Supporting Information for:

The One-Sample PARAFAC Approach Reveals Molecular Size Distributions of Fluorescent Components in Dissolved Organic Matter

Urban J. Wünsch[1*], Kathleen R. Murphy[2], Colin A. Stedmon[1]

[1]National Institute of Aquatic Resources, Section Oceans and Arctic, Technical University of Denmark, Kemitorvet, Building 201, 2800 Kgs. Lyngby, Denmark

[2]Chalmers University of Technology, Water Environment Technology, Sven Hultins Gata 6, 41296 Gothenburg, Sweden

**Corresponding author*: Urban J. Wünsch (urbw@aqua.dtu.dk)

DOI of the main manuscript: 10.1021/acs.est.7b03260

This Supporting Information contains 28 pages, 13 figures, and 2 tables.

Summary of SI contents

S1: Supplementary Methods

Solid-phase extraction of DOM

PPL-based extractions were carried out as described previously.[1,2] Samples were filled inot 1 L pre-combusted (500 °C, 5 hrs) glass bottles and stored for no more than 3 hrs before filtration. 0.7 µm-filtered water was acidified to pH 2 using hydrochloric acid (p.a. grade) and processed directly upon acidification. PPL cartridges (1 g resin, Agilent Bond Elut PPL) were conditioned with methanol, and rinsed with pH 2-water prior to extraction. Afterwards, sample water was applied onto the resin by means of gravity. Cartridges were then rinsed with pH 2-water, dried with N_2, and DOM was eluted with 10 mL methanol. Extracts were stored at -18°C prior to further analyses.

All available details on the XAD8-based extraction samples are listed elsewhere. [3,4] Dried DOM extracts were stored at room temperature in amber glass vials until further analysis.

It should be noted that PPL resins are only able to extract 43 - 65 % of bulk DOC, while XAD-8 may extract 10 – 79 % of DOC. [2] By nature of the resin, extractions (both PPL and XAD-8) preferentially select the more hydrophobic fraction of DOM and thus may inadvertently serve to homogenize the characteristics of the extracted DOM. Furthermore, differences between allochthonous DOM (extracted with PPL resin) and autochthonous DOM (extracted with XAD-8 resin) may be a consequence of the usage of different extraction resins. With regards to DOM

fluorescence, the relative contribution of PARAFAC components may thus be influenced and may not represent the original sample.

High performance size-exclusion chromatography

All buffers were made with ultra-pure water (18.2 MΩ cm, TOC <5ppb). Samples were injected (5 – 25 µL) using a Shimadzu SIL-30AC equipped with a 50 µL injection loop into a mobile phase of 150 mM of ammonium acetate (trace metal grade; Sigma Aldrich) pumped by a Shimadzu LC-20AB at a flow rate of 0.3 mL min^{-1}. For the size-dependent separation of DOM, a TSKgel SuperAWM-H column with a linear range of approx. 10^2 to $2*10^6$ Da was used, heated to a temperature of 30 °C using a Shimadzu CTO-10ASvp.

The eluting DOM was detected with two sequential detectors: First, absorbance was detected between 240 and 600 nm with an increment of 2 nm at a rate of 1.5 Hz using a Shimadzu SPD-M30. Secondly, fluorescence emission was detected between 300 and 600 nm at 5 nm increments using a Shimadzu RF-30Axs at variable excitation wavelengths (240 – 360 nm at 5nm increments, 370 – 450 nm at 10 nm increments) at a rate of 1 Hz. The inter-detector volume was determined and subsequently subtracted as described in the Supplementary Material.

After every run, the column was rinsed for 5 minutes with a mixture of 50 vol% Methanol (LC/MS grade, Sigma Aldrich) in ultrapure water to allow for the elution of any DOM with high column affinity (i.e. showing secondary interaction). After the rinse, the system was equilibrated to allow detectors to return to their respective baseline readings resulting in a total run time of 40 min per sample. Parts of the chromatogram that contained the elution of Methanol-eluted compounds were excluded from further analyses due to the non-linearity of optical properties associated with changing solvent composition. The final chromatogram resulted in approximately 1500 individual absorbance and fluorescence emission scans per sample.

For every sample, a sequence of runs was performed whereby the same sample was injected while instrument parameters were systematically changed between runs (Supporting Information (SI) Figure S1, left panel). In total, the chromatographic run was repeated 35 times to allow the determination of absorbance properties (one run) and fluorescence properties at an EEM-like spectral resolution (34 runs). Variable excitation increments were chosen to save time and provide valuable information in variable regions of the EEM. During the early runs, fluorescence emission was measured at excitation wavelengths varying between 240 and 360 nm with an increment of 5 nm, while during later runs, emission was measured at excitations between 370 and 450 nm with an increment of 10 nm. To ensure chromatographic

stability of the system during the 23-hour period of an experiment, the absorbance detector was used in four-channel mode to track the alignment of chromatograms between different fluorescence emission runs, while the fluorescence detector was used in its four-channel mode when absorbance was subsequently scanned in full spectral resolution. Moreover, a reference sample was analyzed after 17-19 runs to verify system stability and account for possible long-term variability in the light output of the RF-30Axs' Xenon-lamp. The developed approach ensured the rigorous chromatographic alignment of runs across the 35 injections, fulfilling one of the prerequisites for the application of the PARAFAC model. In cases of misalignment, samples were reanalyzed since the proper alignment is a central requirement for PARAFAC. Misalignment may be observed in PARAFAC models when (1) core consistencies for 2-3 component models are atypically low, (2) model residuals show leftover emission spectra for single excitation wavelengths (indicating misaligned data), and (3) PARAFAC takes longer than usual to converge towards a solution. Running a reference sample to confirm chromatic alignment is highly recommended.

