Environmental Science
Water Research & Technology

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Environ. Sci.: Water Res. Technol.*, 2022, **8**, 836

AbspectroscoPY, a Python toolbox for absorbance-based sensor data in water quality monitoring†

C. Cascone, ‡*[a] K. R. Murphy, [b] H. Markensten, [a] J. S. Kern, [c] C. Schleich,[d] A. Keucken[de] and S. J. Köhler [af]

The long-term trend of increasing natural organic matter (NOM) in boreal and north European surface waters represents an economic and environmental challenge for drinking water treatment plants (DWTPs). High-frequency measurements from absorbance-based online spectrophotometers are often used in modern DWTPs to measure the chromophoric fraction of dissolved organic matter (CDOM) over time. These data contain valuable information that can be used to optimise NOM removal at various stages of treatment and/or diagnose the causes of underperformance at the DWTP. However, automated monitoring systems generate large datasets that need careful preprocessing, followed by variable selection and signal processing before interpretation. In this work we introduce AbspectroscoPY ("Absorbance spectroscopic analysis in Python"), a Python toolbox for processing time-series datasets collected by *in situ* spectrophotometers. The toolbox addresses some of the main challenges in data preprocessing by handling duplicates, systematic time shifts, baseline corrections and outliers. It contains automated functions to compute a range of spectral metrics for the time-series data, including absorbance ratios, exponential fits, slope ratios and spectral slope curves. To demonstrate its utility, AbspectroscoPY was applied to 15-month datasets from three online spectrophotometers in a drinking water treatment plant. Despite only small variations in surface water quality over the time period, variability in the spectrophotometric profiles of treated water could be identified, quantified and related to lake turnover or operational changes in the DWTP. This toolbox represents a step toward automated early warning systems for detecting and responding to potential threats to treatment performance caused by rapid changes in incoming water quality.

Received 16th June 2021,
Accepted 16th February 2022

DOI: 10.1039/d1ew00416f

rsc.li/es-water

Water impact

The water treatment sector is increasingly moving toward digitalisation and online sensing, which produces large datasets requiring preprocessing before visualisation and analysis. To this end we have developed an open-source Python toolbox that implements semi-automated processing of spectrophotometric datasets. This will assist in the sustainable management of resources (water and chemicals) during drinking water production.

1. Introduction

Automation plays an essential role in drinking water treatment plants (DWTPs). Many process operation decisions, in both manual and automated systems, are based on data acquired from online sensors. Sensors are increasingly used in drinking water production as a tool for real-time analysis of water quality providing early warning of potential contamination and decision support for process control.[1] Sensors provide either direct measurements of the biological, chemical and physical components of interest (*e.g.*, conductivity, pH, temperature, dissolved oxygen, turbidity, flow cytometry) or measure surrogate parameters that correlate with these.[2–4] Absorbance-based sensors are used worldwide for drinking-, waste-, environmental- and

[a] *Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, SLU, SE 750 07 Uppsala, Sweden. E-mail: claudia.cascone@slu.se, claudia.cascone@gmail.com, hampus.markensten@slu.se, Stephan.kohler@slu.se*
[b] *Department of Architecture and Civil Engineering, Division of Water Environment Technology, Chalmers University of Technology, SE 412 96 Gothenburg, Sweden. E-mail: murphyk@chalmers.se*
[c] *Department of Engineering Mechanics, Royal Institute of Technology, KTH, SE 100 44 Stockholm, Sweden. E-mail: skern@mech.kth.se*
[d] *Vatten & Miljö i Väst AB, SE 311 22 Falkenberg, Sweden. E-mail: Caroline.Schleich@vivab.info, Alexander.Keucken@vivab.info*
[e] *Department of Building and Environmental Technology, Division of Water Resources Engineering, Lund University, SE 221 00 Lund, Sweden*
[f] *Norrvatten AB, Skogsbacken 6, SE 172 41 Sundbyberg, Sweden*
† Electronic supplementary information (ESI) available. See DOI: 10.1039/d1ew00416f
‡ Current affiliation: IVL Swedish Environmental Research Institute Ltd., SE 100 31 Stockholm, Sweden, E-mail: Claudia.Cascone@ivl.se.

industrial water monitoring. These sensors measure total light attenuation in water along a straight light path of defined length, due to it being absorbed by dissolved organic molecules or else scattered by particles.

The coloured or chromophoric fraction of dissolved organic matter (CDOM) is typically the main contributor to light attenuation in natural waters.[5] Although absorbance measurements do not quantify non-absorbing DOM fractions (including labile fractions with a deciding role in biostability), strong linear correlations ($r > 0.9$) between absorption coefficients and dissolved organic carbon (DOC) have been reported for various water bodies.[6–9] As described below, high concentrations of natural organic matter (NOM) in drinking water sources have many negative effects on treated water quality. This issue is gaining urgency because increased concentrations and fluctuations of NOM are occurring in boreal and north European surface waters, in connection with climate variations, reduced acid rain and increased primary production/standing biomass.[10,11]

Insufficient removal of NOM during drinking water treatment is connected to many issues: (i) poor taste and odour, (ii) insufficient removal of bacteria, viruses and parasites and/or bacterial regrowth, (iii) high rates of formation of potentially-carcinogenic disinfection by-products (DBP), due to the reaction of NOM with the disinfectant (*e.g.*, chlorine).[12,13] NOM also has a negative impact on the efficiency of treatment processes. Chlorine demand increases with NOM concentration, and its accumulation on the surface and/or pores of membranes contributes to their fouling, including by irreversible foulants. They cannot be removed by physical cleaning and backwashing but only by expensive chemical cleaning such as clean-in-place (CIP).

Organic matter fractions connected to humic substances (HSs) and biopolymers have been identified as contributors to irreversible fouling.[14] HSs is also the major fraction removed during coagulation, and since HS concentrations correlate well with the UV signal at 254 nm, UV absorbance data from online sensors can be used for real-time adjustments of coagulant dosing.[15,16] Additionally, differential UV absorbance at specific wavelengths (*e.g.*, 272 nm) correlates well with concentrations of DBPs formed after chlorination, so that absorbance-based sensors can be useful for DBP monitoring.[17,18]

The ratio of absorbance at two specific wavelengths ($A_{\lambda 1}/A_{\lambda 2}$) is often used to probe the sources and molecular properties of CDOM. Widely-used ratios have been reported to correlate negatively with aromaticity and molecular weight (MW, A_{250}/A_{365}), to reflect the relative amounts of autochthonous *versus* terrestrial CDOM (A_{254}/A_{436}), and to correlate negatively with the degree of humification (A_{300}/A_{400}).[6,19,20] Another absorbance ratio, A_{220}/A_{254}, correlates negatively with polarity, with higher values of this A_{220}/A_{254} ratio indicating CDOM is more difficult to remove through coagulation–flocculation.[21] Additional spectral metrics in common use include the exponential fit, the slope ratio (S_R) and the spectral slope curve (S_λ).

The UV-vis spectra are commonly modelled with an exponential decreasing function, as in eqn (1):[5,22,23]

$$a_\lambda = a_0 e^{S_e(\lambda-\lambda_0)} + K \quad (1)$$

with a_λ = absorbance value [m^{-1}] at a certain wavelength, a_0 = absorbance value [m^{-1}] at the reference wavelength λ_0, S_e = slope coefficient [nm^{-1}] and K = background constant to offset the baseline shift or attenuation not due to CDOM ("self-absorption"). The amplitude a_0 and the slope S_e are often used as a proxy for concentration and for changes in composition of CDOM, respectively.[24]

S_R is the ratio of the slope at shorter wavelengths ($S_{275-295}$) to the slope at longer wavelengths ($S_{350-400}$). Slope values in the ratio $S_{275-295}/S_{350-400}$ are computed using linear regression of the natural log transformed absorbance spectra. Larger slopes indicate a faster decrease in absorbance with increasing wavelength,[23] which might be used to detect larger changes occurring at shorter wavelengths (275–295 nm) compared to longer wavelength (350–400) or *vice versa*. $S_{275-295}$ is sometimes used to estimate photodegradation. Similar to the ratio A_{250}/A_{365}, S_R negatively correlates to CDOM MW.[20,23]

S_λ is computed from the linear regression of the logarithm of the absorbance spectra over a sliding window applied to the wavelengths.[25] S_λ is the spectral slope (the slope of the linear regression) as function of the wavelength (spectral slope curve) and is used to investigate CDOM biogeochemical processes and sources.[26] In general, various metrics appear to be more or less useful in different studies, and it is necessary to examine the behaviour of a range of different metrics during the data exploration phase.

