Robust analytical methods to comprehensively characterise natural organic matter and related disinfection by-products formation potential

Deliverable D1.2, WP1

Project Number	**101081980**
Project Title	Climate-resilient management for safe disinfected and non-disinfected water supply systems
Project Acronym	SafeCREW
Project Duration	November 2022 – May 2026
Call identifier	HORIZON-CL6-2022-ZEROPOLLUTION-01
Due date of Deliverable	31.10.2023
Final version date	30.10.2023
Updated version date	12.07.2024
Dissemination Level	PU (Public)
Deliverable No.	D1.2
Work Package	WP1
Task	T1.2
Lead Beneficiary	UFZ
Contributing Beneficiaries	DVGW-TUHH, POLIMI, UBA
Report Author	Jon Wullenweber (DVGW-TUHH), Beatrice Cantoni (POLIMI), Francesco Trovo (POLIMI), Ilenia Epifani (POLIMI), Manuela Antonelli (POLIMI), Maolida Nihemaiti (UFZ)
Reviewed by	Thorsten Reemtsma (UFZ), Mathias Ernst (DVGW-TUHH)
Approved by	Anissa Grieb (DVGW-TUHH)

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

History of Versions		
Version	**Publication Date**	**Change**
1	30.10.2023	Submitted to Commission
1.1	12.07.2024	Final review
2	12.07.2024	Submitted to Commission

History of Changes	
Date/section	**Nature of change and reason**
11.07.2024 page 2	To address the reviewers' comment: When updating the deliverable, please include in the history of changes where and what have been added comparing to the initial version of the deliverable. This will facilitate to review the deliverable": Previous "History of Changes" renamed to "History of Versions". "History of Changes" added.
11.07.2024, page 4, Abstract	To address the reviewers' comment: "The abstract should be reviewed to make it readable (4th paragraph)": The last sentence in 4th paragraph was modified as follows: "This approach holds practical significance as it leverages straightforward and simpler methods to draw conclusions about NOM properties typically assessed through more time-consuming and expensive measurement methods."
11.07.2024, page 26, section 4.3	To address the reviewers' comment: "Provide a conclusion to section 4.3: is a correlation possible between NOM and DBP formation?", following paragraph was added: "In summary, we proved the possibility to use absorbance that is a less time and cost-expensive method with respect to fluorescence, for a preliminary NOM characterization also related to the treatment process applied. This finding opens to further insights in linking different methods to characterize NOM to be used for the evaluation of process performance in removing NOM, as precursors of DBPs."
11.07.2024, pages 33, section 6	To address the reviewers comment: "Provide a conclusion to section 5", the relevant section number was added to each paragraph of section 6 "Conclusion", so the readers can easily relate each sub-conclusion to its relevant section in the main text. The conclusion for section 5 was written in the 5th paragraph of section 6 "Conclusion" in the original version of this document.
11.07.2024, page 18 (section 3.1), page 19 (section 3.2), and page 33 (section 6)	The reviewers commented that "Better demonstrate the robustness of the following statement "by correlating the results from the 3 tested analytical methods about NOM with the formation of DBPs, it becomes possible to develop predictive models for anticipating DBP formation in various raw water sources". We believe that this comment was related to original sentences in the 2nd paragraph of section 6 "Conclusion" about the correlation of results from LC-OCD/-UVD/-OND. Following this comment, more explanations were added to this statement in section 6 as follows:

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

	"By correlating LC-OCD/-UVD/-OND results with the formation of DBPs, it becomes possible to associate distinct NOM fractions with the resulting DBPs. Notably, a strong correlation between the humic fraction and THMs has already been demonstrated. The developed method underscores the potential and establishes a foundation for enhancing existing predictive DBP models by incorporating factors such as the humic fraction. The ongoing data collection in this project aims to expand the prediction capabilities based on NOM composition." To further justify this conclusion, a new figure showing the LC-OCD fingerprints of each water matrix and its trihalomethane formation potential was added (Figure 9a) to section 3.1 and following paragraph was inserted in section 3.2: "The results presented in Figure 9 are normalized to 1 mg of dissolved organic carbon (DOC). It is therefore evident that DOC content alone does not provide a reliable prediction of THMs formation. Similarly, the specific ultraviolet absorbance (SUVA) cannot always accurately predict disinfection by-products (DBPs) such as THMs. Empirical models typically incorporate a combination of DOC, UV254, and bromide concentration to predict THMs or other DBPs (Chen and Westerhoff 2010). By employing LC-OCD, we aim to enhance such models and predictions by adding another parameter: the fraction of humic substances. From Figure 9, it is apparent that the highest yield of THMs per mg of DOC was observed in the water matrix where humic substances constituted the largest fraction. The results obtained in these experiments establish the baseline methodology. Through further data collection in subsequent tasks of the project, such as T2.2, T2.5, and WP3, the data gathered will be correlated with DBPs like THMs and novel sulfonated DBPs. Existing predictive models will be revised or expanded to include fractions identified through LC-OCD analysis."
11.07.2024, page 1, report author list	The names of two report authors: Ilenia Epifani (POLIMI), Manuela Antonelli (POLIMI), were missing in the original version of this document. This information is now added.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Abstract

This deliverable describes the development and application of advanced analytical tools, namely liquid chromatography organic carbon detection (LC-OCD) (i), UV absorbance and fluorescence spectroscopy (ii), and Fourier transform-ion cyclotron resonance mass spectrometry (FT-ICR-MS) (iii), that were established in SafeCREW to characterize natural organic matter (NOM). These methods can be used to link NOM characteristics with the formation potential of disinfection by-products (DBPs), depending on the disinfection method applied.

For demonstration purposes, these analytical tools were applied to five different water matrices: three synthetic waters prepared with different NOM extracts and two real drinking water samples collected from case study #1 and #2, were disinfected with chlorine and chlorine dioxide (ClO_2) in lab-scale kinetic experiments. Samples before and after disinfection were analysed.

LC-OCD is a powerful analytical technique for the investigation of different fractions of NOM. These fractions are characterized by their different molecular size and can be clustered into biopolymers, humic substances, building blocks, low molecular weight acids, and neutrals. During the kinetic experiments with chlorine and ClO_2, it was found that larger molecules, especially humic substances, partially break down into smaller compounds such as building blocks. LC-OCD helps to deepen the understanding of NOM changes during chlorine-based disinfection and allows a more precise control of water treatment processes applying disinfection.

Apart from LC-OCD, simpler and widely used measurement methods were explored also, such as UV absorbance and fluorescence spectroscopy. Specifically, both techniques were applied to assess the kinetics of organic matter transformation during chlorine-based disinfection tests. Moreover, a clustering approach to establish connections between basic NOM measurement methods and those derived from more intricate and resource-intensive techniques was developed. This approach holds practical significance as it leverages straightforward and simpler methods to draw conclusions about NOM properties typically assessed through more time-consuming and expensive measurement methods.

In addition, FT-ICR-MS coupled to a LC was applied to obtain the molecular level information of NOM. The ultrahigh mass resolution and mass accuracy of FT-ICR-MS enabled accurately assigning the molecular formula of thousands of NOM components. The elemental ratio plots, i.e., hydrogen to carbon (H/C) *vs.* oxygen to carbon (O/C), as well as the average mass and aromaticity plots, allowed studying the structural changes in NOM composition during chlorine disinfection. FT-ICR-MS data also gave insights into the chemical composition and structures of DBPs, providing the possibility of tracing their potential precursors. Moreover, coupling FT-ICR-MS to a LC allowed direct injection of samples after a simple freeze-drying enrichment, preventing the loss of NOM fractions due to the complicated sample pre-treatments.

The analytical methods developed here will be applied in WP2 and WP3 for investigating the treatment specific changes in NOM and for studying the association of NOM properties with DBP formation potential in drinking water treatment plants and distribution networks at lab, pilot, and full-scale.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Table of contents

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Abbreviations

CDOC	Chromatographic DOC
DBP	Disinfection byproducts
DOC	Dissolved organic carbon
DON	Dissolved organic nitrogen
DOM	Dissloved organic matter
DW	Drinking Water
DWTP	Drinking Water Treatment Plant
EC	European Commission
FT-ICR-MS	Fourier transform-ion cyclotron resonance mass spectrometry
GA	Grant Agreement
HOHLO	Hohlohsee
HAAs	Haloacetic acids
LC	Liquid chromatography
LMW	Low molecular weight
MF	Molecular formula
N-DBPs	Nitrogenous disinfection byproducts
NOM	Natural organic matter
OC	Organic carbon
OCD	Organic carbon detection
OND	Organic nitrogen detection
SRNHO	Mixture of SRNOM and HOHLO
SRNOM	Suwannee River NOM
THMs	Trihalomethanes
UV	Ultraviolet
UV254	UV adsorption at 254 nm
UVD	Ultraviolet detection
WP	Work package

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

1 Introduction

This deliverable will present a comprehensive analytical procedure for characterizing natural organic matter (NOM) in drinking water by combining different analytical tools such as size exclusion chromatography, fluorescence spectroscopy, and Fourier transform-ion cyclotron resonance mass spectrometry (FT-ICR-MS). Three different synthetic water matrices were prepared using different NOM extracts and real drinking water samples were collected from SafeCREW case study #1 and #2. All analytical methods mentioned above were applied on these same samples to thoroughly characterize their NOM.

Additionally, the characteristics of NOM were associated with the formation potential of disinfection by-products (DBPs). This was achieved by disinfecting the tested waters in lab-scale experiments (with chlorine and ClO_2) and developing a correlation analysis based on existing data and trihalomethane (THMs) concentration. Furthermore, an in-depth analysis was conducted using an extensive dataset of raw and treated waters to assess the potential of utilizing a simplified (cost-effective) measurement technique for making inferences about the results obtained from more intricate methods and to develop a method for practical application.

1.1 NOM characterization with LC-OCD/-UVD/-OND

NOM is the main precursor of DBPs (Hua and Reckhow, 2007). Conventional methods for NOM quantification such as dissolved organic carbon (DOC) and UV254 offer basic summarily assessments of organic substances in water treatment. In contrast, Liquid Chromatography Organic Carbon Detection (LC-OCD) combined with UV Detection (-UVD) and Organic Nitrogen Detection (-OND) represents a state-of-the-art approach that provides a more comprehensive characterization of NOM in water samples, offering significant advantages in understanding organic matter behaviour. In the following, it will be collectively referred to as LC-OCD. As described by Huber et al. (2011), _ENREF_10 it separates NOM in its different fractions based on its size and polarity, thus providing a detailed and comprehensive analysis of its composition. The addition of UVD enhances LC-OCD's capabilities by allowing for the quantification of specific organic compounds within the NOM matrix. The chromatograms separate the chromatographic DOC (CDOC) into various fractions, including Iron/Humic Colloids, Polysaccharides, Humic Substances, Building Blocks, Low Molecular Weight Compounds (containing low molecular weight acids), Amphiphilic Substances and Neutral Compounds. In addition, ammonium and dissolved organic nitrogen (DON) are determined in a side stream by the OND (Figure 1).

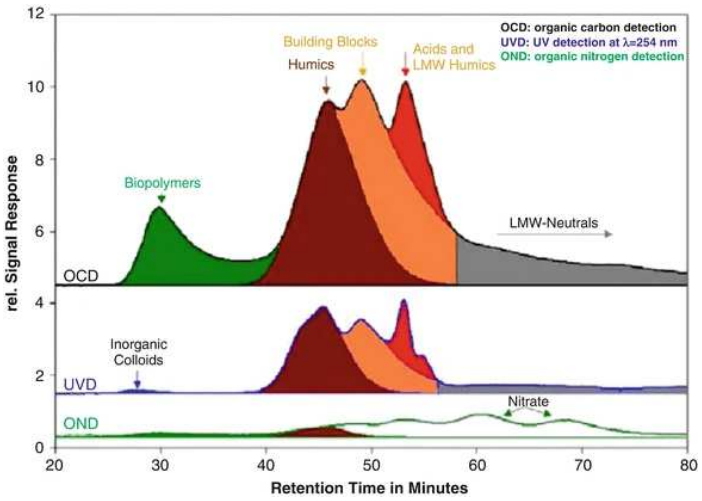


Figure 1. LC-OCD/-UVD/-OND Chromatogram with highlighted fractions (DOC-Labor, 2006)

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

1.2 NOM characterization with absorbance and fluorescence spectroscopy and their link with LC-OCD data

Absorbance and fluorescence are among the most studied and used techniques in the characterisation of NOM in water (Chen and Yu, 2021). NOM is a heterogeneous mixture of compounds, including humic acids, fulvic acids, low molecular weight acids, carbohydrates, proteins and other classes of compounds. As part of its characterisation, fluorescence provides information on the chemical nature according to the structure and functional groups of the organic molecules in the samples (Świetlik and Sikorska, 2004). The absorbance technique, mainly in the 220 - 280 nm wavelength range, is considered to be the most appropriate for measuring NOM. Carboxylic and aromatic chromophores are associated to wavelengths around 220 nm and aromatic groups to 254 nm, thus being useful for assessing the aromaticity of molecules (Matilainen et al., 2011).

The evaluation is done by using water with different types of NOM in order to highlight the possibility of the formation of a different amount of by-products from different loads and types of organic matter. For the analysis of the results, absorbance and fluorescence techniques are used to characterise the raw and disinfected water.

We have thoroughly explored the statistical analysis of spectroscopic measurements of water, encompassing UV-Vis absorbance spectra, fluorescence spectra, and LC-OCD measurements. Our focus is on examining the relationships among these multidimensional and intricate measurements. Given the complex correlation between them, conducting a regression analysis of the data is unfeasible. Therefore, we propose employing a more sophisticated technique based on the hierarchical clustering method, which aids in grouping similar samples.

1.3 NOM characterization with LC-FT-ICR-MS

Mass spectrometric analysis is a method to provide molecular level information of NOM. Ultrahigh resolution mass spectrometry such as Orbitrap and FT-ICR-MS is the state-of-the-art analytical tool to determine the molecular composition of NOM by providing sufficient resolution, mass accuracy, and sensitivity (Reemtsma, 2009). The elemental ratio plots that derived from FT-ICR-MS data, such as the hydrogen to carbon (H/C) *vs.* oxygen to carbon (O/C), could provide insights into the structural characteristic of NOM (e.g., aromaticity, functional groups) (Kim et al., 2003).

FT-ICR-MS has been applied to track the changes in NOM and to study DBP formation mechanism during drinking water production (Lavonen et al., 2015; Milstead and Remucal, 2021; Han et al., 2023). Direct infusion following solid-phase extraction is commonly used as sample introduction during FT-ICR-MS analysis. However, this method can result in significant ion suppression due to the competing ions in complex water matrices. Furthermore, solid-phase extraction cannot fully recover the dissolved organic matter, especially the polar fractions (Jennings et al., 2022). Coupling FT-ICR-MS to a LC can overcome this limitation by chromatographically separating the water matrices from organic matter composition (Han et al., 2021). LC-FT-ICR-MS also allows the direct injection of water sample without losing the dissolved organic matter fractions during sample pre-treatment (Jennings et al., 2022).

