**Global patterns in dissolved organic matter from headwater streams to oceans**

Philippe Massicotte, Eero Asmala, XXX, YYY

**Introduction**

Concentrations of organic carbon (OC) have increasing significantly in freshwater systems in recent decades (REFS). As a result, more OC is being exported via rivers to the global ocean. The fate of this increased carbon load is largely unknown, as the extent of carbon processing during the transport along the hydrological path from streams to rivers, and further via estuaries to coastal seas is highly uncertain.

Dissolved organic matter (DOM) plays a fundamental role in the ecology of aquatic ecosystems (Hansell2014). For example, DOM is the main source of metabolic substrates for heterotrophic bacteria and influences the composition of aquatic microbial communities (Findlay2003). Additionally, the chromophoric fraction of the DOM pool (CDOM) is a major driver of underwater light characteristics (Kirk1994) which modulate many bio-optical processes such as primary production (Thrane2014, Seekell2015) and also constitute a natural screen protecting aquatic organisms against harmful ultraviolet (UV) radiation (refs). For these reasons, the chemical properties (quantity and quality) of the DOM pool transiting in the different aquatic ecosystems are important parameters that determine the rate at which the carbon is processed and further transferred to higher trophic levels (refs).

In recent decades, climate change, eutrophication and changes in land use have contributed to increase the terrestrial DOM inputs from in aquatic ecosystems (Roulet2006, Massicotte2013RSE). This has importance consequences since transformations of even a small fraction of the DOM pool can potentially have large impacts on ecosystem functioning (Prairie2008). Increases in CO2 emissions (Lapierre2013) and reduction in primary production due to light shading (Seekell2015, Thrane2014) already have been evoked as negative consequences of this generalized increase in terrestrial DOM at local and regional scales. However, how aquatic ecosystems will react to this increase in strongly colored DOM remain largely unknown because most studies about the fate and the dynamic of DOM are either site or ecosystem specific and thus do not capture global patterns. Furthermore, generalizing trends from local studies to global scales is a difficult task because during its transport from headwaters lakes to oceans, DOM is influenced by several processes of production and degradation that continuously alter its intrinsic characteristics.

Dissolved organic carbon (DOC, the main constituent of DOM) and CDOM are often used as proxies to characterize the quantity and the quality of the DOM pool. Especially, DOC-normalized UV absorbance at 254 nm (SUVA254) has been used widely to characterize the DOM pool (Weishaar 2003). Normalizing the CDOM absorption by the respective DOC concentration enables the investigation of the DOM quality, independent of its quantity. It was found that CDOM absorption coefficients correlate linearly to the DOC content of the DOM pool (refs), although the relationship vary greatly between studies (refs). One explanation is that the dynamic of DOC and CDOM can be decoupled during its transport from headwater lakes to oceans (Coble2007, Spencer2012a, Helms2013). For example, photochemical processes are likely to have a greater influence on the CDOM fraction as it absorb more UV light compared to the colorless fraction. In addition to carbon, also iron plays a role in the optical properties of DOM, which is especially important in freshwater systems (refs by Vähätalo)

Nowadays, DOC and CDOM properties are routinely measured in most ecological studies. This unique opportunity opens the door to study the fate and the distribution of DOM at larger scales. The aim of this project was to use published data to explore the factors regulating the composition and the fate of the DOM pool during its transition from lakes and oceans. An extensive literature survey was performed to extract datasets containing both DOC and absorption measurements of CDOM to gain insights about the fate and the dynamic of DOM in different aquatic ecosystems. We hypothesized that the CDOM/DOC relationship would be the stronger in freshwater end-members (lakes and rivers) to gradually weaken as the DOM pool is processed during its transition to the oceans.

2. Variation in this relationship can be explained by season, ecosystem type or other biogeochemical variables

3. CDOM-DOC relationship can be used to infer the amount of non-colored DOC

**Methods**

**Literature survey and spatial coverage**

Web of science, Google Scholar as well as public data repositories were screened using “cdom”, “doc”, “dissolved organic carbon”, “absorption” keywords for datasets presenting raw (i.e. not summarized) values of DOC and optical properties of CDOM. The minimum variables required to consider a dataset were DOC, absorbance or absorption of CDOM, geographical coordinates and time of the sampling. For the CDOM data, wavelengths and cuvette size used for the optical measurements were also required. Using these criteria, we compiled 43 datasets containing 10826 unique observations of simultaneous DOC and absorption properties of DOM (Fig. 1, Table 1). A total of 4308 observations with complete CDOM spectra (ie continuous measurements along a range of wavelengths) and 6518 observations with CDOM absorption measurement at discrete wavelengths were extracted. Extracted data where sampled between 1991 and 2014.

