Early warning tool for control and risk management strategies in WDN

Deliverable 3.4

WP 3 DISTRIBUTION – Advanced Monitoring and Prediction



Author: UCY – KIOS

Date:30/09/2026

Funded by the European Union

GRANT AGREEMENT NUMBER			
FULL TITLE / ACRONYM	INNOVATIVE TOOLS TO CONTROL ORGANIC MATTER AND DISINFECTION BYPRODUCTS IN DRINKING WATER / intoDBP		
START DATE	01.12.2022	DURATION	48 months
END DATE	30.11.2026		
PROJECT URL	https://intodbp.eu/		
DELIVERABLE	D3.4		
WORK PACKAGE	WP3		
CONTRACTUAL DATE OF DELIVERY	30.09.2026		
ACTUAL DATE OF DELIVERY	05.10.2026		
NATURE	REPORT/OTHER	DISSEMINATION LEVEL	PUBLIC
LEAD BENEFICIARY	UCY		
RESPONSIBLE AUTHOR	UCY- KIOS		
CONTRIBUTIONS FROM	UCY, ICRA, SCAN, WBL, CANAL, UCT		
ABSTRACT	This deliverable presents the implementation of proposed strategies for early detection, optimal control and risk minimization of DBP formation. A holistic framework that uses Digital Twin (DT) technology to monitor DBP formation, quantify associated exposure risks and evaluate mitigation strategies within WDNs is presented. The framework extends the dbpRisk 2.0 platform by introducing these new modules. Specifically, a digital twin was developed using the calibrated EPANET-MSX model connected to real-time sensor measurements. The methodology enables the prediction of THM concentrations under real-time operating conditions. A Driver Contribution Analysis framework identifies the main factors contributing to elevated predicted THM levels. The resulting algorithms were integrated into the web-based user interface to facilitate real-time monitoring and issue warnings for operators. Potential health risks are calculated by estimating exposure metrics. Mitigation is addressed through optimal placement and dosing schedules for booster chlorination stations, formulated as a multi-objective optimization problem balancing water quality performance and operational cost. Furthermore, an Explainable Artificial Intelligence (XAI) framework has		

	been integrated into the web-based user interface. The XAI framework combines the knowledge and methodologies developed throughout the project with a Large Language Model (LLM) to explain model outputs and provide additional insights to operators.

Disclaimer

Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Copyright message

© intoDBP Consortium, 2024

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation, or both. Reproduction is authorised provided the source is acknowledged.

TABLE OF CONTENTS

LIST OF TABLES 5

LIST OF FIGURES 5

LIST OF ABBREVIATIONS 6

EXECUTIVE SUMMARY 8

1.0 INTRODUCTION 9

2.0 FRAMEWORK 10

2.1 Risk assessment tool 10
2.2 Reaction model validation 11

3.0 ANALYSIS OF THM FORMATION PREDICTORS 15

3.1 Data based approach - Sampling Data Analytics 15
3.1.1 Data Preparation and Variable Definition 15
3.1.2 Regression modelling 16
3.1.3 Performance Indicators 16
3.1.4 Predictor Significance Analysis 19
3.2 Model based approach - Driver contribution analysis 21
3.2.1 Offline Baseline Scenario 21
3.2.2 Assumptions and Limitations 23
3.2.3 Online Application 23

4.0 MITIGATION STRATEGIES 26

4.1 Application Theory 27
4.2 Decision Variables 30
4.3 Objective Functions 30
4.4 Constraints 31

5.0 DBP EXPOSURE RISK ASSESSMENT 34

5.1 Population dynamics 34
5.2 Exposure metrics 35

6.0 CASE STUDY – APPLICATION 36

6.1 DBP Formation Estimation 36
6.2 Mitigation Strategies 37
6.3 Exposure Risk Assessment 41

7.0 EARLY-WARNING SYSTEM PLATFORM AND EXPLAINABLE AI 44

7.1 Real-Time THM Early Warning Module 45
7.2 User Interface Guide 45
7.3 Explainable AI 50

8.0 CONCLUSION 53

9.0 DBPRISK 2.0 - SOFTWARE TOOL FOR SIMULATING DBP FORMATION IN WATER DISTRIBUTION NETWORKS 54

LIST OF TABLES

Table 1: Validation results for the two versions of the case study: Case 1 with the tank serving as the starting point and Case 2 with the WTP. 14
Table 2: Decision Variables of optimal booster placement and dynamic injection scheduling algorithms..... 30

LIST OF FIGURES

Figure 1: Outline of the framework for DBP monitoring and risk assessment..... 11
Figure 2: Limassol case study map with the locations of the four sensor stations... 13
Figure 3: Measured and simulated chlorine residuals considering the measurements at the re-chlorination tank (Tank Outlet) as the ground truth for predicting the residuals at downstream measurement locations of the Limassol Case study. 13
Figure 4: Measured and simulated chlorine residuals considering the measurements at the DWTP (DWTP Outlet) as the ground truth for predicting the residuals at downstream measurement locations of the Limassol Case study. 13
Figure 5: Predicted vs. measured TTHM concentrations, at the DMA Inlet location, from linear regression models using one, two, and three predictors based on data from the Limassol case study. 17
Figure 6: Predicted vs. measured TTHM concentrations, at the D1 location, from linear regression models using one, two, and three predictors based on data from the Limassol case study. 18
Figure 7: Predicted vs. measured TTHM concentrations, at the D2 location, from linear regression models using one, two, and three predictors based on data from the Limassol case study. 18
Figure 8: Predicted vs. measured TTHM concentrations, at the D3 location, from linear regression models using one, two, and three predictors based on data from the Limassol case study. 19
Figure 9: Predicted vs. measured TTHM concentrations, at the D4 location, from linear regression models using one, two, and three predictors based on data from the Limassol case study. 19
Figure 10: Heatmap of statistical significance of individual predictors across TTHM monitoring locations 20
Figure 11: Bar chart comparison of statistical significance of individual predictors across TTHM monitoring locations 20
Figure 12: THM driver contribution analysis for the Limassol case study 22
Figure 13: DMA Consumer demand heatmap indicating areas with high water demand 34
Figure 14: Evolution of THM concentration at a representative network node for the baseline scenario and the early-warning scenario (Left). Corresponding water residence time (Right) 37
Figure 15: Spatial distribution of maximum THM concentrations for two DOC input scenarios under identical chlorine dose and temperature conditions and initial THM level of 45 μg/L at the DWTP outlet: baseline scenario (left) and early warning threshold scenario (right)..... 37
Figure 16: DBP formation control strategies with optimal booster placement across Zone 2 (left) and DMA 230 (right)..... 39

Figure 17: Trade – off curves between tracking error and chemical mass cost for Zone 2 (left) and DMA 230 (right) 39
Figure 18: Diurnal dosing pattern for Zone 2 chlorine boosters 40
Figure 19: Diurnal dosing pattern for DMA 230 chlorine boosters 40
Figure 20: Chlorine residuals at all nodes following the application of the mitigation strategy 41
Figure 21: Simulated THM and HAA concentrations for the baseline and control scenario at the monitoring locations 41
Figure 22: THM daily intake (mg kg^{-1} day^{-1}) frequency distribution under two network operational scenarios based on the real case study conditions 42
Figure 23: THM daily intake (mg kg^{-1} day^{-1}) frequency distribution under two network hypothetical operational scenarios 43
Figure 24: Set of physicochemical data (Temperature, pH, Chlorine, Conductivity) as shown through the initial Grafana interface developed for data visualization 44
Figure 25: Digital Twin Monitoring Dashboard listing predicted parameters at every location of the network 48
Figure 26: DBP Risk Event Summary and Early-Warning web-based platform..... 49

LIST OF ABBREVIATIONS

API: Application Programming Interface

ATL: Ens d'Abastament d'Aigua Ter-Llobregat

CANAL: Canal de Isabel II

CAD: Computer-Aided Design

CP: Critical Point

CSV: Comma-Separated Values file

CY-DBP: Cyprus Disinfection By-Products benchmark network model

D3.3: Deliverable 3.3

DBPs: Disinfection By-Products

DMA: District Metered Area

DOC: Dissolved Organic Carbon

DOM: Dissolved Organic Matter

DT: Digital Twin

DWTP: Drinking Water Treatment Plant

EMT: EPANET-MATLAB Toolkit

EPANET: US EPA water distribution network modeling software

EPANET-MSX: EPANET Multi-Species eXtension

EPyT: EPANET-Python Toolkit

EXPBIO: Exponential biofilm wall-decay model

FCl: Free Chlorine

GIS: Geographic Information System
HAA9: Nine-species Haloacetic Acids group
HAAs: Haloacetic Acids
ICRA: Catalan Institute for Water Research
ICT: Information and Communications Technology
LLM: Large Language Model
MAE: Mean Absolute Error
NOM: Natural Organic Matter
QGIS: Quantum Geographic Information System
R^2: coefficient of determination
Re: Reynolds number
RMSE: Root Mean Square Error
SCADA: Supervisory Control and Data Acquisition
SCAN: (s::can) Water quality sensor manufacturer
THMs: Trihalomethanes
TOC: Total Organic Carbon
TTHMs: Total Trihalomethanes
UCY: University of Cyprus
UNICT: University of Catania
UV: Ultraviolet
UV-Vis: Ultraviolet–Visible
UV254: Ultraviolet absorbance at 254 nm
WBL: Water Board of Limassol
WDN: Water Distribution Network
WP: Work Package
XAI: Explainable Artificial Intelligence

EXECUTIVE SUMMARY

This deliverable presents a holistic framework that uses Digital Twin (DT) technology to support the early detection, exposure risk assessment and mitigation of disinfection by-products (DBPs) in water distribution networks (WDNs), with particular emphasis on trihalomethanes (THMs). The framework extends the dbpRisk 2.0 platform by introducing modules for early warning, exposure risk assessment, mitigation planning and Explainable Artificial Intelligence (XAI), providing water utilities with an integrated decision-support system for monitoring, interpreting and managing DBP formation.

For the early detection of THM formation, a digital twin was developed using the calibrated EPANET-MSX model connected to real-time sensor measurements. The digital twin uses updated hydraulic, operational and water-quality data to predict THM concentrations throughout the distribution network and support early warnings when predicted concentrations approach or exceed predefined thresholds. During routine operation, model predictions are updated daily. Statistical analysis of historical water-quality sampling data supports the identification of relevant THM formation predictors. A Driver Contribution Analysis framework based on counterfactual EPANET-MSX simulations identifies the factors contributing to changes in predicted THM concentrations, helping operators understand the conditions driving elevated levels.

To reduce THM formation while maintaining adequate disinfectant residuals, a two-stage optimization approach determines booster chlorination locations and subsequently their time-varying dosing schedules. The approach balances water quality performance and chemical use while accounting for operational constraints. Application to the Limassol case study demonstrates the early-warning methodology and evaluates the effectiveness of optimized chlorination in limiting additional DBP formation.

Potential risks are estimated using U.S. EPA-based exposure risk metrics, including average daily dose. Population exposure is estimated by allocating the total population served across network nodes in proportion to water demand. These estimates are combined with predicted THM concentrations to assess exposure and compare the potential benefits of mitigation strategies.

The developed algorithms are integrated into a web-based user interface that combines sensor measurements, model predictions, early-warning notifications and explanatory information to support operator decision-making. An XAI framework combines the knowledge and methodologies developed throughout the project with a Large Language Model (LLM). Using simulation outputs, dashboard information and relevant project documents retrieved through Retrieval-Augmented Generation, the LLM provides accessible explanations of predicted conditions, associated risks and potential mitigation actions. This explanatory component helps operators interpret the results produced by the modelling and risk-assessment framework.

1.0 Introduction

This deliverable follows the developments accomplished in D3.3 which included the following:

- the development of the “dbpRisk 2.0” software tool that allows users to predict the formation of DBPs within WDNs,
- the development of a benchmark Water Distribution Network (WDN) model, namely the “CY-DBP”, and
- the exploration of what-if scenarios using the software platform

Following the actions completed in D3.3, the main goal of T3.4 and this deliverable (D3.4) is the development of a holistic framework for the early detection, optimal control and risk minimization of DBPs formation. The main tasks accomplished and presented in D3.4 are listed below:

- the development of model-based and data driven approaches that utilize sensor data for the early detection of DBPs,
- the development of a web-based platform to facilitate real-time monitoring and issue warnings for operators,
- the quantification of DBP risk by utilizing exposure metrics,
- the application of mitigation strategies through optimal placement and scheduling of booster chlorination to reduce the formation and subsequently the impact of DBP formation, and,
- the application of an XAI method to communicate the risks to the decision-makers.

2.0 Framework

2.1 Risk assessment tool

The risk assessment framework illustrated in **Figure 1**, couples different modules aiming to address the potential impact of DBP formation within Water Distribution Networks (WDNs). To that end, the **dbpRisk 2.0** platform was developed in T3.3 to support and evaluate the proposed framework. It is built using the EPANET - Python Toolkit (EPyT)[1] that utilizes the EPANET and EPANET-MSX engines for hydraulic and water quality reaction simulations and visualized using a QGIS customized interface.

A detailed hydraulic model simulates key aspects of the distribution system, including flow and pressure dynamics and water residence times, using computational engines such as EPANET. In parallel, water quality models based on EPANET-MSX are used to simulate disinfection residuals, DBP formation (e.g., THMs and HAAs), and their transport throughout the system. These models incorporate chemical reactions, decay processes, and external environmental influences, such as temperature, driven by the scenarios and informed by sensor data. Furthermore, consumption models simulate the use of water at individual nodes, depending on various conditions.

To address system uncertainties, stochastic methods are required to incorporate variations in water demand, disinfection decay coefficients and reaction rates, as well as fluctuating hydraulic conditions. The outputs are probabilistic distributions that define the likely ranges of DBP concentrations, accounting for model uncertainty. Within these DBP concentration bounds, upper and lower limits can be established to identify the best-case and worst-case exposure to DBPs scenarios. These concentrations can then be coupled with dose-response models to assess the long-term public health impacts. These models are augmented with demographics and population dynamics, considering various parameters such as geographical distribution, age, physiological characteristics, and others. This integration allows for the identification of populations most at risk from DBP exposure under varying system conditions.

The result of this analysis produces the long-term impact probability distribution function, which can then be used to assess the risk. Specifically, to quantify risk, the methodology calculates metrics to capture the expected severity of adverse outcomes. For example, it can evaluate the worst-case scenarios, focusing on the tail end of the impact probability distribution function. The risk metric helps identify the most impactful scenarios, supporting prioritization of control strategies. Finally, decision-makers can evaluate different interventions, such as booster disinfection, valve controls, or disinfection scheduling, to assess the risk reduction.

[1] Kyriakou, M. S., Demetriades, M., Vrachimis, S. G., Eliades, D. G., & Polycarpou, M. M. (2023). EPyT: An EPANET-Python Toolkit for Smart Water Network Simulations. Journal of Open Source Software, 8(92), 5947. https://doi.org/10.21105/joss.05947

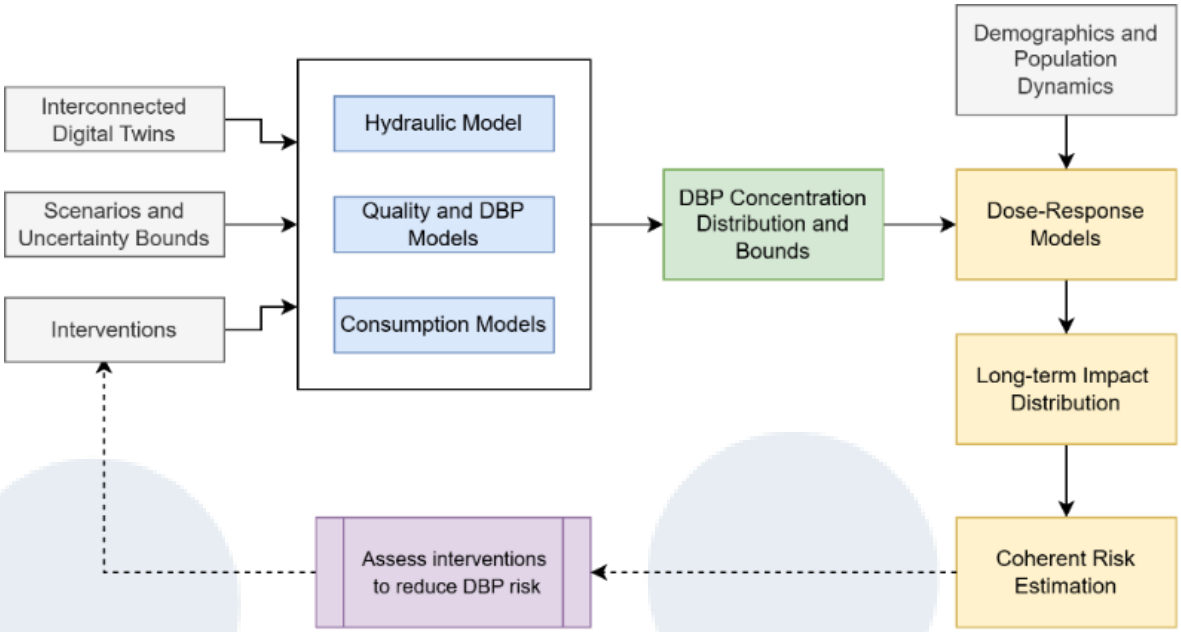


Figure 1: Outline of the framework for DBP monitoring and risk assessment

The approaches are converted into two main modules: a) the "Exposure Risk assessment" module to address the impact from DBP exposure and b) the "Mitigation Strategies" module to assess how the risk is affected. Furthermore, in parallel with the above, real-time modules were developed to support operators to early detect potential DBP formation above acceptable limits.

2.2 *Reaction model validation*

In this subsection the reaction model developed in T3.3 is briefly presented. This reaction model is used to predict the formation of DBPs in WDNs and facilitate the early detection of DBPs formation, and ultimately enable the assessment of mitigation strategies.

At the core of the water quality model is a mechanistic description of chlorine decay and DBP formation driven by reactions with NOM. The chlorine C_{Cl} (mg/L) decay process is described by bulk reactions with NOM fractions, i.e., the fast (FRA) C_{REA-S} (mg/L) and slow-reacting fraction (SRA) C_{REA-S} (mg/L), as well as by reactions with biofilm on the pipe walls:

$$\left(\frac{dC_{Cl}}{dt}\right)_{Bulk} = k_{FRA}\, C_{REA-F}\, C_{Cl} + k_{SRA}\, C_{REA-S}\, C_{Cl},$$

$$\left(\frac{dC_{Cl}}{dt}\right)_{Wall} = (4/D) * r_w,$$

where k_{FRA} and k_{SRA} (L/mgh) the fast and slow reaction rate coefficients, respectively, D the pipe diameter and r_w the wall-surface decay rate.

A defining feature of the model is its explicit support for multiple NOM surrogates. Instead of assuming a single NOM indicator, the model allows to parameterize NOM reactivity using available measurements such as total organic carbon (TOC), dissolved organic carbon (DOC), UV absorbance at 254 nm (UV254), or fluorescence-based indices. This is achieved through the introduction of a surrogate scaling parameter, k_{NOM}, that adjusts the reaction rate equation without modifying the governing equations:

$$k_t = k_{NOM} * \mathrm{k_T} * \exp[-E/R \; x \; (T_{env} - T)/(273 + T_{env})/(273/T)]$$

where k_T the fast or slow reaction rate reference value for a specific temperature T_{env} (e.g. 20°C), k_{NOM} the scaling factor, E the activation energy (Jmol-1), R the universal gas constant (JK-1mol-1) and T the temperature (°C).

