Guidelines to develop and validate soft-sensors for target contaminants, as backbone of an early warning system

Deliverable D4.2, WP4

Project Number	**101081980**
Project Title	Climate-resilient management for safe disinfected and non-disinfected water supply systems
Project Acronym	SafeCREW
Project Duration	November 2022 – May 2026
Call identifier	HORIZON-CL6-2022-ZEROPOLLUTION-01
Due date of Deliverable	Month 36, 31.10.2025
Final version date	Month 36, 30.10.2025
Updated version date	26.06.2026
Dissemination Level	PU (Public)
Deliverable No.	D4.2
Work Package	WP4
Task	T4.2
Lead Beneficiary	POLIMI
Contributing Beneficiaries	EUT, NUWEE
Report Author	POLIMI: Massimiliano Arca, Ilenia Epifani, Francesco Trovò EUT: Danillo Lange Souza Borges Dos Santos, NUWEE: Martynov Serhii, Kupchyk Larysa
Reviewed by	POLIMI: Beatrice Cantoni, Manuela Antonelli
Approved by	DVGW-TUHH: Anissa Grieb
Updated by	EUT: Danillo Lange Souza Borges Dos Santos
Update approved by	DVGW-TUHH: Anissa Grieb

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

History of Versions		
Version	**Publication Date**	**Change**
0.2	18.10.2025	1st version sent to the coordinator
0.5	30.10.2025	Final review
1	30.10.2025	Submitted to commission
1.1	20.06.2026	Correction of errors
2	26.06.2026	Submitted to commission

History of Changes	
Date/Section	**Nature of change and reason**
20.06.2026, section 2.2.1	Text adjusted to better reflect the description of the case study, specifying how many online sensors and locations were used for this task.
20.06.2026, tables 6 and 7	Adjustment made to remove an unintended outlier value of the total Trihalomethanes dataset.

Part. No.	**Participant organisation name**	**Short name**
1 (Coordinator)	German Association for Gas and Water, local research branch at Hamburg University of Technology	DVGW-TUHH
2	Politecnico di Milano	POLIMI
3	Berlin Centre of Competence for Water	KWB
4	BioDetection Systems B.V.	BDS
5	Centre Tecnológic de Catalunya (EURECAT)	EUT
6	German Environment Agency	UBA
7	Helmholtz-Centre for Environmental Research	UFZ
8	Consorci Concessionari d'Aigües per als Ajuntaments i Industries de Tarragona	CAT
9	Tutech Innovation GmbH	TUTECH
10	Metropolitana Milanese SPA	MM
11	Multisensor systems Ltd.	MSS
12	The National University of Water and Environmental Engineering	NUWEE

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Abstract

Deliverable D4.2 provides guidelines for the development and validation of soft-sensors for target contaminants, serving as a backbone of an early warning system within the SafeCREW project. The primary objective is to establish a methodological framework for designing data-driven tools that support the management of smart and climate-resilient Drinking Water Supply Systems (DWSS) and Drinking Water Distribution Networks (DWDN).

The work is based on data collected in three case studies: CS#2 (Milan, Italy – POLIMI), CS#3 (Tarragona, Spain – EUT), and CS#4 (Western Ukraine – NUWEE). Each case integrates low-informative sensors (e.g., pH, temperature, conductivity, UVA254, residual chlorine) and high-informative sensors (e.g., THM-meter, Flow Cytometry – FCM, fluorescence), complemented with network modelling.

By applying incremental and online machine learning (ML) techniques, correlations were established between conventional parameters (e.g., pH, temperature, and absorbance) and target contaminants (e.g., disinfectants, natural organic matter (NOM), disinfection by-products (DBPs), and microbial indicators) to design predictive soft-sensors. These models aim to detect anomalous water quality dynamics and provide early warning signals for proactive DWSS operations. The deliverable presents the methodological framework for input selection, model training, validation, and performance evaluation, along with a comparative analysis of the approaches adopted by POLIMI, EUT, and NUWEE. The results highlight how combining low- and high-informative monitoring data through data-driven modelling significantly enhances the capability to identify failure events and optimize routine DWSS management.

Finally, D4.2 emphasizes the role of soft-sensors as key components in the transition towards integrated and adaptive early warning systems, supporting the objectives of SafeCREW in ensuring safe and climate-resilient drinking water management across diverse European disinfected and non-disinfected supply contexts.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Table of contents

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Abbreviations

GA	Grant Agreement
EC	European Commission
DWSS	Drinking Water Supply Systems
DWDN	Drinking Water Distribution Network
ML	Machine Learning
DBPs	Disinfection by-products
THMs	Trihalomethanes
FCM	Flow cytometry
NOM	Natural Organic Matter
TOC	Total Organic Carbon
MICE	Multiple Imputation by Chained Equations
DWTP	Drinking Water Treatment Plant
ROC	ROC: Receiving Operating Curve
PR	Precision-Recall
LOQ	Limit of Quantification
TTHM	Total Trihalomethanes
KL	Kullback–Leibler
KS	Kolmogorov–Smirnov
LoA	Limits of Agreement
ICC	Intact Cell Count
EDA	Exploratory data analysis
TSS	Total Suspended Solids
MAD	median absolute deviation
ORP	Oxidation-Reduction Potential
COP-kmeans	Constrained Pairwise kmeans
PLS	Partial Least Squares
PVR	Support Vector Regressor
XGBoost	Extreme Gradient Boosting
QRNN	Quantile Regression Neural Networks
LOO	Leave-One-Out
LGBM	LightGBM
LSTM	Long Short-Term Memory
CRISP-DM	Cross-Industry Standard Process for Data Mining
VIF	Variance Inflation Factor
TPR	True Positive Rate
FPR	False Positive Rate
MAPE	Mean Absolute Percentage Error
RFE	Recursive Feature Elimination
SVM	Support Vector Machine
ABS	Absorbance

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Key definitions

Soft-sensor	A computational model, typically data-driven, that estimates complex or costly-to-measure water quality parameters using surrogate variables from online sensors.
Low-informative parameters	Basic operational or physico-chemical indicators easily measurable in real time (e.g., pH, temperature, conductivity, residual chlorine).
High-informative parameters	Advanced indicators providing detailed information on chemical or microbial quality (e.g., fluorescence, flow cytometry, THM-meter measurements).
Natural Organic Matter (NOM)	A mixture of organic compounds in source and treated waters acting as precursors for DBP formation.
Disinfection By-Products (DBPs)	Chemical compounds formed during water disinfection processes, whose concentration must be controlled to minimize human health risks.
Early-warning system	An integrated monitoring and modelling framework designed to detect, predict, and signal potential risks in water quality, enabling preventive management actions.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

1. Introduction and aim of the task

Ensuring the safety and quality of drinking water remains one of the primary challenges for water utilities worldwide. Climate variability, aging infrastructure, and changes in source water quality contribute to increasing complexity in Drinking Water Supply Systems (DWSS) and Drinking Water Distribution Networks (DWDN). To ensure compliance with regulatory standards and maintain consumer trust, water utilities must transition from reactive control approaches—based on scheduled laboratory sampling—to proactive management supported by continuous, data-driven monitoring and predictive analytics. Early-warning systems are crucial to enable this transition. By integrating real-time monitoring, predictive models, and alert mechanisms, they provide timely information on system anomalies or contamination events. However, their effectiveness strongly depends on the availability and interpretability of monitoring data. Traditional water quality monitoring relies heavily on grab sampling and laboratory analyses, which are accurate but infrequent, and on online sensors, which typically measure only a small number of basic parameters such as pH, temperature, or residual chlorine. As a result, the ability to detect rapid changes or localized deterioration in water quality remains limited.

In this context, soft-sensors—also referred to as software sensors—can be the right tool to face these challenges. In fact, soft-sensors use data-driven models, typically based on machine learning (ML) or statistical regression, to estimate parameters that are difficult, costly, or slow to measure directly by exploiting correlations with continuously monitored parameters. When deployed in real time, soft-sensors can act as digital surrogates for complex chemical or microbiological indicators, providing valuable support for an early warning system and enabling smart, climate-resilient water management.

Soft-sensor technologies have long been used in industrial process monitoring and control, and in the last decade, they have gained increasing attention in the water sector, especially for drinking water treatment and distribution systems. Early applications focused on predicting disinfection by-products (DBPs), such as trihalomethanes (THMs) and haloacetic acids, using multi-linear regression and artificial neural networks based on routine process data (e.g., temperature, pH, chlorine dose, and natural organic matter indicators) (Rodriguez, Sérodes and Levallois, 2004; Uyak, Toroz and Meriç, 2005). Recent works have introduced advanced chemometric and machine learning approaches—including support vector regression, random forest, and deep learning—to model DBP formation dynamics with higher accuracy and adaptability (Liang et al., 2025; Peng et al., 2023). Soft-sensors have also been explored for microbial indicators, using data from flow cytometry (FCM), ATP-based methods, or surrogate parameters such as turbidity and temperature to infer microbiological stability in distribution networks (Besmer et al., 2016; Prest et al., 2016). Despite promising results, most existing models remain site-specific, relying on static training datasets that do not adapt to operational variability. Soft-sensors are typically developed for individual plants or networks and show reduced performance when applied to other systems. Few have been validated at full scale or under real operational conditions, limiting confidence in their robustness and generalization.

To address these challenges, the SafeCREW project proposes a harmonized framework for incremental and online soft-sensor development, validated in diverse European case studies. The main objective of this task is to develop, validate, and demonstrate soft-sensors capable of estimating target contaminants—including natural organic matter (NOM), disinfection by-products (DBPs), and microbial indicators—based on combinations of low- and high-informative sensor data collected at full scale. These models form the data-driven backbone of an early warning system designed to detect anomalies and support proactive operation of drinking water systems.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

The framework is tested in three complementary case studies, representing different climatic and management contexts in both Drinking Water Treatment Plants (DWTPs) and Drinking Water Distribution Networks (DWDNs):

- CS#2 (Milan, Italy): combining low- and high-informative data (e.g., UVA254, flow cytometry, THM-meter) for predictive modeling in a disinfected DWSS.
- CS#3 (Tarragona, Spain): integrating network modeling data to predict DBP formation and dynamics under varying operational conditions.
- CS#4 (Western Ukraine): applying the soft-sensor framework in a non-disinfected water supply system to assess generalizability.

The development of soft-sensors within T4.2 relies on accurate and comprehensive water quality from these sites, enabling the validation of the proposed framework across various operational conditions.

The deliverable is structured as follows: Chapter 2 describes the data sources and exploratory analyses for the three case studies: CS#2, Milan, Italy, which includes supply points and the Feltre DWTP; CS#3, Tarragona, Spain; and CS#4, Western Ukraine, Ukraine. Chapter 3 presents the methodological framework for designing, developing, and validating the soft-sensors, while Chapter 4 reports the main results obtained for each case study, along with comparative discussion and cross-case analysis. Finally, Chapter 5 presents the overall conclusions and formulates recommendations for the future integration of soft-sensors into early warning systems for drinking water supply management.

2. Data Sources and Exploratory Data Analysis

This section presents the data sources and exploratory analyses conducted in each case study, outlining data availability, sampling frequency, pre-processing steps, and validation of data consistency prior to model development. Collected data is available in the SafeCREW community of Zenodo (https://zenodo.org/communities/safecrew/).

2.1. CS#2 Milan, Italy

The Milan case study represents a large metropolitan drinking water supply system managed by MM S.p.A., which operates the Integrated Water Service of the Municipality of Milan. For this research, two complementary data sources were selected to capture different aspects of the water supply chain: monitoring data from supply points within the DWDN, and data from the Feltre DWTP. This dual approach enables the development of soft-sensors at multiple stages of the system, from post-treatment monitoring to end-user distribution.

The broader MM infrastructure includes 32 DWTPs (red dots in Figure 1, left), 52 supply points (yellow dots in Figure 1, right), and 661 public fountains, ensuring extensive spatial coverage across the metropolitan area.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Figure 1. Maps of the DWTPs (left) and supply points (right) in Milan.

2.1.1. Supply Points in the DWDN

The supply points dataset focuses on monitoring water quality dynamics within the distribution network, where physical, chemical and microbiological parameters can vary significantly due to residence time, infrastructure conditions, and seasonal variations. A subset of nine supply points, selected from the 52 monitored by MM, was used to ensure spatial representativeness and to capture relevant variability in water quality.
The dataset covers the period from January 2024 to December 2024 and includes:

- **Grab samples**: laboratory analyses performed monthly for seven supply points and weekly for two representative points identified by MM, to capture higher frequency dynamics more accurately.
- **Online sensor data**: each supply point is equipped with sensors that provide continuous 15-minute monitoring of conventional low-informative water quality parameters.

These parameters served as predictors for the development of soft-sensors.

Table 1 summarizes grab-sample data, showing missing values and measurements below the Limit of Quantification (LOQ) for several parameters, notably color, nitrate, free chlorine, and total organic carbon (TOC).

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Table 1. CS#2 supply points grab samples input data summary.

	Color (CU)				Temperature (°C)				pH				Conductivity (uS/cm)			
Supply Point	N° Entries	N° Valid Samples	% Missing	N° < LOQ	N° Entries	N° Valid Samples	% Missing	N° < LOQ	N° Entries	N° Valid Samples	% Missing	N° < LOQ	N° Entries	N° Valid Samples	% Missing	N° < LOQ
Bande Nere	11	9	9.09	1	12	12	0	0	12	12	0	0	11	11	0	0
Berna	11	9	9.09	1	11	11	0	0	11	11	0	0	11	11	0	0
Chiostergi	8	5	12.5	2	8	8	0	0	8	8	0	0	8	8	0	0
Fortunato	8	5	12.5	2	8	8	0	0	8	8	0	0	8	8	0	0
Gramsci	10	6	10	3	10	10	0	0	10	10	0	0	10	10	0	0
Montevideo	40	15	35	11	40	40	0	0	40	40	0	0	40	40	0	0
Prealpi	8	4	12.5	3	8	8	0	0	8	8	0	0	8	8	0	0
Tabacchi	39	20	35.9	5	39	39	0	0	39	39	0	0	39	39	0	0
Tognazzi	9	2	11.11	6	9	9	0	0	9	9	0	0	9	9	0	0

	Nitrate (mg/L)				Free Chlorine (mg/L)				TOC (mg/L)			
Supply Point	N° Entries	N° Valid Samples	% Missing	N° < LOQ	N° Entries	N° Valid Samples	% Missing	N° < LOQ	N° Entries	N° Valid Samples	% Missing	N° < LOQ
Bande Nere	11	11	0	0	12	6	8.33	5	11	11	0	0
Berna	11	11	0	0	11	4	0	7	11	11	0	0
Chiostergi	8	8	0	0	8	5	12.5	2	8	8	0	0
Fortunato	8	8	0	0	8	2	0	6	8	8	0	0
Gramsci	10	10	0	0	10	3	0	7	10	9	10	0
Montevideo	40	18	55	0	40	13	0	27	40	38	5	0
Prealpi	8	8	0	0	8	6	0	2	8	8	0	0
Tabacchi	39	17	56.41	0	39	14	0	25	39	37	5.13	0
Tognazzi	9	9	0	0	9	6	11.11	2	9	9	0	0

Since no direct online sensors for Total Trihalomethanes (TTHMs**)** were available, grab samples provided the reference concentrations required to train the machine learning (ML) models (see Table 2). Soft sensors were then applied to continuous sensor data (Table 3) to produce real-time predictions. No missing data was present in the online dataset.

Table2. CS#2 supply points grab samples output data summary.

	Chloroform (µg/L)				Bromodichloromethane (µg/L)				Dibromochloromethane (µg/L)				Bromoform (µg/L)			
Supply Point	N° Entries	N° Valid Samples	% Missing	N° < LOQ	N° Entries	N° Valid Samples	% Missing	N° < LOQ	N° Entries	N° Valid Samples	% Missing	N° < LOQ	N° Entries	N°Valid Samples	% Missing	N° < LOQ
Bande Nere	11	11	0	0	11	0	0	11	11	6	9.09	4	11	11	0	0
Berna	11	11	0	0	11	0	0	11	11	5	0	6	11	11	0	0
Chiostergi	8	8	0	0	8	0	0	8	8	2	0	6	8	8	0	0
Fortunato	8	8	0	0	8	2	0	6	8	4	0	4	8	8	0	0
Gramsci	10	10	0	0	10	0	0	10	10	3	0	7	10	8	0	2
Montevideo	40	39	2.5	0	40	0	2.5	39	40	14	2.5	25	40	39	2.5	0
Prealpi	8	8	0	0	8	2	0	6	8	4	12.5	3	8	8	0	0
Tabacchi	39	38	2.56	0	39	0	2.56	38	39	30	5.13	7	39	37	2.56	1
Tognazzi	9	9	0	0	9	0	0	9	9	8	11.11	0	9	9	0	0

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Table 3. CS#2 supply points sensor measurements input data summary.