To align detector signals and account for the system's inter-detector volume, a highly concentrated standard with previously determined optical properties was injected (salicylic acid, AppliChem).[5] The dead volume between absorbance and fluorescence detectors was determined to be 68.2 µL.

The effect of peak broadening due to the turbulence created in the different detector cells was investigated using tryptophan (Alfa Aesar): Tryptophan was injected without the analytical column and peak shapes of absorbance (280 nm) and fluorescence (Excitation 280 nm, Emission 400 nm) were compared after the correction of the inter-detector volume. However, it was found that no correction was needed since both peak shapes were highly similar and did not exhibit systematic trends indicative of diffusive processes that would distort peak shapes systematically between detectors (SI Figure S2).

A reference sample was measured approximately every 15th run to ensure the alignment across a high number of injections and between different samples (SI Figure S3). Chromatograms were found to randomly vary in peak position and intensity across all 12 runs.

The analytical column was calibrated using Polystyrene sulfonate (Sigma Aldrich) covering the range of 0.21 to 77 kDa (SI Figure S4). However, the stationary phase of the SuperAWM-H column (Methacrylate) made it necessary to obtain these chromatograms with a mixture of 80 / 20 (v/v) 150 mM Ammonium acetate and methanol. For DOM, this calibration resulted in peak molecular size values in comparable to previous studies.[6–8] However, a low molecular size tail was observed that stretched beyond the linear range of size separation (100 Da), most likely due to secondary interactions. Separations in the tail of chromatograms therefore likely represented a mixture of size separation and secondary hydrophobic retention. While

particularly apparent with the methacrylate stationary phase used in this work, the issue of secondary interactions remain a well-known, unresolved issue of HPSEC in the context of DOM separation.[9] As a consequence of these factors, the column calibration was used only to provide an estimate of the apparent molecular size for peak values.

Data processing

The chromatographic raw data were generated with Shimadzu LabSolutions 5.71A and exported to individual ASCII-files. Subsequent processing was done in Matlab v9.2 (2017a) according to the scheme detailed in SI Figure S1 (left panel). First, all 34 chromatographic runs per physical sample were imported separately to obtain datasets containing fluorescence and absorbance measurements. Next, retention times were converted into elution volumes and the inter-detector volume was accounted for by subtracting the dead volume of 68.2 µL from the elution volumes of fluorescence datasets. Peak broadening due to turbulence in the detector cells was found to be minimal (SI Figure 2), and was thus considered to be negligible. Since fluorescence and absorbance were measured in different retention time intervals, both datasets were interpolated using a 2D linear interpolation function to obtain data for identical elution volumes. As a final preprocessing step, the integrated fluorescence emission at an excitation of 300 nm was compared between reference samples to account for temporal variability in the Xe-lamp output (as done in SI Figure S3). Wherever the total variability exceeded 5 %, fluorescence runs were corrected by applying a correction factor obtained with a spline function to interpolate between individual reference sample runs. This process produced a fully-corrected, three-way HPSEC-EEM dataset of elution volume x emission x excitation for each sample.

Subsequent processing included the inner-filter-effect correction, scatter excision, and outlier removal, following procedures described elsewhere.[10] In particular, the extent of 1st and 2nd order Rayleigh scatter was large in comparison to typical spectrofluorometers. As a result, 30 and 10 nm above and below 1st order Rayleigh scatter, as well as 30 nm above and below 2nd order Rayleigh scatter were excised and not interpolated. As a consequence of the large diagonal areas of undefined numbers, EEM were restricted to Excitation between 260 and 450 nm and Emission between 310 and 550 nm to limited the negative influence that large areas of undefined values may have on least-squares fitting algorithms.

The physical scatter of light in the HPLC fluorescence detector was found to exceed the usual wavelength range at the common excitation wavelength of 350 nm. Since Rayleigh and Raman peaks overlapped the following procedure was developed to obtain the Raman area for signal normalizations (SI Figure S5): (1) The shoulder of the Rayleigh peak influencing the Raman peak was removed by setting all affected values to missing numbers; (2) The leftover part of the Raman peak was used to fit a four-term Gaussian curve (using wavenumbers); (3) The fitted Gaussian was used to extrapolate to previously removed wavenumbers; (4) The Raman area was calculated using the extrapolated Raman peak (using wavelengths in the range of 365 – 435 nm). Examples of size separated EEMs as produced by the HPSEC system described above can be found in SI Figure S6.