DBP formations, such as THMs are estimated utilizing a parallel second order reaction (2R) model which considers the fast and slow reactants for chlorine decay along with fixed yield coefficient Y_{THM} (µg-THM/mg-CL):

$$\frac{dC_{THM}}{dt} = Y_{THM}(\; k_{FRA} \; C_{REA-F} \; C_{Cl} + \; k_{SRA} \; C_{REA-S} \; C_{Cl})$$

Building upon the previously developed reaction model, water temperature is now modelled as an additional bulk species in EPANET-MSX. This provides users with the flexibility to test different temperature modelling formulations within the MSX framework and to couple temperature with other water quality reactions. At this stage, a first-order heat transfer approach reported in the literature was implemented, following the methodology used in drinking water distribution temperature studies[2],[3]. The temperature evolution is described by:

$$\frac{d\mathrm{T}}{dt} = \mathrm{K_T} \; (T_{env} - \mathrm{T}),$$

where T is the bulk water temperature (°C), T_{env} is the surrounding soil/ambient temperature (°C), and $\mathrm{K_T}$ is the first-order heat transfer coefficient. The first-order coefficient is derived from a one-dimensional energy balance and Newton's law of cooling:

$$\mathrm{K_T} = 4 * \frac{h_T}{D \; \rho \; C_p},$$

where h_T is the overall heat transfer coefficient (W m^{-2} K^{-1}), D is the pipe diameter (m), ρ is the water density (kg m^{-3}), and C_p is the specific heat capacity of water (J kg^{-1} K^{-1}). During calibration, a constant effective value of $\mathrm{K_T}$ can be adopted. This value represents an effective heat transfer coefficient that accounts for the combined influence of soil, pipe wall and external heat transfer and was selected to reproduce the observed temperature response.

[2] Blokker, E.J.M. & Pieterse-Quirijns, E.J. (2013). Modeling Temperature in the Drinking Water Distribution System. Journal AWWA.

[3] Blokker, M., Pan, Q., & Van Laarhoven, K. (2024). Validation of an enhanced drinking water temperature model during distribution. Water, 16(19), 2796.

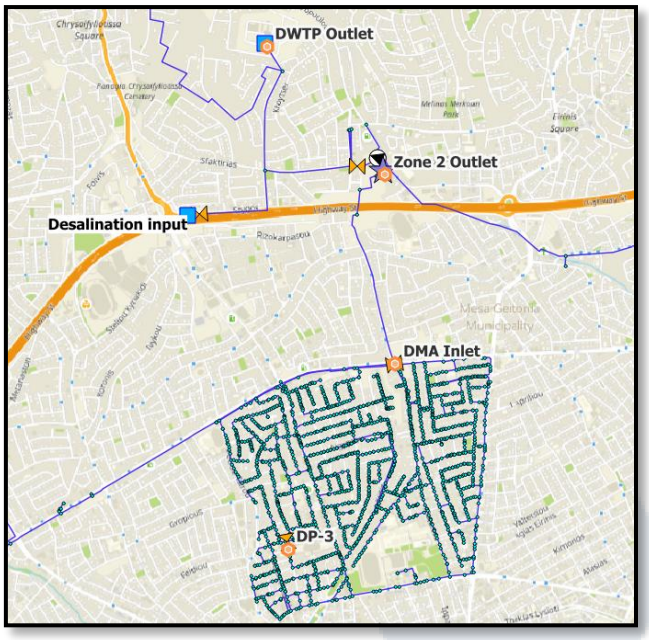


Figure 2: Limassol case study map with the locations of the four sensor stations

Following Deliverable 3.3, the validation of the water quality model of the Limassol case study (**Figure 2**) was re-examined utilizing new approaches to further improve the representation of chlorine dynamics within the system.

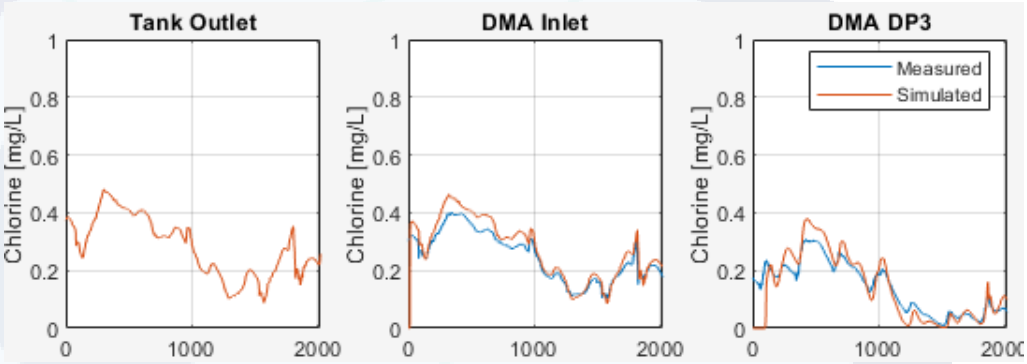


Figure 3: Measured and simulated chlorine residuals considering the measurements at the re-chlorination tank (Tank Outlet) as the ground truth for predicting the residuals at downstream measurement locations of the Limassol Case study.

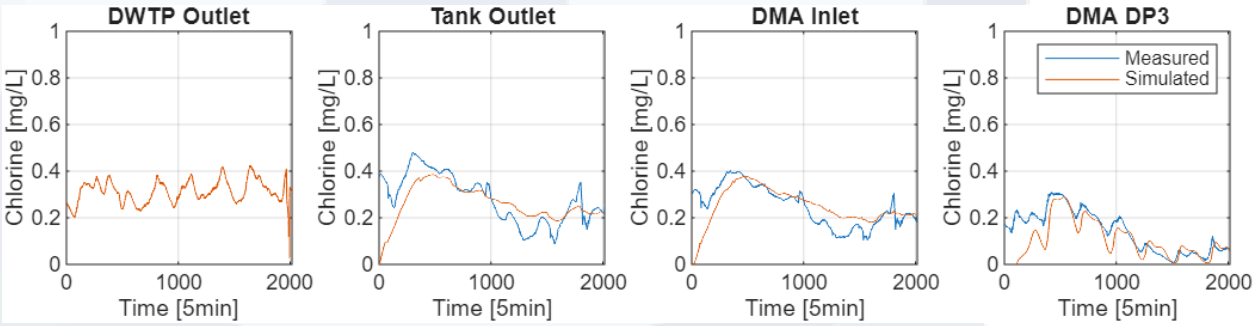


Figure 4: Measured and simulated chlorine residuals considering the measurements at the DWTP (DWTP Outlet) as the ground truth for predicting the residuals at downstream measurement locations of the Limassol Case study.

The validation considers three statistical parameters: the R^2 which indicates how well the model captures the observed variability and trends; values closer to 1 represent

better agreement between simulated and measured data, the RMSE which represents the average magnitude of the difference between simulated and measured values and the MAE indicating the average absolute difference between simulated and measured values.

The metrics are calculated as follows:

- Coefficient of determination (R^2)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

- Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum(y_i - \hat{y}_i)^2}$$

y_i= observed THM concentration

$\hat{y}_i$= predicted THM concentration

$\bar{y}$ = average value of observed values

N= number of observations.

Table 1: Validation results for the two versions of the case study: Case 1 with the tank serving as the starting point and Case 2 with the WTP.

Statistic	Unit	Location 1 – DMA DP3		Location 2 – DMA Inlet	
		Case 1	Case 2	Case 1	Case 2
R^2	–	0.87	0.77	0.86	0.73
RMSE	mg/L Cl_2	0.034	0.044	0.032	0.045
MAE	mg/L Cl_2	0.027	0.031	0.026	0.036

The validation of Case 1 (Zone 2 Tank outlet – DMA 230) remained unchanged, as the model had already demonstrated very good agreement with the measured data. The model achieved R^2 values of 0.87 and 0.86 at Location 1 (DMA DP3) and Location 2 (DMA Inlet), respectively. The associated error metrics were also low, with RMSE values of 0.034 and 0.032 mg/L Cl_2 and MAE values of 0.027 and 0.026 mg/L Cl_2. These results confirm the high accuracy of the calibrated subsystem model and its ability to reproduce the observed chlorine concentration dynamics.

For Case 2 (full system), the re-chlorination process at the Zone 2 Tank was further refined. Instead of relying solely on the average chlorine injection rate provided by the water utility, the applied injection mass rate was adjusted to follow the daily fluctuations observed at the chlorine sensor installed at the Zone 2 Tank outlet. This refinement enabled the model to better reproduce the temporal variability of chlorine concentrations throughout the network. The updated model achieved R^2 values of 0.77 and 0.73 at Location 1 and Location 2, respectively, with RMSE values of 0.044 and 0.045 mg/L Cl_2 and MAE values of 0.031 and 0.036 mg/L Cl_2. Model performance in predicting free chlorine residual remains lower than Case 1 due to the increased complexity and uncertainty associated with representing the full system.

3.0 Analysis of THM Formation Predictors

3.1 Data based approach - Sampling Data Analytics

The objective of this section was to identify statistically significant predictors associated with THM formation for use in the model-based approach for predicting THMs. The analysis was performed using sampling data (provided by UCT and UCY partners), as these measurements are considered more accurate than sensor readings. Since THM formation is influenced by multiple parameters, the statistical analysis was used to identify the most relevant predictors for the model, thereby reducing computational time and uncertainty associated with incorporating all available sensor measurements.

The workflow consists of several analytical steps including data preprocessing, predictor–response model exploration, regression modelling, model performance evaluation and sensitivity analysis:

1. **Data acquisition** Historical water-quality data were collected from upstream stations (DWTP outlet and Z2 tank) and downstream distribution system sampling locations.
2. **Predictor–response definition**
   Upstream parameters (e.g., DOC, TOC, UV254, fluorescence, chlorine, temperature, and pH) were defined as predictors, while THM concentrations at downstream locations were defined as response variables.
3. **Combinatorial regression modelling**
   Linear regression models were constructed using all possible combinations of one, two, and three predictors.
4. **Model evaluation and ranking**
   Models were evaluated using statistical metrics (R^2 and p-values) and ranked according to predictive performance.
5. **Early-warning estimation and sensitivity analysis**
   The selected models can be used to estimate upstream parameter values associated with critical THM concentrations and to quantify parameter sensitivity.

3.1.1 Data Preparation and Variable Definition

The methodology relies on historical water-quality monitoring data collected at the DWTP, the Z2 storage tank, and multiple downstream monitoring points within the distribution network. Laboratory-based measurements are less affected by noise, calibration drift, and signal instability that commonly impact sensor data, ensuring that the observed relationships more accurately reflect the underlying physical and chemical processes. Although fewer in number, these higher-quality observations lead to more robust estimation of model metrics.

Predictor variables were constructed using measurements obtained from the DWTP outlet and the Z2 storage tank. These locations represent key control points where quality of water entering the distribution network can be characterized.

The predictor variables included:

- Organic matter indicators (TOC, DOC)
- Optical surrogate indicators (UV254, fluorescence)
- Disinfectant concentration (free chlorine)

- Physicochemical parameters (temperature, pH)

These variables were selected due to their known influence on chlorine reactivity and THM formation potential.

The response variables correspond to total THM (TTHM) concentrations measured at multiple locations within the distribution network:

- TTHM at DMA inlet
- TTHM at monitoring points D1–D4

These locations represent different hydraulic residence times within the network and therefore capture the spatial evolution of THM formation.

3.1.2 Regression modelling

To systematically identify relationships between predictors and THM concentrations, a combinatorial modelling approach was implemented. For each response variable, linear regression models were constructed using all possible combinations of:

- single predictors
- pairs of predictors
- triplets of predictors

This exhaustive search approach allowed the evaluation of the predictive capability of different variable combinations.

The general form of the regression model is:

$$TTHM_j = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon$$

where:

- $TTHM_j$= total trihalomethane concentration at response location j (µg/L)
- X_i= predictor variable i measured at upstream monitoring locations
- β_0= intercept coefficient
- β_i= regression coefficient associated with predictor X_i
- n= number of predictors included in the model
- ε= residual error term

3.1.3 Performance Indicators

For each candidate model, as before, the following metrics were calculated:

- Coefficient of determination (R^2)
- Root Mean Square Error (RMSE):
- Statistical significance of regression coefficients (p-values)

  For each regression coefficient $\hat{\beta}_j$, a p-value is computed based on the ratio between the estimated coefficient and its standard error:

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

The p-value reflects how strongly each predictor is associated with the response variable. Smaller p-values indicate stronger evidence that the corresponding predictor contributes to explaining variability in the response, whereas larger p-values suggest a weaker or negligible contribution.

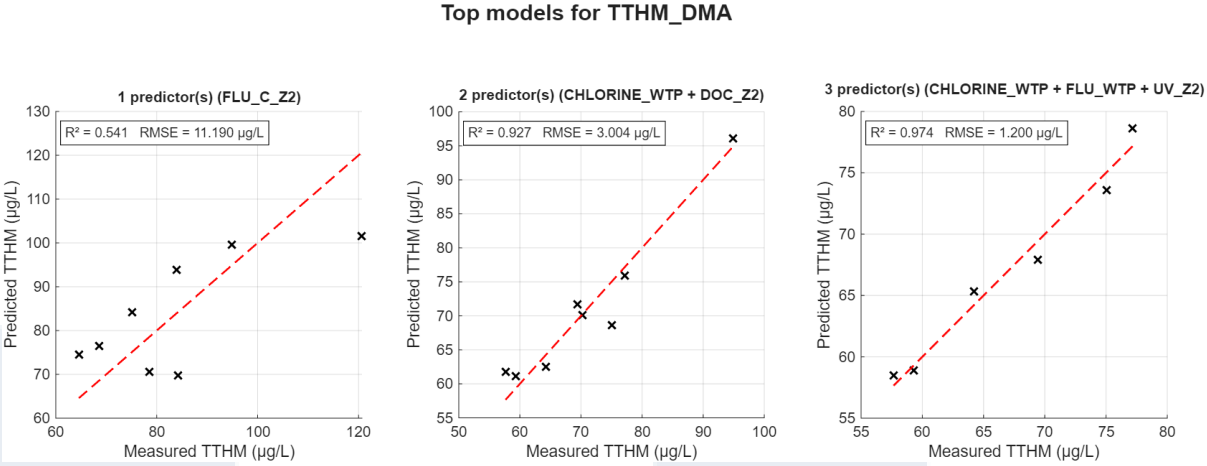


Figure 5: Predicted vs. measured TTHM concentrations, at the DMA Inlet location, from linear regression models using one, two, and three predictors based on data from the Limassol case study.

Figure 5, presents the best-performing regression models for predicting TTHM concentrations at the TTHM_DMA monitoring location using one, two, and three predictor variables derived from measurements collected at the DWTP and the Z2 storage tank. Each panel shows the relationship between measured and model-predicted TTHM concentrations, with the dashed red line representing the 1:1 line indicating perfect agreement.

For the **single-predictor case**, fluorescence measured at the Z2 tank (FLU_C_Z2) was identified as the most informative variable (R^2 = 0.541, RMSE = 11.19 µg/L). Although the predictive performance is moderate, this result indicates that fluorescence at Z2, an indicator related to the presence and reactivity of NOM, is the most effective single parameter among the tested variables for estimating downstream THM formation. This suggests that if only one parameter can be monitored for early indication of THM formation potential, fluorescence measurements at the Z2 tank provide the most useful information.

When **two predictors** are included, the optimal model combines chlorine concentration at the DWTP (CHLORINE_WTP) with dissolved organic carbon at the Z2 tank (DOC_Z2), resulting in a substantial improvement in predictive performance (R^2 = 0.927, RMSE = 3.004 µg/L). This combination captures both key components controlling THM formation: disinfectant availability (chlorine) and the presence of reactive organic precursors (DOC). Importantly, the predictors originate from two different locations (DWTP and Z2), highlighting the benefit of integrating upstream treatment plant conditions with intermediate storage tank water quality.

The **three-predictor model** provides the highest predictive accuracy (R^2 = 0.974, RMSE = 1.200 µg/L) and combines chlorine and fluorescence measurements at the DWTP (CHLORINE_WTP, FLU_WTP) with UV absorbance at the Z2 tank (UV_Z2). This model again integrates measurements from **both monitoring locations**, rather than relying solely on data from a single site. The inclusion of variables from both the treatment plant and the storage tank suggests that accurate prediction of THM formation in the distribution network benefits from capturing both the initial water

quality conditions leaving the DWTP and the evolving organic matter characteristics at Z2.

Overall, the results indicate that predictive capability improves significantly as additional relevant predictors are included. However, they also highlight the importance of combining measurements from different points within the system. While a single parameter measured at Z2 (fluorescence) provides the best standalone indicator, the most accurate predictions are obtained when variables representing both disinfectant levels and organic matter characteristics from multiple locations (DWTP and Z2) are jointly considered. This finding supports the use of distributed monitoring strategies for improving early detection and prediction of THM formation in water distribution networks.

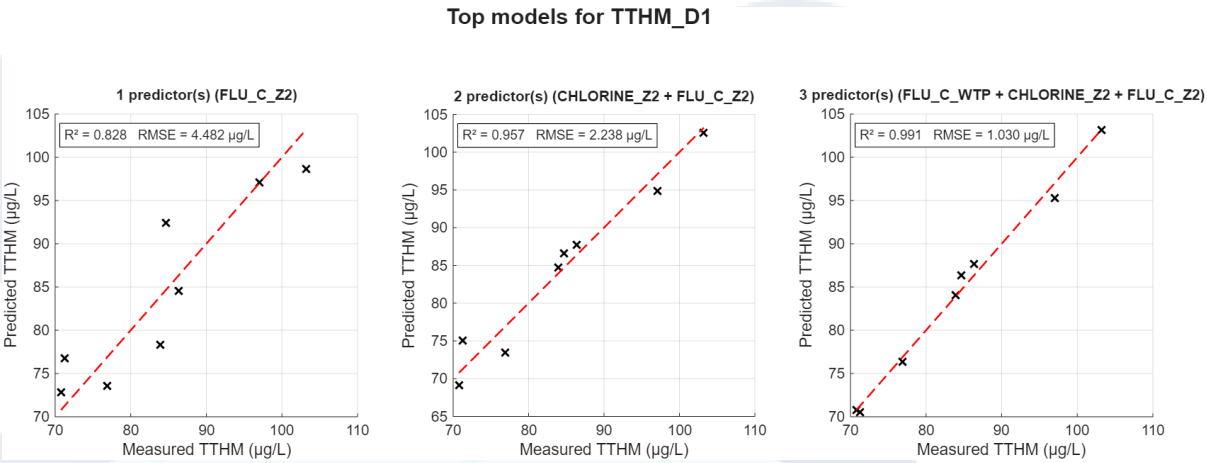


Figure 6: Predicted vs. measured TTHM concentrations, at the D1 location, from linear regression models using one, two, and three predictors based on data from the Limassol case study.

