Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Enhanced statistical evaluation of fluorescence properties to identify dissolved organic matter dynamics during river high-flow events

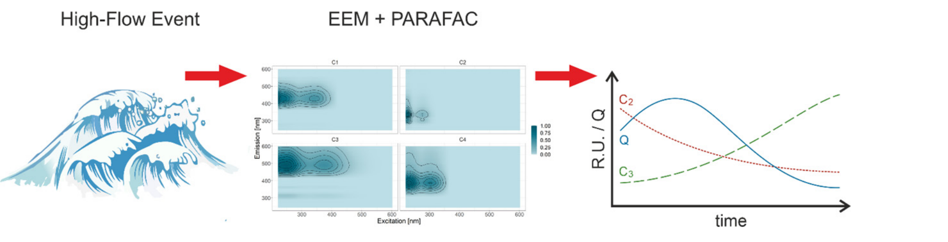
Sandra Peer [*], Anastassia Vybornova, Zdravka Saracevic, Jörg Krampe, Matthias Zessner, Ottavia Zoboli

Institute for Water Quality and Resource Management, TU Wien, Karlsplatz 13/226, 1040 Vienna, Austria

HIGHLIGHTS

- Independent DOM components were discriminated using PARAFAC.
- Varying composition of DOM compounds during high-flow events is revealed.
- sPLS yields specific wavelength pairs serving as proxy parameters.
- Dynamics of water quality parameters can be predicted via EEM and advanced statistics.
- Insight into dominance shifts between point and non-point inputs of organic matter.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Ashantha Goonetilleke

Keywords:
Fluorescence spectroscopy
Excitation-emission matrix
Diffuse emissions
Wastewater effluent
PARAFAC
Sparse PLS

ABSTRACT

Fluorescence spectroscopy has become a widely used technique to characterize dissolved organic matter (DOM) and organic hazardous micro-pollutants in natural and human-influenced water bodies. Especially in rivers highly impacted by municipal and industrial wastewater treatment plants' effluents, the fluorescence signal at low-flow is mainly dominated by these discharges. At river high-flow, their influence decreases due to dilution effects, and at the same time, other compounds of DOM, stemming from diffuse inputs, can increase or even dominate. Therefore, whereas the analysis of DOM is little informative on the changing sources and pathways of emissions, fluorescence spectroscopy can enhance our understanding and our possibilities of monitoring such dynamics in river catchments. This paper analyzed samples from seven high-flow events in an Austrian river. Firstly, independent DOM components were discriminated using a parallel factor analysis (PARAFAC) to show the varying composition of DOM during different phases of high-flow events. Furthermore, partial least squares (PLS) and sparse PLS (sPLS) regression were applied to identify excitation and emission wavelengths, serving as proxy parameters for quantifying dissolved organic carbon (DOC) and chloride. The PLS models show the best prediction accuracy but use the entire excitation-emission matrix in exchange. In selecting predictors, the use of excitation and emission wavelengths adjusted via sPLS is superior to the extracted PARAFAC components. The sPLS model yields 16 wavelength combinations for DOC ($RMSE_{sPLS}$ = 0.41 mg L^{-1}) and 18 wavelength combinations for chloride ($RMSE_{sPLS}$ = 2.21 mg L^{-1}). In contrast to other established optical measurement methods, which require different calibrations for low- and high-flow conditions, these models based on sPLS succeed in quantifying those parameters across the entire range of flow conditions and events of various magnitudes with a relative precision of about 5 %. These results show how the application of multivariate statistical techniques enhances the exploitation of the information provided by fluorescence spectroscopy.

1. Introduction

Fluorescence spectroscopy is an analytical method for water samples from aquatic systems, which is fast, highly sensitive, requires no reagents

[*] Corresponding author.
E-mail address: sandra.peer@tuwien.ac.at (S. Peer).

http://dx.doi.org/10.1016/j.scitotenv.2022.158016
Received 20 May 2022; Received in revised form 18 July 2022; Accepted 9 August 2022
Available online 13 August 2022

and no or very little sample preparation, and is an inexpensive option compared to other analytical methods (Carstea et al., 2016). Its current applications range from detection of contamination events at karst springs (Frank et al., 2017) to wastewater treatment monitoring (Cohen et al., 2014) and pollution source tracking (Cawley et al., 2012). In particular, fluorescence spectroscopy is applicable to characterize dissolved organic matter (DOM) content and organic hazardous micropollutant content of river water (Sgroi et al., 2017). DOM refers to a mixture of variable compositions of organic compounds of different origins. Therefore, if the position and shape of the peaks in the fluorescence fingerprint change, one may assume that this reflects a shift in the composition of the dissolved organic compounds. Especially for the resulting excitation-emission matrix (EEM), different methods have been proposed to define regions that can be primarily assigned to certain fluorescence compounds, e.g., peak picking (Coble, 1996), fluorescence regional integration (Chen et al., 2003), and self-organizing maps (Bieroza et al., 2009). Some recent studies attempt to describe the origin and dynamic of DOM during diverse flow conditions in different water bodies (Yamashita et al., 2008; Harjung et al., 2018). In addition, first approaches to describe the transport dynamics and mixing of DOM from river water, seawater, and effluent wastewater in coastal zones are also emerging (EL-Nahhal et al., 2020, 2021). Findings show that an increase in dissolved organic carbon (DOC) aromaticity accompanies an increasing flow and hence an increase in fluorescence intensity (Vidon et al., 2008; Carstea et al., 2010). It is even possible to observe shifts in fluorescence peaks during and between rainfall events indicating changes of DOM sources (Croghan et al., 2021) or to identify diel periodicity for some fluorophores (Khamis et al., 2020). These findings are predominantly based on in-field fluorescence spectroscopic measurements and thus rely on the limited information of selected wavelength combinations, such as the Coble peaks (Coble, 1996). The potential offered by the information of the whole EEM along with both multivariate and dimension-reducing statistical methods has not yet been fully exploited in this field. Therefore, if it is possible to sample high-flow events appropriately to measure the entire EEM, it might be possible to derive more useful information from it than has been obtained in previous studies with limited on-site methods (Khamis et al., 2020; Croghan et al., 2021).

Furthermore, there is evidence that parallel factor analysis (PARAFAC) (Bro, 1997), a multi-way decomposition method to simultaneously determine and quantify underlying independent fluorescent components (Murphy et al., 2013), is particularly useful in identifying DOM composition and dynamics with a focus on long-term seasonal effects (EL-Nahhal et al., 2021; Retelletti Brogi et al., 2020) or short-term changes during high-flow events (Hong et al., 2012; Austnes et al., 2010; Fellman et al., 2009). For instance, this enables to differentiate the counteracting shift mechanisms in the biodegradability and chemical quality of DOM during storm events in different watersheds (Fellman et al., 2009). Studies in this field have so far strongly focused on linking the shift of DOM composition during storm events to different inputs from terrestrial sources and land use in the river catchment (Nguyen et al., 2013; Yamashita et al., 2011), but no attempts have yet been made to explore changes between DOC from point and diffuse emissions thoroughly. Beyond that, the literature is somewhat limited to extracting PARAFAC components, classifying these as protein- or humic-like, and correlating them with DOC. Multivariate statistical methods could incorporate the comprehensive information offered by multi-parameter monitoring of water quality. Moreover, the interpretation of PARAFAC results is far from being straightforward. Until now, it has not been possible to attribute PARAFAC components to specific chemical compounds, although component scores are assumed to be correlated with actual concentrations (Stedmon and Bro, 2008; Wünsch et al., 2019). In this regard, a crucial aspect is that the extracted components are sensitive to the specific analyzed water body. Accordingly, a cautious choice in modeling is far more important than trying to establish one global model (Pitta and Zeri, 2021). Furthermore, to interpret the variation of PARAFAC components over space and time, the influence of the environment on the fluorescence properties of a compound must first be well-understood (Ishii and Boyer, 2012).

Given current limitations and research gaps, this study aims to investigate the full potential of fluorescence spectroscopy in supporting the identification and understanding of shifts and dynamics of DOM emitted via different point sources and diffuse pathways into rivers during high-flow events. The selected case study, namely a river significantly influenced by both wastewater treatment plant (WWTP) effluents and diffuse emissions, provides ideal conditions for this investigation. The emitters are expected to dominate the signal measured with fluorescence spectroscopy at low flow in such a river. Their emission loads get diluted at high flow, while a shift toward greater visibility, if not even dominance of diffuse emissions in the fluorescence signal, can be expected.

In this paper, the fluorescence properties of DOM during river high-flow events are considered from two newly related perspectives. The first objective is to discriminate PARAFAC components representing the peak shift and variation of DOM compounds during high-flow events. The novelty of this publication is that components are analyzed to show how the composition varies during different phases of highly dynamic flow events rather than just focusing on long-term or seasonal variations. Moreover, events of extremely various magnitudes are compared in terms of components, and similarities as well as differences are explained. The second objective is to establish components or excitation and emission wavelengths serving as proxy parameters for quantifying of water quality parameters such as DOC and chloride (Cl^-). Ordinary least squares (OLS) regression, partial least squares (PLS) regression, and, to the best of the authors' knowledge, for the very first time, sparse PLS (sPLS) regression are applied to address this question. The results are compared regarding their ability to produce reliable predictions and identify relevant combinations of excitation and emission wavelengths, which can serve as credible proxy parameters for quantifying water quality parameters.

