A global fluorescence index for predicting dissolved organic carbon abundance and aromaticity

Kathleen R. Murphy[1,2*]

.

[1]Department of Architecture and Civil Engineering, Chalmers University of Technology; Gothenburg, 41298 Sweden.

[2]Department of Building and Environmental Technology, Lund University; Lund, 22363 Sweden.

ABSTRACT The chemical structures of dissolved organic compounds in natural waters, including degree of aromaticity, affect their physical, chemical, and biological properties and ultimately the fate of carbon in aquatic systems and during water treatment. Herein, a new fluorescence-based aromaticity index named ARIX is shown to link the composition of aquatic dissolved organic matter to its relative aromaticity across diverse aquatic systems. ARIX predicts SUVA, a widely used proxy of aromaticity, as well as the percentage of polycyclic aromatic and polyphenolic molecular formulas determined by FT-ICR MS and the ratio of "humic substances" to "building blocks" fractions determined by LC-OCD. A meta-analysis of nine datasets spanning all seven continents suggests that a single linear relationship predicts SUVA from ARIX in both bulk and extracted DOM. In impounded waterbodies exhibiting decoupling between DOC and absorbance linked to extensive biogeochemical processing, linearity can be restored by accounting for the interaction between absorbance and ARIX. The results deliver new insights on widely discussed trends in DOM optical properties and on the molecular structures underlying optical measurements in aquatic milieu. They further pave the way for new optical technologies capable of real-time predictions of DOC concentration, reactivity and fate.

Keywords: DOM, CDOM, surface waters, reactivity, connectivity, remote sensing, sensors

SYNOPSIS: Algorithms are presented for predicting dissolved organic carbon relative aromaticity and concentration from only spectroscopic measurements.

1. INTRODUCTION

Predicting the fate of dissolved organic matter (DOM) in aquatic systems requires the ability to detect changes in the chemical composition of DOM (*1, 2*). DOM consists of potentially thousands of compounds of varying age and structural complexity including those derived from the degradation of biomass as well as compounds released as by-products of metabolism or chemical processes (*3*). In river systems the molecular characteristics of DOM affect ecosystem health (*4*) and determine whether DOM will leave the water column via biological or photochemical mineralization, flocculation/precipitation or adsorption, or be transported downstream and stored in the deep sea (*2, 5*).

DOM aromaticity is widely studied due to its influence on wide-ranging chemical and biological processes in water. Aromatic compounds resist degradation due to their stable conjugated π-electron systems; when combined with the continuous export of aromatic DOM from land this contributes to the overall carbon storage of aquatic systems over long timescales (*1, 6*). Aromatic compounds control primary production by attenuating light underwater and by binding and retaining nutrients within their molecular structures (*7, 8*). During drinking water treatment, large polyaromatic compounds are more susceptible to flocculation (*9, 10*) and compete more effectively with micropollutants for sites on adsorption filters, leading to their premature saturation (*11, 12*). The selective removal of polyphenolic and other compounds during water treatment further affects the abundance, types, and toxicity of disinfection by-products formed during subsequent reactions with chlorine (*13*).

DOM relative aromaticity, referring to the proportion of carbon atoms associated with aromatic bonds, is a bulk property of the pool of molecules that comprise DOM (*14*). Relative aromaticity determined using Carbon-13 nuclear magnetic resonance, ^{13}C-NMR, is strongly

correlated to the ratio of UV absorption at 254 nm normalized to DOC concentration, termed specific UV absorbance or commonly SUVA (*14, 15*). In natural waters containing low concentrations of dissolved iron, SUVA typically spans the range 1 - 6 $m^2g^{-1}C$ (*16, 17*). Since measuring ^{13}C-NMR aromaticity is relatively impractical, SUVA is routinely measured in its place. However, SUVA has technical shortcomings due to requiring two different instruments (a carbon analyser and a spectrophotometer), which starkly increases measurement costs and negatively affects immediacy, accuracy and precision (*16, 18*).

Fluorescence spectroscopy is widely used to study DOM composition and three fluorescence indices based on simple emission ratios are also widely used proxies of aromaticity (*15*): the "fluorescence index" (FI or FIX), "freshness" or "biological index" (β/α and BIX), and "humification index" HIX (*19-22*). However, their sensitivities to differences in aromaticity are limited (*15, 16*) and a theoretical basis underpinning their variability, and demonstrated links to specific DOM molecular structures, are missing (*15, 23*). Recently, a multispectral fluorescence index ("PARIX") derived using parallel factor analysis (PARAFAC) was used to explain differences in treatment efficiencies among waters with different SUVA subjected to standardized water treatments (*16*). PARIX was defined as the ratio of two PARAFAC components, one with peak emissions above 500 nm, and the other with peak emissions around 400 nm. In treated and untreated drinking water samples from Europe, Asia and North America, PARIX predicted SUVA aromaticity more accurately than FI, HIX or BIX.

While equations using PARAFAC ratios can be useful, it is unclear whether PARAFAC models can be generalized to other datasets or to different types of aquatic systems. Also, the requirement to perform a PARAFAC analysis has practical limitations in monitoring applications since the components are first obtained from modelling sets of multispectral data matrices,

requiring many samples and relatively complicated data processing (*24*). By reanalyzing published datasets spanning the river to ocean continuum, the present study aimed to identify robust algorithms for predicting DOC aromaticity and abundance from in situ spectral measurements. The results provide new insights into the molecular structures underpinning optical measurements and a technical basis for monitoring DOC concentrations and aromaticity in real-time.

2. MATERIALS AND METHODS

2.1 Datasets

Nine published datasets comprising ~1000 samples were re-analysed for this study (Table S1). The datasets were created by eight independent research groups during the past two decades. They span all seven continents, include bulk DOM and extracts, with samples from inland surface waters (rivers, lakes, drinking water plants), groundwater, coastal waters, and the ocean. At minimum each dataset contains SUVA measurements plus fluorescence excitation-emission matrices (EEMs) and/or a PARAFAC model. SUVA was typically determined by the traditional USEPA method, which divides absorbance at 254 nm measured on a spectrophotometer (A_{254}) by DOC concentration measured on a separate TOC analyser (*18*), and sometimes by size-exclusion liquid chromatography with organic carbon detection (LC-OCD), which combines both detectors in a single instrument (*25*). Fluorescence intensities were measured on filtered samples in a 1-cm cell using a scanning excitation-emission (EEM) fluorometer. Absorbance was measured in a 1-cm cell within an *Aqualog* fluorometer or using a separate UV-vis spectrophotometer and a 1-cm or 5-cm cell. SUVA in this article is expressed in units $m^2g^{-1}C$ which is equivalent to L mg_C^{-1} m^{-1}.

2.2 PARIX and ARIX

Two datasets (*Isolates* and *Everglades*) were used to test the generality of PARIX, the PARAFAC-based index developed for estimating percent aromaticity in bulk DOM (*16*). Philibert et al. (*16*) defined PARIX as the ratio of two components referred to as H_{ii} and H_{iii}, where H_{ii} represents a long-wavelength component with peak emissions above 500 nm and H_{iii} represents a shorter-wavelength component with peak emissions near 400 nm. Specifically, it was tested whether PARIX derived from different PARAFAC models and different research groups predicts SUVA in bulk EEMs (*Everglades*) or in DOM extracts (*Isolates*). In both cases, the raw data consisted of PARAFAC loadings reported in published tables and PARIX was calculated as the ratio between reported Fmax values for components similar to H_{ii} and H_{iii}. The *Isolates* dataset included Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) performed on diverse DOM sources (*17*). This dataset was used to test whether PARIX predicts the relative abundance of molecular formulae associated with polycyclic aromatic and polyphenol compounds.