1.4 Association of DBP formation potential with NOM properties

Since LC-OCD can be used for extended NOM characterization, it plays a significant role in understanding DBP formation. LC-OCD allows for a detailed analysis of the composition and nature of organic matter present in the water. This information is crucial in understanding the precursor materials that can react with disinfectants (such as chlorine) during the water treatment process, leading to the formation of DBPs. By characterizing the specific organic components and their concentrations, it becomes possible to predict and assess the potential for DBP formation (Wassink et al., 2011). Certain organic compounds are more prone to react with disinfectants and produce DBPs,

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

which can be harmful to human health. Nitrogen-containing disinfection byproducts (N-DBPs) currently lack regulatory guidelines; however, their potential toxicological implications have been demonstrated (Muellner et al., 2007; Shah and Mitch, 2012). Utilizing Organic Nitrogen Detection (OND) to study changes in dissolved organic nitrogen during disinfection is a crucial approach to gain valuable insights into the formation of N-DBPs. The LC-OCD method helps to identify and manage DBP precursors, allowing for better control and mitigation of DBP formation during the treatment process.

The use of absorbance and fluorescence is not only concerned with the definition of precursors, but also with the analysis of the prediction of DBP formation, an area in which these techniques are still being studied, as the connection between them and the prediction of the amount of DBPs that could be formed, often linked to parallel factor analysis (PARAFAC) analysis, has not yet been fully investigated (Fernández-Pascual et al., 2023). From the literature reviewed, it is evident that studies are mainly concerned with the joint analysis of NOM and DBPs. This is due to the strong link known between organic matter and the formation of chlorinated by-products in the presence of chlorination processes. These studies mainly analysed surface water and drinking water, focusing less on groundwater or water from bank filtration processes. Hypochlorous acid (HClO) is the disinfectant, deriving from processes using mainly Cl_2, most widely used and studied in the field of disinfection, where its tendency to form trihalomethanes (THMs) and haloacetic acids (HAAs) in the presence of organic precursors is now evident. On the other hand, the available literature is less dense with studies on the use of other types of disinfectants such as chlorine dioxide (ClO_2) and ozone. Studies have been carried out with ClO_2, which would lead to the formation of fewer THMs and HAAs at the expense of chlorite and chlorate generation (Yang et al., 2013).

2 Material and Methods

2.1 Chemicals

Sodium Hydroxide (≥ 99 %), Potassium dihydrogen phosphate (≥ 99 %), Sodium phosphate dibasic (> 99 %), Sulfuric Acid (95-97 %) and ortho-Phosphoric acid (≥ 85 %) were purchased from Merck (Darmstadt, Germany), Sodium bisulfate (≥ 37 %) and Sodium hypochlorite (12 %) were purchased from Carl Roth (Karlsruhe, Germany). A 2 $gClO_2/L$ stock solution was generated weekly using the Oxiperm® Pro OCD-162 generator (5 $gClO_2/h$, Grundfos, IT) using sodium chlorite and hydrochloric acid, both technical grade (Chimitex, IT). It was stored in headspace-free amber glass vials at 4 °C. All the chemicals used for ClO_2 measurements were analytical grade and purchased from Sigma Aldrich (USA), except for DPD salt, purchased from Hach Lange (EU) and used through a DPD dispenser (DPD Free Chlorine Reagent, Swiftest™ Dispenser).

2.2 Water samples/NOM extracts

In order to investigate NOM originating from different source waters five different water matrices were used. Suwanee River NOM (SRNOM) represents NOM sourced from rivers, while Hohlohsee (HOHLO) represents NOM derived from lakes. For a more comparative analysis, and to represent a raw water with NOM from various sources, a 1:1 Mixture of SRNOM and HOHLO was used and referred to as “SRNHO”. To represent groundwaters real water samples from Case Study #CS1H and Case Study #CS2M were collected after treatment prior to disinfection.

SRNOM is extensively studied, characterized and their NOM properties are well-documented. IHSS Suwanee River NOM Isolate 2R101N was obtained from DOC Labor Huber (Karlsruhe, Germany). Hohlohsee is a bog lake located in a conservation zone in an upland moor in the Black Forest (Germany). A sample (ID HO31) was provided by the Engler-Bunte-Institute Karlsruhe (Germany) which was collected on 29.07.2021. A comprehensive overview of Hohlohsees water composition is given by

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101081980.

Frimmel et al. (2002). The SRNOM isolate was dissolved in ultra-pure water (Milli-Q®) and gently stirred for one week. #CS1H Tap water is a mixture of deep and shallow groundwater which has been treated by aeration (oxidizer), rapid filters and deacidification. #CS2M water is a deep groundwater which has been treated by granular activated carbon filters with an Empty Bed Contact Time (EBCT) of 11 minutes and sampled before the disinfection step of the full-scale DWTP.

Prior to usage all samples were pre-filtered by 0.45 µm membrane filtration (polyamide, Carl Roth, Germany). The pH was adjusted to 7 for all samples using 0.1 M NaOH and 0.1 M H_2SO_4. SRNOM, HOHLO and SRNHO have been diluted to obtain a final DOC of 3 mg/L. The composition of the used water matrices is listed in Table 1.

Table 1: Overview of the used waters

Composition of the synthetic water	Base matrix	SRNOM Isolate [mg/L]	HOHLO water [ml/L]	DOC [mg/L]	pH
SRNOM	Ultra pure water	7.2		3	7
HOHLO	Ultra pure water		171	3	7
SRNHO 1:1	Ultra pure water	3.6	85.5	3	7
Real waters					
CS1H_DWTP1	DWTP1 treated water before disinfection			3.4	7
CS2M_DWTP1	DWTP1 treated water before disinfection			1.0	7

2.3 Lab-scale chlorination and ClO_2 disinfection

For lab-scale chlorination and DBP formation potential (DBP-FP) tests the five aforementioned water matrices were used. Labware was pre-treated to remove any chlorine demand by soaking the labware in 1 ml/L commercial bleach solution for a minimum of 1 hour. Labware was rinsed afterwards with ultra-pure water. Prior to disinfection each bottle was rinsed with the corresponding sample. Lab-scale kinetics were investigated employing two distinct disinfectants: sodium hypochlorite (NaOCl) and chlorine dioxide (ClO_2). Preliminary tests were conducted to determine the appropriate disinfectant dosage required to achieve a target concentration of 5 mg/L of free chlorine. For ClO_2 tests, the disinfectant solution at the desired ClO_2 dosage concentration (1.92 $mgClO_2$/L that corresponds to 5 $mgCl_{2\text{-}eq}$/L) was prepared diluting stock solution (2 $gClO_2$/L) in ultrapure water. Desired disinfectant dosage was added using a pipettor to prevent highly localized areas of chlorine by submerging the pipet tip under water. Samples were afterwards inverted to mix and incubated in brown glassware according to the test plan (20 ± 5 °C). Samples were taken after five contact times (0.5 h, 2 h, 6 h, 24 h, 48 h). For residual chlorine concentration measurement, an 8-point calibration curve (9 replicates for each point) for NaOCl was determined according to the 4500-Cl.G method (APHA, 2017), using potassium permanganate solution as standard, where 0.891 mg/L potassium permanganate solution corresponds to 1 $mgCl_2$-eq/L as NaOCl in the DPD reaction. Absorbance at 515 nm of each solution for the calibration curve determination was measured immediately. The one for ClO_2 was obtained by the previous one by dividing the slope by 1.9, which is the ratio between the equivalent weights (mg/eq) of ClO_2 and Cl_2 (APHA, 2017). The NaOCl calibration curve covers the range 0.02-0.8 $mgCl_2$-eq/L, while the ClO_2 calibration curve the range 0.038-1.52 $mgClO_2$/L. If no immediate analysis was possible samples where quenched and cooled according to the test plan. For THM analysis 25 mg Ascorbic Acid per vial was used as a quenching agent. For chlorite measurements, nitrogen gas was bubbled for 1 minute as a quenching agent. For a comprehensive FT-ICR-MS analysis after 48 h contact time

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

unquenched samples were prepared and shipped without cooling to UFZ immediately after disinfection and quenched samples (15 µl/L sodium bisulfate) were shipped after 48 h contact time.

2.4 LC-OCD analysis

NOM was analysed by liquid chromatography (LC) coupled with an online organic carbon detection (-OCD) and UV detection (-UVD) at λ = 254 nm and organic nitrogen detection (-OND) at λ = 220 nm (Figure 2). The mobile phase undergoes UV-Cleaning in the DOCOX-reactor before being transferred to an autosampler and then introduced into the chromatographic column. The liquid passes through the UV detector (UVD) followed by the Organic Carbon Detector (OCD). Additionally, a side stream from the UVD is directed into a specialized capillary UV-lamp system (DONOX), where a secondary UV detector measures at 220nm. Within the DONOX-reactor, both Dissolved Organic Nitrogen (DON) and ammonium undergo oxidation, transforming into nitrate for subsequent analysis. Phosphate buffer (KH_2PO_4 2,38 g/L; Na_2HPO_4 1,332 g/L) serves as the mobile phase, while the acidification solution consists of phosphatic acid (ortho-Phosphoric acid 4 mL/L). The quantitative analysis by peak integration is carried out using ChromCALC Software, developed by DOC Labor Huber (Germany). Calibration standards for the OCD, UVD and OND detectors were established using standard compounds commercially available: IHSS SR Fluvic Acid (FA) and IHSS SR Humic Acid (HA) 1S101H.

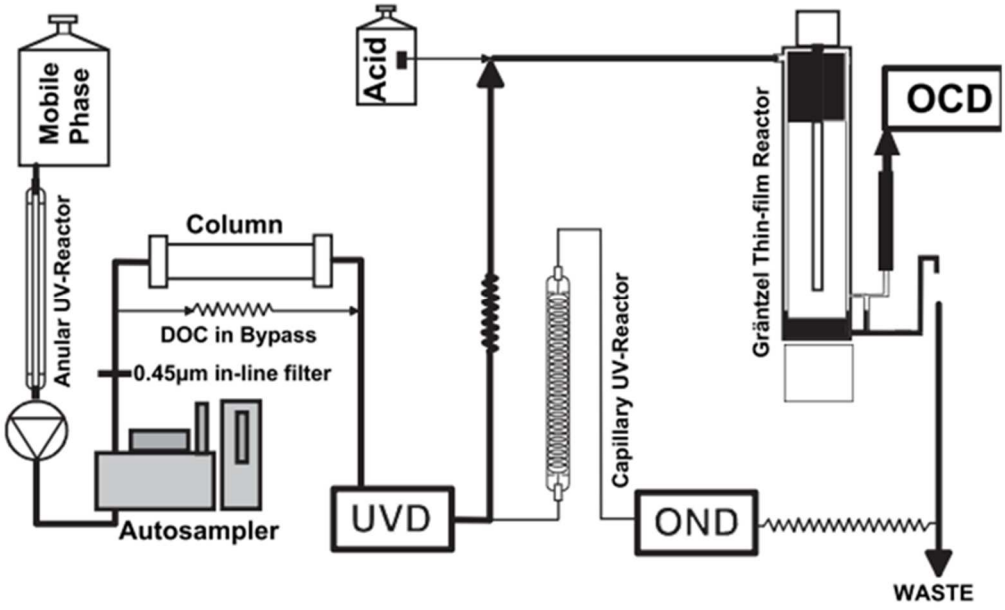


Figure 2. Flow scheme of LC-OCD/-UVD/-OND system (Huber et al., 2011)

2.5 Absorbance and fluorescence analyses

The absorbance spectrum (190-900 nm) was measured using the Hach Lange UV-VIS DR6000 spectrophotometer. For fluorescence measurements, the Agilent Cary Eclipse fluorescence spectrophotometer was used. For both absorbance and fluorescence spectrum measurements, a quartz cuvette with a 1-cm optical path having all four transparent faces was used. For each analysis, the cuvette was rinsed with deionized water and subsequently with the sample before making the measurement. Fluorescence measurement consists of evaluating the emission spectrum of a matrix as the excitation wavelength changes and allows for Excitation Emission Matrix (EEM) and Raman measurement. Ultrapure water was used as a reference for both the EEM Raman measurement and

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

the spectrophotometer absorbance measurement. The parameters for the evaluation of EEM and Raman value are given in Table 2. The final fluorescence assessment was carried out both in terms of total fluorescence (integral of the EEM, corrected for scatters) and in terms of peaks. In fact, using fluorescence values in certain regions of the spectrum, indices can be calculated, which are traditionally linked to organic compounds, such as humic and fulvic acids (A, C, M, D) and proteins (B, T, N) (Wünsch et al., 2019). The location of the peaks is shown in Figure 3. The MATLAB program (version R2022b) was used to process the fluorescence data, specifically the drEEM-0.6.3 package was used, and a code was developed by which Raman and Rayleigh scatters were corrected through the absorbance spectrum performed on the respective sample.

Table 2: Parameters for the evaluation of EEM and Raman value.

Parameter	Unite of Measure	Raman	EEM
Voltage	V	600	600
Corrected Spectra	-	yes	yes
Excitation λ	nm	350	365-450
Emission λ	nm	365-450	270-600
Emission and Excitation Slit	nm	5	5
Scan speed	nm/min	Slowest	1.200 (Fast)

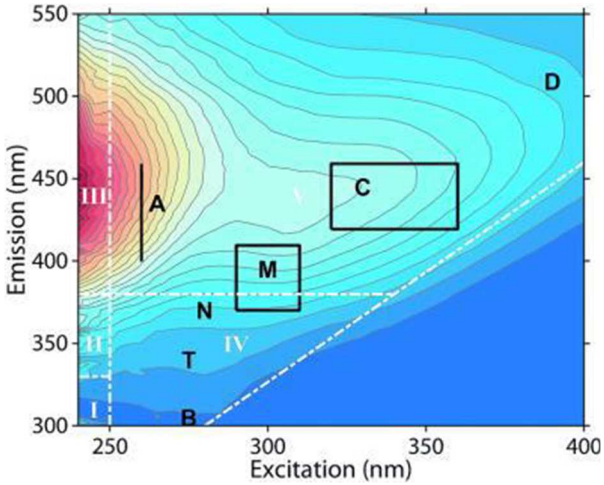


Figure 3: Location of the peaks in a fluorescence EEM matrix (Wünsch et al., 2019)

2.6 Clustering and Correlation Analysis between different NOM characterization techniques

In this section, the employed statistical methodology to correlate results obtained by different NOM characterization techniques is outlined. A detailed description of the data preprocessing steps is provided: the used dissimilarity measures, the hierarchical clustering techniques applied, and the postprocessing strategies employed to derive meaningful insights from these spectroscopic measurements.