2. Materials and methods

2.1. Sampling site and sampling strategy

All samples were taken at the online-monitoring station (N46°55′48″, E16°9′12″) of the project “Sustainable water quality management Rába - Online Monitoring” (NaWas) at the Austrian lowland river Rába (German: Raab). The river has its origin at the foot of the mountain Osser in the municipality of Passail (N47°20′43″, E15°30′55″). It enters the Mosoni Duna, a right-bank tributary of the Danube, in Győr (N47°41′25″, E17°37′49″), Hungary. The catchment of the Rába river, located in the southeast of Austria and characterized by average precipitation of 833 mm y^{-1}, covers an area of 1009 km^2 (Zoboli et al., 2019). It has a mixed land use, with approximately 42 % of the area dedicated to agriculture (25 % arable land and 17 % grassland) and 52 % and 3 % covered by forest and urban areas, respectively. Municipal and industrial wastewater treatment plants (especially three tanneries) contribute with their effluents to ca. 3 % of the total mean flow of the river, which is 9.9 m^3 s^{-1} (Zoboli et al., 2019). The purpose of the NaWas project is the long-term high-resolution monitoring of water quality for supervision of municipal and industrial discharges and the development and testing of new measurement techniques. The river water is constantly pumped into a flow-through tub located in a container at the river bank and equipped with numerous single- and multi-parameter probes for online monitoring, e.g., a turbidity immersion probe measuring total suspended solids (TSS) and a UV–Vis spectrometer measuring total organic carbon (TOC), DOC and nitrate (NO_3-N).

To specifically sample high-flow events in-situ, a commercially available portable sampling device (Bühler 2000) with 24 bottles of 1 L and a cooled sample compartment (∼4 °C) takes water samples out of the previously described tub. All components of the sampling device are routinely thoroughly cleaned and rinsed with sample medium before and after each triggering to prevent contamination. The sampling compartment is stocked with cleaned HDPE bottles approved for organic carbon analysis. In the laboratory, the samples are handled in glass bottles. To limit other confounding influences, the exact same sampling site and same methodology were consistently used throughout the study. The sampling device is triggered remotely either manually or automatically based on complex sampling algorithms derived from online signals to ensure the optimal coverage of increasing and decreasing parts of the hydrograph in real-time. In addition to standard time and volume-based triggering of the auto-sampler, the implemented measurement control software iTUWmon (Winkelbauer et al., 2014) enables the configuration of advanced control strategies. Flow and turbidity are typical parameters influencing the triggering operation. After fine-tuning of triggering conditions, the control strategy is activated, and sampling of recognized events, together with the measurement data of the examined river section, is logged into a central measurement database.

This setup allowed to aim for the sampling of different phases, i.e., increasing and decreasing river high-flow, as well as different magnitudes, i.e., medium-high to annual river high-flow, throughout the seasonal variations of one year. Event A was captured using a time-proportional algorithm. In contrast, a customized algorithm captured events B through D and F based on turbidity and river flow. Events E and G were manually sampled by triggering the sampling device remotely. Fig. 1 shows the river flow of the seven analyzed events and the sampling period within each event. Not all events could be covered entirely and the length of the sampling periods varied noticeably due to the limited number of bottles in the sampling device, sedimentation in the flow-through tub, or server connection failure. Nevertheless, the total of 74 samples covers a wide variety of different configurations of high-flow events. Table 1 specifies the sampling periods, the number of samples per event, river flow and DOC concentration ranges in more detail.

Table 1

Overview of sampling periods, duration, number of samples (n) and, range of discharge and DOC for each river high-flow event.

Event	Start (UTC)	Duration [h]	n	Q [m^{3}s^{-1}]	DOC [mg L^{-1}]
A	2020-06-29 14:30	3.5	8	11.8–127	4.8–7.9
B	2020-07-11 21:55	20.5	14	5.7–32.6	3.5–7.4
C	2020-10-12 02:00	34	14	21.8–92.1	5.6–9.4
D	2020-12-09 09:20	20	10	10–29.1	5.1–6.8
E	2021-01-03 13:20	163.5	10	6–19.6	2.5–3.4
F	2021-02-07 17:30	63.5	9	3.3–24.5	2.7–4.8
G	2021-05-17 05:50	21	9	4.7–18.5	3.5–4

2.2. Analytical methods

Standard water quality analysis of the samples comprised DOC (DIN EN 1484), Cl$^-$ (HPIC according to DIN EN ISO 10304-1), PO_4-P (DIN EN ISO 6878), NH_4-N (DIN EN ISO 11732), and TSS concentrations (DIN 38409-2).

Throughout each event, the samples were cooled in the sampling compartment, retrieved in a timely manner, including in-between collection at more extended events, and filtered with a 0.45 membrane filter upon arrival at the lab. Immediately before the spectroscopic measurement, samples were diluted four-fold with Milli-Q and allowed to warm up to room temperature (~20 °C). Excitation-emission matrices (EEM) have been recorded using the HORIBA Scientific Aqualog® spectrofluorometer equipped with a Xenon lamp. The measurements were performed in a

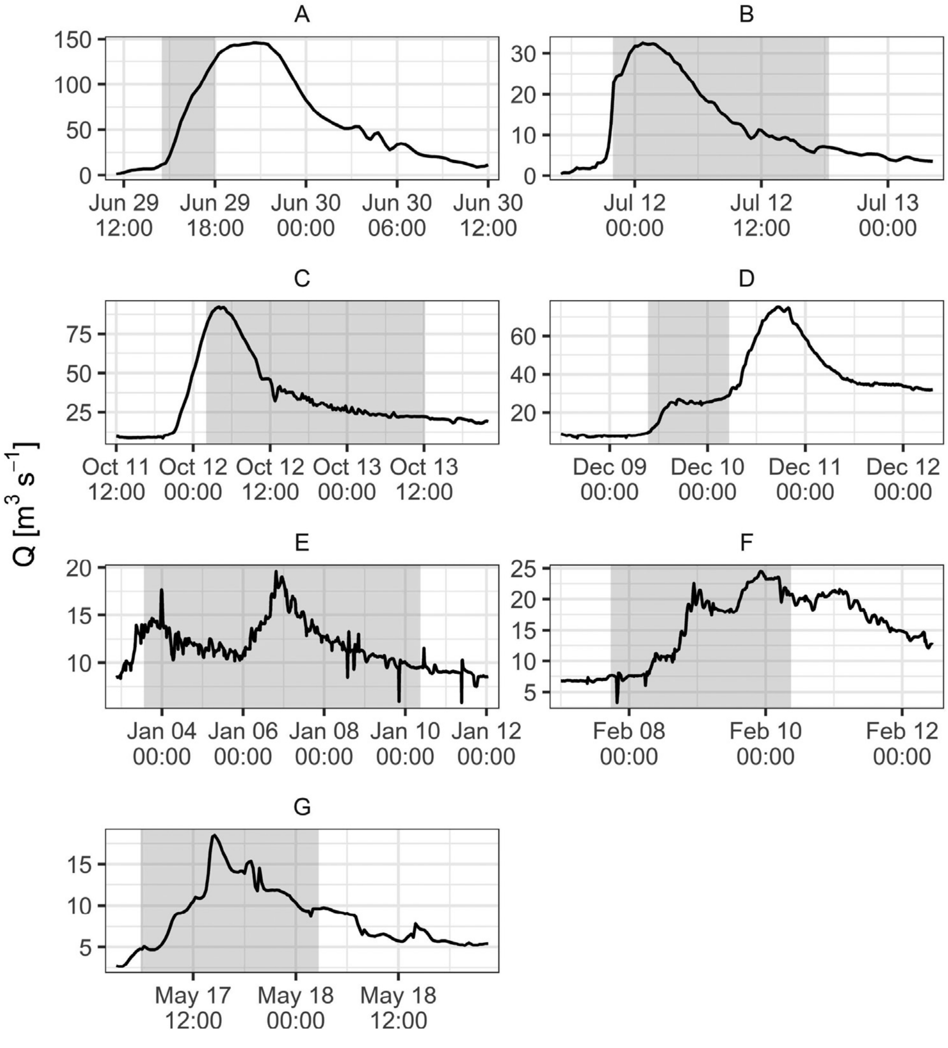


Fig. 1. River-flow time series of analyzed high-flow events A to G with sampling periods highlighted in light-gray.

quartz cell with a 1 cm optical path length for an excitation range of 220 nm 600 nm in 3 nm steps, an emission range of 246 nm 824 nm, a slit width of 5 nm and an integration time of 2 s. Absorbance for each sample has been measured with the same equipment and measurement settings for the same wavelengths as for the fluorescence spectroscopy. Those absorbance spectra were used to correct for inner filter effects following the method proposed by Lakowicz (2006).