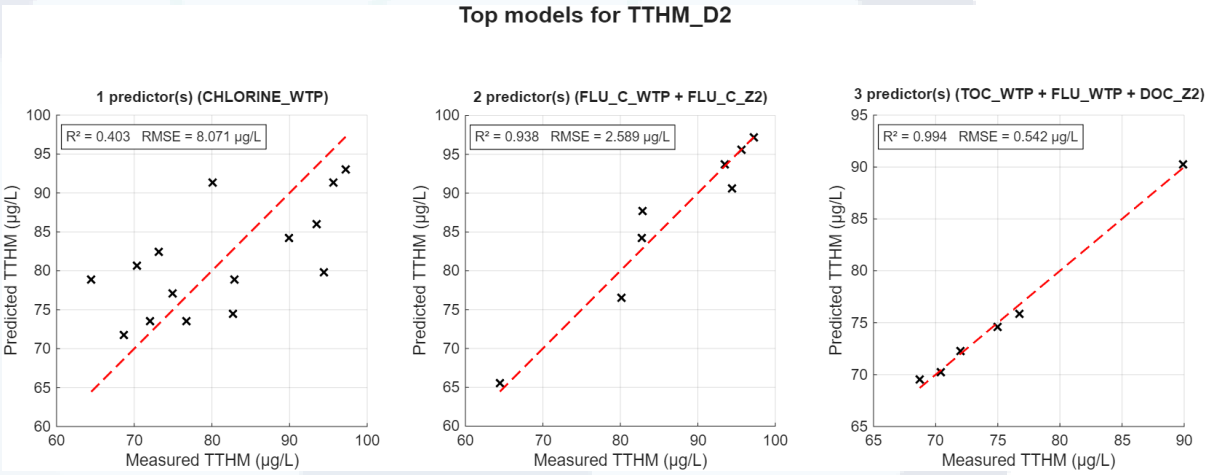


Figure 7: Predicted vs. measured TTHM concentrations, at the D2 location, from linear regression models using one, two, and three predictors based on data from the Limassol case study.

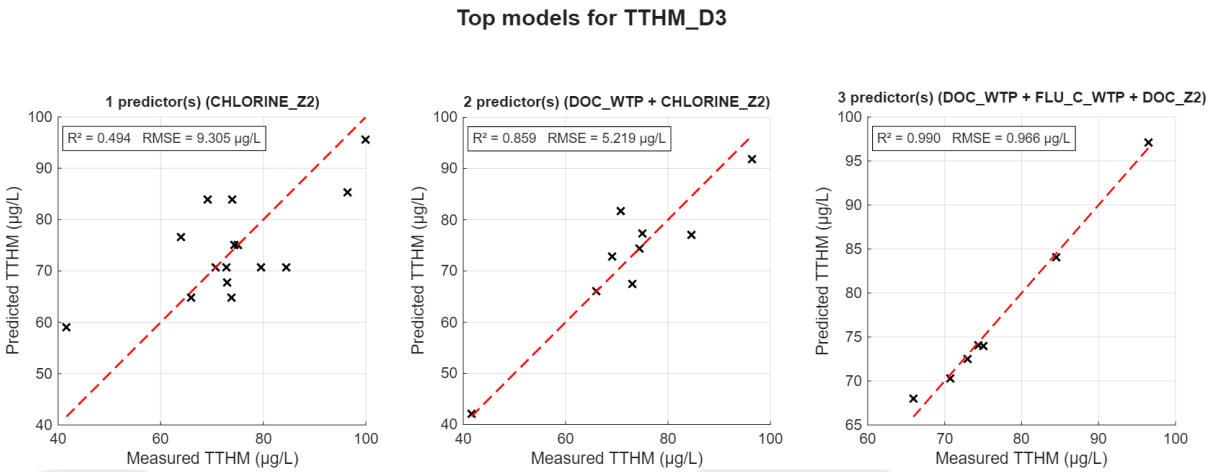


Figure 8: Predicted vs. measured TTHM concentrations, at the D3 location, from linear regression models using one, two, and three predictors based on data from the Limassol case study.

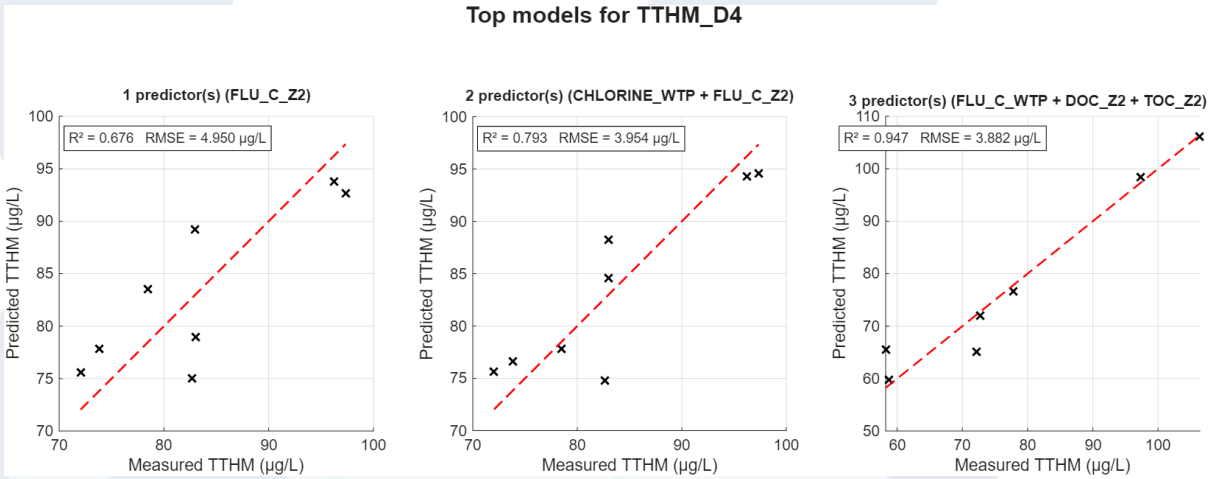


Figure 9: Predicted vs. measured TTHM concentrations, at the D4 location, from linear regression models using one, two, and three predictors based on data from the Limassol case study.

3.1.4 Predictor Significance Analysis

To evaluate the statistical importance of individual predictors across all response locations, p-values obtained from the regression models were analyzed.

The statistical significance of each predictor was visualized using:

- significance heatmaps
- bar charts based on −log10(p-values)

The use of $-log_{10}(p-values)$ provides a more interpretable and visually informative representation of statistical significance. Since p-value values lie in the range 0 - 1 and can be extremely small, direct visualization may compress meaningful differences. The negative logarithmic transformation expands small p-values and compresses larger ones, enhancing contrast in heatmaps. Smaller p-values (indicating stronger statistical evidence) are mapped to larger positive values, making them more visually prominent.

The transformation follows:

$$S = -\log_{10}(p)$$

where S denotes the significance score. Thus, higher values of S correspond to stronger statistical significance. In practice:

- $S < 1$(i.e., $p > 0.1$) → weak or no evidence of association
- $S \approx 1.3$(i.e., $p \approx 0.05$) → marginal significance
- $S \geq 2$(i.e., $p \leq 0.01$) → strong evidence
- $S \geq 3$(i.e., $p \leq 0.001$) → very strong evidence

This transformation enables a clearer spatial comparison of predictor importance, as highly significant regions are emphasized by larger values, while non-significant areas remain visually subdued.

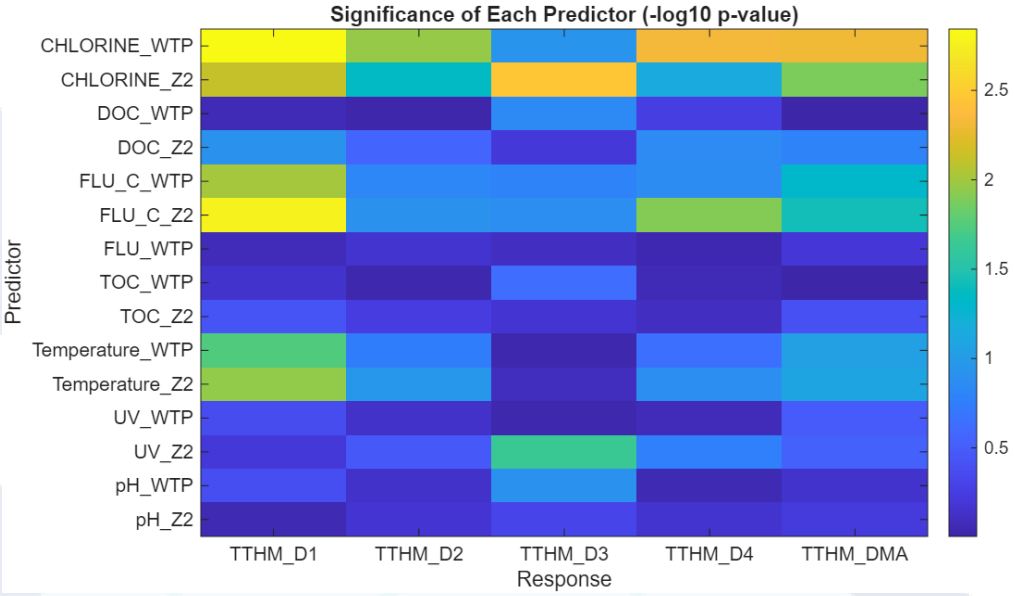


Figure 10: Heatmap of statistical significance of individual predictors across TTHM monitoring locations

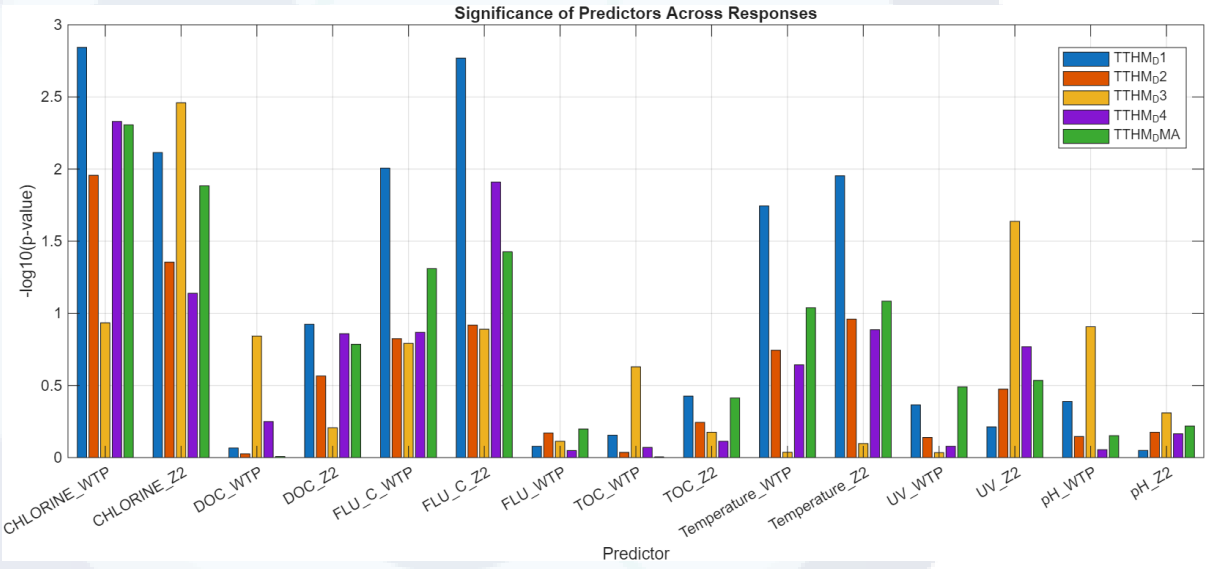


Figure 11: Bar chart comparison of statistical significance of individual predictors across TTHM monitoring locations

Figure 10 presents a heatmap of predictor significance, while **Figure 11** shows the corresponding bar chart representation across all response variables (TTHM_D1–TTHM_DMA). The analysis was performed using sampling measurements, with predictors evaluated individually at each location, where TTHM_DMA represents the

entry point of the distribution system. The examined predictors include chlorine, fluorescence (FLU and FLU_C), DOC, TOC, temperature, UV and pH.

The results indicate that chlorine and fluorescence exhibit the strongest and most consistent associations with TTHM formation across the network. CHLORINE_WTP shows high significance at the inlet (TTHM_DMA) and early distribution points, while CHLORINE_Z2 and FLU_C_Z2 demonstrate increased importance further along the network (e.g., TTHM_D3 and TTHM_D4). Temperature (both WTP and Z2) shows moderate but consistent significance, suggesting a secondary influence on TTHM variability. UV absorbance presents a relatively strong association at TTHM_D3 while DOC, TOC, and pH generally exhibit low significance values, suggesting limited significance when considered independently. Evaluating predictors individually is important as it isolates their direct relationship with the response variables, free from interactions or shared variance with other predictors. This approach enables the identification of key drivers of TTHM formation without the confounding effects of multicollinearity and avoids the computational complexity associated with testing all possible predictor combinations. Consequently, it provides a robust step for screening influential variables.

3.2 Model based approach - Driver contribution analysis

The objective of the Driver Contribution Analysis (DCA) module is to explain why a high THM concentration is predicted by identifying the operational and environmental factors that contributed most to the forecast. The DCA does not directly identify abnormal values in individual contributing parameters. However, when a parameter is identified as a major contributor to an increase in predicted THM concentrations, this indicates that an abnormal change in that parameter relative to previous operating conditions has contributed to the predicted increase. For example, if NOM is identified as a major contributor and UV absorbance is used as its surrogate parameter in the model, the result indicates that the observed UV signal has undergone an abnormal change relative to previous conditions. The user can further track the corresponding measurements through the online platform (see Section 7.0) to investigate recent changes and trends in the underlying measurements. This functionality is intended to support water utility operators by increasing transparency in the prediction process and enabling informed operational decisions.

3.2.1 Offline Baseline Scenario

The proposed methodology is based on a series of EPANET-MSX simulations in which the value of a single model input is increased by, e.g. 20% relative to the baseline operating conditions. The evaluated drivers include:

1. Initial chlorine concentration at the WTP outlet
2. Initial NOM level at the WTP outlet
3. Initial THM concentration at the WTP outlet
4. Water temperature
5. Rechlorination at the Storage Tank

The resulting THM concentration at the monitoring location, e.g. DMA point “DP3”, is extracted from each simulation and compared with the baseline prediction.

Sequential (Consecutive) Simulations:

In this approach, the simulations are executed sequentially. Starting from the baseline operating conditions, the first simulation increases the first driver by 20%. The second simulation retains the previous modification and additionally increases the second driver by 20%. This procedure continues until all selected drivers have been introduced.

The incremental THM increase associated with each driver is calculated as

$$\Delta THM_i = \begin{cases} THM_1 - THM_{base}, i = 1 \\ THM_i - THM_{i-1}, i > 1 \end{cases}$$

where THM_i is the predicted THM concentration after introducing the (i)-th driver.

The total increase is given by

$$\Delta THM_{total} = THM_{final} - THM_{base}$$

and the contribution of each driver is calculated as

$$Contribution_i = 100 * \frac{\Delta THM_i}{\Delta THM_{total}}$$

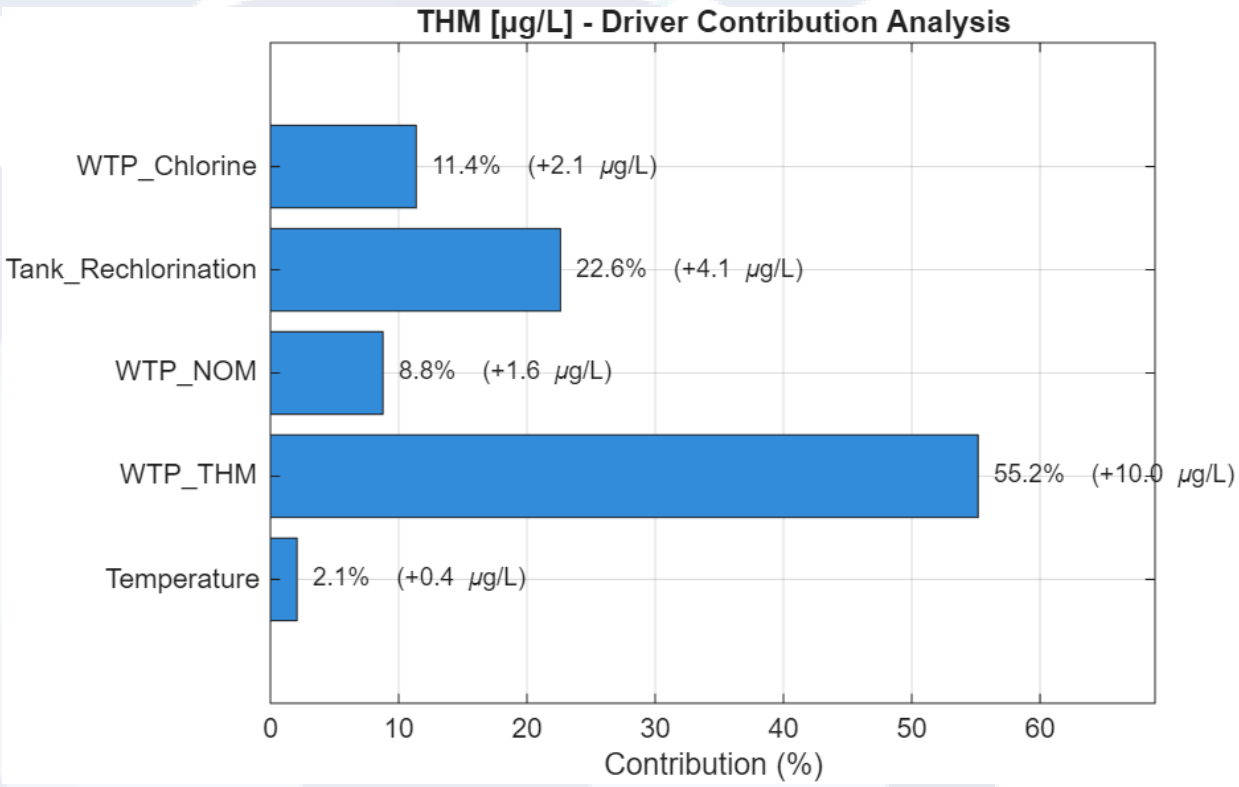


Figure 12: THM driver contribution analysis for the Limassol case study

The results indicate that the initial THM concentration at the WTP is the dominant driver, accounting for approximately 55% of the total THM concentration. This suggests that the quality of the water entering the distribution system has the greatest influence on the final THM levels and highlights the importance of treatment plant performance in controlling THM formation.

Among the distribution system operational parameters, tank rechlorination is identified as the most influential driver, contributing approximately 23% of the total increase. This reflects the significant role of additional chlorine dosing within the network in

promoting THM formation through continued reactions with NOM matter. The initial chlorine concentration at the treatment plant contributes approximately 11%, while the initial NOM concentration accounts for approximately 9%. In contrast, water temperature has a relatively minor contribution (approximately 2%) for the conditions examined, indicating that temperature variations are less influential than chlorine-related processes and source water quality within the analysed period.

It should be emphasized that these contribution percentages are specific to the analysed WDN, its hydraulic characteristics, source water quality, treatment processes and operational conditions. The relative importance of the different drivers is expected to vary between utilities and may also change within the same network under different seasonal or operational conditions. For example, systems with longer water residence times may exhibit a greater influence of water age and temperature, whereas systems supplied with higher background THM concentrations may be dominated by the initial THM concentration entering the distribution network.

Therefore, the DCA should not be interpreted as providing universal contribution factors, but rather as a system-specific explainability tool that identifies the dominant drivers for the particular network and operating conditions represented by the calibrated EPANET-MSX model. Such information enables water utility operators to prioritize operational interventions by focusing on the parameters that exert the greatest influence on THM formation within their own distribution system.

3.2.2 Assumptions and Limitations

This approach has the advantage that the individual contributions sum exactly to the total predicted THM increase. Consequently, it provides an intuitive decomposition of the final prediction and is particularly suitable for dashboard visualization and communication with system operators. **However, because each simulation builds upon the previous one, the calculated contribution depends on the order in which the drivers are introduced**. Drivers evaluated later are assessed under operating conditions that already include the effects of the preceding drivers. Therefore, the results should be interpreted as a sequential attribution rather than a unique physical decomposition.