Eight datasets comprising whole-water DOM samples *(Alaska Rivers* (*26*)*., Australia* (*27*)*, Congo (28) , Everglades* (*29*)*, Horsens* (*30*)*, S. America (31), SUEZ* (*16*)*, Yukon Lakes* (*32*)*)* were used to assess correlations between SUVA and a newly identified fluorescence ratio called ARIX. ARIX is defined as the ratio of emission intensities detected at two fixed emission wavelengths (520 nm / 390 nm) when excited by light at 320 nm. ARIX tracks the ratio of PARAFAC components identified by Philibert et al. (*16*) and referred to as H_{ii} and H_{iii}. Since H_{iii} overlaps spectrally with several ubiquitous PARAFAC components with emission peaks between 400 - 450 nm (*16, 33-35*), the ARIX algorithm was designed to track PARIX using wavelengths that avoid interfering fluorescence from non-target fluorophores.

To estimate ARIX in the *Isolates* dataset, the fluorescence EEMs had to first be reconstructed from the reported PARAFAC model. Kellerman et al. (*17*) tabulated PARAFAC loadings and F_{max} values, which is sufficient information for estimating the original EEMs using published algorithms *(24, 36)*. The reconstructed dataset doesn't account for model residuals associated with lack of fit between the PARAFAC model and the primary dataset. Furthermore, it includes new error sources because the model loadings were reported with low machine precision. Consequently, a lower correlation is expected between SUVA and (reconstructed) ARIX than if the latter were derived from the primary EEM dataset. For the same reason, ARIX values determined for the *Isolates* dataset are not directly comparable to fluorescence indices reported by Kellerman et al. (*17*).

2.3 Regression models

For each dataset, individual regression models were calculated in MATLAB (ver. 2022a) using the *fitlm* function to predict SUVA with model $\hat{y} = \beta_1\,(\mathrm{P})\mathrm{ARIX} + \beta_0$. Models were made with and without MATLAB's robust statistics option that performs automatic outlier exclusion. Regressions were additionally calculated using four widely used fluorescence indices as the independent variable in place of ARIX. These were FI ("fluorescence index"), HIX ("humification index"), β/α ("freshness index") and BIX ("biological index"). FI was calculated as the ratio of emission intensities detected at 470 and 520 nm upon excitation at 370 nm (*19*). HIX was calculated as the sum of emissions at 435-480 nm divided by the sum of emissions from 300–345 and 435–480 nm following excitation at 254 nm (*20*). BIX was calculated as the ratio of emission intensities detected at 430 and 380 nm upon excitation at 310 nm (*37*). β/α was calculated as the ratio of emission detected at 380 nm to the maximum emission detected at 420-

435 nm upon excitation at 310 nm (*21*). The algorithms for BIX and β/α produced negligible differences in regressions with SUVA so BIX alone is reported herein.

To derive a global model linking SUVA and ARIX for all whole-waters, a model II (geometric) regression was calculated using *lsqfitgm* code from MBARI (*38*). Prior to regression analyses, outliers were excluded from three datasets: *Horsens*: samples with SUVA below 2 or above 8 $m^2g^{-1}C$ and (marine) samples from the estuary were excluded; *Australia:* one erroneous data point was excluded, and *Yukon Lakes*: one sample with extremely high DOC (>120 mg/L) was excluded. The *Isolates* dataset was analysed separately using ARIX derived from the reconstructed fluorescence EEMs.

Multiple linear regressions were performed using the *regress* function in MATLAB with A_{254} and ARIX as independent variables and DOC the dependent variable. There was no significant main effect of ARIX in any dataset, but there was a significant main effect on A_{254} in all datasets, and a significant interaction between A_{254} and ARIX in all datasets except *S. America*. Intercepts were also significant in all datasets but *Yukon Lakes*. The final model for each dataset retains only the statistically significant coefficients.

2.4 Sensitivity analysis

Fluorescence, absorbance, and DOC detectors have different inherent sensitivities, and instruments from different manufacturers (and even different versions of the same model) have varying levels of sensitivity and bias. A simulation was performed to estimate how much of the scatter in the relationship between ARIX and SUVA in bulk EEMs might be explained by only random (unbiased) errors. In this simulation, the model II regression equation was assumed to represent the true relationship between ARIX and SUVA. The eight bulk EEM datasets were used to generate a simulated dataset containing identical values of each DOC, SUVA and ARIX.

Only A_{254} was allowed to vary, and it was calculated from Eq. 1 so that the simulated dataset would align perfectly with the regression equation. Finally, a random error residual was added to each variable in each sample, selected from a random distribution defined according to Table S3. The simulation was run 100 times, then the median RMSE value across all 100 runs was recorded.

3. RESULTS AND DISCUSSION

3.1 Predicting DOC composition from fluorescence ratios

PARIX was an unbiased predictor of the proportion of polycyclic aromatic (PA) and polyphenolic (PP) structures in DOM extracts comprising the global *Isolate* dataset, according to molecular formulae measured using FT-ICR MS (Fig. 1a). Data are from Table 1 in Kellerman et al. (*17*), representing diverse freshwater and marine samples, including samples from the Pacific ocean at 21, 240 and 674 m. PARIX is calculated as the ratio of tabulated scores for PARAFAC component C3 divided by C2. Equation 1 estimates the relative abundance of polyphenolic compounds within 5% for samples in which these formulae comprised 2 - 45% of total formulas.

$$\%(\mathrm{PA} + \mathrm{PP}) = 30.28\ \mathrm{PARIX} \qquad \ldots\ldots\ldots \textit{Eq. 1}$$

PARIX also predicted SUVA in the *Isolate* dataset (Fig. 1b) and in bulk EEMs from two brackish river systems in the Florida Everglades (Fig 2a). PARIX in Fig 2a is calculated from Table 2 in Timko et al. (*29*) as the ratio of scores for PARAFAC component C4 divided by C5, which reveals a tight correlation with SUVA spanning both river systems ($R^2 = 0.96$). These two datasets demonstrate linearity between PARIX and relative aromaticity in bulk DOM and extracts, independent of the specific PARAFAC model used to calculate PARIX.

