Electronic Supplementary Material (ESI) for Analytical Methods.

Supplementary Information to:

Emerging patterns in the global distribution of dissolved organic matter fluorescence

Urban J. Wünsch,*[a] Rasmus Bro, [b] Colin Stedmon,[c] Philip Wenig,[d] and Kathleen R. Murphy [a]

a. Chalmers University of Technology, Architecture and Civil Engineering, Water Environment Technology, Sven Hultins Gata 6, 41296 Gothenburg, Sweden

b. University of Copenhagen, Dept. Food Science, 1958 Frederiksberg, Denmark

c. National Institute of Aquatic Resources, Technical University of Denmark, Kemitorvet, 2800 Kgs. Lyngby, Denmark.

d. Lablicate GmbH, Martin-Luther-King Platz 6, 20146 Hamburg, Germany

* Corresponding author: wuensch@chalmers.se

This Supporting Information contains 9 pages and 6 figures.

Supplementary Information contents:

S1: Extraction of OpenFluor component spectra

The 102 database entries stored in OpenFluor were extracted from the website on October 29th, 2018 (`http://openfluor.org`). Emission and excitation spectra were subsequently processed in Matlab (v.9.5, MathWorks Inc.). To compare spectra, the components of all models were interpolated to increments of one nanometre and cut to ranges of 270 -440 nm and 300 -540 nm for excitation and emission, respectively. In some cases, missing numbers (wavelengths not covered in a particular study) were replaced either with zeros for components where no spectral features were present in the missing range or estimates from gaussian fits were used in cases where spectral features were detected on the edges of measured data. In cases where missing wavelengths occurred at wavelengths below the captured spectral range, monotonic extrapolations were used to estimate missing wavelengths. Models with excessive amounts of missing wavelengths were excluded from further analysis (first excitation > 280 nm, first emission > 330 nm, last emission <490 nm). Five models were excluded for such reasons (Søndergaard et al. 2003; Cawley et al. 2012a; b; Bianchi et al. 2014; Tanaka et al. 2014), while an additional six were excluded for non-DOM sample source [data sets containing pure organic compound fluorescence (Wünsch et al. 2015)] or highly autochthonous character [four models of Antarctic ice cores dominated by ultraviolet wavelength range fluorescence (D'Andrilli et al. 2017). All remaining spectra were smoothed using a Savitzky-Golay filter with a window length of 21 nm and a 2nd order polynomial (Savitzky and Golay 1964; Steinier et al. 1972). Spectra were subsequently normalized by division with their Frobenius norm to account for the new, equal wavelength increments. Figure S1 depicts an example of raw and fully processed spectra. The resulting dataset included 90 models and 478 emission and excitation spectra. A list of all models and their corresponding publications is available at `https://openfluor.lablicate.com/of/measurement`.