The DCA evaluates one driver at a time while maintaining all remaining parameters at their baseline values. Consequently, the methodology quantifies the influence of individual parameters but does not explicitly investigate combined operational changes. For example, scenarios involving simultaneous increases in initial chlorine concentration and water temperature, or concurrent changes in NOM concentration and water age, are not considered.

Although such mixed conditions may occur in practice, evaluating every possible combination of drivers would require a substantially larger number of EPANET-MSX simulations and would significantly increase the computational effort. The adopted methodology therefore represents a practical compromise between computational efficiency, interpretability and operational applicability.

3.2.3 Online Application

The DCA described in the previous section provides a quantitative assessment of the relative importance of the different THM formation drivers under the baseline operating conditions. However, during routine operation, the digital twin is executed once every day using updated hydraulic, operational and water quality measurements. Consequently, the objective is no longer to determine which parameters are generally

influential, but rather to identify which parameter(s) are responsible for the change in the predicted THM concentration compared with the previous day's simulation.

To provide an operational explanation of each daily prediction, a real-time DCA is performed after every execution of the digital twin. The proposed methodology is based on counterfactual simulations, which evaluate how the predicted THM concentration would change if one of the updated input variables had remained unchanged while all other conditions were identical. Unlike conventional sensitivity analyses, this approach does not assume linear relationships between the model inputs and THM formation, but instead directly exploits the calibrated EPANET-MSX model to preserve the nonlinear reaction kinetics and hydraulic interactions occurring within the distribution system.

Each day, the digital twin retrieves the most recent measurements from the monitoring system, including the time series of chlorine concentration, source THM concentration, NOM, water temperature, tank rechlorination settings and other available operational variables. Since the digital twin operates using a rolling seven-day simulation window, each model input is represented by a complete seven-day time series consisting of 2016 measurements (5-minute sampling interval). These time series constitute the forcing functions supplied to the EPANET-MSX model.

The baseline daily prediction is obtained by executing the model using the complete set of current seven-day input time series,

$$THM_{today} = f(\mathbf{X}_1^{today}, \mathbf{X}_2^{today}, \ldots, \mathbf{X}_n^{today})$$

where $\mathbf{X}_i^{today}$ denotes the complete seven-day time series of driver i.

To determine the influence of each driver, a series of counterfactual simulations is subsequently performed. For each driver individually, only its seven-day time series is replaced by the corresponding time series used during the previous daily simulation, while all remaining inputs are maintained equal to their current values. Thus, for driver i,

$$THM_i^{cf} = f(\mathbf{X}_1^{today}, \ldots, \mathbf{X}_i^{yesterday}, \ldots, \mathbf{X}_n^{today})$$

where THM_i^{cf} denotes the counterfactual THM prediction obtained by restoring only driver i to its previous-day conditions.

The influence of driver i is then estimated as

$$\Delta THM_i = THM_{today} - THM_i^{cf}$$

which represents the change in the predicted THM concentration resulting solely from the observed variation of the corresponding input between two consecutive daily simulations.

Finally, the relative contribution of each driver is obtained by normalizing the individual influences,

$$Contribution_i = \frac{| \Delta THM_i |}{\sum_{j=1}^{n} | \Delta THM_j |} \times 100$$

where n is the total number of monitored drivers.

Because each counterfactual simulation modifies only one driver while preserving the complete seven-day time series of all remaining variables, the calculated contribution reflects the isolated influence of that particular driver under the actual operating conditions of the current day. Consequently, the methodology preserves all nonlinear interactions represented by the EPANET-MSX model, including chlorine decay, NOM reactions, temperature effects, hydraulic residence time and mixing processes.

The computational requirements remain modest. Following the baseline daily simulation, only one additional EPANET-MSX simulation is required for each monitored driver. Consequently, if five operational variables are considered, the complete explainability analysis requires six model executions (one baseline prediction and five counterfactual simulations). To avoid unnecessary computations and explanations for normal day-to-day fluctuations, the analysis is performed only when a significant change in the predicted THM concentration is observed between two consecutive daily simulations. The daily THM variation is calculated as

$$\Delta THM = THM_{today} - THM_{yesterday}$$

The DCA is activated when

$$| \Delta THM | \geq \Delta THM_{threshold}$$

where $\Delta THM_{threshold}$ is the minimum change considered operationally significant.

Unlike traditional feature importance metrics used in machine learning, the proposed methodology derives these explanations directly from the calibrated EPANET-MSX model and therefore remains physically consistent with the hydraulic and water quality processes governing THM formation within the drinking water distribution system.

4.0 Mitigation strategies

A major challenge in drinking water distribution systems is the inherent trade-off between maintaining sufficient disinfectant residuals to guarantee microbiological safety and minimizing the formation of DBPs. Increasing chlorine dosage generally improves residual disinfectant concentrations but simultaneously promotes DBP formation. Conversely, reducing chlorine dosage can limit DBPs but may result in insufficient residual concentration, particularly in remote sections of the network.

This challenge becomes significantly more complex in large-scale WDNs, where hydraulic conditions, water age, demand patterns, and water quality exhibit considerable spatial and temporal variability. Such variations make it difficult to achieve consistent water quality throughout the network using fixed operational strategies.

Consequently, operational decisions, including treatment plant dosing, booster chlorination scheduling, and hydraulic management, play a critical role in balancing disinfectant residuals and DBP formation. Developing effective mitigation strategies that account for these interacting factors is therefore essential for maintaining regulatory compliance while ensuring safe and high-quality drinking water.

To achieve this objective, the study focuses on the following:

- Identify the operational and water quality parameters within a WDN that influence disinfectant decay and DBP formation.
- Investigate controllable operational variables, including disinfectant dosing, booster chlorination schedules, and hydraulic management strategies.
- Develop optimization approaches that balance effective disinfection with the minimization of DBP formation.
- Ensure compliance with water quality regulations by maintaining adequate disinfectant residual concentrations throughout the distribution network.

Several operational strategies have been proposed to mitigate the formation of DBPs while maintaining adequate disinfectant residuals throughout WDNs. These strategies focus on improving hydraulic performance, reducing water residence times, and optimizing disinfectant application. A summary of potential strategies is presented below:

- Operational optimization to reduce water age: Adjust network operation, including pump scheduling and valve control, improve storage tank turnover, optimize filling and emptying cycles, and minimize dead-end conditions to minimize water residence time and limit the formation of DBPs associated with prolonged disinfectant contact.
- Targeted flushing programs: Periodically flush selected sections of the network, particularly dead-end pipes and low-demand areas, to remove stagnant or aged water and restore acceptable water quality.
- Booster disinfection management: Optimize the location, timing, and dosing rates of booster chlorination stations to maintain adequate disinfectant residuals while avoiding excessive chlorine concentrations that promote DBP formation.
- Disinfectant injection scheduling: Develop dynamic disinfectant dosing schedules that account for temporal variations in water demand, hydraulic conditions, and water quality, enabling more efficient chlorine application than fixed dosing strategies.

Specific conditions of the Limassol case study limited the applicability of several of these approaches. The study area consists of a fully residential water distribution network with a relatively fixed hydraulic configuration and limited operational flexibility. Consequently, large-scale modifications to pump scheduling, valve operation or storage tank management were not considered feasible. Similarly, routine flushing programs were not selected, as they are generally implemented as maintenance activities or corrective measures and would result in unnecessary water losses in a normally operating residential network. Instead, the proposed mitigation strategy **focuses on booster disinfection management and dynamic disinfectant injection scheduling,** as these provide direct operational control over disinfectant residuals while minimizing THM formation without requiring physical modifications to the existing infrastructure. Furthermore, these strategies can be readily integrated into the Digital Twin framework, enabling real-time optimization based on the prevailing hydraulic and water quality conditions. Accordingly, this section presents two sequential optimization problems: (i) the optimal placement of booster chlorination stations, and (ii) the determination of optimal time-varying chlorine injection schedules for the selected booster locations. Both optimization problems exploit the principle of superposition to linearize the relationship between chlorine injections and network-wide disinfectant residuals, enabling efficient solutions using mathematical programming techniques.

Multi-objective optimization provides a mathematical framework for simultaneously considering several competing objectives without combining them into a single performance measure. Instead of producing a single optimal solution, these methods generate a set of Pareto-optimal solutions[4], where no objective can be further improved without causing deterioration in at least one other objective. The provided solutions enable decision-makers to evaluate the trade-offs between different operational strategies and select solutions that best satisfy practical and regulatory requirements.

4.1 Application Theory

The optimization of booster chlorination requires evaluating thousands of different combinations of booster locations, injection times, and chlorine dosages. Directly simulating each candidate strategy with a hydraulic-water quality model (e.g., EPANET) is computationally expensive. Under the assumption of fixed hydraulics and first-order chlorine decay, chlorine transport is a linear system. Consequently, the principle of superposition can be exploited[5]: the response to any arbitrary injection schedule can be reconstructed as a weighted sum of precomputed unit-response simulations, transforming repeated hydraulic-quality simulations into a linear algebra problem, and enabling efficient optimization.

Definition

Let $\mathcal{L}: \mathcal{U} \rightarrow \mathcal{Y}$ be a linear differential operator mapping an input space of booster injection rates $\mathcal{U}$ to an output space of nodal chlorine concentration fields $\mathcal{Y}$. If inputs

[4] Deb, K.: Multi-objective optimisation using evolutionary algorithms: An introduction. In: Multi-objective Evolutionary Optimisation for Product Design and Manufacturing, pp. 3–34. Springer, London (2011)

[5] Vrachimis, S.G., Eliades, D.G. & Polycarpou, M.M. (2024). Disinfection Scheduling in Water Distribution Networks Considering Input Time-Delay Uncertainty. *Journal of Hydroinformatics*, 26(2), 386–396.

$u_a, u_b \in \mathcal{U}$ produce system responses $y_a, y_b \in \mathcal{Y}$ such that $\mathcal{L}\{u_a\} = y_a$ and $\mathcal{L}\{u_b\} = y_b$, then for any scalar coefficients $\alpha, \beta \in \mathbb{R}$:

$$\mathcal{L}\{\alpha \cdot u_a + \beta \cdot u_b\} = \alpha \cdot y_a + \beta \cdot y_b$$

Validity for Chlorine Transport in WDNs

Chlorine dynamics in a WDN with N_N nodes and N_B booster injection points satisfy superposition under two core operational constraints:

1. **Decoupled Hydraulics:** The hydraulic state of the system (flow velocities v and flow paths) is calculated prior to the water quality simulation and remains time-invariant across quality iterations.
2. **First-Order Decay Kinetics:** Chlorine bulk and wall decay reactions are strictly modeled as first-order differential processes $\left(\frac{dC}{dt} = -kC\right)$.

Under these assumptions, the cumulative chlorine residual field Y resulting from multiple simultaneous booster injection strategies is equal to the linear summation of individual impulse-response matrices.

Mathematical Formulation of the Superposition Approach

Let $T_s \in \mathbb{N}^+$ denote the total number of time steps, $N_N \in \mathbb{N}^+$ the number of network nodes, and $N_B \in \mathbb{N}^+$ the number of candidate booster stations.

The network-wide chlorine concentration matrix $Y \in \mathbb{R}^{T_s \times N_N}$ is expressed as:

$$Y = \sum_{b=1}^{NB} \sum_{t=1}^{Ts} a_{tb} H\{t, b\}$$

where*:*

- $\mathrm{Y} \in \mathbb{R}^{\mathrm{T_s} \times \mathrm{N_N}}$ represents the full dynamic network chlorine concentration matrix, where entry $\mathrm{Y_{\tau,n}}$ is the chlorine residual at timestep τ at node n.
- $\mathrm{H} \in \mathbb{R}^{\mathrm{N_B} \times \mathrm{T_s} \times \mathrm{T_s} \times \mathrm{N_N}}$ represents the global impulse-response tensor.
- $\mathrm{H^{\{t,b\}}} \in \mathbb{R}^{\mathrm{T_s} \times \mathrm{N_N}}$ is the dynamic response matrix across all $\mathrm{N_N}$ nodes over $\mathrm{T_s}$ timesteps generated by a unit mass injection impulse applied at timestep t from booster location b.
- $\mathrm{a_{t,b}} \in \mathbb{R}$ is the continuous dosage scaling factor (multiplier) associated with unit mass injection at time t and booster station b.

Objective

The objective is to find the unique chlorine injection matrix Y^* that minimizes the distance to the reference matrix $Y_{\mathrm{ref}} \in \mathbb{R}^{T_s \times N_N}$:

$$Y^* = \arg\min_Y \| Y - Y_{\mathrm{ref}} \| = \arg\min_a \| \sum_{b=1}^{NB} \sum_{t=1}^{Ts} a_{tb} H\{t, b\} - Y_{\mathrm{ref}} \|$$

where $Y^* \in \mathbb{R}^{T_s \times N_N}$denotes the optimal chlorine injection matrix closest to the reference matrix in the chosen norm.

Definition (Vectorization Operator)

Let $\mathrm{vec}: \mathbb{R}^{n\times m} \to \mathbb{R}^{nm\times 1}$ be the linear transformation that converts a matrix $A \in \mathbb{R}^{n\times m}$ into a single column vector by stacking its columns sequentially:

$$\mathrm{vec}(A) = \begin{bmatrix} \mathrm{col}_1(A) \\ \mathrm{col}_2(A) \\ \vdots \\ \mathrm{col}_m(A) \end{bmatrix} \in \mathbb{R}^{nm\times 1}$$

where $\mathrm{col}_i(A)$ denotes the i-th column vector of matrix A.

Implementation

Applying $\mathrm{vec}(\cdot)$ to the matrix sum yield:

$$\mathrm{vec}(Y) = \mathrm{vec}\left(\sum_{b=1}^{N_B}\sum_{t=1}^{T_s} a_{t,b}\, H^{\{t,b\}}\right) = \sum_{b=1}^{N_B}\sum_{t=1}^{T_s} a_{t,b}\cdot \mathrm{vec}\left(H^{\{t,b\}}\right) = A\tilde{a}$$

where

- $\mathrm{vec}(Y) \in \mathbb{R}^{(T_s N_N)\times 1}$ is the vectorized system response vector.
- $A \in \mathbb{R}^{(T_s N_N)\times(T_s N_B)}$ is the block influence matrix whose columns are given by $A_{[:,(b-1)T_s+t]} = \mathrm{vec}\left(H^{\{t,b\}}\right)$.
- $\tilde{a} \in \mathbb{R}^{(T_s N_B)\times 1}$ is the concatenated injection vector:

$$\tilde{a} = [a_{1,1} \quad a_{1,2} \quad \dots \quad a_{1,N_B} \quad a_{2,1} \quad \dots \quad a_{T_s,N_B}]^T$$

- $b = \mathrm{vec}(Y_{\mathrm{ref}}) \in \mathbb{R}^{(T_s N_N)\times 1}$ is the vectorized target concentration field.

Minimization Problem

Let $f(\tilde{a})$ be the unconstrained sum-of-squares objective function

$$f(\tilde{a}) = \|A\tilde{a} - b\|_2^2 = (A\tilde{a} - b)^T(A\tilde{a} - b)$$

Expanding the transposition:

$$f(\tilde{a}) = \tilde{a}^T A^T A\tilde{a} - \tilde{a}^T A^T b - b^T A\tilde{a} + b^T b$$

Since $\tilde{a}^T A^T b \in \mathbb{R}$ is a scalar quantity, $\tilde{a}^T A^T b = (\tilde{a}^T A^T b)^T = b^T A\tilde{a}$. Thus:

$$f(\tilde{a}) = \tilde{a}^T A^T A\tilde{a} - 2b^T A\tilde{a} + b^T b$$

To find the stationary point, take the gradient with respect to parameter vector $\tilde{a}$:

$$\nabla_{\tilde{a}} f(\tilde{a}) = 2A^T A\tilde{a} - 2A^T b$$

Setting $\nabla_{\tilde{a}} f(\tilde{a}) = \mathbf{0}$ yields the normal equations:

$$A^T A\tilde{a} = A^T b$$

If the gramian matrix $A^T A$ is nonsingular (invertible), the unique optimal injection vector $\tilde{a}^*$ is:

$$\tilde{a}^* = (A^T A)^{-1} A^T b$$

If A is square ($T_s N_N = T_s N_B$) and non-singular, this reduces directly to:

$$\tilde{a}^* = A^{-1} b$$

4.2 Decision Variables

The operational optimization framework consists of two sequential phases: (i) Booster Location Placement, and (ii) Dynamic Injection Scheduling as presented in **Table 2**.

Table 2: Decision Variables of optimal booster placement and dynamic injection scheduling algorithms

Variable	Mathematical Domain	Description
Stage 1: Booster Placement		
x_{i}	$\{0,1\}, \forall i \in \{1, \dots, n\}$	Binary indicator variable equal to 1 if node i is selected as a booster facility, 0 otherwise.
a_{i}	$[0,1], \forall i \in \{1, \dots, n\}$	Continuous activation variable governing the proportional injection intensity at node i.
e_{j}	$\mathbb{R}_{\geq 0}, \forall j \in \{1, \dots, m\}$	Continuous slack tracking variable quantifying soft-constraint deviation at sensor node j.
Stage 2: Injection Scheduling		
$u_{b(t)}$	$[0, U_{\max}], \forall b, t$	Continuous chlorine dosing rate $(\mathrm{mg/h})$ at booster b during timestep t.
x	$\mathbb{R}^{T_s N_B \times 1}$	Stacked multi-period decision vector comprising all booster schedules: $\mathbf{x} = \left[u_1(1), \dots, u_{N_B}(T_s)\right]^T$.
$s^{+}_{\mathrm{j}}(t), s^{-}_{\mathrm{j}}(t)$	$\mathbb{R}_{\geq 0}$	Upper and lower bound slack penalty variables at node j and timestep t.

4.3 Objective Functions

Stage 1 — Booster Placement (Mixed-Integer Lexicographic Optimization)

The placement formulation balances network deviation against capital/operational baseline costs:

$$\min_{x,a,e} \quad F_1 = \sum_{j=1}^{m} W_j \cdot e_j$$

$$\min_{x,a,e} \quad F_2 = \sum_{i=1}^{n} a_i \cdot Q_{\text{mass}}$$

where:

- $W_j = \frac{D_j}{\sum_k D_k}$ represents the normalized base demand weighting coefficient at junction node j, with base demand D_j.
- e_j is the absolute residual tracking scalar error at node j.
- Q_{mass} represents the benchmark mass flow scale factor for network dosing.

Optimization Priority: Objectives are solved lexicographically (hierarchically), prioritizing F_1 residual tracking over F_2 mass expenditure. The Pareto optimal boundary is generated by evaluating cardinality boundaries $m \in \{1, \dots, m_{\max}\}$.

