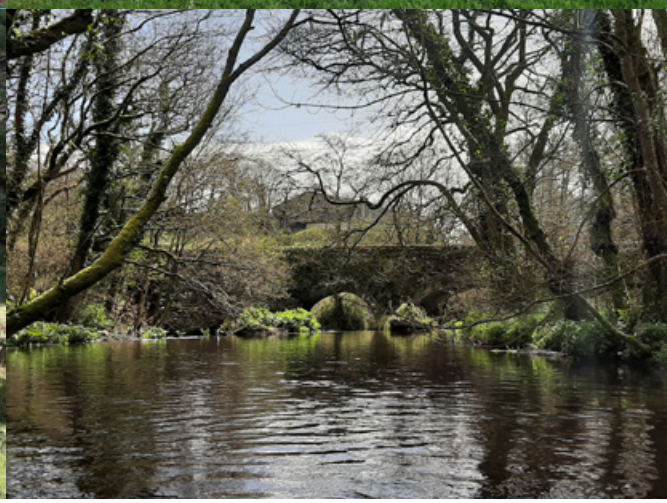
Report No.486

epa Research

Proactive Optical Monitoring of Catchment Dissolved Organic Matter for Drinking Water Source Protection (PRODOM)

Authors: John Weatherill, Boris Droz, Elena Fernández-Pascual, Jean O'Dwyer, Emma Goslan, Connie O'Driscoll and Simon Harrison



www.epa.ie

Rialtas na hÉireann
Government of Ireland

Environmental Protection Agency

The EPA is responsible for protecting and improving the environment as a valuable asset for the people of Ireland. We are committed to protecting people and the environment from the harmful effects of radiation and pollution.

The work of the EPA can be divided into three main areas:

Regulation: Implementing regulation and environmental compliance systems to deliver good environmental outcomes and target those who don't comply.

Knowledge: Providing high quality, targeted and timely environmental data, information and assessment to inform decision making.

Advocacy: Working with others to advocate for a clean, productive and well protected environment and for sustainable environmental practices.

Our Responsibilities Include:

Licensing

- Large-scale industrial, waste and petrol storage activities;
- Urban waste water discharges;
- The contained use and controlled release of Genetically Modified Organisms;
- Sources of ionising radiation;
- Greenhouse gas emissions from industry and aviation through the EU Emissions Trading Scheme.

National Environmental Enforcement

- Audit and inspection of EPA licensed facilities;
- Drive the implementation of best practice in regulated activities and facilities;
- Oversee local authority responsibilities for environmental protection;
- Regulate the quality of public drinking water and enforce urban waste water discharge authorisations;
- Assess and report on public and private drinking water quality;
- Coordinate a network of public service organisations to support action against environmental crime;
- Prosecute those who flout environmental law and damage the environment.

Waste Management and Chemicals in the Environment

- Implement and enforce waste regulations including national enforcement issues;
- Prepare and publish national waste statistics and the National Hazardous Waste Management Plan;
- Develop and implement the National Waste Prevention Programme;
- Implement and report on legislation on the control of chemicals in the environment.

Water Management

- Engage with national and regional governance and operational structures to implement the Water Framework Directive;
- Monitor, assess and report on the quality of rivers, lakes, transitional and coastal waters, bathing waters and groundwaters, and measurement of water levels and river flows.

Climate Science & Climate Change

- Publish Ireland's greenhouse gas emission inventories and projections;
- Provide the Secretariat to the Climate Change Advisory Council and support to the National Dialogue on Climate Action;
- Support National, EU and UN Climate Science and Policy development activities.

Environmental Monitoring & Assessment

- Design and implement national environmental monitoring systems: technology, data management, analysis and forecasting;
- Produce the State of Ireland's Environment and Indicator Reports;
- Monitor air quality and implement the EU Clean Air for Europe Directive, the Convention on Long Range Transboundary Air Pollution, and the National Emissions Ceiling Directive;
- Oversee the implementation of the Environmental Noise Directive;
- Assess the impact of proposed plans and programmes on the Irish environment.

Environmental Research and Development

- Coordinate and fund national environmental research activity to identify pressures, inform policy and provide solutions;
- Collaborate with national and EU environmental research activity.

Radiological Protection

- Monitoring radiation levels and assess public exposure to ionising radiation and electromagnetic fields;
- Assist in developing national plans for emergencies arising from nuclear accidents;
- Monitor developments abroad relating to nuclear installations and radiological safety;
- Provide, or oversee the provision of, specialist radiation protection services.

Guidance, Awareness Raising, and Accessible Information

- Provide independent evidence-based reporting, advice and guidance to Government, industry and the public on environmental and radiological protection topics;
- Promote the link between health and wellbeing, the economy and a clean environment;
- Promote environmental awareness including supporting behaviours for resource efficiency and climate transition;
- Promote radon testing in homes and workplaces and encourage remediation where necessary.

Partnership and Networking

- Work with international and national agencies, regional and local authorities, non-governmental organisations, representative bodies and government departments to deliver environmental and radiological protection, research coordination and science-based decision making.

Management and Structure of the EPA

The EPA is managed by a full time Board, consisting of a Director General and five Directors. The work is carried out across five Offices:

1. Office of Environmental Sustainability
2. Office of Environmental Enforcement
3. Office of Evidence and Assessment
4. Office of Radiation Protection and Environmental Monitoring
5. Office of Communications and Corporate Services

The EPA is assisted by advisory committees who meet regularly to discuss issues of concern and provide advice to the Board.

Proactive Optical Monitoring of Catchment Dissolved Organic Matter for Drinking Water Source Protection (PRODOM)

Authors: John Weatherill, Boris Droz, Elena Fernández-Pascual, Jean O'Dwyer, Emma Goslan, Connie O'Driscoll and Simon Harrison

Lead organisation: University College Cork

Identifying pressures

In Ireland, approximately 82% of public water supplies originate from surface water sources which require disinfection with chlorine to inactivate pathogens and prevent the spread of waterborne disease. The presence of dissolved organic matter (DOM) in source waters can lead to the unintentional formation of potentially harmful disinfection by-products (DBPs) such as trihalomethanes (THMs), of which Ireland has the highest number of reported exceedances for drinking water supplies in the European Union in recent years. The Proactive Optical Monitoring of Catchment Dissolved Organic Matter for Drinking Water Source Protection (PRODOM) project aimed to better understand the role of optically active DOM in the formation and forecasting of a range of DBPs, including the common regulated classes like THMs and haloacetic acids (HAAs) as well as emerging DBPs such as haloketones, haloacetonitriles and halonitromethanes (represented by chloropicrin).
DBP formation was determined from chlorination of surface water and groundwater samples under uniform laboratory conditions with DOM spectroscopic and hydrochemical characterisation of raw water after primary filtration.

Informing policy

The research demonstrated a proof of concept for the application of machine learning tools in the prediction of DBP concentrations in treated water (after primary filtration and chlorination) using DOM spectroscopic variables and common hydrochemical parameters. Both regulated and emerging DBPs could be quantitatively predicted with high confidence. Models which included hydrochemical parameters only marginally improved prediction accuracy, demonstrating a way forward for automated low-cost high sample throughput applications for source water DOM management at water treatment plants. The application of portable in situ fluorimetry (measuring two DBP precursor wavelength pairs simultaneously) was also demonstrated, which could reveal high-frequency DOM export dynamics which govern raw water quality temporal variability in river water sources. The project also ranked landcover sources of DBPs for drinking water catchments in Ireland with tree cover (conifer) on upland peat soils found to have the highest risk for THM and HAA formation. Groundwater (as an alternative water source) was found to be a potential source of emerging nitrogenous DBPs.

Developing solutions

This research highlighted the importance of UV-visible fluorescence and absorbance spectroscopy as low-cost, non-destructive high sample throughput technologies suitable for proactive management of source water DOM quality. This is likely to be useful for drinking water source protection and early warning tool applications at water treatment plants for real-time process control and optimisation. Through the application of machine learning tools, greater automation and optimisation of DBP concentration prediction and forecasting (using DOM spectroscopic variables only) are demonstrated, with a workflow that is suitable for up-scaling to online monitoring for major water treatment plant operations in Ireland such as the proposed Water Supply Project Eastern and Midlands Region. Currently, only UV transmittance is routinely measured from raw water sources in Irish water treatment plants. It is recommended that the acquisition of raw water fluorescence excitation–emission matrices is adopted for public water supplies subject to THM exceedances for improved DBP formation risk management and precursor characterisation across different drinking water catchments in Ireland.

EPA RESEARCH PROGRAMME 2014–2020

Proactive Optical Monitoring of Catchment Dissolved Organic Matter for Drinking Water Source Protection (PRODOM)

(2019-W-MS-43)

EPA Research Report

Prepared for the Environmental Protection Agency

by

University College Cork

Authors:

John Weatherill, Boris Droz, Elena Fernández-Pascual, Jean O'Dwyer, Emma Goslan, Connie O'Driscoll and Simon Harrison

ENVIRONMENTAL PROTECTION AGENCY
An Ghníomhaireacht um Chaomhnú Comhshaoil
PO Box 3000, Johnstown Castle, Co. Wexford, Ireland

Telephone: +353 53 916 0600 Fax: +353 53 916 0699
Email: info@epa.ie Website: www.epa.ie

ACKNOWLEDGEMENTS

This report is published as part of the EPA Research Programme 2014–2020. The EPA Research Programme is a Government of Ireland initiative funded by the Department of the Environment, Climate and Communications. It is administered by the Environmental Protection Agency, which has the statutory function of co-ordinating and promoting environmental research. The Department of Further and Higher Education, Research, Innovation and Science, through the Higher Education Authority, provided additional funding to cover researcher salary costed extensions needed because of disruption to research activities caused by COVID-19.

The authors would like to acknowledge the members of the project steering committee, namely Iain Lake (University of East Anglia), Michelle Roche (EPA), Cáit Gleeson (Uisce Éireann), Aine Butler (Uisce Éireann), Eadaoin Joyce (Uisce Éireann) and Giles Varrault (Université Paris-Est Créteil). The authors would also like to kindly acknowledge the landowners in the study areas who provided access to their land for sampling purposes. Richard Cahill and Denis Hurley are gratefully acknowledged for allowing monitoring instrumentation to be installed on their land. Xie Quishi, Liz Gilchrist, Anna Maria Hogan (University College Cork) and Tammy Crowley (Cranfield University) are acknowledged for their technical analytical support throughout the project. Erica Constant (University College Cork) is acknowledged for the management of the project's financial aspects. The Cork field staff at CDM Smith are acknowledged for their assistance in groundwater sampling. We also thank Kieran Khamis (University of Birmingham) for advice on sensor deployment. We acknowledge Gerrard Jones (Oregon State University) for the fruitful discussions on machine learning data analysis. We also sincerely thank data providers, including the EPA, Met Éireann, Teagasc, Geological Survey Ireland, Ordnance Survey Ireland (now Tailte Éireann) and Copernicus.

DISCLAIMER

This report is based on research carried out/data from 4 November 2019 to 4 December 2023. More recent data may have become available since the research was completed.

The EPA Research Programme addresses the need for research in Ireland to inform policymakers and other stakeholders on a range of questions in relation to environmental protection. These reports are intended as contributions to the necessary debate on the protection of the environment.

EPA RESEARCH PROGRAMME 2014–2020
Published by the Environmental Protection Agency, Ireland

ISBN: 978-1-80009-287-7 May 2025

Price: Free Online version

Project Partners

John Weatherill
School of Biological, Earth and Environmental Sciences
University College Cork
Cork
Ireland
Tel.: +353 (0)21 490 4578
Email: john.weatherill@ucc.ie

Boris Droz
School of Biological, Earth and Environmental Sciences
University College Cork
Cork
Ireland
Email: bodroz@bluewin.ch

Elena Fernández-Pascual
School of Biological, Earth and Environmental Sciences
University College Cork
Cork
Ireland
Email: elenafpascual@gmail.com

Jean O'Dwyer
School of Biological, Earth and Environmental Sciences
University College Cork
Cork
Ireland
Email: jean.odwyer@ucc.ie

Emma Goslan
Cranfield Water Science Institute
Cranfield University
Bedfordshire
United Kingdom
Email: E.H.Goslan@cranfield.ac.uk

Connie O'Driscoll
Ryan Hanley Ltd
Castlebar
County Mayo
Ireland
Email: connieodriscoll@gmail.com

Simon Harrison
School of Biological, Earth and Environmental Sciences
University College Cork
Cork
Ireland
Email: s.harrison@ucc.ie

Contents

List of Figures

List of Tables

Executive Summary

In Ireland, approximately 82% of public water supplies originate from surface water catchments which require disinfection with chlorine. The presence of dissolved organic matter (DOM) in source waters can lead to the formation of potentially carcinogenic disinfection by-products (DBPs), including the sum of four trihalomethanes (THM4), of which Ireland has the highest number of reported exceedances in the EU in recent years. This report summarises the findings of a detailed field and laboratory investigation in the River Lee catchment (Cork, Ireland) supported by predictive modelling using machine learning techniques. Twenty individual DBP compounds, namely four trihalomethanes (THMs), nine haloacetic acids (HAAs), two haloketones (HKs), four haloacetonitriles (HANs) and one halonitromethane (HNM), were considered. The project aimed to better understand the role of optically active DOM in DBP formation and forecasting, measured at low cost and with high sample throughput using fluorescence excitation–emission matrix (EEM) spectroscopy and parallel factor analysis (PARAFAC) together with supporting hydrochemistry. A global literature review was also undertaken in which 218 individual statistically significant linear relationships (e.g. $p \leq 0.05$; coefficient of determination (R^2) ≥ 0.5) were identified from 45 original research articles between ubiquitous environmental PARAFAC components and 41 individual DBP parameters. In 53% of studies, humic-like and fulvic-like components were found to be important for THM and HAA formation, highlighting the utility of these fluorophores as surrogates for regulated DBP formation.

Over 200 field surface water and groundwater samples were collected from February to November 2021 and subjected to chlorination and precursor analysis under standardised conditions. An ensemble machine learning model was developed from an optimal number of predictive variables for eight DBP species with high confidence (average $R^2 = 0.83$, root mean squared percentage error = 29.0%), including trichloromethane (chloroform), bromodichloromethane, dibromochloromethane, dichloroacetic acid, trichloroacetic acid, dichloroacetonitrile, trichloronitromethane and trichloropropanone. A further five compounds could be predicted for presence/absence only, namely tribromomethane (bromoform), trichloroacetonitrile, bromochloroacetonitrile, dibromoacetonitrile and dichloropropanone, with an average accuracy of 96.8%. A DOM optical property-only model (spectral model) was developed, where UV absorbance at 254 nm (A_{254}) could serve as a proxy for dissolved organic carbon, that had comparable performance to the ensemble model for at least five regulated and unregulated DBP classes (THMs, HAAs, HANs, HNMs and HKs).

The DBP and precursor data were evaluated spatially using a nested sub-basin approach to assign samples for comparison according to similar sample source categories, including (i) grassland on mineral soil (58%), (ii) grassland on peatland soil (16%) and (iii) tree cover (14.7%), with groundwater samples considered as a separate category. It was noted that the tree cover category may also include blanket peat soils in upland areas under conifer plantation. Groundwater was found to have between two- and six-times lower formation potential for THM4 and the sum of five haloacetic acids (HAA5) than any surface water landcover category. Conversely, the tree cover category had the highest formation potential for THM4 (mean ± standard deviation) (535 ± 479 µg L^{-1}) and HAA5 (23.6 ± 22.44 µg L^{-1}), significantly exceeding that of any other landcover class, and can therefore be considered the landcover class of highest risk for regulated DBP formation. These insights suggest that the replacement of surface water sources with groundwater is a cost-effective solution for water supplies subject to recurrent THM4 (and potentially HAA5) exceedances. However, groundwater underlying intensive agricultural areas may be a source of unregulated and potentially more harmful nitrogen-containing HAN DBPs. A subset of surface water samples analysed using liquid chromatography–organic carbon detection–organic nitrogen detection confirmed that tree cover category samples contained the largest fraction of hydrophobic organic matter which is likely to play an important role in THM and HAA formation.

The research also explored the application of dual wavelength *in situ* fluorescence sensor technology for high-frequency fluorescent DOM measurement and DBP forecasting using humic-like and tryptophan-like fluorescence with LED excitation at 280 nm. The portable fluorimeter was found to have comparable performance to benchtop EEM measurements over the same wavelength regions. A 10-month *in situ* sensor deployment in surface water in the Dripsey study area was able to identify extreme humic-like DOM export events and other high-frequency dynamics after considerable post-processing of raw sensor data. Using machine learning, the dual wavelengths showed reasonable accuracy for predicting HAA5 ($R^2 = 0.77–0.80$) and THM4 ($R^2 = 0.75$) concentrations but with relatively low precision (average error 48.6%), highlighting the importance of full-spectrum excitation–emission matrix-parallel factor analysis for accurate DBP prediction.

1 Introduction

1.1 Background

In Ireland, approximately 82% of public water supplies originate from surface water catchments which require disinfection to inactivate pathogenic microorganisms and prevent the spread of waterborne diseases (O'Driscoll *et al*., 2018a). Since the 1970s, it has been known that the use of chlorine for disinfection of drinking water can lead to the unintentional formation of potentially harmful halogenated disinfection by-products (DBPs) represented by trihalomethanes (THMs). In recent years, an unacceptably large number of public and private drinking water supplies originating from surface water bodies in Ireland have routinely exceeded the drinking water parametric limit of 100 $\mu g\,L^{-1}$ for the sum of four trihalomethanes (THM4) (O'Driscoll *et al*., 2018a; Rolston and Linnane, 2020). In 2022, a total of 45 public schemes in Ireland failed to achieve THM compliance on at least one occasion, with 23 schemes currently on the EPA remedial action list for THM exceedances, affecting 235,000 consumers (EPA, 2023a). In the same year, 16 private group water schemes failed to achieve THM compliance, affecting approximately 14,000 people (EPA, 2023b). The THM4 parametric limit is the sum of trichloromethane (chloroform) (TCM) ($CHCl_3$), dichlorobromomethane ($CHCl_2Br$), dibromochloromethane (DBCM) ($CHBr_2Cl$) and bromoform ($CHBr_3$). Haloacetic acids (HAAs) are an equally prominent class of DPBs whose regulation in Irish drinking water is imminent, pending the transposition of the recast EU Drinking Water Directive (Directive (EU) 2020/2184) (EU, 2020). The parametric limit of 60 $\mu g\,L^{-1}$ for HAA5 is the sum of five HAAs, namely monochloroacetic acid ($ClCH_2COOH$), dichloroacetic acid (DCAA) ($Cl_2CHCOOH$), trichloroacetic acid (TCAA) (Cl_3COOH), monobromoacetic acid ($BrCH_2COOH$) and dibromoacetic acid ($Br_2CHCOOH$) (EU, 2020).

The formation of DBPs is initiated when hypochlorous acid (HOCl) reacts with dissolved organic matter (DOM) containing dissolved organic carbon (DOC) precursors present at the point of chlorination. DOM is the soluble fraction of natural organic matter (NOM), which is ubiquitous in surface waters globally. DBP formation potential and speciation in public and private drinking water supplies are dependent on (i) source water quality, as indicated by DOM concentration and quality, pH, alkalinity and the presence of the inorganic chloride (Cl^-) and bromide (Br^-) ions (Yang *et al*., 2014), and (ii) treatment plant and distribution network conditions, including disinfectant type, disinfectant dosage, DOM removal extent, temperature, pH, residual chlorine levels and network residence times (Krzeminski *et al*., 2019). Carbonaceous disinfection by-products (CDBPs) such as THMs and HAAs have been shown to have cytotoxic, genotoxic and carcinogenic effects in toxicological studies (Stalter *et al*., 2016). A recent large-scale epidemiological study in EU Member States linked exposure to THMs in drinking water to an increased bladder cancer burden (Evlampidou *et al*., 2020). Regulated CDBPs are believed to represent only a minor fraction of the total organic halogen (TOX) load present in chlorinated drinking water. There are at least 700 known DBPs, with many more non-target compounds yet to be discovered (Richardson and Kimura, 2019). New classes of potentially more harmful nitrogenous disinfection by-products (NDBPs) have been identified in recent years, such as haloacetonitriles (HANs) and halonitromethanes (HNMs), which have received comparatively little attention (Bond *et al*., 2015). NDBPs are formed from DOM in the presence of a nitrogen source or where chloramination is used as a substitute for chlorine disinfection to reduce THH and HAA formation (Bond *et al*., 2011). NDBPs are thought to be considerably more cytotoxic and genotoxic than CDBPs (Plewa *et al*., 2002) and can enhance the overall toxicity of the CDBP species present (Li and Mitch, 2018).