	Color (CU)					Temperature (°C)					pH					Conductivity (uS/cm)				
Supply Point	N° Data	Mean	Std	Start Date	End Date	N° Data	Mean	Std	Start Date	End Date	N° Data	Mean	Std	Start Date	End Date	N° Data	Mean	Std	Start Date	End Date
Bande Nere	38699	0.93	0.54	2023-12-23	2025-01-20	38699	15.59	2.55	2023-12-23	2025-01-20	38699	7.75	0.28	2023-12-23	2025-01-20	38699	516.06	13.70	2023-12-23	2025-01-20
Berna	32792	2.20	1.85	2024-02-01	2025-01-20	32792	15.47	1.23	2024-02-01	2025-01-20	32792	7.52	0.13	2024-02-01	2025-01-20	32792	490.62	39.85	2024-02-01	2025-01-20
Chiostergi	38933	0.97	0.51	2023-12-21	2025-01-20	38933	16.29	3.09	2023-12-21	2025-01-20	38933	7.40	0.23	2023-12-21	2025-01-20	38933	493.56	9.24	2023-12-21	2025-01-20
Fortunato	31317	0.47	0.75	2024-01-25	2025-01-20	31317	16.04	2.70	2024-01-25	2025-01-20	31317	7.55	0.27	2024-01-25	2025-01-20	31317	603.34	7.82	2024-01-25	2025-01-20
Gramsci	34219	0.75	0.62	2024-02-08	2025-01-20	34219	16.54	2.89	2024-02-08	2025-01-20	34219	7.66	0.36	2024-02-08	2025-01-20	34219	521.14	46.68	2024-02-08	2025-01-20
Montevideo	30937	0.05	0.34	2024-01-30	2025-01-08	30937	16.92	3.07	2024-01-30	2025-01-08	30937	7.36	0.19	2024-01-30	2025-01-08	30937	530.28	24.09	2024-01-30	2025-01-08
Prealpi	35057	0.31	0.36	2024-01-30	2025-01-20	35057	16.07	1.55	2024-01-30	2025-01-20	35057	7.54	0.13	2024-01-30	2025-01-20	35057	429.27	13.53	2024-01-30	2025-01-20
Tabacchi	37931	1.53	0.98	2024-01-27	2025-02-27	37931	15.73	2.14	2024-01-27	2025-02-27	37931	7.55	0.18	2024-01-27	2025-02-27	37931	559.55	29.19	2024-01-27	2025-02-27
Tognazzi	36201	0.89	0.46	2024-01-18	2025-01-20	36201	16.46	1.94	2024-01-18	2025-01-20	36201	7.26	0.07	2024-01-18	2025-01-20	36201	696.56	7.89	2024-01-18	2025-01-20

	Nitrate (mg/L)					Free Chlorine (mg/L)					TOC (mg/L)				
Supply Point	N° Data	Mean	Std	Start Date	End Date	N° Data	Mean	Std	Start Date	End Date	N° Data	Mean	Std	Start Date	End Date
Bande Nere	38699	22.10	1.90	2023-12-23	2025-01-20	38699	0.04	0.03	2023-12-23	2025-01-20	38699	0.28	0.14	2023-12-23	2025-01-20
Berna	32792	27.04	2.69	2024-02-01	2025-01-20	32792	0.05	0.08	2024-02-01	2025-01-20	32792	0.45	0.37	2024-02-01	2025-01-20
Chiostergi	38933	30.82	2.25	2023-12-21	2025-01-20	38933	0.03	0.02	2023-12-21	2025-01-20	38933	0.29	0.24	2023-12-21	2025-01-20
Fortunato	31317	30.51	0.57	2024-01-25	2025-01-20	31317	0.06	7.00	2024-01-25	2025-01-20	31317	0.24	0.19	2024-01-25	2025-01-20
Gramsci	34219	28.74	2.03	2024-02-08	2025-01-20	34219	0.02	0.02	2024-02-08	2025-01-20	34219	0.16	0.09	2024-02-08	2025-01-20
Montevideo	30937	24.86	2.12	2024-01-30	2025-01-08	30937	0.08	5.69	2024-01-30	2025-01-08	30937	0.25	0.50	2024-01-30	2025-01-08
Prealpi	35057	28.86	2.15	2024-01-30	2025-01-20	35057	0.03	0.02	2024-01-30	2025-01-20	35057	0.10	0.07	2024-01-30	2025-01-20
Tabacchi	37931	24.66	2.20	2024-01-27	2025-02-27	37931	0.02	0.02	2024-01-27	2025-02-27	37931	0.33	0.25	2024-01-27	2025-02-27
Tognazzi	36201	28.84	2.21	2024-01-18	2025-01-20	36201	0.06	0.06	2024-01-18	2025-01-20	36201	0.24	0.12	2024-01-18	2025-01-20

2.1.2 Data pre-processing summary

Before model development, the raw time-series data underwent a structured pre-processing workflow to ensure consistency and reliability across all monitoring stations. The procedure followed recognized practices for environmental time-series analysis (Imran *et al.*, 2022) and included the following steps:

- Missing-data treatment: missing values missing values were handled using Multiple Imputation by Chained Equations (MICE), van Buuren and Groothuis-Oudshoorn, (2011).
- Outlier detection: identified through expert validation and visual inspection.
- Censored data: censored data were replaced conventionally with LOQ/2.

All computations were performed using Python. In the following, we detail the pre-processing pipeline.

a. Missing values

Given the limited number of grab-sample observations, missing values were handled using Multiple Imputation by Chained Equations (MICE), van Buuren and Groothuis-Oudshoorn, (2011). MICE iteratively models each variable with missing data as a function of the others, generating multiple

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

plausible estimates per missing observation, thereby preserving the inherent variability and reducing bias in downstream analyses. In particular, the method works by modelling each variable with missing values conditionally on the others through a series of regression models. The process involves:

1. Initialization: Assigning initial guesses to missing values using simple imputation (mean values in this case).
2. Iterative modeling: Fitting a regression model for each variable with missing data, using the others as predictors, and updating missing entries with model estimates.
3. Chaining: Repeating the process for all variables in turn, creating a complete cycle of imputations.
4. Multiple datasets: Repeating the entire procedure several times (five imputations in this study) to generate multiple complete datasets that reflect uncertainty in the imputed values.

To assess imputation quality, the Kullback–Leibler (KL) divergence (Shlens, 2014) and Kolmogorov–Smirnov (KS) two-sample test were applied to compare the empirical distributions of raw and imputed data. All parameters across all supply points achieved KL < 0.1 and a p-value > 0.05, confirming that no significant bias was introduced. Figure 2 illustrates an example of the statistical comparison between the two distributions of raw and imputed data using MICE.

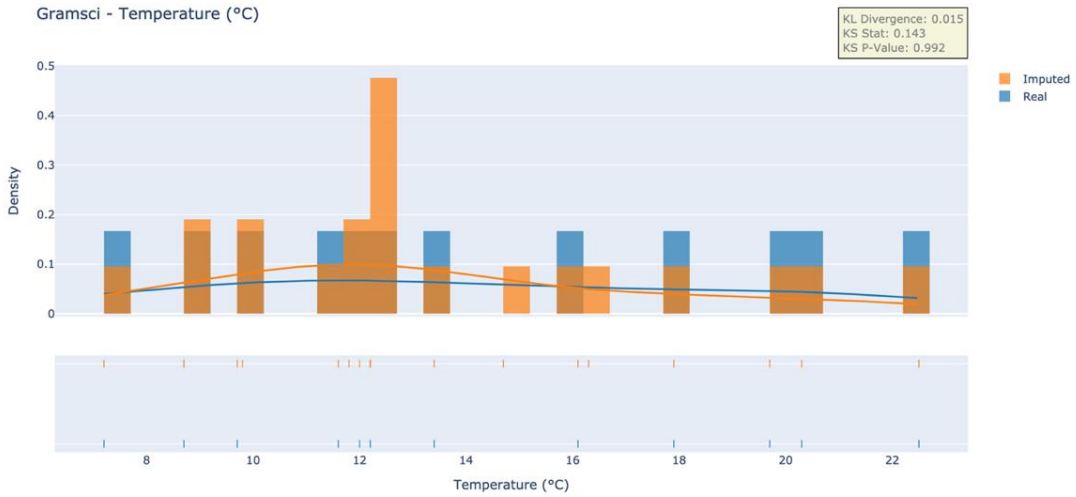


Figure 2. Similarity comparison between real data and imputed data after the MICE technique, for the temperature parameter of the Gramsci supply point. On the x-axis, there is the support of the parameter, and on the y-axis, the density of the measurements. The plot also shows the pdf estimation, together with the results of the tests previously described on the top-right.

Sensor measurements had no missing values; however, after the removal of outliers, the affected timestamps were filled with time interpolation (Pandas library; Mckinney, 2010).

b. Outliers

Outliers were present only in online sensor data and were identified through expert validation and visual inspection. Upper and/or lower thresholds were defined for each parameter and supply point, and values outside these ranges were excluded. The selected thresholds for each parameter are in the following list:

- Gramsci: Turbidity = 1.5 NTU, Conductivity = 400 µS/cm, UVA254 = 1.5 1/m,
- Berna: Turbidity = 1.5 NTU, Temperature = 19.5 °C, Conductivity = 400 µS/cm, UVA254 = 1.5 1/m,
- Bande Nere: Turbidity = 1 NTU, Conductivity = 400 µS/cm, Nitrate = 20 mg/L, UVA254 = 0.4 1/m,

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

- Fortunato: Turbidity = 1 NTU, Conductivity = 400 µS/cm, Free Chlorine = 1.0 mg/L, Nitrate = 25 mg/L, UVA254 = 0.4 1/m, TOC = 1 mg/L,
- Montevideo: Color = 4 CU, Turbidity = 1 NTU, Conductivity = 400 µS/cm, Free Chlorine = 1 mg/L, Nitrate = 20 mg/L, TOC = 1 mg/L, UVA254 = 4 1/m,
- Prealpi: Turbidity = 0.7 NTU, UVA254 = 1.5 1/m,
- Tabacchi: Nitrate = 14 mg/L, Temperature = 10 °C, Conductivity = 400 µS/cm, pH = 6,
- Tognazzi: Conductivity = 400 µS/cm, Free Chlorine = 0.4 mg/L.

All thresholds except those for conductivity and nitrate were upper bounds. The Tabacchi supply point follows a specific rule: for temperature and pH, thresholds act as lower bounds, while for other parameters in Tabacchi, the general rule applies. Figure 3 illustrates the effect of outlier removal on the temperature dataset for Tabacchi, showing improved data stability.

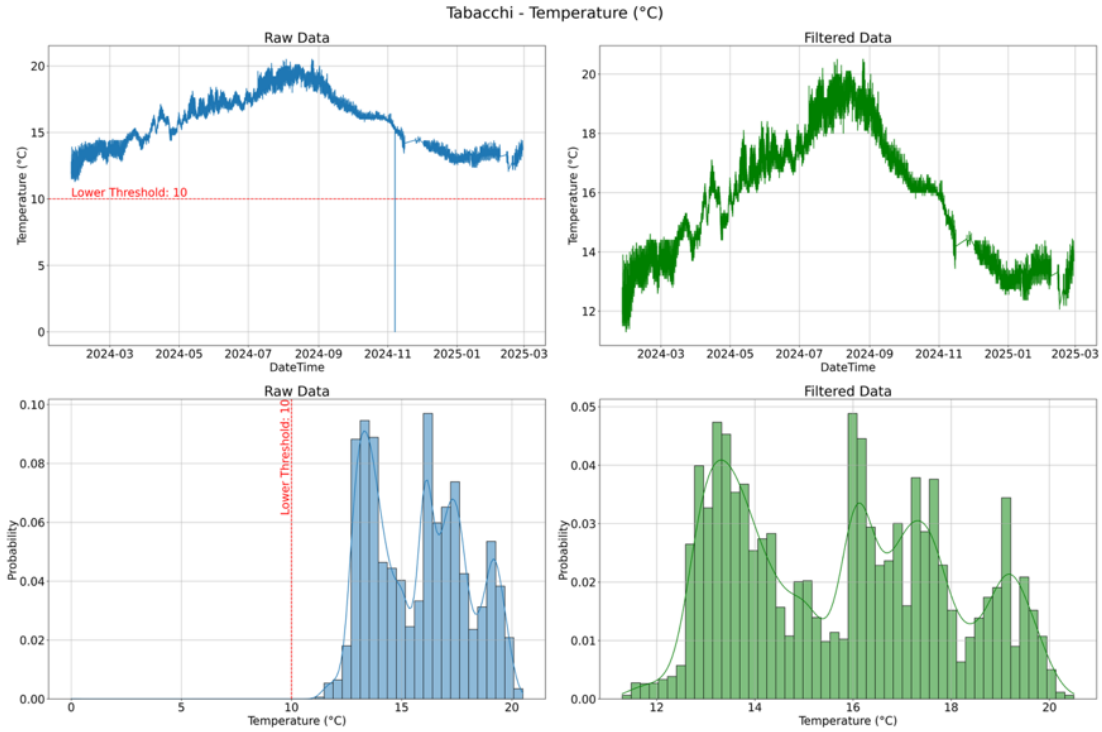


Figure 3. Comparison between the raw data before removing outliers (left) and after the removal (right). The top graphs show the timeseries of the analyzed parameter (temperature for the Tabacchi supply point) with the corresponding threshold on the left graph. The bottom graphs show the probability density function of the analyzed parameter, with the corresponding threshold on the left graph.

c. Censored data

Several grab sample parameters (colour, free chlorine, bromodichloromethane, dibromochloromethane, bromoform) contained censored data below LOQ. An initial attempt using MICE produced implausible results (imputed values above LOQ); therefore, censored data were replaced conventionally with LOQ/2, consistent with standard water-quality practice.

d. Similarity between grab samples and sensor measurements

Since the grab sample data were initially used for model training, and later sensor data for prediction, it was necessary to assess their similarity. For every parameter, each grab sample measurement was

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

paired with the closest sensor observation. To reduce potential noise from individual sensor readings, an hourly median centered around the selected timestamp was computed. Each resulting pair, consisting of the grab sample value and the corresponding hourly median sensor value, was compared through Bland–Altman analysis (Brazdzionyte and Macas, 2007). The method plots the difference between the two techniques versus their mean, with bias and Limits of Agreement (LoA = bias ± 1.96 SD) reported.

All parameters at all supply points showed a robust similarity between the two methods, confirming proper calibration (see Figure 4 for the Bland-Altman evaluation for the temperature example).

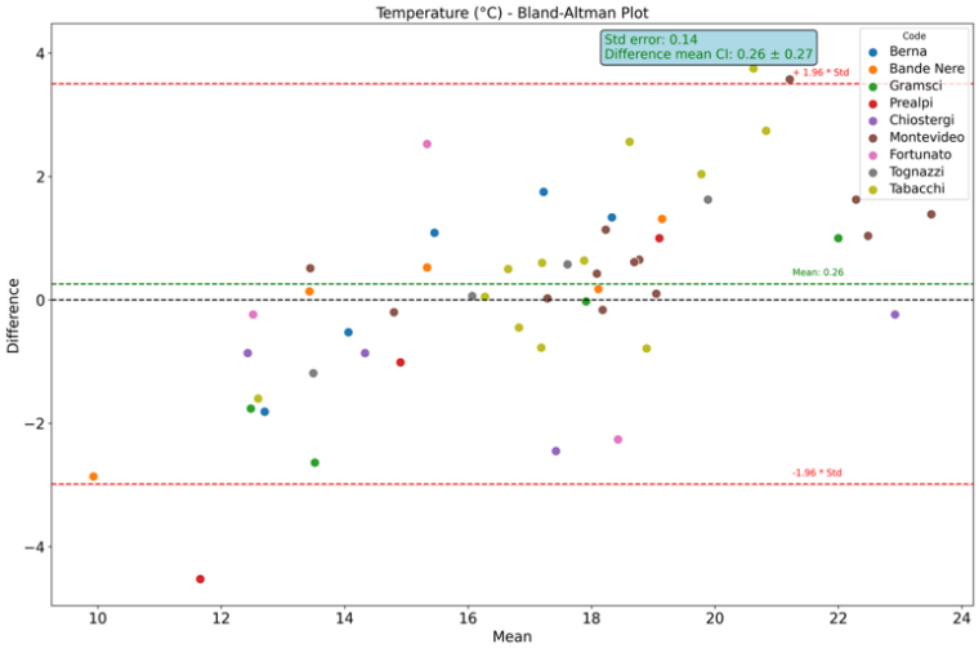


Figure 4. Bland-Altman plot of Temperature (°C). Each color references to a different supply point. On the horizontal axis the mean between the paired observations of the two methods, while on the y axis their difference. The green dotted line represents the mean difference between the measurements, while the two red dotted lines represent the LoAs.

The Bland–Altman approach was adopted to assess the similarity between grab-sample and sensor measurements, as the dataset contained no missing values, allowing direct one-to-one comparison of paired observations. This method could not be applied to imputed data because missing values break the pairing structure, preventing a consistent and statistically valid comparison between the two datasets.

2.1.2. Feltre DWTP

Data were collected at the Feltre DWTP from mid-February 2025 to the end of August 2025, using high-frequency (15-minute) online sensors installed at the post-disinfection stage of the treatment process. These parameters offer key insights into chemical stability, disinfection efficiency, and the dynamics of organic matter within the treatment system. Table 4 lists the monitored low-informative water quality parameters (used as input for the soft-sensors) and their summary statistics.

Table 4. CS#2 DWTP input data summary.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Parameter	Min	Max	Mean	Std
Pressure [atm]	0	2.5	0.53	0.47
TOCeq [mg/l]	0.28	0.58	0.45	0.1
DOCeq [mg/l]	0.13	0.23	0.19	0.03
Turbidity [FTU]	0.63	1.08	0.89	0.14
Conductivity [uS/cm]	635	694	668.79	11.84
Temperature [°C]	14.1	16.5	15.28	0.42
pH	6.78	7.37	7.21	0.13
Free Chlorine [mg/l]	0	0.31	0.07	0.03
Nitrate [mg/L]	12.86	21.08	17.67	1.71
UV254 [1/m]	0.62	1.53	1.15	0.28

Since no online sensors for TTHMs were available, additional microbiological information was obtained using the BactoSense automated flow cytometer, designed for real-time microbial monitoring. BactoSense system performs fully automated sample preparation, staining, and measurement at the sample point, detecting >99.9% of microbial cells larger than 0.1 µm and delivering results within 20 minutes, significantly faster than laboratory methods (hours or days). Sealed cartridges enable safe reagent handling and support up to 1,000 measurements, providing several weeks to months of autonomous operation, depending on the measurement frequency. For this study, Intact Cell Count (ICC) cartridges were used to quantify cells and related parameters, which are considered highly informative for soft-sensor development. The measurement interval was set to 6 hours, balancing operational cost and temporal resolution. This configuration allowed the system to provide timely detection of microbial variations while maintaining operational efficiency. However, the frequency can be configured between 30 minutes and 6 hours. Table 5 shows a summary of the BactoSense high-informative output parameters.

The preprocessing of this sub-case study was more straightforward compared to the supply points, as the dataset did not contain missing values or censored data. Outliers flagged as errors in BactoSense metadata were removed, and the resulting missing values were imputed using time-based interpolation with the Pandas library.

