition of fluorescence data. In consequence deconvolution statistical tools are becoming indispensable to manage and extract information from the large amount of data generated by these instruments. However, sophisticated spectrofluorometers are not widespread and the opportunity to generate large dataset of EEMs is limited in several laboratories. Furthermore, the community recognises limitations of the multivariate tools (Fellman et al., 2009) especially when the dataset is integrated by a heterogeneous pool of EEMs (Stedmon and Bro, 2008). Therefore, a large portion of studies on DOM fluorescence does not implement the deconvolution tools to explore EEMs properties.

Finite mixture models are becoming usual in several scientific disciplines (McLachlan and Peel, 2004). For instance, a recent implementation of FDM, that shows an evident similitude with that described in this note, consists in the identification of peaks/subpopulations in bidimensional cytograms (Boedigheimer and Ferbas, 2008).

FDM analyses individual EEMs. Thus, once the deconvolution is executed and the optimal model for each EEM is found, the implementation of the model is completed. In consequence FDM is not designed to analyse differences among EEMs or to explore if two peaks represent different maxima of the same fluorophore.

From this point forward, an additional data treatment is necessary to analyse these aspects. Inevitably, the post-model analysis makes sense if a relatively large dataset is explored (for instance more than 20 EEMs). However, we remark that number of EEMs that integrate a dataset do not influence the result of the deconvolution of an EEM.

In this note, the information extracted by the deconvolution of our dataset is displayed and analysed in a simple and intuitive way for descriptive purposes only. The approach used to cluster all peaks individuated with the FMD into a set of eleven "characteristic peaks" (Fig. 7) is based on a visual inspection of the positions of the selected peaks into the excitation-emission plane and preliminary information from literature (fundamental in discerning P7). However, the clustering strategy described in this note does not aim to be a standard protocol. Evidently, objectives and experimental design as well as the size of the dataset, modulate the strategy of analysis of the deconvolved EEMs. It emerges that, once the basic model is developed, the next challenge consists of finding a standard technique to deal with the model output

analysis. Under this context, a coupling between FDM outputs and multiway techniques is a pathway that might be explored in the future.

We conclude that the FDM expands the family of deconvolution tools opening the perspective to implement it with datasets composed by extremely different EEMs. The idea underlying the FDM is intuitive and the mathematical language is not excessively complex. Experts on data mining tools remark that sophisticated and complex deconvolution tools can produce similar results to those obtained with a simple peak-picking (Bieroza et al., 2011). In this framework, the approach described here could be viewed as an in-between step between the two extremes because it integrates an improved version of the peak-picking into a relatively simple deconvolution algorithm. This aspect might further help to bring researchers closer to these techniques.

Supplementary material related to this article is available online at: http://www.biogeosciences.net/10/5875/2013/bg-10-5875-2013-supplement.zip.

Acknowledgements. This research is funded by the Spanish Ministry of Education and Science (MEC) (CGL2011-30151-C02-02) and European Community 7th Framework Programme (No. 603629-ENV-2013-6.2.1-Globaqua). Elisabet Ejarque's research was supported by an FPU doctoral scholarship from the MEC (AP2008-03431). Both authors are members of the GRACCIE consortium. We would like to thank Madgalena Bieroza, Jordi Flos and Eusebi'Vazquez for useful comments and suggestions on an earlier version of this manuscript. We also thank S. Ishii and an anonymous reviewer for their stimulating contributions during the review process.

Edited by: G. Herndl

References

Aurenhammer, F. and Klein, R.: Voronoi Diagrams, in: Handbook of Computational Geometry, edited by: Sack, J. R. and Urrutia, J.: North-Holland, Amsterdam, the Netherlands, 201–290, 2000.

Baker, A.: Fluorescence Excitation-Emission Matrix Characterization of River Waters Impacted by a Tissue Mill Effluent, Environ. Sci. Technol., 36, 1377–1382, 2002.

Bentley, J., Stanat, D., and Williams Jr., E.: The Complexity of finding fixed-radius near neighbors, Inform. Process. Lett., 6, 209–213, 1977.

Bieroza, M., Baker, A., and Bridgeman, J.: 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.-Biogeo., 114, G00F07, doi:10.1029/2009JG000940, 2009.

Bieroza, M. Baker, A., and Bridgeman, J.: Classification and calibration of organic matter fluorescence data with multiway analysis methods and artificial neural networks: an operational tool for improved drinking water treatment, Environmetrics, 22, 256–270, 2011.

Boedigheimer, M. J. and Ferbas, J.: Mixture modelling approach to flow cytometry data, Cytom. Part A, 73A, 421–429, 2008.

Boheme, J. R. and Coble, P. G.: Characterization of coloured dissolved organic matter using high energy laser fragmentation, Environ. Sci. Technol., 34, 3283-i3290, 2000.

Coble, P. G., Green, S. A., Blough, N. V., and Gagosian, R. B.: Characterization of dissolved organic-matter in the Black Sea by fluorescence spectroscopy, Nature, 348, 432–435, 1990.

