# Supplementary Information

## Integrating catchment, climate and reservoir drivers to estimate the risk of THM formation at a drinking water treatment plant inlet

**Angela Pedregal-Montes[a,b*], Eleanor Jennings[c], Rafael Marcé[d], Maria José Farré[a]**

[a]Catalan Institute for Water Research (ICRA), Carrer Emili Grahit 101, Parc Científic i Tecnològic de la Universitat de Girona, 17003 Girona, Spain

[b]University of Girona, Plaça de Sant Domènec 3, 17004 Girona, Spain

[c]Centre for Freshwater and Environmental Studies, Dundalk Institute of Technology, A91 K584 Dundalk, Ireland

[d]Centre for Advanced Studies of Blanes (CEAB), Spanish National Research Council (CSIC), 17300 Blanes, Spain

*Corresponding author.

Email address: apedregal@icra.cat (A. Pedregal-Montes)

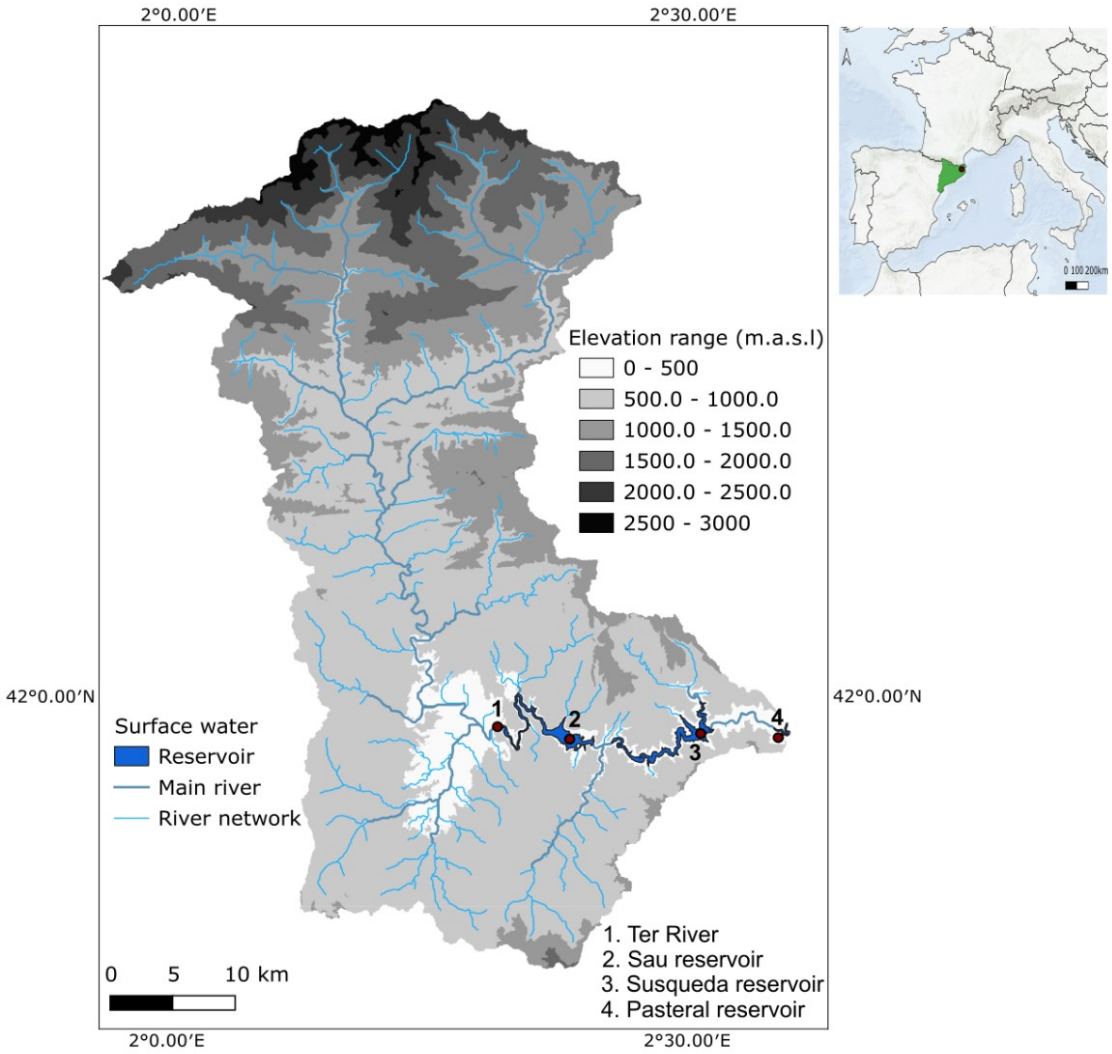


**Figure S1.** Location and physical characteristics of the Ter River catchment in Catalonia (NE Spain). Elevation is shown in ranges (m.a.s.l.), with the river network and main channel overlaid. Reservoirs in the Ter cascade, Sau, Susqueda, and Pasteral, are indicated, together with the Ter River inflow. The inset map shows the regional context within the western Mediterranean.