2.6.1 Data Pre-processing

The pre-processing steps have been designed to address noise, reconstruct missing data, and standardize sampling values across datasets developed by different teams with several measurement tools. We recall that UV-Vis and LC-OCD measurements are represented as vectors, while fluorescence data is represented as a matrix. However, the following methodologies are applicable to both types of data.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

For one-dimensional data such as UV-Vis absorbance spectra and LC-OCD profiles, we smooth them by means of cubic smoothing splines (Green and Silverman, 1993). Cubic smoothing splines provide a flexible and effective means to smooth the data while preserving important features and trends. They help mitigate random fluctuations and outliers, resulting in cleaner and more reliable data for downstream analysis.

Instead, in the case of two-dimensional data, like fluorescence spectra, we use the thin-plate smoothing splines, that make multidimensional data well-suited for smoothing 2D spectra. Thin-plate smoothing splines maintain the integrity of spatial relationships within the data while reducing noise and enhancing the clarity of underlying patterns. An example of this process is presented in Figure 4.

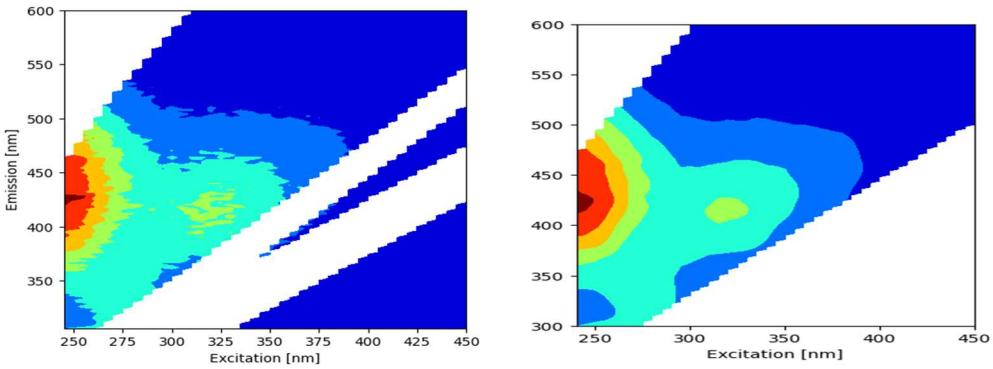


Figure 4: Fluorescence spectrum that has been smoothed and limited to the relevant region for fluorescence spectroscopy for NOM characterization. Raw data (left) and smoothed version (right).

The Generalized Cross-Validation (GCV) technique was used to fine-tune the smoothing parameters required for both spline methods. The GVC method enables us to find the optimal equilibrium between reducing the erratic fluctuations in the output function while preserving the fundamental patterns exhibited by the data. The documentation on the libraries used for this analysis is available in the R software documentation (Nychka et al., 2021; R Core Team, 2021).

Since the output of spline smoothing consists of mathematical functions, there is the flexibility to sample them at any frequency and time. This ensures both uniform sampling across all available datasets, and the computation of spectra at specific points, highlighted as significant in the NOM characterization literature (Matilainen et al., 2011; Wünsch et al., 2019; Li et al., 2020).

The smoothed data are then normalized, dividing them by the area/volume under the curve/surface. Working with such normalized data allows us to focus the analysis solely on the shapes and disregard about the absolute intensity of the data.

2.6.2 Dissimilarity Measure

A hierarchical clustering procedure is based on the choice of a dissimilarity measure between samples. For our spectroscopic data analysis, we opt for the Canberra distance, which is a standard choice in the framework of hierarchical clustering when dealing with positive quantities (Gao et al., 2023). For two points x, y its formal definition is as follows:

$$D_C(x,y) \stackrel{\text{def}}{=} \frac{|x - y|}{|x| + |y|}.$$

So that the dissimilarity between two actual/real absorbance spectra turns out to be:

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

$$D(A_1, A_2) = \int \frac{|A_1(\omega) - A_2(\omega)|}{|A_1(\omega)| + |A_2(\omega)|} d\omega.$$

Whereas its approximated value, computed on the resampled data, becomes:

$$D(A_1, A_2) = \sum_i \frac{|A_1(\omega_i) - A_2(\omega_i)|}{|A_1(\omega_i)| + |A_2(\omega_i)|},$$

(ω_i are the sampled wavelengths of the considered absorbance spectra). An example of a heatmap of the dissimilarities between pairs of absorbance spectra is shown in Figure 5: on the left, we have some different spectra (wavelength vs. intensity), and on the right, each pixel of the map represents a pairwise dissimilarity (the closer the two absorbance spectra are, the lighter the pixel).

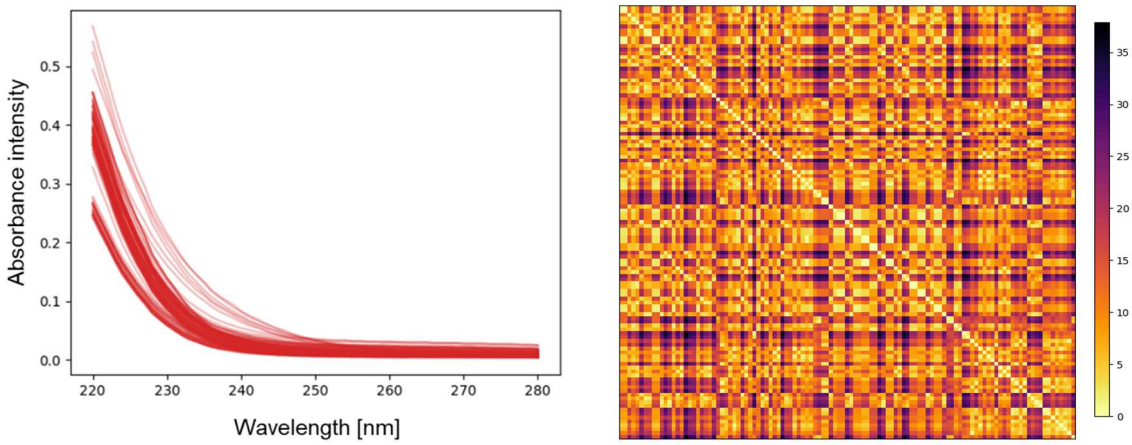


Figure 5: Conversion of a dataset of absorbance spectra (left))to a matrix of pairwise dissimilarities (right).

2.6.3 Hierarchical Clustering

After calculating the distances between different samples' measurements, groups of samples with similar NOM characteristics (measured by each analytical method) were identified. Hierarchical clustering is a powerful data analysis technique developed to address unsupervised learning tasks (Everitt et al., 2011). This technique groups data points into a hierarchical tree-like structure known as dendrogram. It proves particularly valuable when exploring complex datasets with the aim of identifying inherent structures, patterns, or relationships.

An example of a dendrogram is provided in Figure 6 (left). At the bottom of the tree, we find the leaves, each representing a sample. As we move up from the bottom, branching represents the clustering of nearest samples. The distance between two distinct clusters is determined through a process called "linking". In our case, we have chosen an average linking method, which estimates the distance between two clusters by averaging all pairwise distances between the samples in those clusters. This approach allows us to group different clusters together using the same rule at each level. Following this process, we eventually cluster all samples into a single cluster given by the highest part of the tree known as the root.

Finally, the analysis of the results of this procedure involves studying the differences between each cluster while varying the analytical tool used for clustering and the depth at which we cut the tree. An example of cutting the tree to generate 5 clusters, each depicted in distinct colours, is shown on the right in Figure 6.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

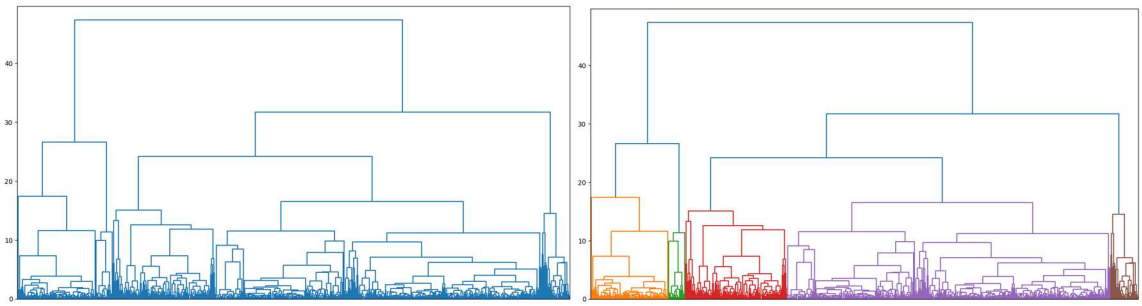


Figure 6: Dendrogram (left) and corresponding clustering generated from a cutting (right).

All computations and graphics for dissimilarity and clustering were coded in the Python library Scipy (Virtanen et al., 2020).

2.6.4 Post-processing

After clustering the samples according to each measurement method, a post-processing is needed to correlate results obtained with different methods. In fact, the post-processing tools are essential for translating the results of our methodology into actionable insights for future research and practical applications. A contingency table was constructed, as it is commonly used for discovering and visualizing relationship between two attributes. This technique organizes data into rows and columns, where each cell in the table represents the intersection of corresponding categories for each attribute (Dodge, 2008). In our context, we used these tables (see two examples in Figure 7 and 8) to identify relationships between pairs of clustering: on the columns we find the different clusters obtained by absorbance spectra, and in the rows the clusters obtained by the fluorescence EEM. In the contingency table, in each cell it is reported the number of samples falling into the intersection between two clusters built with each measurement method. Furthermore, we introduce a transformation that highlights the clusters with higher correlations. This transformation is colour-coded to represent positive correlations in green and negative correlations in red. The transformation is mathematically defined as the difference between the contingency matrix and the tensor product of the marginal empirical distributions, which are visualized as histograms within the contingency matrix.

We present two extreme examples, both to illustrate the construction of the matrices described above and the potential outcomes of this analysis.

The worst-case scenario represents complete independence among clusters, where knowledge of the clustering for one variable (i.e. absorbance) gives no information about the other (i.e. fluorescence); in fact, for each cluster of absorbance there is equal probability to fall in each one of the fluorescence clusters and the correlations (shown in the right graph) are low. An example of this kind of contingency table (left) and the transformed values (right) are presented in Figure 7.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

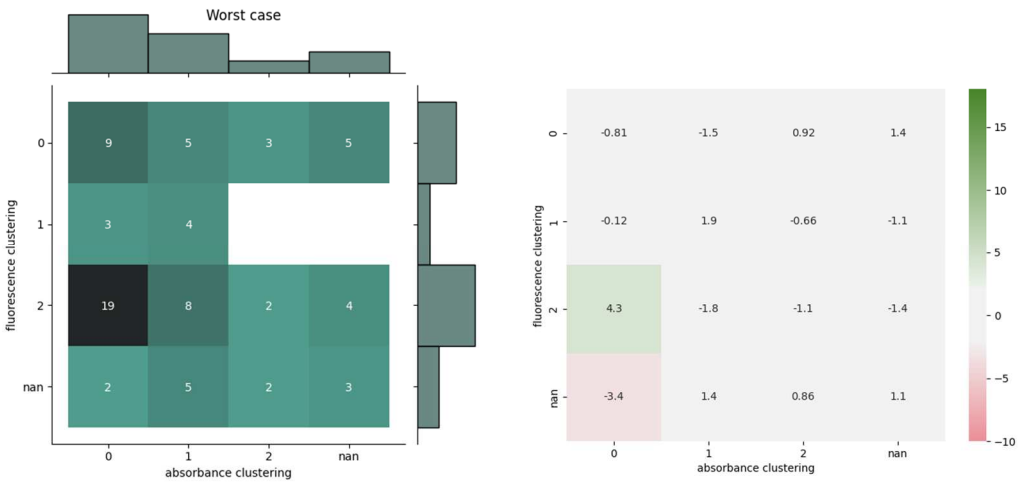


Figure 7: Example of worst-case scenario. Contingency table (left) and transformed values (right).

Conversely, the best-case scenario, exemplified in Figure 8, is when for each cluster of one feature (i.e. absorbance), we have a 100%-correspondence with one cluster for the other (i.e. fluorescence). Therefore, we can identify the cluster of a measurement by only providing the corresponding cluster for the other measurement. In this case, only measuring the absorbance spectrum of one sample, we would know in which absorbance cluster it falls and, consequently we would know also in which fluorescence cluster it would fall and which fluorescence EEM characteristics it has, without the need for fluorescence measurements.

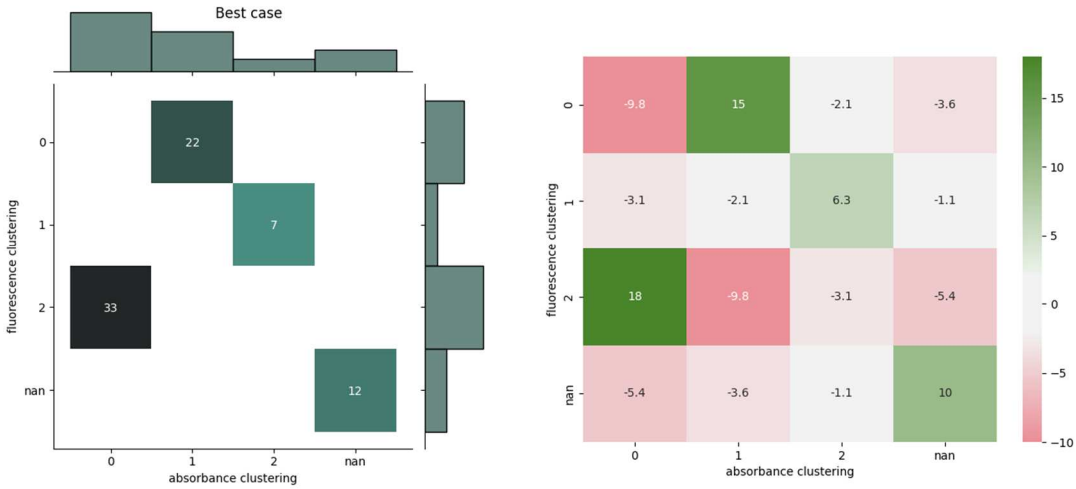


Figure 8: Example of best-case scenario. Contingency table (left) and transformed values (right).