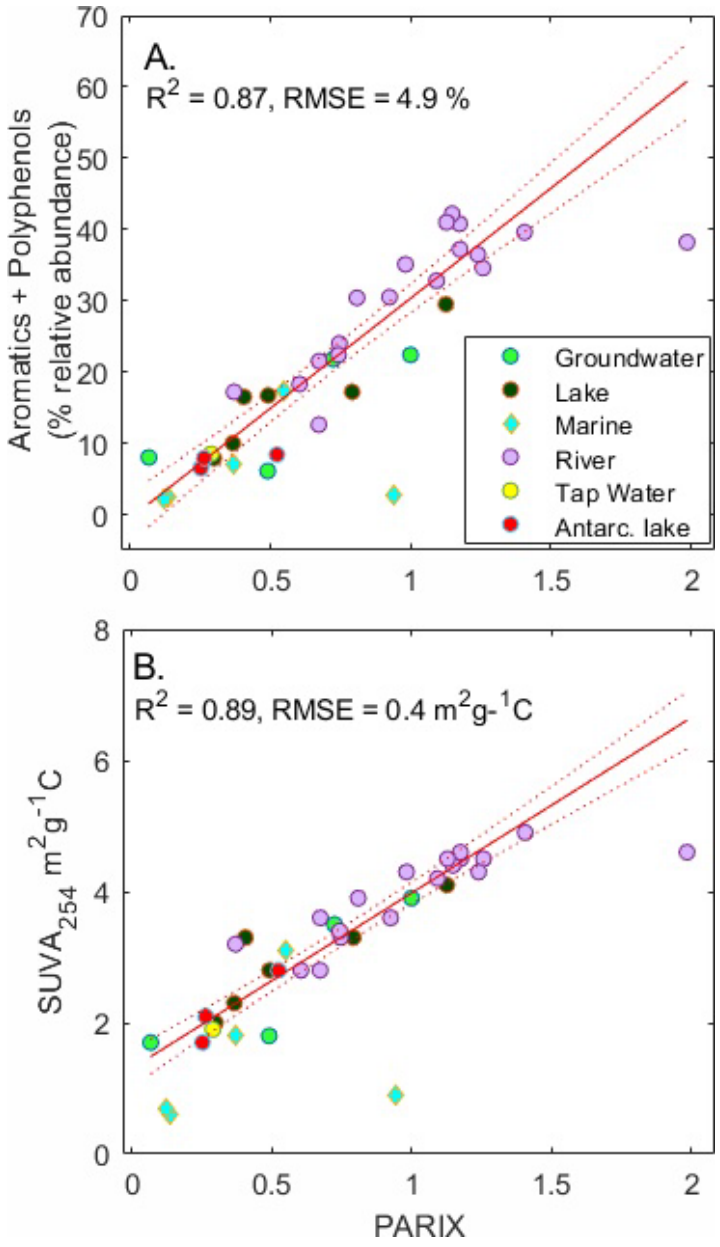

Figure 1. Prediction of relative aromaticity from PARIX in the global *Isolates* dataset. Aromaticity was determined by A: FTICR-MS; or B: SUVA. PARIX was calculated from Table 1 in Kellerman et al. *(19)* as C4/C3.

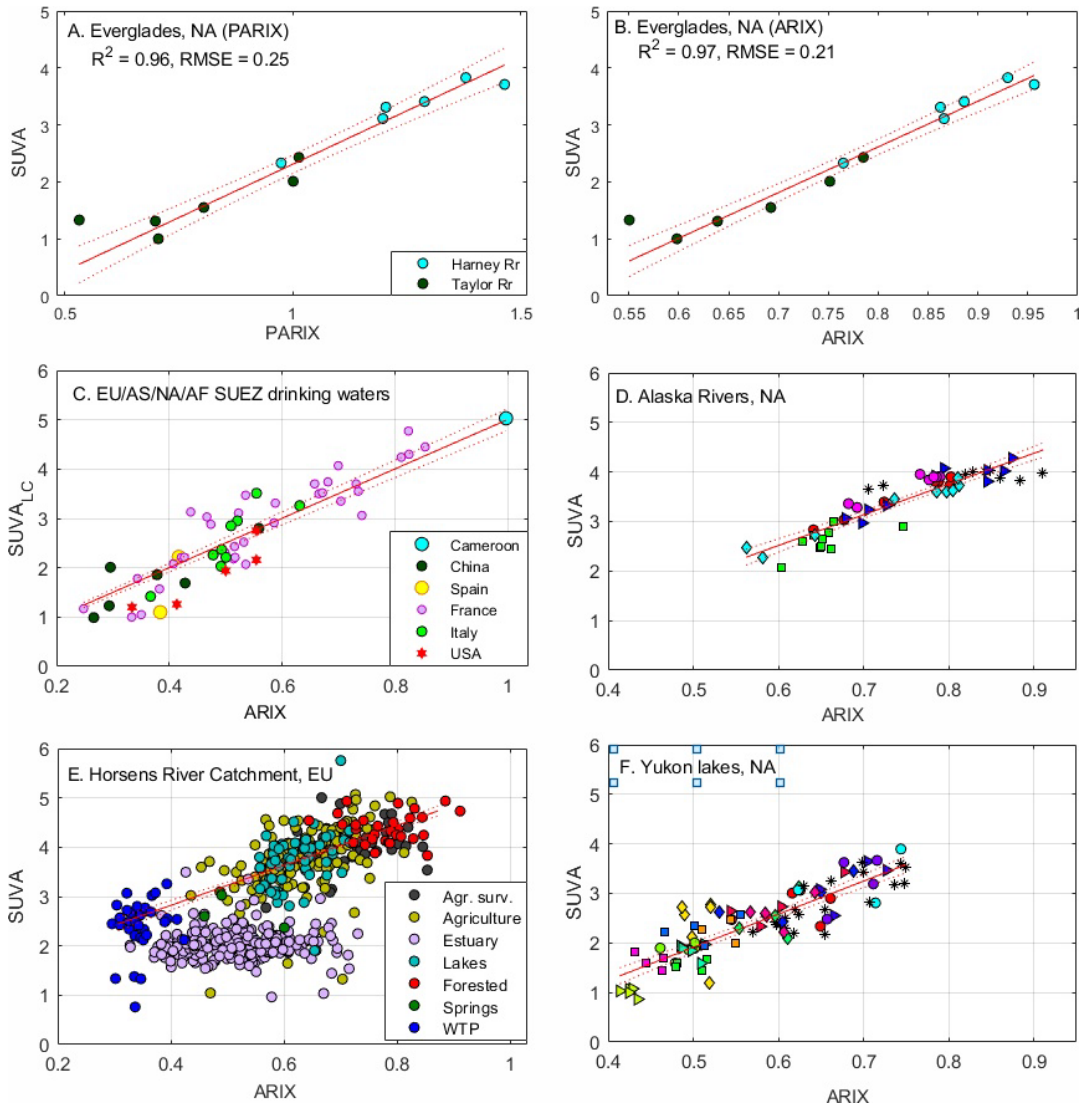

Figure 2. Prediction of SUVA from PARIX and ARIX in bulk surface waters. A: PARIX; B-F: ARIX. Samples are from inland waters and water treatment plants in Europe (EU), North America (NA), Asia (AS) and Africa (AF) *(12, 29, 31-33)*. In A-B, D and F, each new symbol represents a different river or lake.

In the diverse treated and untreated water samples comprising the SUEZ dataset, ARIX was an unbiased predictor of the ratio of humic substance (HS) to building block (BB) fractions determined by LC-OCD (Fig. 3, Eq. 2). While HS is understood to comprise high molecular weight humic substances, the BB fraction represents lower molecular weight weathering and oxidation products of humic substances (*25*).

$$\frac{\mathrm{HS}}{\mathrm{BB}} = 6.1\ \mathrm{ARIX} \qquad \ldots\ldots\ldots \textit{Eq. 2}$$

The relationships in Eq. 1-2 are notable considering that FTIR-MS and LC-OCD characterise all DOC fractions, including coloured and colourless molecules that do not fluoresce. Whether aromatic molecules in the DOC pool exhibit fluorescence depends on their specific structures and electronic properties. The ARIX numerator tracks a long-wavelength fluorescence component identified repeatedly in PARAFAC analyses (*17, 39*) and usually attributed to extensively π-conjugated polyaromatic structures (*40*). The denominator tracks a short-wavelength component with a secondary excitation maximum around 330 nm and emission peak below 400 nm (*16, 35*). This is similar to several oxidised fluorophores consisting of a single aromatic ring with attached carboxy, hydroxy, and methoxy groups, e.g. vanillic acid, syringic acid, and acetovanillone (*41*) albeit with longer absorption, possibly indicating additional substitution and/or the presence of a short, conjugated side chain, as in ferulic acid or coniferyl alcohol, or a conjugated heterocycle, as in coumarin (*5, 41, 42*).