Sensors with high time-resolution allow for tracking rapid changes in water quality and can be integrated into existing supervisory control and data acquisition (SCADA) systems. Membranes are increasingly common at DWTPs, and their effective maintenance requires more highly time-resolved data (on the order of seconds) than for classical treatment processes like coagulation–flocculation. Due to the large amounts of data this generates, DWTPs store only truncated/summarised datasets. In the specific case of absorbance-based sensors, raw data are typically discarded in favour of physical and chemical parameters (*e.g.*, turbidity, DOC) estimated using proprietary algorithms, which risks that valuable information is inadvertently discarded or misinterpreted. A small selection of multispectral CDOM sensors are currently available on commercial markets (*e.g.*, ProPS-UV, Viper (TriOS)), among which the spectro::lyser (s::can Messtechnik GmbH) was used in this study. The spectro::lyser is a UV-vis spectrophotometer probe that measures at a given time-interval attenuated light ("apparent" absorbance, *i.e.*, attenuation measurements due to absorbance and light scattering) in the ultraviolet and visible wavelength range. Published studies involving these instruments typically focus

on using spectral data as proxies for predicting DOC, nutrients or turbidity rather than on interpreting the spectral CDOM data in its own right.[7,27,28]

The aim of this study was three-fold:

1) Identify the main hurdles affecting the processing and interpretation of high-frequency datasets from online absorbance sensors.

2) Develop an open-source toolbox containing routines to efficiently process and visualise absorbance sensor datasets, producing metrics that address drift, random error and redundancy without discarding valuable information.

3) Demonstrate the application of these routines at a drinking water treatment plant, using a sensor dataset to detect anomalies and explain fluctuations in plant performance.

In line with available open source and commercial toolboxes that target the preprocessing and visualisation of non-spectral sensor data[29,30] or that compute metrics from absorption spectra of CDOM,[31] we introduce the AbspectroscoPY toolbox, an open-source toolbox for Python which combines preprocessing operations with specialised spectral analysis of CDOM. Processing is largely automated and requires only a few user-specified input parameters. The toolbox is easily adapted to accommodate other instrument outputs (*e.g.*, turbidity and other sensors where the data are contained in a vector instead of a matrix) across environmental research and management disciplines (*e.g.*, water quality monitoring, colour in aqueous solutions, wastewater, watersheds).[7,27,32] AbspectroscoPY currently contains 13 functions for importing, preprocessing, exploring and analysing absorbance-based sensor data and can be expanded by later users as necessary.

It can be downloaded from GitHub (https://github.com/ClaCasc/AbspectroscoPY), along with an example dataset that can be used to test and explore the functions.

In this paper we provide a tutorial to guide the user through the AbspectroscoPY toolbox, using a case study of a drinking water dataset.

2. Study site and water quality analysis

The drinking water dataset consists of light attenuation measurements collected by three online spectro::lyser spectrophotometers deployed for more than a year (2017–2018) at VIVAB's Kvarnagården DWTP in western Sweden. Fig. 1 shows the full-scale treatment process and placement of the three spectro::lyser units, which coincide with the positions where grab samples were taken during the period March–December 2018. Fig. 1 also reports an example of the obtained fingerprint file from one of the spectro::lyser units with raw attenuation measurements in the UV-vis wavelength range.

The surface water source at the DWTP is Lake Neden, a 3 km^2 slightly acidic (SW, pH 6.7, σ = 60 μS cm^{-1}) oligotrophic lake, surrounded by mixed woodland with an approximately five-year turnover time.[16] With respect to other lakes in the area, Lake Neden is characterised by clear water, low in total and dissolved organic carbon (TOC and DOC, 3.5 mg L^{-1}) and with intermediate specific ultraviolet absorbance (SUVA, 3.2 L mg^{-1} m^{-1}) which indicates a mixture of hydrophobic and hydrophilic fractions of different MW (Table 1). Along the pipeline that transports the water to the DWTP, the water from an alkaline groundwater well (GW, pH 8, σ = 60 μS cm^{-1}, TOC = 0.6 mg L^{-1}) is added to the water from the lake (20% GW/80% SW with 5% variation, *i.e.*, 15% GW/85% SW to 25% GW/75% SW).[16] This results in an incoming water to the DWTP containing relatively low DOC concentrations (~2.9 mg L^{-1}) and SUVA of *circa* 3.1 L mg^{-1} m^{-1} (Table 1).

Timestamp	220 nm	222.5 nm	225 nm	227.5 nm	230 nm
2017-09-29 15:36	20.0779	17.9894	15.9301	14.2964	13.3452
2017-09-29 15:38	20.0547	17.9778	15.9331	14.3061	13.3529
2017-09-29 15:40	20.0654	17.9957	15.9405	14.3142	13.3599

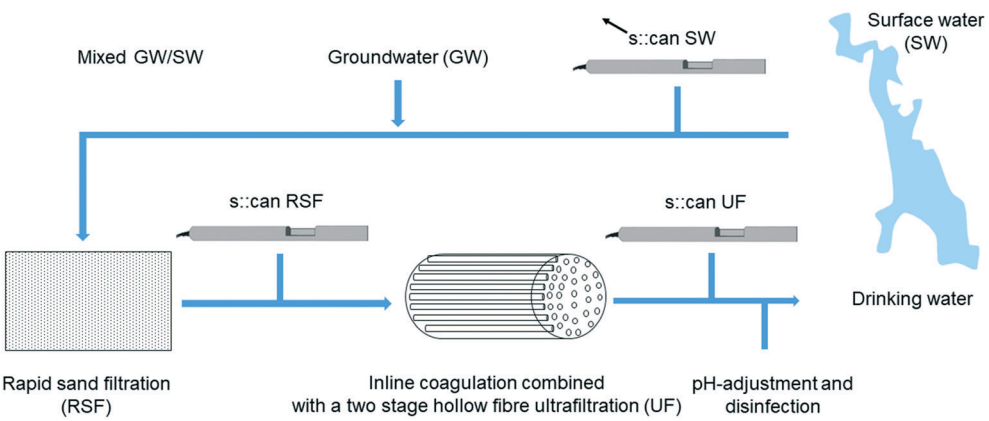

Fig. 1 Treatment steps for the full-scale process at Kvarnagården DWTP, Varberg, Sweden, and placement of the three online spectrophotometers (spectro::lyser, s::can Messtechnik GmbH). The table is an example of the obtained fingerprint file with raw attenuation measurements [absorbance per meter] for a few wavelengths at 2.5 nm interval in the range 200–750 nm. The grab sampling locations coincide with the positions of the spectro::lyser units.

Table 1 Water quality data reported as median value and interquartile range (IQR) of data collected during the period March–December 2018 on n sampling occasions for surface water (SW), rapid sand filtrate (RSF) and ultrafilter permeate (UF). The dilution effect of the groundwater mixed with the surface water (20% GW/80% SW) needs to be considered when evaluating the differences between SW and RSF. The parameters selected include total organic carbon (TOC), dissolved organic carbon (DOC), ultraviolet absorbance at 254 nm (UV_{254}) unfiltered and filtered, specific ultraviolet absorbance (SUVA), humification index (HIX), fluorescence index (FI), freshness index ($\beta:\alpha$), temperature and turbidity

		SW (n = 11)		RSF (n = 16)		UF (n = 16)	
Parameter	Unit	Median	IQR	Median	IQR	Median	IQR
TOC	mg L^{-1}	3.53	0.18	2.96	0.29	2.09	0.13
DOC	mg L^{-1}	3.54	0.21	2.93	0.19	2.08	0.16
[a]$UV_{254unfiltered}$	—[b]	11.8	1.0	9.2	0.6	4.1	0.3
$UV_{254filtered}$	—[b]	11.1	0.7	9.0	1.1	4.4	0.4
SUVA	L mg^{-1} m^{-1}	3.2	0.2	3.0	0.3	2.0	0.3
[c]HIX	—	0.92	0.01	0.92	0.01	0.89	0.02
[c]FI	—	1.44	0.02	1.43	0.03	1.57	0.02
[c]$\beta:\alpha$	—	0.55	0.01	0.54	0.01	0.65	0.02
[a]Temperature	°C	5.0	0.6	7.0	0.9	6.6	0.7
[a]Turbidity	FNU	0.25	0.07	0.18	0.06	0.05	0.03