Table 5. CS#2 DWTP output data summary.

Parameter	Name	Min	Max	Mean	Std
ICC (1/mL)	Intact Cell Count	10	21060	1622.18	1925.17
HNAC (1/mL)	High Nucleic Acid Count	10	9960	540.16	938.06
LNAC (1/mL)	Low Nucleic Acid Count	10	12270	1081.93	1048.71
HNAP (%)	High Nucleic Acid Percentage	0.06	1	0.27	0.13

2.2. CS#3 Tarragona, Spain

The Tarragona case study, coordinated by EUT in collaboration with CAT, represents a disinfected coastal water supply system characterized by extensive online monitoring coverage. This case aims to validate the soft-sensor methodology under Mediterranean climatic conditions and in a complex distribution network, where DBP formation dynamics are influenced by variable temperatures, residence times, and chlorine decay. For this purpose, CAT provided continuous and historical data from its online sensor network, which integrates both high- and low-informative parameters measured through S::CAN multiparameter probes and other monitoring devices connected to the SCADA system.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

The dataset was progressively compiled and harmonized using Python-based workflows compliant with Smart Data Models (SmartWater specification), ensuring interoperability and consistency of time-series data. The following subsections describe the monitoring infrastructure, the dataset structure, and the pre-processing pipeline applied to prepare the data for soft-sensor development in Section 3.2.

2.2.1 Case study monitoring and sensors

Over the past few years, CAT has expanded the number of online water quality sensors throughout its network. This provides valuable information for the SafeCREW project.

The left panel of Figure 5 shows the spatial distribution of monitoring stations across the network, strategically located to maximize water-quality coverage. Currently, 27 water quality stations are active, including sites at the water catchment, the treatment plant, and 21 locations in the distribution network. The central and right panels of Figure 5 display, respectively, the S::CAN multiparameter sensor and an example of its monitoring platform. Although there are over 20 active stations with quality sensors, for the development of this task, two sample sites were used, because they matched the task's requirement and, to better understand the effect of the chlorination and its effects on generating DBPs. The data variables from both sites are identified under "_1" and "_2" (e.g. "CTR_2").

Different types of sensors are available; however, the primary source of data for this study comes from S::CAN instruments. These sensors concentrate most of the signals in their electronics. In this case, the electronic module sends online water quality data to SCADA via MODBUS TCP communications. The remaining sensors use serial communication via the CAT communications channel.

The dataset was derived from historical Citect SCADA data, sampled every 10 seconds, with automatic signal cleaning applied by the sensors. Dataset construction was performed using Python scripts developed in accordance with Smart Data Models (SmartWater/WaterQuality entity type) https://smartdatamodels.org/.

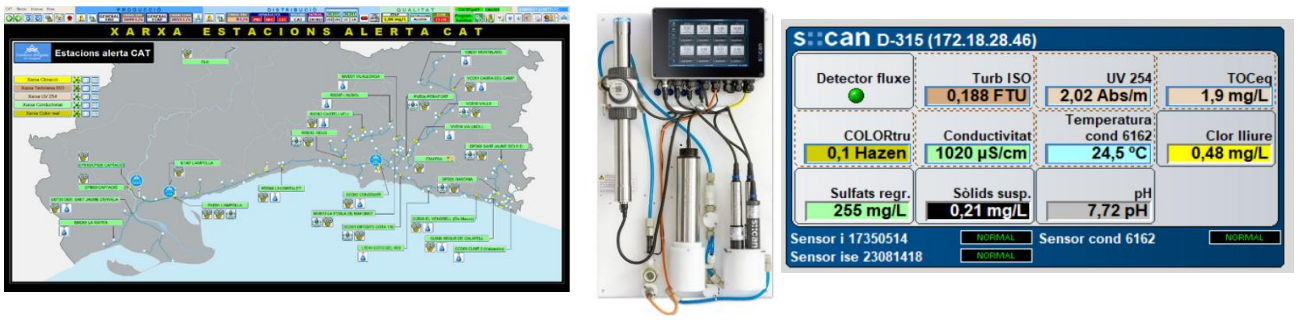


Figure 5. Distribution of online water quality stations (left), automatic S::CAN station (middle), standard monitored parmeters (right).

The online water-quality dataset generation process followed a structured workflow derived from SCADA trend data for the selected period:

- Data extraction: retrieval of analog values together with error and maintenance flags for each parameter, as well as associated flow and tank-level data for each monitoring location.
- Data cleaning: removal of null or duplicated records, invalid timestamps, and other basic inconsistencies.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

- Signal aggregation: computation of minute-averaged values for analog parameters and single-state signals.
- Quality labelling: each record was automatically flagged as *Valid (V)* when both error and maintenance signals equalled 0, and *Faulty (F)* when either flag equalled 1.
- Flow and tank-level validation: values were retained if within the configured min–max range in SCADA. For certain stations, flow data indicate inflow (*IN*) or outflow (*OUT*) depending on operational conditions; this direction attribute was added to the data model.
- Temporal aggregation: hourly averages were calculated using only records marked as *Valid (V)*.
- Final dataset generation: consolidation of validated, time-aligned measurements for subsequent analysis and modelling.

2.2.2 Dataset explanation

The available dataset (provided by CAT) was initially delivered in several batches, reflecting the gradual integration of high-informative sensors throughout SafeCREW project. Table 6 summarises the received raw-data files.

The provided dataset covers the period from 1 May 2024 to 1 September 2025, as additional batches of sensor data were progressively received over time.

The raw datasets, received in .xlsx format from CAT, contain time-stamped measurements of water quality parameters and TTHMs concentrations. The first step involved importing and harmonizing the spreadsheets into a unified time series database, ensuring consistent datetime formats, units, and variable names. Next, the data underwent quality control, which included the removal of duplicates, alignment of sampling intervals, and detection of anomalous or missing values. Outliers were flagged using robust statistical criteria and either corrected or interpolated as appropriate. To improve signal stability, selected variables were further smoothed using a Hampel filter, and derived variables (e.g., ratios, rolling averages, lags) were computed to capture process dynamics. The pre-processed, cleaned, and feature-enriched dataset served as the basis for model training and validation. A general statistical summary of the input data is presented in Table 6.

Table 6. General statistical description of the input data.

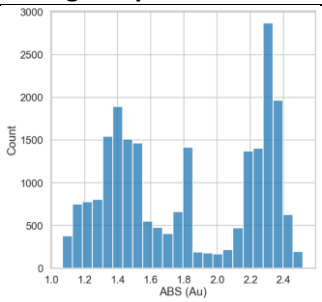
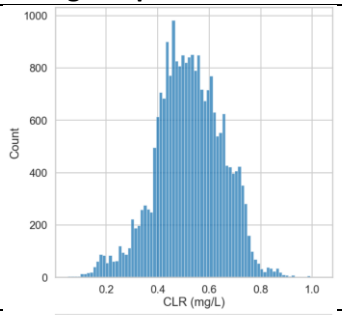
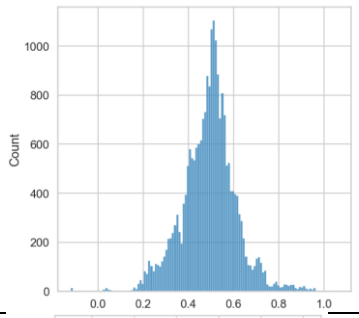
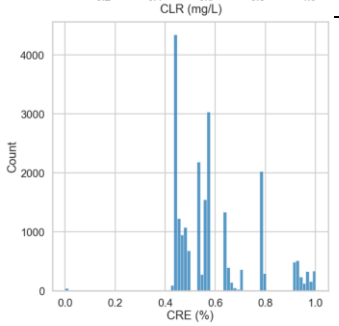
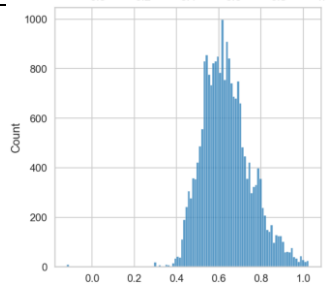
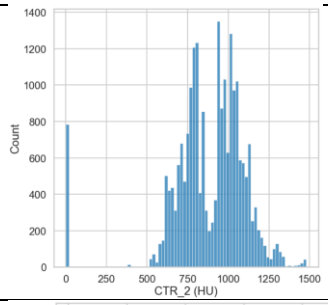
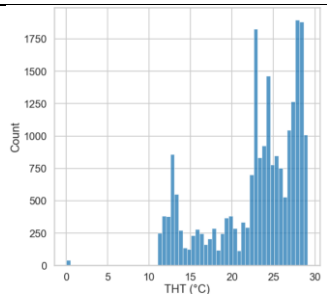
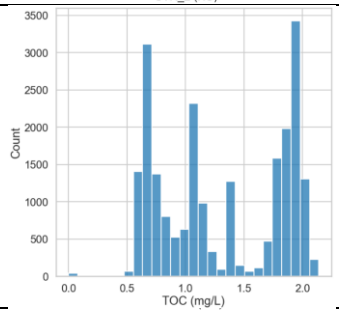
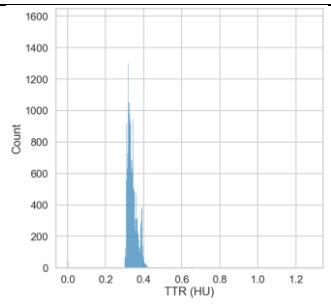
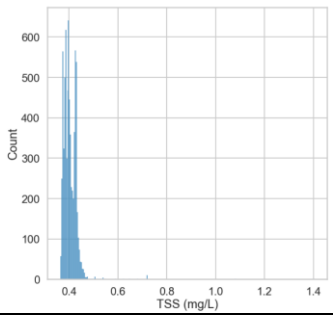
Input name	count	mean	std	min	25%	50%	75%	max	Frequency
ABS (Au)	22419.0	1.811	0.433	1.067	1.414	1.789	2.281	2.518	15 minutes
CLR_2 (mg/L)	22419.0	0.493	0.119	-0.12	0.424	0.501	0.558	1.06	
CLR_2.1 (mg/L)	22419.0	0.637	0.116	-0.12	0.554	0.625	0.704	1.026	
CLR (mg/L)	26835.0	0.525	0.129	0.051	0.439	0.522	0.616	1.029	
CRE (%)	22419.0	0.59	0.158	0.0	0.457	0.559	0.645	1.0	
CTR_2 (HU)	22417.0	877.604	238.121	0.0	764.0	928.0	1032.0	1480.0	
CTR (HU)	22417.0	921.727	174.023	0.0	812.0	912.0	1061.822	1384.0	
THT (°C)	22419.0	22.877	5.191	0.0	19.998	24.281	27.25	29.125	
TTR (HU)	22419.0	0.339	0.032	0.0	0.32	0.332	0.352	1.278	
THM (µg/L)	17372.0	47.688	20.045	0.0	33.500	48.933	63.000	114.000	
TOC (mg/L)	22419.0	1.311	0.529	0.0	0.758	1.176	1.886	2.139	
FTR (HU)	11725.0	53.116	22.686	-0.111	40.454	43.325	73.875	108.665	
TSS (mg/L)	7309.0	0.404	0.03	0.359	0.385	0.4	0.424	1.405	

To better understand the distributional characteristics of the data, exploratory data analysis (EDA) was conducted through histogram plots of the raw variables (see Table 7). These analyses enabled the identification of skewness and extreme values, highlighting that some features, particularly TSS and

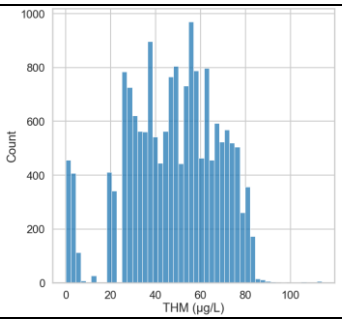
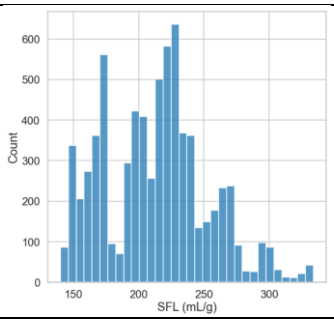
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

THM, exhibited extreme and infrequent values that necessitated dedicated treatment in the pre-processing pipeline.

Table 7. Histogram plots of raw features.

Feature name	Histogram plot	Feature name	Histogram plot
Absorbance (Au)		Chlorine (CLR, mg/L)	
Chlorine site 2 (CLR_2, mg/L)		Color (CRE, %)	
Chlorine site 2, sensor 1 (CLR_2.1, mg/L)		Conductivity site 2 (CTR_2, HU)	
Temperature (THT, °C)		Total Organic Carbon (TOC, mg/L)	
Turbidity (TTR, HU)		Total Solids in Suspension (TSS, mg/L)	

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Total Trihalomethanes (THM, ug/L)		Short Lifetime Fluorescence (SFL, mL/g)	

2.2.3 Data pre-processing summary

Before model development, the raw time-series data underwent a structured pre-processing workflow to ensure consistency and reliability across all monitoring stations. The procedure followed recognized practices for environmental time-series analysis (Imran *et al.*, 2022) and included the following steps:

- Outlier detection: a double-layered approach was implemented to identify and correct anomalies. Global anomalies were flagged using a robust z-score criterion to capture extreme deviations from overall distributional patterns. Local anomalies were corrected using a Hampel filter, which replaces values that deviate excessively from the rolling median within a short temporal window.
- Missing-data treatment: After outlier treatment, all remaining missing values were filled by their rolling window average to ensure consistency of the training dataset.
- Exclusion of unreliable THM data: records associated with known maintenance periods, sensor malfunctions, or data corruption were removed entirely to avoid bias in model calibration and validation.
- Temporal harmonization: finally, the cleaned signals were resampled to a 24-hour interval, harmonizing measurement frequency across parameters and aligning with the daily resolution adopted for model training and validation.

In the following, we detail the pre-processing pipeline.

Time alignment & resampling

The original dataset was available at a 15-minute sampling frequency, which provided high-resolution information but also introduced substantial short-term variability and noise. Initial experiments with a regression model at this fine temporal scale did not yield satisfactory predictive performance, due to the mismatch between high-frequency fluctuations in sensor readings and the slower dynamics of THM formation.

To address this, the data streams were first time-aligned to ensure consistent timestamps across variables and then resampled to a 24-hour interval. This daily aggregation both reduced noise and emphasized the underlying process trends, providing a more stable signal for modeling. On this basis, the predictive framework was redefined as a 24-hour classification task (low, medium, high THM levels), which aligned better with the temporal resolution at which operational decisions are typically made and improved model robustness.

Robust outlier detection

The z-score method was applied as a first step in outlier detection due to its efficiency and simplicity in identifying extreme deviations from the overall distribution of a variable.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

By standardizing each observation relative to the dataset mean and standard deviation, the z-score highlights points that lie unusually far from the expected range. This approach is computationally inexpensive, interpretable, and widely used in environmental data analysis, where anomalies often indicate sensor faults or atypical operating conditions. In our workflow, the z-score $z(t) = \frac{x(t)-\mu}{\sigma}$ served as a global filter, efficiently flagging gross errors and extreme outliers before applying more localized techniques such as the Hampel filter, thereby ensuring that the dataset used for modelling was both statistically robust and representative of normal process dynamics. An observation is considered an outlier if $|z(t)| > z_{\text{thr}}$.

Smoothing time series (Hampel filter)

The Hampel filter was employed as a third step in the preprocessing pipeline to provide robust smoothing of the time series after gross anomalies were identified by the z-score method. Unlike simple moving averages, the Hampel filter replaces a value only when it deviates excessively from the local median within a rolling window, scaled by the median absolute deviation (MAD). This makes it highly resistant to the influence of outliers and well-suited for environmental monitoring data, where isolated spikes or sensor glitches can distort trends. In our approach, the Hampel filter acted as a localized denoising method, ensuring that subtle variations relevant to THM dynamics were preserved while extreme fluctuations were removed. Used in combination with the global z-score detection, it provided a balanced strategy: z-score for efficient anomaly detection and Hampel filtering for robust smoothing, resulting in cleaner and more reliable input signals for the soft-sensor models. Considering the Hampel filter, the equation goes as follows:

$$x*(t) = \{\ m(t),\ \ if\ |\ x(t)\ -\ m(t)\ |\ >\ k\ \cdot\ MAD(t).$$

In this case, the values m(t) and MAD(t) were calculated like that:

$$m(t)\ \ =\ median(\ x[t-w:\ t+w]\ )$$
$$MAD(t) =\ median(\ |x[t-w:\ t+w]-\ m(t)|\ ).$$

Missing data treatment

Short gaps were imputed via rolling-mean interpolation within a symmetric window:

$$\hat{x}(t) = \frac{1}{N_t}\sum_{i=-w}^{w} x(t+i)\ 1_{\{\text{observed}\}},$$

This local approach preserves diurnal and seasonal patterns, thereby reducing bias in the reconstructed series.

2.3. CS#4 Western Ukraine, Ukraine

Water quality monitoring was conducted for three water supply systems located in the west of Ukraine. The following sampling points were selected for the monitoring activities:

- Site #1: (1) water abstraction well, (2) after the rapid filter, (3) after the clean water storage tank, where a sodium hypochlorite solution is dosed, (4–9) six points within the water distribution network (Figure 6).
- Site #2: (1) water abstraction well, (2–5) four points within the water distribution network (Figure 7).
- Site #3: (1) river near the wells, (2–3) two water abstraction wells, (4) after the clean water storage tank, (5–8) four points within the water distribution network (Figure 8).