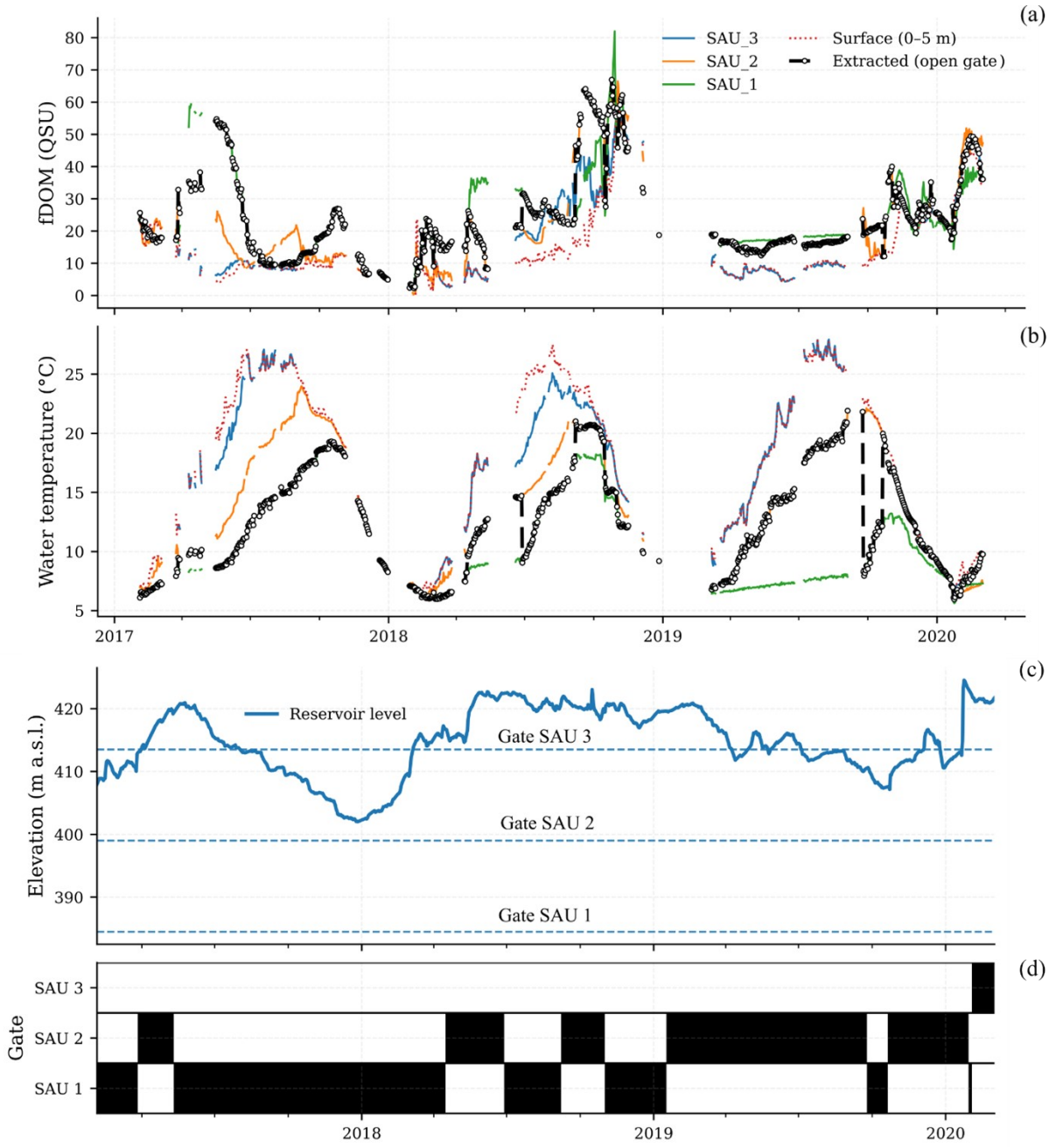


**Figure S2.** Historical records from February 2017 to March 2020 show the comparison of (a) fDOM and (b) water temperature at the Sau reservoir surface, at the withdrawal gate depths, and in the actually extracted water. Panel (c) illustrates the Sau reservoir level as well as its gate elevations. Panel (d) shows the corresponding gate operation records used to determine the active withdrawal depth. Together, the panels illustrate how temporal variations in gate operations control the quality of the extracted water relative to vertical gradients captured by the profiler within the reservoir.