[a] Measured on-site. [b] Absorbance per meter. [c] HIX – Ex: 254, Em: $\sum$(435–480)/($\sum$(300–345) + $\sum$(435–480)).[36] FI – Ex: 370, Em: 470/520.[37] $\beta:\alpha$ – Ex: 310, Em: 380/max(420–435).[38]

At the plant, the treatment process consists of rapid sand filtration, a polyethersulfone hollow fibre ultrafiltration membrane process with in-line coagulation using prepolymerized polyaluminum chloride, pH-adjustment with addition of $Ca(OH)_2/CO_2$, and disinfection with UV irradiation and addition of NH_2Cl. Further details on the treatment process at Kvarnagården DWTP are published elsewhere.[33]

2.1. Organic matter quantification and characterisation

Systematic drift is a common problem affecting sensors, so it is important to calibrate and periodically validate sensor data against grab samples. The grab samples in this study were analysed at the DWTP's own laboratory (unfiltered UV absorbance [Hach DR 5000], temperature and turbidity [Hach 2100N IS]) or at the Swedish University of Agricultural Sciences, SLU (TOC/DOC, filtered UV absorbance, fluorescence) after filtration (pre-combusted glass microfiber filters, GF/F, with a 0.7 μm nominal pore size).

TOC and DOC were measured with a TOC-V_{CPH} carbon analyser (Shimadzu) and DOC had an average coefficient of variation (CV) for replicate measurements of 0.7%. UV absorbance was measured at 254 nm using an AvaSpec-ULS3648 high resolution spectrophotometer (Avantes) in a 5 cm quartz cuvette with CV below 1%. SUVA values were calculated by normalizing the absorbance at 254 nm (UV_{254}) to the DOC concentration.

Fluorescence was measured using an Aqualog spectrofluorometer (Horiba Jobin Yvon) with a 1 cm quartz cuvette connected to a ASX-260 auto sampler (CETAC). The resulting fluorescence excitation emission matrices (EEMs) were preprocessed as discussed by Lavonen and co-workers.[34]

External standards were analysed for quality assurance with each batch of samples (TOC/DOC: ethylenediaminetetraacetic acid, EDTA, 10 mg L^{-1}; absorbance: K-phthalate, 10 mg L^{-1}).[7]

Table 1 displays median value and interquartile range of water quality data from grab samples collected in 2018 from surface water (SW, 11 sampling occasions), rapid sand filtrate (RSF, 16) and ultrafilter permeate (UF, 16). When interpreting differences in water quality between SW and RSF, it is important to account for the dilution with groundwater. Fluorescence indices suggest that the mixing with groundwater did not significantly affect the composition of fluorescent dissolved organic matter (fDOM) in the water in the range of wavelengths used to calculate the indices. Coagulant dosing is controlled in real-time based on attenuation, colour and turbidity measurements from spectro::lyser units located in the sand filtrate and in the permeate.[16] This results in permeate with more stable water quality than would occur without such a control system in place.[35]

2.2. Online spectrophotometer units

The sensors provide attenuation data at excitation wavelengths ranging from 200 to 750 nm at 2.5 nm intervals. Since all sensors were deployed *in situ*, particles could have contributed to apparent absorbance measurements, especially in the surface water where turbidity was greatest.[7]

Measurements were taken every two minutes in SW and every three minutes in RSF and UF. Data were adjusted internally to the correct path length, *i.e.*, 35 mm for the sensors located in the water source and before the ultrafiltration, and 100 mm for the sensor located after the ultrafiltration. During the sampling period local calibrations were performed on the two sensors located in the DWTP. All sensors were subject to regular cleaning and maintenance.[16]

3. AbspectroscoPY: approach, application and evaluation

This section aims to guide the user through the AbspectroscoPY toolbox. We start with an overview of the general data analysis challenges, introduce the specific

toolbox functions created to address these challenges, and end with a discussion of their application for interpreting the case study dataset.

Real-time measurements lead to very large datasets that are challenging to preprocess, visualise and interpret. Pre-treatment typically includes identifying and removing or downweighting erroneous data, including scatter and outliers. When merging datasets from different sensors, further challenges arise when there are mismatching time axes. AbspectroscoPY contains functions for importing, preprocessing and exploring the sensor data as well as plotting spectral metrics to facilitate interpretation (Table 2).

3.1. Import the data files

It is important to download sensor data frequently to prevent data from being over-written, since high-frequency measurements rapidly consume memory. The data can be exported from the instrument and saved as csv-files or preferably as text files. These files can be imported with a function that merges a list of consecutive measurement files into a single dataset (***abs_read***).

For the spectro::lyser, data can be exported with either the Ana::pro software or a spreadsheet program such as Microsoft Excel. In this study, the datasets were *ca.* 0.4–0.5 GB per sensor ($\approx 3 \times 10^5$ measurements × 200 wavelengths).

3.2. Preprocess the dataset

Preprocessing functions in the toolbox are used to prepare the data for plotting.

3.2.1. Assess data quality. The toolbox includes functions to convert the data to the correct category of values (data type) for analysis (***convert2dtype***) and to improve the data quality.

It is possible to handle both missing data (NaN entries, ***nan_check*** and ***dropna***) and duplicates (***dup_check*** and ***drop_duplicates***). Missing data and duplicates are identified and dropped. Dropping missing data should not result in noticeable data loss as long as sampling frequency exceeds the frequency of significant water quality events. Handling duplicates requires caution with interpreting timestamps, since some (but not all) sensors adjust for daylight saving time (DST). For this reason, when removing duplicated data based on timestamp alone, it is important to check dates carefully to avoid deleting data by accident.

3.2.2. Shifted time-axis. Time-series data from different instruments needs to be aligned correctly before their signals can be compared. Even with sensors of the same type, it is crucial to verify that both instruments have comparable time axes. Instruments may have been set up differently in terms of how they treat daylight saving time (DST) or may have systematic time shifts, as in the example in Tables 3 and A1 (ESI†).

The following procedure is recommended:

1. Check whether the sensor automatically adjusts for DST when saving a timestamp; if so, consider shifting the time axis to produce a continuous time-series (***tshift_dst***);

2. Check for other systematic shifts from the local time, for example errors when setting the instrument's internal clock; if these exist then correct the dataset accordingly (***timedelta***).

Table 2 List of analytical steps and substeps implemented in the toolbox AbspectroscoPY and explanation of the aim of the functions

AbspectroscoPY			
Analytical step	Analytical substep	Function name	Aim of the function
Import raw data files	Dataset assembly	abs_read	Import a list of attenuation data files as function of time
Preprocess the dataset	Data type conversion	convert2dtype	Convert one or more categories of values to a different one
	Data quality assessment	nan_check	Quantify missing data in rows and columns
		[b]dropna	Drop rows or columns containing only missing data
		dup_check	Check the occurrence of duplicates
		[b]drop_duplicates	Drop rows or columns which are duplicates
	[a]Time-axis shifting	tshift_dst	Shift the dataset in time one hour forward when the daylight saving time ends
		[b]timedelta	Shift the dataset in time
	Attenuation data correction	abs_pathcor	Correct the attenuation data according to path length
		abs_basecor	Subtract the baseline from the attenuation data
	[a]Data smoothing	[b]rolling	Smooth the absorbance data using a moving median filter
Explore the dataset	Visualisation of data distribution	[b]kdeplot	Visualise the data distribution using Gaussian KDE plot
	[a]Outlier/event identification and removal	outlier_id_drop_iqr	Identify potential outliers and events based on the interquartile (IQR) thresholding strategy and drop them
		outlier_id_drop	Label outliers and events based on user knowledge and drop them
Interpret the results	Absorbance ratios	abs_ratio	Calculate the ratio of absorbance data at two different wavelengths
	Absorbance spectra changes	abs_fit_exponential	Fit an exponential curve to the absorbance data
		abs_slope_ratio	Calculate the slope ratio
		abs_spectral_curve	Generate the spectral slope curve

[a] User decision. [b] Python built-in functions.