In a recent national-scale risk assessment for NOM and THMs (O'Driscoll *et al*., 2018a), it was found that 70% of Irish public water supplies have no or only partial capacity to remove NOM prior to chlorination. Many of these supplies at risk of THM exceedances were identified along the west coast of Ireland in upland settings featuring blanket peatland with high rainfall levels. DOC precursors in these catchments were thought to be recalcitrant, high molecular weight humic and fulvic acids originating from terrestrial

(allochthonous) catchment sources. Although many CDBP formation potential studies have tended to focus on upland peat catchments (e.g. Gough *et al*., 2014; Golea *et al*., 2017), lowland agricultural pastures may also be important sources of DOC precursors (O'Driscoll *et al*., 2018a). The dissolved organic nitrogen (DON) precursors of NDBPs are less well understood but may occur in the fraction of DOM that is readily assimilated by microbes and derived from autochthonous sources (e.g. primary production) such as algae and microbial biomass (Goslan *et al*., 2017). This labile DOM comprises low molecular weight, hydrophilic or neutral amino acids, proteins, amides, purines and polysaccharides, which are rich in DON (Heinz *et al*., 2015). A further major source of labile DOM in Ireland is wastewater derived from intensified dairy agriculture, which can impart high biological oxygen demand and ammonium (NH_4) fluxes in headwater streams (Harrison *et al*., 2019) with large pathogen loads and consequently higher disinfection requirements.

Climate change is predicted to lead to a rise in NOM export from Irish river catchments over the next century and to changes in the timing of water availability and the intensity of rainfall events (O'Driscoll *et al*., 2018b). Rising DOM precursor concentrations in raw water coupled with anticipated population expansion in Ireland will increase pressure on drinking water treatment infrastructure, with the potential for an increased cancer burden in the population exposed to DBP exceedances through drinking water (e.g. Evlampidou *et al*., 2020). Although DOM removal is possible with advanced treatment technologies (e.g. granular activated carbon filtration), process optimisation informed by real-time, raw water DOM quality measurements may represent a more economical solution for small, vulnerable supplies in Ireland. While total organic carbon (TOC) concentration or UV absorbance at 254 nm (A_{254}) is routinely monitored as a surrogate during water treatment, TOC concentrations alone may be of limited value in understanding the molecular reactivity of the DOM pool in response to disinfection with chlorine. In the last two decades, technological advances in UV–visible (UV-vis) and fluorescence excitation–emission matrix (EEM) spectroscopy have offered a unique perspective on DOM quality and process dynamics in freshwater environments (Coble, 1996). EEM spectroscopy is a low-cost, non-destructive and selective optical monitoring approach that can provide rapid information on the concentration, composition and origin of fluorescent dissolved organic matter (FDOM) in aquatic environments (Figure 1.1). When EEM is combined with parallel factor analysis (PARAFAC), complex FDOM admixtures can be "unmixed" to reveal independent precursor surrogates involved in DBP formation, highlighting the utility of excitation–emission matrix-parallel factor analysis (EEM-PARAFAC) in the investigation and management of drinking water chemical quality (O'Driscoll *et al*., 2018a, 2020).

1.2 Literature Review

A systematic literature review and critical evaluation of the potential utility of ubiquitous environmental fluorophores, identified by EEM-PARAFAC, as low-cost, optical surrogates for DBP formation were undertaken by Fernández-Pascual *et al*. (2023) (and references therein) as part of the present research. Out of 378 peer-reviewed studies identified,

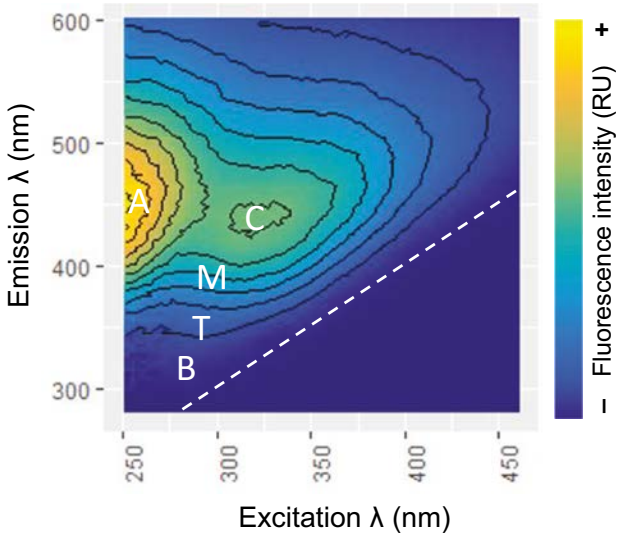


Figure 1.1. Typical freshwater fluorescence EEM showing common environmental fluorescence centres (peaks) where PARAFAC fluorophores are most frequently identified, labelled using the nomenclature of Coble (1996). EEM data from O'Driscoll *et al*. (2017). Peaks A and C, humic- and fulvic-like fluorescence; peak M, microbial humic-like fluorescence (HLF); peak T, tryptophan-like fluorescence; peak B, tyrosine-like fluorescence. The area below the white dashed line indicates the EEM noise region where the excitation wavelength (λ_{ex}) > emission wavelength (λ_{em}). RU, Raman unit.

empirical data were extracted from 45 selected articles matching the search criteria and scope. From these, 218 individual statistically significant linear relationships (e.g. $p \leq 0.05$; coefficient of determination (R^2) ≥ 0.5) were identified with environmental PARAFAC components and 41 individual DBP compounds over 10 DBP classes including the following: four THMs, nine HAAs, two haloketones (HKs), one haloacetaldehyde, one halogenated furanone (X-furanone), six iodinated THMs, one unclassified CDBP, one HNM, four HANs, two haloacetamides, nine N-nitrosamines (NAs) and one cyanide. Within these identified relationships, 135 had strong associations ($R^2 \geq 0.7$) with one or more PARAFAC components. The grouping of statistically significant relationships by DBP class extracted per PARAFAC fluorophore region (Figure 1.1) is shown in Figure 1.2. The majority of studies included (84%) considered composite samples collected from water treatment plants under controlled conditions, whereas the remainder (16%) considered samples collected from the distribution network post chlorination.

The majority of studies (53%) considering the links between PARAFAC components and DBP formation potential identified humic- and fulvic-like fluorophores (e.g. peaks C and A shown in Figure 1.1) and THMs and HAAs, highlighting the potential utility of these fluorophores as surrogates for regulated DBP formation. The formation of unregulated and potentially more harmful NDBP classes (e.g. HNMs, HANs, NAs and cyanogen chloride) was equally associated with both protein-like fluorophores (peaks T and B shown in Figure 1.1) and humic/fulvic-like components, except where autochthonous FDOM sources (e.g. algal or microbial biomass) were dominant. In these cases, protein-like fluorophores (such as tryptophan-like PARAFAC components) alone were strong predictors of NDBP formation. The review highlighted the challenges in transposability of regression models between study areas and different FDOM sources that may exhibit contrasting chemical composition but yield similar fluorescence intensities. Overall, EEM-PARAFAC was demonstrated to provide a much more selective surrogate for DBP formation potential and speciation than single-parameter optical approaches such as A_{254}.

1.3 Objectives

Given that pathogen inactivation is the primary goal of drinking water treatment and that DBPs represent a widespread exposure route to potential carcinogens, achieving a balance between these competing safety drivers presents a formidable challenge for public

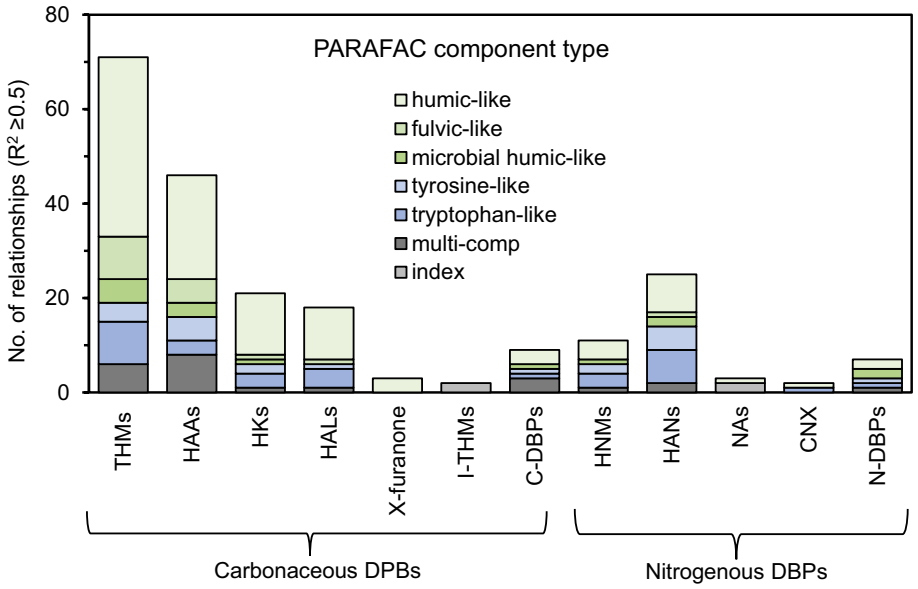


Figure 1.2. Grouping of 218 statistically significant linear relationships ($R^2 \geq 0.5$, $p < 0.05$) according to DBP class and PARAFAC fluorophore region, extracted from 45 peer-reviewed articles that met the scope and search criteria. CNX, cyanogen chloride; I-THM, iodinated trihalomethane. Adapted from Fernández-Pascual *et al.*, 2023; reproduction licensed under CC BY 4.0 DEED (https://creativecommons.org/licenses/by/4.0/deed.en).

health protection in Ireland and around the world. Hence, the overall goal of the present research is to better understand the role of FDOM at the sub-catchment scale in the formation of 20 individual DBP species (namely four THMs, nine HAAs, two HKs, four HANs and one HNM) for the first time in Ireland. This goal will be addressed through the following objectives:

- To develop a laboratory batch chlorination methodology to quantify DBP formation under an excess of chlorine from environmental samples subject to primary filtration only as a single pre-treatment step. This method is suitable for quantifying the maximum DBP yield for a given DOC concentration but may not represent actual concentrations present in consumer tap water from the same raw water source.
- To develop a nested sub-catchment scale, spatiotemporal FDOM and hydrochemical field sampling and analysis programme in two Irish drinking water catchments in order to evaluate DBP formation risk and speciation under a range of landcover types and groundwater found in Ireland.
- To evaluate the potential of FDOM and supporting hydrochemical parameters as predictors for the formation of regulated (THMs and HAAs) and unregulated DBP classes (HKs, HANs and HNMs) using machine learning techniques. Insights from machine learning may be used to inform early warning tool development for raw water quality monitoring at drinking water treatment plant intakes for both quantitative DBP concentration prediction and presence/absence forecasting for water supplies with limited or no DOM removal prior to chlorination.
- To provide a critical appraisal of portable dual-wavelength fluorimetry for continuous *in situ* monitoring applications and DBP formation for deployment in riverine raw water sources typical of small drinking water treatment plants in Ireland.

1.4 Study Areas

Two tributaries of the River Lee in County Cork were selected for detailed sub-catchment-scale investigation, including surface water and groundwater sampling, batch chlorination experiments and a hydrochemical analysis programme. The study areas form two small sub-catchments located 60 km (Bunsheelin River) and 10 km (Dripsey River) west of Cork city, Ireland (51°53′N, 8°28′W) (Figure 1.3). The Dripsey River catchment rises in the Boggeragh Mountains and has an area of 35 km^2, with an approximately linear geometry and an average gradient of 4.6% (±3.2%). The long-term average annual rainfall is 1535 mm (±174 mm) (Banteer Lyre; 52°04′48″N, 8°51′00″W) (Met Éireann, 2023). The Dripsey is a major tributary of the Inniscarra Reservoir, serving 132,000 consumers in Cork city, where THM concentrations ranged between 16 and 82 µg L^{-1}, with an average of 51.6 ± 18.6 µg L^{-1} as observed from 2015 to 2018 in the city distribution network (Cork City

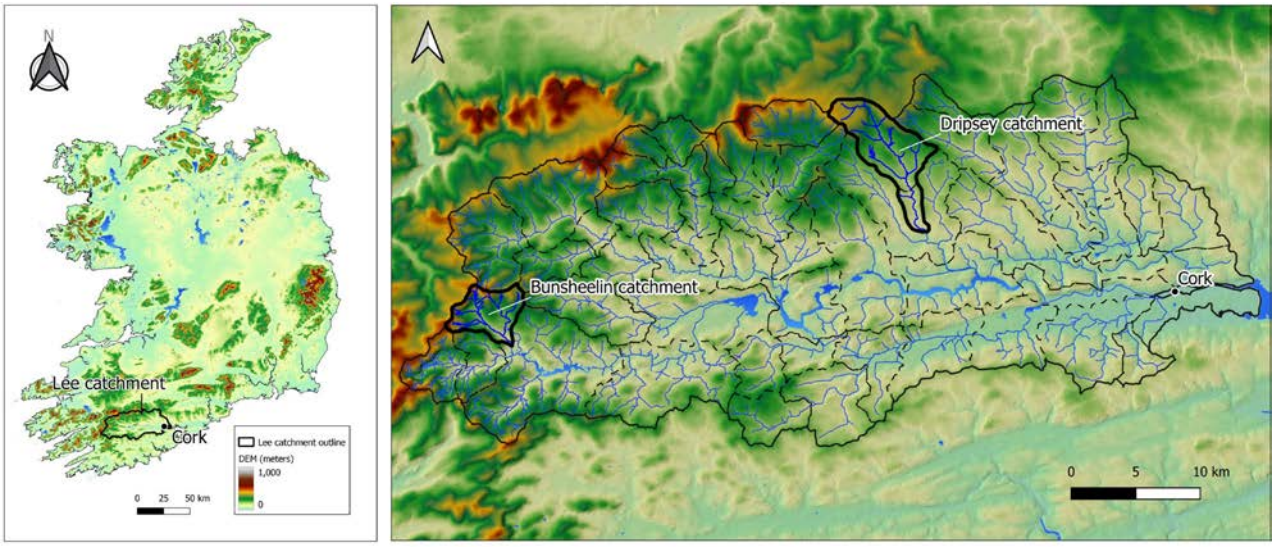


Figure 1.3. Location of the River Lee Basin in County Cork and location of the Dripsey and Bunsheelin study areas.

Council, 2023). Land use in the Dripsey study area is predominantly improved agricultural grassland on mineral soils with areas of peatland and coniferous plantation forestry in the Boggeragh uplands (Harrison *et al*., 2019).

The Bunsheelin River sub-catchment is a steep-sided basin with an average gradient of 9.4% (±7.3%) and a long-term average annual rainfall of 1968 mm (±351 mm) (M. Ballingeary Meelin Mtn.; 51°53′24″N, 9°15′00″W) (Met Éireann, 2023). The Bunsheelin River is the principal public water supply for the village of Ballingeary (*Béal Átha an Ghaorthaidh*), which has a population of 239 people where THM4 concentrations ranged from 46 to 127 µg L^{-1}, with an average of 96 ± 42 µg L^{-1}, as observed from 2015 to 2018 (Cork City Council, 2023). This supply was on the EPA remedial action list due to recurring THM4 exceedances but has since been removed after an upgrade of the treatment plant to include granular activated carbon filtration for removal of THM precursors prior to chlorination (EPA, 2020). The bedrock geology of both catchments comprises Devonian Old Red Sandstone formations, which are considered a locally productive aquifer used for drinking water which may contribute substantial subsurface water flow to stream networks during dry weather periods (Dupas *et al*., 2017). Agricultural intensity (predominantly dairy production) differs for each sub-catchment. In contrast to the Bunsheelin study area, the Dripsey site exhibits a greater density of farmyards where much of the low-lying areas are planted with perennial ryegrass which is intensively managed, often with very high cattle stocking rates (Scanlon *et al*., 2004).

The two sub-catchments were selected for inclusion in the present research based on the following criteria: (i) history of elevated THMs or THM exceedances, (ii) presence of key landcover types found in Ireland (namely peatland, conifer plantation forestry and intensive agriculture on mineral soils) and (iii) geographical proximity to University College Cork laboratories. Neither catchment has significant sources of urban wastewater input; therefore, this potential precursor source is not considered in the current research. The Dripsey study area also hosts a major groundwater monitoring network (Moe *et al*., 2010), which allows DBP formation risk under different landcover types for surface water to be compared with groundwater.

1.5 Source Water Sampling Programme

A total of 24 surface water monitoring points were established in both catchments, broadly following a Strahler stream order approach. Stream orders sampled ranged from 1 to 3 and from 1 to 4 in the Bunsheelin and Dripsey study areas, respectively. Samples were collected approximately each month, between February and November 2021 for Dripsey (DY01–DY12) and between April and November 2021 for Bunsheelin (BS01–BS12). With the exception of the February 2021 monitoring round at Dripsey, all other samples were collected during average or low flow conditions. Groundwater samples from a hillslope transect of three monitoring well clusters (DR1–DR3) (comprising three or four depth-specific piezometers) (Moe *et al*., 2010) and a low flow spring at DY12 (not used for water supply) were collected in July and October 2021.

Grab samples were collected using a 1 L stainless steel pendulum beaker (V2A, Buerkle), which was rinsed three times with sample water prior to collection. Groundwater samples were collected as part of the national groundwater monitoring programme (EPA, 2021). Samples for DBP formation potential, DOM optical properties and hydrochemical analysis were filtered on-site using single-use 0.7 µm GF/F (Whatman GD/X) syringe filters which were immediately chilled to 4°C and stored in constant darkness. Samples for DBP formation potential were collected with zero headspace in pre-cleaned 250 mL amber glass bottles with polytetrafluoroethylene (PTFE)-coated septum caps and adjusted to pH 7.0 ± 0.2 with 6 mL of 0.1 M phosphate buffer. Samples for DOC/dissolved inorganic carbon (DIC) and DOM optical properties were collected in 40 mL and 20 mL amber glass vials with PTFE-coated caps, respectively, with zero headspace. Samples for chloride (Cl^-), bromide (Br^-), total dissolved nitrogen (TDN), ammonium (NH_4-N), nitrate (NO_3-N), nitrite (NO_2-N) and DON determination were filtered on-site using dedicated 0.45 µm cellulose acetate syringe filters and collected in 60 mL Nalgene HDPE bottles, which were immediately chilled to 4°C before analysis or frozen (<–20°C) where analysis could not be completed within 48 hours of collection. Temperature, pH, electrical conductivity and dissolved oxygen were measured on-site using a pre-calibrated multi-parameter sensor (Manta+20, Eureka Water Probes).

2 Predicting Disinfection By-product Formation Using Machine Learning

2.1 Laboratory and Analytical Methods

Fluorescence EEM spectra were acquired using a PerkinElmer LS-50B spectrophotometer with a 1 cm quartz cell. Excitation–emission slits were set to a 5 nm band-pass with excitation wavelengths (λ_{ex}) scanned from 250 to 590 nm at 10 nm intervals with emission from 300 to 650 nm (at 0.5 nm intervals). Blank EEM spectra of ultrapure water were collected daily prior to analysis to monitor instrument stability. UV-vis absorbance spectra were acquired on a UV-vis spectrophotometer (Genesys 10S UV-VIS, Thermo Scientific) over wavelengths of 200–900 nm at 0.5 nm intervals. EEM spectra were corrected for inner filter effects, emission intensities were normalised to the Raman peak of pure water and the independent underlying fluorophores present in each sample were identified using a six-component PARAFAC model in the staRdom R package (Pucher *et al*., 2019) (v1.1.25). Common fluorescence indices and peaks (Figure 1.1) were also calculated using the staRdom R package. All EEM fluorescence and UV-vis absorbance spectra were acquired within 24 hours of sample collection at ambient laboratory temperature (20°C).