2.7 LC-FT-ICR-MS analysis

Samples were enriched using freeze-drying method before LC-FT-ICR-MS analysis. A 40 mL aliquot of water sample was filled in a 50 mL centrifuge tube and stored at -20 ℃ overnight. The frozen sample was then placed in a freeze-dryer (Alpha1-4, Christ, Germany) at 15 °C and 1.25 mbar for 30 h until dryness. The residue was reconstituted in 1 mL of ultrapure water, transferred into an Eppendorf tube, and then centrifuged at 13000 min^{-1} for 10 min. The supernatant was then transferred into glass vials after filtering through 0.2 µm syringe filters (Sartorius Minisart) and kept at -20 °C until analysis. An ultrapure water blank was prepared following the same procedure.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

The enriched samples were analysed with FT-ICR-MS (SolariX XR, Bruker Daltonics Inc., Billerica, MA, USA) coupled to a UHPLC system (UltiMate 3000RS, Thermo Fischer Scientific, Waltham, MA, USA) following a method published previously (Han et al., 2021; Jennings et al., 2022). The chromatographic separation of DOM was achieved using a C18 column (ACQUITY HSS T3, 1.8 µm, 100 Å, 150x3mm, Waters, USA) with a guard column (ACQUITY UPLC HSS T3 VanGuard Pre-column, 100Å, 1.8 µm, 2.1 mm X 5 mm). The column temperature was set to 30 °C. Ultrapure water and LC grade methanol (MeOH; Biosolve, Valkenswaard, Netherlands) were used as mobile phases A and B, respectively. The pH of mobile phase A was adjusted to pH 3 using formic acid and ammonium hydroxide. The same volume of formic acid and ammonium hydroxide were added to mobile phase B. Also, a second pump (LPG-3400SD) delivered a post-column counter gradient to yield consistent solvent compositions for electrospray ionization (50:50 MeOH: water). The ionization was carried out using electrospray ionization source in negative mode with capillary voltage at 4.3 kV, nebulizer gas pressure at 1.0 bar, dry gas temperature at 250 °C, and dry gas flow rate with 8.0 L/min.

Data were recorded with 2MWord (~0.84s FID, full profile mode) in continuous accumulation of selected ions (CASI) mode. Data were segmented and averaged into 1 min wide segments between 5 and 20 min according to established procedures. Each of the 15 segments was internally calibrated (DataAnalysis 5.0, Bruker Daltonics) with a list of commonly detected DOM masses (*m/z* 150-1000, *n* = 275). Molecular formula (MF) were assigned using the elemental composition of $C_{1\text{-}60}$, $^{13}C_{0\text{-}1}$, $H_{1\text{-}122}$, $O_{0\text{-}40}$, $N_{0\text{-}2}$, $S_{0\text{-}1}$, $^{34}S_{0\text{-}1}$, $^{35}Cl_{0\text{-}3}$, $^{37}Cl_{0\text{-}3}$ within the mass tolerance of ± 0.5 ppm. Other details regarding data processing were published previously (Han et al., 2021; Jennings et al., 2022).

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

3 Results and method development for NOM characterization with LC-OCD

3.1 NOM analysis

In the context of the lab-scale kinetic tests, established procedures for investigating the composition of NOM in water have been further developed at DVGW-TUHH. The chromatographic method is a powerful tool for breaking down the hydrophilic CDOC of water into its individual components.

It has been demonstrated that all five waters examined during the kinetic tests contain humic substances, building blocks, and low molecular weight (LMW) acids. As despicted in Figure 9.A regardless of the source water, humic substances consistently constituted the largest proportion. Groundwater CS1H exhibited a higher presence of building blocks compared to river and lake water surrogates. LMW acids were present in all waters but at low levels. Biopolymers, however, were not detected in any of the waters. The LC-OCD analysis for the waters revealed the typical composition of the CDOC for their respective source waters. This comprehensive classification and subdivision of NOM provide a solid foundation for studying changes in NOM quality after the addition of a disinfectant.

The kinetic tests showed that the most significant UV signal decreases in the investigated bog water (HOHLO), correlating with chlorine consumption. UV signal decreased across all identified fractions. However, the OC signal exhibited varying trends in intensity across all waters. For some waters (SRNOM, HOHLO and SRNHO), the signal of higher molecular weight humic substances weakened and shifted towards smaller molecules, indicated by longer elution times. In these regions, the signal increased. Figure 9 illustrates OC and UV signal trends for water SRNHO during the experimental period.

Furthermore, the qualitative examination of the OND revealed a consistent signal decrease across the entire chromatogram. This underscores the potential for expanding the methodology beyond the investigation of organic carbon materials to include the organic nitrogen component. The method developed here establishes the basis for a connection between NOM and (polar) N-DBPs in addition to the regulated carbon-based DBPs such as HAAs and THMs.

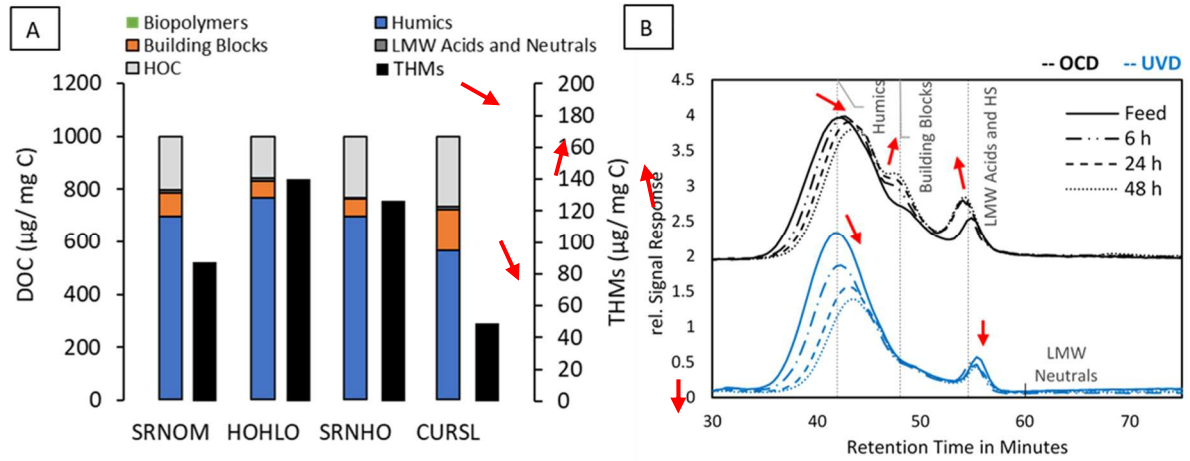


Figure 9: (A) LC-OCD fingerprint of the chromatographable DOC of the four feed water matrices, showing the non-chromatographable hydrophilic organic carbon (HOC) and the fractions of the chromatographable organic carbon (CDOC) with their corresponding fractions, and the formed THMs after 48 hours normalized per mg DOC. (B) Chromatogram of the OCD and UVD signals from the kinetic disinfection test of the SRNHO water matrix

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

3.2 Association with DBP formation

The previously described decrease in the UV signal can be attributed to the degradation and transformation of UV-adsorbing compounds during the disinfection process, resulting in the formation of by-products. The kinetic studies demonstrated an ongoing degradation with increasing retention times of up to 48 hours. The observed changes in NOM via LC-OCD are closely related to chlorine demand and indicate promising initial correlations with the formation of THMs. Humic substances and specific low molecular weight compounds like fulvic acids have been recognized as DBP precursors (Singer, 1999). These fractions are frequently linked to the generation of THMs, HAAs, and other DBPs during disinfection (Nikolaou et al., 2004; Bond et al., 2012a).

The results presented in Figure 9 are normalized to 1 mg of dissolved organic carbon (DOC). It is therefore evident that DOC content alone does not provide a reliable prediction of THMs formation. Similarly, the specific ultraviolet absorbance (SUVA) cannot always accurately predict disinfection by-products (DBPs) such as THMs. Empirical models typically incorporate a combination of DOC, UV254, and bromide concentration to predict THMs or other DBPs (Chen and Westerhoff 2010). By employing LC-OCD, we aim to enhance such models and predictions by adding another parameter: the fraction of humic substances. From Figure 9, it is apparent that the highest yield of THMs per mg of DOC was observed in the water matrix where humic substances constituted the largest fraction. The results obtained in these experiments establish the baseline methodology. Through further data collection in subsequent tasks of the project, such as T2.2, T2.5, and WP3, the data gathered will be correlated with DBPs like THMs and novel sulfonated DBPs. Existing predictive models will be revised or expanded to include fractions identified through LC-OCD analysis.

Additionally, the OND detector proves useful in identifying organic nitrogen-containing compounds, which are also acknowledged as DBP precursors. Nitrogen-containing compounds have the potential to give rise to the formation of nitrogenous DBPs, such as haloacetonitriles (HANs), haloacetamides (HAMs) (Bond et al., 2012b), and nitrogenous sulfonated DBPs (Nihemaiti et al., 2023).

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

4 Results and method development for absorbance and fluorescence spectroscopy and their correlation in NOM characterization and DBP formation prediction

4.1 NOM analysis

Comparison of the absorbance spectra (Figure 10), evaluated in the whole range (190-900 nm), show that both the real waters (CS1H_DWTP1 and CS2M_DWTP1) have absorbance increase for wavelengths lower than 230 nm, that is explained in the literature by interferences from inorganics (Matilainen et al., 2011) that are not present in the three synthetic waters (SRNOM, HOHLO and SRNHO). In the specific range of interest for organics (230-280 nm) spectra highlight similarities between CS1H_DWTP1 and the three synthetic matrices (SRNOM, HOHLO and SRNHO), with values included in a range between 0.17-0.32 cm^{-1}. The CS2M_DWTP1, sampled downstream the activated carbon treatment and immediately upstream of the disinfection, shows lower absorbance values, between 0.02-0.17 cm^{-1}.

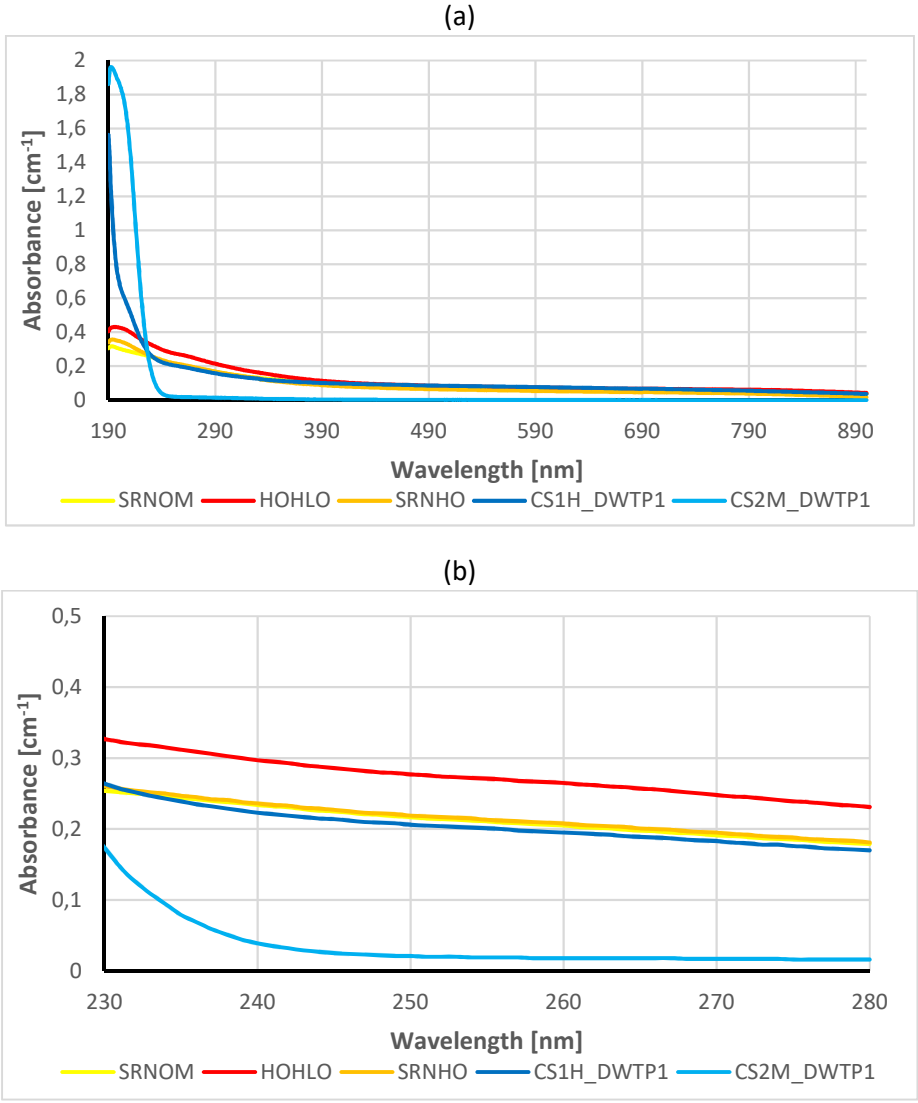


Figure 10: Absorbance spectra, evaluated between wavelength (a) 190-900 nm and (b) 230-280 nm, for each of the five tested feed waters.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Figure 11 shows the 5 feed water fluorescence matrices, reported at the same intensity scale value in order to be easily compared. The CS1H_DWTP1 sample, in the wavelength ranges considered, shows higher values than the three synthetic waters. In CS2M_DWTP1 sample, no peaks are observed in any region of the fluorescence matrix, if compared to the other waters with the same fluorescence intensity range. This might be related to the lower DOC value of CS2M_DWTP1 compared to other waters (1 mg/L vs. ≈ 3 mg/L DOC, Table 1).

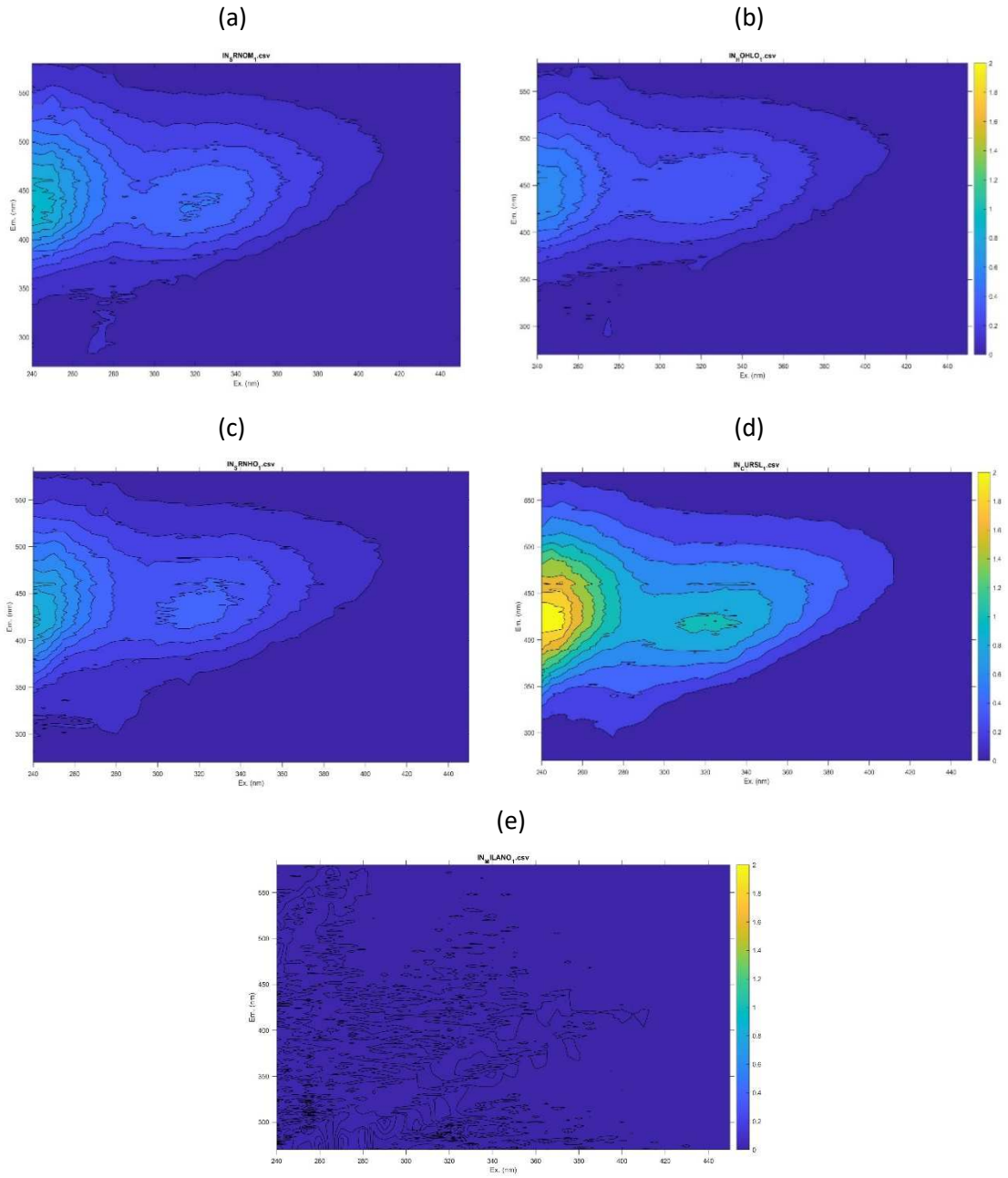


Figure 11: Fluorescence matrices for each of the 5 feed water: (a) SRNOM, (b) HOHLO, (c) SRNHO, (d) CS1H_DWTP1 and (e) CS2M_DWTP1.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

In the context of the lab-scale kinetic tests, established procedures for investigating the composition of NOM in water have been further developed at POLIMI. Studying the trends of both absorbance and fluorescence over contact time (as shown in this paragraph) is an important tool to correlate NOM characterization obtained with the two different measurements (reported in Paragraph 4.3) and to evaluate the corresponding DBPs formation (reported in Paragraph 4.2).