Stage 2 — Injection Scheduling (Regularized Quadratic Program)

Given a fixed subset of selected booster nodes N_B, scheduling is optimized via the following quadratic performance index:

$$\min_{\mathbf{x}, s^+, s^-} \quad \|D\mathbf{x} - r_{\text{ref}}\|_2^2 + \mu\|\mathbf{x}\|_2^2 + \gamma \sum_{b=1}^{N_B} \sum_{t=2}^{T_s} \big(u_b(t) - u_b(t-1)\big)^2 + \lambda \sum_j \sum_t \left(s_j^+(t) + s_j^-(t)\right)$$

where:

- $D \in \mathbb{R}^{(T_s N_N) \times (T_s N_B)}$ the discretized sensitivity/response matrix mapping injection schedules to nodal concentration profiles.
- $r_{\text{ref}} \in \mathbb{R}^{T_s N_N}$ is the spatial-temporal reference residual vector (target chlorine concentrations).
- $\mu > 0$ is the Tikhonov regularization parameter penalizing excessive chemical mass usage.
- $\gamma > 0$ is the temporal variation penalty factor penalizing rapid dynamic swings in dosing pump controls.
- $\lambda \gg 0$ ($\lambda = 10^5$) is the penalty parameter enforcing soft concentration bounds via slacks s^+ and s^-.

4.4 Constraints

To guarantee physical realism, regulatory compliance, and operational feasibility across both stages of the multi-stage optimization framework, a series of structural, logical, and physical bounds are imposed.

Stage 1

The first stage selects candidate nodes for booster station installation. The constraints governing this integer-allocation phase ensure capital budget adherence, operational logic, and target tracking bounds.

1. Cardinality Limit:

$$\sum_{i=1}^{n} x_i = m$$

- **Mathematical Definition:** Restricts the sum of binary selection variables $x_i \in \{0,1\}$ across all n candidate junction nodes to an exact integer value m.
- **Engineering Rationale & Operational Context:** Municipal water utilities face capital expenditure (CapEx) limits, installation footprint constraints, and maintenance budget caps. Setting m limits the total number of installed booster stations (e.g., $m = 2$ for Zone 2, $m = 3$ for DMA 230). This allows decision-makers to evaluate performance across discrete budget levels along the Pareto front.

2. Dosing Linkage Constraints:

$$a_i \leq x_i, \quad \forall i \in \{1, \dots, n\}$$

- **Mathematical Definition:** Couples the continuous baseline dosing intensity $a_i \in [0,1]$ directly to the binary location indicator variable $x_i \in \{0,1\}$.
- **Engineering Rationale & Operational Context:** Enforces fundamental physical logic: a node cannot inject chemical mass ($a_i > 0$) unless a physical booster station has been constructed and activated at that location ($x_i = 1$). If $x_i = 0$, the constraint strictly forces $a_i = 0$, preventing unallocated chemical injection in the mathematical formulation.

3. Soft Residual Target Bounding:

$$\sum_{i=1}^{n} H_{ij}\, a_i - e_j \leq r_{\text{set}}, \quad \forall j \in \{1, \dots, m\}$$

$$-\sum_{i=1}^{n} H_{ij}\, a_i - e_j \leq -r_{\text{set}}, \quad \forall j \in \{1, \dots, m\}$$

- **Mathematical Definition:** Evaluates network response via impulse matrix H_{ij} against target residual concentration r_{set} at sensor node j, utilizing non-negative slack variable $e_j \geq 0$ to capture absolute tracking error $|\cdot|$.
- **Engineering Rationale & Operational Context:** Water distribution networks exhibit complex hydraulic dead-ends and varying residence times. Hard concentration targets across every node often result in infeasible linear programs. By formulating target tracking with soft-slack variables (e_j), the placement optimization penalizes target deviations in the objective function while ensuring the solver converges to an optimal layout.

Stage 2

Once booster locations are fixed (N_B), the second stage calculates optimal dynamic dosing profiles across the simulation horizon (T_s). The following constraints enforce water quality regulations and pump hardware capabilities.

1. Upper/Lower Soft Residual Limits:

$$D\mathbf{x} \leq U_b \cdot \mathbf{1} + s^+$$

$$-D\mathbf{x} \leq -L_b \cdot \mathbf{1} + s^- \Leftrightarrow D\mathbf{x} \geq L_b \cdot \mathbf{1} - s^-$$

- **Mathematical Definition:** Maps dynamic dosing schedule vector $\mathbf{x}$ to nodal residual profiles via sensitivity matrix $\mathbf{D}$. Deviations above upper limit U_b or below lower limit L_b are captured by non-negative slack vectors $\mathbf{s}^+, \mathbf{s}^- \geq 0$.
- Engineering Rationale & Operational Context:
  - **Lower Limit ($L_b \approx 0.20$–0.30 mg/L):** Guarantees minimum secondary disinfection to prevent microbial regrowth and biological contamination in customer taps.
  - **Upper Limit ($U_b \approx 0.50$–0.60 mg/L):** Limits excess disinfectant concentration to prevent aesthetic taste/odor complaints and restrict Disinfection By-Product (DBP) formation (e.g., Trihalomethanes and Haloacetic Acids).
  - **Soft Bounding:** Applying heavily penalized slacks ($\lambda = 10^5$) prevents optimization failure during transient hydraulic shifts (such as sudden demand surges or hydraulic propagation delays) while keeping operational concentrations strictly bounded during normal operation.

2. Operational Bounds & Non-Negativity:

$$0 \leq u_b(t) \leq U_{\max}, \quad \forall b, t$$

$$s_j^+(t) \geq 0, \quad s_j^-(t) \geq 0, \quad \forall j, t$$

- **Mathematical Definition:** Sets strict upper and lower numerical bounds on time-series control inputs $u_b(t)$ and penalty variables $s_j^+(t), s_j^-(t)$.
- Engineering Rationale & Operational Context:
  - **Non-Negativity ($u_b(t) \geq 0$):** Chlorine cannot be physically removed or extracted from water pipes once injected. Dosing pumps can only add chemical mass or turn off.
  - **Maximum Injection Capacity ($U_{\max}$):** Constrains dosing rates to the physical output capacity of installed dosing pumps (e.g., maximum volumetric flow rate in mg/h). This prevents the mathematical algorithm from recommending impractical dosage rates that exceed hardware specifications.

5.0 DBP exposure risk assessment

5.1 Population dynamics

The DBP exposure risk assessment requires an estimate of the population served by each demand node. As detailed demographic information is not available, the nodal population is estimated proportionally to the assigned water demand (**Figure 13**). As described in D3.3, during the hydraulic modelling procedure, to accurately simulate the flow at every pipe and pressure at every node, each consumer/water meter was represented as a junction and integrated to the model. The average hourly demand for each junction was estimated utilizing data from the utility's billing system.

The total serviced population of the study area is known (P_{total} = 10,224 inhabitants), together with the total network demand (Q_{total}) and the base demand allocated to each node (Q_i). The population served by node i is estimated as:

$$P_i = \frac{P_{total} * Q_i}{Q_{total}}$$

where P_i is the estimated population served by node i, P_{total} is the total serviced population, Q_i is the base demand at node i, and Q_{total} is the total network demand.

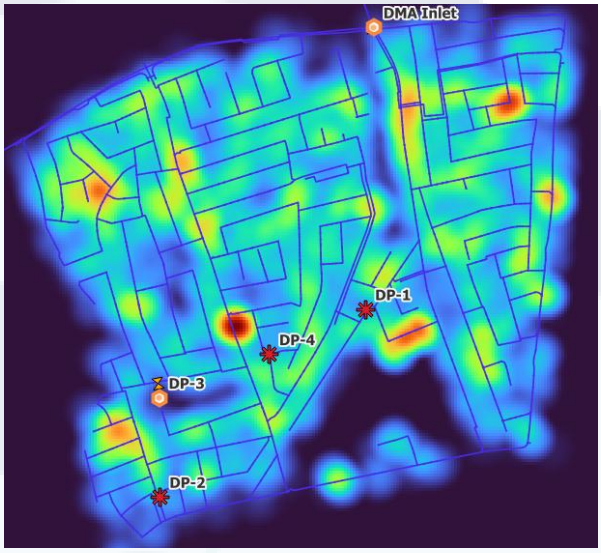


Figure 13: DMA Consumer demand heatmap indicating areas with high water demand

This approach assumes that domestic water consumption is proportional to the number of residents connected to each demand node. Given that the study area consists exclusively of residential consumers, this assumption provides a reasonable approximation of the spatial population distribution. The estimated nodal population is subsequently used to quantify the number of consumers exposed to elevated THM concentrations and to estimate the associated exposure risks. Further, the resident population assigned to each node is assumed to remain constant throughout the simulation period, and the exposure assessment is based on the permanent population distribution. Although daily population dynamics are neglected, the proposed framework can readily accommodate longer-term demographic variations, such as seasonal population changes, residential development or updated census information, by modifying the total population or the nodal demand allocation.

Consequently, the methodology remains applicable to future changes in population distribution without requiring modifications to the risk assessment framework.

The estimated nodal population is combined with the predicted THM concentrations to determine the number of consumers potentially exposed at each node, enabling the spatial quantification of population exposure across the distribution network.

5.2 Exposure metrics

We assess the public health impact utilizing the health risk assessment model developed by the US EPA[6] which estimates the Average Daily Dose (ADD) (mg kg^{-1} day^{-1}):

$$ADD = \frac{C * IR * EF * ED}{BW * AT}$$

where C the pollutant concentration (mg/L), IR the ingestion rate of daily drinking water (L day^{-1}), EF the exposure frequency (days $year^{-1}$), ED the exposure duration (year), BW the body weight (kg) and AT the average time (days).

Monte Carlo simulations can be employed to address uncertainties including ingestion rate, body weight, and exposure duration.

[6] United States Environmental Protection Agency: National Primary Drinking Water Regu-lations: Stage 2 Disinfectants and Disinfection Byproducts Rule. In: Final Rule, pp. 387–493. U.S. Environmental Protection Agency, Washington, DC (2006)

6.0 Case study – Application

The implementation of the proposed framework is illustrated initially utilizing the developed algorithms through the EPANET-MATLAB Toolkit and the CY-DBP benchmark network. Following, the proposed framework along with the developed algorithms which were converted into smart apps were integrated into the updated software platform **dbpRisk 2.0.**

To examine the applicability of the proposed framework, the following three scenarios are evaluated, a) estimation of early indicators to predict the DBP formation before reaching critical levels, b) assessment of the impact associated with DBP exposure and c) evaluation of the effectiveness of mitigation strategies in reducing exposure risk. A seven-day baseline simulation was performed utilizing time series data retrieved from on-the-field sensors. The average initial concentrations at the DWTP outlet were set to 0.4 mg/L for free chlorine, 2 mg/L for DOC, and a variation between 15-55 μg/L for THMs, representing DBPs formed during treatment. Although UV absorbance and fluorescence-based parameters showed greater sensitivity to THM formation, DOC was selected as a representative measure of NOM in this application due to its widespread use for monitoring organic matter and DBP precursors as well as its relatively straightforward measurement and interpretation by water utilities. Re-chlorination is performed at the downstream tank using the mass booster method, adding approximately 0.15 mg/L of chlorine. A uniform water temperature of 20 °C was assumed throughout the simulation.

6.1 DBP Formation Estimation

In this scenario, the DOC and THM threshold at the outlet of the DWTP are examined with all other conditions held constant. To estimate these thresholds, a THM threshold and the percentage of the affected consumption or population required to trigger an alarm must be defined, e.g. at least 10% of the network nodes should exceed the THM threshold of 80 μg/L. The EU regulatory limit for TTHMS is set to 100 μg/L[7],however, since the objective is to provide an early warning and considering that TTHMs can continue to form and increase throughout the distribution network, particularly following re-chlorination, a lower threshold was defined to allow potential exceedances to be identified at an earlier stage. As shown in D3.3, and based on the sampling campaign, TTHM concentrations can increase by more than 20 μg/L in this case study network. The model indicates that an initial DOC concentration exceeding 3 mg/L, in combination with a THM concentration of 45 μg/L at the DWTP outlet will lead to these conditions within approximately 30 hours if measured at the DWTP, or within 10 hours if measured at the storage tank (**Figure 14** and **Figure 15**). Integration of hydraulic and demographic data further indicates that 12% of the network's nodes are affected, which corresponds to about 8% of the population served by the examined area.

[7] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62022CJ0481

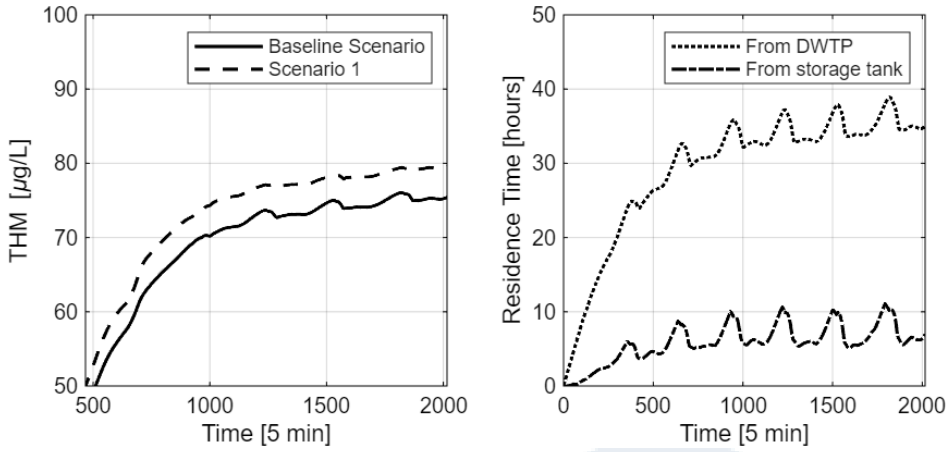


Figure 14: Evolution of THM concentration at a representative network node for the baseline scenario and the early-warning scenario (Left). Corresponding water residence time (Right)

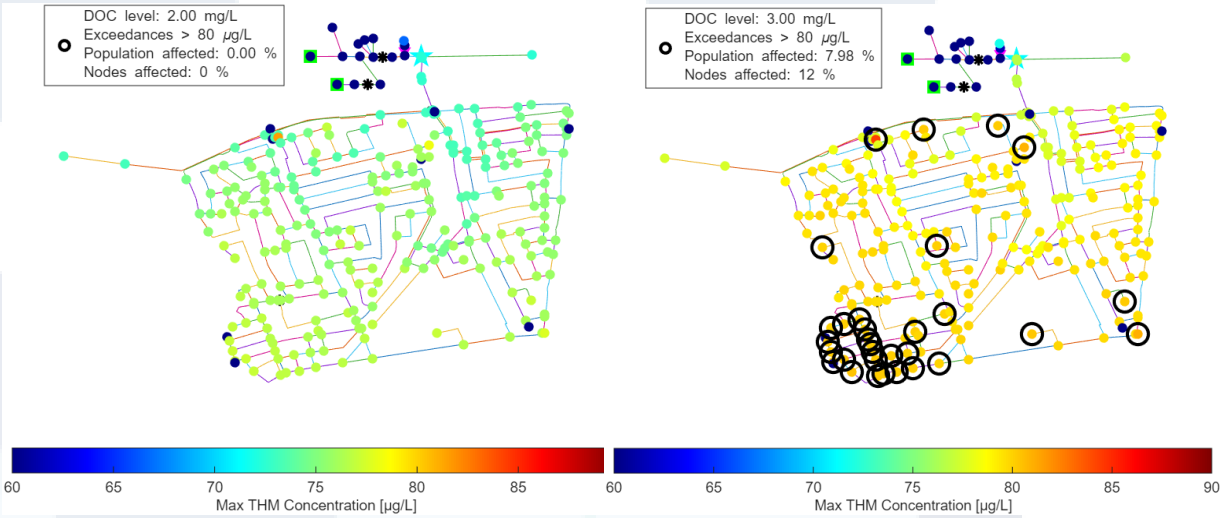


Figure 15: Spatial distribution of maximum THM concentrations for two DOC input scenarios under identical chlorine dose and temperature conditions and initial THM level of 45 µg/L at the DWTP outlet: baseline scenario (left) and early warning threshold scenario (right)

6.2 Mitigation Strategies

This section demonstrates the operational application of the proposed chlorine residual optimization methodology to two distinct levels of the water distribution system:

- Zone 2 Transport Network (Primary Transport System)
- DMA 230 Distribution Network (Downstream Distribution Area)

For both case studies, the impulse-response matrix (H) was generated for all candidate booster locations, and the MILP booster placement problem was solved for m = 1 – 3 booster stations. The optimal configuration was selected through a Pareto trade-off analysis, balancing the demand-weighted chlorine tracking error against the total injected chlorine mass. The selected booster configuration was subsequently used in the dynamic chlorine injection scheduling optimization.

Zone 2:

Following the booster placement optimization, the Pareto Front analysis (**Figure 17**) identified a configuration consisting of two (2) booster stations, located in Area_1 and Area_3, as the preferred solution (**Figure 16** - left). This configuration provided the best compromise between chlorine residual compliance and chemical consumption by ensuring adequate disinfectant coverage throughout the main transmission system without excessive chlorine injection.

Dynamic Scheduling & Performance Evaluation

The dynamic scheduling optimization was executed over a 192-hour (8-day) simulation horizon at a 5-minute resolution ($T_s = 2304$ time steps), with a bulk decay rate $k_b = -1.5$ day^{-1} and a wall decay rate $k_w = -0.5$ m/day. An 8-day horizon was selected to account for initial hydraulic propagation lag; approximately 18 hours are required for chlorine to traverse the transmission network completely.

- **Injection Dynamics (Figure 18):** Initial dosage rates peak sharply during the warm-up phase to establish baseline network residuals. Once steady-state conditions are reached, the injection rates follow a repeating diurnal pattern that closely reflects the system demand and flow variations.
- **Residual Compliance (Figure 20):** Chlorine concentrations across target nodes remain well-bounded within the required 0.30 mg/L $\leq C \leq 0.60$ mg/L regulatory range of the main transport pipelines.

DMA 230:

Based on the booster placement optimization and Pareto Front analysis (**Figure 17**), a configuration comprising three (3) booster stations at nodes dist458, dist1268, and dist1443, was selected (**Figure 16** - right). Compared with the transmission system, the distribution network requires a greater number of booster stations to compensate for localized hydraulic conditions, higher water age and accelerated chlorine decay, particularly in low-demand and dead-end areas.

Dynamic Scheduling & Performance Evaluation

The dynamic scheduling optimization was performed over the same 192-hour ($T_s = 2304$) operational window:

- **Injection Dynamics (Figure 19):** Due to localized residence times, the optimization applies a high initial dose to meet node demand immediately. Following this initial stabilization period, dosing rates settle into steady diurnal oscillations aligned with peak consumer demand hours.
- **Residual Compliance (Figure 20):** Chlorine residuals throughout the distribution network are successfully maintained within the target operating range of 0.10–0.50 mg/L. Minor deviations observed during the initial simulation period are associated with the model warm-up phase and disappear once steady-state conditions are reached.

The baseline scenario assumes an initial THM concentration of 50 µg/L at the DWTP outlet. As water travels through the system, THM concentrations progressively increase due to the continued reaction of residual chlorine with NOM. Re-chlorination at the downstream storage tank further promotes THM formation, while the increased water residence times in the tank and distribution network contribute to the observed accumulation. At the furthest monitoring locations, THM concentrations reached approximately 66 µg/L (**Figure 21**). Following the implementation of the proposed

optimization strategy, the maximum downstream THM concentration decreased to approximately 60 μg/L (**Figure 21**), corresponding to a reduction of around 6 μg/L (~9%) compared with the baseline operation and a ~38% reduction regarding the THM formation downstream of the Zone Tank. Since a substantial proportion of the THMs are already formed at the DWTP outlet (50 μg/L), only a certain reduction range can be achieved within the distribution system. Furthermore, the hydraulic characteristics of the network provide limited opportunity to reduce water age; therefore, the principal mitigation mechanism available is the optimization of chlorine injection and rechlorination practices. The obtained results demonstrate that optimized disinfectant management can effectively suppress additional THM formation while maintaining the required chlorine residuals throughout the network. A similar trend was observed for HAAs (**Figure 21**).