Several correction steps have been applied to the raw EEM data. Subtraction of dark signal, spectral correction, scaling to reference detector, and blank subtraction was done inside the Aqualog software. For blank subtraction, EEM of Milli-Q have been measured on the same day as the sample EEM. The results were then exported and further corrected with the statistical software package R (R Core Team, 2021). Customized functions adapted from the packages *eemR* (Massicotte, 2019) and *staRdom* (Pucher et al., 2019) were applied to import the data and to correct for Rayleigh-masking, Raman-masking and inner filter effects. Furthermore, data were numerically corrected to represent undiluted samples at an integration time of 1 s. Finally, fluorescence intensities of the corrected EEM were converted into Raman Units [R.U.] by normalizing by the daily Raman peak area obtained from Milli-Q 2D-spectrum measured at Ex/Em 350 nm/383-410 nm with the same spectrofluorometer. This step ensures comparability between the sampled events regardless of the measurement settings and changes in lamp intensity due to the temporal distance between measurements.

2.3. Statistical methods

All analyses have been carried out with the statistical software R (R Core Team, 2021). To reduce the dimensionality of its trilinear multi-way data structure, the 74 EEM were decomposed into 2 to 6 underlying components using PARAFAC models with non-negativity constraints following the procedure outlined in the package *staRdom* (Pucher et al., 2019) and validated as described in Stedmon and Bro (2008). Multiple random initializations (Harshman and Lundy, 1994) as well as a split-half and a residual analysis (Murphy et al., 2013) verified the stability of the final four-component model. For the split-half analysis, the 74 samples were randomly divided into four sub-samples and then each pair was combined and analyzed for comparison. The extracted PARAFAC component scores were transformed to Raman Units by compensating for former normalisation in the amount of each component in every sample as described by Pucher et al. (2019) to allow straightforward interpretation, and hence those represent the relative concentration of DOM compounds with similar fluorescent properties.

Spearman correlation was conducted to identify bivariate linear relationships between PARAFAC components and water quality parameters. In the next step, OLS regression (Lai et al., 1979) was used to establish the multivariate relation between PARAFAC components and water quality parameters with log-transformed total suspended solids (TSS) as a covariate to account for the varying magnitude of the events and for the potentially delayed dynamics between changing hydrograph and shifts in DOM emissions. Regression coefficients with $P < .05$ were considered statistically significant.

In order to apply multivariate regression models to the EEM data directly, the data was unfolded into a two-way array, scaled, centered, and wavelength combinations with a sample variance of fluorescence intensity lower than 0.1 were excluded due to their potential to cause numerical problems and non-converging models (Kuhn, 2008). Quantification of water quality parameters was established using two different approaches, whereby the models were compared utilizing the root mean square error (RMSE), which represents the model's error, i.e., the difference between the measured and the predicted concentration in the units of measurement (here mg L^{-1}). Firstly, a kernel PLS regression (Wold, 1985) was conducted with the package *caret* (Kuhn, 2021). The number of components in the PLS model was chosen between 1 and 10 via bootstrap, i.e., fitting the model repeatedly to data randomly sampled with replacement from the original data. The number of components minimizing the RMSE in this validation process was then used to fit the final PLS model. Secondly, an sPLS regression using the SIMPLS algorithm was calculated with the package *spls* (Chung et al., 2019). It is important to note that the PLS includes all variables in the model, even if many contribute only slightly to the prediction, i.e., have regression coefficients close to zero. In contrast, sPLS incorporates a variable selection step. Depending on the choice of the sparsity parameter, a certain number of variables is selected by setting the regression coefficients of non-selected variables to zero. The result is a set of selected variables that contains the wavelength combinations best suited to predict the target water quality parameter. The number of components (1 to 10) and the sparsity parameter (0.95 to 0.995 in 0.005 steps) were chosen by 10-fold cross-validation, i.e., the data is split into ten parts, and each is held out once, in turn, to serve as the validation set. Moreover, expert judgment was used to further reduce the number of variables while maintaining an acceptable RMSE.

3. Results and discussion

3.1. Characteristic peaks and PARAFAC components

A suitable way to visualize the EEM despite its high-dimensional structure is through contour plots. The ranges of excitation and emission wavelengths are plotted on the X and Y axes, respectively, resulting in a grid structure that includes all combinations of excitation and emission wavelengths. The EEM contains the corresponding measured fluorescence intensity in Raman Units (R.U.) for each of these combinations, displayed using a color scale. The darker the shade, the higher the fluorescence intensity in that region. Fig. 2 (a) shows contour plots for two representative EEM of the Rába River at different flow conditions. The quite distinctive position of the highest fluorescence signal for the Rába river is typically located at an excitation of about 228 mn and an emission between 330 mn and 345 mn at low-flow conditions, whereas at high-flow conditions it is still found at the same excitation, but at an emission between 400 nm and 450 nm. The entire fluorescence signal covers an emission range of 250 nm to 550 nm at low-flow and 300 nm to 600 nm at high-flow. Additional contour plots of several samples of all events with separate color scales can be found in Fig. S1 and Fig. S2 in the Supplementary Material to illustrate the variety of the fluorescence signal at different flow conditions more comprehensively.

PARAFAC analysis of all 74 samples resulted in a final model with four mutually independent areas of the fluorescence signal (components) that reflect the events' variability in terms of flow and seasonality. Leverage was under 0.25 for all samples. The results of the split-half analysis regarding the validity and stability of the established model can be found in the supplementary material (Fig. S3). Fig. 2 (b) provides a graphical overview of the four components of the final PARAFAC model, and Table 2 summarizes the exact peak location of the components and their common interpretation.

Hereafter, the extracted components (C1-C4) of the final PARAFAC model are described according to published models in the OpenFluor database (Murphy et al., 2014) specifically focusing on surface waters as well as variations or deviations in the extracted components. Components C1 and C3 lie within the area of Coble peak A, which is predominantly attributed to humified material (Coble, 1996). Both are universal components that have been described in numerous PARAFAC models. The primary and secondary peaks of C1 have both been described as a fulvic-like compound with terrestrial sources (Retelletti Brogi et al., 2020; Hong et al., 2012; Yamashita et al., 2011; Austnes et al., 2010). C3 also originates from terrestrial sources but resembles a humic-like compound (Retelletti Brogi et al., 2020; Nguyen et al., 2013; Hong et al., 2012; Austnes et al., 2010), for which it is suggested that these arise from older soil organic matter than fulvic-like compounds (Yamashita et al., 2011). Despite its large dimension, the primary peak of C4 falls into the region of Coble peak C, indicating the presence of humified material (Nguyen et al., 2013) possibly of biological, microbial, or terrestrial origin (Hong et al., 2012; Yamashita et al., 2008). However, it should be mentioned that C4 partially extends into the region of Coble peak T_2. Unfortunately, the initial expectation that the very large