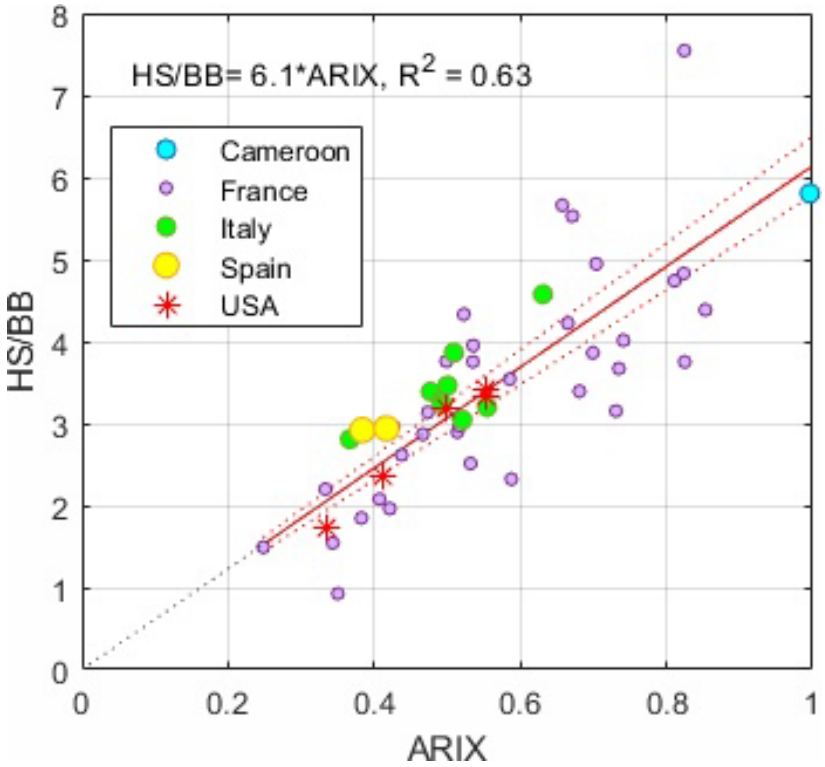

Figure 3. Prediction of LC-OCD composition from ARIX. Samples represent treated and untreated surface and groundwaters from the *SUEZ* dataset (HS: humic substances, BB: building blocks).

3.2 Relationships between SUVA and (P)ARIX

In individual datasets, ARIX and PARIX were reliable predictors of SUVA by linear regression (Figure 2, Table S2). In the *Everglades* dataset representing two river systems draining a tropical wetland, ARIX predicted SUVA more accurately than PARIX with RMSE = 0.21 $m^2g^{-1}C$. Low prediction errors (0.24-0.36 $m^2g^{-1}C$) were also observed for two high-latitude datasets consisting of six rivers (Fig. 2d) and fifteen hydrologically isolated lakes (Fig. 2f) in the Yukon basin, Alaska (*26, 32*). In the isolate dataset and in seven whole-water datasets, ARIX and/or PARIX outperformed traditional fluorescence indices when predicting SUVA. Average prediction errors ($m^2g^{-1}C$) in increasing order were: ARIX (0.39) < β/α = BIX (0.44) < HIX (0.55) < FI (0.60) (Table S2 and Fig. S1-S4). Thus, although FI and HIX are the two fluorescence indices used most frequently to predict DOM aromaticity, both were poorer predictors of SUVA than ARIX and the "biological" indices.

In Fig. 1, ARIX predicted percent aromatic molecular formulas in DOM isolates, including in extracts from 21 m and 240 m depths in the Pacific Ocean. For whole-water samples in the *Horsens* dataset (*30*) correlations between ARIX and SUVA disintegrate in the bay (Fig. 2e). Horsens river flows past relatively pristine sites in its upper reaches, then through an agriculturally impacted landscape, and ultimately past a wastewater treatment plant near the entrance to the bay. This progression is seen by decreasing SUVA and ARIX while moving downstream. Bay sites had a marine character, with high salinities (32 ppt) and low SUVA (1.6-2.4) as is typical for marine samples (*43*). In the estuary, ARIX varied randomly and independently from SUVA, indicating fluorescence intensities dropped below the limits for quantifying ARIX.

Variations in water chemistry affect the prediction of relative aromaticity from fluorescence ratios. Changes in pH affect some wavelengths more than others (*39, 44*), which could manifest as different slope terms for datasets with different pH, or a lower coefficient of determination (R^2) if pH varies between samples. Fe(II) and Fe(III) cations interfere with SUVA due to light absorption by aqueous iron complexes (*45*), and both species as well as several other metals (e.g. Cu, Hg, Al) reduce fluorescence via quenching reactions (*46*). In the presence of quenching metals, non-linearities would be expected to arise between SUVA and ARIX because A_{254} increases with increasing Fe concentration, whereas ARIX will decrease due to the preferential quenching of long-wavelength fluorescence (*47-49*).

Overall, the regression slope terms (β_1) for predicting SUVA from ARIX varied between datasets, with the North American datasets having steeper slope terms than their European, Australian, and African counterparts (Table S2). Comparing regressions equations for pairs of datasets, β_1 was not statistically different in *Alaska Rivers* (6.36 ± 0.41) vs. *Yukon Lakes* (7.03 ± 0.45) or *SUEZ* (5.8 ± 0.36), although the latter dataset is dominated by European samples (*16, 26, 32*). Also, β_1 for *Yukon Lakes* was not statistically different to *Everglades* (8.01 ± 0.48). Three datasets had lower slopes with β_1 not significantly different to each other: *Congo* (3.71 ± 0.50), *Horsens* (Denmark) (4.05 ± 0.19), and *S. America* (4.51 ± 0.58) (*28, 30, 31*). The *Australia* dataset (*27*) had considerable scatter and significantly lower β_1 (2.08 ± 0.19).

3.3 Global models for predicting DOM aromaticity

The *SUEZ* and *Isolates* datasets each span several continents and multiple biomes yet produced similar or higher R^2 than several geographically restricted datasets. For the *Isolate* dataset, fluorescence measurements were performed on extracted DOM, which probably limited

interfering matrix effects and improved signal/noise especially for marine DOM. However, the SUEZ dataset was measured on whole-water EEMs and still indicates a single regression for predicting SUVA from ARIX regardless of sample origin.

Plotting all eight whole-water datasets together (Fig. 4) showed ARIX to be confined within the range 0.15-1.0, with most samples falling between 0.25 and 0.9. Except for *Everglades*, all datasets were reasonably well captured by a single regression equation. Eq. 3 is derived from a geometric (Model II) regression since ARIX and SUVA were measured on multiple instruments (*50*) and has $R^2 = 0.48$ and RMSE = 0.83 $m^2g^{-1}C$.