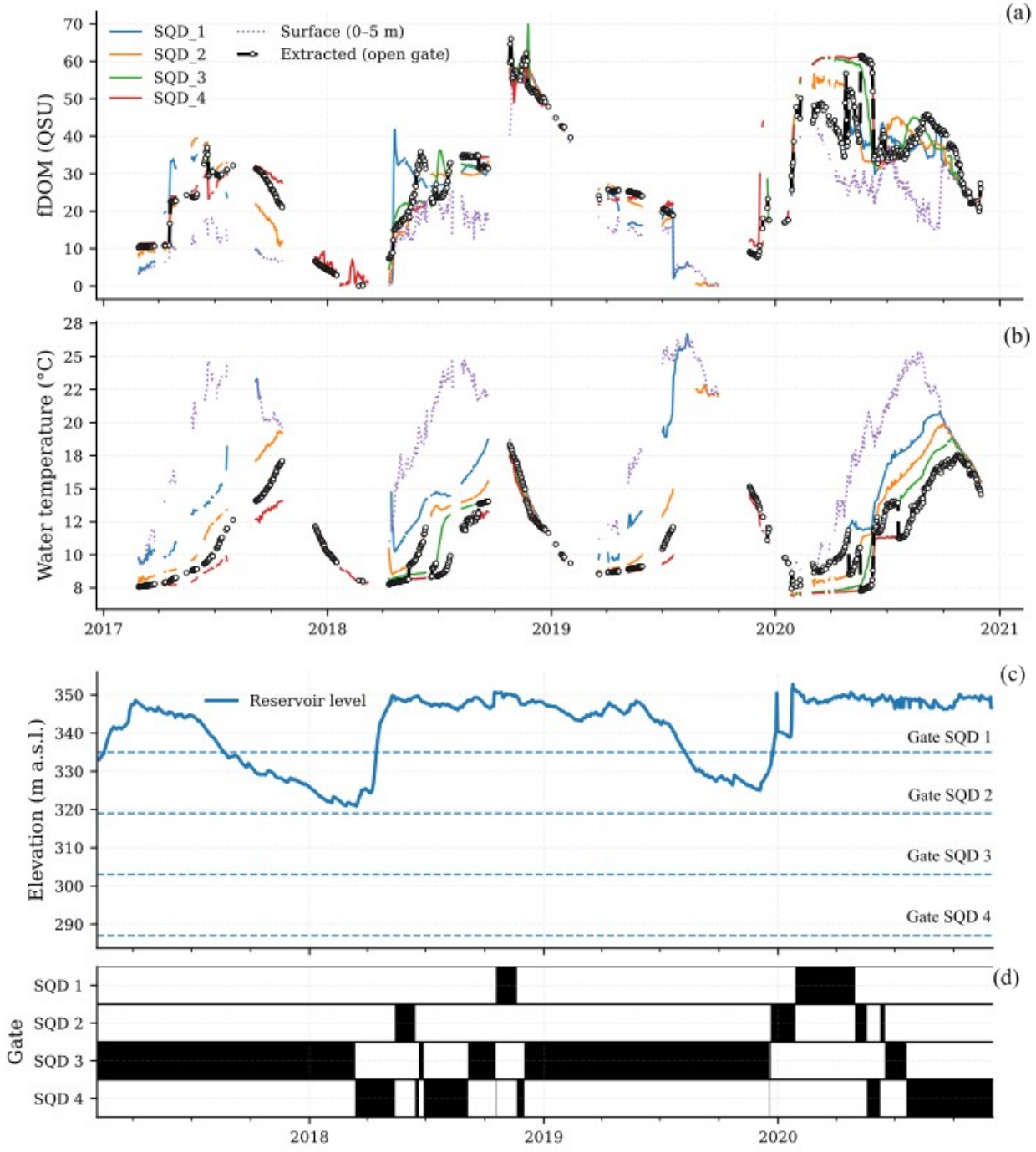


**Figure S3.** Historical records from February 2017 to November 2020 show the comparison of (a) fDOM and (b) water temperature at the Susqueda reservoir surface, at the withdrawal gate depths, and in the actually extracted water. Panel (c) illustrates the Susqueda reservoir level as well as its gate elevations. Panel (d) shows the corresponding gate operation records used to determine the active withdrawal depth. Together, the panels illustrate how temporal variations in gate operations control the quality of the extracted water relative to vertical gradients captured by the profiler within the reservoir.

**Table S1.** Physical and hydrological characteristics of the Sau, Susqueda, and Pasteral reservoirs in the Ter River cascade.

| | **Sau** | **Susqueda** | **Pasteral** |
|---|---|---|---|
| **Capacity ($hm^3$)** | 166 | 233 | ~1.5 |
| **Surface area (ha)** | ~570 | 463-466 | 34-35 |
| **Average depth (m)** | ~25 | ~50 | ~5.7 |
| **Maximum depth (m)** | ~65 | 125-135 | ~33 |
| **Mean residence time (HRT)** | ~3.6 months | ~3.3 months | 1-2 days |

**Text S1. Data preprocessing and feature construction for machine learning models.**

A structured preprocessing pipeline was applied following stablished predictive modeling best practices to ensure consistency across predictors and target variables and to support robust model training and evaluation[60].

(1) **High-frequency data processing**. High frequency profiler data (2-min) were aggregated to daily resolution. fDOM values were first corrected for water temperature effects using linear regression[31, 52], and then aggregated together with the remaining profiler variables. Depending on the operational status of the withdrawal gates, the water quality time series corresponding to the active withdrawal depth (integrated over 5 m vertical intervals) were extracted to ensure consistency between measured conditions and actual withdrawal operations. Surface water quality was obtained by integrating the upper 5 m of the water column. As a result, daily time series were generated for both the surface (0-5 m) and withdrawal depths at the Sau and Susqueda reservoirs.

(2) **DWTP data**. UV absorbance at 254 nm (UV254) was measured with an 8453 UV-vis Spectrophotometer (Agilent, USA). Dissolved organic carbon (DOC) was analyzed as non-purgeable organic carbon (NPOC) using a TOC-$V_{CSN}$ analyzer (Shimadzu). For DOC and UV254 measurements, samples were filtered at 0.45μm to ensure analysis of dissolved organic matter (DOM) fraction. The specific ultraviolet absorbance (SUVA) was calculated as UV254 divided by DOC. Total trihalomethanes (THMs) were extracted and analyzed following standard methods[61].