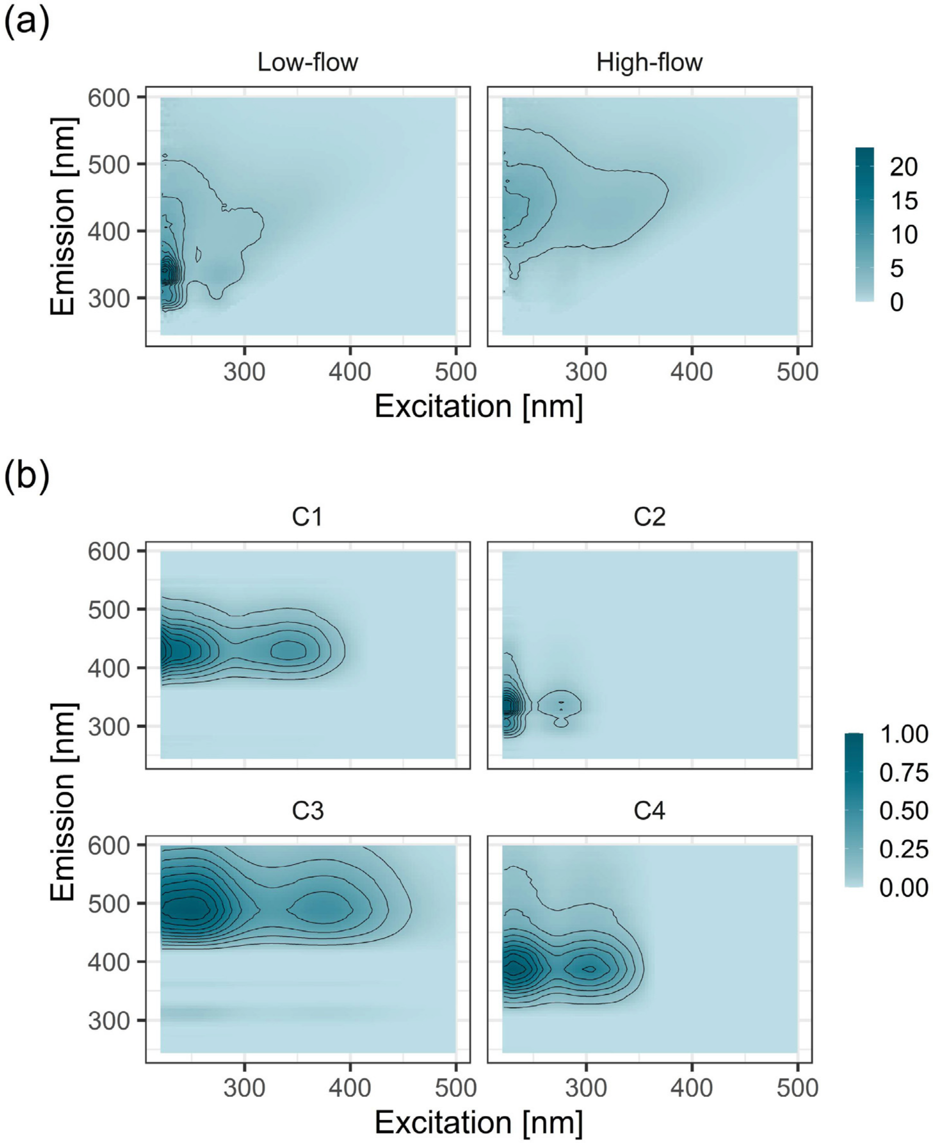


Fig. 2. (a) Representative contour plots of river water EEM at low-flow (3.3 $m^3 s^{-1}$) and high-flow (92.1 $m^3 s^{-1}$) conditions. The color scale indicates the fluorescence intensity [R.U.]. (b) Contour plots of the four PARAFAC components C1, C2, C3, and C4 normalized to the maximum fluorescence intensity of each component.

component C4 could be further decomposed into narrower, clearly separated components proved disappointing, even with more components in the PARAFAC model. Still, since this only applies to a minor peak, the presented PARAFAC model shall stick to the attribution of the unambiguous primary peak of C4 as humic-like. Component C2 is located within the region of Coble peak T_2, which is indicative of the presence of protein-associated dissolved organic compounds (Coble, 1996). The interpretation of C2 as protein-associated is somewhat controversial since this mainly applies to its secondary peak (λ_{ex}: 275/λ_{em}: 341) (Retelletti Brogi et al., 2020; Hong et al., 2012), whereas no clear association has yet been confirmed for the primary peak of C2. Most studies nevertheless suggest the characterization of component C2 as tryptophan-like (Harjung et al., 2018; Yamashita et al., 2011), whereby no clear distinction regarding the origin either from autochthonous or anthropogenic production is possible (Hong et al., 2012; Hudson et al., 2007) and requires to be inferred in the present samples via additional background information. Due to the regional proximity, C2 is nevertheless tentatively classified as such here. Considering the characteristics of the catchment and specifically the discharging industrial WWTPs, the findings subsequently presented suggest that C2 might be a component particularly influenced by industrial treated wastewater and thus representative of synthetic compounds emitted into the Rába (Fig. 4). The overall appearance of our PARAFAC model closely resembles those reported by Pitta and Zeri (2021), who studied the influence of sample augmentation on the resulting global PARAFAC model and therein also reported individual models for different watersheds.

Table 2
Overview of PARAFAC components found in the samples of river high-flow events. Peak labels and descriptions follow the convention of Coble (2007) and Hudson et al. (2007). A comparison to components of PARAFAC models published on OpenFluor (Murphy et al., 2014) is given in the description.

Component	$\lambda_{ex}/\lambda_{em}$ [nm]	Peak	Description
C1	222/429	A	Humic-like
C2	225/341	T_2	Tryptophan-like
C3	249/489	A	Humic-like
C4	231/387	C	Humic-like

3.2. Shift of PARAFAC components during events

Fig. 3 shows the time series of the discharge and the four PARAFAC components for each of the seven sampled events. It is evident that as the flow increases, the fluorescence intensity of C2 decreases rapidly, indicating an attenuation of the signal due to a dilution of the organic compound concentration. Depending on the event's strength, this dilution occurs with a time lag of 8 to 15 h (B, C, and D) due to the so-called piston flow, i.e., the flood wave carries the water in the river before the event along its path (Sophocleous, 2002). Thus, samples continue to be taken from this base-flow situation before the newly discharged water reaches the monitoring station. Only from this point on the changing water composition due to high-flow situation can be measured. Regardless of the piston flow, the effect of combined sewer overflows should at least be considered. It also occurs with a delay and is accompanied by a sudden increase in the organic load as soon as its storage capacity has been exceeded. Especially in the case of massive storm events, this might further contribute to an additional increase in the DOM. Differences in the time lag are explained by the exact location of the storm event in the spacious catchment area of the river and how remote it is from the measuring station. Subsequently, the signal of C2 remains stable at an evidently lower level for some time with decreasing discharge (B, C, and G) before it increases with some delay until the initial level is regained (E).

Hong et al. (2012) attribute this pattern to the inflow of anthropogenic sources, such that there is a dilution of the protein-like compounds during a high-flow event if the inflow has not increased proportionately. In contrast, the signal's dynamic of C1 is analogous to that of the discharge, but also with a similar time lag and a gradual decrease, which could be the result of the piston flow effect. In particular, the signal increases evidently for events with discharge greater than 30 m^3 s^{-1} (A to D), while this is significantly less pronounced for events with lower discharge. This result is in perfect accordance with the findings of Nguyen et al. (2013) and Fellman et al. (2009). At high flow, C1 may nominally obtain a signal nearly as high as C2 (10 R.U. to 12 R.U. at event A), or C1 may even temporarily supersede C2 as the highest peak (B to D). However, there is strong evidence that this may not necessarily occur as well, especially in unaffected rivers (Nguyen et al., 2013). Correspondingly, the components C3 and C4 follow the course of component C1, but the range of the absolute fluorescence signal is much narrower. Especially the apparent shift toward a higher concentration of humic-like compounds indicates the input of near-surface soil layers as the key contributor of DOM during high-flow events (Fellman et al., 2009). The altered ratio of humic- to fulvic-like compounds, i.e., components C3 and C1, may be attributed to this as well (Fellman et al., 2009).

Retelletti Brogi et al. (2020) indicate that, in general, protein-like compounds predominate in summer. The available data for the Rába River show a corresponding variation in the base-flow signal for component C2. For this reason, it can be concluded that the fluorescence spectroscopic signal of C2 at base-flow conditions is attributable to the influence of

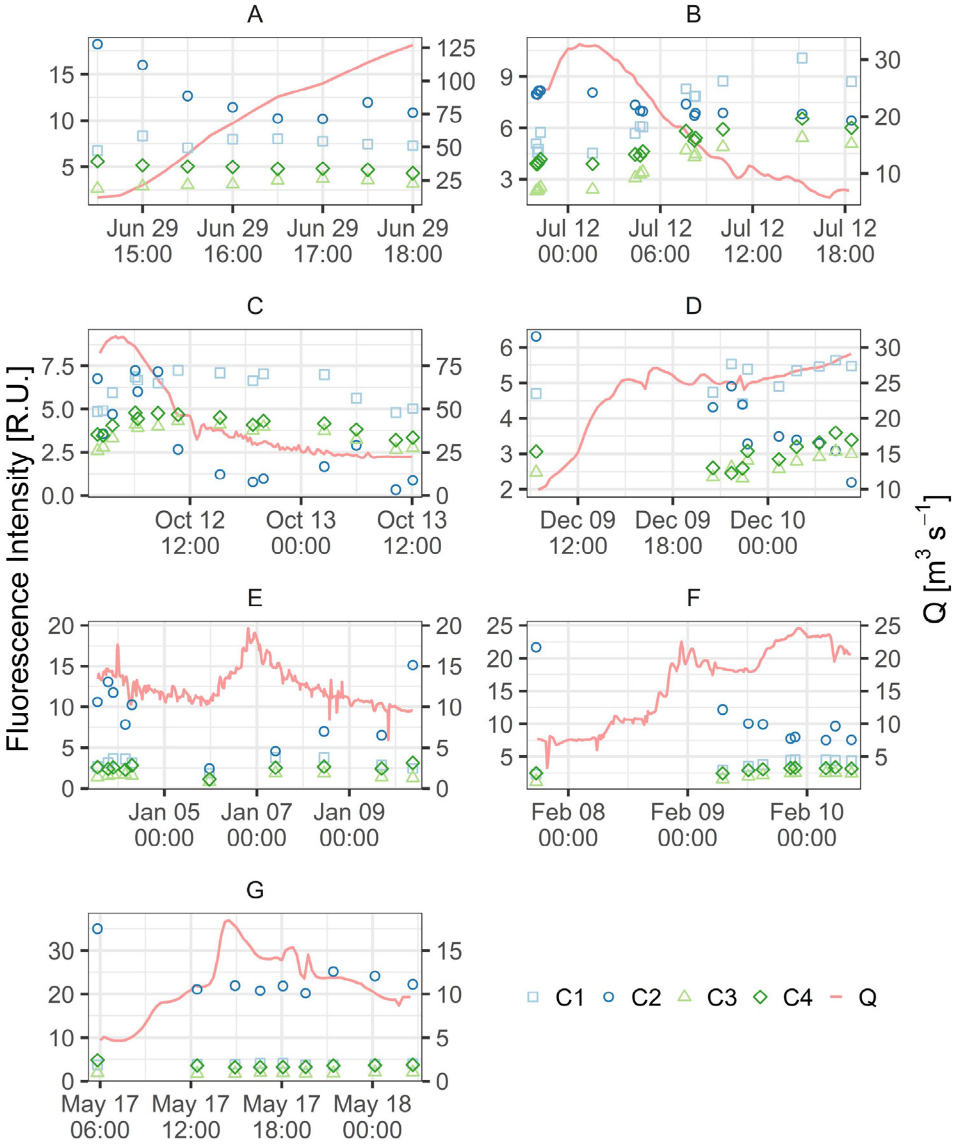


Fig. 3. Time series of the discharge and the four PARAFAC components C1, C2, C3, and C4 during the seven sampled events.