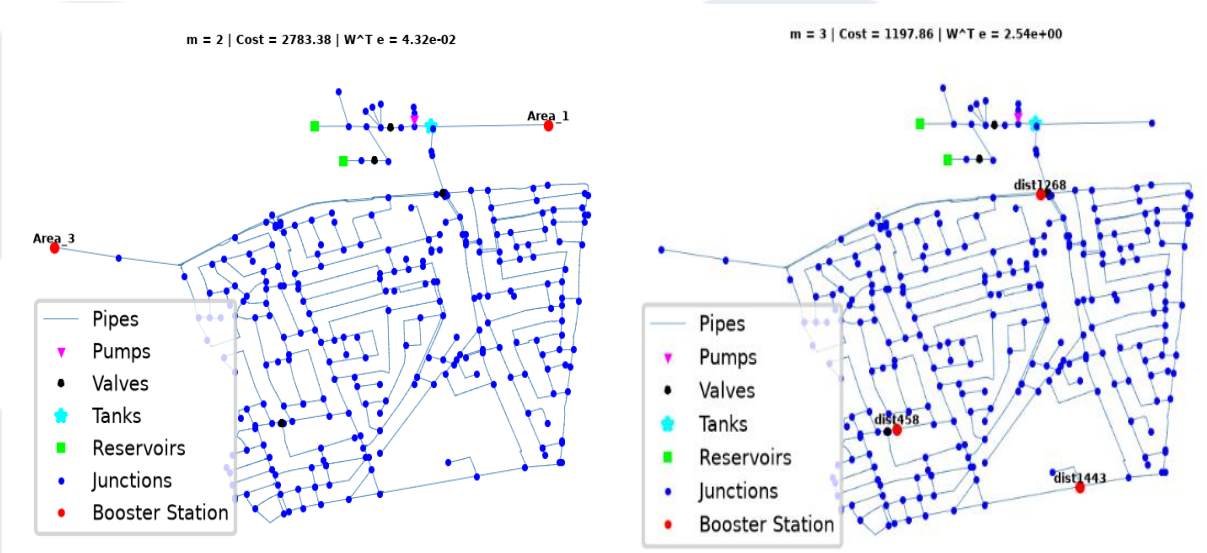


Figure 16: DBP formation control strategies with optimal booster placement across Zone 2 (left) and DMA 230 (right)

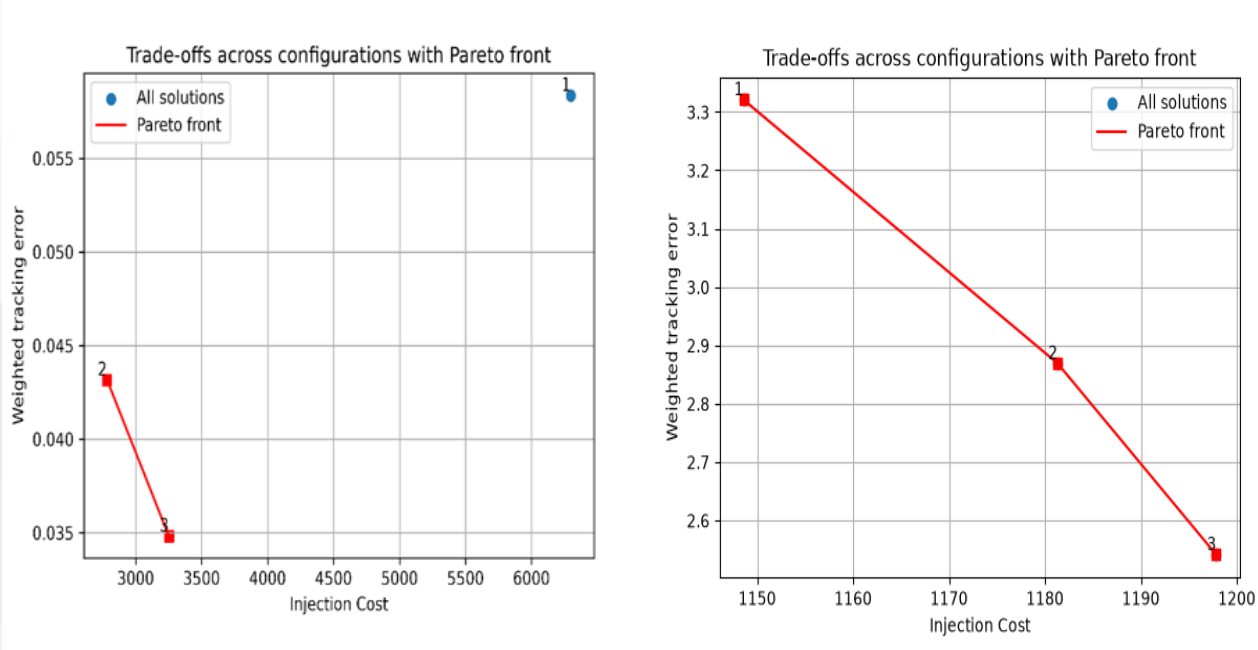


Figure 17: Trade – off curves between tracking error and chemical mass cost for Zone 2 (left) and DMA 230 (right)

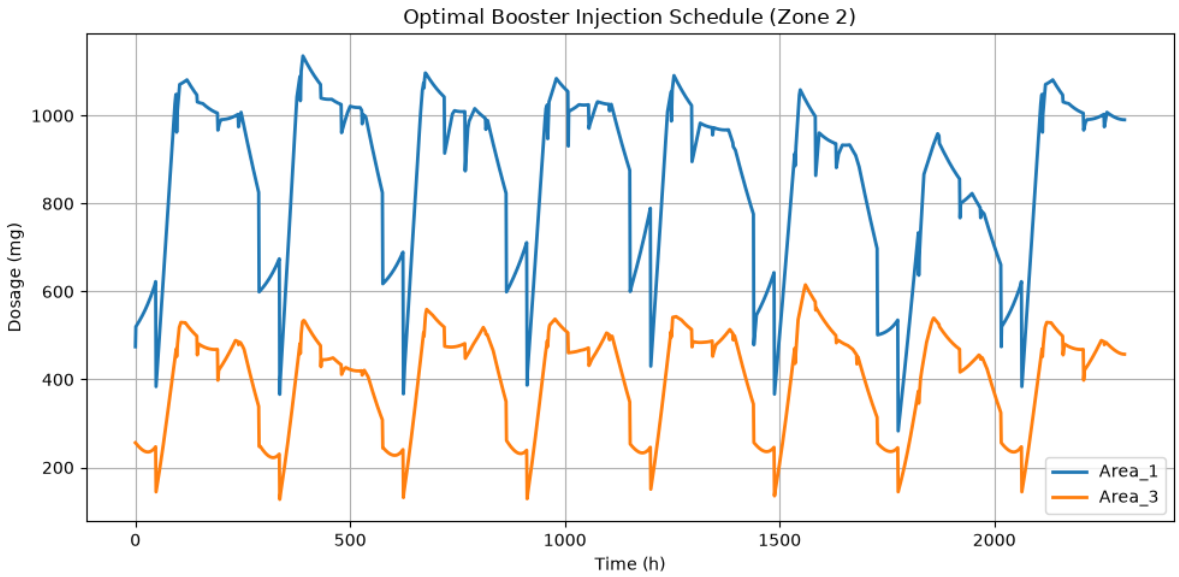


Figure 18: Diurnal dosing pattern for Zone 2 chlorine boosters

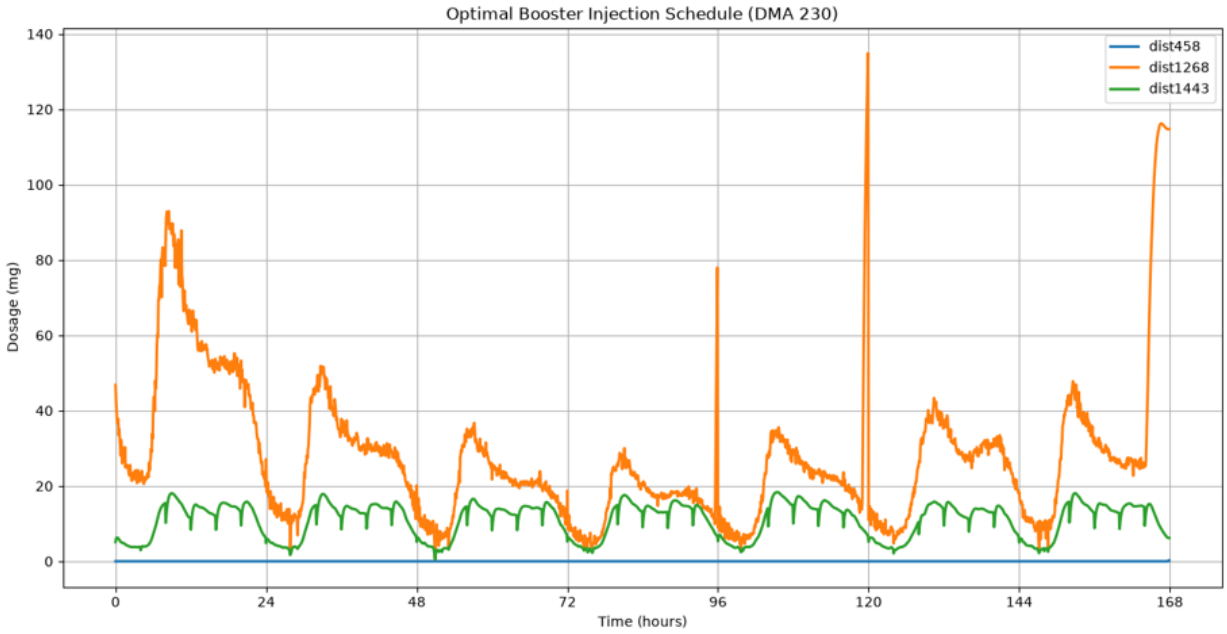


Figure 19: Diurnal dosing pattern for DMA 230 chlorine boosters

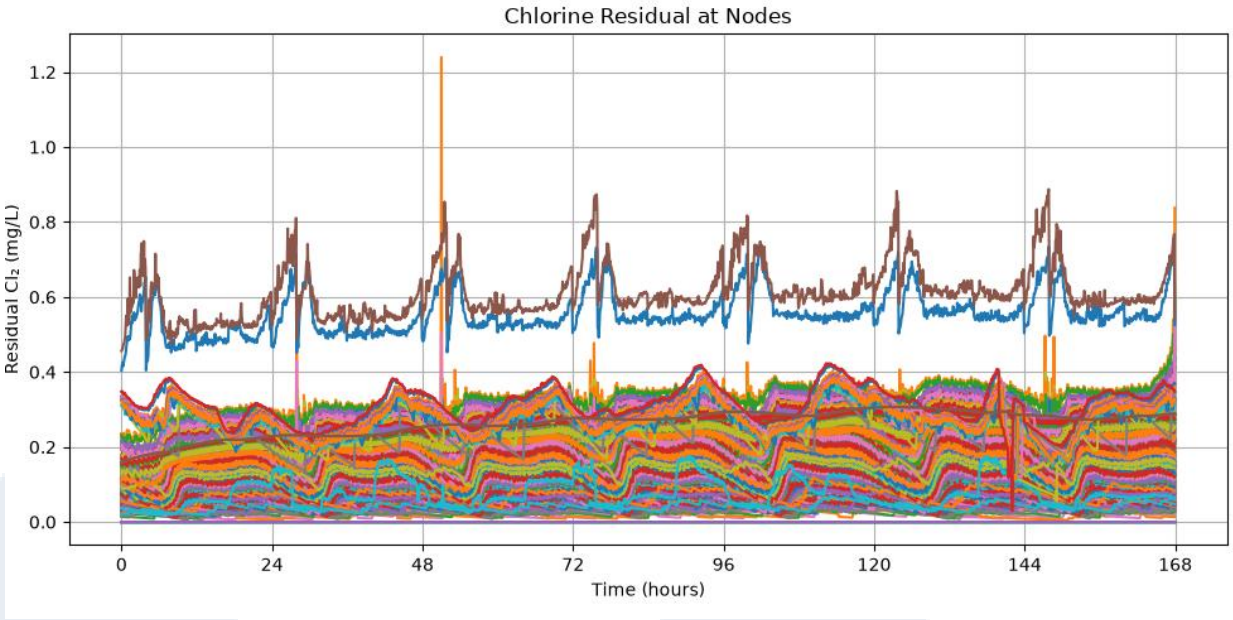


Figure 20: Chlorine residuals at all nodes following the application of the mitigation strategy

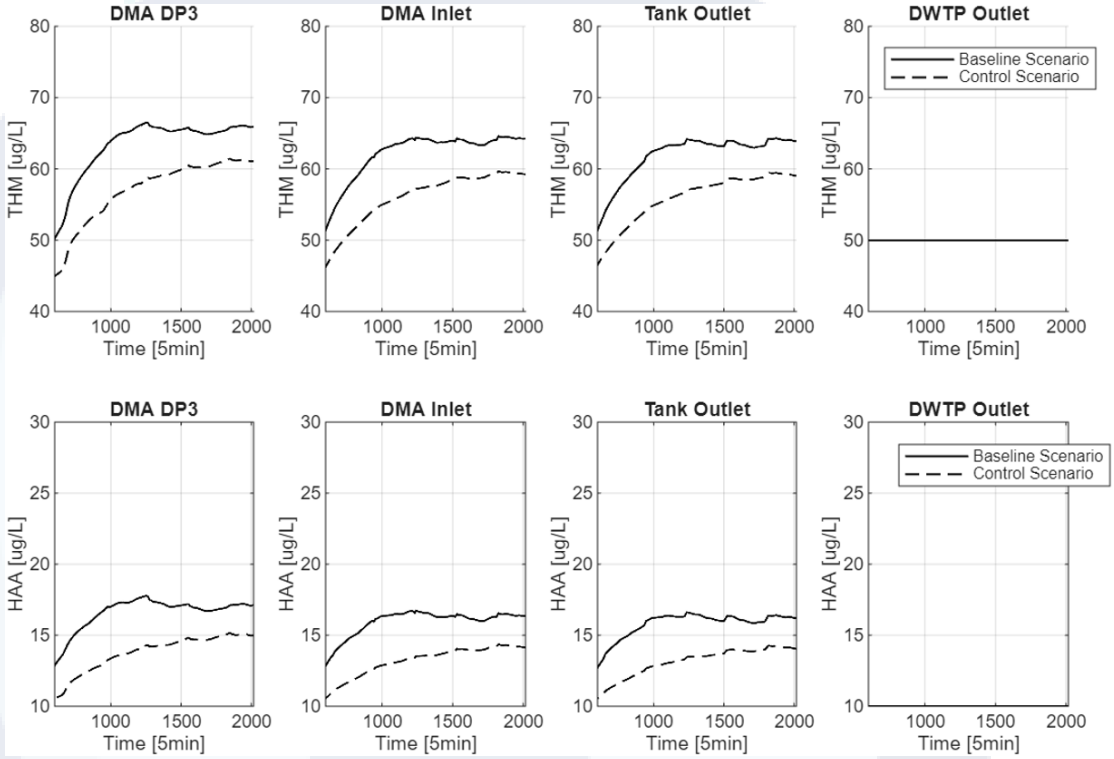


Figure 21: Simulated THM and HAA concentrations for the baseline and control scenario at the monitoring locations

6.3 Exposure Risk Assessment

Following the application of the mitigation strategies, the applicability of the exposure risk assessment framework is presented below. Uncertainty in exposure parameters was addressed using Monte Carlo simulation with 10,000 iterations, accounting for variability in ingestion rate and body weight under lifetime exposure. **Figure 22**

presents the frequency distribution of the estimated ADD of THMs. As expected, the optimized chlorination strategy resulted in only a modest reduction in THM exposure. This is because the majority of THMs originated from the DWTP outlet, while the limited hydraulic flexibility of the distribution system restricted the potential for further reduction through operational control alone.

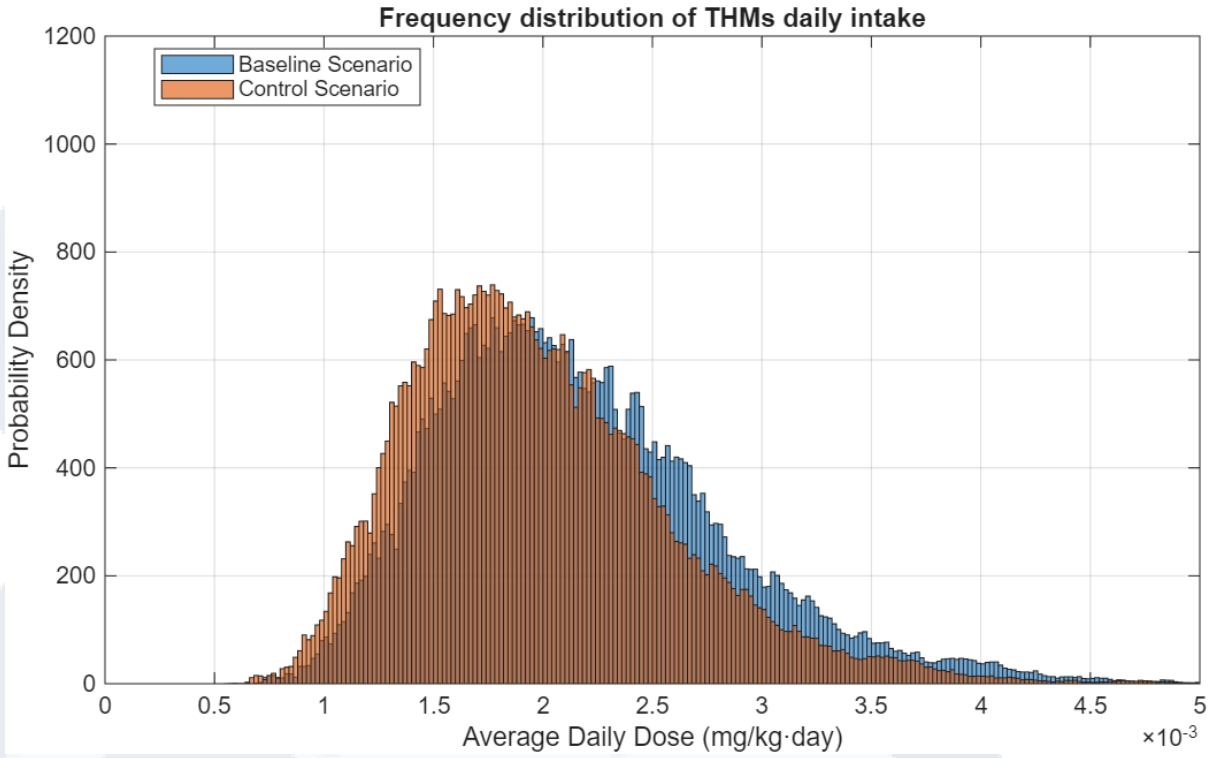


Figure 22: THM daily intake (mg kg^{-1} day^{-1}) frequency distribution under two network operational scenarios based on the real case study conditions

To better demonstrate the capability of the proposed mitigation framework, an additional hypothetical scenario was evaluated. In this scenario, a fixed THM concentration of 15 µg/L was assumed at the DWTP outlet in addition to higher NOM and chlorination levels. Under baseline operation, THM concentrations increased to approximately 45 µg/L within the downstream distribution system. The application of the proposed mitigation strategy reduced the maximum THM concentration from 45 µg/L to 25 µg/L. This reduction translated directly into a lower ADD and a corresponding decrease in the estimated risk, demonstrating the potential benefits of optimized booster chlorination and dynamic disinfectant scheduling.