An important fraction of the dataset comes from an extensive sampling survey made by the UCSB Global CDOM Project that occurred in the ocean interior between xxx and yyy (Nelson2013). These observations were collected in the Orinoco River outflow in the Caribbean, North American continental shelf surface waters, Indian Ocean abyssal waters, North Pacific main thermocline waters and South Pacific subtropical surface waters (Nelson2013). Another important subset of the dataset comes from the XXX and YYY environmental monitoring initiatives taking place on the East Coast and along the Gulf of Mexico in the USA (Fig. 1). Observations of inland water (lakes and rivers) in North America originate mostly from the Global Lake Ecological Observatory Network (GLEON) and the Long Term Ecological Research (LTER) monitoring programs of North temperate lakes.

1. North Europe (Denmark, Finland, etc.)
2. Africa

Maybe an appendix with boxplot of the main variables (per study?).

**Ecosystem classification**

Each observations were assigned to a defined ecosystem using either the sampling location or the salinity when available (supplementary Fig. 1). Observations presenting salinity values were classified as follows: river (salinity < 0.5), estuary (0.5 < salinity <= 5), coastal (5 < salinity <= 30), ocean (salinity > 30). Based available information, observations were classified as follows: river (*n* = 4527), ocean (*n* = 3143), estuary (*n* = 1273), wetlands (*n* = 954), coastal (*n* = 866), lake (*n* = 627), pond (*n* = 84), brines (*n* = 58) and sewage (*n* = 32). Brines and sewage have categorized as ecosystems because they represent systems which are outside the typical hydrological continuum from headwater streams to oceans.

**Data processing**

Absorbance by CDOM were converted to absorption coefficients and expressed per meter using equation 1 (Kirk1994):

EQ1 HERE

where aCDOM(λ) is the absorption coefficient (m−1) at wavelength λ, A(λ) the absorbance at wavelength λ and L the path length of the optical cell in meters. Given that UV–visible absorption spectra of CDOM decrease approximately exponentially with increasing wavelength, a simple exponential model (equation 2) has been used to extract to extract quantitative information about optical properties of CDOM (Jerlov1968, Bricaud1981, Stedmon2001).

EQ2 HERE

where aCDOM is the absorption coefficient (m−1), λ is the wavelength (nm), λ0 is a reference wavelength (nm), K is a background constant (m−1) accounting for scatter in the cuvette and drift of the instrument. S is the spectral slope (nm−1) that describes the approximate exponential rate of decrease absorption with increasing wavelength. For the reminder of the text, S300-600 denotes the spectral slope calculated between 300 and 600 nm, S275-295 the spectral slope calculated between 275 and 295 nm and S350-400 the spectral slope calculated between 350 and 400 nm. The slope ratio (SR) is the ratio between S275-295 and S350-400 which is a proxy for DOM molecular weight (Helms2008). Specific ultra-violet absorbance (SUVA254) was calculated by dividing absorbance at 254 nm by DOC content in mgC (Weisharr2003). For other analysis, DOC were expressed in µmol C assuming that one mol of carbon weight 12 grams.

Given the wide range of wavelengths used in each study, absorption spectra were filtered to keep measurement between 250 and 600 nm at 1 nm increment. For the Nelson et al. (20xx) dataset (Table 1), absorption was only available between 275 and 600 nm. Four criterion were used for controlling the quality of absorption spectra: (1) SUVA254 had to be smaller or equal to 6 as suggested by XXX (ref Eero, I think you told me about this value of 6), (2) the spectral slope (S, equation 2) needed to be smaller than 0.08 nm-1, (3) the determination coefficient of the fit (R2, equation 1) needed to be at least 0.95 and (4) aCDOM(440) needed to be positive. Based on these criteria, a total of 119 absorption spectra were discarded from further analyzes.

**Spectral slope curve analysis**

Spectral slope curves were analyzed according to Loiselle.

**Estimation of aCDOM(350)**

We found out that a wide range of different wavelengths were used to report absorption coefficients of CDOM (Table 2). To make absorption coefficient comparable among studies, an interpolation procedure was used to estimate absorption coefficient at 350 nm (aCDOM(350)) independently of the wavelength used in each study. This choice was motivated because absorption at 350 nm was among the most reported wavelength in the available data. To achieve this we used available complete absorption spectra to predict the value of aCDOM(350) from observations measured at other wavelengths (Fig. 2). This was done by regressing all the values at a specific wavelength (ex. aCDOM(254)) against aCDOM(350). Then, the slope and the intercept of the linear regression were used to predict aCDOM(254) to aCDOM(350). Based on reported wavelengths, a total of 17 linear models were made (Table 2). A minimum value of 0.98 for the determination coefficient (*R2*) was used as a threshold to discard observations that were “too far” from the targeted wavelength of 350 nm (Table 2). Note that the complete absorption spectra from the Nelson2002 were not used in the interpolation procedure because of their reduced spectral range at lower wavelengths (275-700).

**Statistical analysis**

All statistical analysis were performed in R 3.3.1 (RCoreTeam2016). The segmentation analysis has been performed using the **segmented** R package (Muggeo2003, Muggeo2008). Principal component analysis (PCA) was performed using the **vegan** R package (Oksanen2016).

