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. Keuckende 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.

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. 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. Absorbance-based sensors are used worldwide for drinking-, waste-, environmental- and...
Environmental Science: Water Research & Technology

Paper

common use include the exponential decreasing function, as in eqn (1):\textsuperscript{5,22,23}

\[ a_i = a_0 e^{x(i-\lambda_0)} + K \]  

with \( a_i = \text{absorbance value} \ [\text{m}^{-1}] \) at a certain wavelength, \( a_0 = \text{absorbance value} \ [\text{m}^{-1}] \) at the reference wavelength \( \lambda_0 \), \( S_i = \text{slope coefficient} \ [\text{m}^{-1}] \) and \( K = \text{background constant} \) to offset the baseline shift or attenuation not due to CDOM ("self-absorption"). The amplitude \( a_0 \) and the slope \( S_i \) are often used as a proxy for concentration and for changes in composition of CDOM, respectively.\textsuperscript{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,\textsuperscript{23} which might be used to detect larger changes occurring at shorter wavelengths (275–295 nm) compared to longer wavelength (350–400 nm) 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.\textsuperscript{20,23}

\( S_J \) is computed from the linear regression of the logarithm of the absorbance spectra over a sliding window applied to the wavelengths.\textsuperscript{25} \( S_J \) 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.\textsuperscript{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 spectrOlyser (s: can Messtechnik GmbH) was used in this study. The spectrOlyser 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.\textsuperscript{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\textsuperscript{29,30} or that compute metrics from absorption spectra of CDOM,\textsuperscript{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).\textsuperscript{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

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\textsuperscript{2} slightly acidic (SW, pH 6.7, $\sigma = 60 \, \mu S \, cm^{-1}$) oligotrophic lake, surrounded by mixed woodland with an approximately five-year turnover time.\textsuperscript{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\textsuperscript{-1}) and with intermediate specific ultraviolet absorbance (SUVA, 3.2 L mg\textsuperscript{-1} m\textsuperscript{-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, $\sigma = 60 \, \mu S \, cm^{-1}$, TOC = 0.6 mg L\textsuperscript{-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).\textsuperscript{16} This results in an incoming water to the DWTP containing relatively low DOC concentrations ($\sim 2.9$ mg L\textsuperscript{-1}) and SUVA of \textit{circa} 3.1 L mg\textsuperscript{-1} m\textsuperscript{-1} (Table 1).
To C and DOC were measured with a TOC-VCPH carbon analyser (Shimadzu) and DOC had an average coefficient 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 (UV254) to the DOC concentration.

Fluorescence was measured using an Aqualog spectrofluorometer (Horiba Jobin Yvon) with a 1 cm quartz cuvette connected to an 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 (≈3 × 10^3 measurements × 200 wavelengths).

### 3.2. Preprocess the dataset

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

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

<table>
<thead>
<tr>
<th>Analytical step</th>
<th>Analytical substep</th>
<th>Function name</th>
<th>Aim of the function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Import raw data files</td>
<td>Dataset assembly</td>
<td><code>abs_read</code></td>
<td>Import a list of attenuation data files as function of time</td>
</tr>
<tr>
<td>Preprocess the dataset</td>
<td>Data type conversion</td>
<td><code>convert2dtype</code></td>
<td>Convert one or more categories of values to a different one</td>
</tr>
<tr>
<td></td>
<td>Data quality assessment</td>
<td><code>nan_check</code></td>
<td>Quantify missing data in rows and columns</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>dropna</code></td>
<td>Drop rows or columns containing only missing data</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>dup_check</code></td>
<td>Check the occurrence of duplicates</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>drop_duplicates</code></td>
<td>Drop rows or columns which are duplicates</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>tshift_dst</code></td>
<td>Shift the dataset in time one hour forward when the daylight saving time ends</td>
</tr>
<tr>
<td>Explore the dataset</td>
<td>Attenuation data correction</td>
<td><code>abs_pathcor</code></td>
<td>Shift the dataset in time one hour forward when the daylight saving time ends</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>abs_basecor</code></td>
<td>Shift the dataset in time one hour forward when the daylight saving time ends</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>kdeplot</code></td>
<td>Visualise the data distribution using Gaussian KDE plot</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>outlier_id_drop_iqr</code></td>
<td>Identify potential outliers and events based on the interquartile (IQR) thresholding strategy and drop them</td>
</tr>
<tr>
<td>Interpret the results</td>
<td>Absorbance ratios</td>
<td><code>abs_ratio</code></td>
<td>Calculate the ratio of absorbance data at two different wavelengths</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>abs_fit_exponential</code></td>
<td>Fit an exponential curve to the absorbance data</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>abs_slope_ratio</code></td>
<td>Calculate the slope ratio</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>abs_spectral_curve</code></td>
<td>Generate the spectral slope curve</td>
</tr>
</tbody>
</table>