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

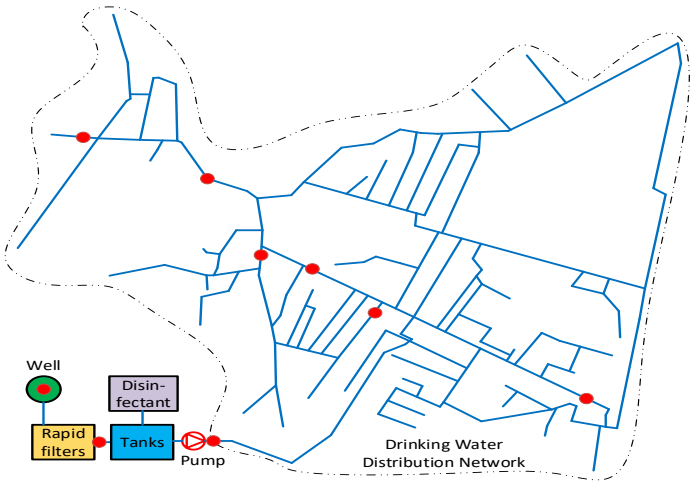


Figure 6. Sampling points (site #1)

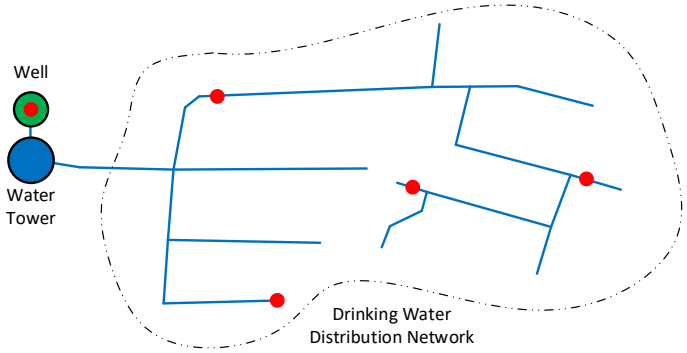


Figure 7. Sampling points (site #2)

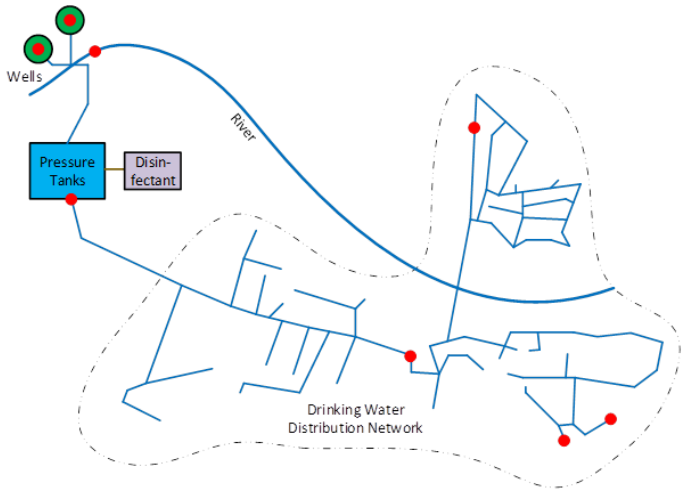


Figure 8. Sampling points (site #3)

Measured Parameters

The following water quality parameters were measured:

- Chemicals: Free chlorine, Bound chlorine, Chlorites (periodically since 2025), Chlorates (periodically since 2025), Bromides (periodically since 2025), Iron, Manganese, Calcium, Magnesium, Ammonium, Chlorides, Sulphates, Nitrates, Phosphates, Silicates, pH, Alkalinity,

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Permanganate oxidation ($KMnO_4$), Conductivity, UV Absorbance 254 nm (since 2025), Trihalomethanes, Temperature, Potassium + Sodium, ORP (since 2025), Turbidity (since 2025).

- Microbial: Total coliforms, *E. coli*, Enterococci, Coliphages, Salmonella, *Pseudomonas aeruginosa*, Eggs, larvae of helminths, Oocysts of intestinal pathogenic protozoa, Cysts of intestinal pathogenic protozoa.

According to Ukrainian regulations, microbial parameters – except for the Total microbial count – are determined qualitatively (Found/Not Found). The following parameters have been measured quantitatively: Total coliforms, *E.coli*, Enterococci, Coliphages, Salmonella, and *Pseudomonas aeruginosa*.

Sampling and Analytical Methods

The parameters pH, Conductivity, Temperature, and ORP have been measured on-site at the sampling locations. The parameters Free chlorine, Bound chlorine, Iron, Manganese, Calcium, Magnesium, Ammonium, Chlorides, Sulfates, Nitrates, Phosphates, Silicates, Alkalinity, Permanganate oxidation, UV Absorbance (254 nm), and Turbidity have been analyzed in the NUWEE hydrochemical laboratory. The parameters Trihalomethanes, Bromides, Chlorites, Chlorates, and microbial indicators have been analysed in the laboratories of the Rivne Regional Centre for Disease Control and Prevention.

The following equipment and methods (indicated in brackets near each parameter) have been used to measure water quality parameters:

- pH, ORP (EZODO 8200M)
- Temperature, Conductivity (GMH 3410)
- Chlorine (DSTU ISO 7393-1:2003 or Mi404)
- Iron (KFK-3 / DR 6000, MPM N 081/12-175-05)
- Manganese (KFK-3, MPM N 081/12-0107-03)
- Calcium (KFK-3 / DR 6000, DSTU ISO 6058:2003)
- Magnesium (KFK-3 / DR 6000, MPM N 081/12-0107-03)
- Ammonium (KFK-3 / DR 6000, MPM N 081/12-0106-03)
- Chlorides (KFK-3 / DR 6000, MPM N 081/12-0644-09)
- Sulphates (KFK-3 / DR 6000, MPM N 03-06-12)
- Nitrates (KFK-3 / DR 6000, MPM N 081/12-0644-09)
- Phosphates (KFK-3 / DR 6000, MPM N 081/12-0005-01)
- Silicates (KFK-3 / DR 6000, MPM N 081/12-0015-01)
- Alkalinity (DSTU ISO 9963-1:2007)
- Permanganate oxidation (MPM N 081/12-0016-01)
- UV Absorbance 254 nm (DR 6000)
- Potassium + Sodium (calculation method)
- Turbidity (KFK-3 / DR 6000, MPM N 03-06-17)
- Trihalomethanes (DSTU ISO 10301-2004)
- Chlorites, Chlorates, Bromides (DSTU ISO 10304-1:2003)
- Total coliforms (MPM N 10.2.1-113-2005)
- E. coli (MPM N 10.2.1-113-2005)
- Enterococci (MPM N 2285-81)
- Coliphages (MPM N 10.2.1-113-2005)
- Salmonella (MPM N 10.2.1-113-2005)

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Sampling Frequency

Water quality parameters for CS#4 have been measured monthly since July 2024, for a period of one year. An additional set of measurements is planned for each of the three monitored sites to expand the temporal coverage of the dataset.

Data Preprocessing

The procedure used for processing the data of CS#4 is the same as the ones described in Section 2.1.1, since, as will be explained in the following, the use of this data will be limited to validation purposes and will be fed into the same models developed for CS#1 in Milan.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

3. Methodological Framework for Soft-Sensor Development

In this section, building on data collected from the three case studies, Milan (CS#2), Tarragona (CS#3) and Western Ukraine (CS#4), we present the methodological framework that integrates statistical and ML techniques to transform monitoring data into predictive, actionable information, establishing relationships between conventional indicators (e.g., pH, temperature, conductivity, chlorine) and target contaminants such as disinfectants, NOM, and DBPs.

Across all sites, a common workflow was adopted for processing and modelling the data:

1. Exploratory Data Analysis: to align grab-sample and online sensor data, remove of outliers, treat missing values and censored data, and define input–output parameters, i.e., identify conventional low-informative parameters as predictors and selected contaminants as targets.
2. Model design and training: through the selection of suitable machine learning algorithms based on data structure and temporal resolution.
3. Evaluation: assessment of predictive accuracy, uncertainty, and feature relevance to validate model robustness.

To ensure reproducibility and interoperability across the case studies, all analyses were performed in a Python-based open-source environment, using a consistent suite of data science and ML libraries for preprocessing, model training, explainability, and performance evaluation. The main tools and their purposes within the framework are summarized in Table 8.

Table 8. Employed libraries. ROC: Receiver Operating Curve, PR: Precision-Recall.

Category	Libraries / Tools	Purpose
Exploratory Data Analysis	*Python, NumPy, Pandas, Matplotlib, Seaborn*	Data cleaning, transformation, visualization, and statistics
Model Design	*Scikit-learn, XGBoost/LightGBM, SHAP*	Model development, hyperparameter tuning, and explainability
Evaluation	*scikit-learn.metrics, Matplotlib, Seaborn*	Performance metrics, ROC/PR curves, and confusion matrices

While this structure was consistently applied, specific adaptations were required to reflect the operational and data contexts of each case study, that is summarized in Table 9. In fact, in Milan (CS#2), the focus was on disinfected systems, combining online and grab-sample data from both distribution network points and the Feltre DWTP, where additional microbiological data from flow cytometry (BactoSense) were available. In Tarragona (CS#3), the framework incorporated high-frequency online sensor data from a Mediterranean coastal system, emphasizing DBP (TTHMs) dynamics and network modelling for classification-based prediction of daily THM risk levels. In contrast, Western Ukraine (CS#4) applied the same workflow to a non-disinfected system, relying mainly on chemical and microbiological parameters collected through laboratory analyses, to test the generalizability of the models in lower-data-density contexts.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Table 9. Summary of the three case-studies context and the specific parameters for the soft-sensors development.

Case Study	Disinfection (Yes/No; Disinfectant)	Soft-sensor Location	Low-informative Parameters (source)	High-informative Parameters Modelled
CS#2 Milan (Italy)	Yes; Sodium hypochlorite	Treatment plant Distribution network	Temperature (sensor), pH (sensor), Conductivity (sensor), Nitrate (sensor), Free chlorine (sensor), Turbidity (sensor), UVA254 (sensor), Color (grab), TOC (grab), flow cytometry (sensor)	Total Trihalomethanes (TTHMs: chloroform, bromodichloromethane, dibromochloromethane, bromoform) at supply points; Microbial indicators (Intact Cell Count, High/Low Nucleic Acid Cells, HNAP%) at DWTP
CS#3 Tarragona (Spain)	Yes; Sodium hypochlorite	Distribution network	Temperature (sensor), Conductivity (sensor), pH (sensor), Turbidity (sensor), Chlorine residual (sensor), TOC (sensor), Absorbance (sensor), Color (sensor)	Total Trihalomethanes (TTHMs – DBPs)
CS#4 Western Ukraine (Ukraine)	No	Treatment plant Distribution network	Temperature (grab), Conductivity (grab), pH (grab), Nitrate (grab), Iron (grab), Manganese (grab), UV254 (grab), ORP (grab)	Microbial indicators (E. coli, Total coliforms, Enterococci, Coliphages, Salmonella, P. aeruginosa) and selected chemical indicators (e.g., permanganate oxidation, THMs for validation)

In the next paragraphs the detailed information related to the application of this common framework to build case specific soft-sensors is explained.

3.1. CS#2 Milan, Italy

For the Milan case study, two complementary methodological approaches were developed, reflecting the existence of two sub-case studies previously described: (i) the supply points in the DWDN based on grab samples and online sensor measurements data, and (ii) the Feltre DWTP where high-frequency online monitoring data were integrated with microbiological indicators from BactoSense. These two contexts differ in data availability, sampling frequency, and target variables, which required the design of distinct modeling strategies for soft-sensor development.

3.1.1. Supply Points in the DWDN

A critical challenge in developing soft-sensors for the Milan distribution network was the limited availability of grab sample data for most supply points. Training an individual model for each point was unfeasible, as it would result in severe data sparsity and overfitting. Conversely, using a single global model for all supply points would require assuming that water quality dynamics are homogeneous across the entire network, which is unrealistic and could lead to poor generalization due to heterogeneity and noise in the dataset.

To address this issue, we adopted a clustering strategy to group supply points exhibiting similar water quality behavior. This approach allowed us to train one soft-sensor per cluster, balancing the need for sufficient training data with the requirement to capture local variability. Clustering was performed using both input and target variables, leveraging all available information. The chosen method was Constrained *k*-means (COP-kmeans) (Bradley, Bennett and Demiriz, 2000), a variant of the traditional *k*-means algorithm that incorporates pairwise constraints to guide the clustering process:

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

- Must-link constraints: two points must belong to the same cluster.
- Cannot-link constraints: two points must belong to different clusters.

In our case, the must-link constraint was applied to ensure that all samples from the same supply point were assigned to the same cluster, maintaining operational and physical consistency. This converts the problem from clustering individual samples to clustering groups of samples (supply points), where inter-group distances were computed using aggregated feature means.

The optimal number of clusters *k* was selected through the elbow method, applied to the weighted within-cluster variance:

$$WCV(k) = \left(\frac{1}{N}\right)\sum_{c=1}^{k}\sum_{x_i \in C_c} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_c\|^2,$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ are the multivariate observations and $\bar{\boldsymbol{x}}_c$ is the empirical centroid of cluster *c*. After determining the optimal *k*, the COP-kmeans algorithm was applied to partition the supply points accordingly.

For each cluster, ML models were trained using an 80/20 train-test split, with input parameters defined in Table 1 and the target variable (TTHMs) in Table 2. Input variables parameters were scaled into [0, 1] interval, to ensure that they have the same importance in magnitudes and prevent dominance of high-scale parameters (e.g., conductivity), ensuring that low-scale (e.g., free chlorine) were equally represented during training.

Different ML models were tested to identify the one that best fits and generalizes the datasets. Four models were analyzed, from simpler to more advanced approaches: Partial Least Squares, Support Vector Regressors, XGBoost, and Quantile Regression Neural Networks.

Partial Least Squares (PLS) (Pirouz, 2012) reduces the dimensionality of correlated predictors by projecting them, together with the target variable, into a latent space that maximizes their covariance. The procedure identifies components (latent variables) that explain the most variance in the predictors PLS is well suited for water quality datasets because many predictor variables are correlated, and the number of samples is relatively small. By working with latent variables, PLS avoids problems of multicollinearity and produces more robust predictive models.

Support Vector Regressor (SVR) (Smola, Scholkopf and Scholkopf, 2004) is the regression counterpart of Support Vector Machines (SVMs). Instead of classifying data, SVR estimates a function that approximates the relationship between input variables and a continuous target variable, within a defined error margin ε. By using kernel functions, SVR efficiently models both linear and non-linear dependencies and is robust to noise and high-dimensional inputs.

Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016) is an ML algorithm based on the principle of boosting, where multiple weak learners are combined sequentially to form a strong predictive model. While XGBoost is most associated with decision trees as base learners, it also allows the use of linear models (linear regression or logistic regression) as boosters.

Quantile Regression Neural Networks (QRNNs) (Jantre, Bhattacharya and Maiti, 2020) extend conventional neural networks by estimating conditional quantiles (e.g., 0.1, 0.5, 0.9) given predictors, rather than single mean predictions. QRNNs enable not only point predictions of contaminants (e.g., TTHMs) but also uncertainty estimates, which are crucial in water quality monitoring. Unlike linear quantile regression, QRNNs can capture nonlinear relationships between predictors and the response,

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

and they also provide flexibility, as the same model can output multiple quantiles simultaneously, offering a richer description of possible outcomes.
Each of the previous ML models contains external configuration variables, called hyperparameters, which must be set before training. They are manually set, and the process of finding the optimal values for each hyperparameter is referred to as hyperparameter tuning.
The main hyperparameters tuned for each model were:

- PLS**:** number of latent variables; tolerance used as the convergence criterion in the power method.
- SVR: kernel type; regularization parameter (*C*); epsilon-tube width (ε) defining the no-penalty region; and the kernel coefficient (*γ*) for RBF, polynomial, and sigmoid kernels.
- XGBoost: number of estimators (boosting rounds); learning rate; L1 alpha (α) and L2 (λ) regularization terms; and the liner model algorithm updater.
- QRNN: number of layers; number of units per layer; activation function of the neurons; and batch size per iteration.

To enable an unbiased comparison across models and configurations, and to mitigate data scarcity, the Leave-One-Out (LOO) cross-validation was adopted, with the Root Mean Squared Error (RMSE) as the evaluation metric. It is a special case of *k*-fold cross-validation where the number of folds equals the number of samples. In this method, each observation is used once as a test case while all remaining data are used for training. The process is repeated for all samples, and the final RMSE is obtained by averaging the prediction errors across iterations. This strategy ensures that all samples contribute to both training and testing, providing an almost unbiased estimate of predictive performance. An independent test dataset was subsequently used to evaluate the models on unseen data. Predictive uncertainty was quantified by computing 95% confidence intervals around the estimated values. Model interpretability was further investigated through a feature importance analysis using the SHAP framework (Lundberg and Lee, 2017). SHAP, based on Shapley values from cooperative game theory, quantifies the marginal contribution of each feature to the model output, providing both global and local interpretability.

Finally, once validated, the trained models were applied to real-time sensor data, generating TTHMs predictions every fifteen minutes. The same preprocessing steps adopted for grab-sample data were applied to the online sensor inputs to ensure consistency.

3.1.2. Feltre DWTP

For the Feltre DWTP, the soft-sensor was developed to predict ICC (mg/L) as the target variable. Unlike the DWDN case, the larger volume and higher frequency of data enabled additional feature-engineering and modeling steps. To address temporal dependencies inherent to high-frequency monitoring data, additional predictors were generated through time-lagged and rolling-window transformations (up to 6 hours). The resulting dataset comprised ≈ 80 input variables. A feature-selection step based on Granger causality (Shojaie and Fox, 2021) was applied to retain only variables whose lagged values provided statistically significant predictive information on the target. This reduced the feature set to 22 inputs**.** The central idea is that if the past values of a variable *X* improve the prediction of another variable *Y*, beyond what can be achieved using only past values of *Y*, then *X* is said to Granger-cause *Y*. The dataset was split into 60% training and 40% testing subsets.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

In addition to the previously adopted models (XGBoost, QRNN), two further algorithms were implemented to exploit the richer temporal structure of the data: LightGBM and LSTM.