Table 3 Difference in time between the time information displayed on the three spectro::lyser units ($t_{s::can}$) in Fig. 1 and the local time (t_{CEST}) for two specific dates during the periods of daylight saving time (DST, 03/10/2018) and standard time (ST, 27/11/2018). This information is required to use the functions ***tshift_dst*** and ***timedelta*** in the AbspectroscoPY toolbox. The table is an example of how to prove whether different sensors in surface water (SW), rapid sand filtrate (RSF) and ultrafilter permeate (UF) take in account DST and show any systematic shift from the local time. Table A1 (ESI†) reports how to account for the differences in time of the sensors in SW, RSF and UF

		Time [hh:mm:ss]				
Sample	Period	$t_{s::can}$	t_{CEST}	Δt_{tot}	Δt_{DST}	$\Delta t_{s::can}$
SW	DST	06:48:00	08:16:00	− 01:28:00	− 01:00:00	− 00:28:00
	ST	07:38:00	08:06:00	− 00:28:00		
RSF	DST	10:06:00	09:23:00	+ 00:43:00	00:00:00	+ 00:43:00
	ST	10:25:00	09:42:00	+ 00:43:00		
UF	DST	10:05:00	09:21:00	+ 00:44:00	00:00:00	+ 00:44:00
	ST	10:23:00	09:39:00	+ 00:44:00		

Δt_{tot} = total time difference ($t_{s::can} - t_{CEST}$). Δt_{DST} = time difference due to DST ($\Delta t_{tot\ DST} - \Delta t_{tot\ ST}$). $\Delta t_{s::can}$ = time difference due to other reasons ($\Delta t_{tot\ DST} - \Delta t_{DST}$).

If working with more than one sensor and the aim is to compare across sensors, it is important to:

3. Synchronise the clocks, by defining one sensor as a reference and shifting the time axes of all other sensors accordingly (***timedelta***);

4. Account for time lags while water travels between two sensors, by using one sensor's time axis as a reference, then correcting the timestamp of the other sensors to account for the lag (***timedelta***).

The toolbox allows the user to perform time alignment even when the degree of time lag changes over time; time alignment is essential for understanding whether an event in one part of the treatment plant is attributable to something that occurred at an earlier stage. For example, a change in attenuation data detected by the sensor in RSF can be due to altered coagulant dose, in response to attenuation data measured by the sensor in SW.

Table 3 illustrates an example of correcting the time lag between the internal clock ($t_{s::can}$) of the three spectro::lyser units in Fig. 1 and the local time (t_{CEST}) during periods of DST and standard time (ST). The two sensors in the DWTP automatically adjusted for DST, unlike the sensor in SW, as shown from the constant time difference between the internal clock and the local time during DST and ST periods. Therefore, according to step 1 in the procedure, the time axis for the DWTP sensors was shifted forward by 1-hour after the summertime ended. These three sensors also had systematic offsets from the local time unrelated to DST and, in line with step 2, their time axes were each shifted accordingly. Table A1† indicates how to quantify the time lag between the three sensors using the time shifted datasets. According to step 3, the time axis of the sensor in SW was shifted forward by 1-hour. Additionally, using user knowledge of the time taken for a parcel of water to travel between the SW and DWTP sites, the time axis of the sensor in SW was shifted forward by 11-hours in line with step 4.

In cases where data frequencies vary between sensors, a decision must be made about whether to interpolate low-frequency data or conversely, discard some high-frequency data. For example, at Kvarnagården DWTP the transmembrane pressure (TMP) which tracks membrane permeability is measured every 5 s whereas absorbance is measured every 3 minutes. Whether it is preferable to interpolate or discard data depends on the measurement frequency in relation to the time scale of actionable changes in the observed data. If after discarding data the measurement frequency is high compared to the how quickly the spectral data change, then it was probably safe to discard. If not, then it might have been better to interpolate. Either way, interpolation will be most accurate when applied to data that change either slowly or predictably; for example, by following a cyclic pattern that can be modelled during interpolation.

3.2.3. Correct attenuation data. Despite careful sensor calibration, signal output may drift over time affecting the interpretation of the dataset. For this reason, post-calibration of the instruments should be performed, especially when the user suspects systematic deviations. For the absorbance spectrophotometers in this study, the signal is internally calibrated using a dual beam which minimises instrument electronic drift but not the optical drift (*i.e.*, scratched windows, insufficient cleaning). This problem can be addressed by performing the baseline correction.

The AbspectroscoPY toolbox contains several functions for correcting the attenuation data of the clean and aligned dataset. First, the data may need to be normalised by the optical path length (***abs_pathcor***) unless this happens automatically as for the spectro::lyser. Then the median of the absorbance values at a chosen wavelength range (in our example, 700–735.5 nm, but a different range can be set, ***abs_basecor***) is subtracted from the absorbance data to account for the instrumental baseline drift.[26] The toolbox allows for visualising the median and the noise level (three standard deviations). At wavelengths above 700 nm, absorbance from CDOM and chlorophyll is negligible and signals are due to turbidity combined with random electronic noise.[39,40] By averaging across a range of wavelengths, the random noise is removed, leaving only turbidity. To determine an appropriate wavelength range for the baseline, the attenuation spectra should be plotted for a range of samples (covering the temporal variability of the data) and

checking their shift from zero. If baseline shifts occur they can be handled with this function, which can be applied to either the whole dataset or specific portions of it. In addition to this, this function allows to multiply/sum/subtract the whole dataset or part of it by a certain value to perform necessary calibrations or to account for interferences of anions and cations (*e.g.*, nitrate, iron[20]).

For the DWTP example in this paper, it is relevant to examine whether there may be systematic biases in apparent absorbance measured by the sensor, compared with apparent (unfiltered) absorbance measured using a desktop spectrophotometer. Fig. A1† shows the unfiltered UV_{254} data from grab samples (*x*-axis) for SW *versus* the scatterplot of the UV_{255} data from the spectro::lyser (*y*-axis; due to the 2.5 nm wavelength resolution this is the nearest wavelength to UV_{254}). Considering the instrumental error of the laboratory analyses, the data from the sensor seem to be slightly biased.

Once the data reliability is assessed, the next step is to visualise the data. Fig. 2 shows the plot of the preprocessed time-series of the UV absorbance values at 255 nm from the three spectro::lyser units indicated in Fig. 1. Five periods were distinguished using the SW time-series as reference and taking into account that Lake Neden is a dimictic lake: a comparatively stable period (P1, end of summer stagnation), two periods with considerable temporal fluctuation (P2 and P5, autumn circulation, Fig. A2, ESI†) and two periods with increasing and decreasing absorbance trends (P3, end of autumn circulation and winter stagnation and P4, spring circulation and summer stagnation, respectively). Three events related to changes in the lake and adjustment of the coagulant dosing in the DWTP are indicated by the arrows in Fig. 2 (compared to Fig. A3, ESI†). Events 1 and 3 are caused by the autumn lake circulation in two consecutive years. Event 2 indicates a challenging period for the DWTP in connection with the spring lake circulation, characterised by a prolonged period of decreasing membrane permeability that ultimately required CIP of the UF membrane.

3.2.4. Smooth noisy data. Python has a number of built-in functions to smooth data and reduce noise variability (*e.g.*, *rolling*, *lowess*). Herein we demonstrate the use of a median filtering using the function ***rolling***. Median filtering is a simple and robust smoothing technique that works well when there are sporadic outliers. The user specifies a window size for the median filter, depending upon data frequency and the aim of the filtering. With median filtering, it is essential to visualize the data to decide on an appropriate smoothing window. A smaller window size leads to noisy data but it is preferred to keep narrow spikes whereas a larger window will smooth out cyclical peaks, to emphasize trends rather than oscillations. It is probably better to under-smooth than over-smooth to avoid removing important information.

Outliers in sensor datasets may be caused events of interest for deeper study, in which case they need to be retained (*e.g.*, abrupt changes in coagulant dosing, Fig. A4, ESI†) or known artefacts that are easily identified and can be ignored (*e.g.*, maintenance operations of membranes and sensors). Additional methods for handling outliers are discussed in section 3.3.

Fig. 3 demonstrates the application of the smoothing function to the data in Fig. 2 period P1. A 60-min window size was chosen since it is wide enough to capture both the trend and oscillations. Raw RSF data feature daily cycles often with a double peak, probably related to changes in flow rate due to changes in demand. The UF data show a cyclic behavior due to backwashing cycles which occur approximately every two hours. The UF signal also reports narrow spikes that are smoothed out by using a 60-min window for the rolling median filter. A smaller window size of 15-min will retain these features in the filtered signal.