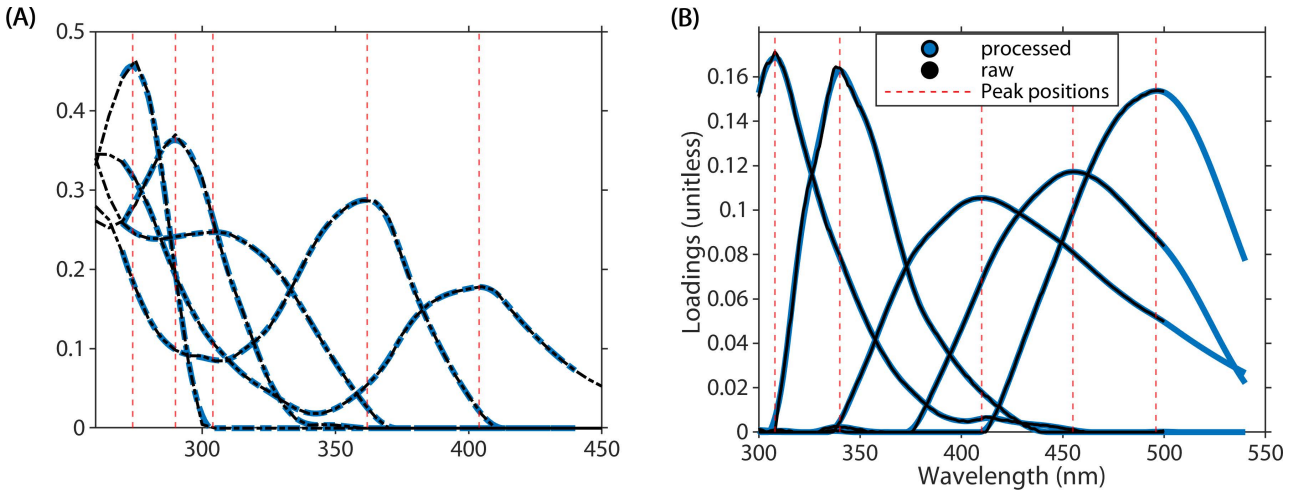

Figure S1: Example of raw and processed OpenFluor PARAFAC spectra (Yamashita et al. 2013)**. (A):** Excitation spectra **(B)**: Emission spectra.

While it is common to report multiple excitation maxima for PARAFAC components (Stedmon et al. 2003), we aimed to calculate the Stokes shift of spectra and thus determined the wavelength at which the maximum fluorescence emission and first fluorescence excitation occurred (Reynolds 2014). The Stokes shift was then calculated as the difference between first excitation (when plotted against wavenumber, i.e. the "last" peak when plotted against wavelength) and emission peak in wavenumber and expressed in electron volts (eV). Fig. S1 shows examples of determined peak positions.

S2 Sensitivity analysis of the shift- and shape sensitive congruence

The Tucker congruence coefficient (TCC) was modified to increase its sensitivity to shape differences and peak shifts. The shift- and shape-sensitive congruence (SSC) was based on the classic definition of TCC (Tucker 1951; Lorenzo-Seva and ten Berge 2006):

$$TCC\ (x,y) = \frac{\sum x\, y}{\sqrt{\sum x^2 \sum y^2}}, \text{ (1)}$$

where *x* and *y* are loadings of two factors with identical x-axis-scale. TCCs are traditionally calculated for emission and excitation spectra (TCC_{ex} and TCC_{em}) to form the overall $TCC_{em\ x\ ex}$ (Murphy et al. 2014). While excitation spectra of chromophores are typically very distinct and respond to changes in molecular structure sensitively, emission spectra do not have similarly unique features and vary less in comparison (Wünsch et al. 2015). To quantify the divergence of the two emission spectra *x* and *y* with regards to peak wavelength differences, the penalty term α was defined as:

$$\alpha(x,y) = \frac{|\lambda_x - \lambda_y|}{\lambda_{max} - \lambda_{min}}, \text{ (2)}$$

where $|\lambda_x - \lambda_y|$ is the difference between peak positions of spectrum *x* and *y* in nanometres and $\lambda_{max} - \lambda_{min}$ represents the range of observed wavelengths in nanometres. For example, a 10 nm difference between peaks of spectrum x and y when x and y were measured between 300 and 550 nm results α = 0.04. Moreover, SSC was introduced to allow a more sensitive quantification of peak area differences. We defined the penalty term β, which quantifies the differences between the areas of two peaks as follows:

$$\beta(x,y) = \frac{|\int x - \int y|}{\lambda_{max} - \lambda_{min}}, \text{ (3)}$$

where $|\int x - \int y|$ is difference between the integrals of spectra *x* and *y* normalized by their uniform norm. SSC was then calculated based on the subtraction of the two penalty terms β and α from TCC:

$$SSC\ (x,y) = TCC - \sum \alpha, \beta. \text{ (4)}$$

The penalty terms α and β are used to sensitize Tuckers congruence coefficient (TCC) to the occurrence of peak shifts and shape differences between compared spectra. To quantify how the resulting shift / shape sensitive congruence differs from TCC, TCCs between a reference spectrum and modified spectra were compared to TCC- α and TCC- β (Fig. S2). The result of our analysis indicated that SSCs are significantly more sensitive towards shifts in peak wavelength and changes in the broadness between peaks since the band of highly similar spectra identified by TCC compared to TCC-α or TCC- β was narrower (Fig. S2B, D). Importantly, α and β behaved predictively (Fig. S2A, C).