SUVA = 7.2 ARIX- 1.2 …… Eq. 3

Fig. 4 shows only bulk DOM datasets; the *Isolate* dataset was excluded due to the unavailability of original fluorescence EEMs. However, after estimating the *Isolate* EEMs from the published PARAFAC model (Fig. S4-6), the regression coefficients were not significantly different to Eq. 3. For the *Isolates* dataset, the slope was 7.77 ± 0.57 with intercept -1.15 ± 0.33, producing Eq. 4:

SUVA = 7.8 ARIX - 1.1 …… *Eq. 4*

Extending Weishaar's equation linking ^{13}C-NMR aromaticity to SUVA in DOM extracts (*14*):

Percent aromaticity = 6.52 SUVA+ 3.63 = 50.7 ARIX - 3.9 …… *Eq. 5*

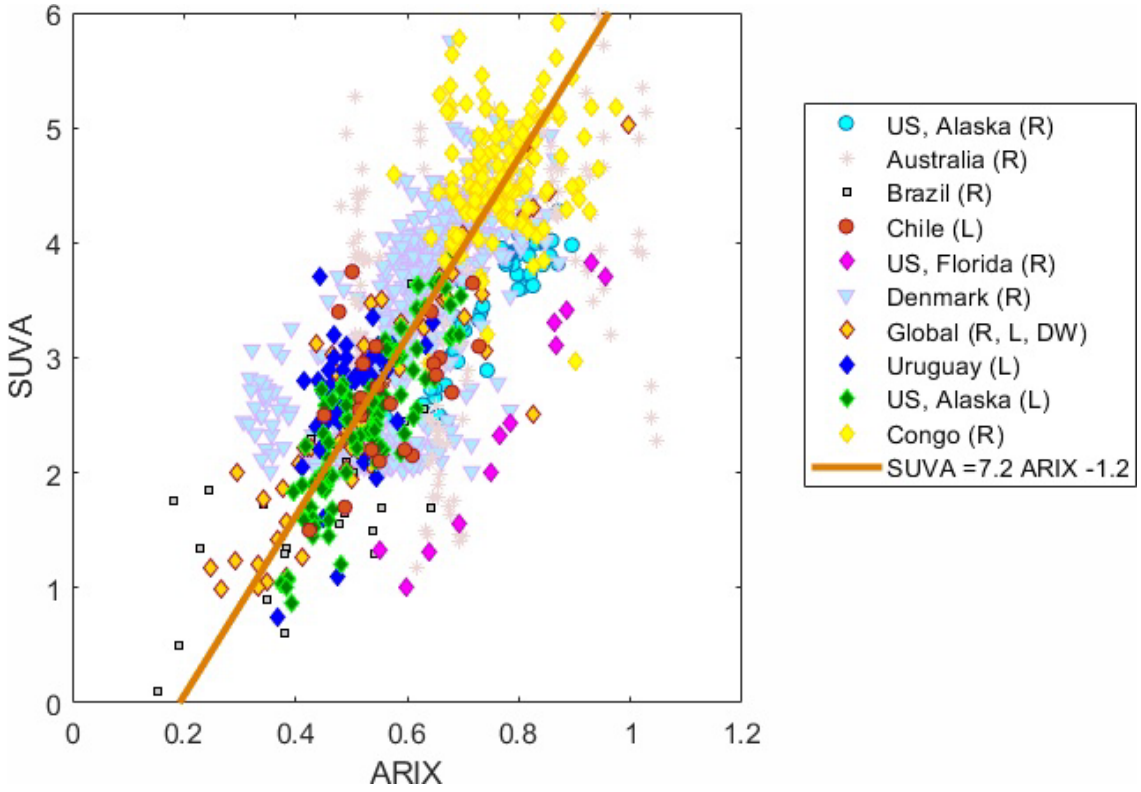

Figure 4. Prediction of SUVA from ARIX in bulk DOM from global surface and groundwaters. Samples represent fresh and brackish waters (R: rivers, L: lakes, DW: treated and untreated drinking water). The orange line represents a Model II regression with $R^2 = 0.48$.

Equations 1-4 support the hypothesis that a single slope factor links ARIX to SUVA across all DOM whole-water samples and extracts. Even the *Everglades* dataset shows the same slope with ARIX transposed by 0.3 units to the right. A right shift indicates a high background of colorless DOC, also seen by *Everglades* having an unusually large intercept in the regression of DOC vs absorbance (Section 3.4). Some of variability between datasets probably arises from artefacts; for example, the instruments used to measure ARIX and SUVA varied between datasets/regions, as did some protocols including whether samples were frozen prior to analysis.

Inter-laboratory comparison exercises often highlight biases arising from slightly different procedures and analytical instruments, including for fluorescence spectroscopy (*51*), DOC and SUVA (*18*), and FTIR-MS (*52*). SUVA measured by the USEPA method is the ratio of measurements derived from a spectrophotometer and a carbon analyser, both with different inherent sensitivities and biases, making it highly susceptible to both random and systematic errors (*18*). In developing the USEPA standardized method for SUVA analysis, Potter and Wimsatt (*18*) compared SUVA measured on duplicate samples using five different commercial DOC analysers placed in the same laboratory. Despite these efforts to standardize measurement conditions, and the use of a single spectrophotometer to measure absorbance in a 10-cm cell, a high standard deviation (~0.3 $m^2g^{-1}C$) was observed across all measurements.

In the current study, a sensitivity analysis indicated that ~29% of variability in Fig. 4 can be explained by purely random measurement errors under realistic assumptions about the precision of fluorescence, absorbance, and DOC detectors (Fig. S7). Furthermore, since seven different laboratories and 21 different detectors were used to derive Eq. 3, systematic biases related to different instruments and measurement protocols are unavoidable. In datasets where SUVA was measured using both the USEPA method and LC-OCD, stronger correlations were always

obtained with $SUVA_{LC}$. In the *SUEZ* dataset, biases could be traced to the absorbance detector because DOC measured by LC-OCD was perfectly estimated by DOC measured using a carbon analyzer ($\beta_1 = R^2 = 1$, RMSE = 0.01 mgL^{-1}) yet SUVA was 11% lower than $SUVA_{LC}$ ($\beta_1 = 0.89$, $R^2 = 0.85$). In two other datasets, traditional SUVA correlated only weakly to $SUVA_{LC}$. Individual datasets encompassing geographically diverse samples may provide more realistic estimates of the variability to be expected when a single set of precise detectors is used to predict SUVA from (P)ARIX across systems. For the *SUEZ* (whole-water) and *Isolates* (extract) datasets, RMSE was 0.42-0.44 $m^2g^{-1}C$ (Table S2).

3.4 Improving DOC predictions from absorbance measurements

Aquatic systems dominated by terrestrial DOM typically show tight correlations between CDOM absorption and DOC, allowing DOC concentrations to be predicted from A_{254}. However, numerous studies show that these parameters diverge in concert with decreasing hydrologic connectivity with the landscape (*43, 53*). Whereas terrestrial DOC is dominated by polyphenolic and condensed aromatic compound classes (*54*), in-situ production combined with the photodegradation of impounded DOM generates abundant unsaturated and phenolic molecular compounds that may come to represent a significant portion of the DOC pool (*26, 55*). High abundances of degraded aromatic structures result in the decoupling of A_{254} and DOC, preventing the accurate prediction of DOC concentration from A_{254} (*26, 53*). This phenomenon is extensively described and presents a significant hinder to predicting DOC concentrations from optical measurements in ecological, biogeochemical and remote-sensing studies (*43*).

For all but a single dataset in this study (*S. America*), the correlation between DOC and A_{254} was significantly improved by adding a term to represent the interaction of A_{254} with ARIX (Eq. 6):

$$DOC = b_0 + b_1 A_{254} + b_2 A_{254} ARIX \quad \text{…… } Eq.\ 6$$

Across the datasets, b_0 coefficients were positive (0-2.9) and highest for *Everglades*, indicating variable contributions by colorless DOC. The coefficient b_1 was positive and spanned an order of magnitude (15-120). The coefficient b_2 ranged between small and large negative values (-6 to -123) indicating that the effect of A_{254} on DOC generally decreased as SUVA increased; *Congo* was the sole exception with a small positive b_2 value (Table S3). *Everglades* and *Yukon Lakes* had very large b_1 and b_2 coefficients and by using Eq. 6 to predict DOC, RMSE decreased by 47% and 35% respectively relative to models with $b_2 = 0$ (Fig. 5, Table S3). These results indicate the potential to significantly improve DOC predictions from *in situ* spectroscopic measurements particularly in aquatic systems with low hydrologic connectivity, as exemplified by isolated boreal lakes (e.g. *Yukon Lakes*) and tropical wetlands (e.g. *Everglades*).