S2: Supplementary Figures

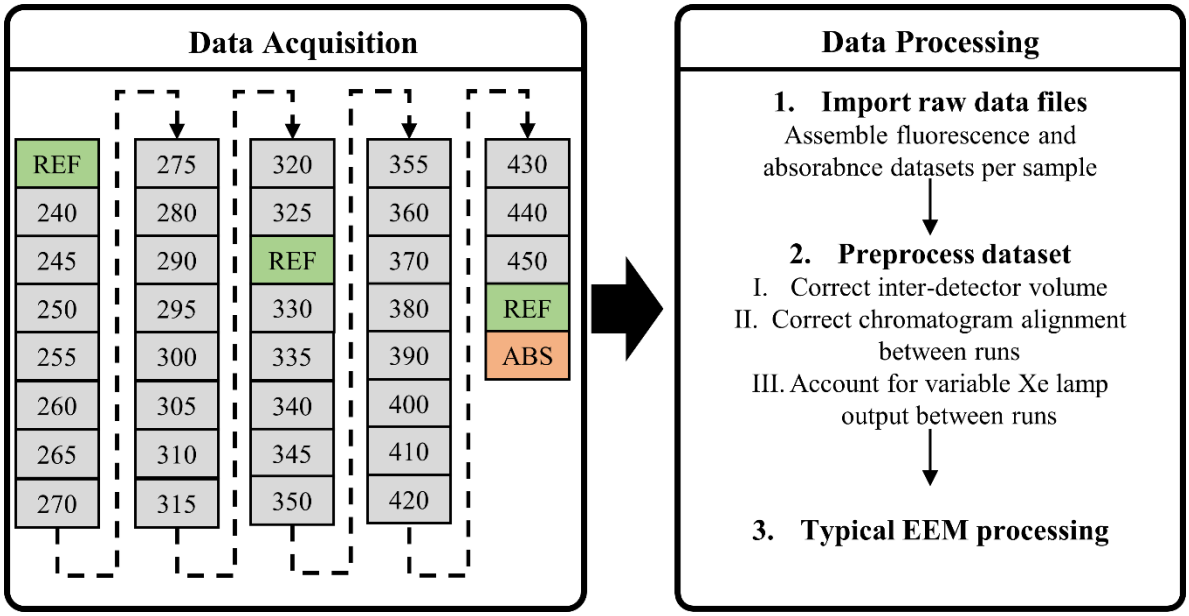

Figure S1. Schematic illustration of HPSEC data acquisition (left) and data processing (right) required to obtain size-dependent absorbance spectra and fluorescence EEMs of DOM samples. Grey rectangles depict the measurement of fluorescence emission at the stated excitation wavelength, green rectangles depict the measurement of a reference sample for alignment purposes, the orange rectangle depicts the measurement of CDOM absorbance spectra.

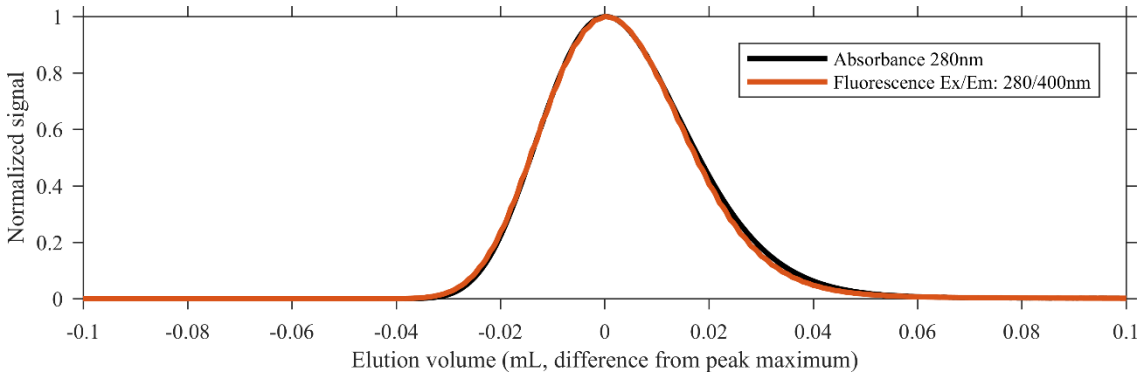

Figure S2: Visual representation of the alignment between absorbance (black line) and fluorescence detector (orange line) using an injection of Tryptophan without an analytical column. Both detectors have been aligned by the subtraction of the dead volume from the elution volume of the fluorescence detector (0.06 mL) and centered so that the peak maximum is located at an elution volume of 0 mL for both detectors.

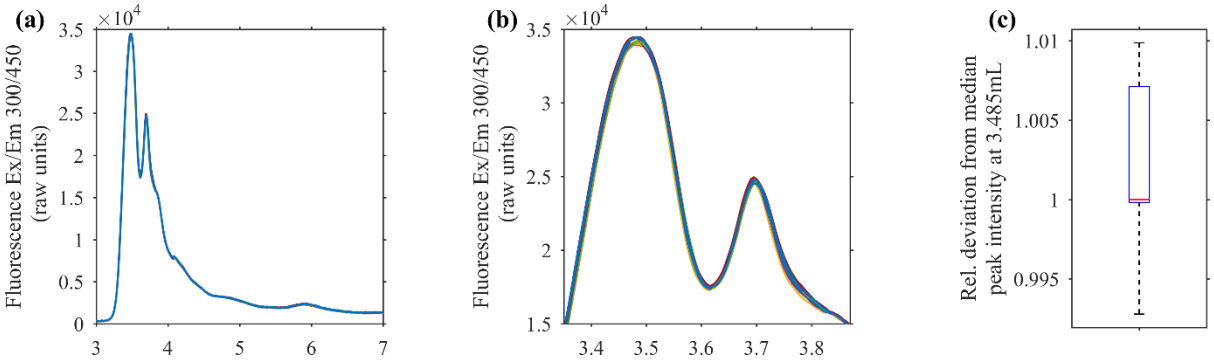

Figure S3: 12 Chromatograms of a reference DOM sample at excitation 300 nm, emission 450 nm (a). (b) focusses on the chromatogram between elution volumes of 3.3 and 3.9 mL (main peak). A boxplot of the relative deviation of raw peak intensities from the median raw peak intensity (peak intensity divided by median peak intensity) of the main peak at 3.485 mL in shown in (c). Chromatograms were measured in a time span of 172 hours (7 days and 4 hours, ~14 hours between individual runs) and were approximately equally spaced across this time.