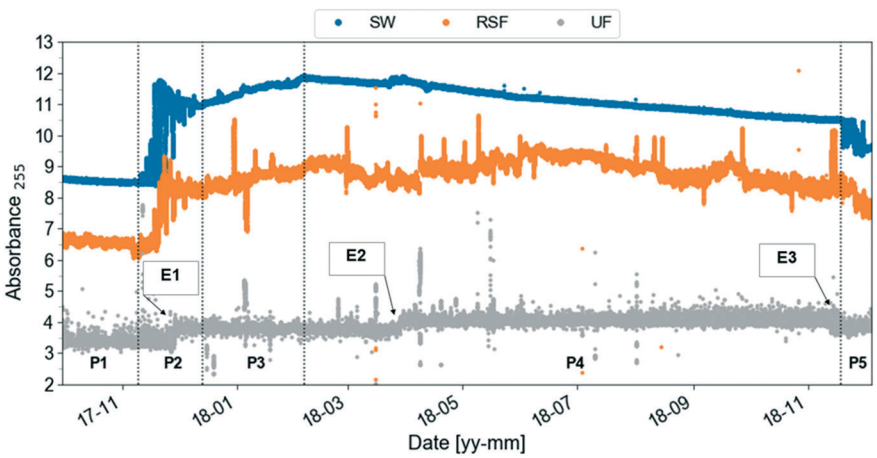

Fig. 2 Preprocessed UV absorbance at 255 nm (absorbance per meter) time-series obtained from the spectro::lysers in surface water (SW, frequency of sampling, 2 min), rapid sand filtrate and ultrafilter permeate (RSF, UF, frequency of sampling, 3 min) in the period September 2017–December 2018. Five periods (P) are identified using the surface water time-series as reference: each period is defined by two consecutive vertical dashed lines. Three events related to changes in the lake and adjustment of the coagulant dosing at Kvarnagården DWTP are indicated by the arrows: events 1 and 3 are caused by the autumn lake circulation in two consecutive years. Event 2 indicates the starting point of a prolonged period of decrease in membrane permeability lasting until June 2018. Compare to Fig. A3 (ESI†).

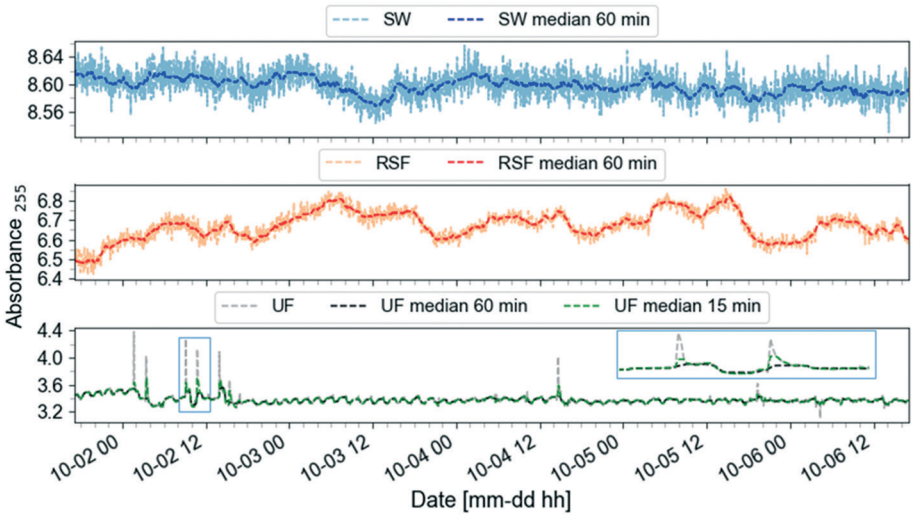

Fig. 3 Rolling median smoothing function with a 60-min window size applied on UV absorbance at 255 nm (absorbance per meter) time-series (October 2017) of surface water (SW), rapid sand filtrate (RSF) and ultrafilter permeate (UF). Comparison between specifying a window size of 60-min (black) *vs.* 15-min (green) in the function *rolling* in the AbspectroscoPY toolbox for the UF time-series, with a close-up showing narrow spikes in the light blue rectangle.

3.3. Explore the dataset

Several functions for exploring the dataset are included in the toolbox.

3.3.1. Identify and remove outliers. Outliers in the data can be labelled using user defined events and outliers associated with specific event categories can be automatically removed (***outlier_id_drop***). For example, for membrane benchmarking it is important to exclude periods when performance deviations are explained by extrinsic factors such as power outages or unscheduled maintenance work. High quality records of WTP operations such as maintenance of the sensor or the plant, *e.g.*, using a logbook, can give valuable information to help distinguish between artefacts and anomalies in the data.

Fig. 4 shows an example of application of the ***outlier_id_drop*** function to the SF and UF absorbance data in Fig. 2. Symbols on the plot indicate times when there was no feed water to RSF and UF (no feed event, data not shown; these data for RSF were not available before June 2018) and coagulant dose was changed (Al dose event, Fig. A3, ESI†). Symbols indicate the approximate location of the event in time for visualisation purposes. To label known events, the user needs to specify in a csv-file the start and end dates, the type of event and its label reference. The event can then be dropped using the label reference (Fig. A5, ESI†).

Functions to identify potential outliers and unexplained events and potentially to remove them (***outlier_id_drop_iqr***) are provided. The user first needs to specify periods (*e.g.*, P1–P5 in Fig. 2) then outlier identification is based on the

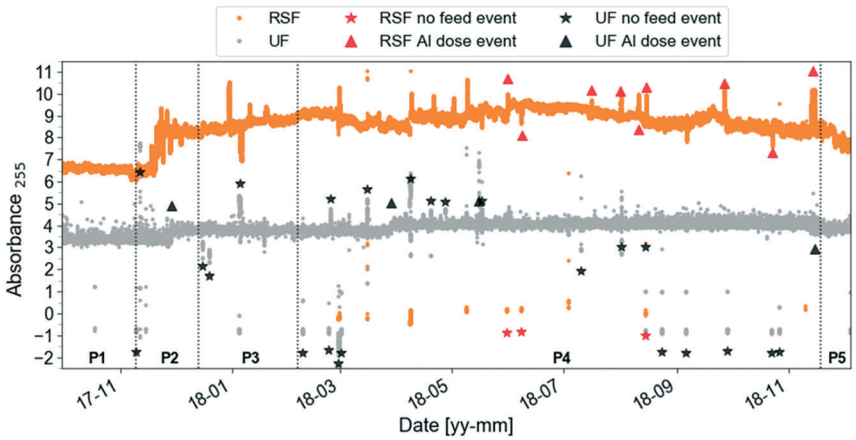

Fig. 4 Same preprocessed UV absorbance at 255 nm (absorbance per meter) time-series as in Fig. 2 (zoomed out) with two types of events labelled by the user using the function ***outlier_id_drop*** in the AbspectroscoPY toolbox for rapid sand filtrate (RSF) and ultrafilter permeate (UF). The symbol identifies the whole event period, using the average timestamp of the event as *x*-axis coordinate and the median absorbance value at 255 nm plus–minus one absorbance unit offset as *y*-axis coordinate.

interquartile (IQR) thresholding strategy. The multiplication factor for IQR was set to 1.5.[41] The IQR method was tested on slope ratio data since slopes are sensitive to outliers.

The slope ratio data in this case were obtained from the SW absorbance data preprocessed as in 3.2 except for baseline correction and median smoothing and on the fully preprocessed dataset (Fig. A6, ESI†). The data indicate that the slope ratios for period P1 are statistically different from periods P3 and P4.

3.3.2. Visualise data distribution. The kernel density estimate (KDE) is an approach to estimate the underlying probability density function of a dataset, similar to a histogram, but with greater flexibility due to the possibility to calculate it differently by specifying different kernel types. The built-in Python function ***kdeplot*** assumes an underlying Gaussian distribution at the location of each data point. In Fig. A7 (ESI†), it is used to visualise how the distribution of absorbance values varies in terms of density (height of the curve at each point) when the observation wavelength is changed.[42]

KDE plots of RSF and UF data showed sharper peaks than SW, indicating a smaller range of absorbance measurements, and each wavelength shorter than 327.5 nm had a three-pointed distribution. This is a natural consequence of the automatic coagulant dosing at the DWTP that aims to reach specific UF permeability targets. It shows that three distinct permeability targets were applied in the DWTP, resulting in step changes in water quality (compare Fig. A7 to Fig. A5, ESI†).