discharges from several industrial WWTPs and that the fluctuation is much more subject to the production cycle than to seasonal conditions. Under base-flow conditions, the components C1 and C3 are consistently similar around 2 R.U. to 6 R.U. and reveal only minor differences, which are well explained by seasonal variations. Especially in summer, the greatly enhanced primary productivity in the river and the increased transfer of soil organic matter via more frequent rainfall events result to higher DOC concentrations than in winter (see summer event B vs. winter event E in Fig. 3) (Harjung et al., 2018; Yamashita et al., 2011). However, it is evident that the seasonal cycle depends on the respective catchment. Studies in other rivers, for instance, report an increase in humic-like components in May and June (Hong et al., 2012). Due to the high-flow season in autumn, humic-like DOM, mainly from terrestrial sources, may well be input to the river at this time, too (Retelletti Brogi et al., 2020). Together, these mechanisms lead to a reasonably low variance of the fluorescence spectroscopic signal of components C1 and C3 at base-flow in the Rába River. Last but not least, discharges of municipal WWTPs and WWTPs of some industrial sectors (e.g., meat processing) also add humic-like substances (Rodríguez-Vidal et al., 2020; Li et al., 2014). However, this is a constant input from point sources, which is not affected by production-related cycles.

3.3. Correlation between EEM and water quality parameters

The interpretation of C1, C3, and C4 as mainly linked to diffuse inputs during high-flow and C2 as influenced by industrial emissions is further confirmed by pairwise correlations between the components and relevant water quality parameters (Fig. 4). C1, C3, and C4, which are highly correlated among themselves ($r \geq .84$), are also medium to highly associated with DOC ($.60 \leq r \leq .89$), i.e., these serve as proxy parameters for organic pollution. The fact that three components show this correlation simultaneously strengthens the assumption that different wavelength areas in the EEM represent different DOM compounds. DOC does not distinguish these as a sum parameter, so the EEM or the PARAFAC components bring a more nuanced insight about changes to which the DOM composition is subject during the high-flow event. Standard water quality parameters provide this only to a much lower extent, or in many cases, not at all. For this reason, a perfect correlation between DOC and the components C1, C3, or C4 is not to be expected.

However, C2, the component which covers a whole different wavelength region compared to the other components, shows a strong positive correlation with Cl^- ($r = .75$), the input of which at the Rába River is mainly due to industrial discharges. It is crucial to know that industrial wastewater treatment plants usually do not have a sewer and little or no storage facilities. This leads to relatively stable effluent volume flows and concentrations, which fluctuate only due to changes in production qualities and quantities. This is also confirmed by the dilution of the chloride concentration during high-flow events and suggests that constant point discharges largely influence it in addition to a natural background concentration. Their contribution indeed decreases compared to the total runoff during an event, but if the concentration is set in relation to the flow rate, an increase in the chloride load can be seen. This finding particularly indicates diffuse chloride inputs during high-flow events since the chloride load from point sources is independent of river flow and is rather constant over time. Fluctuations in the input of chloride load due to changes in the production cycles of several months do not matter due to the shorter duration of the high-flow events and can therefore be disregarded. The mentioned diffuse input sources, besides their geogenic background, are most likely de-icing salt run-off from roads (mainly NaCl and $CaCl_2$) and leaching of certain agricultural fertilizers from surrounding agricultural soils since some fertilizers contain Cl^-. Furthermore, the components C1, C3, and C4 result highly correlated ($.65 \leq r \leq .70$) with orthophosphate (PO_4-P). PO_4-P exists in dissociated form in water; thus, no absorbance spectrum is available for this compound in the UV/Vis-range (Linstrom and Mallard, 2022). However, the increased phosphorus transfer from agricultural soils into surface waters during rainfall and erosion events is a well-known process that can explain this positive correlation (Sims and Sharpley, 2005).

It is reasonable to assume that not all of the numerous data points in the EEM contribute equally to the prediction of standard water quality parameters. Three different statistical models are compared regarding their ability

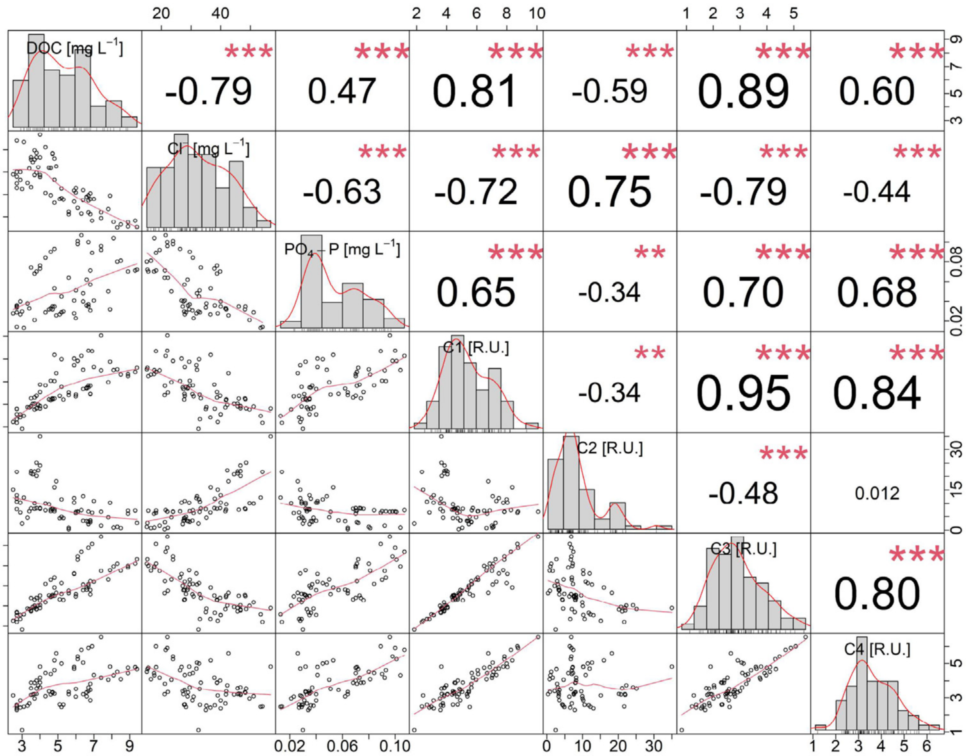


Fig. 4. Distribution of and pairwise correlation between water quality parameters and the four PARAFAC components C1, C2, C3, and C4. Correlation coefficients are calculated according to Spearman. Asterisks represent statistical significance (** indicates $P < .01$, *** indicates $P < .001$).

to verify wavelengths that are sufficiently capable proxies to accurately represent the pattern described in the previous section throughout different discharge conditions and annual seasons. The OLS regression analyzes the relationship between DOC resp. Cl⁻ and the previously described PARAFAC components. To avoid multicollinearity, component C3 has been used to represent the highly correlated carbon-associated components (C1, C3, and C4), as it exhibits the highest correlation with the parameters of interest (Fig. 4). The resulting regression equations are shown in Eqs. (1) and (2).

$$\widehat{\mathrm{DOC}}_{\mathrm{OLS}} = 1.33 - 0.09 \cdot \mathrm{C2} + 1.02 \cdot \mathrm{C3} + 0.32 \cdot \log(\mathrm{TSS}) \tag{1}$$

$$\widehat{\mathrm{Cl}^-}_{\mathrm{OLS}} = 42.27 + 0.87 \cdot \mathrm{C2} - 5.53 \cdot \mathrm{C3} - 0.25 \cdot \log(\mathrm{TSS}) \tag{2}$$