LightGBM (LGBM) (Ke *et al.*, 2017) is a tree-based gradient boosting framework that builds decision trees in a leaf-wise manner, prioritizing the splits that lead to the largest reduction in loss. Linear trees were employed as a strategy to mitigate the extrapolation problem commonly associated with decision trees. This way, the model better handles scenarios where the relationships between input variables and the target variable extend beyond the observed data range. The training process minimizes a regularized loss function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \Omega(\text{Tree}),$$

where $l(y_i, \hat{y}_i)$ is the loss function (squared error in this case), and $\Omega(\text{Tree})$ penalizes overly complex trees to avoid overfitting. It provides high accuracy by capturing non-linear relationships and naturally handles missing values and categorical features, making it robust for real-world monitoring data. Being a tree-based method, it inherently performs feature selection, avoiding the above-described feature selection step.

Long Short-Term Memory (LSTM) Networks (Staudemeyer and Morris, 2019) are a class of Recurrent Neural Network (RNN) designed to learn temporal dependencies, by maintaining an internal memory state. Each LSTM unit is built around a cell state $\boldsymbol{c}_t$ (the memory) and a hidden state $\boldsymbol{h}_t$ (the output at time *t*). Information flow is controlled by three gates: forget, input, and output**,** which regulate the flow of information over time. LSTMs are particularly suited for time-series data, such as water quality sensor measurements, where past conditions influence current outcomes. They can capture long-term dependencies) better than classical RNNs. In the context of soft sensors, LSTMs enable the direct integration of temporal patterns into predictive models, making them more robust for forecasting.

The hyperparameters for both models were optimized through the same tuning strategy described in Section 3.1.1. Like the supply points case study, cross-validation was employed to minimize bias in the comparison between models. In this case, a 5-fold cross-validation was applied to the training dataset, using the TimeSeriesSplit method from *scikit-learn* (Pedregosa *et al.*, 2011) to account for the temporal ordering of the data and avoid information leakage. As a final evaluation step, a test dataset was set aside to assess model performance on unseen data. To incorporate predictive uncertainty, 95% confidence intervals were computed alongside the point estimates, providing both an expected value and a range of uncertainty for each prediction.

3.2. CS#3 Tarragona, Spain

Based on the dataset and monitoring infrastructure described in Section 2.2, the Tarragona case study (coordinated by EUT in collaboration with CAT) focused on the development and validation of a data-driven soft-sensor for estimating and classifying daily future DBPs expressed as TTHMs risk levels. This work emphasizes the use of state-of-the-art online water quality sensors combined with advanced data-driven modelling techniques to try to capture the complex dynamics of THMs formation. Aside from an accurate prediction, the research aims to ensure the timeliness of alerts and minimize false alarms.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

3.2.1 Methodological framework

The modelling workflow followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology (Ncr and Clinton, 1999), ensuring a transparent and replicable structure for data-driven model development across six iterative phases: business and data understanding, preparation, modelling, evaluation, and deployment, ensuring methodological transparency and consistency. The process begins with defining objectives and understanding the problem context, followed by data acquisition, exploration, and transformation to ensure analytical quality. Subsequent phases involve model construction and performance evaluation to verify alignment with project goals, culminating in the deployment of validated models for operational use. Its iterative nature allows continuous refinement as new insights or data become available.

Early tests with regression models for predicting continuous THM concentrations were discarded, as reliable forecasts beyond one hour required inclusion of past THM values in training. Following consultation with CAT, the approach shifted toward a classification-based soft-sensor, estimating daily THM risk levels (low, medium, high) to enable timely early warning actions. (Bergman *et al.*, 2016). This design achieved 24-hour predictive capability while maintaining interpretability, reproducibility, and a balanced trade-off between sensitivity and false-alarm control, in line with good ML practice for environmental monitoring.

3.2.2 Input and data preprocessing and aggregation

Raw sensor data were obtained at sub-hourly frequency and harmonized through a multi-step cleaning process. All available input variables were initially retained to train the models and to achieve a consistent estimation of THMs over time.

An exploratory Pearson correlation analysis was carried out to assess both the linear association between each candidate variable and the TTHMs (µg/L) and the mutual relationships among predictors. The correlation coefficients between individual variables and THMs are reported in Table 10, while the full pairwise correlation structure is shown in the heatmap (Figure 9)**.** Coefficients were computed on the complete daily-resampled dataset before model development, providing a descriptive overview of relationships within the data. Considering the correlation mapping and aiming to exploit all available information, feature selection was postponed until the end of the feature-engineering phase; therefore, no features were discarded at this stage.

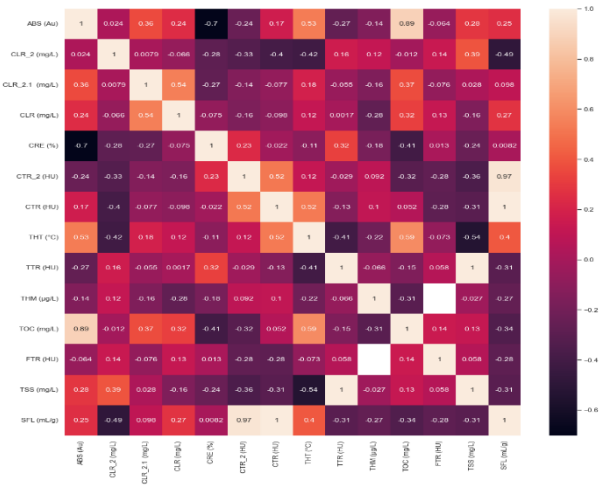


Figure 9. Pearson correlation heatmap.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Table 10. Pearson correlation value of all variables against THMs.

Candidate	THM (µg/L) Pearson correlation
ABS (Au)	-0.14
CLR_2 (mg/L)	0.12
CLR_2.1 (mg/L)	-0.16
CLR (mg/L)	-0.28
CRE (%)	-0.18
CTR_2 (HU)	0.09
CTR (HU)	0.10
THT (°C)	-0.22
TTR (HU)	-0.07
THM (µg/L)	1.0
TOC (mg/L)	-0.31
TSS (mg/L)	-0.03
SFL (mL/g)	-0.2727

Inclusion criteria

During model design, the inclusion of features followed these criteria:

- Features must improve out-of-sample performance under forward-chaining (time-series) cross-validation.
- Contributions are assessed with a nonlinear classifier (gradient-boosted trees / random forest) to capture interactions and non-additive effects.
- Selection is performed via recursive feature elimination to control overfitting by repeatedly refitting the model on reduced subsets.
- The final 50 features are chosen at the RFE plateau that maximizes validation macro-F1 (ties broken by model parsimony). Post-selection, we sanity-check importance with permutation importance on a chronological hold-out to ensure the selected features provide real predictive signal rather than spurious correlation. This procedure yields a statistically defensible subset with better bias–variance trade-off, lower multiple-testing risk (via Cross validation), and improved transferability.

Target definition

We discretized continuous THM measurements into Low/Medium/High using the empirical tertiles of the training data. This data-driven approach avoids arbitrary cutoffs when site-specific or regulatory thresholds vary, helping to balance class sizes, which in turn improves model stability and recall across classes. The thresholds are recalculated when new data are received to reflect seasonal and operational changes.

While this method improves balance and robustness, it may compromise interpretability outside the site-specific context and necessitate model recalibration when transferred to other systems.

3.2.4 Feature engineering

During model development, we observed that using the raw sensor variables as they were did not provide sufficient predictive power, motivating the need for a more sophisticated feature engineering strategy.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Guided by state-of-the-art research in water quality modelling and time-series prediction, we expanded the input space with features designed to capture both process dynamics and seasonal patterns. The following transformations were applied:

- Interaction features (products of key water quality parameters) to account for nonlinear relationships influencing THM formation.
- Lag features (at 3, 6, 12, and 24-hour offsets) to incorporate temporal dependencies and delayed effects.
- Rolling statistics, such as rolling standard deviation (for 3-, 6-, 12- and 24-hour windows), alongside mean rolling values (for 3-, 6-, 12- and 24-hour windows), to reflect variability and short-term fluctuations.
- Feature differences (e.g., successive changes between consecutive measurements) in the form of subtraction and process-informed indicators like chlorine decrease and decay rates, which are critical drivers of THM formation potential.
- Calendar/time features such as sine-cosine transformations of day, month indicators, and other seasonal encodings, to model cyclic and long-term patterns.

All features were normalized to [0, 1] to prevent scale dominance. Multicollinearity was checked through updated Pearson and Variance Inflation Factor (VIF**)** analysis before model training. The final engineered dataset consisted of roughly 90 variables, subsequently reduced to ≈ 50 through the RFE procedure described above.

3.2.5 Model design

Previous iterations

In an earlier iteration, we experimented with a regression-based approach to directly predict continuous THM concentrations from the 15-minute resolution data, as shown in Figure 10. This model also included past values of THM concentrations inside its training data, which was a measure taken to improve the model's performance, more specifically, the *LightGBM*.

While this provided reasonable short-term fits, the method proved inadequate for our operational requirements, since the goal of the early warning system was to provide at least 24-hour predictions.

The regression model struggled to generalize at this longer horizon, amplifying noise and short-term fluctuations rather than capturing the broader process dynamics that drive daily THM formation. The discarded regression model was only able to achieve a good prediction performance at the next 15 or 30 minutes.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

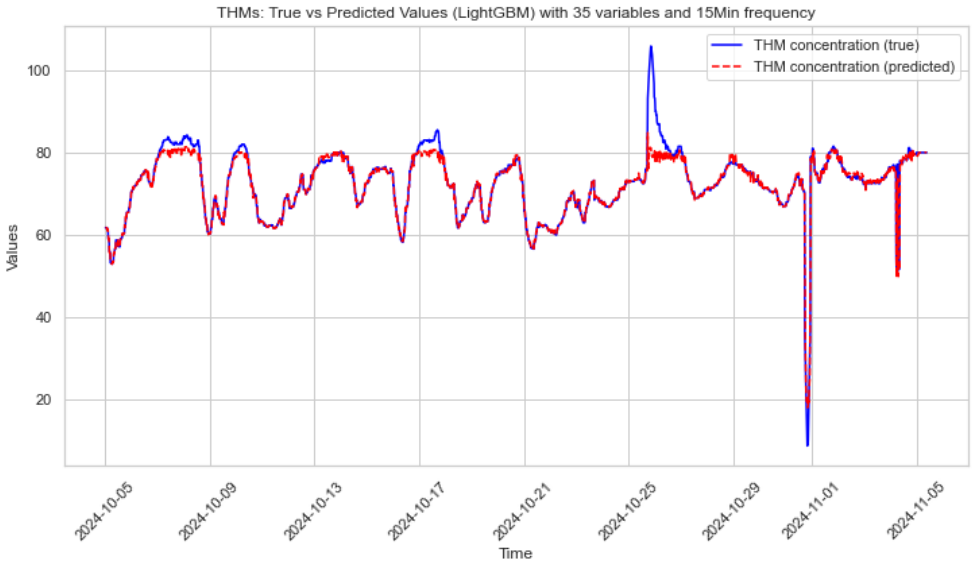


Figure 10. Previous iteration regression model results.

This limitation motivated a shift toward a 24-hour classification framework, where inputs were aggregated and engineered to reflect daily patterns, and outputs were simplified into categorical risk levels (low, medium, high). This redesign improved robustness, interpretability, and alignment with the decision-making timescale of plant operators.

Training protocol

Model development followed a rigorous protocol ensuring reproducibility, robustness, and fair comparison across algorithms. All processing steps–data cleaning, aggregation, feature generation, and model fitting–were embedded in a unified pipeline to guarantee reproducibility and prevent data leakage. Multiple classifiers (logistic regression, SVM, random forest, gradient boosting, LightGBM) were compared using time-aware cross-validation (expanding-window scheme). Hyperparameters were tuned through Bayesian optimisation to minimise the validation log-loss while monitoring macro-F1. Preprocessing stages (imputation, scaling, feature construction) were fitted exclusively on training folds and applied to validation/test data within the pipeline. The Gradient Boosting Classifier(Friedman, 2001) was selected as the operational model due to its capacity to capture non-linear dependencies, handle heterogeneous features, and maintain robustness on tabular water-quality data.

Model structure

The Gradient Boosting Classifier is an ensemble learning algorithm that builds a series of weak learners, usually decision trees, in a sequential manner. Each new tree is trained to correct the errors (residuals) made by the previous ensemble, and the final model is obtained by combining all trees into a weighted sum. By iteratively reducing the loss function through gradient descent, gradient boosting achieves high predictive accuracy and can capture nonlinear relationships and complex interactions between features.

Although gradient boosting could be considered a “black box” model, explainability tools make its predictions more transparent. Feature importance (based on split frequencies or permutation tests) reveals which variables contribute most to classification decisions, while SHAP values provide local explanations showing how each feature influences individual predictions. This focus on explainability ensures that the model is not only accurate but also interpretable, enabling operators to understand

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101081980.

why a given water quality profile is classified as low, medium, or high THM risk. A diagram showing the internal structure of working of the Gradient Boosting is shown in Figure 11.

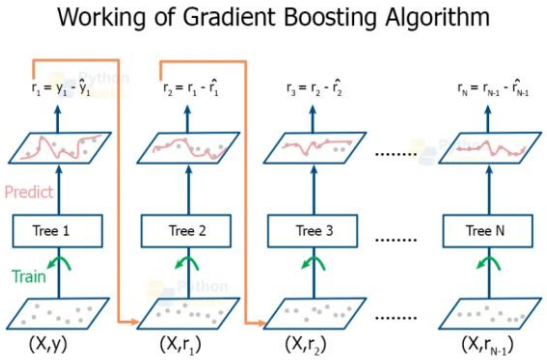


Figure11. Structure of Gradient Boosting Algorithm.

Explainability

Explainability (Linardatos, Papastefanopoulos and Kotsiantis, 2020) was a central aspect of the model design, as early warning systems must be both accurate and transparent to gain trust from operators and stakeholders. To this end, we placed strong emphasis on analyzing feature importance, which provided valuable insights into the variables driving the classification of THM risk levels. Global importance measures (such as impurity-based and permutation importance from the gradient boosting model) highlighted that parameters related to chlorine decay, absorbance, and temperature interactions were consistently among the most influential predictors. This confirmed the scientific plausibility of the model, as these variables are well-documented in the literature to govern THM formation.

Beyond global patterns, we also explored local explanations using SHAP, which allowed us to understand how individual features contributed to specific predictions. This capability is particularly valuable in operational contexts, as it enables operators to see not only *what* the model predicts but also *why*. For example, a high THM classification could be linked to elevated UV254 absorbance and reduced chlorine residuals, aligning with process expectations.

Overall, feature importance analysis and explainability tools proved essential for validating the model's credibility, supporting both the scientific rationale of the soft-sensor and the practical confidence needed to use it as guidelines in a water treatment early warning system. This will be discussed further in the sections related to the results and conclusions.

Feature importance

Feature importance (Kaneko, 2023) is an explainability technique that quantifies the contribution of each input variable to a model's predictions. In tree-based models such as gradient boosting, importance can be measured by how frequently and effectively a feature is used to split the data, or by the impact on model performance when the feature is perturbed.

This analysis is particularly valuable in the context of THM prediction, as it helps to pinpoint the water quality parameters and operational factors that most strongly influence THM formation. Identifying these key drivers not only increases trust in the soft-sensor but also provides actionable insights for operators, who can focus monitoring and control efforts on the variables that matter most for preventing exceedances. Since it is a multi-categorical model, it is possible to extract the importance

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

per kind of output, as shown in the feature importance summary plot for each label (see Table 18 in Section 4.2).

3.2.6 Validation & performance evaluation

Model validation was conducted using a time-series cross-validation approach with chronological folds, ensuring that temporal dependencies were preserved and avoiding information leakage. This scheme follows a forward-chaining (expanding-window) configuration, in which the model is trained on progressively increasing data segments and tested on subsequent unseen periods. This strategy reflects realistic operational deployment conditions, where future predictions are made based solely on past information.

To ensure robustness and fair comparison among candidate algorithms, all preprocessing steps (data cleaning, scaling, feature engineering) were fitted exclusively on the training folds and then applied to validation and test data within the same pipeline. This strict separation prevented data leakage and ensured the reliability of the performance assessment.

Given the multiclass classification setting (Low, Medium, High THM-risk), the model's performance was evaluated using a comprehensive suite of standard metrics (Chen *et al.*, 2024) using a suite of standard metrics:

- Accuracy: overall proportion of correctly classified instances.
- Precision (per class): fraction of predicted class instances that were correct, reflecting false alarm control.
- Recall (per class): fraction of true class instances correctly identified, critical for minimizing missed high-risk events.
- F1-score (per class): harmonic mean of precision and recall, balancing both aspects.
- Macro-averaged F1: treats all classes equally, ensuring balanced performance across Low, Medium, and High.
- ROC-AUC and PR-AUC (one-vs-rest): probability-based measures to evaluate class separation and robustness under class imbalance.

Among these metrics, recall for the High-risk category and macro-F1 were the principal evaluation indicators, as they best capture the safety-critical and balanced aspects of the soft-sensor's objectives.

Confusion matrix

The confusion matrix was used as the main diagnostic tool to evaluate classification performance. It summarizes the number of correct and incorrect predictions across the three THM categories (Low, Medium, High), offering a detailed view of how the model performs for each class. This matrix highlights not only the overall accuracy but also the specific types of misclassifications. In this case, the primary objective was to ensure that "High" THM risk events were correctly identified, since missing these critical cases could compromise the reliability of the early warning system. Therefore, while balanced performance across all classes was important, the evaluation placed particular emphasis on achieving high recall for the High class, even at the cost of accepting a slightly higher false-alarm rate in the other categories.

ROC curves

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

The ROC curve (Fawcett, 2006) analysis complemented the confusion matrix by providing a probabilistic evaluation of class separability. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across varying probability thresholds, while the Area Under the Curve (AUC) quantifies the overall discriminative capability. The ROC curve is particularly important in the context of an early warning system, as it allows us to evaluate the trade-off between sensitivity (recall) and false alarms, ensuring that high-risk events are reliably detected without overwhelming operators with unnecessary alerts.