The kinetic tests (Figure 12) showed that, regardless of the source water, absorbance at 254 nm and the total fluorescence decrease with chlorine consumption. The absorbance of HOHLO lake water, with higher UVA_{254} at the beginning of the test, was reduced by 45%. The river NOM (SRNOM), the river and lake waters mixture (SRNHO) and the real water from CS1H_DWTP1, all starting from UVA_{254} of 0.21 cm^{-1}, showed a decrease of 40 to 50%. The CS2M_DWTP1 sample starting at a very low UVA_{254} compared to the other waters, was subjected to a lower removal rate.

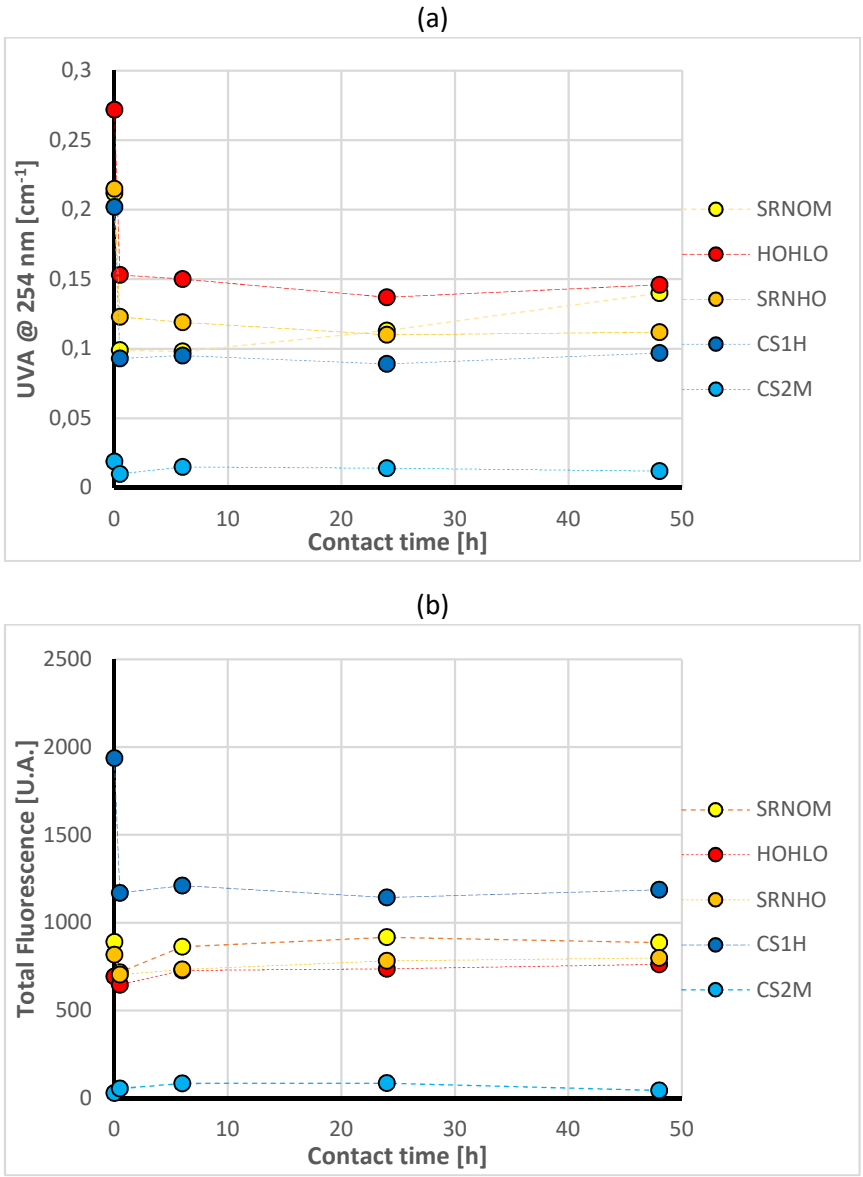


Figure 12: Trends over time for (a) UVA_{254} and (b) Total Fluorescence, for each of the five disinfected waters.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

4.2 Association with DBPs formation

The previously described decrease in the UV absorbance and total fluorescence signals can be attributed to the degradation and transformation of UV-adsorbing compounds during the disinfection process, resulting in part in the formation of DBPs.

As for the formation of regulated DBPs, THMs trend (trichloromethane, bromodichloromethane, dibromochloromethane, tribromomethane) in the different water matrices at different contact times is reported in Figure 13. The data represented refer only to chloroform, since it is the only THM of interest detected in these tests. The generated THMs concentration, regardless the water matrix and the contact time, are always lower than 1 µg/L and result to be not significant compared to the imposed limits in the EU DW Directive 2184/2020, where the limit for the sum of trihalomethanes (THMs) is 100 µg/L. This is also not significant looking at the Italian DW regulation enforced in 2023 (D. Lgs. 18/2023) that sets a 30 µg/L. Since the disinfectant concentration tested in these tests (5 $mg_{Cl2\text{-}eq}$/L) and the contact times (until 48 hours) are greater than the typical disinfectant concentrations and contact times applied in drinking water, this is a confirmation that the use of ClO_2 as disinfectant in a wide range of drinking water matrices does not result in THMs above regulatory levels.

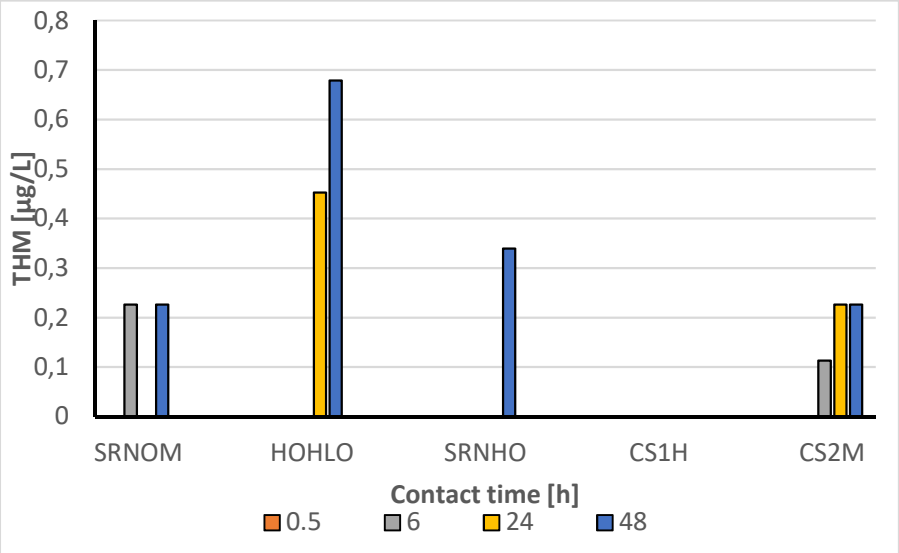


Figure 13: THM (Chloroform) formation for each of the five disinfected water at different contact time.

The second regulated DBP evaluated in these tests is chlorite, whose limit in the EU DW Directive is set to 700 µg/L. In fact, when chlorine dioxide is employed, it can react with both organic and inorganic compounds to form chlorite and chlorate, which can have negative effects on human health (Sorlini et al., 2014). The chlorite time behaviour is reported in Figure 14. There is evidence of a gradual increase in chlorite concentration from 400 up to 950 µg/L for the two real water matrices (CS1H and CS2M), as well as for synthetic matrix representing the river water (SRNOM). This shows that, in case of ClO_2 disinfection, the main regulated DBPs to be assessed are chlorite and chlorate, compared to THMs. However, the trend for SRNHO shows an early formation of chlorite that is no longer found at higher contact times. For the bog water (HOHLO) chlorite is not detected (concentration below the LOQ=75 µg/L) at any time. This result is interesting because it is possible to evaluate which NOM characteristic is present in the SRNOM but absent in the HOHLO sample, allowing to identify the specific precursor of chlorite in DW.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

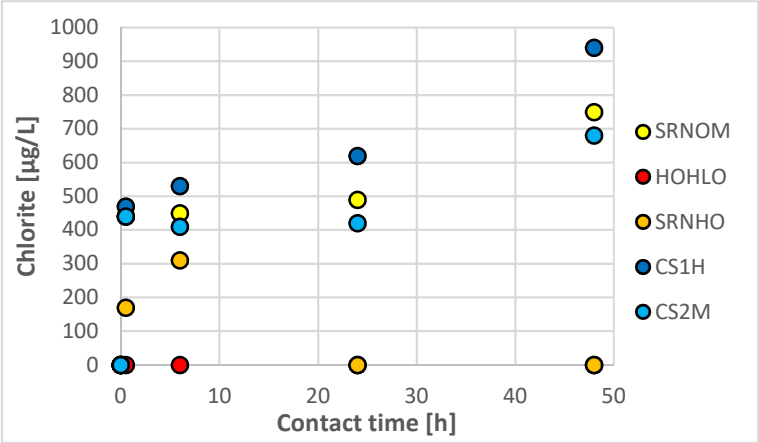


Figure 14: Chlorite time behaviour for each of the five disinfected water at different contact time.

4.3 Correlation between different NOM characterization methods

The last goal of the simultaneous measurement of NOM through different techniques was to evaluate the possibility to correlate the NOM characteristics found with each measuring method. In fact, if the correlation would exist, it was/would be possible to apply only the less time and cost-expensive methods (e.g. absorbance) to have a full characterization of NOM, without the need more complex techniques (e.g. fluorescence or LC-OCD/UVD). The developed method has the aim to evaluate whether it is possible to find such correlation among different measurements. To describe the developed procedure, we show an example based on a historically available dataset where NOM has been characterized by both UV absorbance spectra and fluorescence EEM. This dataset consists of 113 water samples taken from a DWTP with river source water at three different times between June and July 2021. The dataset was assembled by recording absorbance and fluorescence spectra for samples subjected to various treatment combinations, encompassing an ozonation treatment and 48-hour contact with activated carbon. Absorbance spectra and the average fluorescence are presented in Figure 15.

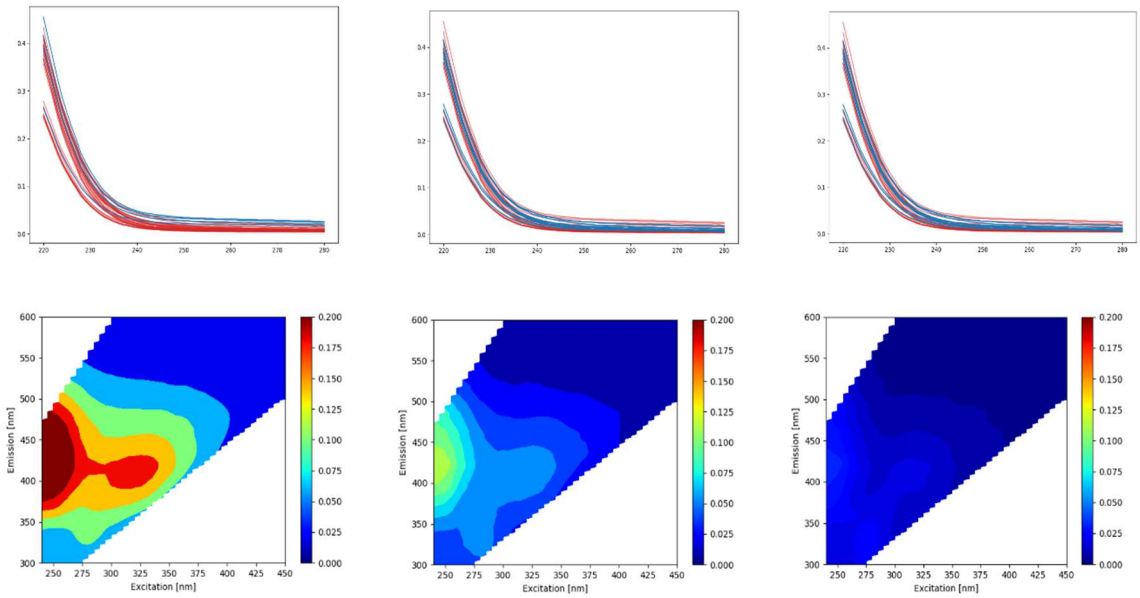


Figure 15: Top absorbance spectra, bottom the average fluorescence spectra, from left to right: before the ozonation, after the ozonation, after the activated carbon treatment.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Figure 16 presents the absorbance spectra (left) and the corresponding centres of the clusters (right) obtained by applying hierarchical clustering (see paragraph 2.6.3). We can already discern a complex structure in the data, revealing at least two distinct clusters for the absorbance spectra. Conversely, we can directly observe the impact of advancing the treatment through the diminishing intensity of the spectra, it is important to note that the observed clustering is not directly linked to the treatment effect. This complexity aligns with what we discussed in Section 2.6, thus requiring the need for more sophisticated statistical tools unveiling non-linear correlations among the two measurement sets.