3.4. Interpret the results

Once the data are cleaned and ready for analysis, AbspectroscoPY provides tools to investigate spectral changes. Here, the aim is to identify typical profiles and detect spectral anomalies related to changes in organic matter character. In our DWTP example, the autumn lake circulation is an example of such an anomaly. Similar to the "cdom" package for the R software environment,[31] functions to calculate common metrics from absorbance spectra of CDOM are implemented in the AbspectroscoPY toolbox, including S, S_R and S_λ, as well as ratios between absorbance values at specific wavelengths.

3.4.1. Absorbance ratios. In order to investigate the sources and molecular properties of CDOM, a well-known metric is the ratio of absorbance at two specific wavelengths ($A_{\lambda 1}/A_{\lambda 2}$) which is calculated with the algorithm (***abs_ratio***).

For the current dataset, it was interesting to compare the maximum change of absorbance ratios (in percent, using averaged values of the last week of period P4) to the averaged values of absorbance ratios on the first week of period P4. This gave a maximum increase of 5.4%, 16.8%, 3.1% and 7.1% in period P4 for the ratios A_{250}/A_{365}, A_{254}/A_{436}, A_{300}/A_{400} and A_{220}/A_{254}, respectively. Behaviour of the ratios A_{250}/A_{365}, A_{254}/A_{436} and A_{300}/A_{400} were consistent with each other, suggesting a decrease of aromaticity and MW of CDOM and an increase of the relative abundance of autochthonous *versus* terrestrial CDOM during period P4. During the same period the results obtained for the ratio A_{220}/A_{254} pointed to a decrease of polarity that suggested that DOM would be more difficult to remove. These findings are in accordance with other studies of Swedish surface waters. For instance, in Lake Tämnaren the ratio A_{250}/A_{365} increased during the summer period reaching its maximum values in September[26] and in the river Fyris, the fDOM also decreased during the spring–summer period.[7] This was attributed to the shift of MW distribution to lower MW by photodegradation.[23,43]

Considering the A_{254}/A_{436} ratio in Fig. A8 (ESI†), the ratio showed an abrupt increase at the end of March and middle of June 2018 coinciding with the spring circulation of the lake. This signal was more prominent when using longer wavelengths in the ratio (*e.g.*, compare A_{250}/A_{365} and A_{254}/A_{436} ratios in Fig. A8, ESI†). The sudden increase in this period indicated a sudden increase of autochthonous CDOM. During the same period, event 3 (decrease in membrane permeability) occurred in the DWTP.

3.4.2. Exponential fits. Fig. A9 (ESI†) shows an example of fitting the absorbance spectra from the spectro::lyser in SW to a single exponential decay function at a specific date (***abs_fit_exponential***) at the reference wavelength 350 nm, according to eqn (1); this model is dependent on the wavelength range used in the fit.[24]

3.4.3. Slope ratio. Fig. A6 (ESI†) shows the slope ratio time-series in SW (***abs_slope_ratio***). The decrease of S_R during periods P2 and P3 compared to period P1 indicated that SW was mainly composed of terrestrial CDOM with higher MW. When comparing S_R to the time-series of the absorbance ratio A_{250}/A_{365} in Fig. A8 (ESI†), the two spectral metrics showed a similar trend during the periods P2, P3, P5 and beginning of P4. In contrast, during period P1 S_R displayed only a small increase during period P1 and during period P4 a quite continuous increase from April 2018 until reaching its maximum in August 2018. Over the same period, the ratio A_{250}/A_{365} showed a much larger increase during both period P1 and P4, with step increases during period P4.

3.4.4. Spectral slope curve. This study used a sliding window with a width of 21 nm, which is similar to previous studies,[25,31] applied to the wavelengths 220–697.5 nm at 1 nm resolution. Since the absorbance data from the spectro::lyser have a 2.5 nm resolution originally the data were resampled at 1 nm increments using a cubic spline interpolation[31] and then filtered using a correlation coefficient threshold of R^2 of 0.98 (***abs_spectral_curve***). Instead of the original negative slope, we report the absolute value of the slopes since positive numbers are easier to discuss. Since absorbance slopes are generally negative, this does not introduce ambiguity. The absorbance is constant at high wavelengths throughout (*i.e.*, there is no translation over time), and therefore all variations of the absorbance curves (in both shape and magnitude) are directly reflected in the data for the spectral slope curve. The spectral slope curve analysis allows for a much easier identification of the wavelength regions where greatest variability occurs in

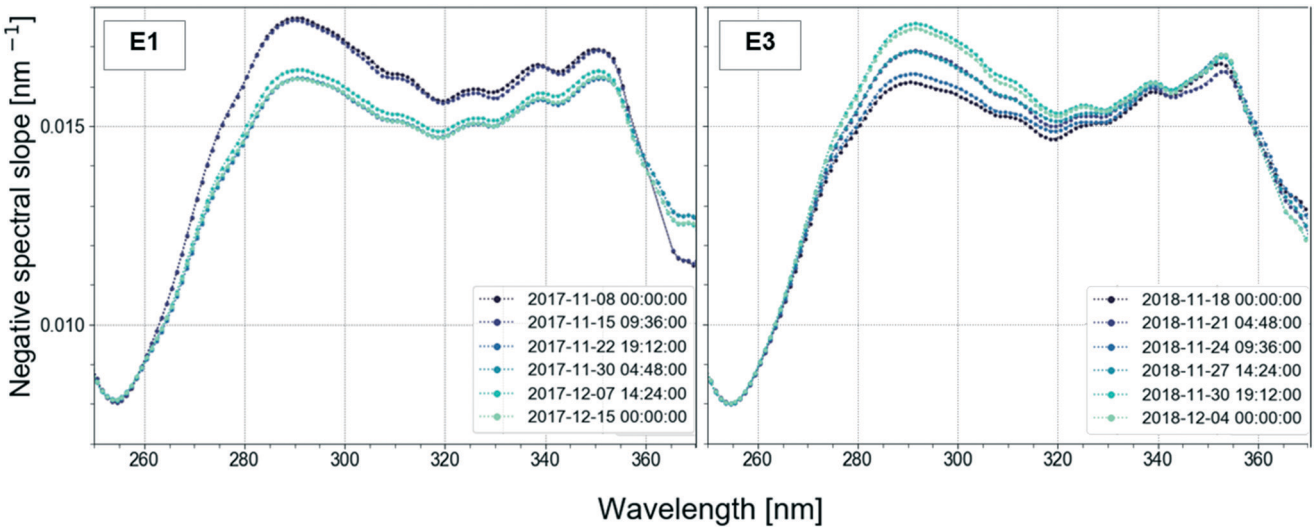

Fig. 5 Spectral slope curve of the spectro::lyser absorbance data in surface water as a function of wavelength calculated using the ***abs_spectral_curve*** function in the AbspectroscoPY toolbox. The selected dates cover the period before, during and after the autumn lake circulation event in 2017 (left plot, November–December, event 1) and 2018 (right plot, November–December, event 3).

comparison to the analysis of absolute changes of absorbance (Fig. A10, ESI†).

In the study, the aim was to compare a typical profile to the autumn lake circulation (events 1 and 3). First for these events, the largest change of the spectral slope was observed at 290.5 nm (Fig. 5). Then, the variation in spectral slope was computed at that wavelength over the course of the two lake circulation events. In order to have a reference of a typical profile, the same analysis was repeated for periods without events throughout the year for the same time interval. In 2017, event 1 was associated with a 7.4% decrease of the slope at 290.5 nm over the duration of the event (*ca.* 5 weeks), while the slope increased by 8.3% during event 3 (*ca.* 2.5 weeks) in 2018. A typical slope variation over a 3-week period without events was well below 1%. Apart from the shift in the magnitude of the spectral slope in the wavelength range 270–350 nm during the circulation events indicating large changes in the absorbance, both in magnitude and shape, the overall variations of the profiles with the wavelength are similar in all periods. The low variability of the profile shape is probably due to the long residence time of Lake Neden.[20] In the period between the end of March and the middle of June 2018 (event 2), the spectral slope increased by 1.3%. Changes in spectral slope could be used to decide when to take grab samples in order to answer specific questions with more targeted analyses.