DOC and DIC concentrations were determined on a Shimadzu TOC-VPCH/CPN analyser within 48 hours of collection with a limit of quantification (LQ) of 0.12 and 0.15 mg C L^{-1}, respectively. TDN, NH_4-N, NO_3-N and NO_2-N concentrations were determined using automated Lachat Quick-Chem 8000 flow injection analysis colorimetric methods with LQs of 0.083, 0.002, 0.01 and 0.02 mg N L^{-1}, respectively. DON was calculated as TDN – (NO_3-N + NO_2-N + NH_4-N) with an LQ of 0.09 mg N L^{-1}. All nitrogen species analyses were carried out within 48 hours of sample collection with 10% of the samples collected analysed in duplicate. Cl^- was determined using the ferricyanide method on a Shimadzu Pharma Spec UV-1700 UV-vis spectrophotometer. Br^- was quantified using ion chromatography (Quick-Chem method 10-510-00-1-A) using a 2.2 mM $NaHCO_3$/2.8 mM Na_2CO_3 eluent with a flow rate of 2.0 mL min^{-1} and an LQ of 0.002 mg Br L^{-1}. The relative standard deviation for all hydrochemical analytical methods was < 5%.

Field filtered and pH-adjusted (pH = 7.0 ± 0.2) 250 mL water samples for chlorination were spiked with a sodium hypochlorite solution (EMPLURA, Merck) using a 5/1 Cl_2/DOC ratio so that free chlorine was in excess of DOC throughout chlorination (Goslan *et al*., 2017). DBP formation potential experiments were carried out over a 72 ± 2 hour period, representing the typical contact time of small Irish water treatment plants, in constant darkness at 25.0°C ± 0.2°C using a temperature-controlled incubator (Incubator IN30, Memmert). Residual free chlorine was determined at the start and end of each experiment using the *N*,*N*-diethyl-*p*-phenylenediamine method on a portable colorimeter (Pocket Colorimeter II, Hach). Samples containing no residual chlorine at the end of the experiment were excluded from further analysis. Chlorinated samples were quenched with L-ascorbic acid solution using a 5/1 L-ascorbic acid/Cl_2 ratio to arrest further DBP formation. Experimental blanks containing ultrapure water with phosphate pH buffer were chlorinated at 16 ± 8 mg Cl_2 L^{-1} for every sample batch of 12 and processed in the same manner as field samples. After quenching, aliquots were carefully removed for DBP analysis. THMs were analysed in duplicate with aliquots collected in 40 mL amber glass vials (containing 0.5 mL of 0.1% of sodium thiosulfate solution) filled without headspace and closed with PTFE-coated septum caps. THM samples were stored at 4°C and analysed within 14 days of quenching by an accredited contract laboratory (Eurofins, Shannon, Ireland) following United States Environmental Protection Agency (USEPA) method 524.2 with an LQ per species of 0.9 µg L^{-1}. Aliquots for HAAs were collected in 60 mL HDPE bottles filled without headspace and immediately frozen (< –20°C) until extraction and analysis according to a modified version of USEPA method 552.3 with an LQ per HAA species of 1 µg L^{-1}. HAN, HK and HNM aliquots were extracted within 14 days of collection and stored at < –20°C prior

to analysis according to USEPA method 551.1 with an LQ per species of 0.1 µg L^{-1}.

2.2 Disinfection By-product Formation and PARAFAC Components

The overall results of the chlorination experiments for the two surface water catchments (Bunsheelin, BS01–BS12, and Dripsey, DY01–DY12) and groundwater (DR1–3) are presented in Figure 2.1 to illustrate chlorine demand as a function of measured DOC concentration. Overall, the apparent chlorine demand was found to be relatively constant over the range of DOC concentrations observed (e.g. Cl_2/DOC = 3.47 ± 0.07 mg Cl_2 mg C^{-1}), suggesting that the natural oxidant demand was dominated by DOC for all field samples. DOC concentrations could be reproduced accurately (R^2 = 0.94) from A_{254} measurements, as widely reported in the literature (Fernández-Pascual *et al*., 2023). Of the 20 individual DBP species considered in this study, THMs (n = 199), HAAs (n = 175), and HAN, HK and HNM (n = 160) classes exhibited an overall concentration range from < LQ to 2132 µg L^{-1}, with a broad distribution in concentrations observed (Table 2.1). Eight individual DBP species were noted to occur in over 75% of samples, namely TCM (244.64 ± 283 µg L^{-1}), bromodichloromethane (BDCM) (28.5 ± 15 µg L^{-1}), DBCM (4.3 ± 5.5 µg L^{-1}), DCAA (5.1 ± 4.6 µg L^{-1}), TCAA (6.8 ± 9.4 µg L^{-1}), dichloroacetonitrile (DCAN) (2.7 ± 1.2 µg L^{-1}), trichloronitromethane (TCNM) (1.7 ± 2.2 µg L^{-1}) and trichloropropanone (TCP) (4 ± 2.5 µg L^{-1}). Tribromomethane (bromoform) (TBM), trichloroacetonitrile (TCAN), bromochloroacetonitrile (BCAN), dibromoacetonitrile (DBAN) and dichloropropanone (DCP) were present in less than 50% of treated field samples.

A six-component PARAFAC model was derived from 205 EEM spectra, with an average leverage of 0.029 (± 0.031) which exhibited a convergence of 253 models over 1000 runs (Figure 2.2). The six-component model exhibited a very good average residual error of the measured signal of 0.091 ± 0.122. The split half analysis showed a good similarity of the splits, as reported by a Tucker's congruence coefficient (TCC) close to 1 (0.92 ± 0.22 and 0.95 ± 0.17 for excitation and emission, respectively). The first three components (C1–C3) constituted 98% of the model variable importance (Table 2.2). Components C1 and C2 are well known terrestrially derived (allochthonous) humic-like fluorophores with maximum fluorescence intensity (F_{max}) at λ_{ex} = 250 nm and emission wavelength (λ_{em}) = 431 nm, and λ_{ex} = 250 nm and λ_{em} = 482 nm, respectively, which are widely described in the literature in freshwater environments (Yamashita *et al*., 2011; Peleato *et al*., 2016). Component C3 is a tryptophan-like fluorophore with an F_{max} at λ_{ex} = 250 nm and λ_{em} = 280 nm, representing low molecular weight proteinaceous and nitrogenous DOM which is also well known in the literature (e.g. Shutova *et al*., 2014; Marcé *et al*., 2021). Components C4 to C6 did

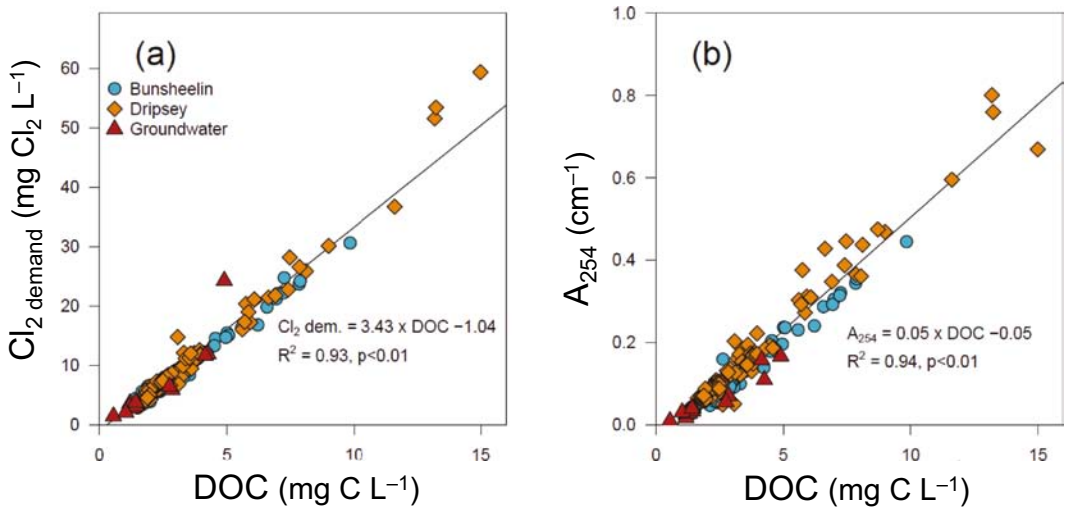


Figure 2.1. Relationship between observed DOC concentrations and (a) measured chlorine demand after 72-hour chlorination and (b) A_{254}. The three study areas are considered as a single group in the linear regression analysis.

Table 2.1. Measured concentrations and occurrence of 15 DBP parameters after 72 hours, under an excess of chlorine where samples were above the method LQ

Disinfection by-product	Concentration (µg L^{-1})		Occurrence (%)
	Range (minimum–maximum)	Mean	
THMs			
TCM	1.55–2132	244.64	100
BDCM	<LQ–101.39	28.45	92.8
DBCM	<LQ–55.37	4.28	78.9
TBM	<LQ–9.87	0.24	7.1
THM4	1.55–2159.12	277.76	100
HAAs			
DCAA	<LQ–28.60	5.07	97.1
TCAA	<LQ–64.16	6.82	94.9
HAA5	1.00–91.19	12.13	100
HANs			
TCAN	<LQ–4.16	0.10	24.5
DCAN	<LQ–9.37	2.66	98.6
BCAN	<LQ–4.25	0.57	42.4
DBAN	<LQ–2.8	0.16	16.2
HNMs			
TCNM	<LQ–24.19	1.71	87.5
HKs			
DCP	<LQ–2.77	0.13	11.3
TCP	<LQ–22.96	3.99	99.4

not return any matches in the OpenFluor database (Murphy *et al*., 2014) and were not considered further.

2.3 Machine Learning Approach

Fluorescence indices, common absorbance parameters, PARAFAC component F_{max} values and hydrochemical parameters of source water samples were all considered as potential predictive variables. The final predictive variables were selected for each DBP species or class to limit the collinearity between predictive variables, discard noise variables and optimise model predictions as described in Xiao *et al*. (2018) using the following procedures: linear correlation, principal component analysis and input cancellation. The selected predictive variables were then used to train a species-specific DBP machine learning model written in the R package (version 4.2.0). DBP machine learning was completed using an ensemble, nested cross-validation approach and repeated 100 times, which produced robust and unbiased performance estimates regardless of the sample size (Vabalas *et al*., 2019). The total sample dataset was randomly split into 80% and 20% for the training and validation datasets, respectively. The training data were classified to account for DBP presence or absence (e.g. <LQ) using a support vector machine (SVM). For samples exhibiting >50% of a specific DBP presence, quantitative training data were used to train three independent machine learning techniques, i.e. bagging tree (BAG), generalised boosted regression models (GBMs) and neural networks (NNETs), in parallel. Finally, model performance and prediction parameters were taken from the average of 300 independent models (100 repetitions for each technique) resulting in a higher predictive accuracy and less bias towards one predictive technique than a single model (Xiao *et al*., 2018; Shtein *et al*., 2019). The averaged relative importance was determined and a sensitivity analysis (SA) was undertaken to evaluate how the selected predictive variables influence individual DBP species predictions.

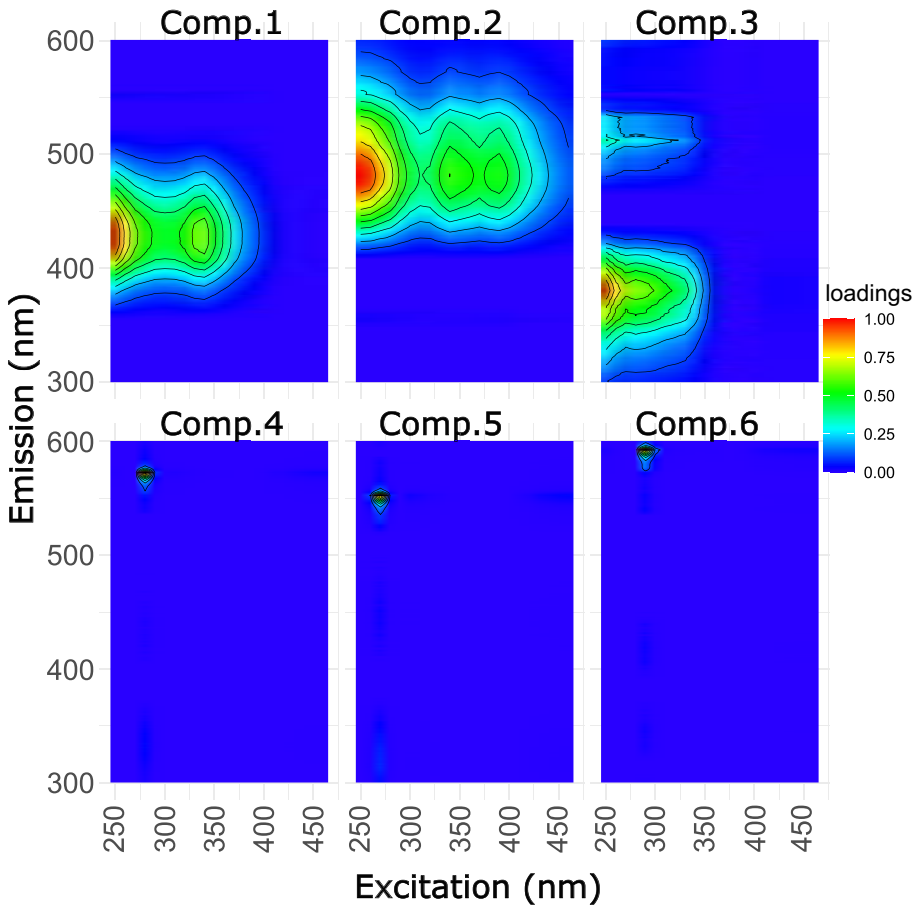


Figure 2.2. Modelled PARAFAC component loadings for components C1 to C6 derived from 205 individual EEM spectra acquired in the present study.

Table 2.2. PARAFAC component C1–C3 wavelength pair and fluorophore identification by comparison with the OpenFluor database of published EEM-PARAFAC spectra

Component	$\lambda_{ex,\ max}$	$\lambda_{em,\ max}$	Variable importance (%)	Type	Number of matches in OpenFluor[a]	Best match reference[b]
C1	250	431	55.8	Humic-like	50 (5)	Peleato *et al*., 2016; Yamashita *et al*., 2011
C2	250	482	36.9	Humic-like	75 (19)	Peleato *et al*., 2016
C3	250	380	5.3	Tryptophan-like	8 (0)	Marcé *et al*., 2021; Shutova *et al*., 2014
C4	280	573	0.9	Unknown	No match	Unknown
C5	270	553	0.6	Unknown	No match	Unknown
C6	290	593	0.4	Unknown	No match	Unknown

Component refers to PARAFAC component, $\lambda_{ex,\ max}$ and $\lambda_{em,\ max}$ refer to the wavelength corresponding to the F_{max} of the excitation and emission, respectively. Variable importance indicates variable importance in the PARAFAC model.

[a]Accessed on 25 July 2022 with a TCC≥0.95 and in parentheses with a TCC≥0.98. The OpenFluor database is described in Murphy *et al*. (2014).

[b]Best match using with PARAFAC model associated with similar freshwater environments, TCC≥0.98 for C1 and C2, and TCC≥0.95 for C3.

2.4 Predictive Variable Selection

From the exploratory data analysis and variable selection procedures, some variables were removed from the potential selection, as they contained little information or were correlated with the variables ultimately selected (Figure 2.3). NO_2-N presented many samples below the LQ of 0.002 mg N L^{-1} (44.6%) and was therefore not considered a representative predictive variable within the dataset. NO_3-N, TDN and NH_4-N were used to calculate DON and, from the data cancellation procedure results, DON and NH_4 were considered the best representatives for nitrogen species predictive variables and consequently retained in the model. Moreover, the selection of NH_4-N as a predictive variable may be supported from a mechanistic point of view where free chlorine can react quickly with NH_4-N to form the monochloramine that may participate in NDBP formation (Bond *et al*., 2011). Data cancellation and principal component analysis findings both suggest that between 10 and 12 variables were optimal for the prediction of observed DBPs. During cancellation, the root mean squared percentage error (RMSPE) declined after removing one variable and became asymptotic after removal of more than 12 variables. Furthermore, the input cancellation procedure results suggest that PARAFAC components were better predictive variables than traditional fluorescence "peak picking" (e.g. peaks B, T, A, M and C from Coble, 1996) (Figure 1.1), highlighting the importance of independent fluorophores in DBP formation.

Ten variables, namely DIC, DOC, NH_4^+, DON and Br^- concentrations, the humification index (HIX), PARAFAC components C1–C3 and the proportion of component C3 relative to the sum of components C1–C6 (% C3), were ultimately selected as the optimal variables to predict species-specific DBP concentrations using an ensemble model (Table 2.3).

2.5 Ensemble Model and Interpretation

Of the 20 individual DBP species investigated, the rate of occurrence of some species, namely

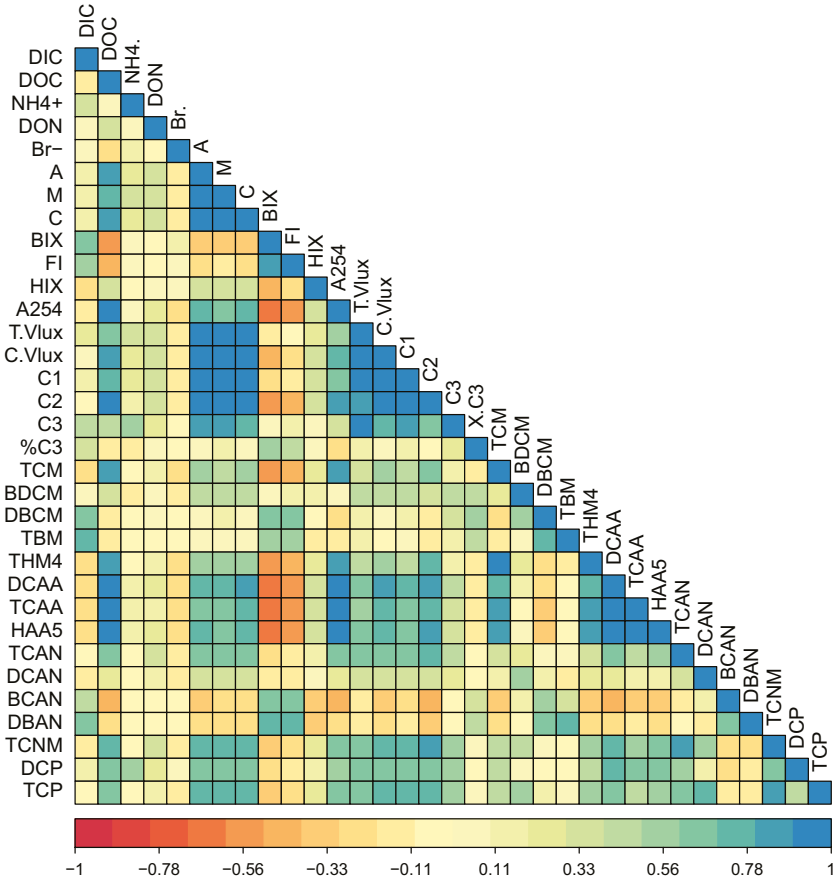


Figure 2.3. Correllogram illustrating Pearson's correlation coefficients between all candidate predictive variables and observed DBPs.