Next, TCC and SSC were calculated for a set of 14 emission spectra of pure organic substances (Wünsch et al. 2015). To perform a sensitivity analysis, TCCs and SSCs between all unique comparisons (N = 273) were calculated. Since no two spectra are identical, any quantification of spectral similarity suggesting identity of two spectra presents a type I error (false positive).

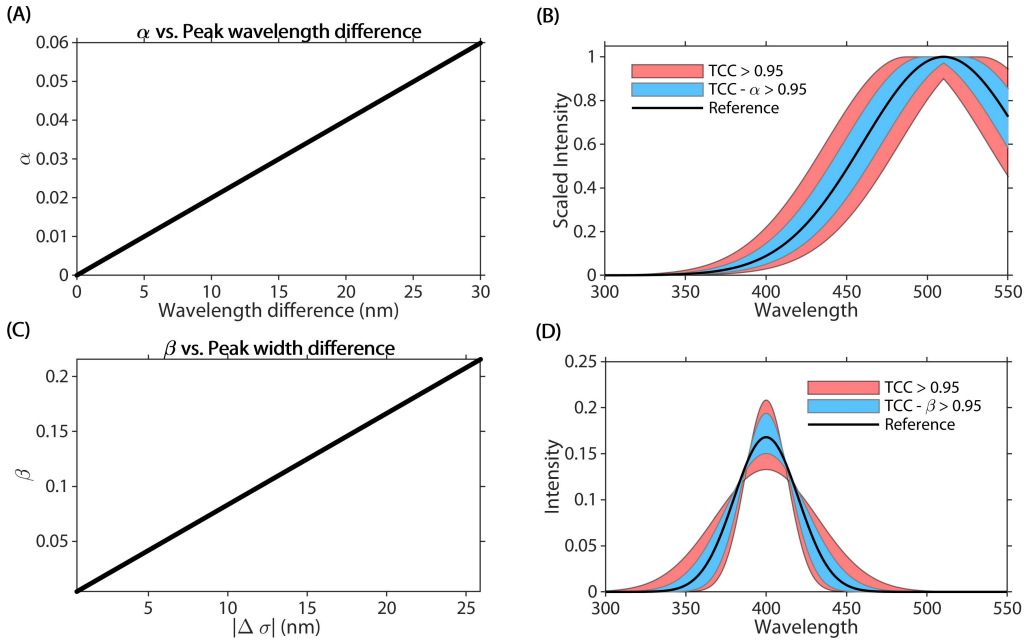

Figure S2: Assignment of spectral identity by Tuckers congruence and shift- and shape sensitive congruence. (A): Response of shift-penalty α to divergence in peak positions without shape changes. **(B):** Band of shifted peaks classified as highly similar according to Tuckers congruence (red) and TCC- α (blue). **(C)**: Response of the shape penalty β to shape changes (difference between peak widths σ). **(D)**: Band of peaks with varying peak broadness classified as highly similar by TCC (red) and TCC-β (blue).