Coble, P. G., Del Castillo, C. E., and Avril, B.: Distribution and optical propoerties of CDOM in the Arabian Sea during the 1995 SW moonson, Deep-Sea Res. Pt. II, 45, 2195–2223, 1998.

Cory, R. M. and McKnight, D. T.: Fluoescence spectroscopy reveals ubiquitous presence of oxidized and reduced quinones in dissolved organic matter, Environ. Sci. Technol., 39, 8142–8149, 2005.

Fellman, J. B., Hood, E., and Spencer, R. G. M.: Fluorescence opens new windows into dissolved organic matter dynamics in freshwater ecosystems: a review, Limnol. Oceanogr., 55, 2452–2462, 2010.

Frjguhwirth-Schnatter, S.: Finite mixture and Markov switching models, Springer Series in Statistics, doi:10.1007/978-0-387-35768-3, 492 p., 2006.

Ganza, V. G. and Vorozhtsov, E. V.: Numerical Solution for Partial Differential Equations, Problem Solving Using Mathematica, CRC Press, 347 pp., 1996.

Goletz, C., Wagner, M., Gruebel, A., Schmidt, W., Korf, N., and Werner, P.: Standardization of fluorescence excitation-emission-matrices in aquatic milieu, Talanta, 85, 650–656, doi:10.1016/j.talanta.2011.04.045, 2011.

Her, N., Amy, G., McKnight, D., Sohn, J., and Yoon, Y.: Characterization of DOM as a function of MW by fluorescence EEM and HPLC-SEC using UVA, DOC, and fluorescence detection, Water Res., 37, 4295–4303, 2003.

Horst, R. and Pardalos, P. M.: Handbook of Global Optimisation. Kluwer. Dordrecht, the Netherlands, 900 pp., 1995,

Hudson, N., Baker, A., and Reynolds, D.: Fluorescence analysis of dissolved organic matter in natural, waste and polluted waters – a review, River Res. Applic., 23, 631–649, 2007.

Ishii S. K. and Boyer T. H.: Behavoir or reoccurring PARAFAC components in fluorescent dissolved organic matter in natural and engineering systems: a critical review, Environ. Sci. Technol., 46, 2006–2017, 2012.

Kato, T., Omachi, S., and Aso, H.: Asymmetric Gaussian and Its Application to Pattern Recognition In Structural, Syntactic, and Statistical Pattern Recognition, in: Lecture Notes in Computer Science, edited by: Caelli, T., Amin, A., and Dvin, R. P. W.,, SSPR&SPR, LNCS 2396, 405–413, doi:10.1007/3-540-70659-3, 2002.

Korshin, G. V. Kumke, M. U., Li, C. W., and, Frimmel, F. H.: Influence of Chlorination on Chromophores and Fluorophores in Humic Substances, Environ. Sci. Technol., 33, 1207–1212, 1999.

Lakowicz, J. R.: Principles of Fluorescence Spectroscopy, 3rd Edn., Springer, 1255 pp., 2006.

Lawaetz, A. J. and Stedmon, C. A.: Fluorescence intensity calibration using the Raman scatter peak of water, Appl. Spectrosc., 63, 936–940, 2009.

McLachlan, G. and Peel, D.: Finite Mixture Models, in: Wiley series in probability and statistics: Applied probability and statistics, Wiley-Interscience, 456 pp., 2004,

Nelder, J. A. and Mead, R.: A simplex method for function minimization, Comput. J., 7, 308–313, 1965

Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math., 20, 53–65, 1987.

Saadi, I., Borisover, M., Armon, R., and Laor, Y.: Monitoring of effluent DOM biodegradation using fluorescence, UV and DOC measurements, Chemosphere, 63, 530–539, 2006.

Schwarz, G.: Estimating the dimension of a model, Ann. Stat., 6, 461–464, 1978.

Stedmon, C. A. and Bro, R.: Characterizing dissolved organic matter fluorescence with parallel factor analysis: a tutorial. Limnol. Oceanogr.-Meth., 6, 572–579, 2008.

Stevenson, P. G., Mnatsakanyan, M., Guichon, P. G., and Shalliker, R. A.: Peak picking and the assessment of separation performance in two-dimensional high performance liquid chromatography, Analyst, 135, 1541–1550, 2010.

Traganza, E. D.: Fluorescence excitation and emission spectra of dissolved organic matter in sea water, B. Mar. Sci., 9, 897–904, 1969.

Westerhoff, P., Chen, W., and Esparza, M.: Fluorescence Analysis of a Standard Fulvic Acid and Tertiary Treated Wastewater, J. Environ. Qual., 30, 2037–2046, 2001.

Yamashita, Y. and Tanoue, E.: Chemical characterization of protein-like fluorophores in DOM in relation to aromatic amino acids, Mar. Chem., 82, 255–271, 2003.