To evaluate the quality of the different statistical modeling approaches, the predicted DOC and Cl⁻ concentrations are compared to the actual measured concentrations in Fig. 5. Although the models perform quite satisfactorily by means of the RMSE ($RMSE_{\mathrm{OLS}} = 0.9$ mg L^{-1} for DOC and $RMSE_{\mathrm{OLS}} = 5.31$ mg L^{-1} for Cl⁻), the OLS regression over- or underestimates the concentrations of some high-flow events throughout. Retelletti Brogi et al. (2020) found that the correlation between DOC and PARAFAC components is different when calculated separately for different seasons. Since only univariate regression models were considered, the results are comparable only to a limited extent. Yet, there is no evidence that the systematic under- or overestimation is due to the seasonal occurrence of high-flow events in the study at hand. The underestimated and overestimated events for both DOC and Cl⁻ even cannot be grouped according to their maximum flow. For example, events B and D both have a maximum flow close to 30 m^3 s^{-1}. However, the OLS model systematically overestimated the DOC for Event B, while that for Event D is underestimated. The same is true for the comparison of Events E and G. Adding covariates to the model, e.g., log TSS or the concentration of different phosphorus fractions, could perhaps improve the performance in this respect by allowing to account for different phases of each event, different flow conditions and hysteresis effects. In our study, adding these covariates did not improve the performance significantly. Still, it would be worth testing this hypothesis in the future with a larger data set covering a higher number of high-flow events. Furthermore, the assumption of a linear relationship between the components and DOC or Cl⁻ is not valid across all events, so this method cannot aim for using the PARAFAC components as proxy parameters. For Event F, this is strikingly evident when the predicted and measured Cl⁻ concentrations are compared.

In general, sPLS and PLS performed very successfully for DOC ($RMSE_{\mathrm{sPLS}} = 0.41$ mg L^{-1} and $RMSE_{\mathrm{PLS}} = 0.11$ mg L^{-1}) as well as Cl⁻ ($RMSE_{\mathrm{sPLS}} = 2.21$ mg L^{-1} and $RMSE_{\mathrm{PLS}} = 1$ mg L^{-1}). Both methods considerably exceed the OLS regression in terms of RMSE and fit very well for all sampled high-flow events, regardless of their differences in flow and seasonal occurrence. Whereas PLS clearly outperforms sPLS with regard to its predictive capability, the entire EEM has to be provided for this purpose. On the other hand, sPLS requires only 18 pairs of emission-excitation wavelengths for Cl⁻ (Eq. (3)) and 16 pairs of wavelengths for DOC (Eq. (4)), respectively.

$$\begin{aligned}\widehat{\mathrm{Cl}^-}_{\mathrm{sPLS}} = {} & 2.80 \cdot \lambda_{222/373} + 1.49 \cdot \lambda_{222/573} + 1.42 \cdot \lambda_{225/592} \\ & + 1.20 \cdot \lambda_{243/296} + 1.09 \cdot \lambda_{225/269} + 1.06 \cdot \lambda_{228/269} \\ & + 1.00 \cdot \lambda_{264/287} + 0.78 \cdot \lambda_{225/373} + 0.64 \cdot \lambda_{246/296} \\ & - 0.23 \cdot \lambda_{222/260} - 0.43 \cdot \lambda_{222/517} - 0.62 \cdot \lambda_{222/364} \\ & - 0.73 \cdot \lambda_{222/264} - 1.25 \cdot \lambda_{228/278} - 1.42 \cdot \lambda_{228/251} \\ & - 1.85 \cdot \lambda_{222/569} - 2.75 \cdot \lambda_{222/531} - 5.85 \cdot \lambda_{378/517}\end{aligned} \tag{3}$$

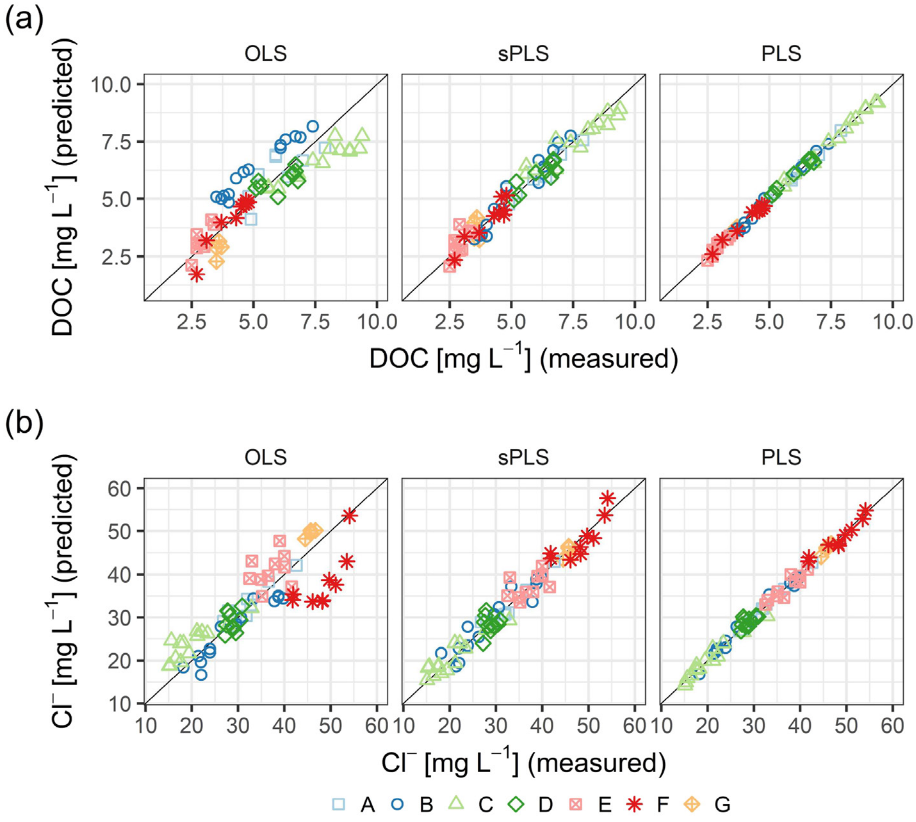


Fig. 5. Predicted versus measured (a) DOC and (b) Cl⁻ concentration in event samples. Prediction is based on OLS regression (OLS), sparse PLS regression (sPLS) and PLS regression (PLS).

$$\begin{aligned}\widehat{\mathrm{DOC}}_{\mathrm{sPLS}} = \; & 0.69 \cdot \lambda_{243/592} + 0.55 \cdot \lambda_{225/550} + 0.45 \cdot \lambda_{222/273} \\ & + 0.42 \cdot \lambda_{402/508} + 0.37 \cdot \lambda_{408/517} + 0.32 \cdot \lambda_{399/517} \\ & + 0.26 \cdot \lambda_{231/255} + 0.26 \cdot \lambda_{393/517} + 0.23 \cdot \lambda_{222/264} \\ & - 0.20 \cdot \lambda_{225/260} - 0.27 \cdot \lambda_{399/442} - 0.28 \cdot \lambda_{222/255} - 0.31 \cdot \lambda_{222/573} \\ & - 0.35 \cdot \lambda_{228/273} - 0.40 \cdot \lambda_{225/373} - 0.52 \cdot \lambda_{222/564}\end{aligned} \tag{4}$$

Most of the by sPLS selected wavelength pairs for modeling Cl^- are located in area d, which expands strongest toward area c when the relaxation of the sparsity parameter increases the number of variables in the model. Area c itself is very narrow but also well specific for the highest peak in the fluorescence signal at low-flow conditions (cf. Fig. 2). This result is in remarkable alignment with the PARAFAC component C2, which is highly correlated with Cl^-. For DOC, region f is the one that spans the most pairs of wavelengths. As the number of variables is increased, region f is found to expand fastest toward an excitation wavelength of 320 nm. Hence, this region plays a crucial role in enabling the EEM to serve as a proxy for DOC. This is particularly interesting since this region is contained in component C3 but only as a secondary peak. Although both peaks, primary and secondary, are essential for characterization as a terrestrial humic-like compound, the information of the secondary peak is more relevant in the model of the sPLS. Moreover, the suitability of peak C, which is located in this area, as a proxy parameter for DOC has been well demonstrated (Carstea et al., 2020). But the fact that the applicability of peak C as the only universal proxy for DOC is controversial (Baldwin and Valo, 2015), speaks in favor of the approach chosen here to include several wavelength combinations in the prediction. This advantage of the multi-parameter approach applies especially if peak C is not the dominant feature in the EEM (Saraceno et al., 2009). Regions a and e provide information about the highest emission wavelengths at which a fluorescent signal is still present. Especially with increasing flow, it can be observed how the fluorescence fingerprint stretches to emission wavelengths of 500 nm and higher. Fig. 6 visualizes the relevant wavelength combinations selected by sPLS.