The methodological framework for validation followed the principles of reproducible ML evaluation for environmental time-series modelling (Ferrer, Scharenborg and Bäckström, 2024), ensuring that the reported performance reflects genuine generalization ability rather than random or transient patterns. All quantitative performance results, together with visual evaluation outputs (confusion matrices, ROC and PR curves), are presented and discussed in Section 4.2

3.3. CS#4 Western Ukraine, Ukraine

Since the amount of data available for CS#4 was very limited and sparse in time, the only feasible approach was to use a soft-sensor pretrained on data having similar measurements and relationships and apply it to the CS#4 grab samples data. To foster this approach, soft-sensors from CS#2 (supply points subcase study) and the related data were utilized.

The first step involved understanding and quantifying the similarity between the three Western Ukraine sites and the Milan supply points clusters. This was achieved by labeling each site measurement to one of the two supply points' clusters according to the Mahalanobis distance (De Maesschalck, Jouan-Rimbaud and Massart, 2000) between parameters in CS#4 sampling points and the distribution of Milan clusters.

The Mahalanobis distance is a multivariate distance metric that measures the distance of a point from the mean of a distribution, considering the correlations among variables. Unlike the Euclidean distance, which treats all variables as independent and equally scaled, the Mahalanobis distance adapts to the covariance structure of the data. For a point x and a distribution with mean vector μ and covariance matrix Σ, the Mahalanobis distance is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}.$$

If variables are uncorrelated and have the same variance, the Mahalanobis distance reduces to the Euclidean distance.

After assessing the similarity with the clusters, the data were scaled using the same procedure applied in CS#2, and inference was performed with the soft-sensor previously trained on the clusters that the data were classified into.

4. Results

This section presents the outcomes of soft-sensor development and validation across the case studies, demonstrating both the potential and limitations of these approaches for water quality monitoring. The results are interpreted in the context of practical water system management, highlighting how soft-sensors can enhance monitoring capabilities, reduce operational costs, and support proactive decision-making.

The soft-sensor approach offers several key advantages for water utilities: continuous monitoring of parameters that traditionally require laboratory analysis, early detection of quality changes that might indicate contamination or treatment failures, and cost-effective expansion of monitoring coverage

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

across large distribution networks. However, the results also reveal important limitations and considerations for operational deployment.

4.1. CS#2 Milan, Italy

4.1.1. Supply Points in the DWDN

As previously explained in Section 3, the first step in the pipeline for soft-sensor development involved clustering. COP-kmeans was the chosen approach, an extension of k-means where constraints about the data can be defined.

In this setting, only Must-Link constraints were enforced to ensure that all the samples from the same supply point were assigned to the same cluster. This constraint converts the problem from clustering individual samples to clustering groups of samples (supply points), where the distance between groups is computed using aggregated feature statistics (in this case, the mean).

The clustering analysis using COP-kmeans successfully partitioned the 9 supply points into 3 distinct clusters based on their water quality characteristics. The elbow method indicated that k = 3 was the optimal number of clusters, as shown in Figure 12, where the addition of further clusters provided diminishing returns in variance reduction.

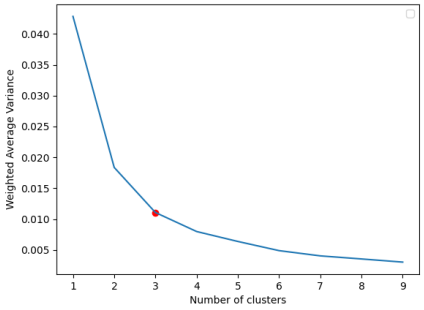


Figure 12. Elbow method for determining the optimal number of clusters based on the weighted average variance. The red point indicates the selected value (k=3), where adding more clusters yields marginal improvement in variance reduction.

The final cluster configuration resulted in Cluster 1 containing five supply points (Chiostergi, Fortunato, Gramsci, Montevideo, Prealpi), Cluster 2 with three supply points (Bande Nere, Berna, Tabacchi), and Cluster 3 consisting solely of Tognazzi. Due to insufficient data availability, soft-sensor development for Cluster 3 was not feasible. Figure 13 illustrates the spatial distribution of these clusters across the Milan distribution network.

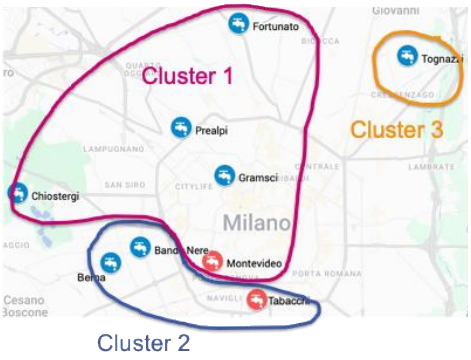


Figure 13. Map of the supply points clustered by COP-kmeans.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Table 11 reveals the water quality patterns that drove the clustering solution. Cluster 3 is distinguished primarily by significantly higher conductivity values (mean 690.83 µS/cm) compared to Clusters 2 and 1 (approximately 510 µS/cm), suggesting different source water characteristics or treatment processes. Nitrate concentrations also show cluster-specific patterns, with Cluster 2 exhibiting lower values (with a mean of 26.185 mg/L).

Importantly, TTHMs levels vary substantially between clusters, with Cluster 1 showing the highest concentrations (mean 6.823 µg/L) and greatest variability, while Cluster 2 displays lower and more consistent levels (mean 1.157 µg/L). This pattern suggests that different network zones have distinct DBP formation potentials, likely related to water age, residual chlorine levels, and organic matter content.

Table 11. Summary of the supply points' clusters.

	Cluster 1				Cluster 2				Cluster 3			
	Mean	Std	5%	95%	Mean	Std	5%	95%	Mean	Std	5%	95%
Color (CU)	0.977	0.82	0.25	2.7	1.188	0.785	0.25	2.6	0.908	0.76	0.25	1.85
Temperature (°C)	16.521	3.681	11.05	23	15.723	4.272	8.6	20.8	16.05	3.11	12.525	20
pH	7.585	0.28	7.2	8.045	7.588	0.268	7.1	7.9	7.567	0.103	7.425	7.675
Conductivity (uS/cm)	507.038	51.541	429.75	610.35	512.78	32.492	457	549	690.833	4.021	685	693.75
Nitrate (mg/L)	29.1	3.184	23.42	32.56	26.185	4.486	19.7	31.8	29.75	0.777	29.025	30.75
Free Chlorine (mg/L)	0.032	0.03	0.01	0.079	0.029	0.026	0.01	0.07	0.062	0.031	0.018	0.088
TOC (mg/L)	0.308	0.264	0.131	0.749	0.366	0.287	0.18	1.06	3.313	1.817	0.785	5.092
TTHMs (µg/L)	6.823	2.393	3.355	10.7	3.329	1.157	1.7	5.5	2.617	0.471	2.125	3.225

Model selection and performance

The ML model comparison revealed cluster-specific optimal approaches. As shown in Table 12, for Cluster 1, the QRNN achieved the best cross-validation performance with an RMSE of 1.566 µg/L. This result suggests that the greater diversity within Cluster 1 benefits from the uncertainty quantification capabilities of QRNN, which can better handle the variability in DBP formation patterns.

For Cluster 2, XGBoost performed optimally with an RMSE of 0.847 µg/L, indicating that the ensemble approach effectively captures the more consistent water quality relationships in this cluster. The superior performance for Cluster 2 reflects its more homogeneous water quality characteristics and lower inherent variability.

Table 12. RMSE (µg/L) results of the LOO cross-validation. The lower the error, the more accurate the model. For each cluster, the lower score is highlighted.

Models	Cluster 1	Cluster 2
PLS	1.782	0.870
SVR	1.737	0.936
QRNN	1.566	0.884
XGBoost	1.782	0.847

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Validation results on test data

The hold-out test evaluation demonstrated satisfactory performance for both clusters (Figure 14). Overall, the model's performance on Cluster 2 shows less uncertainty, as the confidence intervals are narrower than those on Cluster 1. This is acceptable since Cluster 2 contains three supply points while Cluster 1 has five, leading to a potentially higher diversity. The mean prediction of the models is coherent with the actual measurements, showing for Cluster 1 a RMSE of 1.659 (µg/L) and a Mean Absolute Percentage Error (MAPE) of 24%, while for Cluster 2 the model performs a RMSE of 0.811, with a MAPE of 14%. Even if the uncertainty intervals of Cluster 1 are quite large, they do not cross 30 µg/L, which is the D.Lgs. 31/2001, meaning they could be used to assess if a parameter is above or below that threshold with confidence.

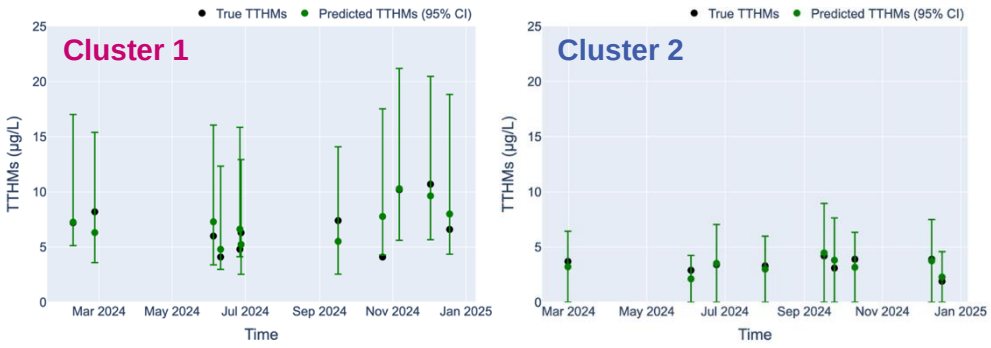


Figure 14. Evaluation of the best models on a test set for Cluster 1(left) and Cluster 2 (right). Black dots correspond to the real grab samples measurements, while green ones represent the estimation of the model, together with a 95% confidence interval.

The continuous deployment of soft-sensors using 15-minute sensor data provided mixed results that offer important insights for practical implementation. For Cluster 1 supply points (Figure 15), the soft-sensor predictions showed good agreement with grab sample measurements at three locations (Prealpi, Chiostergi, and Fortunato), demonstrating successful translation from laboratory reference data to continuous monitoring.However, systematic underestimation was observed at Montevideo and Gramsci, suggesting that local conditions at these sites may not be fully captured by the cluster-averaged model. This performance variation within clusters highlights a fundamental limitation: clustering reduces individual site variability but cannot eliminate all location-specific effects.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

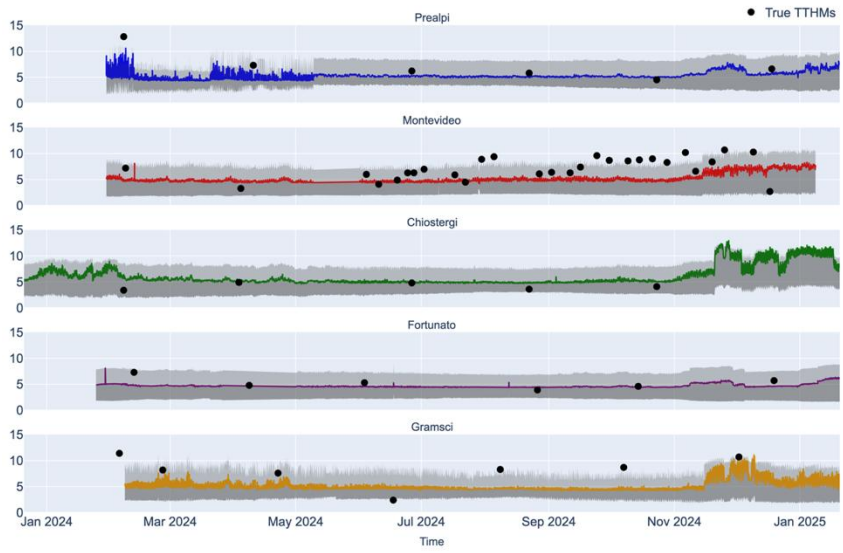


Figure 15. Soft-sensor predictions of TTHMs for supply points belonging to Cluster 1.

For Cluster 2 (Figure 16), an interesting temporal pattern emerged. Initial predictions showed consistent overestimation across all three supply points, but accuracy improved substantially toward the end of the monitoring period. This improvement suggests either seasonal effects not fully captured in the training data or gradual model adaptation as operational conditions stabilized.

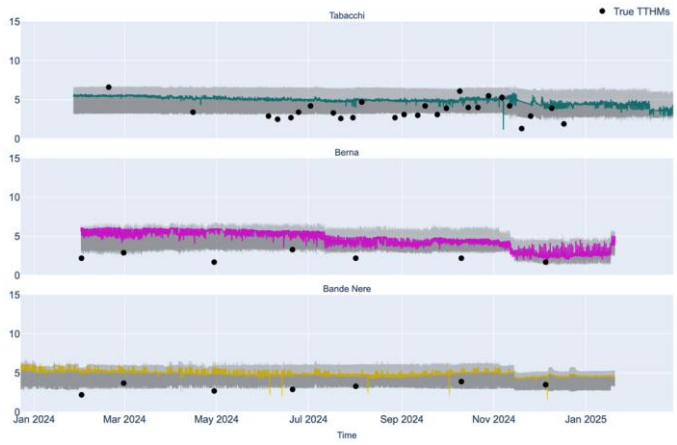


Figure 16. Soft-sensor predictions of TTHMs for supply points belonging to Cluster 2.

Insights into feature importance for each cluster are presented in Figures 17 and 18 for Cluster 1 and Cluster 2, respectively.

For Cluster 1, conductivity emerges as the most influential predictor by a large margin, indicating that variations in conductivity are strongly associated with changes in the predicted TTHMs concentrations. Other parameters, such as TOC, temperature, and color, have a more moderate but still relevant contribution, while nitrate, free chlorine, and pH show relatively minor influence on the model output. This result suggests that the water quality dynamics in the supply points grouped under Cluster 1 are largely captured by conductivity trends, with other parameters adding secondary refinements.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

In contrast, Cluster 2 exhibits a more balanced contribution across multiple parameters. Here, nitrate and pH stand out as the most important predictors, followed by free chlorine and TOC, indicating that DBP formation dynamics in this cluster are driven by a more complex interplay of chemical and physical parameters rather than a single dominant factor. The more homogeneous feature importance profile may also explain the better generalization ability observed in this cluster, as the model relies on multiple complementary signals instead of a single driving variable.

These findings provide valuable insights into the site-specific drivers of DBP formation, supporting more targeted monitoring strategies: for Cluster 1, conductivity should be prioritized as a key indicator, whereas in Cluster 2, attention should be distributed across several water quality parameters.

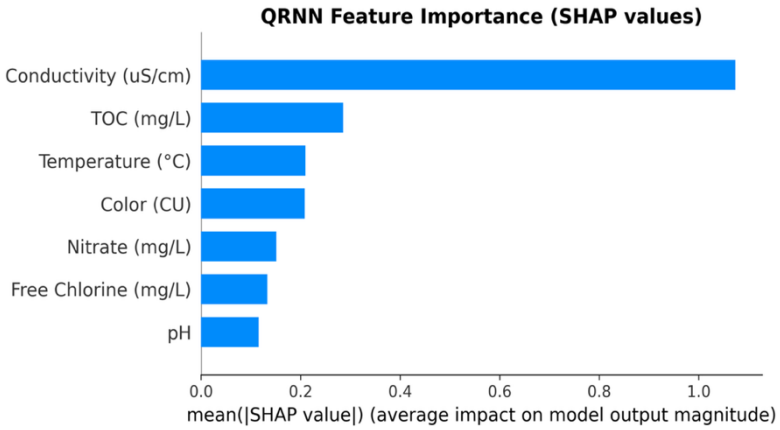


Figure 17. Mean SHAP values for Cluster 1.

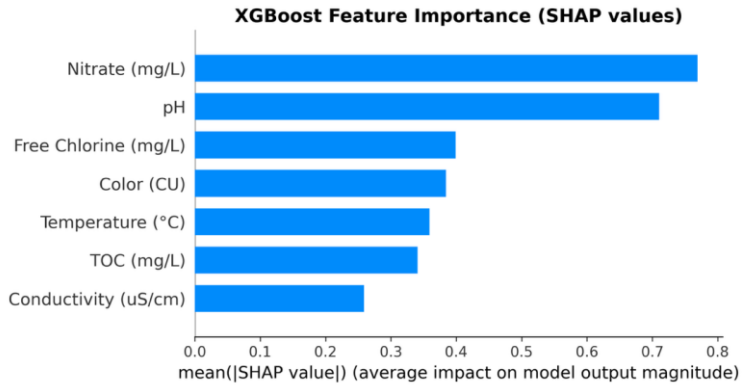


Figure 18. Mean SHAP values for Cluster 2.

These results demonstrate that soft-sensors can provide valuable continuous monitoring capabilities for DBP levels in distribution networks, but with important caveats. The cluster-based approach successfully balances data availability with local specificity, providing reasonable prediction accuracy for most locations. However, the performance variation within clusters suggests that periodic validation with grab samples remains necessary to verify the accuracy of soft-sensors and detect model drift.

The ability to predict TTHMs levels every 15 minutes represents a significant improvement over monthly grab sample monitoring, enabling early detection of elevated DBP formation and supporting proactive management responses. Even with the observed uncertainties, the soft-sensors provide

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

sufficient accuracy to support compliance monitoring and identify trends that may indicate emerging water quality issues.

4.1.2. Feltre DWTP

The model comparison for the Feltre DWTP case study utilized 5-fold cross-validation with the TimeSeriesSplit method from *scikit-learn* (Pedregosa FABIANPEDREGOSA *et al.*, 2011) to preserve temporal data integrity. The LGBM network demonstrated superior performance compared to other approaches, achieving RMSE values of 0.65 (1/mL) for ICC predictions (Table 13), suggesting that the relationships between conventional water quality parameters and microbial indicators are well-captured by piece-wise linear models rather than requiring complex non-linear transformations.

Table 13. RMSE (1/mL) results of the k-fold cross-validation. The lower the error, the more accurate the model. The lower score is highlighted.