(3) **Data validation.** Missing, physically implausible, or invalid values (negative concentrations or sensor artifacts) were identified and removed prior to any interpolation.

(4) **Temporal alignment and gap-limited interpolation**. All variables were aligned to a common daily timeline spanning 4 February 2017 to 30 November 2020, corresponding to the period of Susqueda profiler availability. Variables already available at daily resolution (meteorology, catchment simulations, reservoir operational data, and DWTP inlet observations) were aligned to this grid without temporal interpolation. Interpolation was applied only to profiler-derived variables to support daily alignment and sequence-base modeling. For profiler time series, short gaps (≤14 days) were filled using linear interpolation between bounding observations[62]. Longer gaps were treated as missing and were not bridged by interpolation.

To preserve information about data availability during long gaps, missingness was encoded explicitly in the ML predictors using (i) binary availability flags and (ii) time-since-last-observation variables[63]. Missing profiler values during long gaps were assigned neutral placeholder values solely to allow construction of continuous LSTM input sequences; the accompanying missingness features allow the models to condition predictors on the reliability of profiler inputs.

(5) **Normalization.** Each variable was rescaled to the unit interval using min-max normalization. This transformation places all predictors on a comparable numerical scale and improves numerical stability during training by reducing the influence of differences in magnitude and units. Normalization parameters were computed from the modeling period and applied consistently for model training and scenario simulations.

(6) **Smoothing**. To reduce the influence of short-duration noise and outliers (e.g. sensor spikes), a nine-day median filter was applied. For time series containing missing segments, filtering was applied only within continuous observed (non-missing) stretches to avoid propagating information across long data gaps. Median filtering is well suited to environmental time series because it suppresses isolated extremes while preserving overall temporal structure more effectively than moving averages.

**Text S2. Machine learning model configuration and training settings**

Machine learning models were implemented using a conssitent architecture, training protocol, and evaluation procedure across all modeled targets, with predictor sets tailored to each target as described in Table 1 and Text S1.

**LSTM configuration.** LSTM models were formulated as sequence-to-one predictors using a fixed lookback window of 14 days, such that each predcition at day *t* was generated from the preceding 14 days of predictor values (*t-14* to *t-1*). The network architecture consisted of a single LSTM layer with 80 units (linear activation), followed by a dropout layer (dropout rate = 0.2) and fully connected dense output layer with one neuron. Models were trained using the RMSprop optimizer and mean squared error loss. Training was run for a maximum of 500 epochs with a batch size of 16. Early stopping was applied to prevent overfitting, monitored on the loss for the chronologically held-out period, with a patience of 50 epochs, a minimum improvement threshold of $1x10^{-6}$, and restoration of the best-performing model weights.

**Chronological splitting and reproducibility.** To represent time ordering and avoid information leakage, model training and evaluation used a chronological split with the first 80% of the time series used for training and the remaining 20% reserved as a held-out test period. Reproducibility was promoted by fixing random seeds (Python, NumPy, and TensorFlow) and enabling deterministic computation where supported.

**Random Forest configuration and feature ranking.** RF models were trained as non-sequential baselines and to support driver ranking. RF models were fit using 1000 trees, bootstrap sampling, and a minimum leaf size of 2. Predictor relevance was first quantified using RF permutaiton importance under time-series cross-validation (five folds). Within each fold, permutaiton importance was repeated 40 times per predictor, and fold-wise importance scores were averaged to obtain mean and standard deviation estimates across folds.

**Permutation importance for LSTM**. LSTM permutation importance was computed on the held-out period to quantify the sensitivity of predictive skill to each selected predictor. For each predictor, values were permuted repeatedly across test samples while preserving within-sequence temporal structure of the remaining predictors. Importance was expressed as the mean decreased in $R^2$ relative to the unpermuted model ($\Delta R^2$) and estimated from 50 permutation repeats per predictor.

**Implementation details**. Model implementation was performed in Python using scikit-learn for RF models and TensorFlow/Keras for LSTM models. All settings described aboved were kept fixed across modeled targets to facilitate consistent comparison of model performance and driver attribution outcomes.