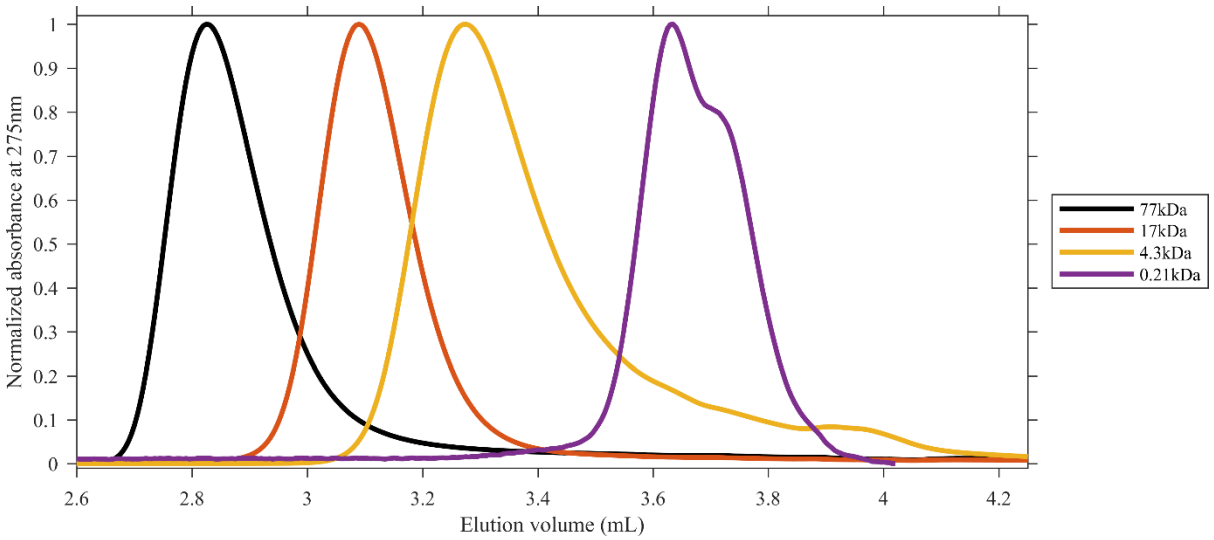

Figure S4: Chromatograms of the molecular size calibrants (Polystyrene sulfonate). Chromatograms have been normalized to the maximum peak intensity at 275 nm. Log_{10}(molecular size in Da) and elution volume (in mL) correlated with an R^2 of 0.987. Regression coefficients were -0.31 and 4.38 for slope and intercept, respectively.

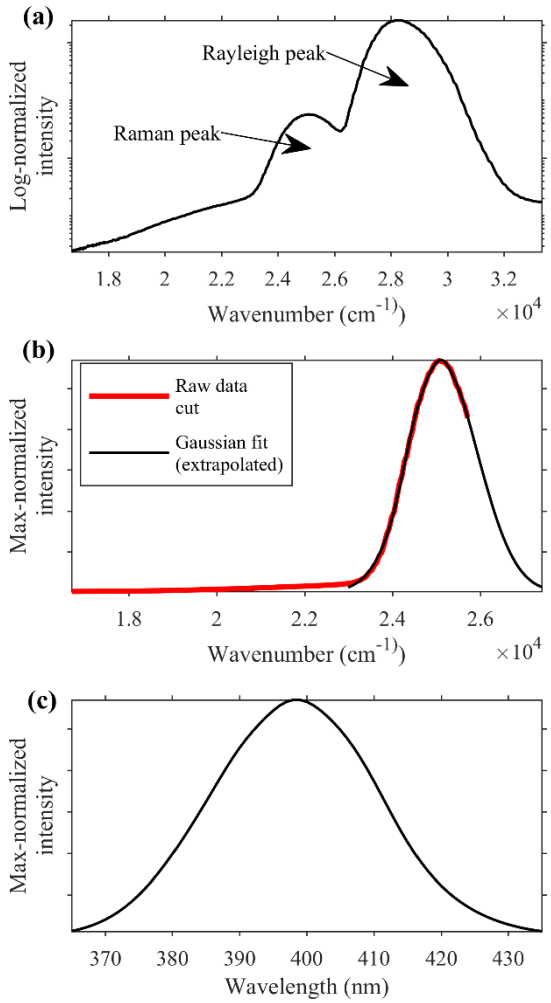

Figure S5: Processing of Emission scans at Excitation 350 nm for the extraction of Raman areas. Raw emission scans at excitation 350 nm contain overlapping Rayleigh and Raman scatter (a). In order to obtain representative Raman areas, the Rayleigh affected area of the Raman peak was deleted (red line in (b)). This data was then used to fit a four-term Gaussian using wavenumbers to ensure the proper Gaussian peak shapes (black line in (b)).The final extrapolated Raman peak (scaled to wavelengths) is shown in (c).