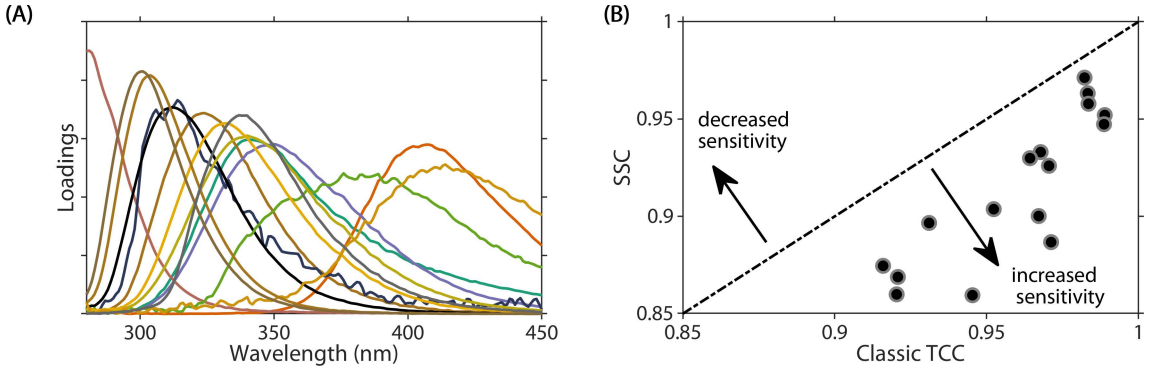

Figure S3: Tuckers congruence and shift / shape sensitive similarity between emission spectra of pure organic compounds. (A): Emission spectra of 14 organic compounds. (B): Comparison of Tuckers congruence (TCC) between all unique comparisons and the Shape / Shift sensitive congruence (SSC). The dashed line represents the line of equal sensitivity. Notations of increased or decreased sensitivity refer to SSC in comparison to TCC.

Figure S3 shows the comparison between TCC and SCC. In all comparisons, SSCs were lower compared to corresponding TCCs. Assuming a threshold of 0.95 as evidence for sufficient spectral similarity to assume identity of spectra, TCC classified 12.1% of spectra as indistinguishable, while the SSC only classified 4.4 % as indistinguishable. Considering all unique comparisons, the SSC reduced the rate of type I errors by 63 %. This demonstrates that the SSC is more sensitive than TCC to spectral differences.

S3: Peak position of under-specified PARAFAC models

To document the influence of under-specification (fitting too few components) in PARAFAC models describing DOM fluorescence, a dataset describing the molecular-size distribution of fluorescent DOM in Pony Lake (Antarctica, International Humic Substances Society standard 1R109F) was investigated. The particular dataset was generated from a chromatographic separation of a single sample, producing 252 EEMs from which a six-component model was developed and validated (Wünsch et al. 2017). Here, we investigated how the peak position of the component with the longest fluorescence emission would be impacted by assuming the presence of fewer components. Figure S4 shows how the peak position of the longest-emitting component exhibits emission maxima from ~510 nm in the validated six-component model. When fitting fewer components, the final emission peak occurs at shorter and shorter wavelengths until it occurs at 465 nm in the two-component model. The apparent change in emission maximum wavelength can be explained by spectral averaging of multiple peaks that occurs when PARAFAC models are under-specified.