In addition to statistical tools included in the R-based *cdom* package, the AbspectroscoPY Python toolbox includes the possibility to obtain a time-series of the local information of the spectral slope curve, *i.e.*, the negative spectral slope at a specific wavelength (*e.g.*, 290.5 nm), using eqn (2):

$$\left[-\frac{\delta a}{\delta \lambda}|\lambda = 290.5\ \text{nm}\right](t) \quad (2)$$

The algorithm computes percentage changes in comparison to the averaged spectral slope results obtained on a reference day for a chosen wavelength. Fig. 6 shows percentage changes in spectral slopes in SW, RSF and UF for the lake circulation event in 2018. For the current dataset, profiles were similar for SW, RSF and UF except for a plateau in the UF data on November 12–17th 2018. This was probably caused by an abrupt increase in coagulant dosing (Fig. A11, ESI†).

Different wavelengths produce different views of spectral slope changes. Fig. A12 (ESI†) displays the time-series of spectral slope at 254.5 nm. Compared to the plot at 290.5 nm, variations were much less prominent. This might indicate a different removal of organic components at different wavelengths. The trends in the temporal variation of the spectral slope were very similar at 272.5 nm and 290.5 nm. Since it has been shown that the wavelength 272 nm is related to DBPs, the analysis of the time-series could be relevant for DBP monitoring and used as an early warning system.

3.5. Archive scripts, data and plots

Data can be exported from the toolbox as csv-files, or plots of desired format and resolution, using a range of scripts available on GitHub.

4. Conclusions

Absorbance (UV/vis) spectroscopy is widely used for monitoring natural organic matter in water treatment due its low cost, high sensitivity and speed. Sensors take this technique to the next level allowing for continuous measurements to catch rapid changes in water quality. However, large datasets need to be carefully preprocessed including *e.g.*, time axis correction, filtering and outlier

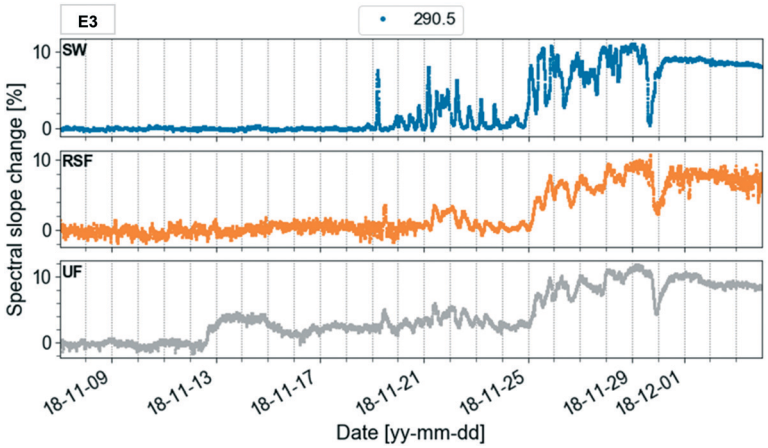

Fig. 6 Time-series of spectral slope percent changes in spectro::lyser absorbance data at 290.5 nm. The plots refer to surface water (SW), rapid sand filtrate (RSF) and ultrafilter permeate (UF). The selected dates cover the period before, during and after the autumn lake circulation event (November–December 2018, event 3).

identification. Thereafter, it is crucial to apply spectral metrics that facilitate and guide interpretation.

The Python toolbox AbspectroscoPY addresses some of the main issues that hamper the processing of sensor data, by handling duplicates, systematic time shifts, baseline correction and outliers. It also provides a selection of metrics for data interpretation including absorbance ratios, exponential fits, slope ratios and spectral slope curves. In addition, it contains functions to visualise changes in metrics over time. The general workflow includes elements such as:

a) Plot absorbance ratios to get an overview of time periods undergoing large changes in CDOM sources and molecular properties.

b) Compute the rate of change of absorbance with respect to wavelength (spectral slope) to detect wavelength ranges with significant temporal variability in the absorbance slopes. The analysis can be focused on periods based on (a) or periods of particular interest to the user *e.g.*, lake circulation events or decreases in membrane permeability.

c) For specific wavelength ranges identified in (b), plot the time-series of the spectral slope changes (%) to investigate the temporal evolution of the absorbance curves. The time-series could be used as an early warning system by identifying correlations with important events.

The AbspectroscoPY toolbox combines these tools in a general purpose open-source Python environment that can be applied to different data sources in a variety of fields, including drinking or wastewater treatment and the food industry.

The capabilities of the toolbox were showcased using optical sensor data collected at Kvarnagården WTP using Lake Neden as water source. Based on trends in the attenuation data, five different periods were identified in a dataset spanning 15 months that were well correlated with natural events in the lake such as seasonal circulation. Despite the very stable water quality, these events as well as changes in the WTP such as changes in the coagulant dosing or a decrease in membrane permeability can be detected using the spectral metrics provided in the toolbox.

New features can easily be added to the toolbox due to its open-source format, potentially including:

a) Particle compensation algorithms, for implementation wherever there are continuous turbidity measurements. Turbidity corrections increase the accuracy of absorbance measurements in surface waters.

b) Algorithms for subtracting the spectra of interfering compounds absorbing in the same wavelength range as DOM.

c) Advanced tools for outlier identification.

d) Algorithms to calculate indices that water producers can use as decision support tools, such as the absorbance slope index (ASI).[44]

Author contributions

CC, KRM, AK and SJK conceptualised the study. AK was responsible for resources, AK and SJK were in charge of funding acquisition. CC and CS were in charge of the investigation. CC, HM and JSK were responsible for the software development and validation. CC was in charge of data curation, formal analysis, methodology and visualisation. CC, KRM, JSK and SJK wrote the article. HM, CS and AK commented on draft versions of the article. All authors approve the final article.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We would like to thank the Geochemical laboratory at the Department of Aquatic Sciences and Assessment at the Swedish University of Agricultural Sciences, SLU. In particular, we would like to acknowledge Nilofar Åkerlund, Christian Demandt, Sofia

Firpo, Johannes Kikuchi, Ingrid Nygren and Karin Wallman for helping with laboratory analysis. CC, HM and SJK acknowledge funding by Svenskt Vatten (SVU 16-103), DRICKS (SVU 20-121) and VIVAB. KRM acknowledges funding by the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS grant 2017-00743).