Table 2.3. Observed ranges and occurrence of hydrochemical parameters and DOM optical properties including the HIX and PARAFAC components C1–C3 selected as the predictive variables for the ensemble machine learning model

Predictive variable	Range (minimum–maximum)	Mean	Occurrence (%)
DIC (mg C L^{-1})	1.57–62.2	8.66	100
DOC (mg C L^{-1})	0.55–14.99	3.19	100
NH_4^+ (µg N L^{-1})	3–2127	30	100
DON (mg N L^{-1})	<LQ–1.971	0.31	96
Br^- (µg L^{-1})	<LQ–117	29	73.7
HIX [-]	0.602–0.978	0.92	100
C1 (RU, F_{max})	0.06–1.94	0.42	100
C2 (RU, F_{max})	0.02–1.12	0.25	100
C3 (RU, F_{max})	0.05–0.82	0.13	100
% C3 (RU %)	2.77–24.61	10	100

RU, Raman unit.

TBM, TCAN, BCAN, DBAN and DCP, was very low, with these species occurring in only 7%, 25%, 42%, 16% and 11% of samples, respectively (Table 2.1). As a result, only binary presence/absence SVM models were trained for these species. Table 2.4 presents the results of the training and validation of the optimal models. The average performance of the optimal models on the training dataset was strong for predicting species presence (SVM accuracy = 96.8% ± 3.0%) and high for predicting the concentration (R^2 = 0.83 ± 0.06, precision = 12.6% ± 2.6%, RMSPE = 29.0% ± 9.9%) of DBPs, where BAGs and GBMs demonstrate similar performance and NNETs overperformed for all DBPs. Overall, the optimal models slightly underpredicted DBPs by 4% ± 2% (calculated on the validation regression using the average observed value). Finally, model calibration and validation using the

Table 2.4. Average model performance for the prediction of DBP species using the optimal ensemble model

	Training dataset				Validation dataset			
DBP	SVM accuracy (%)	R^2	Precision (%)	RMSPE (%)	SVM accuracy (%)	R^2	Precision (%)	RMSPE (%)
TCM	n.p.	0.90	13.9	27.2	n.p.	0.76	20.1	46.4
BDCM	n.p.	0.85	10.0	18.9	n.p.	0.45	15.4	29.1
DBCM	94.2 ± 1.0	0.77	17.4	45.0	89.6 ± 5.4	0.43	26.4	64.3
TBM	99.8 ± 0.3				95.7 ± 3.3			
THM4	n.p.	0.90	11.9	23.6	n.p.	0.73	18.1	40.1
DCAA	n.p.	0.91	8.3	24.7	n.p.	0.80	12.6	39.9
TCAA	n.p.	0.94	10.3	25.4	n.p.	0.86	14.8	38.2
HAA5	n.p.	0.93	11.2	26.5	n.p.	0.86	17.5	40.1
TCAN	92.9 ± 1.3				85.9 ± 5.9			
DCAN	99.6 ± 0.4	0.80	11.4	24.6	96.7 ± 1.1	0.07	19.2	47.8
BCAN	91.5 ± 1.6				85.6 ± 5.6			
DBAN	95.8 ± 0.9				90.5 ± 5.2			
TCNM	94.9 ± 0.9	0.76	13.4	40.9	91.5 ± 4.8	0.33	20.0	70.8
DCP	97.9 ± 1.0				92.9 ± 4.4			
TCP	n.p.	0.80	10.4	22.8	n.p.	0.33	15.9	43.4

Accuracy: accuracy of classification computed as the percentage of predicted true value divided by the measured true value; n.p.: not performed. R^2 values were calculated using the linear regression of observed data as a function of the predicted. Precision corresponds to the standard deviation of the three machine learning techniques over the 100 replicate predictions.

surface water data show comparable performance to the optimal models (maximum percentage difference considering all species had an $R^2 < 5\%$ and RMSPE < 26%), suggesting that the optimal models properly handle the difference between surface water and groundwater DOM predictive variables.

From Table 2.4, it can be inferred that predicted concentrations were overfitted ($R^2_{training} > R^2_{validation}$), so multiple lines of evidence were explored to validate the modelling approach. First, the variable-to-sample-size ratio was much lower (e.g. 10/214 = 0.04) than the thresholds identified by Vabalas *et al*. (2019), who reported that high thresholds can introduce dimensionality issues and degrade model accuracy. Second, the one-way analysis of covariance showed no statistical differences ($p \geq 0.65$) between the correlation of predicted and observed DBPs for the training dataset compared with the validation dataset. Third, the bias and variance comparison reported in Zhang and Ling (2018) showed greater bias than the small variance, which tends to indicate that the optimal models were not overfitted. Finally, an analysis using duplicate data augmentation showed no significant differences for training data R^2 and RMSPE values, but did show increases for validation data, suggesting that a machine learning dataset should ideally be very large in size (≥ 1200 data points). From the findings arising from the validation steps undertaken above, it was concluded that the performance of the optimal models was not significantly affected by overfitting but that a larger sample size would be desirable in future studies.

Interrogation of the optimal models was undertaken to understand how the predictive variables affect individual DBP concentrations and to provide insights into possible pathways of DBP formation. An SA of the optimal models is presented in Figure 2.4, in which all predictive variables were individually interrogated across the observed range of values for all DBP species (Table 2.1). The SA percentage difference indicates how each variable influences prediction. The SA results were dominated by the influence of DOC across all THM and HAA species (apart from BDCM), which is consistent with previous models (e.g. Sadiq and Rodriguez, 2004). Apart from DOC and PARAFAC components C1 and C2, most individual variables did not strongly influence the model; however, the machine learning techniques employed demonstrate that all selected predictive variables were important for DBP prediction and that models were more strongly influenced by predictive variable combinations than any one single variable alone. As a result, the

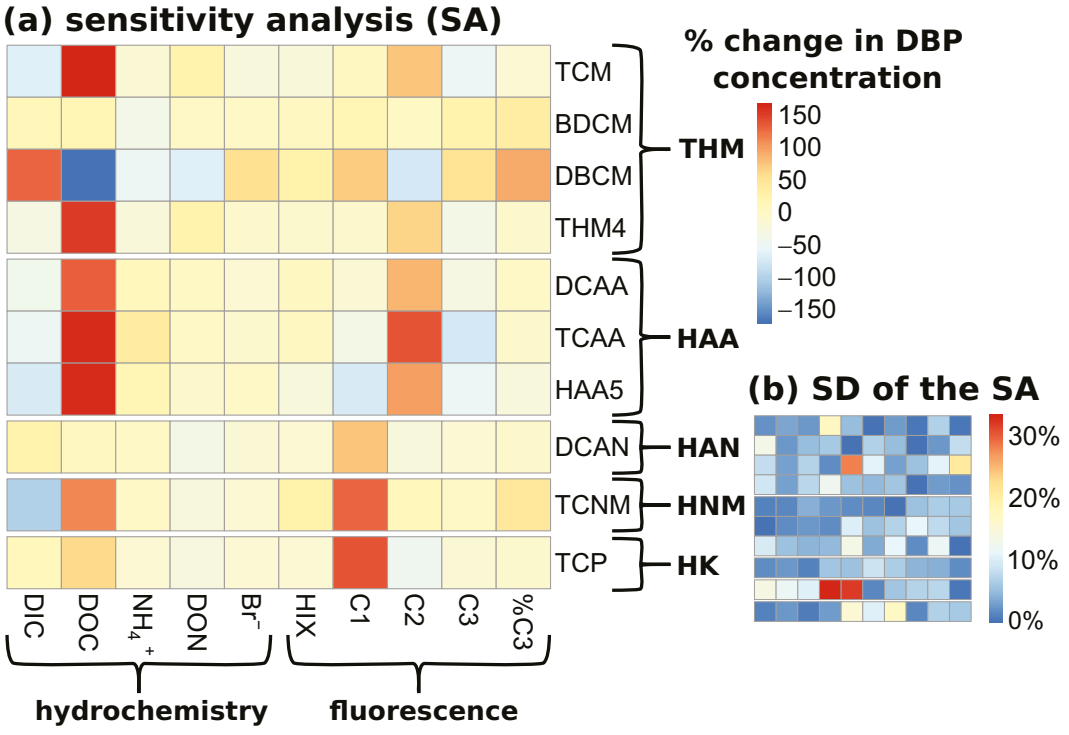


Figure 2.4. (a) Results of the SA performed on the ensemble model and (b) standard deviation (SD) of the SA results.

performance was similar when a model was trained after removing any individual variable (maximum percentage differences considering all DBP species were $R^2 < 3\%$ and RMSPE < 9%) but declined when two variables were removed by an average difference RMSPE of 10% ± 10% (maximum < 19%).

It is interesting that the NDBP species (DCAN and TCNM) were more influenced by DOC and C1 (a humic-like component identified by PARAFAC) than DON and C3 (a tryptophan-like PARAFAC component) or potentially chloramine (not determined in the present study), which is believed to be indirectly produced from the reaction between HOCl and NH_4-N (Bond *et al*., 2011). C3 was not strongly represented in the sample dataset (the percentage of C3 ranged between 2.77% and 24.61%, Table 2.3). Other studies have reported an association between C3 and NDBP formation and wastewater and microbial DOM sources (Fernández-Pascual *et al*., 2023). A relatively weak effect on the SA results was induced by Br^-, which may be explained by the relatively low Br^- concentrations compared with DOC concentrations (Br/DOC = 0.011 ± 0.013), for which a ratio below 0.01 is known to generate a lower proportion of brominated species than of chlorinated DBPs (Francis *et al*., 2010). Consequently, the potential influence of Br^- on the prediction of brominated DBPs will be outside the range of model training in the present study. The influence of carbonate species (e.g. DIC) observed in the SA on brominated species (DBCM) follows previous observations (Francis *et al*., 2010) and is supported by the correlation results (Figure 2.3), in which all brominated species were moderately positively correlated with DIC ($R^2 \geq 0.5$, $p \leq 0.01$). Carbonate species have been thought to catalyse the oxidation of Br^- by HOCl (with rate constants up to 1000 times greater), which generates brominated species in preference to chlorinated DBPs (Brodfuehrer *et al*., 2020).

2.6 Dissolved Organic Matter Spectral Model

The optimal ensemble model represents the most suitable combination of predictive variables to quantify a broad spectrum of DBPs in raw drinking water sources prior to chlorination. However, as some hydrochemical variables require offline wet chemical techniques for determination (e.g. DIC), the optimal model has some limitations when potentially applied for online early warning monitoring applications. Consequently, a spectral machine learning model was trained using predictive variables that could be monitored using online spectrometry alone (spectral model). To this end, the spectral models used fluorescence “peak picking” (A, M, C), HIX, A_{254} (as a DOC surrogate) and two wavelength pair regions representing humic-like and tryptophan-like fluorescence (TLF) (Figure 2.3), which could be measured by either a fluorimeter connected to a flowthrough cell or a commercial *in situ* sensor (Fox *et al*., 2022). A_{254} was shown to be a highly robust surrogate (Figure 2.1) for DOC concentrations across all samples (e.g. $R^2 = 0.94$, $p < 0.01$), which is well established. The spectral model (Table 2.5) showed moderate differences compared with the ensemble model (Table 2.4), with R^2 ranging from –28% to –4.5% and RMPSE ranging from 12% to 39% in the training dataset for the different DBP species. Overall, the spectral model findings suggest that TCM, DBCM, DCAA, TCAA and TCNM species would be particularly suitable (e.g. R^2 difference < 10%) for online monitoring applications using fluorescence measurements combined with A_{254} as a surrogate for DOC.

Table 2.5. Comparison between the optimal model performance (Table 2.4) and the spectral model suitable for online-only measurements at source water intakes expressed as percentage difference

	Training dataset (% difference)			Validation dataset (% difference)		
DBP species	SVM accuracy (%)	R^2	RMSPE (%)	SVM accuracy (%)	R^2	RMSPE (%)
TCM	n.p.	–8.9	39.0		–6.7	13.3
BDCM	n.p.	–27.9	36.1		–72.5	24.6
DBCM	2.0	–8.9	23.4	3.8	–45.9	33.3
TBM	0.7			9.8		
THM4	n.p.	–8.6	36.9		–3.7	12.9
DCAA	n.p.	–4.5	23.7		–3.5	3.5
TCAA	n.p.	–6.2	21.7		–7.5	4.5
HAA5	n.p.	–4.6	24.1		–5.7	5.2
TCAN	2.5			10.7		
DCAN	0.0	–26.6	27.4	0.2	–17.8	–1.3
BCAN	4.0			10.7		
DBAN	0.4			2.5		
TCNM	1.4	–9.3	12.2	0.5	10.7	–10.6
DCP	1.2			3.5		
TCP	n.p.	–15.8	27.5		–2.4	2.3

Accuracy: accuracy of classification computed as the percentage of predicted true value divided by the measured true value; n.p., not performed. R^2 values were calculated on the linear regression of observed as a function of the predicted.

3 Spatial Variability of Disinfection By-products and Precursors at Sub-catchment Scale

3.1 Landcover Analysis and Nested Sub-basin Approach

The two sub-catchments selected for field investigation (Figure 3.1) have comparable land cover, with a total area of 51.1 km^2 encompassing 23.7% tree cover (predominantly managed conifer plantations) and 75.0% agricultural grassland (mainly improved pasture) (Zanaga *et al*., 2022). A further 29.0% and 20.5% of the Bunsheelin and Dripsey catchment areas, respectively, are underlain by thin cutaway peat or peaty podzol soils (hereafter classified as "peatland" soil (PS)) in upland areas, with well-drained mineral soils underlying lowland areas classified as podzolic soil (Teagasc and Cranfield University, 2014). Peatland areas in Ireland are known to be important contributors to surface water DOM fluxes and THM precursors (O'Driscoll *et al*., 2018a), but mineral soils and conifer plantation DOM contributions to DBP precursor fluxes are less well understood.

To investigate the spatial associations between DBP precursors and potential sources, a nested sub-basin approach was employed so that the contributing sub-basin area upstream of each monitoring point was considered an independent catchment for landcover analysis, with BS12 and DY12 representing the entire Bunsheelin (17 km^2) and Dripsey (35 km^2) sub-catchment areas, respectively. The contributing sub-basin areas for monitoring points were computed using the European Digital Elevation Model (Copernicus Land Monitoring Service, 2016) and the Irish river network shapefile (EPA, 2017) by way of an automated procedure in the Geographic Resources Analysis Support System (GRASS) version 8.2 (van der Kwast and Menke, 2022). BS05 was a small unmapped field drain with an unknown catchment so was excluded. The sub-basins were classified into three landcover categorical variables (grassland on mineral soil, grassland on peatland and tree cover), where a specific land cover represented more than 50% of the mapped catchment area (Figure 3.1) using WorldCover at 10 m resolution (Zanaga *et al*., 2022). PS was defined as the sum of the area covered by peatland as shown in the Derived Irish Peat Map (Connolly and Holden, 2009), the CORINE Land Cover model (Copernicus Land Monitoring Service, 2016) and the National Subsoils Map (Fealy and Stuart, 2009). Most of the samples were collected from sub-catchments (Figures 3.2 and 3.3) on grassland on mineral soils (57.6%) followed by grassland on PSs (16.1%) and tree cover (14.7%). PSs and tree cover tended to be associated with headwater (first order) stream reaches of both catchments in areas of high topography, where elevated rainfall is expected. Overlap between tree cover and PSs is expected given that plantation forestry occurs extensively at higher elevations in the Dripsey study area (Figure 3.3).

In addition, groundwater data, accounting for 11.5% of the samples, were treated as a separate category, representing an alternative to surface water as a source of drinking water. The locations of piezometer clusters DR1–3 and the low flow spring are indicated in Figure 3.1c. The groundwater monitoring network of DR1–3 comprised 11 piezometers, with a depth range of 1.8 to 70.1 m below ground level, with response zones installed in the transition zone, and shallow and deep bedrock pathways in Old Red Sandstone (Devonian Ballytrasna Formation) arranged along a hillslope transect (Moe *et al*., 2010).

3.2 Spatial Variability of Disinfection By-product Precursors

Surface water samples were categorised between three landcover classes and, based on the nested sub-basin approach, are compared with each other and groundwater using a paired Wilcoxon signed-rank test in Figure 3.4. Groundwater had a significantly higher DIC concentration ($37.97 \pm 18.13\ mg\ C\ L^{-1}$) and a lower DOC concentration ($2.65 \pm 3.03\ mg\ C\ L^{-1}$) than any surface water landcover category sample, suggesting that DOC mineralisation is an important biogeochemical process in the groundwater system underlying the lower Dripsey catchment. Groundwater was also elevated with respect to NO_3 ($1.87 \pm 1.61\ mg\ N\ L^{-1}$) and PARAFAC component C3 ($15.70\% \pm 4.71\%$), which is interpreted as a

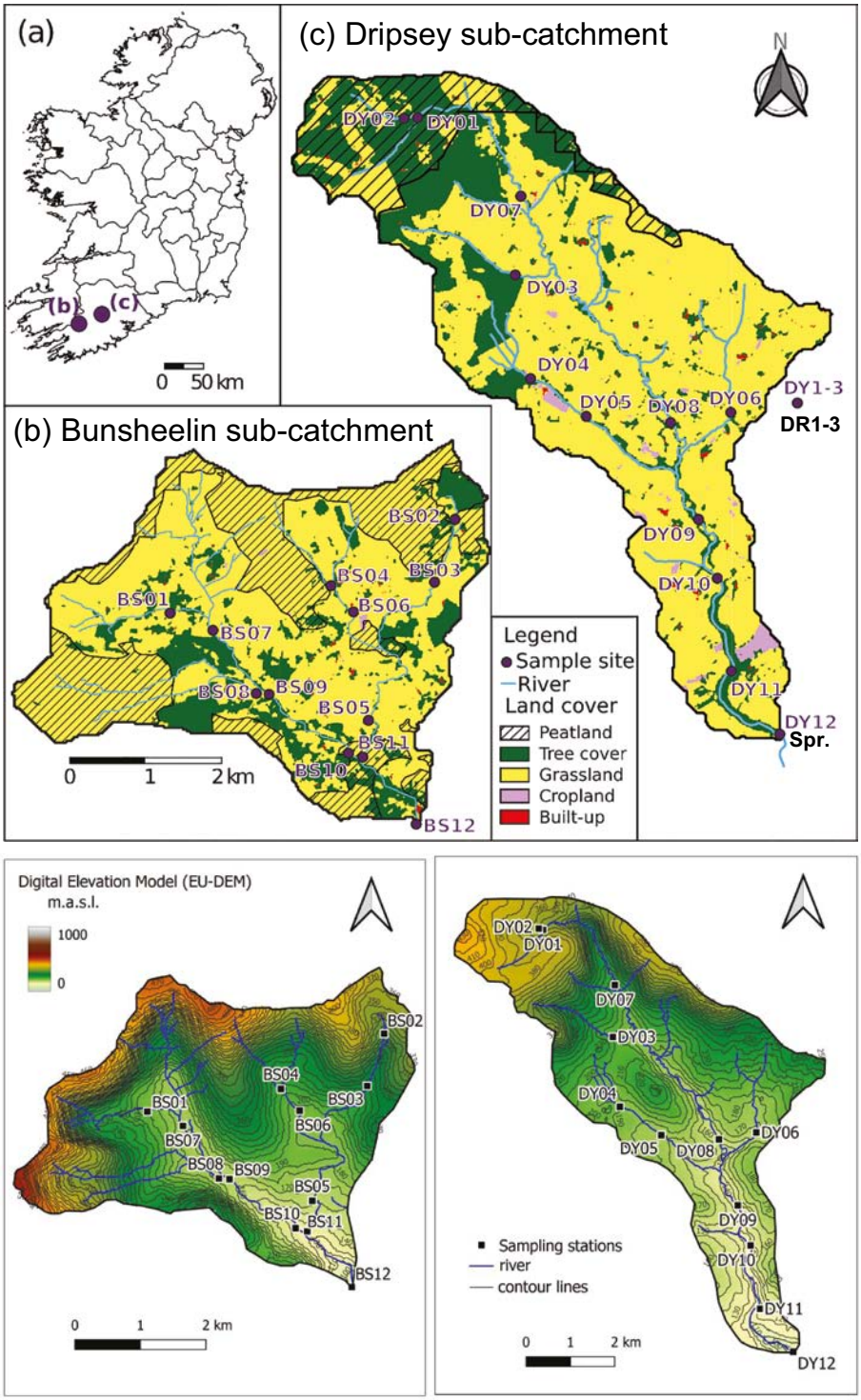


Figure 3.1. (a) Locations of (b) Bunsheelin and (c) Dripsey sub-catchment study areas with mapped land cover. The lower panels illustrate the topography of and monitoring points within each of the sub-catchments. Groundwater monitoring locations are indicated on panel (c), with DR1–3 and a natural spring not used for water supply (Spr.) highlighted in bold.