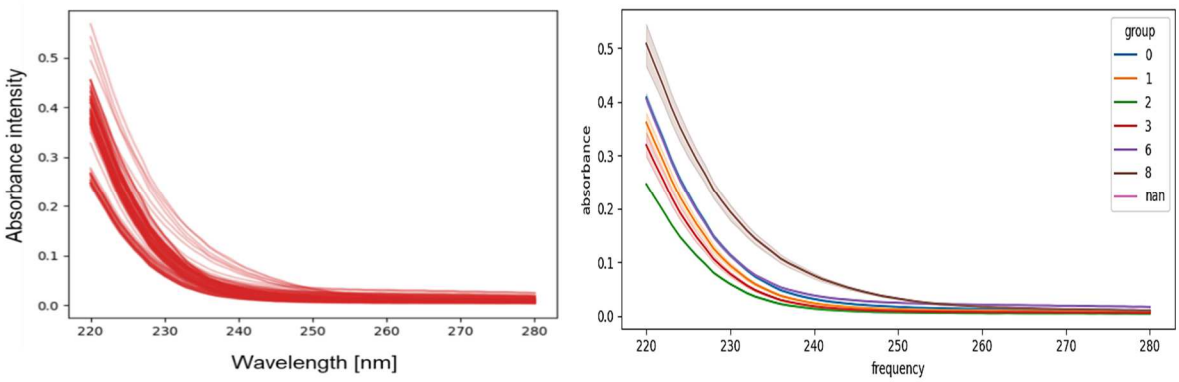


Figure 16: Clusters produced by cutting the absorbance spectra hierarchical tree at 10 clusters (clusters with fewer than 5 samples are grouped together under the 'nan' cluster) Left: original dataset, Right: average absorbance spectrum per cluster and corresponding pointwise 90% confidence interval

The contingency table for the absorbance vs fluorescence clustering is presented in Figure 17 (left) as well as the corresponding transformed values (right). We compared the clusters we obtain through absorbance spectra and those of fluorescence spectra. We have 6 clusters created from the absorbance spectra, one per columns, and along the rows the 8 clusters corresponding to the fluorescence spectra. Cells with darker colours indicate a higher number of shared samples, while blank spaces represent clusters with no shared samples.

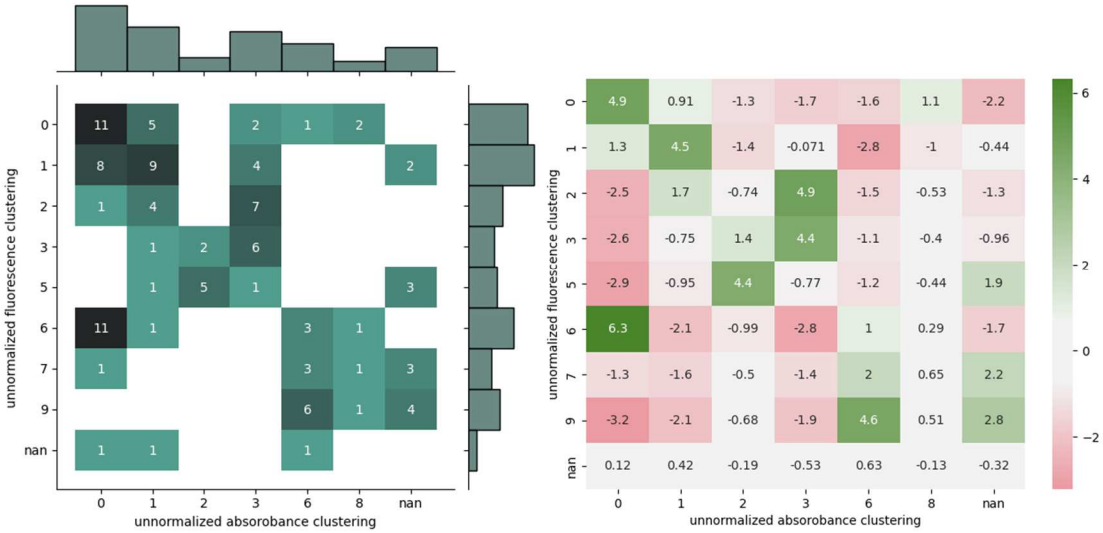


Figure 17: Clustering with respect to absorbance spectra vs fluorescence spectra

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Clusters 0, 1, 3, and 6 are the most numerous for the unnormalized absorbance clustering. Focusing on them we can already extrapolate some information, for instance, the white regions in the matrix in the left part of Figure 17, tell us that being in cluster 3 for the absorbance implies that such a sample cannot belong to clusters 6-9 for the fluorescence. This allows having only a limited variety of fluorescence for this data. Similarly, if the sample belongs to cluster 0 for the absorbance, it will most likely also belong to cluster 0 or 6 for the fluorescence.

The green regions in the right matrix of Figure 17 show the associative pattern between the clusters. For example, being in cluster 6 for the absorbance sharply increases the likelihood of being in cluster 9 for the fluorescence.

Further analysis can be conducted by looking at the average fluorescence spectra of the clusters a sample might belong to. For instance, cluster 3 for the absorbance is associated with clusters 2 or 3 for the fluorescence In Figure 18, we report the normalized fluorescence spectra for these two clusters (cluster 2 on the left and 3 on the right), this allows us to focus on the shape difference between the two spectra. In particular, the two clusters have similar shapes except that Cluster 3 has a peak in the area of the typical peak T (corresponding to protein-like NOM).

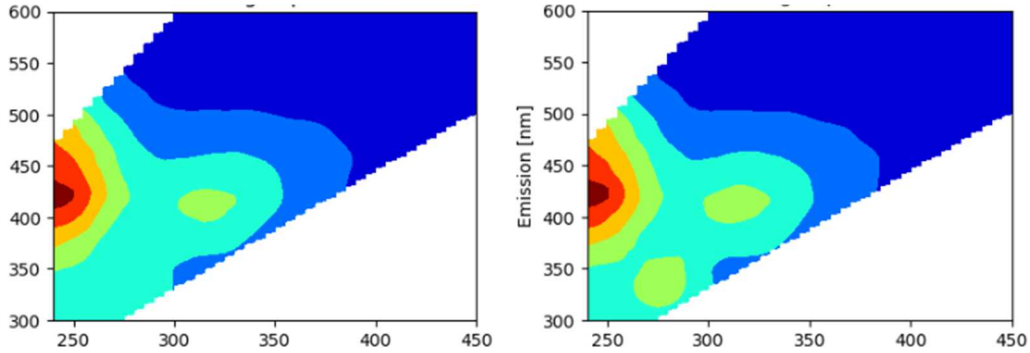


Figure 18: Normalized fluorescence spectra for clusters 2 (left) and 3 (right) of the fluorescence, the shape difference shows that moving from cluster 2 to cluster 3 corresponds to a redistribution of fluorescence intensity from region III, towards regions IV a V of the regional integration approach commonly used in fluorescence spectroscopy analysis

For this dataset also the metadata corresponding to the treatment intensity were available, allowing to compare the clusters produced by the treatment class to those obtained in an unsupervised fashion.

We can see in Figure 19 that the absorbance is a good predictor of the treatment class, for example clusters 0, 1 and 2 are all associated to high level of ozonation but increasing level of AC dosage. While cluster 7 shows low levels of both ozonation and activated carbon treatment.

In summary, we proved the possibility to use absorbance that is a less time and cost-expensive method with respect to fluorescence, for a preliminary NOM characterization also related to the treatment process applied. This finding opens to further insights in linking different methods to characterize NOM to be used for the evaluation of process performance in removing NOM, as precursors of DBPs.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

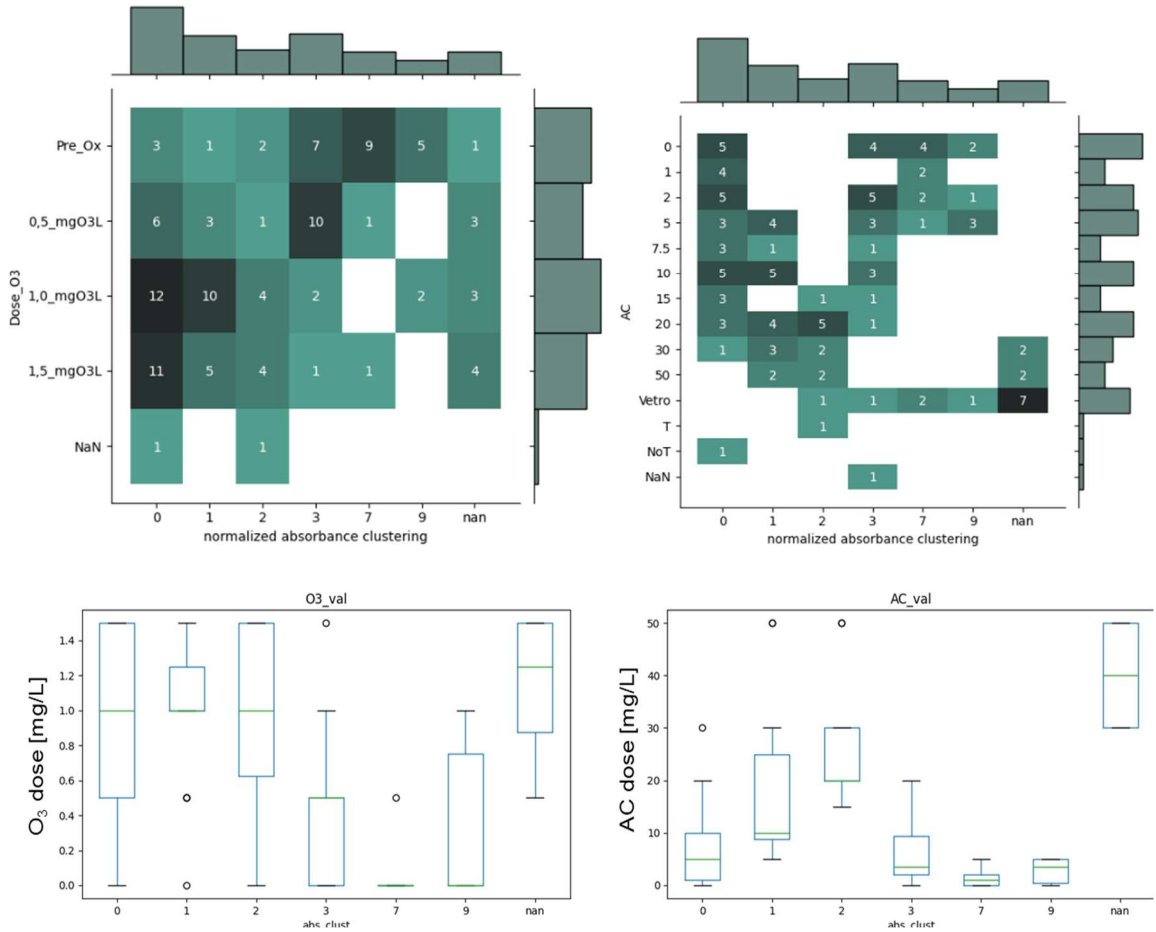


Figure 19: Top: Normalized absorbance clustering against the levels of treatments: ozonation (left) and activated carbon adsorption (right). Bottom: boxplots of the treatment levels grouped by normalized absorbance clustering.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

5 Results and method development for NOM characterization with LC-FT-ICR-MS

5.1 NOM analysis

A workflow of sample analysis and data processing with FT-ICR-MS is established and operational at the UFZ. The generic method of DOM analysis using FT-ICR-MS is based on non-target direct infusion as sample introduction after solid-phase extraction. This method, however, can easily overlook a large fraction of DOM due to their low recovery during sample extraction (especially the polar fractions) and matrix suppression during ionization (Patriarca et al., 2018). To overcome this limitation, we couple FT-ICR-MS to a liquid chromatography (LC) and directly inject the whole sample or after a simple freeze-drying enrichment. LC-FT-ICR-MS has been applied by the UFZ team to study the change in effluent organic matter during wastewater ozonation (Jennings et al., 2022) and to characterize DOM and DBPs in disinfected drinking water (Han et al., 2023). In addition to the information on individual chemical composition and structures (e.g., H/C and O/C ratio, unsaturation degree and aromaticity), LC coupled FT-ICR-MS can also provide insight into the polar fractions of DOM and enable the better separation of structural isomers (Han et al., 2021; Jennings et al., 2022).

The five water matrices in Table 1: synthetic waters with SRNOM, HOHLO, SRNHO, and real drinking waters before disinfection from case study#1 (CS1H_DWTP1) and #2 (CS2M_DWTP1), were disinfected by chlorine or ClO_2 for 48 hours in the lab (section 2.3). Samples before and after disinfection were analysed using LC-FT-ICR-MS after freeze-drying enrichment. Table 3 shows the examples of the number of unique molecular formulas (MFs) per sample. No Cl was included in elemental composition during MF assignment, as this element is not considered part of NOM. This first stage of data analysis did not aim at detecting halogenated DBPs, but rather studied the effect of disinfection on non-halogenated NOM.

Thousands of unique MFs were detected in each sample, with the general trend of chlorination reducing the number of unique MFs. This change likely was due to the formation of new MFs with Cl in post-chlorination samples.

Table 3: Unique Molecular Formulas per Sample [a]

Samples	CHO	CHNO	CHNOS	CHOS	CHN	CHNS	CHS	other
SRNOM_pre	3512	1711	225	643	7	13	10	7
SRNOM_post	3009	1557	178	461	6	13	4	3
CS1H_DWTP1_pre	3639	3133	1116	1929	3	10	3	7
CS1H_DWTP1_post	2975	2069	498	1191	3	6	6	9

[a] "pre" and "post" refer to before and after chlorination, respectively. Cl was not included in elemental composition during molecular formula assignment.