For the prediction of both DOC and Cl^-, those regions that typically exhibit low signal (c versus g and b versus f) are also included in the sPLS model in each case albeit with negative coefficients. This is likely to be specific to the situation at the Rába River, where the peak shift described earlier means that both peaks never present simultaneously with similar high fluorescence signals. Compared to the peaks of the four PARAFAC components, the sPLS indeed contains areas that do not correspond to the primary peaks. Instead, it rather exploits the entire fluorescence spectroscopic fingerprint by, for instance, also accounting for peripheral spectroscopic zones. Thus, the prediction performance improves significantly by using slightly more information from the EEM than just the PARAFAC components. This is in accordance with the results of Yin et al. (2021), who showed that a boosting regression tree yielded a better prediction for DOC than the corresponding PARAFAC model. As such, fluorescence spectroscopy coupled with multivariate statistics becomes not only a promising proxy parameter for standard water quality parameters but also provides important insights into the qualitative change of the DOM in the riverine system.

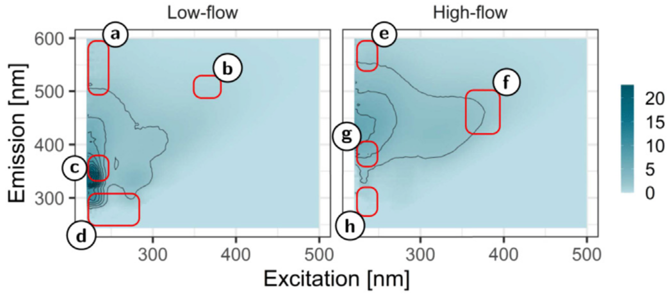


Fig. 6. Wavelength combinations selected by sPLS. The rectangles a to d mark areas where 18 wavelength combinations for modeling Cl^- are located. The rectangles e to g mark areas where 16 wavelength combinations for modeling DOC are located. Note that each rectangle spans over one or more wavelength combinations. The exact excitation-emission wavelengths of the combinations are specified in Eq. (3) for Cl^- and Eq. (4) for DOC.

In contrast to other established optical measurement methods, which require different calibrations for both low- and high-flow conditions, quantification using EEM even succeeds across the entire range of flow conditions and events of various magnitudes. Yet the use of combinations of excitation and emission wavelengths adjusted via sPLS is clearly superior to the extracted PARAFAC components. As expected, the prediction accuracy increases with the number of wavelength combinations included in the model. This in fact also applies to OLS regression (Carter et al., 2012). However, this is not the case for all available water quality parameters (e.g., NH_4-N, data not shown), while some can be at least estimated through a mediating relationship (e.g., PO_4-P, data not shown).

4. Conclusions

This paper shows that fluorescence spectroscopy is well suited for identifying and predicting water quality dynamics during river high-flow events. In particular, the combination with multivariate statistical techniques precisely reflects known phenomena of flood dynamics, such as the piston flow, much better than previous strategies like peak picking could do. This result fosters the hope that fluorescence spectroscopic online measurements can monitor water quality even more effectively on-site, provided that preceding filtration is implemented.

Although combining the extracted PARAFAC components with OLS is a widely accepted approach, these models are not universally comparable as no globally applicable PARAFAC model has been found yet. Using the whole EEM via PLS is advantageous as it incorporates all the information. Still, at the same time, it is a disadvantage because of the high-dimensional data structure. Consequently, the use of sPLS represents a novel and promising solution, with the benefit of combining the selection of wavelength combinations with quantifying water quality parameters all at once. This provides a considerably better quantification than PARAFAC components and an equally satisfying quantification as PLS but requires significantly fewer wavelength combinations. For this reason, the sPLS model is highly recommended for fluorescence spectroscopy with in-field instruments with a local calibration. The future goal is to extend this to water quality parameters beyond DOC and Cl^- as this opens a massive gateway to expand the targeted water quality monitoring by fluorescence spectroscopy.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank the present and past members of the NaWas project team, especially project leader Steffen Kittlaus, for the great effort during the sampling campaigns, Andreas Winkelbauer, Lukas Kornfeind and Roland Fuiko for stimulating discussions about online-monitoring and automatic sampling, Ferdinand Zinn-Zinnenburg for his additional work in the lab and the Office of the Provincial Government of Burgenland for providing high-resolution discharge data.

This research was funded by the Austrian Federal Ministry of Agriculture, Regions and Tourism, and the TU Wien via the competitive funding program for innovative projects. The authors acknowledge TU Wien Bibliothek for financial support through its Open Access FundingProgramme.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.scitotenv.2022.158016.

References

Austnes, K., Evans, C.D., Eliot-Laize, C., Naden, P.S., Old, G.H., 2010. Effects of storm events on mobilisation and in-stream processing of dissolved organic matter (DOM) in a welsh peatland catchment. Biogeochemistry 99, 157–173. https://doi.org/10.1007/s10533-009-9399-4.

Baldwin, D.S., Valo, W., 2015. Exploring the relationship between the optical properties of water and the quality and quantity of dissolved organic carbon in aquatic ecosystems: strong correlations do not always mean strong predictive power. Environ. Sci.-Proc. Imp. 17, 619–630. https://doi.org/10.1039/C4EM00473F.

Bieroza, M., Baker, A., Bridgeman, J., 2009. Exploratory analysis of excitation-emission matrix fluorescence spectra with self-organizing maps as a basis for determination of organic matter removal efficiency at water treatment works. J. Geophys. Res. 114, G00F07. https://doi.org/10.1029/2009JG000940.

Bro, R., 1997. PARAFAC. Tutorial and Applications. 38. Chemometr. Intell. Lab, pp. 149–171. https://doi.org/10.1016/S0169-7439(97)00032-4.

Carstea, E.M., Baker, A., Bieroza, M., Reynolds, D., 2010. Continuous fluorescence excitation–emission matrix monitoring of river organic matter. Water Res. 44, 5356–5366. https://doi.org/10.1016/j.watres.2010.06.036.

Carstea, E.M., Bridgeman, J., Baker, A., Reynolds, D.M., 2016. Fluorescence spectroscopy for wastewater monitoring: a review. Water Res. 95, 205–219. https://doi.org/10.1016/j.watres.2016.03.021.

Carstea, E.M., Popa, C.L., Baker, A., Bridgeman, J., 2020. In situ fluorescence measurements of dissolved organic matter: a review. Sci. Total Environ. 699. https://doi.org/10.1016/j.scitotenv.2019.134361.

Carter, H.T., Tipping, E., Koprivnjak, J.-F., Miller, M.P., Cookson, B., Hamilton-Taylor, J., 2012. Freshwater DOM quantity and quality from a two-component model of UV absorbance. Water Res. 46, 4532–4542. https://doi.org/10.1016/j.watres.2012.05.021.

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

Chen, W., Westerhoff, P., Leenheer, J.A., Booksh, K., 2003. Fluorescence excitation-emission matrix regional integration to quantify spectra for dissolved organic matter. Environ. Sci. Technol. 37, 5701–5710. https://doi.org/10.1021/es034354c.

Chung, D., Chun, H., Keles, S., 2019. spls: Sparse Partial Least Squares (SPLS) Regression and Classification URL: https://CRAN.R-project.org/ package=spls r package version 2.2-3.

Coble, P.G., 1996. Characterization of marine and terrestrial dom in seawater using excitation-emission matrix spectroscopy. Mar. Chem. 51, 325–346. https://doi.org/10.1016/0304-4203(95)00062-3.

Coble, P.G., 2007. Marine optical biogeochemistry: the chemistry of ocean color. Chem. Rev. 107, 402–418. https://doi.org/10.1021/cr050350+.

Cohen, E., Levy, G.J., Borisover, M., 2014. Fluorescent components of organic matter in wastewater: efficacy and selectivity of the water treatment. Water Res. 55, 323–334. https://doi.org/10.1016/j.watres.2014.02.040.

Croghan, D., Khamis, K., Bradley, C., Van Loon, A.F., Sadler, J., 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. Sci. Total Environ. 755, 142731. https://doi.org/10.1016/j.scitotenv. 2020.142731.

EL-Nahhal, I., Redon, R., Raynaud, M., EL-Nahhal, Y., Mounier, S., 2020. Characterization of the fate and changes of post-irradiance fluorescence signal of filtered anthropogenic effluent dissolved organic matter from wastewater treatment plant in the coastal zone of Gapeau river. Environ. Sci. Pollut. Res. 27, 23141–23158. https://doi.org/10.1007/s11356-020-08842-w.

EL-Nahhal, I., Redon, R., Raynaud, M., EL-Nahhal, Y., Mounier, S., 2021. Modelling of impact of presence/absence of suspended particulate organic matter from river and sea and effluent wastewater on fluorescence signal in the coastal area of Gapeau River. Environ. Sci. Pollut. Res. 28, 36707–36726. https://doi.org/10.1007/s11356-021-13265-2.

Fellman, J.B., Hood, E., Edwards, R.T., D'Amore, D.V., 2009. Changes in the concentration, biodegradability, and fluorescent properties of dissolved organic matter during stormflows in coastal temperate watersheds. J. Geophys. Res. Biogeosci. 114. https://doi.org/10.1029/2008JG000790.