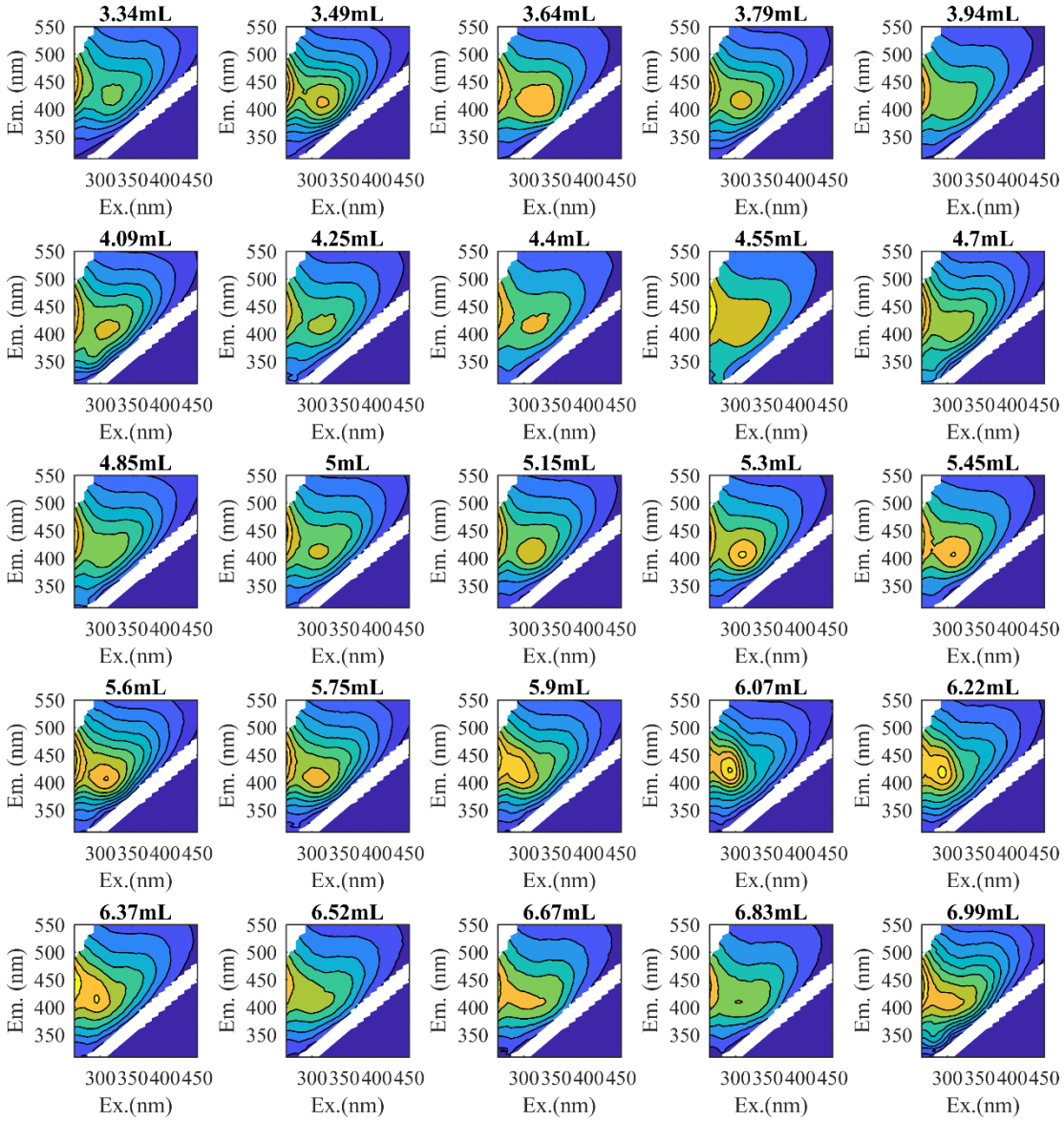

Figure S6: Example of 25 (of 1500) size separated EEMs obtained with the HPSEC system for the Pony Lake XAD-8 DOM extract. Plot titles represent the elution volume of the respective EEM.

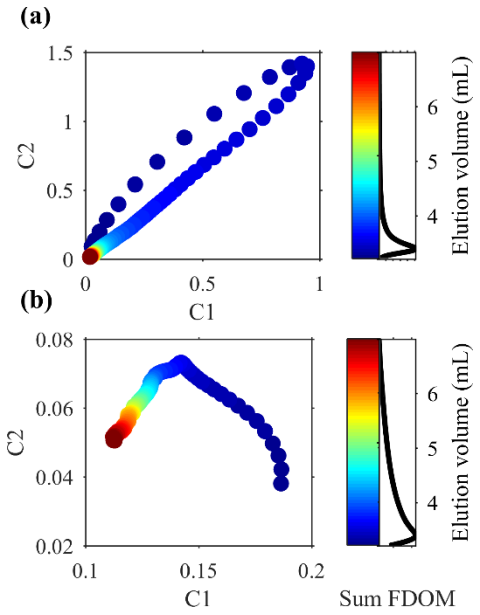

Figure S7: Comparison of the correlation between two components derived by PARAFAC modelling. Scores in (a) were obtained using a Raman normalized dataset, scores in (b) were obtained from the same dataset that was normalized by applying $\log_{10}$ to EEMs prior to modeling. To facilitate the chromatographic location of individual scores in the HPSEC chromatogram, the total fluorescence chromatogram is plotted vertically along the colorbar, representing the elution volume.

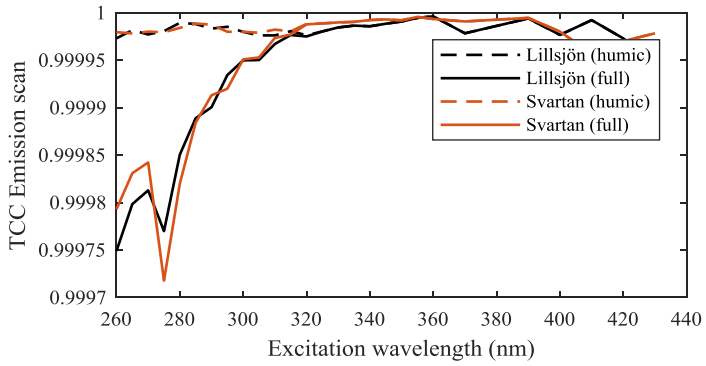

Figure S8: Tucker congruence coefficient (TCC) between the sum of size-separated EEMs and its corresponding bulk EEM across all measured excitation wavelengths. Dashed lines represent the TCCs for only the humic fluorescence emission (400-550 nm), solid lines represent the TCC of the full fluorescence emission (330-550 nm).