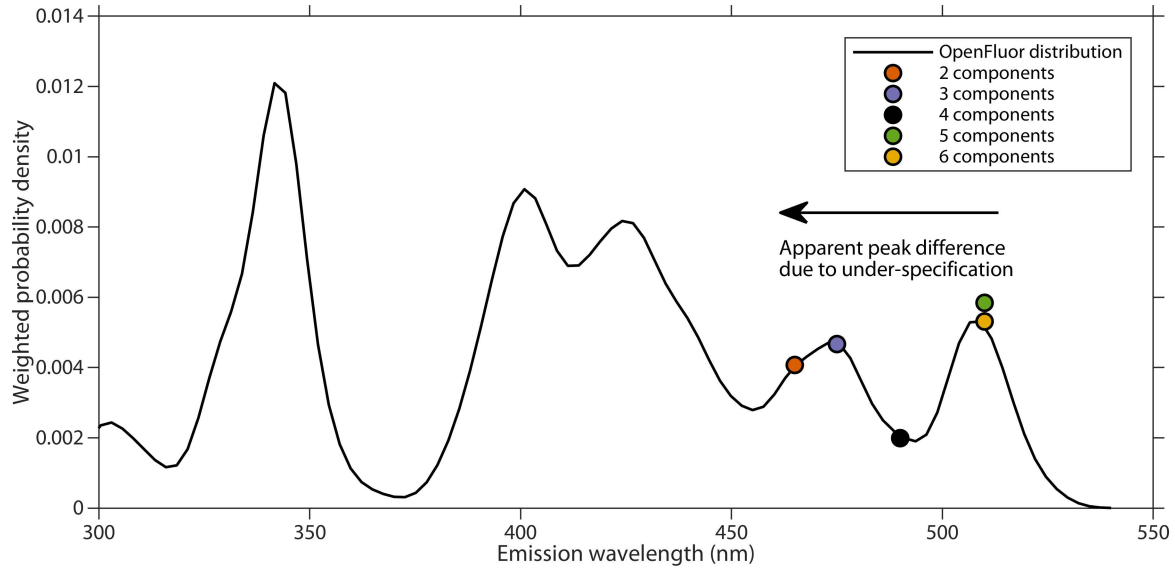

Figure S4: Effect of under-specification on the longest emitting PARAFAC component. The black line shows the weighted probability distribution of emission peaks in the OpenFluor database. Superimposed are the emission positions of the PARAFAC component with the longest emission maximum in two- to six component models of an EEM dataset from Pony Lake (Antarctica, IHSS standard 1R109F) (Wünsch et al. 2017). When fewer components are fitted, the peak maximum of the longest-emitting component moves to shorter wavelengths.

S4: Meta-analysis of component- model similarity

To investigate the primary factors responsible for model similarity, a meta-analysis of matching models was carried out. First, the best matching model for all 90 entries analysed in this study was identified. Secondly, four meta-variables (primary sample source, number of components, instrument, and number of components with emission maxima > 400 nm) were compared for each matched model. For all comparisons and meta-variables, the results were plotted in heatmaps (Fig. S5). If a particular meta-variable is a strong predictor for the similarity between models, high counts should be visible along the diagonal in the heatmap. On the other hand, scatter would indicate that a particular meta-variable is not a strong predictor of the similarity between models.