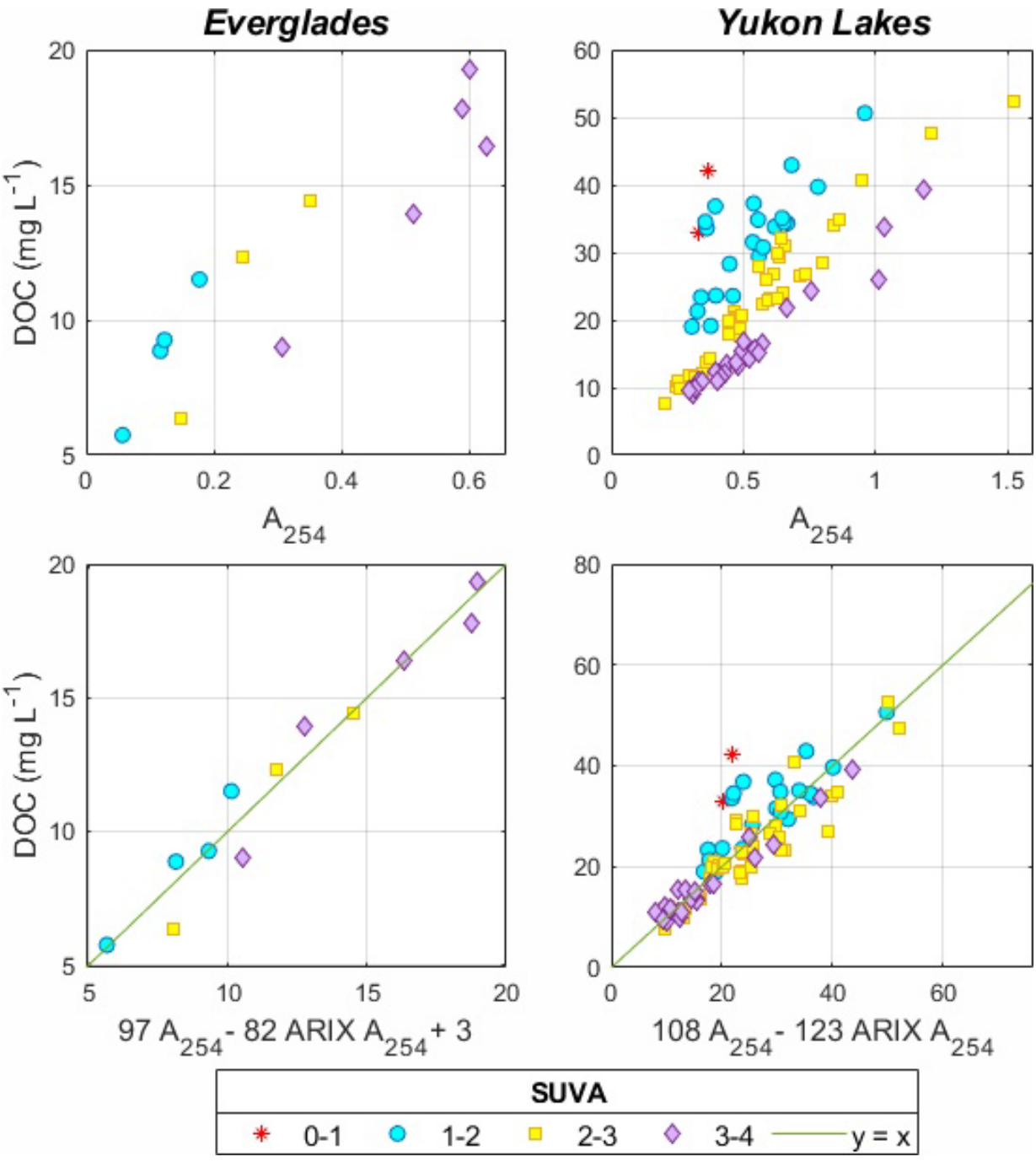

Figure 5: Prediction of DOC concentrations from absorbance and ARIX. Top row: DOC vs A_{254} in (A) *Everglades*; (B) *Yukon Lakes* datasets. Bottom row: Measured DOC vs DOC predicted from Eq. 6 for (C) *Everglades*; (D) *Yukon Lakes*.

Equation 6 includes two terms reflecting different CDOM components, along with an intercept representing non-coloured DOC. The first term (b_1*A_{254}) represents the contribution to DOC of fresh terrestrial CDOM and is independent of aromaticity, indicating that for the molecules comprising this pool, absorbance increases in proportion to the number of carbon atoms. The second term (b_2*A_{254}*$ARIX$) and represents the contributions to DOC of terrestrial breakdown products and autochthonous DOM. This term is generally negative and decreasing with increasing aromaticity, because the conjugated polyphenolic structures represented by the ARIX numerator are more efficient at absorbing light per carbon atom than the degraded phenolic structures represented by the ARIX denominator. For terrestrially dominated ecosystems with strong hydrologic connectivity only the first term is needed to predict DOC concentration with reasonable accuracy, whereas to model optically complex waters including impounded catchments with substantial *in situ* biogeochemical processing, both terms are required.

3.5 Benefits for water quality monitoring

Global surface waters face a changing climate with greater variability in both the quantity and quality of DOM. Warmer temperatures are increasing rates of litter decomposition in soils and rates of biological production in water at the same time as changes in land use and altered rainfall patterns are changing hydrological regimes and the connectivity between landscapes and DOM (*56, 57*). This affects correlations between DOC and spectroscopic measurements and decreases the accuracy of predicting DOC concentrations from absorbance measurements obtained in the laboratory or from remotely sensed imagery (*58*). Also due to the link between aromaticity and chemical reactivity, any lack of predictability in surface water composition negatively affects drinking water treatment by increasing the risk of chemical over- or underdosing (*16*).

Algorithms developed in this study improve the prediction of DOC abundance and reactivity from spectroscopic measurements which is especially important in waterbodies with high optical complexity. New *in situ* spectroscopic instruments could leverage these results to deliver currently missing data and provide real-time predictions of DOC abundance, reactivity and fate. Such instruments could simplify ground-truthing of remote sensing algorithms and enhance research on carbon cycling. Further applications include drinking water treatment, whereby real-time optical data could be used to adjust chemical doses in response to changing DOC composition, facilitating the sustainable removal of DOC compounds and their associated micropollutants.

Supporting Information. The following files are available free of charge. A pdf document containing Figs. S1 to S7 and Tables S1 to S4.

Corresponding Author: murphyk@chalmers.se

Author Contributions: KRM designed the study, collated datasets from primary sources in cited references, performed the data analyses, and wrote the manuscript.

Funding Sources: KRM acknowledges support from the Swedish research council Formas, grants 2022-01974 and 2023-01972.

Competing interests: KRM is the listed inventor of Swedish patent WO2023163638 and is a board member of Valys AB.

Data and materials availability: (Note that data and code will be made publicly available upon acceptance of this article) The data needed to reproduce these results are forthcoming in the Dryad database (https://doi.org/10.5061/dryad.x69p8czt8). MATLAB code for extracting ARIX and other fluorescence indices from EEMs is available via the pickpeaks function of the drEEM software package at www.drEEM.org.

ACKNOWLEDGMENT

Open science is a cornerstone of achieving rapid scientific progress as is needed to address the world's water quality challenges. The author acknowledges the efforts of many scientists who enabled this study by making their data available for this research, especially Saman Acharya, Cristina Romera Castillo, Daniel Graeber, Sarah Johnston, Tibault Lambert, Marc Philibert, and Colin Stedmon. Colin Stedmon commented on an early version of this manuscript.