**Results**

**Estimation of aCDOM(350)**

Between 250 and 400 nm, the R2 of the linear regressions to predict aCDOM(350) remained above 0.98 (Fig. 2A). After this point, R2 dropped rapidly to reach 0.86 at 500 nm. The slopes of the regressions increased almost exponentially and ranged between 0.28 to 6.99 (Fig. 2B). The intercepts increased linearly and ranged between -1.36 and 1.51 (Fig. 2C). Before 350 nm, slopes presented negative values and became positive after this point. Regression coefficients used to interpolate aCDOM(λ) in this study are presented in Table 2. Note that absorption coefficients at 420, 440 nm and 443 nm were discarded because of the R2 bellow the selected threshold of 0.98. A heat map plot showing the R2 of the regressions between all possible pairs of wavelengths between 250 nm and 500 nm is presented in supplementary Fig. 2. An R data frame containing regression coefficients for all wavelengths is also provided as a supplementary appendix rda file. The coefficient of correlation between 254 and 350 was very high, 0.9X, and this allows us to use absorbance at 350 nm to calculate SUVA350 instead of SUVA254, due to better data coverage at 350 nm.

**DOC and aCDOM distribution**

Absorption coefficients at 350 nm varied by three orders of magnitude between ocean and wetland ecosystems (Fig. 3A). In oceans, median aCDOM(350) was around 0.1 m-1 and reached approximately 100 m-1 in wetlands ecosystems. TODO

**Global relation between DOC and aCDOM(350)**

A strong positive log-linear relationship was found between aCDOM(350) and DOC (Fig. 4A, n = 11562, R2 = 0.93, p < 0.0001). At low value of DOC (~35 µmol), predicted value of aCDOM(350) was 0.03 m-1. As DOC increased to a maximum of 44600 µmol, predicted aCDOM(350) reached 1097 m-1 (Fig. 4A). The derived equation from the log-linear model indicated that aCDOM(350) increases by xxx m-1 for each unit of increase in DOC. The robustness of the global relationship was found to vary greatly among the different ecosystems and R2 averaged 0.63 (Fig. 4B, supplementary Fig. 3).

Rivers, brines, ponds, estuaries and wetlands presented the highest R2.

**Segmentation analysis**

The linear trend of the specific UV absorbance at 254 nm (SUVA254) along the salinity gradient was modeled using a piecewise regression where two different breakpoints at salinity 8.66 and 26.84 were found (Fig. 5, R2 = 0.74, p < 0.0001). Between salinity 0 and 8.66, the slope of the linear regression was -0.3 indicating that SUVA254 decreased by this amount for each unit increase in salinity. Between salinity 8.66 and 26.84, SUVA254 remained stable and the slope of the regression was not significantly different from 0. Another significant slope with a value of -0.09 was found after salinity 26.84.

**Principal component analysis**

The first two principal components of the PCA explained 67.5% of the total variance in the dataset (Fig. 6). The first principal component (45.76%) was mostly capturing variation in SUVA254, S, salinity, S275-295 and SR. The second principal component (21.74%) was mostly associated to S350-400, DOC, S275-295 and SR. Brines, oceans and coastal observations clustered together and were associated to high spectral slope values (S, S275-295, S350-400) and SR. On the other hand, lakes, rivers and estuaries correlated positively with high SUVA254 and DOC concentration.

**Discussion**

Additionally, absorption coefficients of CDOM are often reported at different wavelength (254, 350, 400, 440) which make cross-studies comparisons difficult given the non-linear pattern in aCDOM absorption (equation 1).

Bleaching of DOM from lakes → open oceans

**Link with salinity and breakpoint**

Lapierre2013 found a positive relationship between aCDOM(440) and the concentrations of biologically and photochemically degradable DOC in boreal aquatic ecosystems. Based on their results we can conclude the highest rates of bio and physical processing are occurring when DOC is highly colored. Here we found that estuaries-rivers-wetlands were dominated with highly colored DOC suggesting that these ecosystems are strongly contributing to highly process DOC before it reaches the open ocean.

**Outlook and future research**

From the analysis of spectral CDOM absorption and its relationship with DOC, it is evident that the level of detail acquired from single wavelength measurements is considerably inferior to the use of spectral information. As the computational capabilities are not likely to present obstacles for utilizing spectral analyses, we strongly recommend researchers to use and develop methods that use the full potential contained in the CDOM spectra.

Collecting the dataset used in this study has shown that there are serious shortcomings in our current practices in making the scientific data openly available. We fully acknowledge the considerable amount of work and funds used to acquire the data, but the fact that there have been a lot of resources spent to get the data just emphasizes the rationale to make the best possible use of it. One of the first steps to make the data available (after a reasonable period of exclusive use) would be to use the existing data portals (Pangaea etc.) for uploading and storing the data. After making the data available in any of the public repositories, new cross-platform solutions could be used by the community to use and analyse the stored data.