tryptophan-like fluorophore (Table 2.2). In addition, the samples from grassland on mineral soil exhibited elevated NO_3 ($1.67 \pm 1.3\,mg\,N\,L^{-1}$), comparable to that of groundwater. These findings suggest that groundwater pathways and grassland mineral soils may be potential sources of nitrogen precursors for NDBPs (Bond *et al.*, 2011), particularly in the lower Dripsey catchment, where the export of inorganic

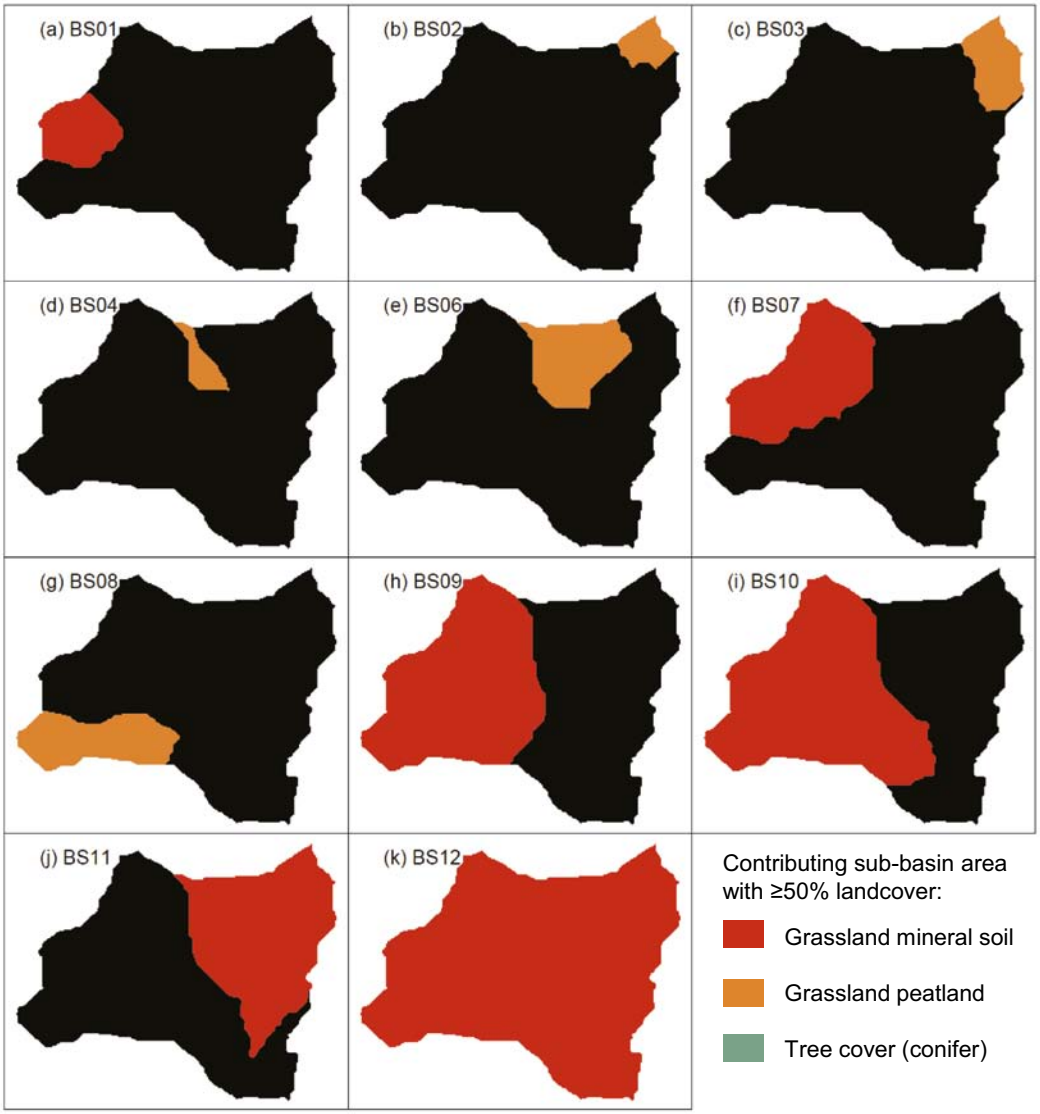


Figure 3.2. Interpreted predominant landcover classes (≥50% area) for 11 (a–k) nested sub-basins (BS01–BS04 and BS06–BS12) of the Bunsheelin sub-catchment (total area 17 km²). The sub-basin for BS05 was not mapped and is not included.

nitrogen to surface water via subsurface pathways may be favoured.

DOC concentrations were highest at monitoring points draining tree cover (5.43 ± 3.36 mg C L^{-1}), with the lowest concentrations associated with the grassland categories (3.04–3.72 mg C L^{-1}). Conversely, DON concentrations were generally similar for all landcover categories apart from grassland compared to tree cover ($p \leq 0.05$). Moreover, DON did not follow a similar distribution to the tryptophan-like PARAFAC component C3. The finding that the highest DOC concentration was associated with tree cover sub-basins contrasts with the findings of large-scale studies that suggest that DOC and DON export generally increases with agricultural intensification, which is associated with grassland on mineral soil sub-basins (Stanley and Maxted, 2008; Graeber *et al*., 2012; Heinz *et al*., 2015). The difference in these findings may be the result of peaty podzolic or organic-rich soils also being present in sub-basins classified as tree cover (managed plantation conifer), which occur in first-order tributaries in the headwaters of the Dripsey River (DY01, DY02, DY03, DY07) at a relatively high elevation in the Boggeragh Mountains (Figures 3.1 and 3.3). Artificially managed drainage of organic-rich soils associated with commercial forestry operations may also play a role in increased DOC export from these headwaters (Pschenyckyj *et al*.,

Figure 3.3. Interpreted predominant landcover classes (≥50% area) for 12 (a–l) nested sub-basins (DY01–DY12) of the Dripsey sub-catchment (total area 35 km^2).

2023). It is also interesting to note that the humic-like and fulvic-like PARAFAC components (C1 and C2) did not follow a similar distribution to DOC among landcover categories, with the highest contributions originating from grassland on PS (42.45% ± 9.46% and 23.84% ± 6.06% for C1 and C2, respectively), suggesting a lower fluorescent component of DOM in headwaters draining tree cover. As an important

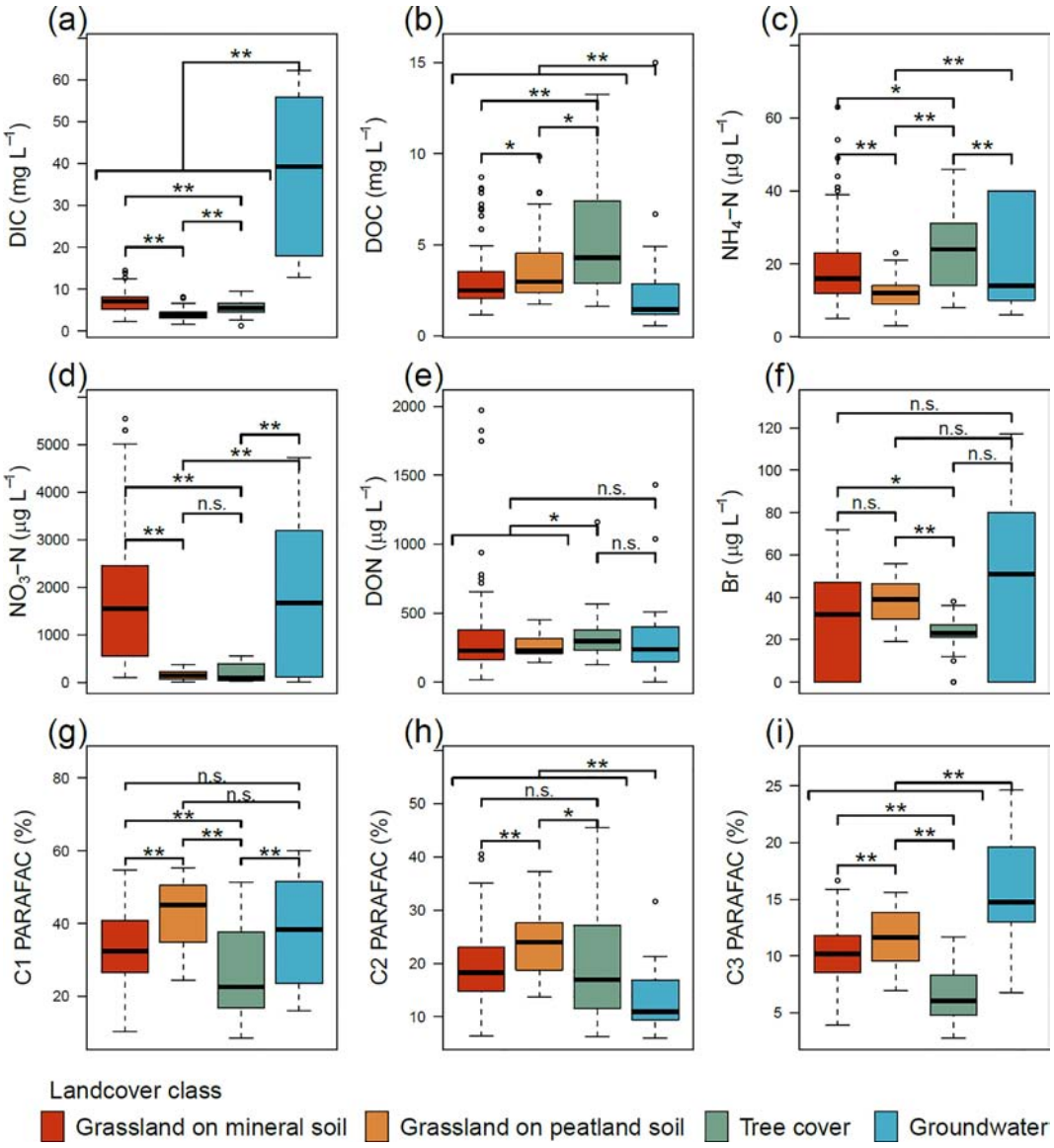


Figure 3.4. Distribution and comparison of DBP precursor and hydrochemical parameters over three landcover categories for surface water and groundwater (a–i). Horizontal bars denote paired, Wilcoxon signed-rank test variables, with statistically significant differences being marked with asterisks (*$p \leq 0.05$; **$p \leq 0.01$) and n.s. indicating $p > 0.05$ (i.e. not significant). NH_4-N data for groundwater (c) were highly skewed, so outliers have been omitted for clarity.

precursor for brominated DBPs, Br^- concentrations were quite variable, with a broad range observed in groundwater, which was not statistically distinct ($p > 0.05$) from the surface water landcover categories.

3.3 Disinfection By-product Formation Potential and Dissolved Organic Carbon Yield

As in the preceding section, surface water samples grouped between three landcover classes are compared with one another and with groundwater (using paired Wilcoxon signed-rank tests) in Figure 3.5 for measured total DBP formation concentrations and yield (with respect to DOC) for regulated and unregulated DBP classes. CDBPs include THMs (THM4), HAAs (HAA5) and HKs (sum of two HKs (HK2)), of which THM4 and HAA5 are regulated. NDBPs include HANs (sum of four HANs (HAN4)) and, as a single representative of the HNM class, TCNM. For the groundwater samples, THM4 (92.6 ± 102.8 µg L^{-1}), HAA5 (6.42 ± 17.21 µg L^{-1}) and TCNM (1.19 ± 4.83 µg L^{-1}) had formation concentrations that were much lower than those of

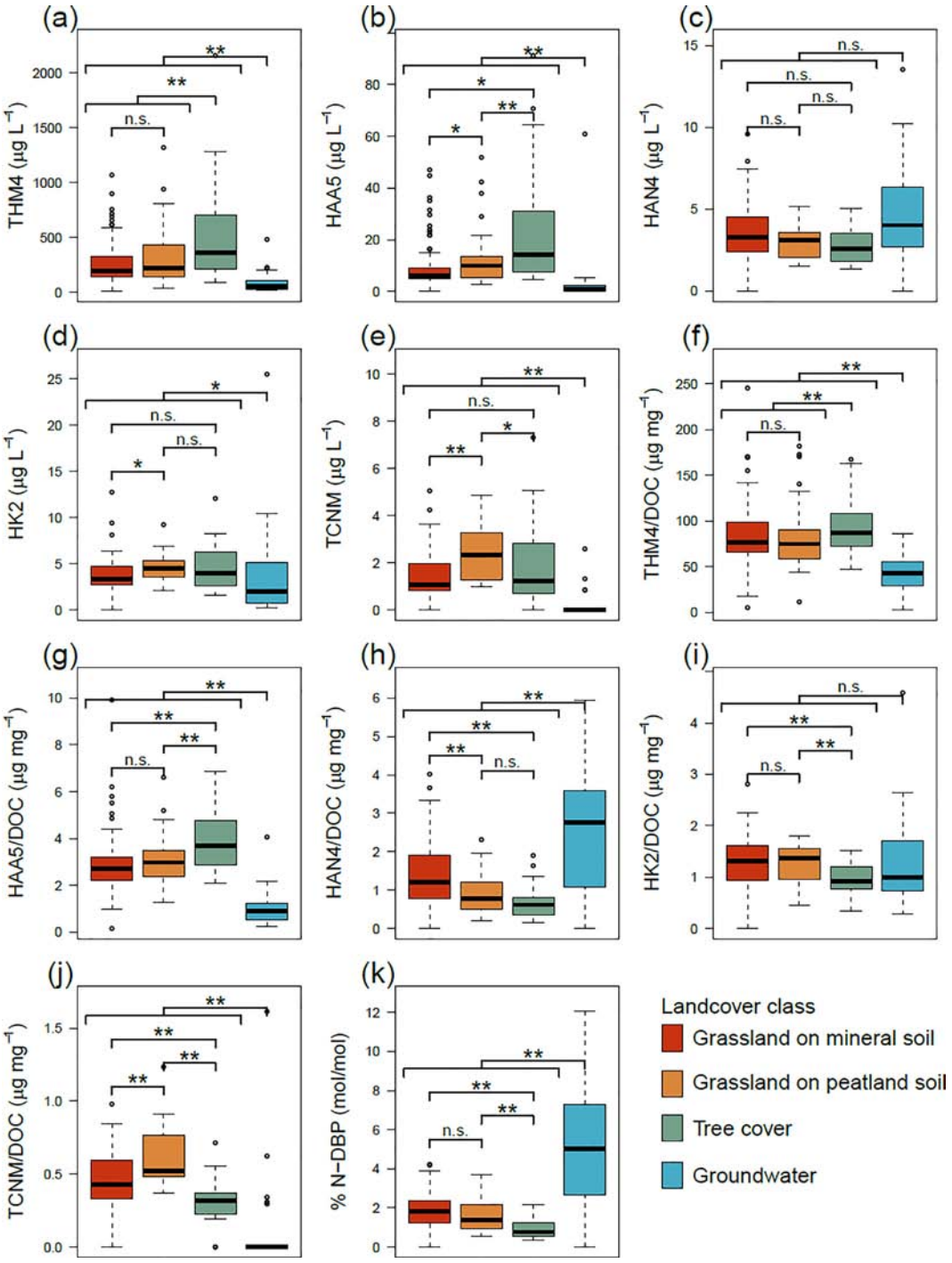


Figure 3.5. Distribution and comparison of the total concentrations of DBPs formed (a–e) and DOC yields (f–k) for regulated and unregulated DBP classes over three landcover categories for surface water and groundwater. Horizontal bars denote paired, Wilcoxon signed-rank test variables, with statistically significant differences being marked with asterisks (*$p \leq 0.05$, **$p \leq 0.01$) and n.s. indicating $p > 0.05$.

any of the landcover classes and were between two and six times lower than the highest surface water concentration. This may be explained by the low DOC concentrations (2.65 ± 3.03 mg C L^{-1}) reported from the groundwater samples in comparison with samples from surface water (Figure 3.4). The tree cover category, comprising samples from the Dripsey uplands (DY01–DY03, DY07), covering an altitude of 200 m to over 350 m above sea level (Figures 3.1 and 3.2), had the highest concentrations of THM4 (535 ± 479 µg L^{-1}) and HAA (23.6 ± 22.44 µg L^{-1}), significantly exceeding those of any other landcover

class and groundwater samples. This is consistent with the highest DOC concentrations (5.43 ± 3.36 mg C L^{-1}) also being reported for this category (Figure 3.4). The highest concentrations of TCNM (2.36 ± 1.15 µg L^{-1}) were measured in samples from grassland on PS, whereas HK2 and HAN4 concentrations showed little or no statistically significant spatial variability among landcover categories.

DBP yield (µg mg C^{-1}) is an important indicator of DOC precursor reactivity towards specific DBP formation pathways and is calculated by dividing the DBP concentration (µg L^{-1}) by the measured DOC concentration (mg C L^{-1}) (Goslan *et al*., 2017) (Figure 3.5). Groundwater samples exhibited much lower reactivity for DOC than any of the surface water landcover categories for THM4 (42.5 ± 22.0 µg mg^{-1}), HAA5 (1.2 ± 1 µg mg^{-1}) and TCNM (0.13 ± 0.35 µg mg^{-1}) formation. DOC from samples originating from the tree cover category in the Dripsey headwaters (DY01–DY03, DY07) (Figure 3.3) was found to be the most reactive for the formation of regulated DBPs, including THM4 (94.0 ± 29.7 µg mg^{-1}) and HAA5 (3.9 ± 1.25 µg mg^{-1}). These sub-basin types can be considered a high risk for THM and HAA exceedances, in terms of both DOC precursor export and reactivity under the chlorination conditions adopted in this study. Although HAN4 formation concentrations did not exhibit statistically significant differences between any category pairs, the HAN4 yield was much higher for groundwater samples (2.54 ± 1.73 µg mg^{-1}) than for the highest surface water category of grassland on mineral soil (1.406 ± 0.844 µg mg^{-1}). Moreover, the molar fraction of NDBPs to total DBPs (mol mol^{-1}) was between two and four times greater for groundwater than for any surface water category. These findings suggest a potentially significant role for groundwater as a source of unregulated NDBPs where DOC concentrations are relatively low but have greater reactivity for HAN4 formation with elevated inorganic (NO_3-N) and organic (PARAFAC C3%) nitrogen precursors (Figure 3.4) available. It is interesting to note that groundwater was not a potential source of TCNM, another NDBP of the HNM class with formation concentrations and DOC yields greater for all surface water landcover categories (Figure 3.4).

3.4 Molecular Fractionation of Dissolved Organic Matter in Surface Water

Liquid chromatography–organic carbon detection–organic nitrogen detection (LC-OCD-OND) was used to quantify and differentiate specific molecular fractions of DOM (as DOC and DON) (Huber *et al*., 2011) in surface water samples from the November 2021 (n = 24) monitoring round. Samples were collected in 60 mL amber glass vials with PTFE-coated septum caps with zero headspace and kept at 4°C in darkness until analysis within 7 days of collection. LC-OCD-OND analysis was undertaken using a DOC Labour GmbH instrument (model V9) with a typical LQ of 0.01 mg L^{-1} for individual DOM fractions with a < 4% and < 18% relative standard deviation between measurements for DOC and DON, respectively (Heinz *et al*., 2015). The molecular fractions of DOM obtained (comprising both DOC and DON) are described in Table 3.1.