REFERENCES

1. D. A. Hansell, Recalcitrant Dissolved Organic Carbon Fractions. *Ann Rev Mar Sci* **5**, 421-445 (2013).
2. M. Berggren, F. Guillemette, M. Bieroza, I. Buffam, A. Deininger, J. A. Hawkes, D. N. Kothawala, R. LaBrie, J.-F. Lapierre, K. R. Murphy, E. S. Al-Kharusi, M. P. D. Rulli, G. Hensgens, H. Younes, U. J. Wünsch, Unified understanding of intrinsic and extrinsic controls of dissolved organic carbon reactivity in aquatic ecosystems. *Ecology* **103** e3763 (2022).
3. A. M. Kellerman, T. Dittmar, D. N. Kothawala, L. J. Tranvik, Chemodiversity of dissolved organic matter in lakes driven by climate and hydrology. *Nature Communications* **5**, 3804 (2014).
4. A. J. Tanentzap, J. A. Fonvielle, Chemodiversity in freshwater health. *Science* **383**, 1412-1414 (2024).
5. C. A. Stedmon, Y. Yamashita, in *Biogeochemistry of Marine Dissolved Organic Matter, 3rd edition,* C. A. Carlson, D. A. Hansell, Eds. (2024), chap. 5.
6. T. S. Catala, I. Reche, A. Fuentes-Lema, C. Romera-Castillo, M. Nieto-Cid, E. Ortega-Retuerta, E. Calvo, M. Alvarez, C. Marrase, C. A. Stedmon, X. A. Alvarez-Salgado, Turnover time of fluorescent dissolved organic matter in the dark global ocean. *Nat Commun* **6**, 5986 (2015).
7. R. G. Qualls, C. J. Richardson, Factors controlling concentration, export, and decomposition of dissolved organic nutrients in the Everglades of Florida. *Biogeochemistry* **62**, 197-229 (2003).
8. J. T. O. Kirk, in *Light and Photosynthesis in Aquatic Ecosystems,* J. T. O. Kirk, Ed. (Cambridge University Press, Cambridge, 2010), pp. 50-97.
9. J. A. Leenheer, J.-P. Croué, Peer Reviewed: Characterizing Aquatic Dissolved Organic Matter. *Environ Sci Technol* **37**, 18A-26A (2003).
10. M. Edwards, Predicting DOC removal during enhanced coagulation. *Journal - American Water Works Association* **89**, 78-89 (1997).
11. B. Schreiber, T. Brinkmann, V. Schmalz, E. Worch, Adsorption of dissolved organic matter onto activated carbon - The influence of temperature, absorption wavelength, and molecular size. *Wat Res* **39**, 3449-3456 (2005).
12. F. Zietzschmann, C. Stützer, M. Jekel, Granular activated carbon adsorption of organic micro-pollutants in drinking water and treated wastewater – Aligning breakthrough curves and capacities. *Wat Res* **92**, 180-187 (2016).
13. A. Andersson, E. Lavonen, M. Harir, M. Gonsior, N. Hertkorn, P. Schmitt-Kopplin, H. Kylin, D. Bastviken, Selective removal of natural organic matter during drinking water production changes the composition of disinfection by-products. *Environ. Sci. Water Res. Technol.* **6**, 779-794 (2020).
14. J. L. Weishaar, G. R. Aiken, B. A. Bergamaschi, M. S. Fram, R. Fujii, K. Mopper, Evaluation of Specific Ultraviolet Absorbance as an Indicator of the Chemical Composition and Reactivity of Dissolved Organic Carbon. *Environ Sci Technol* **37**, 4702-4708 (2003).
15. J. A. Korak, G. McKay, Critical Review of Fluorescence and Absorbance Measurements as Surrogates for the Molecular Weight and Aromaticity of Dissolved Organic Matter. *Environmental Science: Processes & Impacts*, (2024).

16. M. Philibert, S. Luo, L. Moussanas, Q. Yuan, E. Filloux, F. Zraick, K. R. Murphy, Drinking water aromaticity and treatability is predicted by dissolved organic matter fluorescence. *Wat Res* **220**, 118592 (2022).
17. A. M. Kellerman, F. Guillemette, D. C. Podgorski, G. R. Aiken, K. D. Butler, R. G. M. Spencer, Unifying Concepts Linking Dissolved Organic Matter Composition to Persistence in Aquatic Ecosystems. *Environ Sci Technol* **52**, 2538-2548 (2018).
18. B. B. Potter, J. C. Wimsatt, "Method 415.3 - Measurement of total organic carbon, dissolved organic carbon and specific uv absorbance at 254 nm in source water and drinking water.," (U.S. Environmental Protection Agency, Washington, DC, 2005).
19. R. M. Cory, M. P. Miller, D. M. McKnight, J. J. Guerard, P. L. Miller, Effect of instrument-specific response on the analysis of fulvic acid fluorescence spectra. *Limnol Oceanogr Meth* **8**, 67-78 (2010).
20. T. Ohno, Fluorescence inner-filtering correction for determining the humification index of dissolved organic matter. *Environ Sci Technol* **36**, 742–746 (2002).
21. E. Parlanti, K. Wörz, L. Geoffroy, M. Lamotte, Dissolved organic matter fluorescence spectroscopy as a tool to estimate biological activity in a coastal zone submitted to anthropogenic inputs. *Org. Geochem.* **31**, 1765-1781 (2000).
22. H. F. Wilson, M. A. Xenopoulos, Effects of agricultural land use on the composition of fluvial dissolved organic matter. *Nat. Geosci.* **2**, 37-41 (2009).
23. J. A. Korak, A. D. Dotson, R. S. Summers, F. L. Rosario-Ortiz, Critical analysis of commonly used fluorescence metrics to characterize dissolved organic matter. *Wat Res* **49**, 327-338 (2014).
24. K. R. Murphy, C. A. Stedmon, D. Graeber, R. Bro, Fluorescence spectroscopy and multi-way techniques. PARAFAC. *Anal Meth* **5**, 6557-6566 (2013).
25. S. A. Huber, A. Balz, M. Abert, W. Pronk, Characterisation of aquatic humic and non-humic matter with size-exclusion chromatography - organic carbon detection - organic nitrogen detection (LC-OCD-OND). *Wat Res* **45**, 879-885 (2011).
26. S. E. Johnston, J. C. Carey, A. Kellerman, D. C. Podgorski, J. Gewirtzman, R. G. M. Spencer, Controls on Riverine Dissolved Organic Matter Composition Across an Arctic-Boreal Latitudinal Gradient. *J. Geophys. Res.-Biogeosci.* **126**, e2020JG005988 (2021).
27. S. Acharya, A. Holland, G. Rees, A. Brooks, D. Coleman, C. Hepplewhite, S. Mika, N. Bond, E. Silvester, Relevance of tributary inflows for driving molecular composition of dissolved organic matter (DOM) in a regulated river system. *Wat Res* **237**, 119975 (2023).
28. T. Lambert, S. Bouillon, F. Darchambeau, P. Massicotte, A. V. Borges, Shift in the chemical composition of dissolved organic matter in the Congo River network. *Biogeosciences* **13**, 5405-5420 (2016).
29. S. A. Timko, C. Romera-Castillo, R. Jaffé, W. J. Cooper, Photo-reactivity of natural dissolved organic matter from fresh to marine waters in the Florida Everglades, USA. *Environmental science. Processes & impacts* **16**, 866-878 (2014).
30. C. A. Stedmon, S. Markager, M. Søndergaard, T. Vang, A. Laubel, N. H. Borch, A. Windelin, Dissolved organic matter (DOM) export to a temperate estuary: seasonal variations and implications of land use. *Estuar. Coast.* **29**, 388-400 (2006).
31. D. Graeber, J. Gelbrecht, M. T. Pusch, C. Anlanger, D. von Schiller, Agriculture has changed the amount and composition of dissolved organic matter in Central European headwater streams. *Sci Tot Env* **438**, 435-446 (2012).