Figure 23 shows the frequency distribution of daily THM intake under the baseline and control scenarios. The control scenario exhibits a lower average daily dose compared to the baseline, with most values concentrated around 0.8 × 10^{-3} mg/kg·day, while the baseline scenario shows a wider distribution extending up to ~2.5 × 10^{-3} mg/kg·day.

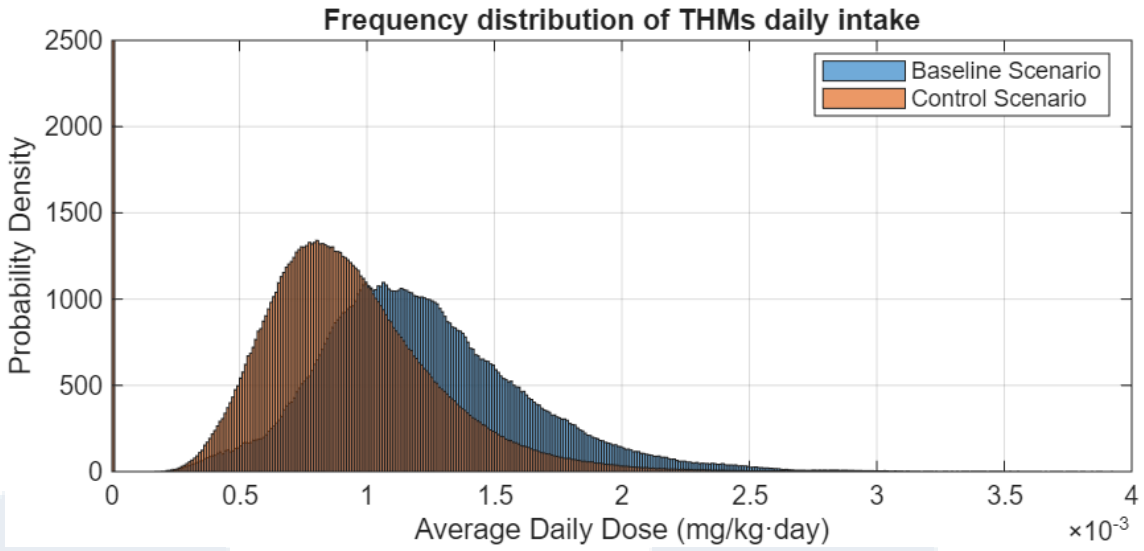


Figure 23: THM daily intake (mg kg^{-1} day^{-1}) frequency distribution under two network hypothetical operational scenarios

7.0 Early-warning system platform and Explainable AI

The developed early detection methodologies for THMs were implemented as modules within a web-based platform, which is provided to water utility operators for real-time monitoring and early warning purposes.

The platform is based on Grafana[8], an open-source data visualization and monitoring software widely used for real-time analytics, interactive dashboards, and alerting. It enables users to integrate multiple data sources, visualize time-series data, and configure automated alerts, making it particularly suitable for operational monitoring and decision support in water systems.

Web-based dashboards were initially deployed during the development and implementation of the monitoring system in the Limassol case study (Task 3.2), where four sensor stations were installed across the distribution network. The dashboards served as the primary user interface, enabling water operators to visualize incoming sensor data and monitor system performance (**Figure 24**).



Figure 24: Set of physicochemical data (Temperature, pH, Chlorine, Conductivity) as shown through the initial Grafana interface developed for data visualization

Following the development of DBP estimation methodology, a new web-based Dashboard was created and published by KIOS to act as an Early Warning Platform. This Dashboard is available to operators of the Limassol District Local Government Organization (previously named WBL) who manage the case study system.

In this configuration, the predictive models are implemented as custom computational modules that perform the following tasks:

1. Retrieve real-time sensor data from the Grafana data source.
2. Perform automated data preprocessing and quality control (data validation).
3. Compute THM predictions using the proposed models.
4. Generate alerts when predicted concentrations approach regulatory thresholds.
5. Visualize both sensor measurements and predicted THM concentrations on operational dashboards.

[8] https://grafana.com/

7.1 Real-Time THM Early Warning Module

The predictive module can be implemented using a computational service (e.g., Python or MATLAB scripts) that communicates with the software platform backend through API calls or time-series databases, such as InfluxDB. During routine operation, sensor measurements are acquired and monitored in real time at 5-min intervals, allowing continuous monitoring and alerting. The Digital Twin can update its predictions whenever a new set of measurements is received; however, the THM prediction and DCA are performed once daily for computational efficiency. This is also appropriate because the hydraulic and water-quality conditions of the network are not expected to change substantially over short periods of a few hours. Thus, real-time sensor monitoring is maintained, while the computationally intensive THM prediction and DCA are updated daily. At this stage, false-positive and false-negative rates, missed events and precision cannot be assessed, as measured THM concentrations are not yet available for comparison with the model predictions. In future applications, the Digital Twin should be complemented by daily THM sampling to enable systematic validation of the predictions and assessment of detection performance. During this project, the few THM measurements from the sampling campaign were used to calibrate the model.

The workflow for the application of the THM early warning module is presented below:

1. Acquire sensor data
    - Retrieve real-time measurements from DWTP and Z2 tank sensors
    - Parameters: UV254, DOC, TOC, fluorescence, chlorine, temperature, pH
2. Preprocess measurements
    - Remove missing values
    - Apply smoothing filter (moving average)
3. Apply model-based approach using the reaction model (hydraulic and quality modelling)
4. Evaluate prediction quality
    - Check prediction limits
    - Detect abnormal spikes
    - Compare with historical statistics
5. Trigger early warning alarm

$$\widehat{TTHM}(t) \geq TTHM_{threshold}$$

6. Store results
    - Save predicted THM values to time-series database.
7. Update models periodically
    - Compare predictions with laboratory THM measurements (if available)
    - Recalibrate model coefficients if residual errors increase.

7.2 User Interface Guide

Figure 25 and **Figure 26** illustrate the developed web-based dashboards. **Figure 25** presents the main monitoring interface, which displays the daily simulation results together with an interactive network map showing the spatial distribution of the selected parameter. **Figure 26** presents the DBP Risk Event Summary, which is

automatically generated every day. This dashboard consolidates the outputs of the Digital Twin into several dedicated modules, including predicted water quality conditions, exposure risk assessment, sensor status and impact on uncertainty of estimation, Driver Contribution Analysis, recommended mitigation actions, scenario evaluation, regulatory context and an XAI system overview.

A brief description of each dashboard component is provided below.

1. Network Information

This panel summarizes the location of the event within the WDN, including the corresponding DMA, operational zone and associated asset (e.g., storage tank). It enables operators to rapidly identify the affected area.

2. Regulatory Context

The applicable regulatory and operational limits are summarized for reference, including the maximum allowable THM concentration, minimum chlorine residual and health risk reference values. This enables rapid comparison between predicted conditions and regulatory requirements.

3. DMA Exposure Assessment

This panel summarizes the estimated exposure impact of the predicted THM concentration. It presents the calculated ADD together with the estimated population exposed within the affected area. These metrics provide an indication of the potential health implications associated with the predicted water quality conditions.

4. Simulation Conditions

The predicted hydraulic and water quality conditions responsible for the estimated THM concentration are summarized in this panel. Displayed variables include predicted THM concentration, chlorine residual, water age, water temperature, NOM representatives and source THM concentration.

5. Sensor Status

The operational status of the monitoring network is presented through sensor health indicators and data availability metrics. For each monitored parameter, the dashboard reports the percentage of successfully received measurements together with recent average values, enabling operators to assess data completeness and identify missing or unreliable measurements before interpreting the model predictions.

6. Driver Contribution Analysis

The DCA ranks the principal operational and water quality drivers according to their estimated contribution to the predicted THM concentration. For each driver, the dashboard reports the relative contribution together with the supporting evidence extracted from the Digital Twin simulations, providing operators with an interpretable explanation of the prediction.

7. Uncertainty Assessment

This section summarizes the principal sources of uncertainty affecting the prediction, including missing sensor measurements and outdated sampling data. An overall confidence indicator is also presented to assist operators in evaluating the reliability of the prediction.

8. Recommended Actions

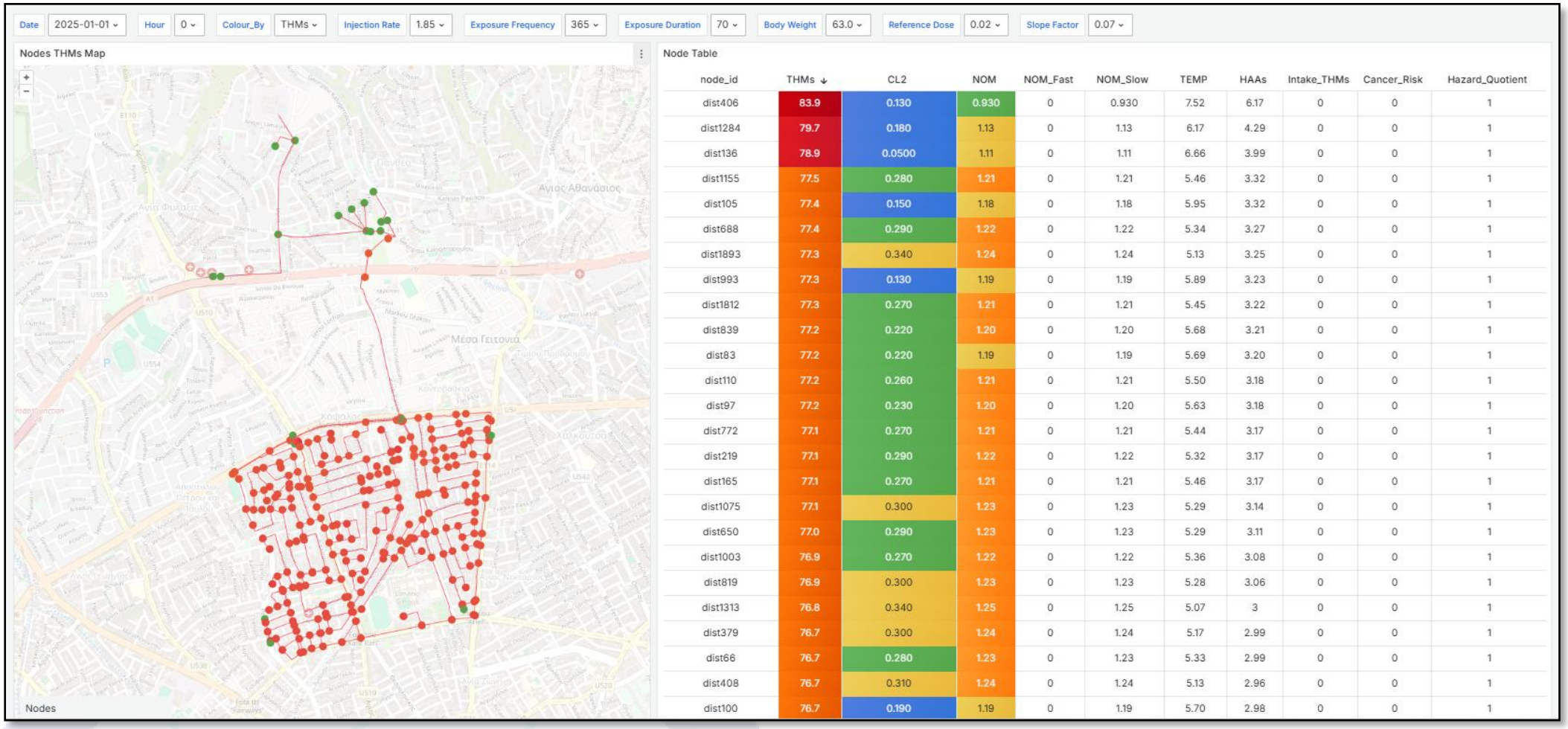
The dashboard provides operational recommendations for mitigating the DBP formation risk (adjusted per case study).

9. Scenario Evaluation

This panel compares the predicted THM concentration from the baseline and optimization scenarios, allowing operators to evaluate the potential benefits of implementing the recommended actions.

10. System Overview and Explainable AI

The lower section of the dashboard provides a simplified graphical representation of the case study, illustrating the progression of water from the treatment plant through the booster chlorination point and storage tank to the distribution network. The diagram highlights the predicted THM risk location together with the prevailing water quality conditions along the system. A dedicated ‘LLM SUMMARY’ panel summarizes the most influential drivers identified by the DCA, providing operators with a concise explanation of the event and supporting rapid interpretation of the Digital Twin predictions. Together, these components constitute the XAI layer of the decision support system by translating complex model outputs into intuitive, actionable information.



node_id	THMs ↓	CL2	NOM	NOM_Fast	NOM_Slow	TEMP	HAAs	Intake_THMs	Cancer_Risk	Hazard_Quotient
dist406	83.9	0.130	0.930	0	0.930	7.52	6.17	0	0	1
dist1284	79.7	0.180	1.13	0	1.13	6.17	4.29	0	0	1
dist136	78.9	0.0500	1.11	0	1.11	6.66	3.99	0	0	1
dist1155	77.5	0.280	1.21	0	1.21	5.46	3.32	0	0	1
dist105	77.4	0.150	1.18	0	1.18	5.95	3.32	0	0	1
dist688	77.4	0.290	1.22	0	1.22	5.34	3.27	0	0	1
dist1893	77.3	0.340	1.24	0	1.24	5.13	3.25	0	0	1
dist993	77.3	0.130	1.19	0	1.19	5.89	3.23	0	0	1
dist1812	77.3	0.270	1.21	0	1.21	5.45	3.22	0	0	1
dist839	77.2	0.220	1.20	0	1.20	5.68	3.21	0	0	1
dist83	77.2	0.220	1.19	0	1.19	5.69	3.20	0	0	1
dist110	77.2	0.260	1.21	0	1.21	5.50	3.18	0	0	1
dist97	77.2	0.230	1.20	0	1.20	5.63	3.18	0	0	1
dist772	77.1	0.270	1.21	0	1.21	5.44	3.17	0	0	1
dist219	77.1	0.290	1.22	0	1.22	5.32	3.17	0	0	1
dist165	77.1	0.270	1.21	0	1.21	5.46	3.17	0	0	1
dist1075	77.1	0.300	1.23	0	1.23	5.29	3.14	0	0	1
dist650	77.0	0.290	1.23	0	1.23	5.29	3.11	0	0	1
dist1003	76.9	0.270	1.22	0	1.22	5.36	3.08	0	0	1
dist819	76.9	0.300	1.23	0	1.23	5.28	3.06	0	0	1
dist1313	76.8	0.340	1.25	0	1.25	5.07	3	0	0	1
dist379	76.7	0.300	1.24	0	1.24	5.17	2.99	0	0	1
dist66	76.7	0.280	1.23	0	1.23	5.33	2.99	0	0	1
dist408	76.7	0.310	1.24	0	1.24	5.13	2.96	0	0	1
dist100	76.7	0.190	1.19	0	1.19	5.70	2.98	0	0	1

Figure 25: Digital Twin Monitoring Dashboard listing predicted parameters at every location of the network

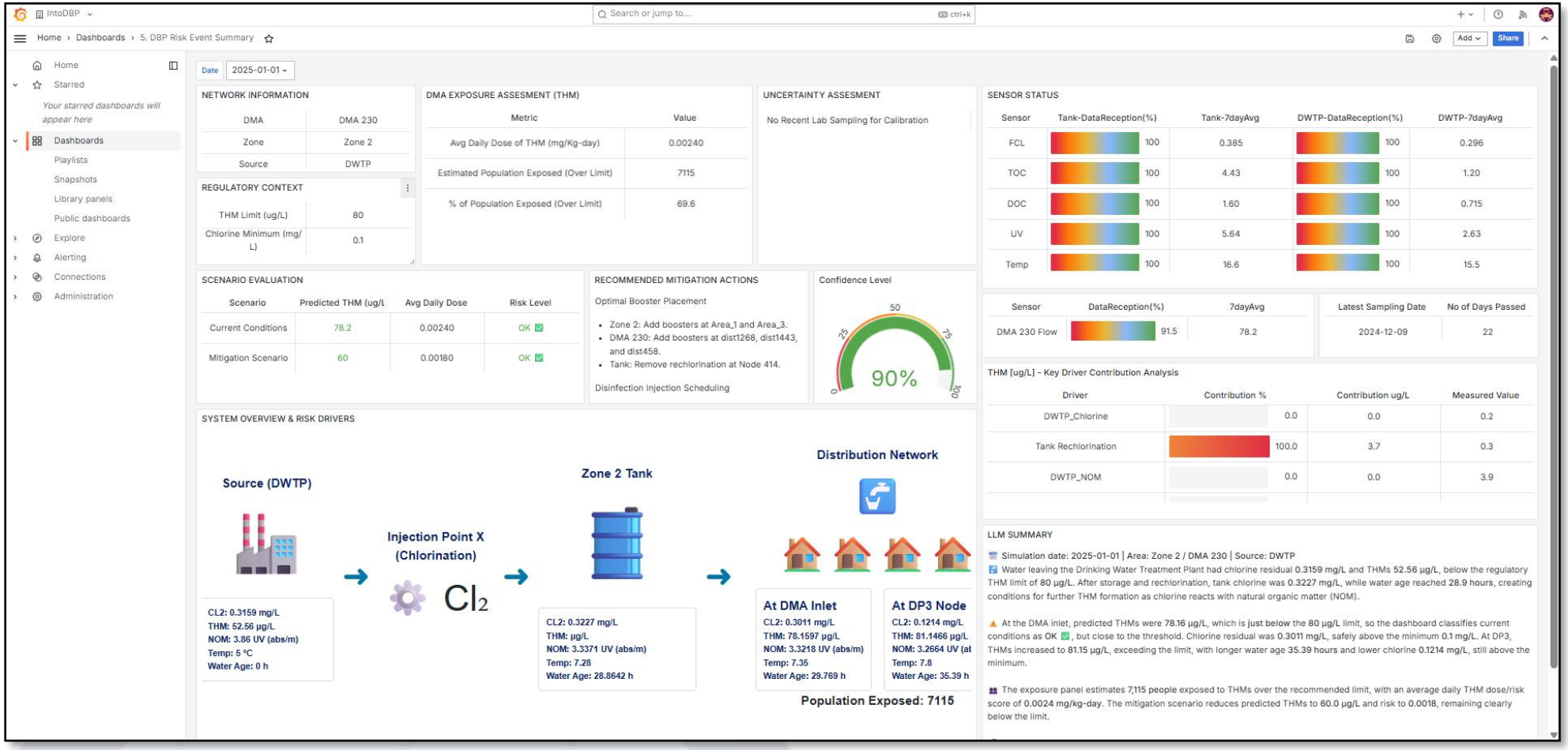


Figure 26: DBP Risk Event Summary and Early-Warning web-based platform

7.3 Explainable AI

To improve the accessibility and interpretability of the results produced by the intoDBP modelling and early-warning framework, an Explainable Artificial Intelligence (XAI) component based on a Large Language Model (LLM) was integrated into the web-based user interface.

This component complements the numerical and graphical information presented by the dashboard by providing users with a natural-language interpretation of simulation outputs. This functionality is particularly useful because the intoDBP framework combines several layers of information, including hydraulic conditions, water-quality parameters, predicted THM concentrations, exposure indicators, and operational thresholds which may be difficult to interpret collectively without specialist knowledge.