From the LC-OCD-OND analysis of surface water samples (Figure 3.6) (BS01–BS12 and DY01–DY12), the hydrophobic organic carbon (HOC) fraction appears to be the most important DOC fraction in tree cover sub-basin samples (71.9% ± 17.3%), which support higher aromaticity values associated with enrichment of more aromatic recalcitrant humified organic material (Kalbitz *et al*., 2002). The second largest DOM fraction found in grassland samples is that of building blocks (BBs) and low molecular weight substances (LMWSs), which together are in agreement with a larger C3 PARAFAC component, reflecting a greater contribution of microbially

Table 3.1. DOM molecular fractions assigned from LC-OCD-OND analysis following Huber *et al*. (2011)

DOM fraction	Molecular weight (g mol^{-1})	Fraction composition
Hydrophobic organic carbon	–	All organic matter retained by the column
Non-humic high molecular-weight substances	> 20,000	Biopolymers, polysaccharides and proteins
Humic substances	~1000	Humic or fulvic acid
Building blocks	300–450	Breakdown products of humic fraction
Low molecular weight substances	< 350	Organic acid, alcohols, aldehydes and ketones

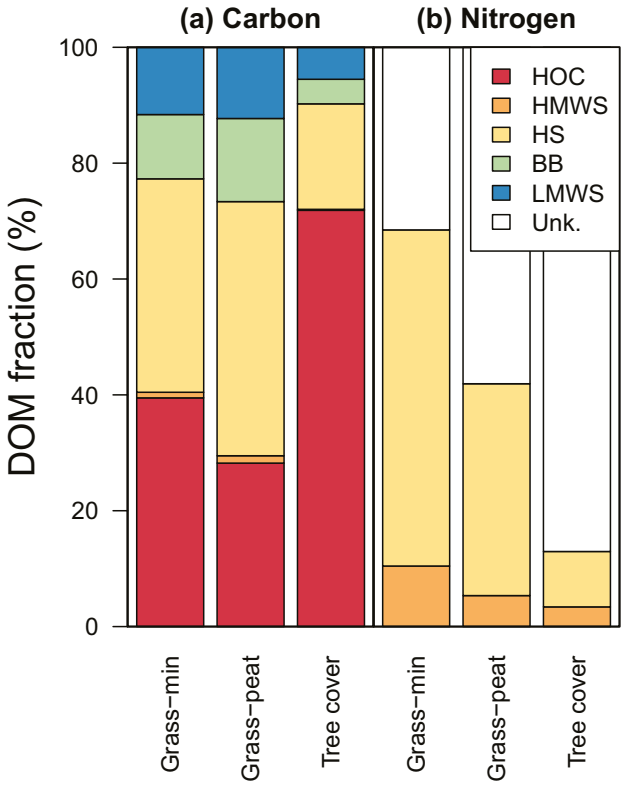


Figure 3.6. DOC (a) and DON (b) molecular fractionation according to landcover categories of grassland on mineral soil ($n = 15$), grassland on peatland ($n = 5$) and tree cover ($n = 4$). HMWSs have a molecular weight of > 20,000 g mol^{-1}, HSs of ~ 1000 g mol^{-1}, BBs of 300–450 g mol^{-1} and LMWSs of < 350 g mol^{-1}. Grass-min, grassland on mineral soil; Grass-peat, grassland on peatland; Unk., unknown.

transformed DOM and fresh allochthonous plant material in grassland samples. Finally, the ratio between the hydrophilic fractions (e.g. the sum of non-humic high molecular weight substances (HMWSs), humic substances (HSs), BBs and LMWSs) and the HOC fraction was largest for the grassland on peatland (2.56 ± 0.31) samples, followed by the grassland on mineral soil (1.86 ± 0.91) and tree cover (0.47 ± 0.45) sub-basin samples (data not shown).

4 Application of Dual Wavelength *In Situ* Fluorescence Spectroscopy

4.1 Introduction

This study explored the application of *in situ* fluorescence spectroscopy as a low-cost approach for the online monitoring of FDOM and potential DBP formation risk. The system selected for miniaturised *in situ* application was the VLux multi-parameter sensor (Chelsea Technologies, UK). This system is configured for *in situ* detection of chlorophyll, chromophoric DOM (reported as "FDOM" by the manufacturer) and TLF using a single-wavelength (280 nm) LED excitation source. The FDOM channel corresponds to the peak A/C region of Coble (1996) (Figure 1.1), which is also known as humic-like fluorescence (HLF). The sensor automatically corrects the measured fluorescence signal intensity for "inner filter effects" using UV absorbance at 280 nm (A_{280}) as well as turbidity and temperature effects (Fox *et al*., 2022). Each of the fluorescence channels had been calibrated at the factory (Table 4.1). The chlorophyll channel is not relevant for DBP research and is not considered further. All the fluorescence data are reported in quinine sulfate units (QSU) to facilitate data interpretation and benchtop comparison.

4.2 Dual Wavelength *In Situ* Fluorimeter Performance

A series of experiments were designed to compare the performance of the *in situ* fluorimeter with the benchtop EEM fluorescence spectrophotometer used elsewhere in this study. All solutions were prepared using ultrapure water (> 18 MΩ cm) and American Chemical Society (ACS)-grade reagents or those of higher purity. Standard reference materials for DOM, including Suwannee River Fulvic Acid (SRFA) standard III (3S101F) and Suwannee River Humic Acid (SRHA) standard III (3S101H), were obtained from the International Humic Substances Society. SRFA and SRHA stock solutions were prepared at 100 mg L^{-1}, with measured DOC concentrations of 45.7 and 32.7 mg C L^{-1} for SRFA and SRHA, respectively. L-tryptophan was obtained at TraceCert quality from Sigma-Aldrich (batch No. BCBW2043) and prepared at 0.1 g L^{-1} (DOC 63 mg C L^{-1}). Quinine sufate monohydrate (Sigma-Aldrich) stock solution was prepared at 1 g L^{-1} in 0.5 M H_2SO_4. A PS sample was collected from upstream of the DY02 location in the Dripsey catchment area (Figure 3.1) to create an artificial DOM-rich soil solution such as in Lee *et al*. (2015). Homogenised organic matter was wet sieved (< 0.2 mm) and mixed at 50 g L^{-1} with stream water obtained locally and was then agitated continuously for 24 hours. Finally, the mixture was pre-filtered through qualitative filter papers (Grade 1, Whatman) and stored in darkness at 4°C until use.

Reference materials were filtered using 0.7 µm GF/F (Whatman GD/X) prior to analysis and a series of dilutions of 0.07–45.7, 0.52–32.7, 0.35–30.0 and

Table 4.1. Manufacturer specifications for the multi-wavelength *in situ* spectrophotometer employed in this study

Parameter	λ_{ex} (nm)	λ_{em} (nm)	Dynamic range	Resolution	Substance used for factory calibration
Chlorophyll	280	682 ± 15	0–700 mg L^{-1}	0.01 mg L^{-1}	Basic blue 3
Tryptophan (TLF)	280	365 ± 25	0–600 QSU	0.02 QSU	L-tryptophan
			0–1200 mg L^{-1}	0.04 mg L^{-1}	
FDOM (HLF)	280	450 ± 25	0–600 QSU	0.02 QSU	1,3,6,8-Pyrenetetrasulfonic acid
			0–800 mg L^{-1}	0.04 mg L^{-1}	
A_{280}	280	–	0–3.5 cm^{-1}	0.002 cm^{-1}	–
Turbidity	860	–	0–1000 FNU	0.01 FNU	Formazin standard

Three fluorescence emission wavelengths are detected with a 280 nm LED excitation source, with UV absorbance also being measured at 280 nm. FNU, formazin nephelometric units.

0.003–6.3 mg C L^{-1} were prepared. EEM fluorescence and UV-vis absorbance spectra were acquired on benchtop PerkinElmer LS-50B luminescence and Thermo Fisher Scientific Genesys 10S UV-vis spectrophotometers, respectively. Blank EEM fluorescence spectra of ultrapure water were also acquired daily prior to sample analysis. The stability of the benchtop fluorimeter Xe excitation source was monitored daily using quinine sulfate at 1 µg L^{-1} (intensity at $\lambda_{ex}/\lambda_{em}$ = 350/450 nm: 9.40 ± 5.87 arbitrary units, n = 5). EEM fluorescence and UV-vis absorbance spectra were processed using the staRdom R package (version 1.1.25) (Pucher *et al*., 2019) for inner filter correction and Raman normalisation as described previously. The VLux sensor acquires fluorescence intensity data in QSU and it was therefore necessary to develop a conversion factor for direct comparison with Raman units (RU), which are standard units for corrected EEM fluorescence intensity data. An instrument-specific conversion factor for converting QSU to RU for both fluorimeters was obtained as follows: quinine sulfate stock was successively diluted with 0.5M H_2SO_4 to prepare a large dilution series, with concentrations ranging from 0.05 to 100 mg L^{-1}, similar to that used in Murphy *et al*. (2010). The intensity of EEM fluorescence in RU at $\lambda_{ex}/\lambda_{em}$ = 350/450 nm was plotted against the quinine sulfate dilution series (Figure 4.1). The slope corresponds to the conversion factor (f) for converting RU to QSU according to Lawaetz and Stedmon (2009).

For comparison with the benchtop fluorimeter and measured DOC concentrations, the VLux sensor was exposed to the same serial dilutions of reference materials (SRFA, SRHA, PS and tryptophan), filtered through 0.7 µm GF/F (Whatman GD/X) as described previously. Laboratory sensor measurements were carried out in a 500 mL glass cylinder, shielded from ambient light with aluminium foil, using the Hawk handheld readout module, logging between 3 and 8 minutes per sample. Prior to each measurement, a sample blank was measured using ultrapure water and used as the offset measurement for the series of samples. The measured values were then averaged, and the precision of the sensor signal estimated using the coefficient of variation (CV) of the measurement. The precision and accuracy of benchtop and VLux sensor measurements were evaluated based on the R^2 using linear regression. Comparisons of the benchtop fluorimeter with the VLux sensor's HLF and

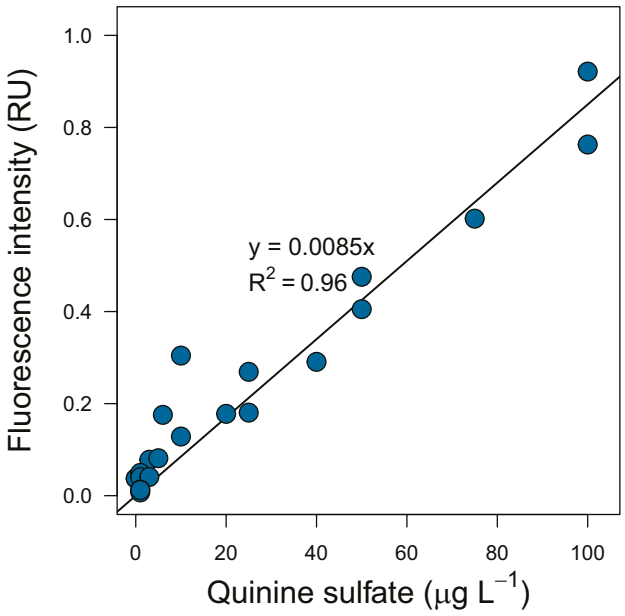


Figure 4.1. Raman-corrected fluorescence intensity of a standard series of quinine sulfate dilutions (n=23). Data were acquired using a PerkinElmer LS-50B luminescence spectrophotometer on 6 different days.

TLF channels are presented in Figure 4.2. The HLF channel of the VLux sensor showed good accuracy for PS and SRHA dilutions (R^2 ≥ 0.95) containing highly humified material but lower accuracy for SRFA dilutions (R^2 = 0.7). The TLF channel showed reasonable accuracy (R^2 = 0.87) compared with the benchtop fluorimeter for tryptophan dilutions, with some deviations observed at higher concentrations.

Comparisons between VLux sensor HLF and A_{280} channels with measured DOC concentrations are given in Figure 4.3, with comparisons between TLF and A_{280} channels with measured DOC concentrations shown in in Figure 4.4. The CV between replicate measurements for the FDOM and A_{280} channels for the humic reference materials were 0.123 ± 0.646 QSU and 0.038 ± 0.210 cm^{-1}, respectively. In comparison with measured DOC concentrations, the SRFA samples had the highest R^2 value (0.95) and the PS sample had the lowest R^2 value (0.82). The A_{280} channel showed excellent accuracy for the SRFA sample (R^2 = 0.98), but lower accuracy for the SRHA (R^2 = 0.94) and lower accuracy still for PS (R^2 = 0.87) samples. From Figure 4.4, it can be seen that TLF concentrations measured by the sensor could reproduce DOC concentrations for L-tryptophan dilutions with reasonable accuracy (R^2 = 0.93),

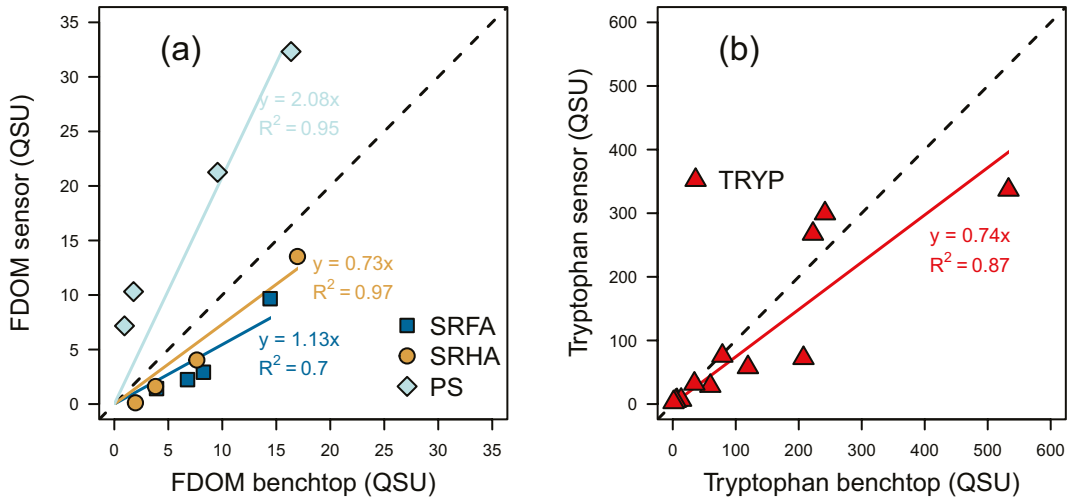


Figure 4.2. Relationship between VLux sensor measurements and benchtop measurements for (a) the HLF channel with humic reference materials and (b) the TLF channel with tryptophan as a reference material. The dashed line expresses the 1:1 relationship.

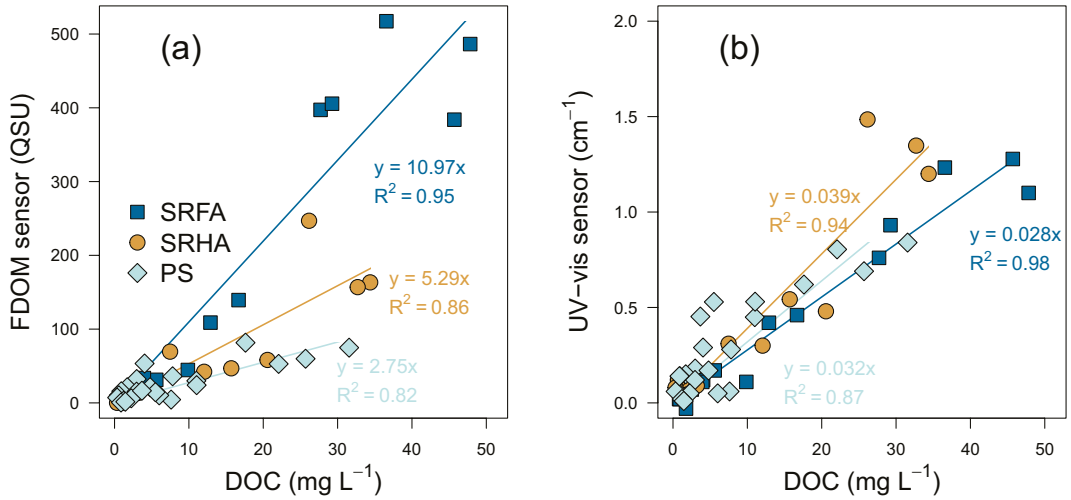


Figure 4.3. Relationship between VLux sensor measurements and measured DOC concentrations for (a) the HLF and (b) the A_{280} channels for humic reference materials.

particularly for higher concentrations of DOC. On the other hand, the A_{280} channel exhibited low accuracy ($R^2 = 0.55$) for DOC concentrations measured from tryptophan dilutions. The CV for the TLF and A_{280} channel measurements were 0.0219 ± 0.0334 QSU and 0.0029 ± 0.0056 cm^{-1}, respectively.

4.3 High-frequency Online Monitoring of Fluorescent Dissolved Organic Matter

To investigate high-frequency FDOM dynamics, an autonomous monitoring station was established adjacent to the DY12 monitoring point (Figure 3.1). The site selected for deployment was a low-lying reach of the Dripsey catchment with riparian access on private land, where the risk of theft or vandalism of field equipment was minimal. The sensor was deployed in the field at the outlet of the Dripsey catchment (DY12) between 5 October 2021 and 31 July 2022 with data logged at 15-minute intervals after a 3-minute sensor "warm-up" period. The VLux sensor was integrated with a portable real-time monitoring system (ESNET2, Meteor Communications, UK) using an SDI-12 interface with telemetry-enabled data link to an online portal. The sensor was housed in a vertical aluminium stilling well of 10 cm diameter with 1 cm perforations, allowing exchange with the

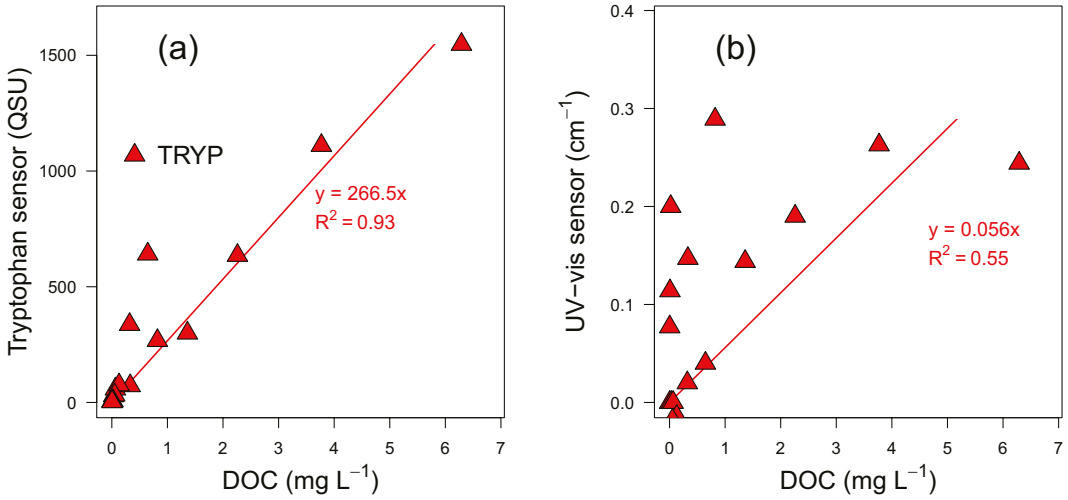


Figure 4.4. Relationship between VLux sensor measurements and DOC concentrations for (a) TLF and (b) the A_{280} channels for tryptophan reference material. Tryptophan DOC concentrations were calculated from the measured DOC value at 0.1 g L^{-1}.

surrounding river water column. A water level recorder (Diver, Van Essen Instruments, the Netherlands) was housed with the sensor to provide continuous river stage information. Initial test deployments indicated a tendency for particulate accumulation over the optical windows and the sensor was wrapped in a 2 mm nylon mesh screen to reduce particulate ingress for continuous measurement deployment. The sensor was removed for cleaning once every 2 to 3 weeks. The cleaning regime consisted of replacing the nylon mesh screen and stray light guard and soaking the instrument in warm soapy water for a minimum of 20 minutes. The optical windows were then cleaned directly using cotton swabs and a 50% (ACS-grade) isopropyl alcohol solution to remove the biofilms and associated organic matter adhering to the optical surfaces.