32. S. E. Johnston, R. G. Striegl, M. J. Bogard, M. M. Dornblaser, D. E. Butman, A. M. Kellerman, K. P. Wickland, D. C. Podgorski, R. G. M. Spencer, Hydrologic connectivity determines dissolved organic matter biogeochemistry in northern high-latitude lakes. *Limnol Oceanogr* **65**, 1764-1780 (2020).
33. U. J. Wünsch, K. R. Murphy, C. A. Stedmon, The one-sample PARAFAC approach reveals molecular size distributions of fluorescent components in dissolved organic matter. *Environ Sci Technol* **51**, 8 (2017).
34. U. J. Wünsch, R. Bro, C. A. Stedmon, P. Wenig, K. R. Murphy, Emerging patterns in the global distribution of dissolved organic matter fluorescence. *Anal Meth* **11**, (2019).
35. N. Moona, A. Holmes, U. J. Wünsch, T. J. R. Pettersson, K. R. Murphy, Full-Scale Manipulation of the Empty Bed Contact Time to Optimize Dissolved Organic Matter Removal by Drinking Water Biofilters. *ACS ES&T Water* **1**, 1117-1126 (2021).
36. C. A. Andersson, R. Bro, The *N*-way Toolbox for MATLAB. *Chemom Intell Lab* **52**, 1-4 (2000).
37. A. Huguet, L. Vacher, S. Relexans, S. Saubusse, J. M. Froidefond, E. Parlanti, Properties of fluorescent dissolved organic matter in the Gironde Estuary. *Org. Geochem.* **40**, 706-719 (2009).
38. MBARI. (Moss Landing, CA, 2022), vol. 2023.
39. K. R. Murphy, S. A. Timko, M. Gonsior, L. Powers, U. Wünsch, C. A. Stedmon, Photochemistry illuminates ubiquitous organic matter fluorescence spectra. *Environ Sci Technol* **52**, 11243-11250 (2018).
40. G. McKay, Emerging investigator series: critical review of photophysical models for the optical and photochemical properties of dissolved organic matter. *Environmental Science: Processes & Impacts* **22**, 1139-1165 (2020).
41. U. J. Wünsch, K. R. Murphy, C. A. Stedmon, Fluorescence quantum yields of natural organic matter and organic compounds: Implications for the fluorescence-based interpretation of organic matter composition. *Front Mar Sci* **2**, (2015).
42. C. A. Stedmon, N. Nelson, in *Biogeochemistry of Marine Dissolved Organic Matter, 2nd ed.,* D. A. Hansell, C. A. Carlson, Eds. (Elsevier, Eastbourne, UK, 2015), pp. 481–508.
43. P. Massicotte, E. Asmala, C. Stedmon, S. Markager, Global distribution of dissolved organic matter along the aquatic continuum: Across rivers, lakes and oceans. *Sci Tot Env* **609**, 180-191 (2017).
44. S. A. Timko, M. Gonsior, W. J. Cooper, Influence of pH on fluorescent dissolved organic matter photo-degradation. *Wat Res* **85**, 266-274 (2015).
45. A. Stefánsson, Iron(III) Hydrolysis and Solubility at 25 °C. *Environ Sci Technol* **41**, 6117-6123 (2007).
46. N. Senesi, T. M. Miano, M. R. Provenzano, G. Brunetti, Characterization, differentiation, and classification of humic substances by fluorescence spectroscopy. *Soil Sci.* **152**, 259-271 (1991).
47. B. A. Poulin, J. N. Ryan, G. R. Aiken, Effects of iron on optical properties of dissolved organic matter. *Environ Sci Technol* **48**, 10098-10106 (2014).
48. Y. Yamashita, R. Jaffé, Characterizing the interactions between trace metals and dissolved organic matter using excitation-emission matrix and parallel factor analysis. *Environ Sci Technol* **42**, 7374-7379 (2008).

49. M. Heibati, C. A. Stedmon, K. Stenroth, S. Rauch, J. Toljander, M. Säve-Söderbergh, K. R. Murphy, Assessment of drinking water quality at the tap using fluorescence spectroscopy. *Wat Res* **125**, 1-10 (2017).
50. R. R. Sokal, F. J. Rohlf, *Biometry: The principles and practice of statistics in biological research*. (W.H. Freeman & Co., New York, ed. 3, 1995).
51. K. R. Murphy, K. D. Butler, R. G. M. Spencer, C. A. Stedmon, J. R. Boehme, G. R. Aiken, The measurement of dissolved organic matter fluorescence in aquatic environments: An interlaboratory comparison. *Environ Sci Technol* **44**, 9405–9412 (2010).
52. J. A. Hawkes, J. D'Andrilli, J. N. Agar, M. P. Barrow, S. M. Berg, N. Catalán, H. Chen, R. K. Chu, R. B. Cole, T. Dittmar, R. Gavard, G. Gleixner, P. G. Hatcher, C. He, N. J. Hess, R. H. S. Hutchins, A. Ijaz, H. E. Jones, W. Kew, M. Khaksari, D. C. Palacio Lozano, J. Lv, L. R. Mazzoleni, B. E. Noriega-Ortega, H. Osterholz, N. Radoman, C. K. Remucal, N. D. Schmitt, S. K. Schum, Q. Shi, C. Simon, G. Singer, R. L. Sleighter, A. Stubbins, M. J. Thomas, N. Tolic, S. Zhang, P. Zito, D. C. Podgorski, An international laboratory comparison of dissolved organic matter composition by high resolution mass spectrometry: Are we getting the same answer? *Limnol Oceanogr Meth* **18**, 235-258 (2020).
53. R. G. M. Spencer, K. D. Butler, G. R. Aiken, Dissolved organic carbon and chromophoric dissolved organic matter properties of rivers in the USA. *J. Geophys. Res.-Biogeosci.* **117**, (2012).
54. A. M. Kellerman, D. N. Kothawala, T. Dittmar, L. J. Tranvik, Persistence of dissolved organic matter in lakes related to its molecular characteristics. *Nature Geosci* **8**, 454-457 (2015).
55. C. Grasset, K. Einarsdottir, N. Catalán, L. J. Tranvik, M. Groeneveld, J. A. Hawkes, K. Attermeyer, Decreasing Photoreactivity and Concurrent Change in Dissolved Organic Matter Composition With Increasing Inland Water Residence Time. *Global Biogeochemical Cycles* **38**, e2023GB007989 (2024).
56. N. Catalán, C. Rofner, C. Verpoorter, M. T. Pérez, T. Dittmar, L. Tranvik, R. Sommaruga, H. Peter, Treeline displacement may affect lake dissolved organic matter processing at high latitudes and altitudes. *Nature Communications* **15**, (2024).
57. D. L. Swain, B. Langenbrunner, J. D. Neelin, A. Hall, Increasing precipitation volatility in twenty-first-century California. *Nature Climate Change* **8**, 427-433 (2018).
58. P. L. Brezonik, L. G. Olmanson, J. C. Finlay, M. E. Bauer, Factors affecting the measurement of CDOM by remote sensing of optically complex inland waters. *Remote Sensing of Environment* **157**, 199-215 (2015).