Frank, S., Goeppert, N., Goldscheider, N., 2017. Fluorescence-based multi-parameter approach to characterize dynamics of organic carbon, faecal bacteria and particles at alpine karst springs. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2017.09.095.

Harjung, A., Sabater, F., Butturini, A., 2018. Hydrological connectivity drives dissolved organic matter processing in an intermittent stream. Limnologica 68, 71–81. https://doi.org/10.1016/j.limno.2017.02.007.

Harshman, R.A., Lundy, M.E., 1994. PARAFAC: parallel factor analysis. Comput. Stat. Data Anal. 18, 39–72. https://doi.org/10.1016/0167-9473(94)90132-5.

Hong, H., Yang, L., Guo, W., Wang, F., Yu, X., 2012. Characterization of dissolved organic matter under contrasting hydrologie regimes in a subtropical watershed using parafac model. Biogeochemistry 109, 163–174.

Hudson, N., Baker, A., Reynolds, D., 2007. Fluorescence analysis of dissolved organic matter in natural, waste and polluted waters—a review. River Res. Appl. 23, 631–649. https://doi.org/10.1002/rra.1005.

Ishii, S.K.L., Boyer, T.H., 2012. Behavior of reoccurring PARAFAC components in fluorescent dissolved organic matter in natural and engineered systems: a critical review. Environ. Sci. Technol. 46, 2006–2017. https://doi.org/10.1021/es2043504.

Khamis, K., Bradley, C., Hannah, D.M., 2020. High frequency fluorescence monitoring reveals new insights into organic matter dynamics of an urban river, Birmingham, UK. Sci. Total Environ. 710, 135668. https://doi.org/10.1016/j. scitotenv.2019.135668.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26. https://doi.org/10.18637/jss.v028.i05.

Kuhn, M., 2021. caret: Classification and Regression Training URL: https://CRAN.R-project.org/package=caret r package version 6.0-90.

Lai, T., Robbins, H., Wei, C., 1979. Strong consistency of least squares estimates in multiple regression II. J. Multivar. Anal. 9, 343–361. https://doi.org/10.1016/0047-259X(79)90093-9.

Lakowicz, J.R., 2006. Principles of Fluorescence Spectroscopy. 3rd ed. Springer, US https://doi.org/10.1007/978-0-387-46312-4.

Li, W.-T., Chen, S.-Y., Xu, Z.-X., Li, Y., Shuang, C.-D., Li, A.-M., 2014. Characterization of dissolved organic matter in municipal wastewater using fluorescence PARAFAC analysis and chromatography multi-Excitation/Emission scan: a comparative study. Environ. Sci. Technol. 48, 2603–2609. https://doi.org/10.1021/es404624q.

Linstrom, P., Mallard, W. (Eds.), 2022. NIST Chemistry Webbook: NIST Standard Reference Database Number 69. National Institute of Standards and Technology https://doi.org/10.18434/T4D303.

Massicotte, P., 2019. eemR: Tools for Pre-Processing Emission-ExcitationMatrix (EEM) Fluorescence Data https://CRAN.R-project.org/ package=eemR r package version 1.0.1.

Murphy, K.R., Stedmon, C.A., Graeber, D., Bro, R., 2013. Fluorescence spectroscopy and multi-way techniques. PARAFAC. Anal. Methods 5, 6557. https://doi.org/10.1039/c3ay41160e.

Murphy, K.R., Stedmon, C.A., Wenig, P., Bro, R., 2014. OpenFluor– an online spectral library of auto-fluorescence by organic compounds in the environment. Anal. Methods 6, 658–661. https://doi.org/10.1039/C3AY41935E.

Nguyen, H.V.-M., Lee, M.-H., Hur, J., Schlautman, M.A., 2013. Variations in spectroscopic characteristics and disinfection byproduct formation potentials of dissolved organic matter for two contrasting storm events. J. Hydrol. 481, 132–142. https://doi.org/10.1016/j.jhydrol.2012.12.044.

Pitta, E., Zeri, C., 2021. The impact of combining data sets of fluorescence excitation - emission matrices of dissolved organic matter from various aquatic sources on the information retrieved by PARAFAC modeling. Spectrochim. Acta A 258, 119800. https://doi.org/10.1016/j.saa.2021.119800.

Pucher, M., Wünsch, U., Weigelhofer, G., Murphy, K., Hein, T., Graeber, D., 2019. staRdom: versatile software for analyzing spectroscopic data of dissolved organic matter in R. Water 11, 2366. https://doi.org/10.3390/w11112366.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Retelletti Brogi, S., Balestra, C., Casotti, R., Cossarini, G., Galletti, Y., Gonnelli, M., Vestri, S., Santinelli, C., 2020. Time resolved data unveils the complex DOM dynamics in a Mediterranean river. Sci. Total Environ. 733, 139212. https://doi.org/10.1016/j.scitotenv.2020.139212.

Rodríguez-Vidal, F.J., García-Valverde, M., Ortega-Azabache, B., González-Martínez, A., Bellido-Fernández, A., 2020. Characterization of urban and industrial wastewaters using excitation-emission matrix (EEM) fluorescence: searching for specific fingerprints. J. Environ. Manag. 263, 110396. https://doi.org/10.1016/j.jenvman.2020.110396.

Saraceno, J.F., Pellerin, B.A., Downing, B.D., Boss, E., Bachand, P.A.M., Bergamaschi, B.A., 2009. High-frequency in situ optical measurements during a storm event: assessing relationships between dissolved organic matter, sediment concentrations, and hydrologic processes. J. Geophys. Res.-Biogeo. 114. https://doi.org/10.1029/2009JG000989.

Sgroi, M., Roccaro, P., Korshin, G.V., Vagliasindi, F.G.A., 2017. Monitoring the behavior of emerging contaminants in wastewater-impacted Rivers based on the use of fluorescence excitation emission matrixes (EEM). Environ. Sci. Technol. 51, 4306–4316. https://doi.org/10.1021/acs.est.6b05785.

Sims, J., Sharpley, A., 2005. Phosphorus: Agriculture and the Environment. American Society of Agronomy, Agronomy Series.

Sophocleous, M., 2002. Interactions between groundwater and surface water: the state of the science. Hydrogeol. J. 10, 52–67. https://doi.org/10.1007/s10040-001-0170-8.

Stedmon, C.A., Bro, R., 2008. Characterizing dissolved organic matter fluorescence with parallel factor analysis: a tutorial. Limnol. Oceanogr. Methods 6, 572–579. https://doi.org/10.4319/lom.2008.6.572.

Vidon, P., Wagner, L.E., Soyeux, E., 2008. Changes in the character of DOC in streams during storms in two Midwestern watersheds with contrasting land uses. Biogeochemistry 88, 257–270. https://doi.org/10.1007/s10533-008-9207-6.

Winkelbauer, A., Fuiko, R., Krampe, J., Winkler, S., 2014. Crucial elements and technical implementation of intelligent monitoring networks. Water Sci. Technol. 70, 1926–1933. https://doi.org/10.2166/wst.2014.415.

Wold, H., 1985. Partial least squares. In: Kotz, S., Johnson, N.L. (Eds.), Encyclopedia of Statistical Sciences. Multivariate Analysis to Plackett and Burman Designs6. John Wiley & Sons, New York, pp. 581–591.

Wünsch, U.J., Bro, R., Stedmon, C.A., Wenig, P., Murphy, K.R., 2019. Emerging patterns in the global distribution of dissolved organic matter fluorescence. Anal. Methods 11, 888–893. https://doi.org/10.1039/C8AY02422G.

Yamashita, Y., Jaffé, R., Maie, N., Tanoue, E., 2008. Assessing the dynamics of dissolved organic matter (DOM) in coastal environments by excitation emission matrix fluorescence and parallel factor analysis (EEM-PARAFAC). Limnol. Oceanogr. 53, 1900–1908. https://doi.org/10.4319/lo.2008.53.5.1900.

Yamashita, Y., Kloeppel, B.D., Knoepp, J., Zausen, G.L., Jaffé, R., 2011. Effects of watershed history on dissolved organic matter characteristics in headwater streams. Ecosystems 14, 1110–1122. https://doi.org/10.1007/s10021-011-9469-z.

Yin, H., Wang, K., Liu, Y., Huang, P., Yu, J., Hou, D., 2021. Fluorescence excitation-emission matrix spectroscopy and boosting regression tree model to detect dissolved organic carbon in water. Water 13, 3612. https://doi.org/10.3390/w13243612.

Zoboli, O., Clara, M., Gabriel, O., Scheffknecht, C., Humer, M., Brielmann, H., Kulcsar, S., Trautvetter, H., Kittlaus, S., Amann, A., Saracevic, E., Krampe, J., Zessner, M., 2019. Occurrence and levels of micropollutants across environmental and engineered compartments in Austria. J. Environ. Manag. 232, 636–653. https://doi.org/10.1016/j.jenvman.2018.10.074.