Raw sensor fluorescence data were corrected automatically for inner filter effects (absorbance), turbidity and temperature (using internal algorithms), and were further processed to remove noise, drift and uncalibrated data. The method detection limits (MDLs) calculated using the blank method were similar for the HLF (< 3.21 QSU) and TLF (< 3.21 QSU) channels. The MDL was defined as the minimum concentration of a substance that can be measured and for which it can be reported with 99% confidence that the analyte concentration is greater than zero (USEPA, 2016). As part of post-processing, data values below the MDL were set to 0, with values above sensor saturation (> 600 QSU) removed and values interpolated from adjacent values. Finally, drift and data noise were smoothed using a rolling mean procedure. A number of technical issues were encountered with sensor deployment, which required the instrument to be returned to its UK manufacturer on three occasions. These issues incurred significant delays and meant that the continuous data acquisition period was shorter than intended. The final processed continuous dataset for the 10-month deployment period is presented in Figure 4.5, alongside information on river stage and hourly regional rainfall totals.

Over the 10-month continuous monitoring period, hydrometeorological conditions were characterised by wet weather and associated high river levels for October 2021, December 2021 to January 2022 and early March 2022. A recession period was noted between mid-March and mid-June 2022. From the continuous *in situ* fluorescence data, HLF intensity was found to closely follow river stage with a number of major HLF export events lasting 12–24 hours, as observed in late October, early December 2021 and early March 2022, where fluorescence intensity was often in excess of 400 QSU when rainfall intensity was ≥ 5 mm h^{-1}. From the regression equation for the SRFA reference material in Figure 4.3a, peak DOC concentrations during these events were likely to be in the region of 30–50 mg C L^{-1}, indicating an extreme risk for DBP formation during these events in the Dripsey catchment. During the recession period in 2022, HLF intensity declined to a steady baseline of < 20 QSU. TLF intensity showed similar peak intensity to HLF at

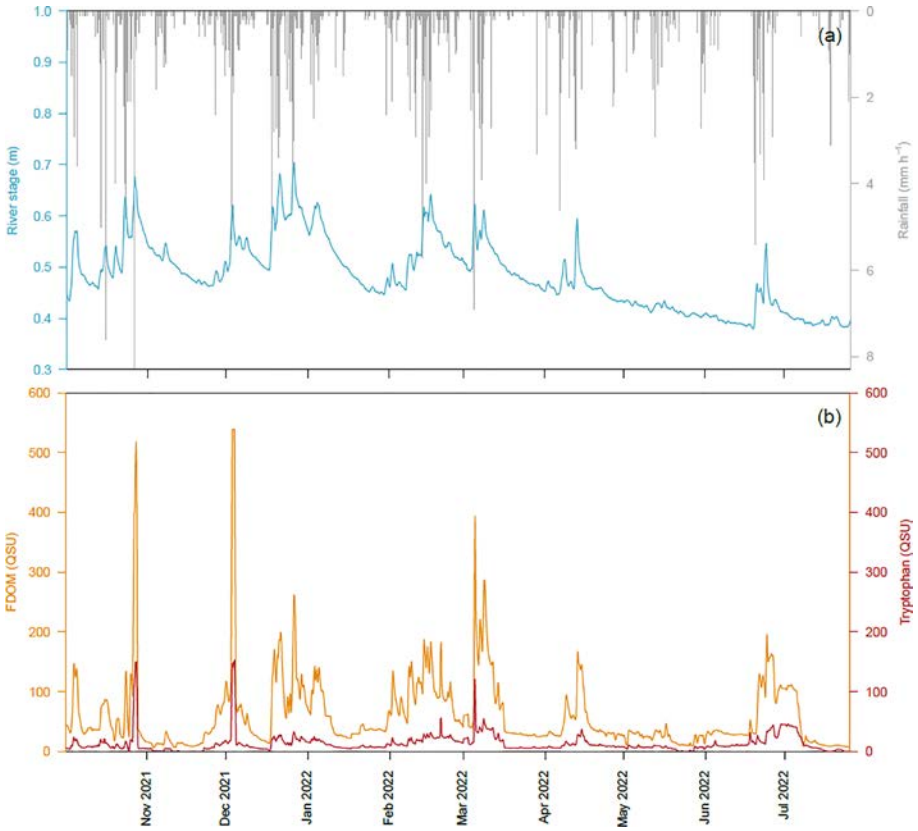


Figure 4.5. (a) River stage at DY12 and hourly regional rainfall data for Cork Airport (Met Éireann, 2023) and (b) corrected and post-processed HLF (FDOM) and TLF wavelengths from 5 October 2021 to 31 July 2022, with data recorded at 15-minute intervals.

the start of river flow events, but values were quicker to decline after the initial event, suggesting exhaustion of tryptophan-like or proteinaceous DOM sources after continuous rainfall (Croghan *et al*., 2021).

4.4 Disinfectant By-product Formation from Dual Wavelength Fluorimetry

The corrected Dripsey River and Bunsheelin River EEM datasets containing surface water samples used for DBP formation potential earlier in the project were processed to extract HLF (λ_{ex} 280 nm, λ_{em} 450 ± 25 nm) and TLF (λ_{ex} 280 nm, λ_{em} 365 ± 25 nm) fluorescence intensities, as measured by the VLux sensor (Figure 2.3). These data were used to re-calibrate a machine learning model, as reported in Chapter 2. Only DBP species containing few samples (< 10%) lower than the LQ were evaluated and only the DBP parameters exhibiting a satisfactory relationship between observed and predicted ($R^2 \geq 0.7$) concentrations were retained. The VLux DBP machine learning model was repeated 100 times using three different techniques, namely BAGs, GBMs and NNETs, with 300 independent models used to calculate model performance and prediction parameters. The machine learning model was calibrated as described in Section 2.2, with the one-way analysis of covariance showing no statistically significant differences between validation and training datasets ($p > 0.05$). The average model performance for the HLF and TLF channels is presented in Table 4.2. Overall, for the DOC range investigated in the surface water EEM samples (0–15 mg C L^{-1}), the VLux wavelength pairs showed reasonable accuracy for predicting HAAs ($R^2 = 0.77$–0.80) and THM4 ($R^2 = 0.75$) but low precision (average RMSPE 48.6%). These results show that dual-wavelength sensor approaches for predicting DBP formation may suffer from considerable uncertainty in reproducing DBP concentrations in comparison with full EEM spectra (Section 2.6).

Table 4.2. Average machine learning model performance for training and validation data using HLF and TLF wavelength pair intensity extracted from the surface water EEM dataset

Training data					Validation data			
DBP	Slope	Intercept	R^2	RMSPE (%)	Slope	Intercept	R^2	RMSPE (%)
TCM	1.42±0.06	–58.3±17.3	0.75	50.9	1.35±0.07	–42.4±20.9	0.65	61.8
THM4	1.38±0.06	–67.1±18.9	0.75	42.9	1.33±0.07	–51.4±22.6	0.66	50.4
DCAA	1.22±0.05	–0.67±0.30	0.77	36.8	1.18±0.06	–0.44±0.36	0.68	45.2
TCAA	1.31±0.05	–0.95±0.46	0.80	46.0	1.30±0.06	–0.77±0.54	0.73	58.1
HAA5	1.28±0.05	–1.67±0.69	0.80	42.2	1.27±0.06	–1.41±0.83	0.73	51.7

5 Conclusions and Recommendations

5.1 Conclusions

This research project developed a methodology for determining the concentrations of DBPs formed for 20 individual DBP species, namely four THMs, nine HAAs, two HKs, four HANs and one HNM, for the first time in Ireland. These compounds include THM4 and HAA5, which are regulated chemical parameters in drinking water, as well as unregulated CDBPs and NDBPs, of which the latter may be more harmful to human health than the regulated compounds. The measured DBP concentrations represent the maximum concentrations obtained under an excess of chlorine disinfectant at 25°C and pH 7 for a contact time of 72 hours where pre-treatment prior to chlorination is limited to primary filtration only. The present study considered over 200 environmental surface water and groundwater samples, all containing ubiquitous environmental fluorophores (e.g. PARAFAC components C1 to C3) and a wide range of DOC (< LQ to 14.99 mg C L^{-1}), DIC (1.24 mg C L^{-1} to 62.2 mg C L^{-1}), NH_4-N (3 µg N L^{-1} to 2,359 µg N L^{-1}), DON (< LQ to 1.97 mg N L^{-1}) and Br^- (< LQ to 117 µg L^{-1}) concentrations, with associated CDBP and NDBP species spanning four orders of magnitude. The panel of environmental source waters considered is representative of the surface water quality at the sub-catchment scale under average to low flow conditions, where land cover ranges from plantation forestry to peatland and agricultural grassland as well as groundwater underlying intensive agricultural areas.

Through the use of machine learning techniques, it was possible to develop an ensemble predictive model for eight DBP species (TCM, BDCM, DBCM, DCAA, TCAA, DCAN, TCNM and TCP), the concentrations of which could be quantitatively predicted with high confidence (average $R^2 = 0.83$, RMSPE = 29.0%); the model was also able to accurately classify a further five species (TBM, TCAN, BCAN, DBAN and DCP) in terms of their presence or absence (average accuracy = 96.8%). This ensemble model framework provides a way forwards towards forecasting both THM4 and HAA5 concentrations from raw water quality at drinking water source intakes with a high degree of accuracy and precision using both EEM fluorescence spectra and common wet chemical parameters. A DOM optical property-only model (spectral model) was developed without the need for additional wet chemical parameters. With the benefit of considering 20 individual DBP species, the spectral model may assist in achieving the goal of protecting human health through continuous online and near real-time applications in areas where the concentrations of at least five regulated and unregulated DBP classes (i.e. THMs, HAAs, HANs, HNMs and HKs) can be estimated with reasonable accuracy for the first time, greatly surpassing the capability of traditional linear regression models, which rely solely on A_{254} as a surrogate for DOC.

A nested sub-basin approach was used to evaluate the influence of land cover on the formation of DBPs and their potential precursor sources in comparison with groundwater samples from the lower reaches of the Dripsey River. The 51 km^2 study area, comprising the Bunsheelin and Dripsey River sub-catchments of the River Lee, was divided into sub-basins containing > 50% of one landcover class, including grassland on mineral soils (57.6%) followed by grassland on PSs (16.1%) and tree cover (14.7%). For the tree cover category, comprising samples from the Dripsey uplands (DY01–DY03, DY07), some overlap with peatland or organic soils was noted. This landcover category was found to have the highest formation concentrations of THM4 (535 ± 479 µg L^{-1}) and HAA (23.6 ± 22.44 µg L^{-1}), significantly exceeding those of any other land cover and groundwater. Hence, upland conifer plantation forestry (on mostly organic soils) can be considered the land cover associated with the highest risk of exceedances, based on the results from this study. Conversely, groundwater samples had far lower formation concentrations for THM4 (92.6 ± 102.8 µg L^{-1}) and HAA5 (6.42 ± 17.21 µg L^{-1}), which were between two and six times lower than the highest surface water concentrations. These findings suggest that the replacement of surface water sources subject to frequent THM4 (and potentially HAA5) exceedances with local alternative groundwater sources may be a cost-effective solution, where major upgrades to surface water treatment plants are cost prohibitive. However, groundwater underlying

intensive agricultural regions may be a source of unregulated and potentially more harmful nitrogen-containing HANs, with groundwater DOC being particularly reactive in the formation of these NDBPs in the presence of nitrogen precursors (NO_3-N and PARAFAC component C3). A subset of surface water samples were subject to LC-OCD-OND analysis for molecular fractionation of DOM, which confirmed that tree cover samples contained the greatest fraction of hydrophobic organic matter, likely to play an important role in THM and HAA formation.

This study has provided a critical appraisal of modern *in situ* fluorescence sensor technology (VLux sensor) using two wavelength regions for HLF (λ_{ex} 280 nm, λ_{em} 450 ± 25 nm) and TLF (λ_{ex} 280 nm, λ_{em} 365 ± 25 nm) using an LED excitation source. The sensor system automatically compensates for matrix effects (inner filter effect, turbidity and temperature), which affect fluorescence signal integrity under turbid or highly coloured water conditions. The sensor was found to have generally good performance in comparison with benchtop EEM measurements over the same wavelength regions under laboratory conditions. The findings from a 10-month *in situ* sensor deployment at the outlet of the Dripsey River study area (DY12) provided insight into the high-frequency temporal dynamics of HLF and TLF, with extreme humic-like DOM export events being identified during peak river flows. Using the machine learning approach developed earlier in the project, the VLux wavelength pairs showed reasonable accuracy for predicting HAA5 (R^2 = 0.77–0.80) and THM4 (R^2 = 0.75) but relatively low precision (average error 48.6%) under average or low flow conditions. The findings demonstrate that dual-wavelength sensor approaches for predicting DBP formation are associated with more uncertainty in reproducing DBP concentrations for DOM admixtures than a collection of full EEM spectra and the application of PARAFAC to identify the independent underlying fluorophores which are involved in DBP formation.

5.2 Recommendations

- The machine learning models presented in the present study identify key DOM optical properties and hydrochemical parameters (namely DOC, DIC, DON, Br^-, NH_4, A_{254}, HIX and PARAFAC components C1–C3), required to accurately predict up to eight individual DBP species of which THM4 and HAA5 are relevant for the implementation of the recast EU Drinking Water Directive (Directive (EU) 2020/2184) in Ireland. Monitoring programmes at water treatment plants should include analysis of these parameters at different stages of the treatment process chain to identify selective removal/persistence of key precursors (e.g. PARAFAC components C1 and C2) during water treatment. TOX concentrations should be measured in treated water samples to quantify the total concentrations of halogenated DBPs produced from chlorination, of which many non-target compounds have yet to be described in the literature.
- Spatiotemporal sampling programmes in drinking water catchments should include the same DOM and hydrochemical parameters as the present study. The predictive modelling framework developed here may be used to inform preventative source protection measures aimed at reducing DBP precursor loads in raw water prior to entering water treatment plants. Consideration should be given to coastal catchments where elevated Br^- concentrations may produce a higher proportion of brominated DBPs, which may be more hazardous to human health than chlorinated DBPs.
- Further research involving online or near real-time FDOM monitoring should include the application of new fluorescence technology such as absorbance–transmittance EEM spectroscopy, through which both fluorescence and absorbance measurements are determined simultaneously on the same sample. With this approach, DOC concentrations can be automatically predicted from A_{254} measurements on the same sample, and the spectral model developed in this study may be suitable for quantitative predictions of both THM4 and HAA5 concentrations. The automation of raw water sampling, filtration and spectral acquisition should also be explored.
- The present study investigated predominantly humic DOM sources (e.g. PARAFAC components C1 and C2). Algal/microbial biomass and urban, domestic and agricultural wastewater point source inputs were not significant in the study areas. Future DBP formation research should consider these sources, as they are often important drivers of water body status deterioration in Ireland

as well as being potential sources of NDBP precursors, as shown elsewhere in the literature.

- Further sampling and DBP formation potential studies on groundwater sources used for drinking water are recommended. The present study included only a limited number of groundwater samples from a single aquifer type (Old Red Sandstone). Further DBP formation studies (including TOX analysis) are needed on different aquifer flow regimes and lithologies that may be prone to DOM mobilisation, such as karst, to verify that groundwater sources represent a safer alternative than surface water for the protection of human health from DBP exceedances.

References

Bond, T., Huang, J., Templeton, M.R. and Graham, N., 2011. Occurrence and control of nitrogenous disinfection by-products in drinking water – a review. *Water Research*, 45, 4341–4354.

Bond, T., Templeton, M.R., Kamal, N.H.M., Graham, N. and Kanda, R., 2015. Nitrogenous disinfection byproducts in English drinking water supply systems: occurrence, bromine substitution and correlation analysis. *Water Research*, 85, 85–94.

Brodfuehrer, S.H., Wahman, D.G., Alsulaili, A., Speitel Jr, G.E. and Katz, L.E., 2020. Role of carbonate species on general acid catalysis of bromide oxidation by hypochlorous acid (HOCl) and oxidation by molecular chlorine (Cl_2). *Environmental Science & Technology*, 54, 16186–16194.

Coble, P.G., 1996. Characterization of marine and terrestrial DOM in seawater using excitation–emission matrix spectroscopy. *Marine Chemistry*, 51, 325–346.

Connolly, J. and Holden, N.M., 2009. Mapping peat soils in Ireland: updating the derived Irish peat map. *Irish Geography*, 42, 343–352.

Copernicus Land Monitoring Service, 2016. *European Digital Elevation Model (EU-DEM v1.1)*. Available online: https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1 (accessed 20 January 2023).

Cork City Council (2023). *Public Drinking Water*. Available online: https://www.corkcoco.ie/en/resident/environment/public-drinking-water (accessed 7 July 2023).

Croghan, D., Khamis, K., Bradley, C., Van Loon, A.F., Sadler, J. and Hannah, D.M., 2021. Combining *in-situ* fluorometry and distributed rainfall data provides new insights into natural organic matter transport dynamics in an urban river. *Science of the Total Environment*, 755, 142731.

Dupas, R., Mellander, P.E., Gascuel-Odoux, C., Fovet, O., McAleer, E.B., McDonald, N.T., Shore, M. and Jordan, P., 2017. The role of mobilisation and delivery processes on contrasting dissolved nitrogen and phosphorus exports in groundwater fed catchments. *Science of the Total Environment*, 599, 1275–1287.

EPA (Environmental Protection Agency), 2017. *Water Flow Network*. Available online: https://gis.epa.ie/geonetwork/srv/eng/catalog.search#/metadata/c4043e19-38ec-4120-a588-8cd01ac94a9c (accessed14 June 2023).

EPA (Environmental Protection Agency), 2020. *Site Visit Report – SV20448 – Irish Water*. EPA, Johnstown Castle, Ireland.

EPA (Environmental Protection Agency), 2021. *Ireland's National Water Framework Directive Monitoring Programme 2019–2021*. EPA, Johnstown Castle, Ireland.

EPA (Environmental Protection Agency), 2023a. *Drinking Water Quality in Public Supplies 2022*. EPA, Johnstown Castle, Ireland.

EPA (Environmental Protection Agency), 2023b. *Drinking Water Quality in Private Group Schemes and Small Private Supplies 2022*. EPA, Johnstown Castle, Ireland.

EU (European Union), 2020. Directive (EU) 2020/2184 of the European Parliament and of the Council of 16 December 2020 on the quality of water intended for human consumption (recast). OJ L 435, 23.12.2020, p. 1−62.

Evlampidou, I., Font-Ribera, L., Rojas-Rueda, D., Gracia-Lavedan, E., Costet, N., Pearce, N., Vineis, P., Jaakkola, J.J., Delloye, F., Makris, K.C. and Stephanou, E.G., 2020. Trihalomethanes in drinking water and bladder cancer burden in the European Union. *Environmental Health Perspectives*, 128, 017001.

Fealy, R. and Stuart, G. 2009. *Teagasc-EPA Soils and Subsoils Mapping Project: Final report*. EPA, Johnstown Castle, Ireland.

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, 1997–2008.

Fox, B.G., Thorn, R.M.S., Dutta, T.K., Bowes, M.J., Read, D.S. and Reynolds, D.M., 2022. A case study: the deployment of a novel *in situ* fluorimeter for monitoring biological contamination within the urban surface waters of Kolkata, India. *Science of the Total Environment*, 842, 156848.

Francis, R.A., Van Briesen, J.M. and Small, M.J., 2010. Bayesian statistical modeling of disinfection byproduct (DBP) bromine incorporation in the ICR database. *Environmental Science & Technology*, 44, 1232–1239.

Golea, D.M., Upton, A., Jarvis, P., Moore, G., Sutherland, S., Parsons, S.A. and Judd, S.J., 2017. THM and HAA formation from NOM in raw and treated surface waters. *Water Research*, 112, 226–235.

Goslan, E.H., Seigle, C., Purcell, D., Henderson, R., Parsons, S.A., Jefferson, B. and Judd, S.J., 2017. Carbonaceous and nitrogenous disinfection by-product formation from algal organic matter. *Chemosphere*, 170, 1–9.