Models	ICC (1/mL)
XGBoost	0.79
LGBM	0.65
QRNN	0.7
LSTM	0.66

The test dataset evaluation revealed important insights about soft-sensor capabilities for treatment process monitoring. Figure 19 shows that the LGBM model successfully captures baseline variations and normal operational fluctuations in microbial parameters, demonstrating effective learning of typical treatment process dynamics. However, a critical limitation emerged in the model's ability to predict sudden spikes in microbial counts. These events, which could indicate treatment failures or contamination incidents, represent the most important conditions for an early warning system to detect. The failure to predict these spikes suggests that the conventional water quality parameters monitored may not contain sufficient information to explain extreme microbial events.

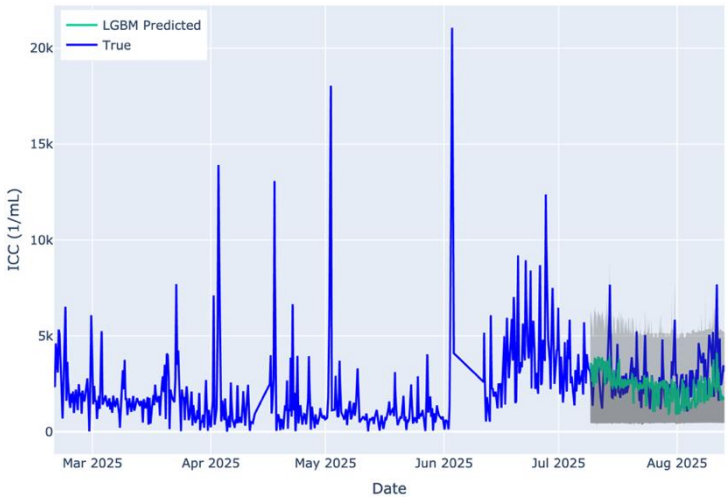


Figure 19. Soft-sensor prediction vs real measurements of ICC (1/mL). Blue line represents the real measurements of the BactoSense, while the red one the soft-sensor predictions, with the light-red area indicating the 95% confidence interval.

Insights into feature importance for the Feltre DWTP are presented in Figure 20, which displays the normalized importance of the candidate features with the first 20 largest importance. The results indicate that the prediction of ICC is largely driven by a limited subset of highly informative parameters.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Notably, the use of a 6-hour rolling mean transformation appears to have been effective in generating features that carry strong predictive power, as these transformed variables dominate the ranking. Furthermore, lagged features with longer lags (4–6 hours) exhibit a higher impact compared to shorter lags, suggesting that the microbial dynamics captured by ICC are more strongly influenced by slower, cumulative changes in water quality. Among the base parameters, conductivity, free chlorine, nitrate, and pH emerge as the most relevant predictors, indicating that these variables play a key role in shaping the temporal evolution of ICC.

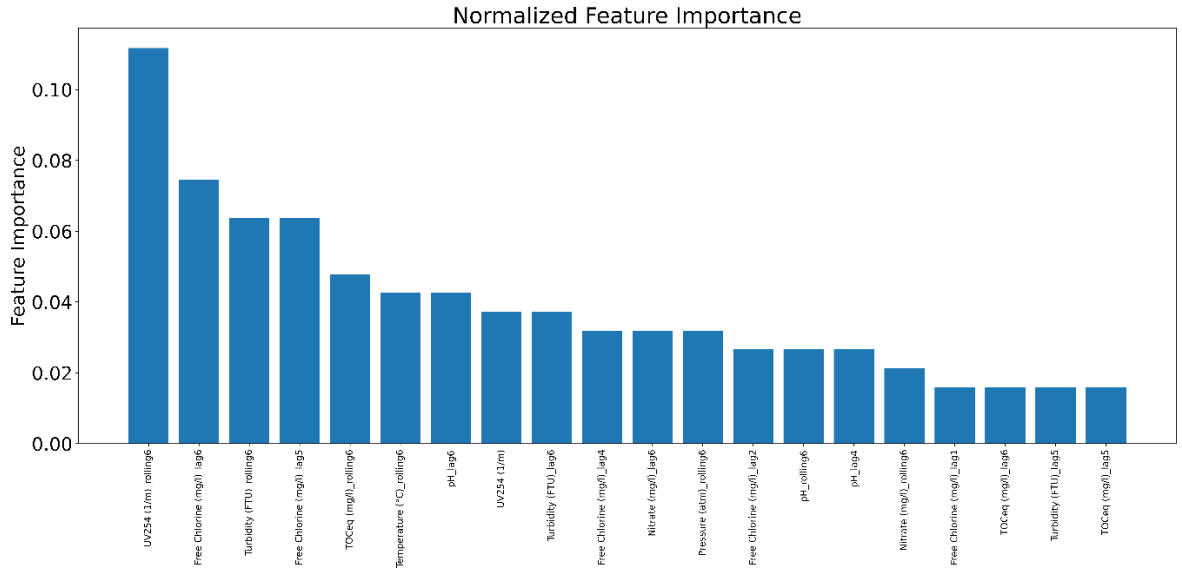


Figure 20. LGBM normalized feature importance.

4.2. CS#3 Tarragona, Spain

The Tarragona case study aimed to develop and validate a data-driven soft-sensor capable of providing daily predictions of THM-risk levels (Low, Medium, High). The modelling strategy and pipeline described in Section 3.2 were applied to the historical and online data collected at the water treatment plant, with the objective of evaluating the system's capacity to anticipate DBP formation trends and support operational decision-making.

4.2.1 Exploratory analysis and feature relevance

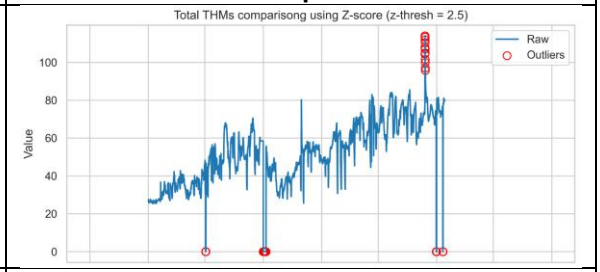
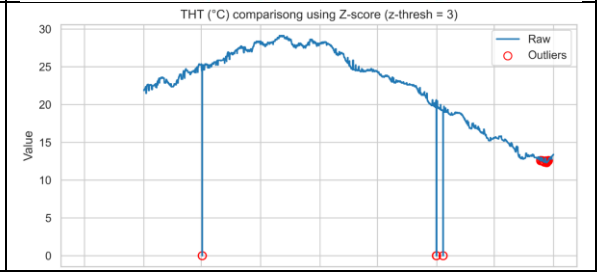
The initial exploratory assessment of correlations between conventional water-quality indicators and measured THM concentrations provided a first insight into the relationships within the dataset. Table 10 in Section 3.2.2 reports the Pearson correlation coefficients between each input parameter and TTHMs, while Figure 9 in Section 3.2.2 shows the complete correlation heatmap. Parameters associated with organic content (TOC, SFL) and chlorine-related variables (CLR, CLR_2, CLR_2.1) exhibited the strongest correlations with THM formation. Other parameters, such as temperature (THT) and color indices (CTR, CTR_2), displayed weaker but consistent relationships. This analysis confirmed that a multivariate approach, integrating multiple predictors, was required, as no single variable provided sufficient explanatory power.

4.2.2 Data pre-processing and feature construction

The outlier detection strategy effectively removed anomalous peaks caused by sensor errors or maintenance operations. Examples of this step are summarized in Table 14.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Table 14. Outlier detection using z-score.

Variable	Outlier detection example
Total THMs	
THT	

After filtering, the Hampel smoothing procedure yielded continuous and stable time series suitable for feature generation. Feature engineering expanded the original variable space by introducing lag, rolling, and interaction terms, resulting in a total of more than 90 features (Table 15).

Table 15. Complete list of engineered features.

Feature name	Feature description	Feature name	Feature description
ABS (Au)	Raw treated feature	CLR_2_lag_1	Lag-derived feature
CLR_2 (mg/L)		CLR_lag_1	
CLR_2.1 (mg/L)		CLR_2_lag_3	
CLR (mg/L)		CLR_2.1_lag_3	
CRE (%)		CLR_lag_3	
CTR_2 (HU)		CLR_2_lag_6	
CTR (HU)		CLR_2.1_lag_6	
THT (°C)		CLR_lag_6	
TTR (HU)		TTR_lag_6	
TOC (mg/L)		CLR_2_lag_12	
TSS (mg/L)		CLR_2.1_lag_12	
SFL (mL/g)		CLR_lag_12	
hour_sin	Time-derived feature	CTR_2_lag_12	
hour_cos		TTR_lag_12	
day_sin		CLR_2_lag_24	
day_cos		CLR_2.1_lag_24	
month_sin		CLR_lag_24	
month_cos		CTR_2_lag_24	
day_of_week_sin		TTR_lag_24	
day_of_week_cos		ABS_CLR_2_interaction	Feature interaction
is_weekend		ABS_CLR_2.1_interaction	Feature interaction
		CTR_hs_diff	Feature difference
CLR_2_rolling_std_3	Statistical rolling std/mean	ABS_rolling_std_12	Statistical rolling std/mean
CLR_2.1_rolling_std_3		CLR_2_rolling_std_12	
CLR_rolling_std_3		CLR_2.1_rolling_std_12	
CTR_2_rolling_std_3		CLR_rolling_std_12	
CTR_rolling_std_3		CTR_2_rolling_std_12	
THT_rolling_std_3		CTR_rolling_std_12	
TTR_rolling_std_3		THT_rolling_std_12	
TSS_rolling_std_3		TTR_rolling_std_12	

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

SLF_rolling_std_3		TSS_rolling_std_12	
ABS_rolling_std_6		SLF_rolling_std_12	
CLR_2_rolling_std_6		CLR_2_rolling_mean_24	
CLR_2.1_rolling_std_6		CLR_rolling_mean_24	
CLR_rolling_std_6		ABS_rolling_std_24	
CTR_2_rolling_std_6		CLR_2_rolling_std_24	
CTR_rolling_std_6		CLR_2.1_rolling_std_24	
THT_rolling_std_6		CLR_rolling_std_24	
TTR_rolling_std_6		CRE_rolling_std_24	
TSS_rolling_std_6		CTR_2_rolling_std_24	
SLF_rolling_std_6		CTR_rolling_std_24	
CLR_2_rolling_mean_12		THT_rolling_std_24	
CLR_rolling_mean_12		TSS_rolling_std_24	
ABS_rolling_std_12		SLF_rolling_std_24	
ABS_diff	Feature difference	CLR2_5h_decrease	CLR decrease (5h)
CLR_diff		CLR1_6h_decrease	CLR decrease (6h)
CRE_diff		CLR_decay_rate	CLR decay rate
TTR_diff		THM (µg/L)	Raw treated data
SLF_diff		THM_value	Statistical category

Recursive Feature Elimination (RFE) reduced this set to the fifty most informative variables (Table 16), mainly related to chlorine dynamics, absorbance, and short-term variability descriptors. Those variables were selected to train the final model with only the most informative features.

Table 16. Selected most informative features to train the definitive classification model.

Feature name			
month_cos	ABS (Au)	CTR_2 (HU)	CTR_2_rolling_std_6
TOC (mg/L)	THT_rolling_std_24	TTR (HU)	CTR_2_rolling_std_24
THT (°C)	CTR_2_lag_24	CLR_2_rolling_mean_12	day_sin
CTR (HU)	CLR_2_lag_3	CTR_hs_diff	ABS_rolling_std_6
CLR_2_lag_24	THT_rolling_std_6	THT_rolling_std_3	TTR_lag_6
CLR_2_rolling_mean_24	CLR_2.1_rolling_std_3	CLR_2_lag_6	CLR_2.1_lag_24
CLR_2 (mg/L)	CTR_2_rolling_std_3	CLR_2_lag_12	CLR_lag_24
CLR_cota_lag_1	TTR_rolling_std_12	CRE_diff	CTR_rolling_std_12
ABS_CLR_2.1_interaction	CLR_diff	CLR_decay_rate	SLF_rolling_std_6
TTR_lag_24	CLR2_5h_decrease	TTR_lag_12	month_sin
CTR_2_lag_12	CLR_2_rolling_std_24	CTR_rolling_std_6	CRE_rolling_std_24
ABS_CLR_cota_interaction	CTR_rolling_std_24	CRE (%)	ABS_rolling_std_12
ABS_diff	CLR_2.1_rolling_std_12		

4.2.3 Model comparison and selection

Several classification algorithms were tested to identify the most suitable model for predicting daily THM risk. The comparative results are summarized in Table 17.

Table 17. Classification training candidates.

Model	**Precision**	**Recall**	**F1-score**	**Accuracy**
SVM	0.69	0.82	0.67	0.78
Logistic Regression	0.70	0.76	0.61	0.75
KNN	0.80	0.71	0.75	0.75
Gradient Boosting	**0.94**	**0.94**	**0.91**	**0.92**
LightGBM classifier	0.89	0.88	0.82	0.88

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Among the tested models — Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbours (KNN), LightGBM, and Gradient Boosting — the Gradient Boosting Classifier achieved the highest overall performance, with an accuracy above 90% and balanced precision–recall across classes. The LightGBM model also showed good predictive ability but slightly higher variability across folds, while simpler models such as Logistic Regression and SVM underperformed on the minority (High-risk) class. The superiority of Gradient Boosting was attributed to its capacity to capture nonlinear relationships and its resilience to noisy environmental data.

4.2.4 Classification performance and diagnostic analysis

The detailed classification outcomes are illustrated through the confusion matrix (Figure 21) and the ROC curves (Figure 22).

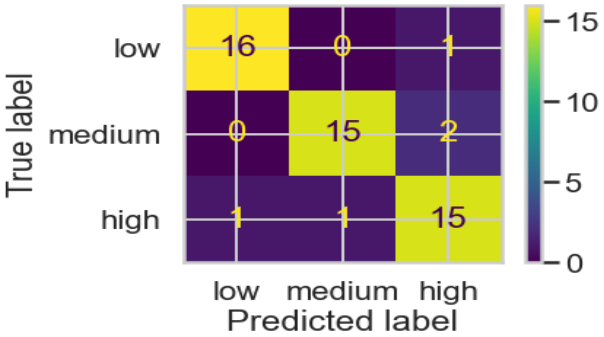


Figure 21. Confusion matrix.

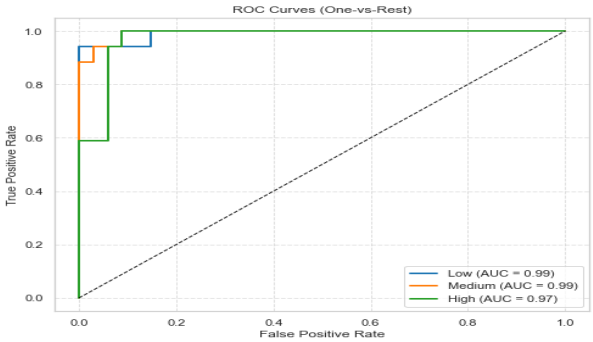


Figure 22. ROC curve.

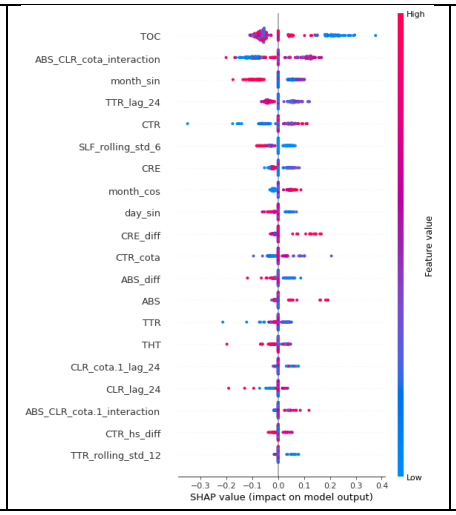
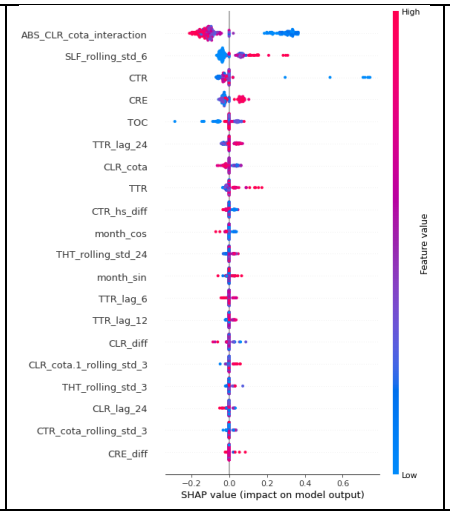
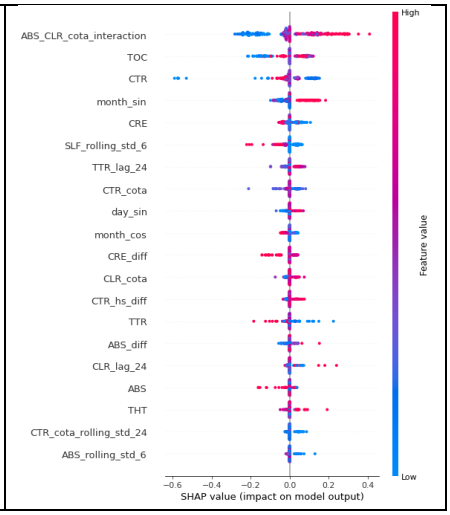
The confusion matrix revealed strong agreement between predicted and observed classes, with the largest accuracy achieved for the Low and High-risk levels. The High THM-risk category, which is operationally critical for early warning, was correctly identified in nearly all validation cases, confirming the model's sensitivity to adverse conditions. Occasional misclassifications mainly occurred between adjacent categories (Medium and High), which is typical in threshold-based classification tasks. The ROC curve analysis further demonstrated the model's discriminative ability, with all classes achieving AUC values well above 0.9. The steep shape of the curves confirmed that the classifier maintained high sensitivity while effectively controlling false alarms.