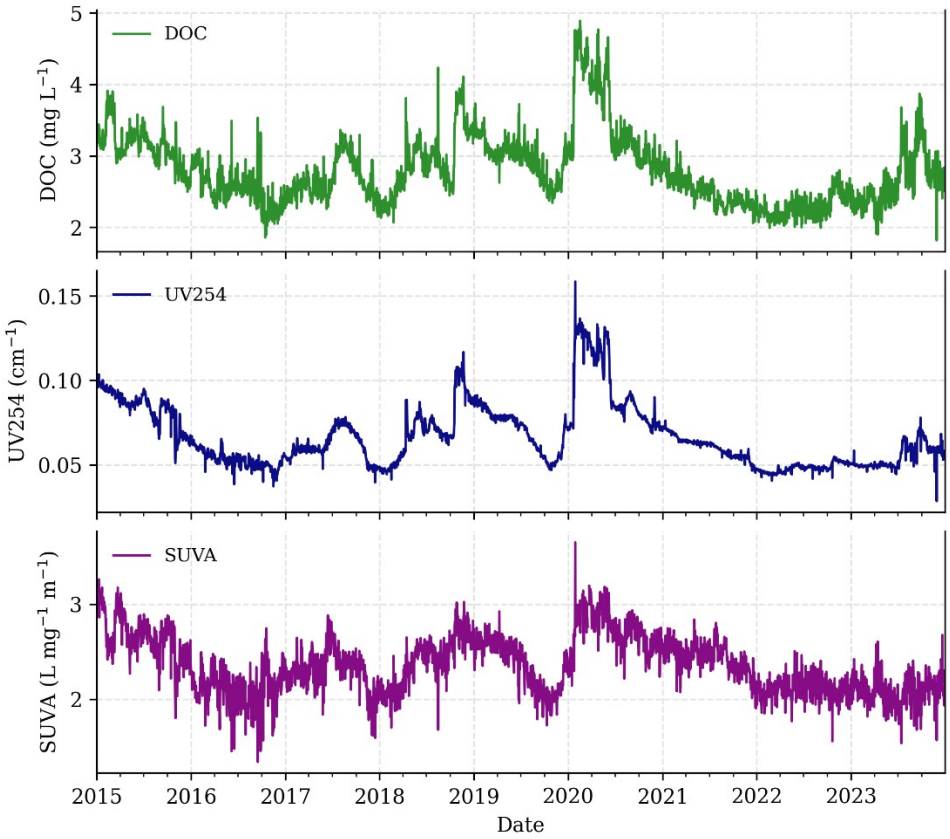


**Figure S4.** Daily time series of dissolved organic carbon (DOC), UV254 absorbance, and SUVA in the DWTP inlet from January 2015-December 2023. DOC represents the concentration of dissolved organic matter, UV254 reflects aromatic carbon absorbance, and SUVA (UV254/DOC x 100) indicates the aromatic character of DOM.

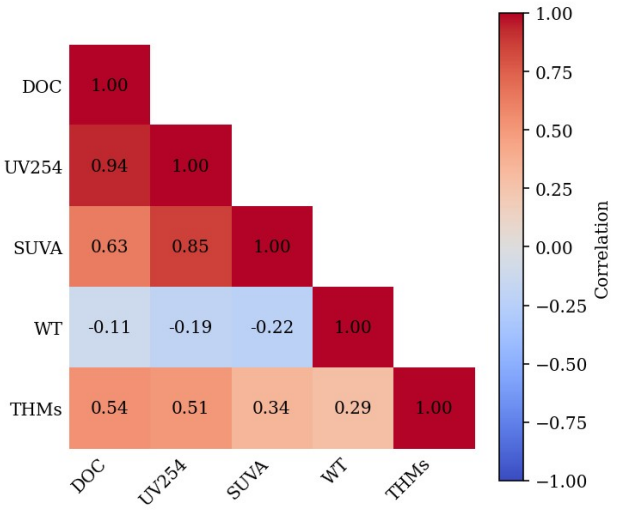


**Figure S5.** Pearson correlation matrix showing relationships among dissolved organic carbon (DOC), UV254 absorbance, SUVA and water temperature (WT) at the inlet, and total THMs concentration at the DWTP outlet. The analysis was based on 418 paired measurements collected between January 2015 and December 2023.