Gough, R., Holliman, P.J., Willis, N. and Freeman, C., 2014. Dissolved organic carbon and trihalomethane precursor removal at a UK upland water treatment works. *Science of the Total Environment*, 468, 228–239.

Graeber, D., Gelbrecht, J., Pusch, M.T., Anlanger, C. and von Schiller, D., 2012. Agriculture has changed the amount and composition of dissolved organic matter in Central European headwater streams. *Science of the Total Environment*, 438, 435–446.

Harrison, S., McAree, C., Mulville, W. and Sullivan, T., 2019. The problem of agricultural "diffuse" pollution: getting to the point. *Science of the Total Environment*, 677, 700–717.

Heinz, M., Graeber, D., Zak, D., Zwirnmann, E., Gelbrecht, J. and Pusch, M.T., 2015. Comparison of organic matter composition in agricultural versus forest affected headwaters with special emphasis on organic nitrogen. *Environmental Science & Technology*, 49, 2081–2090.

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 Research*, 45, 879–885.

Kalbitz, K. and Geyer, S., 2002. Different effects of peat degradation on dissolved organic carbon and nitrogen. *Organic Geochemistry*, 33, 19–32.

Krzeminski, P., Vogelsang, C., Meyn, T., Köhler, S.J., Poutanen, H., de Wit, H.A. and Uhl, W., 2019. Natural organic matter fractions and their removal in full-scale drinking water treatment under cold climate conditions in Nordic capitals. *Journal of Environmental Management*, 241, 427–438.

Lawaetz, A.J. and Stedmon, C.A., 2009. Fluorescence intensity calibration using the Raman scatter peak of water. *Applied Spectroscopy*, 63, 936–940.

Lee, E.J., Yoo, G.Y., Jeong, Y., Kim, K.U., Park, J.H. and Oh, N.H., 2015. Comparison of UV–Vis and FDOM sensors for *in situ* monitoring of stream DOC concentrations. *Biogeosciences*, 12, 3109–3118.

Li, X.F. and Mitch, W.A., 2018. Drinking water disinfection byproducts (DBPs) and human health effects: multidisciplinary challenges and opportunities. *Environmental Science & Technology*, 52, 1681–1689.

Marcé, R., Verdura, L. and Leung, N., 2021. Dissolved organic matter spectroscopy reveals a hot spot of organic matter changes at the river–reservoir boundary. *Aquatic Sciences*, 83, 67.

Met Éireann, 2023. *Historical Climate Data*. Available online: https://www.met.ie/climate/available-data/historical-data (accessed 12 June 2023).

Moe, H., Craig, M. and Daly, D. 2010. *Poorly productive aquifers: monitoring installations and conceptual understanding*. CDM Smith and EPA, Dublin.

Murphy, K.R., Butler, K.D., Spencer, R.G., Stedmon, C.A., Boehme, J.R. and Aiken, G.R., 2010. Measurement of dissolved organic matter fluorescence in aquatic environments: an interlaboratory comparison. *Environmental Science & Technology*, 44, 9405–9412.

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

O'Driscoll, C., Xiao, L., Zhan, X., Misstear, B. and Pilla, F., 2017. *Assessment of Natural Organic Matter (NOM) and Ptaquiloside in Irish Waters*. EPA, Johnstown Castle, Ireland.

O'Driscoll, C., Sheahan, J., Renou-Wilson, F., Croot, P., Pilla, F., Misstear, B. and Xiao, L., 2018a. National scale assessment of total trihalomethanes in Irish drinking water. *Journal of Environmental Management*, 212, 131–141.

O'Driscoll, C., Ledesma, J.L., Coll, J., Murnane, J.G., Nolan, P., Mockler, E.M., Futter, M.N. and Xiao, L.W., 2018b. Minimal climate change impacts on natural organic matter forecasted for a potable water supply in Ireland. *Science of the Total Environment*, 630, 869–877.

O'Driscoll, C., McGillicuddy, E., Croot, P., Bartley, P., McMyler, J., Sheahan, J. and Morrison, L., 2020. Tracing sources of natural organic matter, trihalomethanes and metals in groundwater from a karst region. *Environmental Science and Pollution Research*, 27, 12587–12600.

Peleato, N.M., McKie, M., Taylor-Edmonds, L., Andrews, S.A., Legge, R.L. and Andrews, R.C., 2016. Fluorescence spectroscopy for monitoring reduction of natural organic matter and halogenated furanone precursors by biofiltration. *Chemosphere*, 153, 155–161.

Plewa, M.J., Kargalioglu, Y., Vankerk, D., Minear, R.A. and Wagner, E.D., 2002. Mammalian cell cytotoxicity and genotoxicity analysis of drinking water disinfection by-products. *Environmental and Molecular Mutagenesis*, 40, 134–142.

Pschenyckyj, C., Donahue, T., Kelly-Quinn, M., O'Driscoll, C. and Renou-Wilson, F., 2023. An examination of the influence of drained peatlands on regional stream water chemistry. *Hydrobiologia*, 1–27.

Pucher, M., Wünsch, U., Weigelhofer, G., Murphy, K., Hein, T. and Graeber, D., 2019. staRdom: versatile software for analyzing spectroscopic data of dissolved organic matter in R. *Water*, 11, 2366.

Richardson, S.D. and Kimura, S.Y., 2019. Water analysis: emerging contaminants and current issues. *Analytical Chemistry*, 92, 473–505.

Rolston, A. and Linnane, S., 2020. Drinking water source protection for surface water abstractions: an overview of the group water scheme sector in the Republic of Ireland. *Water*, 12, 2437.

Sadiq, R. and Rodriguez, M.J., 2004. Disinfection by-products (DBPs) in drinking water and predictive models for their occurrence: a review. *Science of the Total Environment*, 321, 21–46.

Scanlon, T.M., Kiely, G. and Xie, Q., 2004. A nested catchment approach for defining the hydrological controls on non-point phosphorus transport. *Journal of Hydrology*, 291, 218–231.

Shtein, A., Kloog, I., Schwartz, J., Silibello, C., Michelozzi, P., Gariazzo, C., Viegi, G., Forastiere, F., Karnieli, A., Just, A.C. and Stafoggia, M., 2019. Estimating daily $PM_{2.5}$ and PM_{10} over Italy using an ensemble model. *Environmental Science & Technology*, 54, 120–128.

Shutova, Y., Baker, A., Bridgeman, J. and Henderson, R.K., 2014. Spectroscopic characterisation of dissolved organic matter changes in drinking water treatment: from PARAFAC analysis to online monitoring wavelengths. *Water Research*, 54, 159–169.

Stalter, D., O'Malley, E., Von Gunten, U. and Escher, B.I., 2016. Fingerprinting the reactive toxicity pathways of 50 drinking water disinfection by-products. *Water Research*, 91, 19–30.

Stanley, E.H. and Maxted, J.T., 2008. Changes in the dissolved nitrogen pool across land cover gradients in Wisconsin streams. *Ecological Applications*, 18, 1579–1590.

Teagasc and Cranfield University, 2014. *Irish National Soils Map, 1:250,000k, V1b*. Available online: http://gis.teagasc.ie/soils/downloads.php (accessed 7 July 2023).

USEPA (United States Environmental Protection Agency), 2016. *Definition and Procedure for the Determination of the Method Detection Limit Method, Rev. 2*. USEPA, Washington, DC, USA.

Vabalas, A., Gowen, E., Poliakoff, E. and Casson, A.J., 2019. Machine learning algorithm validation with a limited sample size. *PloS One*, 14, 0224365.

Van der Kwast, H. and Menke, K., 2022. *QGIS for Hydrological Applications: Recipes for Catchment Hydrology and Water Management*. Locate Press, Chugiak, AK, USA.

Xiao, Q., Chang, H.H., Geng, G. and Liu, Y., 2018. An ensemble machine-learning model to predict historical $PM_{2.5}$ concentrations in China from satellite data. *Environmental Science & Technology*, 52, 13260–13269.

Yamashita, Y., Kloeppel, B.D., Knoepp, J., Zausen, G.L. and Jaffé, R., 2011. Effects of watershed history on dissolved organic matter characteristics in headwater streams. *Ecosystems*, 14, 1110–1122.

Yang, Y., Komaki, Y., Kimura, S.Y., Hu, H.Y., Wagner, E.D., Mariñas, B.J. and Plewa, M.J., 2014. Toxic impact of bromide and iodide on drinking water disinfected with chlorine or chloramines. *Environmental Science & Technology*, 48, 12362–12369.

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S. and Lesiv, M., 2022. *ESA WorldCover 10 m 2021 v200*. https://doi.org/10.5281/zenodo.7254221.

Zhang, Y. and Ling, C., 2018. A strategy to apply machine learning to small datasets in materials science. *npj Computational Materials*, 4, 25.

Abbreviations

A_{254}	Ultraviolet absorbance at 254 nm
A_{280}	Ultraviolet absorbance at 280 nm
ACS	American Chemical Society
BAG	Bagging tree
BB	Building block
BCAN	Bromochloroacetonitrile
BDCM	Bromodichloromethane
CDBP	Carbonaceous disinfection by-product
CV	Coefficient of variation
DBAN	Dibromoacetonitrile
DBCM	Dibromochloromethane
DBP	Disinfection by-product
DCAA	Dichloroacetic acid
DCAN	Dichloroacetonitrile
DCP	Dichloropropanone
DIC	Dissolved inorganic carbon
DOC	Dissolved organic carbon
DOM	Dissolved organic matter
DON	Dissolved organic nitrogen
EEM	Excitation–emission matrix
EEM-PARAFAC	Excitation–emission matrix-parallel factor analysis
F_{max}	Maximum fluorescence intensity
FDOM	Fluorescent dissolved organic matter
GBM	Generalised boosted regression model
HAA	Haloacetic acid
HAA5	Sum of five haloacetic acids
HAN	Haloacetonitrile
HAN4	Sum of four haloacetonitriles
HIX	Humification index
HK	Haloketone
HK2	Sum of two haloketones
HLF	Humic-like fluorescence
HMWS	Non-humic high molecular weight substance
HNM	Halonitromethane
HOC	Hydrophobic organic carbon
HOCl	Hypochlorous acid
HS	Humic substance
LC-OCD-OND	Liquid chromatography–organic carbon detection–organic nitrogen detection
LMWS	Low molecular weight substance
LQ	Limit of quantification
MDL	Method detection limit
NA	N-nitrosamine
NDBP	Nitrogenous disinfection by-product
NNET	Neural network
NOM	Natural organic matter

PARAFAC	Parallel factor analysis
PS	Peatland soil
PTFE	Polytetrafluoroethylene
QSU	Quinine sulfate unit
R^2	Coefficient of determination
RMSPE	Root mean squared percentage error
RU	Raman unit
SA	Sensitivity analysis
SRFA	Suwannee River Fulvic Acid
SRHA	Suwannee River Humic Acid
SVM	Support vector machine
TBM	Tribromomethane (bromoform)
TCAA	Trichloroacetic acid
TCAN	Trichloroacetonitrile
TCC	Tucker's congruence coefficient
TCM	Trichloromethane (chloroform)
TCNM	Trichloronitromethane
TCP	Trichloropropanone
TDN	Total dissolved nitrogen
THM	Trihalomethane
THM4	Sum of four trihalomethanes
TLF	Tryptophan-like fluorescence
TOC	Total organic carbon
TOX	Total organic halogen
USEPA	United States Environmental Protection Agency
UV-vis	Ultraviolet-visible
λ_{em}	Emission wavelength
λ_{ex}	Excitation wavelength

An Ghníomhaireacht Um Chaomhnú Comhshaoil

Tá an GCC freagrach as an gcomhshaol a chosaint agus a fheabhsú, mar shócmhainn luachmhar do mhuintir na hÉireann. Táimid tiomanta do dhaoine agus don chomhshaol a chosaint ar thionchar díobhálach na radaíochta agus an truaillithe.

Is féidir obair na Gníomhaireachta a roinnt ina trí phríomhréimse:

Rialáil: Rialáil agus córais chomhlíonta comhshaoil éifeachtacha a chur i bhfeidhm, chun dea-thorthaí comhshaoil a bhaint amach agus díriú orthu siúd nach mbíonn ag cloí leo.

Eolas: Sonraí, eolas agus measúnú ardchaighdeáin, spriocdhírithe agus tráthúil a chur ar fáil i leith an chomhshaoil chun bonn eolais a chur faoin gcinnteoireacht.

Abhcóideacht: Ag obair le daoine eile ar son timpeallachta glaine, táirgiúla agus dea-chosanta agus ar son cleachtas inbhuanaithe i dtaobh an chomhshaoil.

I measc ár gcuid freagrachtaí tá:

Ceadúnú

- Gníomhaíochtaí tionscail, dramhaíola agus stórála peitril ar scála mór;
- Sceitheadh fuíolluisce uirbigh;
- Úsáid shrianta agus scaoileadh rialaithe Orgánach Géinmhodhnaithe;
- Foinsí radaíochta ianúcháin;
- Astaíochtaí gás ceaptha teasa ó thionscal agus ón eitlíocht trí Scéim an AE um Thrádáil Astaíochtaí.

Forfheidhmiú Náisiúnta i leith Cúrsaí Comhshaoil

- Iniúchadh agus cigireacht ar shaoráidí a bhfuil ceadúnas acu ón GCC;
- Cur i bhfeidhm an dea-chleachtais a stiúradh i ngníomhaíochtaí agus i saoráidí rialáilte;
- Maoirseacht a dhéanamh ar fhreagrachtaí an údaráis áitiúil as cosaint an chomhshaoil;
- Caighdeán an uisce óil phoiblí a rialáil agus údaruithe um sceitheadh fuíolluisce uirbigh a fhorfheidhmiú
- Caighdeán an uisce óil phoiblí agus phríobháidigh a mheasúnú agus tuairisciú air;
- Comhordú a dhéanamh ar líonra d'eagraíochtaí seirbhíse poiblí chun tacú le gníomhú i gcoinne coireachta comhshaoil;
- An dlí a chur orthu siúd a bhriseann dlí an chomhshaoil agus a dhéanann dochar don chomhshaol.

Bainistíocht Dramhaíola agus Ceimiceáin sa Chomhshaol

- Rialacháin dramhaíola a chur i bhfeidhm agus a fhorfheidhmiú lena n-áirítear saincheisteanna forfheidhmithe náisiúnta;
- Staitisticí dramhaíola náisiúnta a ullmhú agus a fhoilsiú chomh maith leis an bPlean Náisiúnta um Bainistíocht Dramhaíola Guaisí;
- An Clár Náisiúnta um Chosc Dramhaíola a fhorbairt agus a chur i bhfeidhm;
- Reachtaíocht ar rialú ceimiceán sa timpeallacht a chur i bhfeidhm agus tuairisciú ar an reachtaíocht sin.

Bainistíocht Uisce

- Plé le struchtúir náisiúnta agus réigiúnacha rialachais agus oibriúcháin chun an Chreat-treoir Uisce a chur i bhfeidhm;
- Monatóireacht, measúnú agus tuairisciú a dhéanamh ar chaighdeán aibhneacha, lochanna, uiscí idirchreasa agus cósta, uiscí snámha agus screamhuisce chomh maith le tomhas ar leibhéil uisce agus sreabhadh abhann.

Eolaíocht Aeráide & Athrú Aeráide

- Fardail agus réamh-mheastacháin a fhoilsiú um astaíochtaí gás ceaptha teasa na hÉireann;
- Rúnaíocht a chur ar fáil don Chomhairle Chomhairleach ar Athrú Aeráide agus tacaíocht a thabhairt don Idirphlé Náisiúnta ar Ghníomhú ar son na hAeráide;
- Tacú le gníomhaíochtaí forbartha Náisiúnta, AE agus NA um Eolaíocht agus Beartas Aeráide.

Monatóireacht & Measúnú ar an gComhshaol

- Córais náisiúnta um monatóireacht an chomhshaoil a cheapadh agus a chur i bhfeidhm: teicneolaíocht, bainistíocht sonraí, anailís agus réamhaisnéisiú;
- Tuairiscí ar Staid Thimpeallacht na hÉireann agus ar Tháscairí a chur ar fáil;
- Monatóireacht a dhéanamh ar chaighdeán an aeir agus Treoir an AE i leith Aeir Ghlain don Eoraip a chur i bhfeidhm chomh maith leis an gCoinbhinsiún ar Aerthruailliú Fadraoin Trasteorann, agus an Treoir i leith na Teorann Náisiúnta Astaíochtaí;
- Maoirseacht a dhéanamh ar chur i bhfeidhm na Treorach i leith Torainn Timpeallachta;
- Measúnú a dhéanamh ar thionchar pleananna agus clár beartaithe ar chomhshaol na hÉireann.

Taighde agus Forbairt Comhshaoil

- Comhordú a dhéanamh ar ghníomhaíochtaí taighde comhshaoil agus iad a mhaoiniú chun brú a aithint, bonn eolais a chur faoin mbeartas agus réitigh a chur ar fáil;
- Comhoibriú le gníomhaíocht náisiúnta agus AE um thaighde comhshaoil.

Cosaint Raideolaíoch

- Monatóireacht a dhéanamh ar leibhéil radaíochta agus nochtadh an phobail do radaíocht ianúcháin agus do réimsí leictreamaighnéadacha a mheas;
- Cabhrú le pleananna náisiúnta a fhorbairt le haghaidh éigeandálaí ag eascairt as taismí núicléacha;
- Monatóireacht a dhéanamh ar fhorbairtí thar lear a bhaineann le saoráidí núicléacha agus leis an tsábháilteacht raideolaíochta;
- Sainseirbhísí um chosaint ar an radaíocht a sholáthar, nó maoirsiú a dhéanamh ar sholáthar na seirbhísí sin.

Treoir, Ardú Feasachta agus Faisnéis Inrochtana

- Tuairisciú, comhairle agus treoir neamhspleách, fianaise-bhunaithe a chur ar fáil don Rialtas, don tionscal agus don phobal ar ábhair maidir le cosaint comhshaoil agus raideolaíoch;
- An nasc idir sláinte agus folláine, an geilleagar agus timpeallacht ghlan a chur chun cinn;
- Feasacht comhshaoil a chur chun cinn lena n-áirítear tacú le hiompraíocht um éifeachtúlacht acmhainní agus aistriú aeráide;
- Tástáil radóin a chur chun cinn i dtithe agus in ionaid oibre agus feabhsúchán a mholadh áit is gá.

Comhpháirtíocht agus Líonrú

- Oibriú le gníomhaireachtaí idirnáisiúnta agus náisiúnta, údaráis réigiúnacha agus áitiúla, eagraíochtaí neamhrialtais, comhlachtaí ionadaíocha agus ranna rialtais chun cosaint chomhshaoil agus raideolaíoch a chur ar fáil, chomh maith le taighde, comhordú agus cinnteoireacht bunaithe ar an eolaíocht.

Bainistíocht agus struchtúr na Gníomhaireachta um Chaomhnú Comhshaoil

Tá an GCC á bainistiú ag Bord lánaimseartha, ar a bhfuil Ard-Stiúrthóir agus cúigear Stiúrthóir. Déantar an obair ar fud cúig cinn d'Oifigí:

1. An Oifig um Inbhunaitheacht i leith Cúrsaí Comhshaoil
2. An Oifig Forfheidhmithe i leith Cúrsaí Comhshaoil
3. An Oifig um Fhianaise agus Measúnú
4. An Oifig um Chosaint ar Radaíocht agus Monatóireacht Comhshaoil
5. An Oifig Cumarsáide agus Seirbhísí Corparáideacha

Tugann coistí comhairleacha cabhair don Ghníomhaireacht agus tagann siad le chéile go rialta le plé a dhéanamh ar ábhair imní agus le comhairle a chur ar an mBord.

EPA Research

Webpages: www.epa.ie/our-services/research/
LinkedIn: www.linkedin.com/showcase/eparesearch/
Twitter: @EPAResearchNews
Email: research@epa.ie