4.2.5 Feature importance and explainability

The feature-importance analysis highlighted chlorine-related variables, particularly those describing chlorine decay and residual concentration, as the strongest predictors of THM risk. Absorbance (ABS) and temperature (THT) also contributed significantly, confirming their influence on the formation of disinfection by-products. The detailed ranking of the most relevant predictors for each class is presented in Table 18, which illustrates the distinct influence patterns across the *Low*, *Medium*, and *High* THM-risk categories. The SHAP analysis further confirmed that combinations of high absorbance with low chlorine residuals increased the probability of *High-risk* events, in line with expected process behavior and operational experience.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Table 18. Feature importance summary plots.

low		medium		high	

4.2.6 Summary of the results

The developed classification model achieved a precision of 92%, confirming its ability to provide dependable THM risk predictions daily. The F1-score was also high, reflecting a balanced performance across precision and recall. The confusion matrix indicated that the model was generally successful in correctly classifying all three THM risk levels (Low, Medium, High) most of the time.

Analysis of feature importance revealed consistent patterns across classes, with differences in ranking. For the High-risk label, the most influential features were the absorbance × chlorine interaction, TOC, and indicators such as CTR, CRE, and time-related variables. For the Medium class, important drivers included CLR, TTR, the same absorbance-chlorine interaction, and THT. In the Low class, relevant predictors were TOC, the interaction term, SLF rolling, time features, and THT (without lags). Overall, the results highlighted that TOC, THT, and the interaction between absorbance and chlorine consistently played a central role in TTHMs risk classification.

The ROC curves demonstrated strong discriminative performance, with AUC values significantly above the random baseline. Comparison with alternative algorithms showed that other models achieved similar levels of performance, indicating that the results were robust and not tied to a single classifier. Feedback from CAT partners was positive, as they recognized the potential of this model to provide a 24-hour early warning system that could support the identification of potential THM exceedances.

A real-world application of the current framework would provide:

- an early-warning soft-sensors based on routine water quality measurements that provide 24-hour prior alerts of THM risks, supporting proactive management in DWSS and DWDN;
- decision-support value, by highlighting key drivers such as TOC, THT, and chlorine–absorbance interactions, the model offers interpretable insights that could guide operators in focusing monitoring and control efforts where they matter most.

While the results are promising, the model is currently a research prototype; further validation, transferability testing, and integration into operational workflows are necessary before it can be deployed in real-world settings. Indeed, among the known limitations, we mention the site-specificity

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

and the data dependency. Specifically, the model is not directly transferable to other DWSS or DWDN contexts, and it must be retrained with local data to account for different source water characteristics and operational regimes. Moreover, the model accuracy is constrained by the amount and frequency of available data; resampling to daily intervals reduces temporal detail, and additional long-term datasets are needed to further improve robustness.

4.3. CS#4 Western Ukraine, Ukraine

The transfer learning approach for CS#4 provided important insights into the limitations and potential applications of cross-system soft-sensor deployment. The similarity analysis using Mahalanobis distance revealed that nearly all Western Ukraine measurements (38 of 40) showed greater similarity to Milan's Cluster 2 than to Cluster 1, suggesting some consistency in water quality characteristics between the systems.

Cluster assignment results

The similarity analysis revealed that nearly all measurements from the three Western Ukraine sites were more closely aligned with Milan's Cluster 2 characteristics rather than Cluster 1, with only two out of 40 measurements from Site 3 showing closer proximity to Cluster 1. For consistency, all measurements were assigned to Cluster 2. The PCA visualization (Figure 23) confirmed this assignment, showing Sites 1 and 2 clustered near Milan's Cluster 2 data, while Site 3 measurements were more distant but still closer to Cluster 2 than Cluster 1.

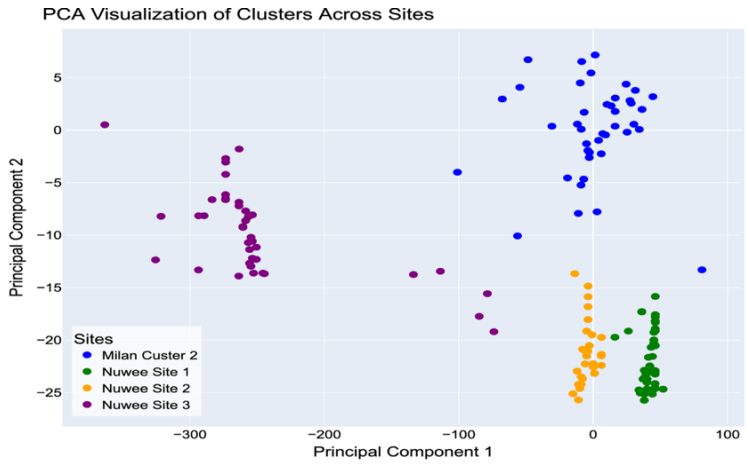


Figure 23. Scatter plot of the first two components of PCA trained on Milan Cluster 2 and applied to the three Western Ukraine sites.

Transfer learning performance

The direct application of the Milan Cluster 2 soft-sensor to Western Ukraine data yielded poor prediction performance across all sites (Figure 24), as it tends to predict an almost constant value across all three sites. This behavior can be explained by the strong assumptions made when transferring models between CS#2 and CS#4, particularly the assumption of comparable water quality dynamics, which does not fully hold in practice.

In terms of performance, the soft-sensor achieved a MAPE of 99.13% with an RMSE of 2.72 µg/L at Site 1, a MAPE of 104.75% with an RMSE of 5.53 µg/L at Site 2, and a MAPE of 294% with an RMSE of 7.83 µg/L at Site 3. These metrics confirm the trends shown in the plots: there is a substantial difference in performance between the Western Ukraine sites and Cluster 2, as well as between the sites themselves. Site 3 shows the poorest performance, likely due to larger discrepancies in water quality parameters compared to Milan's Cluster 2, whereas Sites 1 and 2 are somewhat closer in behavior.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

These results highlight the limitations of directly transferring soft-sensors across case studies without site-specific calibration or additional local data.

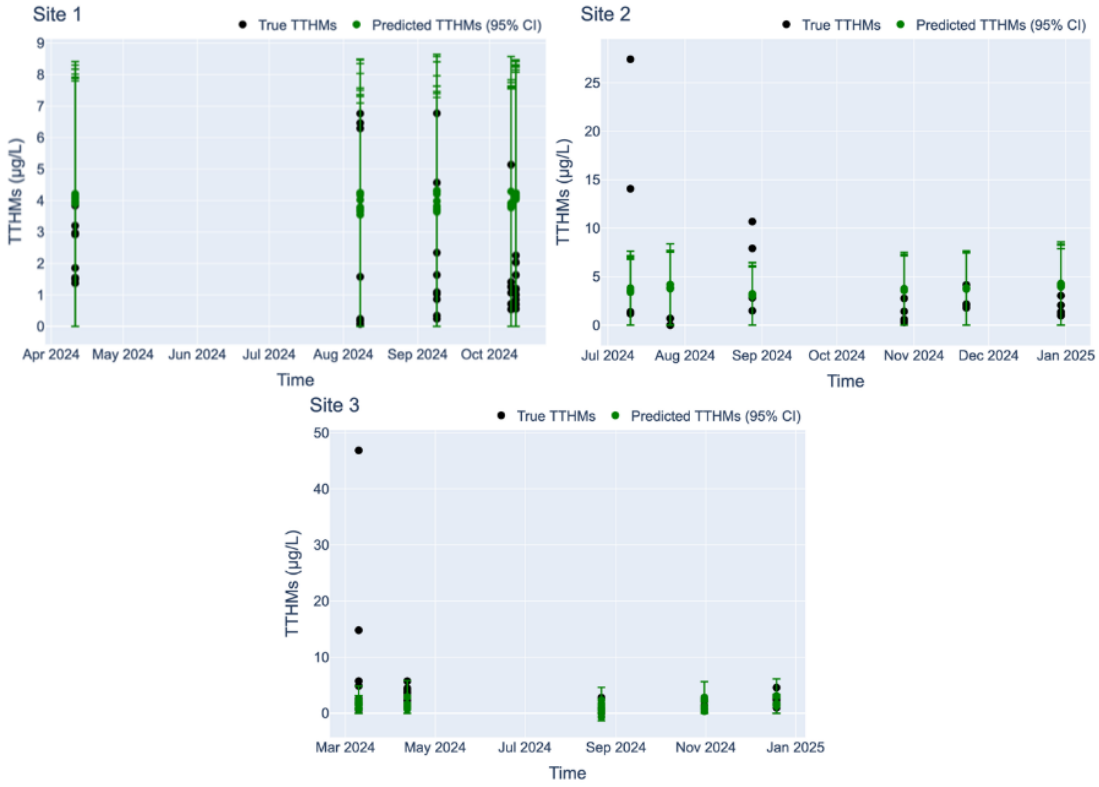


Figure 24. Prediction of CS#2 Cluster 2 soft-sensor on Western Ukraine sites. The black dots are the real measurements, while the green ones are the predictions, together with the corresponding 95% confidence interval.

5. Conclusions

The results from the three case studies—Milan, Tarragona, and Western Ukraine—demonstrate that data-driven soft-sensors can effectively estimate key water-quality indicators and capture complex relationships between routine parameters and target contaminants, such as disinfection by-products and microbial indicators. Across all sites, the joint use of low- and high-informative data sources significantly enhanced predictive accuracy, enabling near-real-time assessment of water quality under different climatic and operational contexts.

In Milan, soft-sensors successfully reproduced the behavior of DBPs in both a treatment plant and the distribution system, highlighting their potential for continuous online estimation. In Tarragona, robust and explainable ML models proved capable of classifying daily THM risk levels and supporting operational decisions. In Western Ukraine, the application of the framework demonstrated the promising transferability of models developed elsewhere, while highlighting the need for local calibration to account for regional variability.

Methodologically, the workflow developed under Task 4.2 ensures transparent data management, reproducible feature engineering, and consistent model validation using open-source Python tools. The integration of explainable AI techniques, particularly SHAP analysis, enhances the understanding of variable influence and fosters operator confidence in model outputs. The three case studies collectively demonstrate that a single, harmonized framework can guide utilities in planning

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

monitoring campaigns and developing data-driven soft-sensors using both existing and new sensor infrastructures. Despite differences in data availability and operational settings—ranging from the disinfected and highly instrumented systems of Milan and Tarragona to the non-disinfected and data-limited systems of Western Ukraine—the same stepwise approach proved applicable and effective. The approaches and procedures we described can be used as guidelines for the development of new soft-sensors in future deployments of such algorithms. The CS#4 highlighted that the approach is not universal, but that some preliminary feasibility studies should be performed before extending the provided approach to a new setting, particularly where available data are scarce.

Soft-sensors emerge as digital surrogates for complex laboratory measurements, providing early warning signals of deviations in treatment performance and network stability. When embedded in supervisory control systems, they can help utilities detect anomalies, optimize disinfectant dosing, and reduce the frequency and cost of laboratory testing.

This document has been prepared by data technicians with expertise in data processing, modelling, and the interpretation of sensor outputs. While it provides technical guidelines and methodological insights from a data-driven perspective, the true applicability and integration of these sensors within DWDN and DWSS require the knowledge and experience of water sector specialists. The contribution presented here is intended to serve as a foundation for further discussion and decision-making, while operational feasibility, regulatory compliance, and domain-specific considerations must ultimately be validated and refined by professionals in the water field.

Future work will focus on extending validated models to additional pilot sites and testing long-term robustness under diverse hydraulic and climatic conditions. Soft-sensors represent proactive tools for DWSS and DWDN, combining sensor data with advanced ML techniques to support informed operational decisions, reduce the risk of regulatory breaches, and enhance consumer protection. These developments will contribute to SafeCREW's overarching goal of ensuring safe, adaptive, and resilient drinking-water management across European and non-disinfected supply contexts.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081980.

Bibliography

Bergman, L.E. *et al.* (2016) “Application of Classification Trees for Predicting Disinfection By-Product Formation Targets from Source Water Characteristics,” *https://home.liebertpub.com/ees*, 33(7), pp. 455–470. Available at: https://doi.org/10.1089/EES.2016.0044.

Besmer, M.D. *et al.* (2016) “Online flow cytometry reveals microbial dynamics influenced by concurrent natural and operational events in groundwater used for drinking water treatment,” *Scientific Reports 2016 6:1*, 6(1), pp. 1–10. Available at: https://doi.org/10.1038/srep38462.

Bradley, P.S., Bennett, K.P. and Demiriz, A. (2000) “Constrained K-Means Clustering.”

Brazdzionyte, J. and Macas, A. (2007) “Bland-Altman analysis as an alternative approach for statistical evaluation of agreement between two methods for measuring hemodynamics during acute myocardial infarction.,” *Medicina (Kaunas, Lithuania)*, 43(3), pp. 208–214. Available at: https://doi.org/10.3390/medicina43030025.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011) “mice: Multivariate Imputation by Chained Equations in R,” *Journal of Statistical Software*, 45(3), pp. 1–67. Available at: https://doi.org/10.18637/JSS.V045.I03.

Chen, T. and Guestrin, C. (2016) “XGBoost: A Scalable Tree Boosting System,” *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016, pp. 785–794. Available at: https://doi.org/10.1145/2939672.2939785.

Chen, W. *et al.* (2024) “Machine Learning-Based Water Quality Classification Assessment,” *Water 2024, Vol. 16, Page 2951*, 16(20), p. 2951. Available at: https://doi.org/10.3390/W16202951.

Fawcett, T. (2006) “An introduction to ROC analysis,” *Pattern Recognition Letters*, 27(8), pp. 861–874. Available at: https://doi.org/10.1016/J.PATREC.2005.10.010.

Ferrer, L., Scharenborg, O. and Bäckström, T. (2024) “Good practices for evaluation of machine learning systems.” Available at: https://arxiv.org/abs/2412.03700v1 (Accessed: October 18, 2025).

Friedman, J.H. (2001) “Greedy function approximation: A gradient boosting machine.,” *https://doi.org/10.1214/aos/1013203451*, 29(5), pp. 1189–1232. Available at: https://doi.org/10.1214/AOS/1013203451.

Imran *et al.* (2022) “A Survey of Datasets, Preprocessing, Modeling Mechanisms, and Simulation Tools Based on AI for Material Analysis and Discovery,” *Materials (Basel, Switzerland)*, 15(4). Available at: https://doi.org/10.3390/MA15041428.

Jantre, S.R., Bhattacharya, S. and Maiti, T. (2020) “Quantile Regression Neural Networks: A Bayesian Approach,” *Journal of Statistical Theory and Practice*, 15(3). Available at: https://doi.org/10.1007/s42519-021-00189-w.

Kaneko, H. (2023) “Interpretation of Machine Learning Models for Data Sets with Many Features Using Feature Importance,” *ACS Omega*, 8(25), pp. 23218–23225. Available at: https://doi.org/10.1021/ACSOMEGA.3C03722/ASSET/IMAGES/LARGE/AO3C03722_0007.JPEG.

Ke, G. *et al.* (2017) “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” *Advances in Neural Information Processing Systems*, 30. Available at: https://github.com/Microsoft/LightGBM. (Accessed: September 8, 2025).

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101081980.

Liang, Y. *et al.* (2025) “Machine learning-guided prediction of chlorinated/chloraminated disinfection by-product formation in drinking water treatment,” *Water Research*, 283, p. 123849. Available at: https://doi.org/10.1016/J.WATRES.2025.123849.

Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S. (2020) “Explainable AI: A Review of Machine Learning Interpretability Methods,” *Entropy (Basel, Switzerland)*, 23(1), pp. 1–45. Available at: https://doi.org/10.3390/E23010018.

Lundberg, S.M. and Lee, S.I. (2017) “A Unified Approach to Interpreting Model Predictions,” *Advances in Neural Information Processing Systems*, 2017-December, pp. 4766–4775. Available at: https://arxiv.org/abs/1705.07874v2 (Accessed: September 22, 2025).

De Maesschalck, R., Jouan-Rimbaud, D. and Massart, D.L. (2000) “The Mahalanobis distance,” *Chemometrics and Intelligent Laboratory Systems*, 50(1), pp. 1–18. Available at: https://doi.org/10.1016/S0169-7439(99)00047-7.

Mckinney, W. (2010) *Data Structures for Statistical Computing in Python*.

Ncr, (and Clinton, J. (1999) “CRISP-DM 1.0 Step-by-step data mining guide.”

Pedregosa FABIANPEDREGOSA, F. *et al.* (2011) “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, 12(85), pp. 2825–2830. Available at: http://jmlr.org/papers/v12/pedregosa11a.html (Accessed: September 8, 2025).

Peng, F. *et al.* (2023) “Predicting the formation of disinfection by-products using multiple linear and machine learning regression,” *Journal of Environmental Chemical Engineering*, 11(5), p. 110612. Available at: https://doi.org/10.1016/J.JECE.2023.110612.

Pirouz, D.M. (2012) “An Overview of Partial Least Squares,” *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/SSRN.1631359.

Prest, E.I. *et al.* (2016) “Long-Term Bacterial Dynamics in a Full-Scale Drinking Water Distribution System,” *PLOS ONE*, 11(10), p. e0164445. Available at: https://doi.org/10.1371/JOURNAL.PONE.0164445.

Rodriguez, M.J., Sérodes, J.B. and Levallois, P. (2004) “Behavior of trihalomethanes and haloacetic acids in a drinking water distribution system,” *Water Research*, 38(20), pp. 4367–4382. Available at: https://doi.org/10.1016/J.WATRES.2004.08.018.

Shojaie, A. and Fox, E.B. (2021) “Granger Causality: A Review and Recent Advances,” *Annual Review of Statistics and Its Application*, 9, pp. 289–319. Available at: https://doi.org/10.1146/annurev-statistics-040120-010930.

Smola, A.J., Sch¨olkopf, B. and Sch¨olkopf, S. (2004) “A tutorial on support vector regression *,” *Statistics and Computing*, 14, pp. 199–222.

Staudemeyer, R.C. and Morris, E.R. (2019) “Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks.” Available at: https://arxiv.org/abs/1909.09586v1 (Accessed: September 8, 2025).

Uyak, V., Toroz, I. and Meriç, S. (2005) “Monitoring and modeling of trihalomethanes (THMs) for a water treatment plant in Istanbul,” *Desalination*, 176(1–3), pp. 91–101. Available at: https://doi.org/10.1016/J.DESAL.2004.10.023.

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101081980.