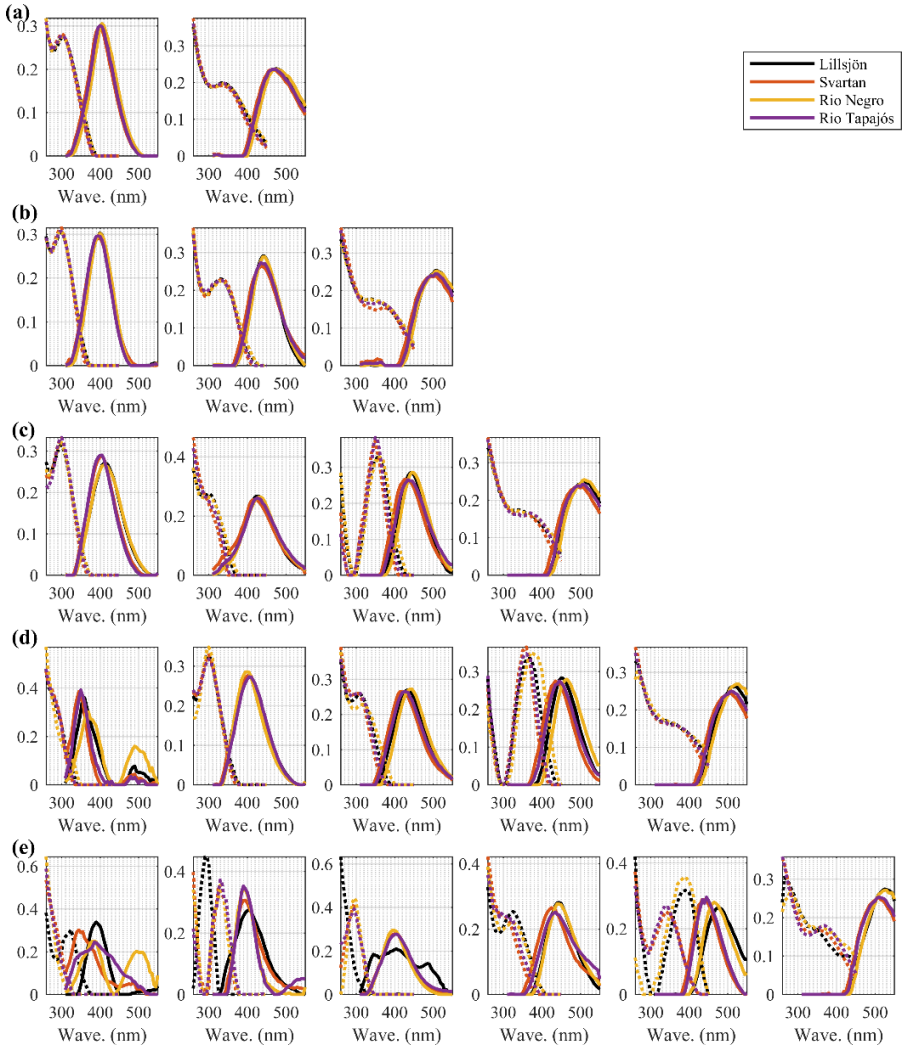

Figure S9: Spectral loadings of PARAFAC models with two to six components (a-d). Different colors represent the spectral loadings of components obtained from the four different freshwater samples. Components are ordered by increasing emission maximum to facilitate comparisons, rather than the model contribution. Dashed lines represent the excitation spectra, filled lines the emission spectra. Please note that model loadings in panel (e) depict models that could not be validated.