Figure 20 illustrates the relative abundance of each formula class distributed in LC segments. The CS1H_DWTP1 samples, which are the real drinking water samples from case study#1, appeared to have less CHO formula class, but much higher CHOS, CHNOS, and CHNO compared to the synthetic water with SRNOM. These results suggested that real water samples have more diversity in chemical composition, thus can be more interesting to study in the future as compared to synthetic waters prepared with NOM extracts. This higher diversity may be explained by the limited extraction efficacy of any isolation process for NOM.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

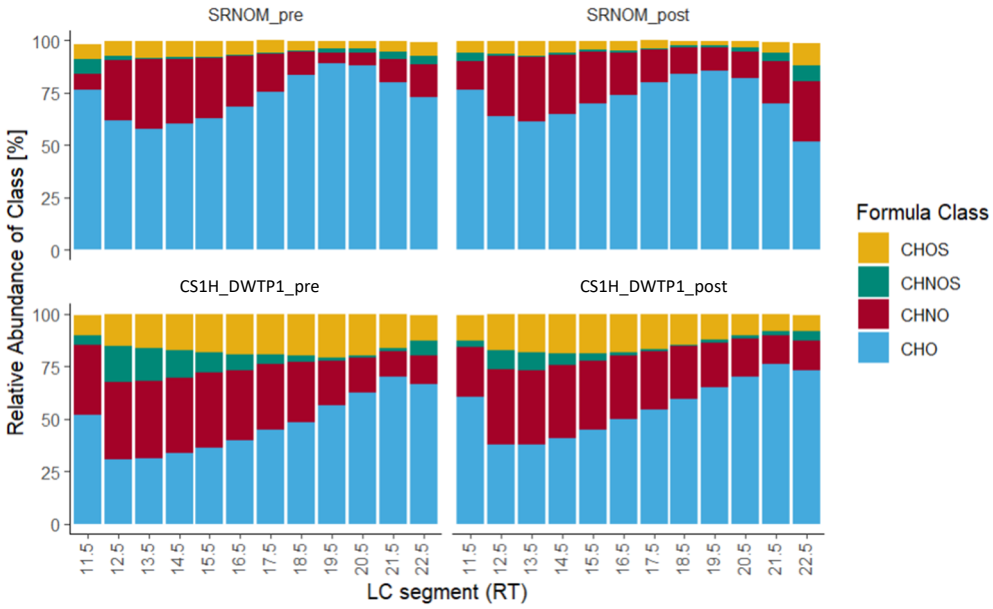


Figure 20: The relative abundance of formula classes in each LC segment. "pre" and "post" mean before and after chlorination, respectively. Cl was not included in elemental composition during molecular formula assignment.

Figure 21 shows the weighted average of MFs detected in real drinking water sample from case study #1 before and after chlorination. It is visible that both aromaticity and *m/z* decrease by chlorination, indicating that chlorine is highly reactive towards those compounds with high aromaticity (e.g., aromatic rings, unsaturated bonds) and breaks down the large molecules into smaller ones as already seen by LC-OCD results.

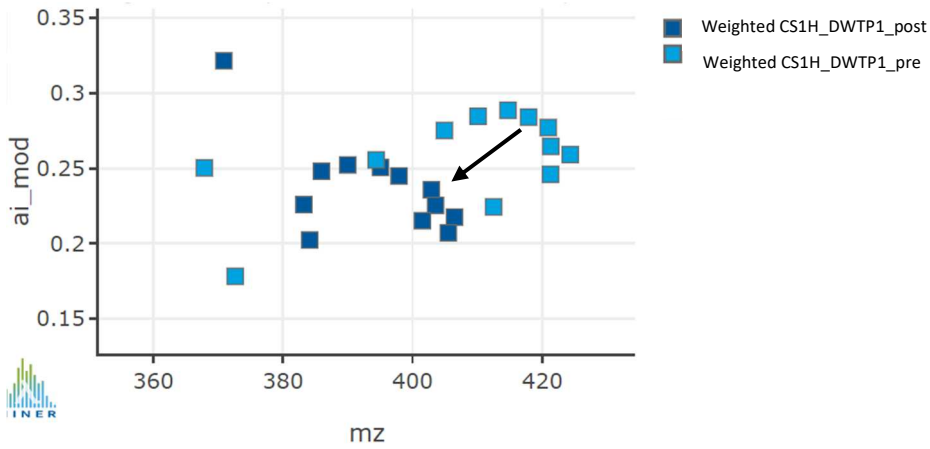


Figure 21: Intensity weighted average aromaticity (ai_mod) versus mass (mz) of all MFs detected in CS1H_DWTP1 sample before and after chlorination. "-pre" and "-post" refer to the sample before and after chlorination, respectively.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

5.2 DBPs formation potential

5.2.1 Non-halogenated N-DBPs

The LC-FT-ICR-MS method applied here can also be useful to characterize DBPs in drinking water. As an example, Figure 22 shows the H/C *vs*. O/C ratio of CHNO formula class in real drinking water samples collected from case study#1 before and after chlorination. The red dots are those non-halogenated nitrogen containing compounds newly generated after chlorination (intensity increase), thus are considered as non-halogenated nitrogenous DBPs (N-DBPs). These N-DBPs appeared to have higher polarity (high O/C ratio) compared to other CHNO compounds. This is possibly related to the strong reactivity of chlorine towards electron-rich polar precursors (e.g., aromatic rings) to produce oxygen rich N-DBPs (Milstead and Remucal, 2021).

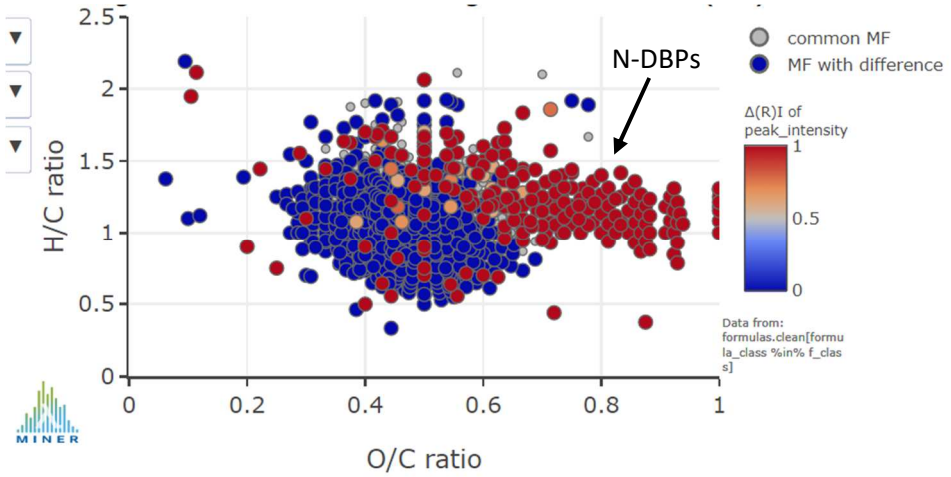


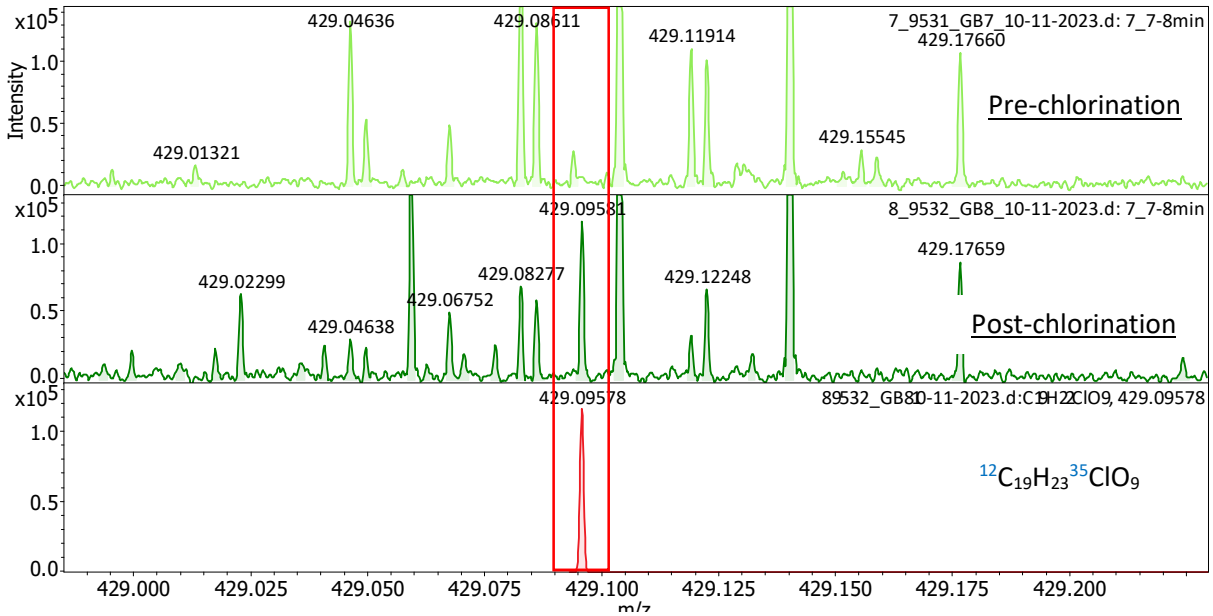
Figure 22: The molecular hydrogen to carbon (H/C) vs. oxygen to carbon (O/C) ratios of nitrogen-containing molecular formulas (CHNO-MFs) in CS1H_DWTP1 samples in the chlorinated sample. A group of CHNO-MFs appear in the chlorinated sample (red), but were undetected before chlorination. These MF are characterized by above-average O/C ratios as compared to the bulk of CHNO-MF present before chlorination (blue).

5.2.2 Halogenated DBPs

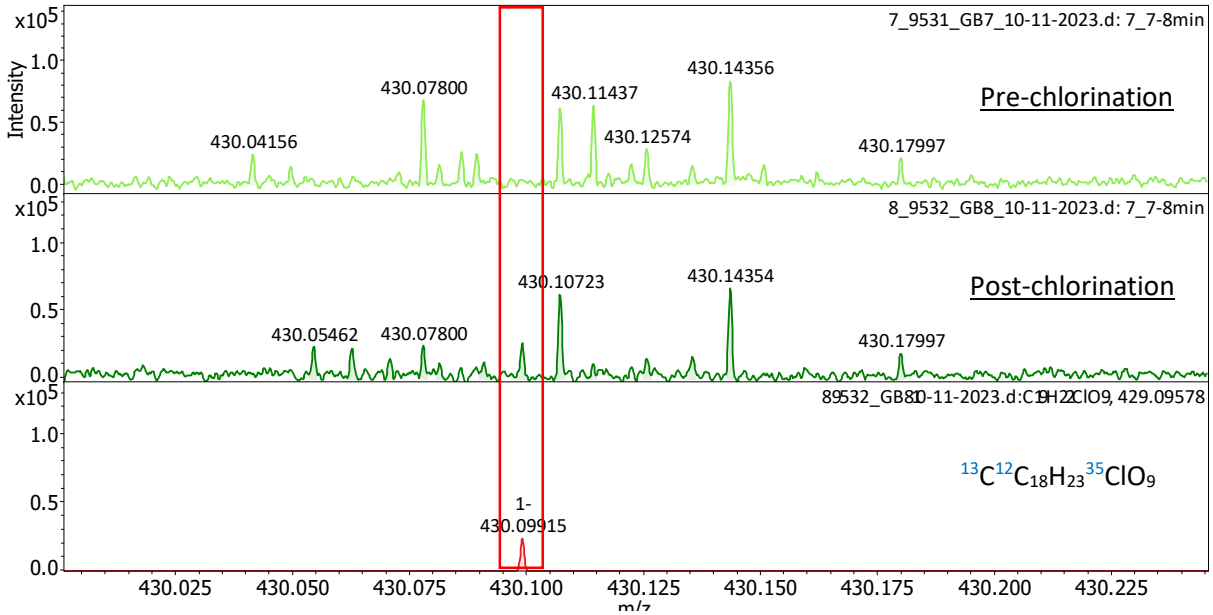
As a second stage of data analysis, chlorine (Cl) was included in elemental composition when assigning MFs in order to identify the chlorinated DBPs in disinfected water. Figure 23 shows identification of a new chlorinated DBP with MF of $C_{19}H_{23}ClO_9$ in real drinking water sample. A new peak was visible in a mass spectrum in 7-8 min LC segment after chlorination, which was absent in the pre-chlorinated sample, and thus considered as DBP. According to the accurate mass, *m/z* 429.09578, its MF was assigned as $C_{19}H_{23}ClO_9$. Figures (a), (b), and (c) illustrate its carbon and chlorine isotope peaks (i.e., $^{12}C^{35}Cl$, $^{13}C^{35}Cl$, and $^{12}C^{37}Cl$, respectively), further supporting the proposed chemical composition.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

a)



b)



This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

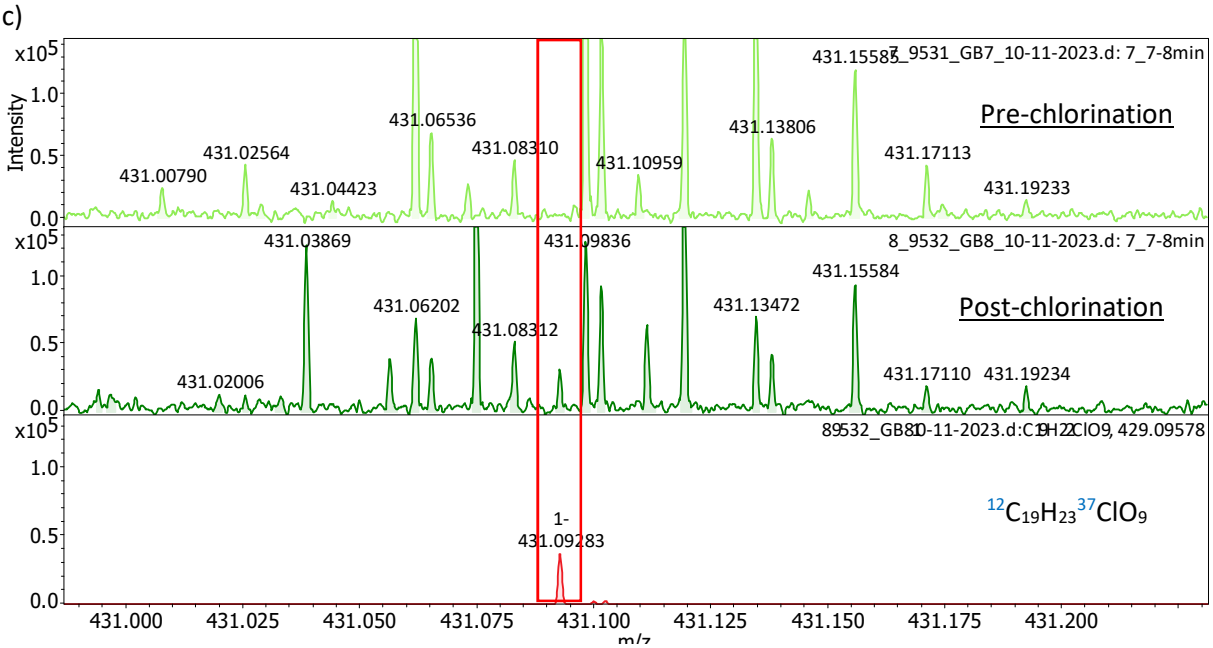


Figure 23: Identification of a DBP ($C_{19}H_{23}ClO_9$) in drinking water sample after chlorination (CS1H_DWTP1 in Table 1). (a), (b), and (c) show the carbon and chlorine isotope peaks (i.e., $^{12}C^{35}Cl$, $^{13}C^{35}Cl$, and $^{12}C^{37}Cl$, respectively).

5.3 DBP precursors

LC-FT-ICR-MS data can also provide information on the precursors of DBPs. As an example, the potential precursor of $C_{19}H_{23}O_9Cl$ shown in Figure 23 can be traced back according to an assumption that this DBP was produced by Cl substitution for H. A potential precursor peak with MF of $C_{19}H_{24}O_9$ was visible in the LC chromatogram (Figure 24). Such method can be extended during SafeCREW project to identify and to structurally characterize the DBP precursors.