---

User decision. Python built-in functions.
Table 3  Difference in time between the time information displayed on the three spectrolyser units \((t_{\text{scan}})\) in Fig. 1 and the local time \((t_{\text{CEST}})\) 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 \(\text{tshift}\_\text{DST}\) and \(\text{timedelta}\) 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.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Period</th>
<th>(t_{\text{scan}})</th>
<th>(t_{\text{CEST}})</th>
<th>(\Delta t_{\text{tot}})</th>
<th>(\Delta t_{\text{DST}})</th>
<th>(\Delta t_{\text{scan}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>SW</td>
<td>DST</td>
<td>06:48:00</td>
<td>08:16:00</td>
<td>-01:28:00</td>
<td>-01:00:00</td>
<td>-00:28:00</td>
</tr>
<tr>
<td></td>
<td>ST</td>
<td>07:38:00</td>
<td>08:06:00</td>
<td>-00:28:00</td>
<td>+00:43:00</td>
<td>+00:44:00</td>
</tr>
<tr>
<td>RSF</td>
<td>DST</td>
<td>10:06:00</td>
<td>09:23:00</td>
<td>+00:43:00</td>
<td>+00:43:00</td>
<td>+00:44:00</td>
</tr>
<tr>
<td></td>
<td>ST</td>
<td>10:25:00</td>
<td>09:42:00</td>
<td>+00:44:00</td>
<td>+00:44:00</td>
<td>+00:44:00</td>
</tr>
<tr>
<td>UF</td>
<td>DST</td>
<td>10:05:00</td>
<td>09:21:00</td>
<td>+00:44:00</td>
<td>+00:44:00</td>
<td>+00:44:00</td>
</tr>
<tr>
<td></td>
<td>ST</td>
<td>10:23:00</td>
<td>09:39:00</td>
<td>+00:44:00</td>
<td>+00:44:00</td>
<td>+00:44:00</td>
</tr>
</tbody>
</table>

\(\Delta t_{\text{tot}}\) = total time difference \((t_{\text{scan}} - t_{\text{CEST}})\). \(\Delta t_{\text{DST}}\) = time difference due to DST \((\Delta t_{\text{tot\ DST}} - \Delta t_{\text{tot\ ST}})\). \(\Delta t_{\text{scan}}\) = time difference due to other reasons \((\Delta t_{\text{tot\ DST}} - \Delta t_{\text{DST}})\).

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 \((\text{timedelta})\);

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 \((\text{timedelta})\).

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 SW can be attributed 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 \((t_{\text{scan}})\) of the three spectrolyser units in Fig. 1 and the local time \((t_{\text{CEST}})\) 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 \((\text{abs\ pathcor})\) unless this happens automatically as for the spectrolyser. 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, \(\text{abs\ basecor}\) 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\textsuperscript{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\textsuperscript{†} shows the unfiltered UV\textsubscript{254} data from grab samples (x-axis) for SW versus the scatterplot of the UV\textsubscript{255} data from the spectro\textsuperscript{®}lyser (y-axis; due to the 2.5 nm wavelength resolution this is the nearest wavelength to UV\textsubscript{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\textsuperscript{®}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\textsuperscript{†}) 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\textsuperscript{†}). 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\textsuperscript{†}) 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.
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. 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.

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.\textsuperscript{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, ESIt). 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 \texttt{kdeplot} assumes an underlying Gaussian distribution at the location of each data point. In Fig. A7 (ESIt), 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.\textsuperscript{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, ESIt).

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,\textsuperscript{31} functions to calculate common metrics from absorbance spectra of CDOM are implemented in the AbspectroscoPY toolbox, including $S$, $S_R$ and $S_I$ 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 (\texttt{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 $A_{250}/A_{365}$, $A_{254}/A_{436}$, $A_{360}/A_{400}$ and $A_{220}/A_{254}$, respectively. Behaviour of the ratios $A_{250}/A_{365}$, $A_{254}/A_{436}$ and $A_{360}/A_{400}$ 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 $A_{250}/A_{354}$ 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 Tämaren the ratio $A_{250}/A_{365}$ increased during the summer period reaching its maximum values in September$^{26}$ and in the river Fyris, the IDOM also decreased during the spring-summer period.\textsuperscript{7} This was attributed to the shift of MW distribution to lower MW by photodegradation.$^{23,43}$

Considering the $A_{254}/A_{365}$ ratio in Fig. A8 (ESIt), 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 (\textit{e.g.}, compare $A_{250}/A_{365}$ and $A_{252}/A_{436}$ ratios in Fig. A8, ESIt). 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 (ESIt) shows an example of fitting the absorbance spectra from the spectro::lyser in SW to a single exponential decay function at a specific date (\texttt{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 (ESIt) shows the slope ratio time-series in SW (\texttt{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 (ESIt), 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 (\texttt{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 (\textit{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
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):

\[
\frac{-\delta \alpha}{\delta \lambda}_{\lambda=290.5\,\text{nm}}(t)
\]

\((2)\)

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
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
Environmental Science: Water Research & Technology

References


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.


17 G. Stéphanie and D. Caetano, Real-Time Estimation of Disinfection By-Products through Differential UV Absorbance, Water, 2020, 12, 2536.


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.