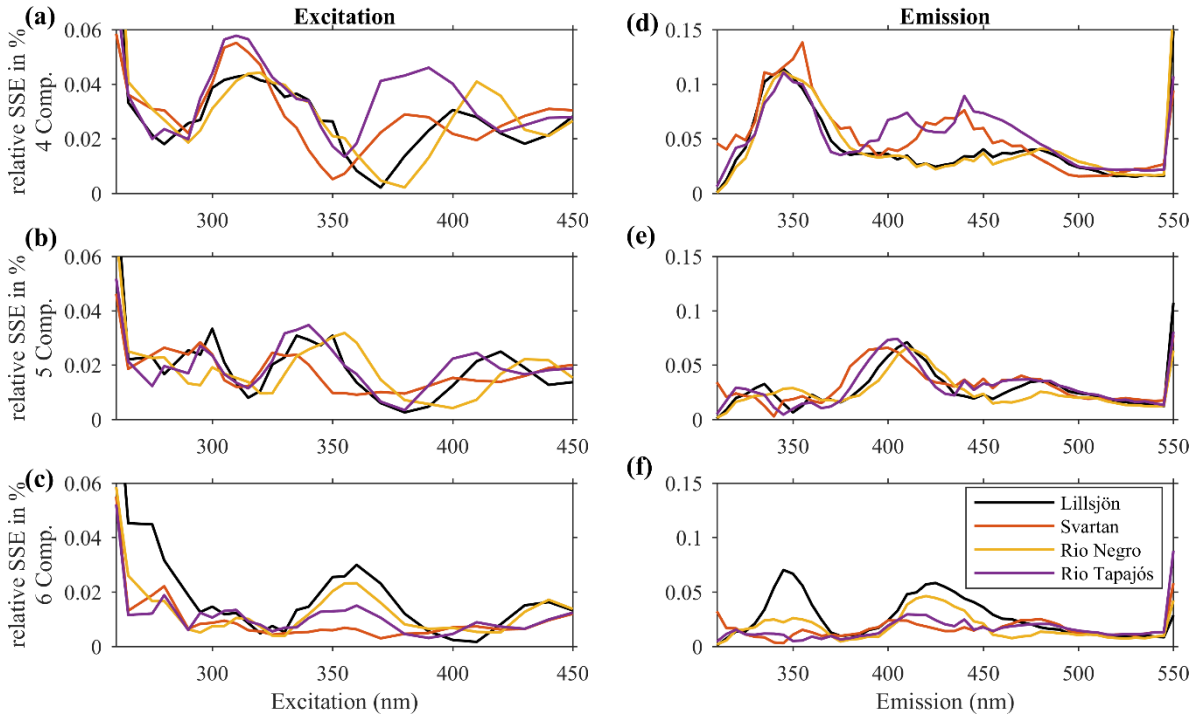

Figure S10: Sum of squared error (SSE) overview for four [top panel, (a) and (d)], five [middle panel, (b) and (e)], and six [bottom panel, (c) and (f)] component PARAFAC models of four freshwater DOM samples (colored differently according to the provided legend). The left panel depicts the SSE for excitation loadings, emission SSE's are shown in the right panel. All SSEs were divided by the maximum of the squared fluorescence for every sample to account for different overall sample fluorescence and are thus given as percent. The middle panel detailing the 5 component model depicts SSEs of the model shown in the main manuscript. The scale of y-axis in was set equal to demonstrate the development from four (top) to six (bottom) model components. Please note that squared residuals are always positive and hence combine negative and positive residuals across all samples. Moreover (c) and (f) depict SEEs for models that could not be validated using common criteria.

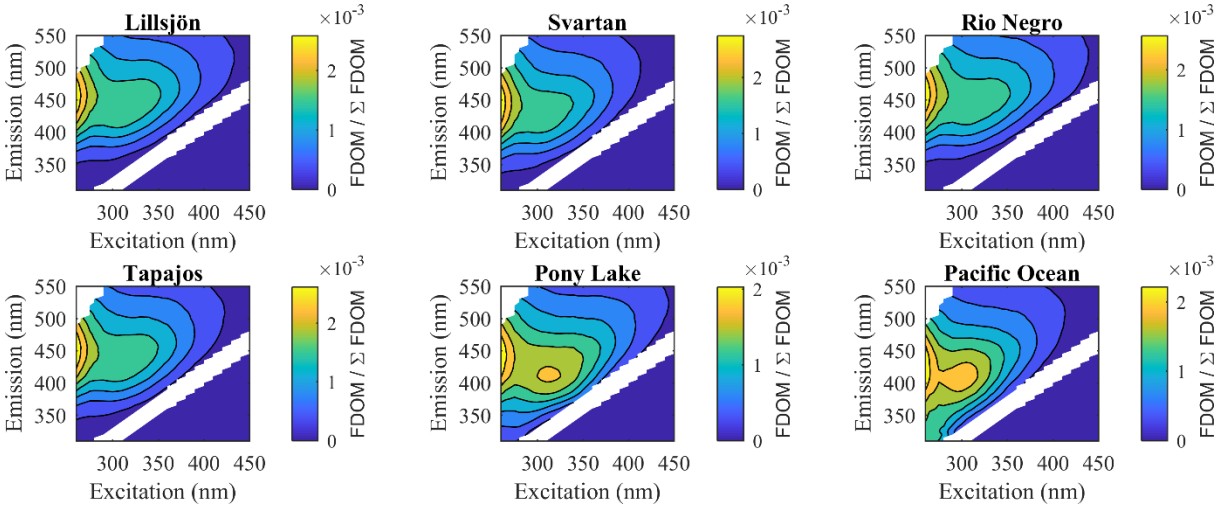

Figure S11: Bulk EEM of all six DOM samples. EEMs were produced by summing up all HPSEC-EEMs and normalized by the sum of fluorescence.

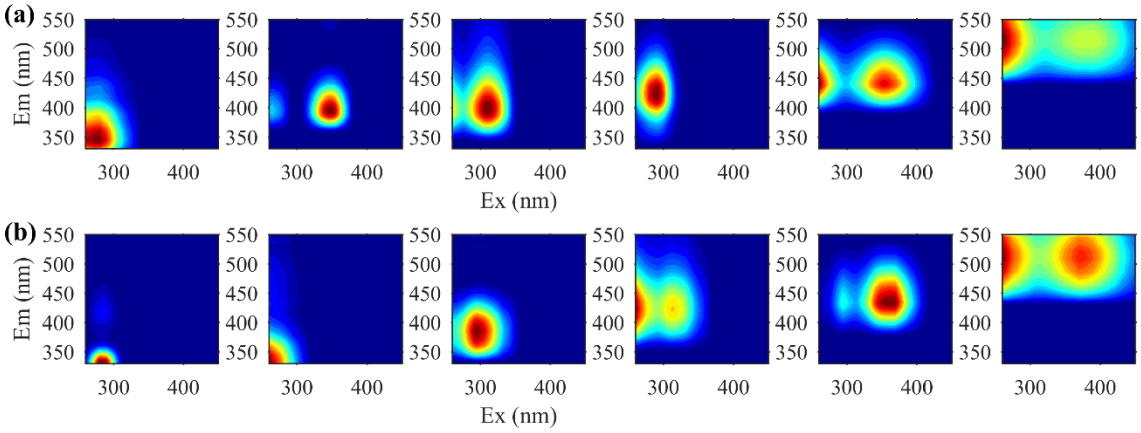

Figure S12: Contour plots of six PARAFAC components describing HPSEC EEMs from Pony Lake (a) and the Pacific Ocean sample (b). Model statistics for (a) and (b), respectively: Core consistency: 5 and 3 %, percent explained variance: 99.94 and 99.94 %, Sum of Squares Error (SSE): 0.0084 and 0.0175.