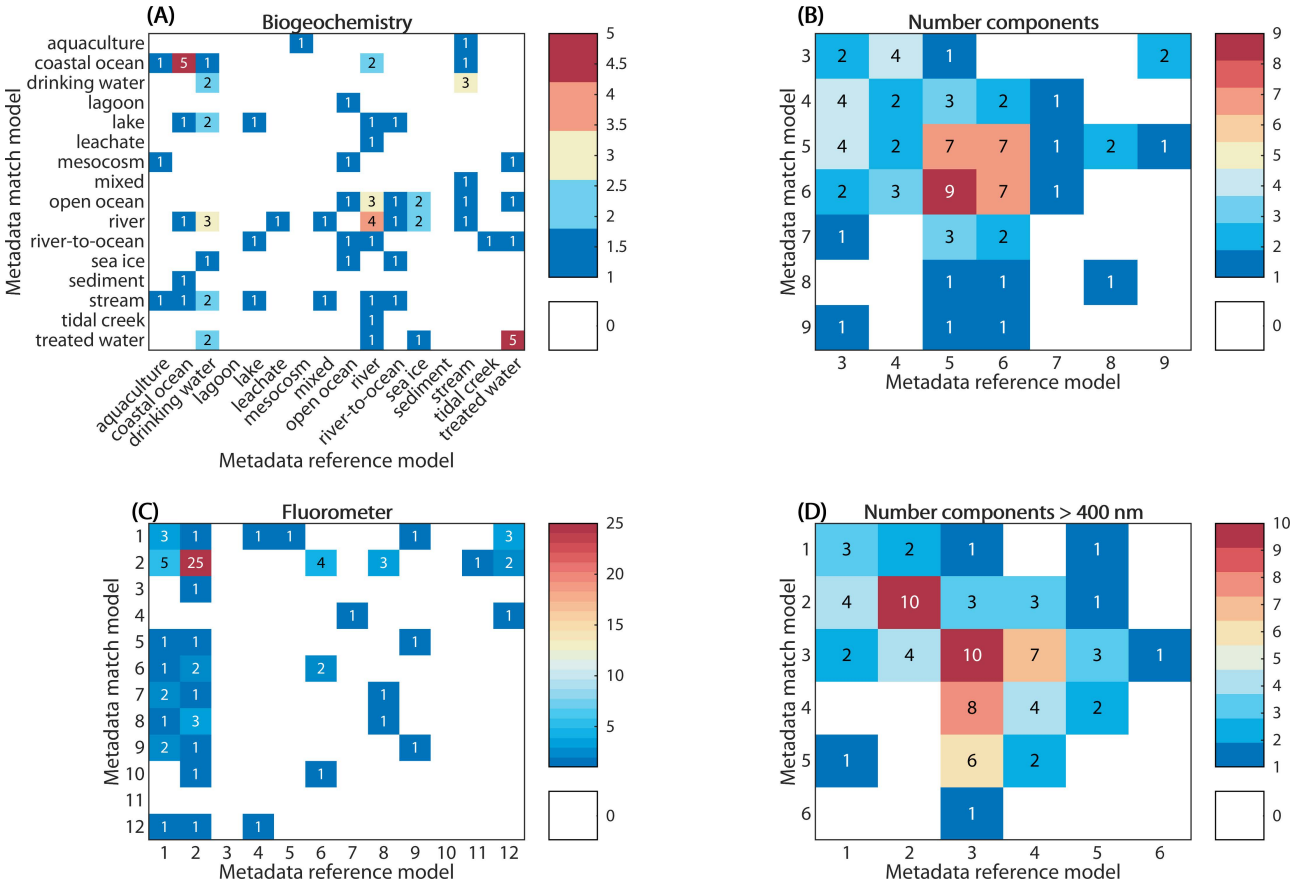

Figure S5: Meta-analysis of OpenFluor PARAFAC model similarity. The four plots show number of most-similar models as a function of primary sample source (A), number of components in the models (B), fluorometer (C), and number of components with emission maxima > 400 nm (D). In (C), variables were anonymized to avoid associations with commercial products.

The meta-analysis of the PARAFAC models in the OpenFluor database revealed that the similarities between models appears to be driven by the total number of PARAFAC components in the model (Fig. S5B), especially the number of visible-wavelength components with emission maxima > 400 nm (Fig. S5D). On the other hand, the high similarity between e.g. riverine or treated water studies (Fig. S5A) were most likely driven my methodological similarities (multiple models derived in a single study). Patterns with regards to the fluorometer used in PARAFAC studies may only reflect the fact that most studies used one of two popular fluorometers (Fig. S5C).

With regards to the overall number of components, the following patterns were observed (Fig. S5B, D). There was a tendency for models with similar numbers of components to be spectrally well matched, since a total of 30 models with five or six components best matched models with the same

number of components. Moreover, a similar trend was observed when the number of visible-wavelength components was considered (Fig. S5D). Models with the same or similar number of components emitting light at > 400 nm often were best matched with each other.

To investigate the occurrence of source-specific components, all 478 components were grouped according to the primary sample source (excluding source categories containing less than three models, Fig. S6). Source-specific components could not be visually identified. Rather revealed similar distributions across samples sources. For example, Open Ocean models showed similar component distributions compared to models describing treated (fresh)water.