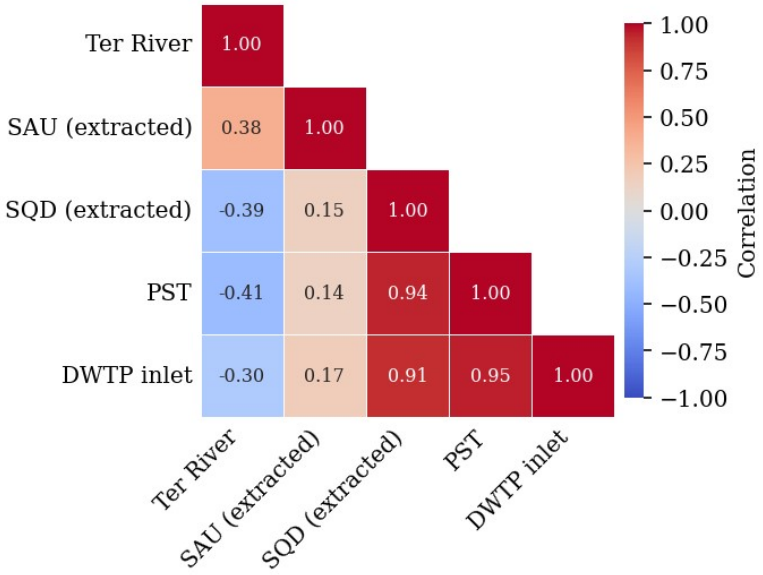


**Figure S6.** Pearson correlation matrix showing the spatiotemporal relationships among monthly DOC concentrations along the river-reservoir-DWTP continuum. The analysis was based on 50 paired monthly measurements collected between January 2015 and December 2023 at five monitoring sites: Ter River, Sau (extracted withdrawal depth), Susqueda (extracted withdrawal depth, SQD), Pasteral intake (PST), and at the drinking water treatment plant (DWTP) inlet.