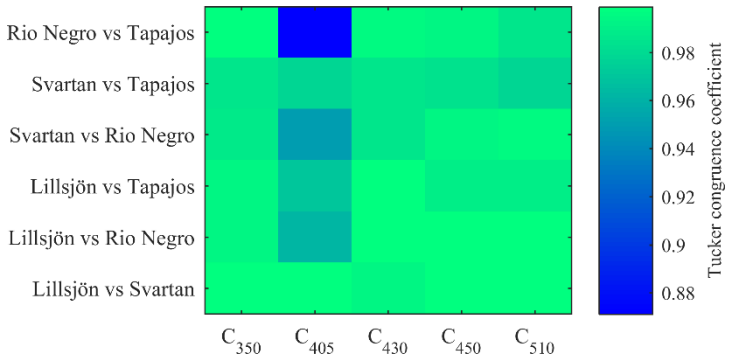

Figure S13: Heatmap of Tucker congruence coefficients showing the similarity between molecular size distributions of PARAFAC components in all allochthonous samples.

S3: Supplementary Tables

Table S1: Information on samples used in this study.

Sample origin	Latitude	Longitude	Sampling time	Extraction resin	Description
Rio Negro	-2.613	-60.942	4/2013	PPL	Black water river in the Amazon basin, surface sample
Rio Tapajós	-2.351	-54.769	4/2013	PPL	Clear water river in the Amazon basin, surface sample
Svartan River	58.440	15.457	2/2012	PPL	Boreal river, surface sample
Lake Lillsjön	58.659	16.144	6/2011	PPL	Boreal lake, sampling depth 7 m

Pacific Ocean[11]	Unspecified, 170 km south of Honolulu, Hawaii		2/1986	XAD-8	240 m depth, no terrestrial OM sources
Pony Lake[3]	-77.55°	166.15	1/2006	XAD-8	Antarctic lake, no terrestrial OM sources, no depth information available

Table S2: Tucker congruence between PARAFAC components from Lake Lillsjön FDOM, compared to Svartan River, Rio Negro, and Rio Tapajós.

Component	Lillsjön x Svartan	Lillsjön x Rio Negro	Lillsjön x Tapajós
C_{350}	0.85	0.94	0.95
C_{405}	0.99	0.99	1.00
C_{430}	0.97	0.98	0.99
C_{450}	0.90	0.93	0.97
C_{510}	0.98	1.00	0.99

S4: Supplementary Information References

(1) Timko, S. A. Photochemistry of Dissolved Organic Matter: Reactivity and application in constructed treatment wetlands, University of California, Irvine, 2015.

(2) Dittmar, T.; Koch, B. P.; Hertkorn, N.; Kattner, G. A simple and efficient method for the solid-phase extraction of dissolved organic matter (SPE-DOM) from seawater. *Limnol. Ocean. Methods* **2008**, *6*, 230–235.

(3) Cawley, K. M.; McKnight, D. M.; Miller, P.; Cory, R.; Fimmen, R. L.; Guerard, J.; Dieser, M.; Jaros, C.; Chin, Y.-P.; Foreman, C. Characterization of fulvic acid fractions of dissolved organic matter during ice-out in a hyper-eutrophic, coastal pond in Antarctica. *Environ. Res. Lett.* **2013**, *8* (4), 1–10.

(4) Thurman, E. M.; Malcolm, R. L. Preparative isolation of aquatic humic substances. *Environ. Sci. Technol.* **1981**, *15* (4), 463–466.

(5) Wünsch, U. J.; Murphy, K. R.; Stedmon, C. A. Fluorescence Quantum Yields of Natural Organic Matter and Organic Compounds: Implications for the Fluorescence-based Interpretation of Organic Matter Composition. *Front. Mar. Sci.* **2015**, *2* (November), 1–15.

(6) Helms, J. R.; Stubbins, A.; Ritchie, J. D.; Minor, E. C.; Kieber, D. J.; Mopper, K.

Absorption spectral slopes and slope ratios as indicators of molecular weight, source, and photobleaching of chromophoric dissolved organic matter. *Limonology Oceanogr.* **2008**, *53* (3), 955–969.

(7) Huang, H.; Sawade, E.; Cook, D.; Chow, C. W. K.; Drikas, M.; Jin, B. High-performance size exclusion chromatography with a multi-wavelength absorbance detector study on dissolved organic matter characterisation along a water distribution system. *J. Environ. Sci. (China)* **2015**, *44*, 235–243.

(8) Allpike, B. P.; Heitz, A.; Joll, C. A.; Kagi, R. Size Exclusion Chromatography To Characterize DOC Removal in Drinking Water Treatment. *Environ. Sci. Technol.* **2005**, *39* (7), 2334–2342.

(9) Sandron, S.; Rojas, A.; Wilson, R.; Davies, N. W.; Haddad, P. R.; Shellie, R. A.; Nesterenko, P. N.; Kelleher, B. P.; Paull, B. Chromatographic methods for the isolation, separation and characterisation of dissolved organic matter. *Environ. Sci. Process. Impacts* **2015**, *17* (9), 1531–1567.

(10) Murphy, K. R.; Stedmon, C. A.; Graeber, D.; Bro, R. Fluorescence spectroscopy and multi-way techniques. PARAFAC. *Anal. Methods* **2013**, *5* (23), 6557.

(11) Malcolm, R. L. The uniqueness of humic substances in each of soil, stream and marine

environments. *Anal. Chim. Acta* **1990**, *232* (C), 19–30.