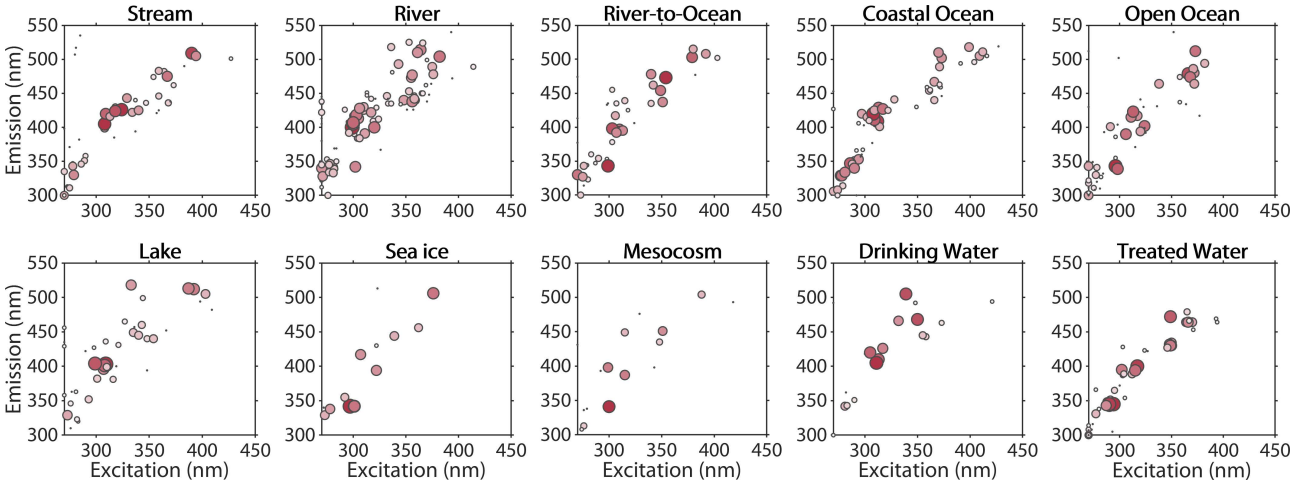

Figure S6: Peak positions of OpenFluor components by primary sample source. Dots are coloured and sized by the frequency they match other components in the database (TCC_{Ex} > 0.95, SSC_{Em} > 0.95).

S5: Supplementary Information References

Bianchi, T. S., C. Osburn, M. R. Shields, S. Yvon-Lewis, J. Young, L. Guo, and Z. Zhou. 2014. Deepwater horizon oil in Gulf of Mexico waters after 2 years: Transformation into the dissolved organic matter pool. Environ. Sci. Technol. **48**: 9288–9297.

Cawley, K., Y. Ding, J. Fourqurean, and R. Jaffe. 2012a. Characterising the sources and fate of dissolved organic matter in Shark Bay, Australia: A preliminary study using optical properties and stable carbon isotopes. Mar. Freshw. Res. **63**: 1098–1107.

Cawley, K. M., K. D. Butler, G. R. Aiken, L. G. Larsen, T. G. Huntington, and D. M. McKnight. 2012b. Identifying fluorescent pulp mill effluent in the Gulf of Maine and its watershed. Mar. Pollut. Bull. **64**: 1678–1687.

D'Andrilli, J., C. M. Foreman, M. Sigl, J. C. Priscu, and J. R. McConnell. 2017. A 21000-year record of fluorescent organic matter markers in the WAIS Divide ice core. Clim. Past **13**: 533–544.

Lorenzo-Seva, U., and J. M. F. ten Berge. 2006. Tucker's congruence coefficient as a meaningful index of factor similarity. Methodology **2**: 57–64.

Murphy, K. R., C. A. Stedmon, P. Wenig, and R. Bro. 2014. OpenFluor- an online spectral library of auto-fluorescence by organic compounds in the environment. Anal. Methods **6**: 658–661.

Reynolds, D. M. 2014. The Principles of Fluorescence, p. 3–34. *In* P. Coble, J. Lead, A. Baker, D.M. Reynolds, and R.G.M. Spencer [eds.], Aquatic Organic Matter Fluorescence. Cambridge University Press.

Savitzky, A., and M. J. E. Golay. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. **36**: 1627–1639.

Søndergaard, M., C. A. Stedmon, and N. H. Borch. 2003. Fate of terrigenous dissolved organic matter (DOM) in estuaries: Aggregation and bioavailability. Ophelia **57**: 161–176.

Stedmon, C. A., S. Markager, and R. Bro. 2003. Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy. Mar. Chem. **82**: 239–254.

Steinier, J., Y. Termonia, and J. Deltour. 1972. Comments on Smoothing and Differentiation of Data by Simplified Least Square Procedure. Anal. Chem. **44**: 1906–1909.

Tanaka, K., K. Kuma, K. Hamasaki, and Y. Yamashita. 2014. Accumulation of humic-like fluorescent dissolved organic matter in the Japan Sea. Sci. Rep. **4**: 1–7.

Tucker, L. R. 1951. A method for synthesis of factor analysis studies, *In* Personell Research Section Report No. 984. Department of the Army.

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

Wünsch, U. J., K. R. Murphy, and C. A. Stedmon. 2017. The one-sample PARAFAC approach reveals molecular size distributions of fluorescent components in dissolved organic matter. Environ. Sci. Technol. **51**: 11900–11908.

Yamashita, Y., J. N. Boyer, and R. Jaffé. 2013. Evaluating the distribution of terrestrial dissolved organic matter in a complex coastal ecosystem using fluorescence spectroscopy. Cont. Shelf Res. **66**: 136–144.