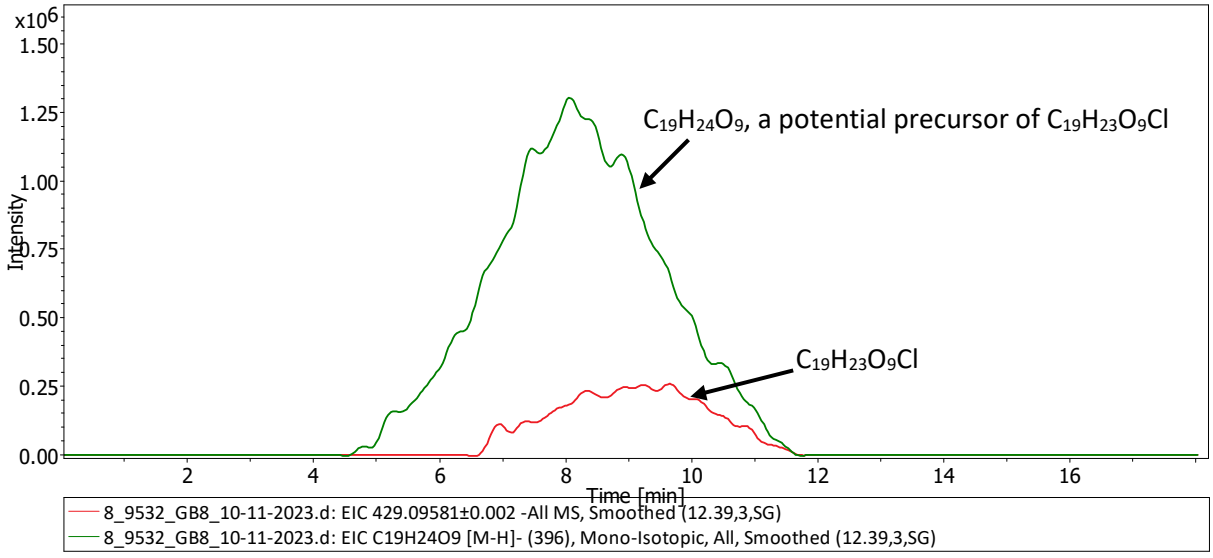


Figure 24: The extracted ion chromatogram of $C_{19}H_{23}ClO_9$ and its potential precursor $C_{19}H_{24}O_9$ found in chlorinated drinking water sample (CS1H_DWTP1 in Table 1).

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

6 Conclusions

The data shown here for the five water matrices before and after disinfection by chlorine or ClO_2 illustrate the complementary nature of the three analytical approaches and of the statistical exploitation established in SafeCREW.

LC-OCD/UVD/OND analysis (section 3) can play a very useful role in unravelling the dynamics of NOM interaction with chlorine during disinfection process. This analytical approach not only allows for a comprehensive understanding of NOM transformation and breakdown but also enables the investigation of reaction kinetics. By correlating LC-OCD/-UVD/-OND results with the formation of DBPs, it becomes possible to associate distinct NOM fractions with the resulting DBPs. Notably, a strong correlation between the humic fraction and THMs has already been demonstrated. The developed method underscores the potential and establishes a foundation for enhancing existing predictive DBP models by incorporating factors such as the humic fraction. The ongoing data collection in this project aims to expand the prediction capabilities based on NOM composition. Moreover, the ability to fingerprint raw waters based on their NOM profiles offers valuable insights into the unique characteristics of different water sources and their evolving composition. This method can play an important role in supporting process optimization, quality control, and regulatory compliance.

In addition, the comparison of absorbance and fluorescence spectra in various water sources (section 4) provides valuable insights. Real waters exhibited distinct absorbance patterns, possibly due to inorganic interferences absent in synthetic waters, with a different NOM content and compositions among the five tested waters. The kinetic tests demonstrated how absorbance and fluorescence evolve over contact time, showing variations in NOM degradation and DBP formation. Notably, the use of ClO_2 as a disinfectant in different drinking water matrices did not lead to THMs above regulatory level. Chlorite formation, a critical concern, was observed to be variable among different water sources. This result can help identify boundary conditions of chlorite formation in drinking water. Therefore, this study highlights the importance of understanding water composition and its implications for disinfection practices, emphasizing the need for tailored strategies to mitigate DBP formation in drinking water systems.

Moreover, the simultaneous measurement of NOM using different techniques aimed to establish correlations between NOM characteristics determined by each method. The goal was to assess the possibility of using less time and cost-intensive methods, such as absorbance, to fully characterize NOM without resorting to more complex techniques like fluorescence or LC-OCD/UVD. Comparing absorbance and fluorescence clustering unveiled associations and patterns, allowing for the prediction of treatment effects. This underscores the potential of using absorbance as a reliable predictor. Overall, these findings emphasize the importance of employing multiple techniques to comprehensively understand NOM and improve water treatment processes.

LC-FT-ICR-MS analysis (section 5) provided the molecular level information of NOM in different water matrices. The real drinking water sample was found to have more diverse chemical composition compared to synthetic waters prepared with NOM extracts. The hydrogen to carbon (H/C) *vs*. oxygen to carbon (O/C) plot as well as the aromaticity (ai_mod) *vs*. mass (mz) plot indicated that chlorine appeared to be highly reactive towards NOM compounds with high aromaticity, and breaks down the large molecules into smaller ones. FT-ICR-MS data also provided information about DBP formation. Such as chlorination led to the formation of some oxygen-rich non-halogenated N-DBPs. A preliminary attempt was made to trace the precursor of a chlorinated DBP based on the assumption that this DBP was generated by Cl substitution for H. A potential precursor was also observed from LC chromatogram

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

with expected mass. Such work can be further extended in the future to study the precursors of the DBPs of interest if needed.

The analytical methods for NOM characterization established in this WP and the concept of associating NOM characteristics with DBP formation potential mentioned here will be applied in the course of the project. Specifically in WP2, these methods will be used to study the treatment specific changes in NOM composition and its link with DBP formation potential during full-scale treatment processes applied in three case studies (Task 2.5). In Task 2.2, 2.3, and 2.4, these methods will be used to explore the efficacy of alternative treatment processes (e.g., membrane electro-sorption, adsorption, advanced oxidation) for NOM removal. Moreover, these analytical methods will be useful in WP3 to monitor DBP precursors in drinking water distribution network.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

7 Bibliography

APHA (2017) Standard Methods for the Examination of Water and Wastewater (23rd ed.), American Public Health Association., Washington DC, USA.

Bond, T., Goslan, E.H., Parsons, S.A. and Jefferson, B. 2012a. A critical review of trihalomethane and haloacetic acid formation from natural organic matter surrogates. Environmental Technology Reviews 1(1), 93-113.

Bond, T., Templeton, M.R. and Graham, N. 2012b. Precursors of nitrogenous disinfection by-products in drinking water—A critical review and analysis. Journal of Hazardous Materials 235-236, 1-16.

Chen, W. and Yu, H.-Q. 2021. Advances in the characterization and monitoring of natural organic matter using spectroscopic approaches. Water Research 190, 116759.

DOC-Labor (2006) LC-OCD – Liquid chromatography Organic Carbon Detection, Information Brochure 1/2006.

Dodge, Y. (2008) Contingency Table. In The Concise Encyclopedia of Statistics, pp. 110-111, Springer New York, New York, NY.

Everitt, B.S., Landau, S., Leese, M. and Stahl, D. (2011) Cluster Analysis, John Wiley & Sons, Ltd.

Fernández-Pascual, E., Droz, B., O'Dwyer, J., O'Driscoll, C., Goslan, E.H., Harrison, S. and Weatherill, J. 2023. Fluorescent Dissolved Organic Matter Components as Surrogates for Disinfection Byproduct Formation in Drinking Water: A Critical Review. ACS ES&T Water 3(8), 1997-2008.

Frimmel, F.H., Abbt-Braun, G., G. Heumann, K., Hock, B., Lüdemann, H.-D. and Spiteller, M. (2002) Refractory Organic Substances in the Environment, Wiley-VCH, Weinheim, Germany.

Gao, C.X., Dwyer, D., Zhu, Y., Smith, C.L., Du, L., Filia, K.M., Bayer, J., Menssink, J.M., Wang, T., Bergmeir, C., Wood, S. and Cotton, S.M. 2023. An overview of clustering methods with guidelines for application in mental health research. Psychiatry Research 327, 115265.

Green, P.J. and Silverman, B.W. (1993) Nonparametric Regression and Generalized Linear Models: A roughness penalty approach (1st ed.), Chapman and Hall/CRC, New York.

Han, L., Kaesler, J., Peng, C., Reemtsma, T. and Lechtenfeld, O.J. 2021. Online Counter Gradient LC-FT-ICR-MS Enables Detection of Highly Polar Natural Organic Matter Fractions. Analytical Chemistry 93(3), 1740-1748.

Han, L., Lohse, M., Nihemaiti, M., Reemtsma, T. and Lechtenfeld, O.J. 2023. Direct non-target analysis of dissolved organic matter and disinfection by-products in drinking water with nano-LC-FT-ICR-MS. Environmental Science: Water Research & Technology 9(6), 1729-1737.

Hua, G. and Reckhow, D.A. 2007. Characterization of Disinfection Byproduct Precursors Based on Hydrophobicity and Molecular Size. Environmental Science & Technology 41(9), 3309-3315.

Huber, S.A., Balz, A., Abert, M. and Pronk, W. 2011. Characterisation of aquatic humic and non-humic matter with size-exclusion chromatography--organic carbon detection--organic nitrogen detection (LC-OCD-OND). Water Res 45(2), 879-885.

Jennings, E., Kremser, A., Han, L., Reemtsma, T. and Lechtenfeld, O.J. 2022. Discovery of Polar Ozonation Byproducts via Direct Injection of Effluent Organic Matter with Online LC-FT-ICR-MS. Environmental Science & Technology 56(3), 1894-1904.

Kim, S., Kramer, R.W. and Hatcher, P.G. 2003. Graphical Method for Analysis of Ultrahigh-Resolution Broadband Mass Spectra of Natural Organic Matter, the Van Krevelen Diagram. Analytical Chemistry 75(20), 5336-5344.

Lavonen, E.E., Kothawala, D.N., Tranvik, L.J., Gonsior, M., Schmitt-Kopplin, P. and Köhler, S.J. 2015. Tracking changes in the optical properties and molecular composition of dissolved organic matter during drinking water production. Water Research 85, 286-294.

Li, L., Wang, Y., Zhang, W., Yu, S., Wang, X. and Gao, N. 2020. New advances in fluorescence excitation-emission matrix spectroscopy for the characterization of dissolved organic matter in drinking water treatment: A review. Chemical Engineering Journal 381, 122676.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Matilainen, A., Gjessing, E.T., Lahtinen, T., Hed, L., Bhatnagar, A. and Sillanpää, M. 2011. An overview of the methods used in the characterisation of natural organic matter (NOM) in relation to drinking water treatment. Chemosphere 83(11), 1431-1442.

Milstead, R.P. and Remucal, C.K. 2021. Molecular-Level Insights into the Formation of Traditional and Novel Halogenated Disinfection Byproducts. ACS ES&T Water 1(8), 1966-1974.

Muellner, M.G., Wagner, E.D., McCalla, K., Richardson, S.D., Woo, Y.-T. and Plewa, M.J. 2007. Haloacetonitriles vs. Regulated Haloacetic Acids: Are Nitrogen-Containing DBPs More Toxic? Environmental Science & Technology 41(2), 645-651.

Nihemaiti, M., Icker, M., Seiwert, B. and Reemtsma, T. 2023. Revisiting Disinfection Byproducts with Supercritical Fluid Chromatography-High Resolution-Mass Spectrometry: Identification of Novel Halogenated Sulfonic Acids in Disinfected Drinking Water. Environmental Science & Technology 57(9), 3527-3537.

Nikolaou, A.D., Golfinopoulos, S.K., Lekkas, T.D. and Kostopoulou, M.N. 2004. DBP Levels in Chlorinated Drinking Water: Effect of Humic Substances. Environmental Monitoring and Assessment 93(1), 301-319.

Nychka, D., Furrer, R., Paige, J. and Sain, S. 2021. fields: tools for spatial data." R package version 14.1.

Patriarca, C., Bergquist, J., Sjöberg, P.J.R., Tranvik, L. and Hawkes, J.A. 2018. Online HPLC-ESI-HRMS Method for the Analysis and Comparison of Different Dissolved Organic Matter Samples. Environmental Science & Technology 52(4), 2091-2099.

R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reemtsma, T. 2009. Determination of molecular formulas of natural organic matter molecules by (ultra-) high-resolution mass spectrometry: status and needs. J Chromatogr A 1216(18), 3687-3701.

Shah, A.D. and Mitch, W.A. 2012. Halonitroalkanes, Halonitriles, Haloamides, and N-Nitrosamines: A Critical Review of Nitrogenous Disinfection Byproduct Formation Pathways. Environmental Science & Technology 46(1), 119-131.

Singer, P.C. 1999. Humic Substances as Precursors for Potentially Harmful Disinfection By-Products. Water Science and Technology 40(9), 25-30.

Sorlini, S., Gialdini, F., Biasibetti, M. and Collivignarelli, C. 2014. Influence of drinking water treatments on chlorine dioxide consumption and chlorite/chlorate formation. Water Research 54, 44-52.

Świetlik, J. and Sikorska, E. 2004. Application of fluorescence spectroscopy in the studies of natural organic matter fractions reactivity with chlorine dioxide and ozone. Water Research 38(17), 3791-3799.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A.P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C.N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D.A., Hagen, D.R., Pasechnik, D.V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G.A., Ingold, G.-L., Allen, G.E., Lee, G.R., Audren, H., Probst, I., Dietrich, J.P., Silterra, J., Webber, J.T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J.L., de Miranda Cardoso, J.V., Reimer, J., Harrington, J., Rodríguez, J.L.C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N.J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P.A., Lee, P., McGibbon, R.T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T.J., Robitaille, T.P., Spura, T., Jones, T.R., Cera, T., Leslie, T.,

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y.O., Vázquez-Baeza, Y. and SciPy, C. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 17(3), 261-272.

Wassink, J.K., Andrews, R.C., Peiris, R.H. and Legge, R.L. 2011. Evaluation of fluorescence excitation–emission and LC-OCD as methods of detecting removal of NOM and DBP precursors by enhanced coagulation. Water Supply 11(5), 621-630.

Wünsch, U.J., Bro, R., Stedmon, C.A., Wenig, P. and Murphy, K.R. 2019. Emerging patterns in the global distribution of dissolved organic matter fluorescence. Analytical Methods 11(7), 888-893.

Yang, X., Guo, W., Zhang, X., Chen, F., Ye, T. and Liu, W. 2013. Formation of disinfection by-products after pre-oxidation with chlorine dioxide or ferrate. Water Research 47(15), 5856-5864.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.