References

1 T. Bartrand, W. Grayman and T. Haxton, *Drinking Water Treatment Source Water Early Warning System State of the Science Review*, U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-17/405, 2017.
2 C. H. Clausen, M. Dimaki, C. V. Bertelsen, G. E. Skands, R. Rodriguez-Trujillo, J. D. Thomsen and W. E. Svendsen, Bacteria Detection and Differentiation Using Impedance Flow Cytometry, *Sensors*, 2018, **18**, 3496.
3 C. Gruden, S. Skerlos and P. Adriaens, Flow cytometry for microbial sensing in environmental sustainability applications: current status and future prospects, *FEMS Microbiol. Ecol.*, 2004, **49**, 37–49.
4 K. Patil, S. Patil, M. Patil and M. Patil, Monitoring of Turbidity, PH & Temperature of Water Based on GSM, *International journal for research in emerging science and technology*, 2015, **2**, 16–21.
5 A. Bricaud, A. Morel and L. Prieur, Absorption by dissolved organic matter of the sea (yellow substance) in the UV and visible domains, *Limnol. Oceanogr.*, 1981, **26**, 43–53.
6 P. Li and J. Hur, Utilization of UV-Vis spectroscopy and related data analyses for dissolved organic matter (DOM) studies: A review, *Crit. Rev. Environ. Sci. Technol.*, 2017, **47**, 131–154.
7 S. Hoffmeister, K. R. Murphy, C. Cascone, J. L. J. Ledesma and S. J. Köhler, Evaluating the accuracy of two in situ optical sensors to estimate DOC concentrations for drinking water production, *Environ. Sci.: Water Res. Technol.*, 2020, **6**, 2891–2901.
8 E. I. Prest, F. Hammes, M. C. M. van Loosdrecht and J. S. Vrouwenvelder, Biological Stability of Drinking Water: Controlling Factors, Methods, and Challenges, *Front. Microbiol.*, 2016, **7**, 45.
9 A. Avagyan, B. R. K. Runkle and L. Kutzbach, Application of high-resolution spectral absorbance measurements to determine dissolved organic carbon concentration in remote areas, *J. Hydrol.*, 2014, **517**, 435–446.
10 A. Lepistö, P. Kortelainen and T. Mattsson, Increased organic C and N leaching in a northern boreal river basin in Finland, *Global Biogeochem. Cycles*, 2008, **22**, 1–10.
11 C. Forsberg and R. C. Petersen, A darkening of Swedish lakes due to increased humus inputs during the last 15 years, *SIL Proceedings, 1922-2010*, 1990, **24**, 289–292.
12 M.-G. Kang, Y.-H. Ku, Y.-K. Cho and M.-J. Yu, Variation of dissolved organic matter and microbial regrowth potential through drinking water treatment processes, *Water Sci. Technol.: Water Supply*, 2006, **6**, 57–66.
13 G. V. Korshin, W. W. Wu, M. M. Benjamin and O. Hemingway, Correlations between differential absorbance and the formation of individual DBPs, *Water Res.*, 2002, **36**, 3273–3282.
14 R. H. Peiris, H. Budman, C. Moresoli and R. L. Legge, Fluorescence-based fouling prediction and optimization of a membrane filtration process for drinking water treatment, *American Institute of Chemical Engineers*, 2012, **58**, 1475–1486.
15 S. J. Köhler, E. Lavonen, A. Keucken, P. Schmitt-Kopplin, T. Spanjer and K. Persson, Upgrading coagulation with hollow-fibre nanofiltration for improved organic matter removal during surface water treatment, *Water Res.*, 2016, **89**, 232–240.
16 A. Keucken, G. Heinicke, K. M. Persson and S. J. Köhler, Combined Coagulation and Ultrafiltration Process to Counteract Increasing NOM in Brown Surface Water, *Water*, 2017, **9**, 697.
17 G. Stéphanie and D. Caetano, Real-Time Estimation of Disinfection By-Products through Differential UV Absorbance, *Water*, 2020, **12**, 2536.
18 N. Beauchamp, C. Dorea, C. Bouchard and M. Rodriguez, Multi-wavelength models expand the validity of DBP-differential absorbance relationships in drinking water, *Water Res.*, 2019, **158**, 61–71.
19 R. Jaffé, J. N. Boyer, X. Lu, N. Maie, C. Yang, N. M. Scully and S. Mock, Source characterization of dissolved organic matter in a subtropical mangrove-dominated estuary by fluorescence analysis, *Mar. Chem.*, 2004, **84**, 195–210.
20 M. Erlandsson, M. N. Futter, D. N. Kothawala and S. J. Köhler, Variability in spectral absorbance metrics across boreal lake waters, *J. Environ. Monit.*, 2012, **14**, 2643–2652.
21 G. V. Korshin, C.-W. Li and M. M. Benjamin, Monitoring the properties of natural organic matter through UV spectroscopy: A consistent theory, *Water Res.*, 1997, **31**, 1787–1795.
22 C. A. Stedmon, S. Markager and H. Kaas, Optical Properties and Signatures of Chromophoric Dissolved Organic Matter (CDOM) in Danish Coastal Waters, *Estuarine, Coastal Shelf Sci.*, 2000, **51**, 267–278.
23 J. R. Helms, A. Stubbins, J. D. Ritchie, E. C. Minor, D. J. Kieber and K. Mopper, Absorption spectral slopes and slope ratios as indicators of molecular weight, source, and photobleaching of chromophoric dissolved organic matter, *Limnol. Oceanogr.*, 2008, **53**, 955–969.
24 M. S. Twardowski, E. Boss, J. M. Sullivan and P. L. Donaghay, Modeling the spectral shape of absorption by chromophoric dissolved organic matter, *Mar. Chem.*, 2004, **89**, 69–88.
25 S. A. Loiselle, L. Bracchini, A. M. Dattilo, M. Ricci, A. Tognazzi, A. Cózar and C. Rossi, Optical characterization of chromophoric dissolved organic matter using wavelength distribution of absorption spectral slopes, *Limnol. Oceanogr.*, 2009, **54**, 590–597.
26 R. A. Müller, D. N. Kothawala, E. Podgrajsek, E. Sahlée, B. Koehler, L. J. Tranvik and G. A. Weyhenmeyer, Hourly, daily, and seasonal variability in the absorption spectra of chromophoric dissolved organic matter in a eutrophic, humic lake, *J. Geophys. Res.: Biogeosci.*, 2014, **119**, 1985–1998.

27 S. S. Ruhala and J. P. Zarnetske, Using in-situ optical sensors to study dissolved organic carbon dynamics of streams and watersheds: A review, *Sci. Total Environ.*, 2017, **575**, 713–723.
28 G. Langergraber, N. Fleischmann and F. Hofstädter, A multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater, *Water Sci. Technol.*, 2003, **47**, 63–71.
29 J. S. Horsburgh, S. L. Reeder, A. S. Jones and J. Meline, Open source software for visualization and quality control of continuous hydrologic and water quality sensor data, *Environ. Model. Softw.*, 2015, **70**, 32–44.
30 MATLAB, *Signal Processing Toolbox Release 2021a*, 2018.
31 P. Massicotte and S. Markager, Using a Gaussian decomposition approach to model absorption spectra of chromophoric dissolved organic matter, *Mar. Chem.*, 2016, **180**, 24–32.
32 W. Boënne, N. Desmet, S. Van Looy and P. Seuntjens, Use of online water quality monitoring for assessing the effects of WWTP overflows in rivers, *Environ. Sci.: Processes Impacts*, 2014, **16**, 1510–1518.
33 A. Keucken, *Ph.D. Thesis*, Lund University, 2017.
34 E. E. Lavonen, D. N. Kothawala, L. J. Tranvik, M. Gonsior, P. Schmitt-Kopplin and S. J. Köhler, Tracking changes in the optical properties and molecular composition of dissolved organic matter during drinking water production, *Water Res.*, 2015, **85**, 286–294.
35 S. Xia, X. Li, Q. Zhang, B. Xu and G. Li, Ultrafiltration of surface water with coagulation pretreatment by streaming current control, *Desalination*, 2007, **204**, 351–358.
36 T. Ohno, Fluorescence Inner-Filtering Correction for Determining the Humification Index of Dissolved Organic Matter, *Environ. Sci. Technol.*, 2002, **36**, 742–746.
37 D. M. McKnight, E. W. Boyer, P. K. Westerhoff, P. T. Doran, T. Kulbe and D. T. Andersen, Spectrofluorometric characterization of dissolved organic matter for indication of precursor organic material and aromaticity, *Limnol. Oceanogr.*, 2001, **46**, 38–48.
38 E. Parlanti, K. Wörz, L. Geoffroy and M. Lamotte, Dissolved organic matter fluorescence spectroscopy as a tool to estimate biological activity in a coastal zone submitted to anthropogenic inputs, *Org. Geochem.*, 2000, **31**, 1765–1781.
39 H. M. Sosik and B. G. Mitchell, Light absorption by phytoplankton, photosynthetic pigments and detritus in the California Current System, *Deep Sea Res., Part I*, 1995, **42**, 1717–1748.
40 S. C. Johannessen, W. L. Miller and J. J. Cullen, Calculation of UV attenuation and colored dissolved organic matter absorption spectra from measurements of ocean color, *J. Geophys. Res.: Oceans*, 2003, **108**, 3301.
41 J. Yang, S. Rahardja and P. Fränti, presented in part at the *Proceedings of the International Conference on Artificial Intelligence*, Information Processing and Cloud Computing, Sanya, China, 2019.
42 Y.-C. Chen, A tutorial on kernel density estimation and recent advances, *Biostatistics & Epidemiology*, 2017, **1**, 161–187.
43 S. Bertilsson and L. J. Tranvik, Photochemically produced carboxylic acids as substrates for freshwater bacterioplankton, *Limnol. Oceanogr.*, 1998, **43**, 885–895.
44 G. Korshin, C. W. Chow, R. Fabris and M. Drikas, Absorbance spectroscopy-based examination of effects of coagulation on the reactivity of fractions of natural organic matter with varying apparent molecular weights, *Water Res.*, 2009, **43**, 1541–1548.