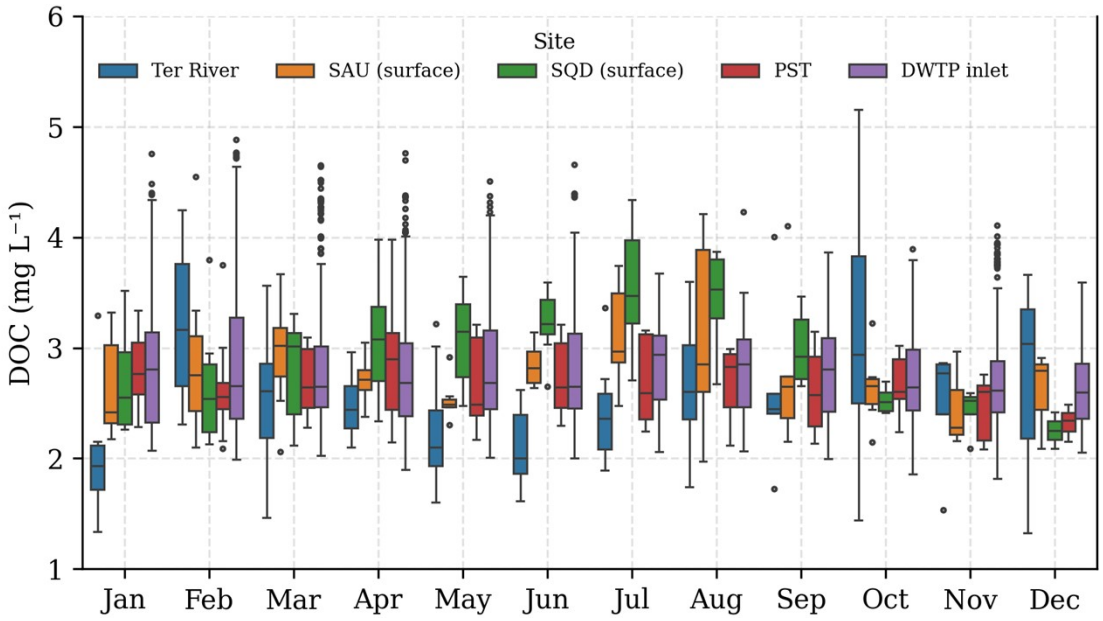


**Figure S7.** Seasonal distribution of DOC concentrations along the river-reservoir-DWTP continuum for the period January 2015-December 2023. Monthly boxplots are shown for the Ter River, Sau (surface), Susqueda (surface, SQD), Pasteral intake (PST), and the DWTP inlet ($n_{Ter}$ = 69, $n_{Sau}$ = 74, $n_{Susqueda}$ = 70, $n_{Pasteral}$ = 81, $n_{DWTP}$ = 108).

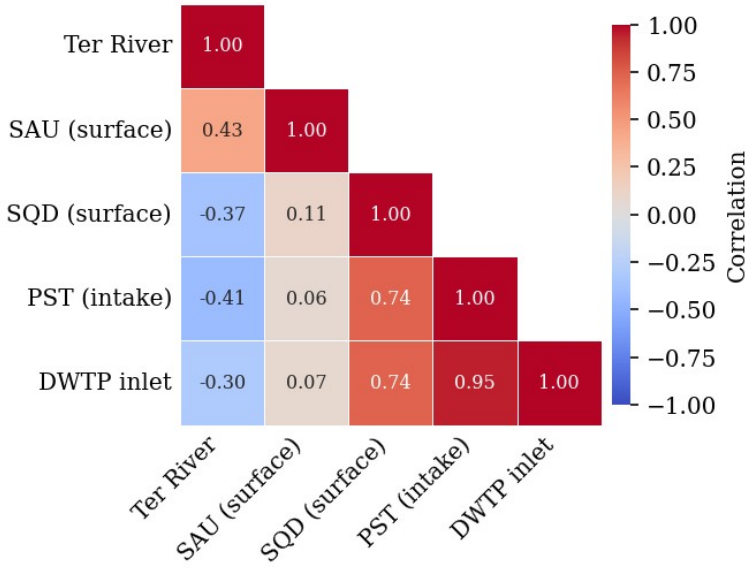


**Figure S8.** Pearson correlation matrix showing the spatiotemporal relationships among monthly DOC concentrations along the river-reservoir-DWTP continuum. The analysis was based on 50 paired monthly measurements collected between January 2015 and December 2023 at five monitoring sites: Ter River, Sau (surface), Susqueda (surface, SQD), Pasteral intake (PST), and at the drinking water treatment plant (DWTP) inlet.

**Table S2.** Thresholds defined by Godó et al., 2021[25] to classify raw water DOC and water temperature into low, medium, and high indicator levels for THM formation risk assessment.

| Raw water variable | Range | Classification |
|---|---|---|
| DOC (mg $L^{-1}$) | 0.0 – 2.7 | Low |
| | 2.7 – 3.4 | Mid |
| | >3.4 | High |
| Water temperature (ºC) | 0 – 10 | Low |
| | 10 – 15 | Mid |
| | >15 | High |

**Table S3.** THM formation risk matrix proposed by Godó et al., 2021[25], assigning combined risk classes (Minimum-Maximum) from the cross-classification of DOC (low-high) and water temperature (low-high) categories.

| **THM formation risk class** | | **Water temperature classification** | | |
|---|---|---|---|---|
| | | **Low** | **Mid** | **High** |
| **DOC classification** | **Low** | Minimum | Low | Mid |
| | **Mid** | Low | Mid | High |
| | **High** | Mid | High | Maximum |