The web-based dashboards act as the operational interface through which monitoring data, predictions and warnings obtained by the developed algorithms are communicated to water utility operators.

The LLM-based component extends this functionality from conventional visualization towards an explanatory decision-support environment. Rather than requiring users to independently interpret individual dashboard panels, numerical values and model outputs, the LLM receives a structured representation of the current simulation results and is instructed to explain these results in the context of the simulated water distribution system. In particular, the system prompt establishes the role of the LLM as an expert analysing a simulation of THM formation and the potential risk to the population. It also explicitly describes the principal system topology represented by the simulation.

The information supplied dynamically to the LLM consists of three principal components:

I. The **numerical results generated by the current simulation**, thereby ensuring that the explanation refers to the conditions currently being presented to the user rather than to generic operating conditions.
II. A **glossary** which provides **definitions and contextual information** for the variables included in the simulation results, allowing the LLM to interpret technical abbreviations, parameters and indicators consistently.
III. The **parameter definitions** of the corresponding panels were extracted from the models. The inclusion of the parameter definitions is particularly important because it supplies contextual metadata describing what the displayed values represent and, where applicable, the thresholds against which they should be evaluated.

The dashboard itself presents predicted hydraulic and water-quality conditions including THM concentration, chlorine residual, water age, water temperature, NOM and source THM concentration, together with applicable operational or regulatory thresholds. It also contains public-health information, sensor-status indicators, DCA, uncertainty information, recommended actions, scenario comparisons and regulatory context. Consequently, the LLM does not receive isolated numerical values, but a structured combination of values, definitions and dashboard semantics that enables it to produce an explanation that more closely reflects what the operator is viewing.

The instruction requires the LLM to generate a concise summary intended for users who may not be familiar with the raw simulation outputs and therefore emphasizes interpretation rather than simple repetition of the numerical results. The LLM is specifically instructed to use the dashboard definitions to explain the meaning of the values and their relationship to the corresponding thresholds, and to present the information in a readable, well-structured form.

This is especially important for an early-warning application, where the objective is not merely to report a predicted concentration but to communicate whether the predicted conditions indicate normal operation, an approaching threshold, or a potentially significant event requiring further attention.

The early-warning module itself retrieves sensor information, performs preprocessing and quality control, computes THM predictions, generates alerts as concentrations approach regulatory thresholds, and communicates measurements and predictions through the operational dashboard. The LLM therefore acts as an explanatory layer above these analytical components rather than as the underlying THM prediction model. This distinction is important: quantitative predictions originate from the calibrated modelling and risk-assessment framework, while the role of the LLM is to synthesize, contextualize and communicate those outputs in language that supports operator understanding and decision-making.

An important feature of the implemented approach is the incorporation of Retrieval-Augmented Generation (RAG). In addition to the current simulation results, glossary and parameter definitions, relevant background information is retrieved from a curated collection of intoDBP project documents and supplied to the LLM as contextual knowledge for each user query. The knowledge base includes project material addressing DBP formation and modelling, monitoring-system deployment, hydraulic and water-quality modelling, Digital Twin development, early-warning methodologies, exposure risk assessment and mitigation strategies.

This is particularly valuable because DBP formation is influenced by interacting water-quality and hydraulic factors; the intoDBP modelling review identifies, among others, disinfectant conditions, pH, temperature and organic or inorganic constituents, as well as network characteristics such as velocity, retention time, water age and pipe properties.

Similarly, the monitoring framework integrates strategically positioned sensor measurements with hydraulic and water-quality models to improve understanding and prediction of DBP formation and to support proactive management. RAG enables the LLM to draw upon this project-specific knowledge when explaining why particular conditions may contribute to increased THM formation. Importantly, the retrieved material is selected according to its relevance to the question rather than inserting the complete documentation into every request. This provides the LLM with focused contextual evidence while reducing unnecessary information in the prompt and maintaining the connection between the generated explanation and the methodologies developed within intoDBP.

The resulting architecture therefore combines deterministic numerical simulation with generative AI in a complementary manner: the Digital Twin and associated analytical modules remain responsible for calculating hydraulic behaviour, water-quality

evolution, THM formation, exposure and risk indicators, while the LLM provides a human-oriented explanatory interface to these results.

This reflects the overall objective of the XAI layer described for the platform, in which project-generated information is synthesized to produce clear and context-aware explanations. The approach is therefore intended to reduce the cognitive burden associated with interpreting complex, multidisciplinary simulation outputs while preserving the underlying model as the authoritative source of quantitative predictions.

By combining current simulation outputs, explicit dashboard semantics and definitions, and dynamically retrieved project knowledge, the LLM provides a mechanism through which operators can move from observing what the Digital Twin predicts to understanding the context and significance of those predictions. In this way, the LLM-based XAI and RAG functionality transforms the web-based dashboard from a primarily visualization-oriented interface into a more comprehensive decision-support environment, facilitating communication of complex model results and supporting informed interpretation of THM formation, associated population exposure and potential mitigation actions within the intoDBP framework.

8.0 Conclusion

Overall, the work demonstrates the potential of digital twins to support the proactive management of DBP formation in water distribution networks. The developed framework integrates monitoring data, THM prediction, early warning, exposure-risk assessment, driver analysis, and mitigation optimization within a common decision-support environment. The Limassol case study served as the primary operational demonstration, providing a real-world setting for evaluating the complete framework, while the methodologies and software components developed are transferable to other networks and case studies with appropriate calibration and monitoring data.

The main outcome is an integrated set of tools and methodologies that enables water utilities to move from reactive DBP monitoring towards proactive risk management. The developed digital twin, dbpRisk 2.0, mitigation algorithms, exposure assessment methods, and operator interface provide reusable components for predicting DBP formation, understanding its drivers, assessing potential exposure risks, and evaluating mitigation strategies under different operational conditions.

The main reusable outputs of the work include:

- Benchmark network and virtual testbed for DBP scenario modelling and testing
    - *Pavlou et al. (2024). A Comprehensive Virtual Testbed for Modeling Disinfection Byproduct Formation in Water Distribution Networks.*

- Open reaction model (dbpMODEL) for simulating DBP formation in water distribution networks.
- dbpRisk 2.0 for network simulation, DBP risk analysis, and scenario assessment.
    - (dbpRisk 2.0. QGIS plugin. https://plugins.qgis.org/plugins/dbpRisk2/)

- Risk – management framework
    - *Eliades et al. (2025). Risk-Based Management of Disinfection By-Products using Drinking Water Digital Twins.*
    - *Pavlou et al. (2026). A Framework for Early Warning and Control of Disinfection By-Products in Water Distribution Networks Using Digital Twins*

9.0 dbpRisk 2.0 - Software tool for simulating DBP formation in Water Distribution Networks

Accompanying this Deliverable is a separate User Manual that provides a detailed guide to dbpRisk 2.0 including the new features presented in this study. The manual offers clear and concise instructions for installing the application and utilizing its full range of functionalities. It presents step-by-step guidance on installing the software and navigating the interface. Users are guided through the configuration of the mechanistic model, the use of the scenario builder, and the execution of simulations. The manual also explains how to interpret and explore simulation results and concludes with a practical use case. It is designed to assist both new and experienced users in applying dbpRisk 2.0 effectively for decision-making and analysis.

intoDBP will create innovative tools and strategies to improve water quality management for safe human use and a healthy environment. It focuses on catchment protection and forecasting, transformative drinking water treatment, and real-time monitoring to combat the effects of climate and global change. In particular, intoDBP focuses on pollution and risks related to disinfection byproducts (DBPs). By developing and applying advanced, integrated, and cost-effective sensors and analytical methods, intoDBP will expand knowledge on water quality and DBP precursors to better understand its formation and human exposure in Europe. intoDBP monitoring results will feed into numerical forecasting tools to predict source water changes and formulate climate change adaptation pathways at catchment and treatment scale. intoDBP also develops transformative options for advanced and cost-effective upgrade of water treatment and disinfection.

In the intoDBP consortium researchers, small and large enterprises, communication experts and public services join forces to generate interdisciplinary solutions, that will generate a renewed perspective of drinking water surveillance, support decision-making and governance, and increase system resilience. intoDBP will implement and validate its cross-cutting products in four complementary case studies from three European countries where compliance with DBP regulation currently is an acknowledged challenge. The direct and visible positive impact of intoDBP in the case studies will foster rapid product adoption at a European and global scale, thus strengthening Europe's position and role in the global water market.

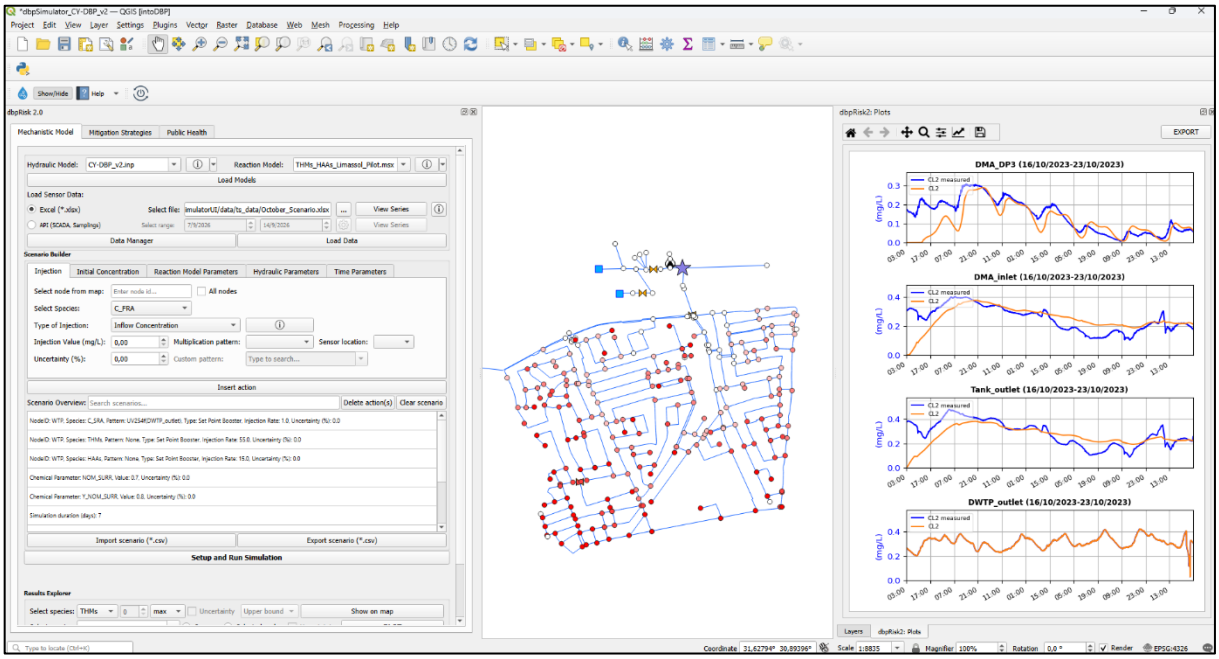
Reaching out beyond the water sector itself, intoDBP will directly engage society through surveys to analyze exposure to DBPs, collect data about catchment protection initiatives, create awareness and promote sustainable consumer behavior such as reducing bottled water consumption.



intoDBP

Funded by
the European Union

dbpRisk 2.0 User Manual



Last Update:	29/09/2026
Version:	0.4

KIOS Research & Innovation Center of Excellence

Nomenclature 7
1 Introduction 8
2 Installation Instructions 10
2.1 Operating system requirements 11
3 User interface 12
3.1 Mechanistic model 15
3.1.1 Select hydraulic model 15
3.1.2 Select reaction model 17
3.1.3 Load GIS on map 18
3.1.4 Data loading 18
3.1.5 Data manager 20
3.1.6 Calibrate and validate models 20
3.2 Scenario builder (Mechanistic Model) 20
3.2.1 Injection 21
3.2.2 Initial concentration 22
3.2.3 Reaction model parameters 23
3.2.4 Hydraulic parameters 23
3.2.5 Time parameters 23
3.2.6 Insert action into the scenario simulator 24
3.3 Scenario simulator (Mechanistic Model) 26
3.4 Results explorer 27
3.5 Mitigation Strategies 32
3.6 DBP Exposure 33
4 Use case 35
4.1 Baseline scenario: 36
4.2 Optimization scenario 45

4.3 Scenarios with uncertainties: 53

List of figures

Figure 1: Installing the dbpRisk 2.0 plugin in QGIS 10
Figure 2: Installation of dbpRisk 2.0 in QGIS using a .ZIP file 10
Figure 3: Loading the installed dbpRisk 2.0 plugin in QGIS 11
Figure 4: User interface 12
Figure 5: Menu bar 12
Figure 6: Toolbar 13
Figure 7: Main interface 14
Figure 8: Mechanistic model configuration interface 15
Figure 9: Select hydraulic model 15
Figure 10: Hydraulic model Information, EPANET, Notepad buttons 16
Figure 11: Hydraulic model information 16
Figure 12: View network in EPANET 16
Figure 13: Select reaction model 17
Figure 14: Reaction model Information and Notepad buttons 17
Figure 15: Selected MSX file 17
Figure 16: Load models 18
Figure 17: Map view displaying the loaded GIS data 18
Figure 18: Data loading interface 19
Figure 19: Sensor data loading and visualization interface 19
Figure 20: Data manager button 20
Figure 21: Model calibration button 20
Figure 22: Scenario builder interface 21
Figure 23: Injection parameters 21
Figure 24: Injection types 22
Figure 25: Initial concentration parameters 22
Figure 26: Reaction model parameters 23
Figure 27: Hydraulic parameters 23
Figure 28: Time parameters 24
Figure 29: Insert action into scenario simulator 24
Figure 30: Scenario simulator interface 25
Figure 31: Delete action 25
Figure 32: Clear scenario 26
Figure 33: Import scenario button 26
Figure 34: Export scenario button 26

Figure 35: Setup and Run Simulation button 26
Figure 36: Results explorer interface 27
Figure 37: Select species 27
Figure 38: Choose hour option 28
Figure 39: Show on map button 28
Figure 40: Simulated values on map 28
Figure 41: Show minimum values 29
Figure 42: Select species 29
Figure 43: Plot species at sensor locations 29
Figure 44: Plots of species at sensor locations 30
Figure 45: Plot species at selected nodes 30
Figure 46: Plots of species at selected nodes 31
Figure 47: Inputs required for the optimization algorithm to control DBP formation 33
Figure 48: Inputs required for the DBP exposure assessment 34
Figure 49: Select hydraulic model 35
Figure 50: Select reaction model 35
Figure 51: Load models 35
Figure 52: Map view displaying the loaded GIS data 36
Figure 53: Load sensor data 36
Figure 54: Simulated maximum chlorine concentration values plotted on the map along with the corresponding time series at sensor locations compared to the field sensor data (baseline scenario) 38
Figure 55: Simulated maximum chlorine concentration values plotted on the map with the corresponding legend (baseline scenario) 39
Figure 56: Simulated maximum NOM (UV) concentration values plotted on the map with the corresponding legend (baseline scenario) 39
Figure 57: Simulated maximum water age values plotted on the map with the corresponding legend (baseline scenario) 40
Figure 58: Simulated maximum temperature concentration values plotted on the map with the corresponding legend (baseline scenario) 41
Figure 59: Simulated maximum THM concentration values plotted on the map with the corresponding legend (baseline scenario) 42
Figure 60: Chlorine and THM concentration time series for a specific node within the distribution zone (baseline scenario) 43
Figure 61: DBP Exposure dashboard with nodes above the defined THM threshold level highlighted 44

Figure 62: DBP Exposure dashboard indicating the THM Average Daily Dose (mg kg-1 day-1) frequency distribution under (baseline scenario)..... 44
Figure 63: DBP formation control strategies with optimal booster placement across Zone 2 (left) and DMA 230 (right) 46
Figure 64: Diurnal dosing pattern for Zone 2 chlorine boosters..... 46
Figure 65: Diurnal dosing pattern for DMA 230 chlorine boosters 47
Figure 66: Chlorine residuals at all nodes following the application of the mitigation strategy 47
Figure 67: Simulated maximum chlorine concentration values plotted on the map with the corresponding legend (control scenario) 50
Figure 68: Simulated maximum THM concentration values plotted on the map with the corresponding legend (control scenario) 50
Figure 69: Chlorine and THM concentration time series for specific nodes within the distribution zone (control scenario)..... 51
Figure 70: Chlorine and THM concentration time series for a specific node within the distribution zone (control scenario)..... 51
Figure 71: DBP Exposure dashboard indicating the THM Average Daily Dose (mg kg-1 day-1) frequency distribution under (baseline and control scenario)..... 52
Figure 72: THMs concentration timeseries with uncertainty bounds concerning the THM Yield coefficient..... 53

List of Tables

Table 1: Baseline scenario disinfection strategy..... 37
Table 2: Control scenario disinfection strategy 48

Nomenclature

QGIS: Quantum Geographic Information System

EPANET: A software application for modeling water distribution piping systems, developed by the United States Environmental Protection Agency (US EPA)

MSX: EPANET-MSX (Multi-Species eXtension) is a specialized extension to the EPANET hydraulic modeling engine that enables simulation of multiple interacting chemicals within water distribution systems.

DMA: District Metered Area

UI: User Interface

EPyT: EPANET-Python Toolkit, a Python package for simulating water networks using EPANET

GIS: Geographical Information System

INP File: Input file format used by EPANET for hydraulic model simulations

KIOS CoE: KIOS Research & Innovation Center of Excellence

1 Introduction

This manual provides a comprehensive guide to the *dbpRisk 2.0* software application, offering users clear instructions for accessing and utilizing the tool's functionalities.

The document begins with an overview of the IntoDBP project, followed by step-by-step guidance on how to install the application and navigate the user interface. Users are guided through configuring the mechanistic model and using scenario builder and simulator. The document also explains how to explore simulation results. It concludes with a practical use case. This manual is designed to support both new and experienced users in effectively applying *dbpRisk 2.0* for decision-making and analysis.

IntoDBP is an EU-funded project that will develop, test, scale-up, validate, and benchmark innovative tools and strategies to improve water quality management for safe human use and a healthy environment. It focuses on catchment protection and forecasting, transformative drinking water treatment, and real-time monitoring to combat the effects of climate and global change.

In particular, the project aims to develop a comprehensive approach from source to tap for an optimum drinking water surveillance strategy, to foster AI sensor deployment methodologies and algorithms in water distribution networks (WDN) and create a new open and ready-to-use workflow using a combination of dynamic and statistical models. In addition, it aims to increase the understanding of human exposure, to generate and implement models and recommendations to impact policy and decision-making in Europe, as well as to provide guidance to decision makers to formulate optimized and future-proofed climate change adaptation pathways to successfully tackle emerging water quality threats.

The project will develop its cross-cutting solutions on 4 complementary case studies combining rural and dense urban areas, from 3 European countries (Spain, Cyprus, and Ireland), where disinfection by-products (DBPs) are a scientific, technological, and political challenge.

The consortium consists of 16 partners from Europe, including researchers, small and large enterprises, communication experts, and public services. On behalf of the University of Cyprus, NIREAS International Water Research Center and KIOS Center of Excellence are participating as project partners.

The KIOS research team contributes to the project's outcomes by implementing *dbpRisk 2.0* application, a mathematical and modelling tool for the formation of DBPs, and for investigating strategies for optimal control and minimization of their formation risk. This tool represents a significant advancement in the field of water quality management for safe human use.

To make effective use of the *dbpRisk 2.0* application, users are guided through a series of structured steps. These are detailed in the following sections of this manual and include:

1. Installation: Set up the application and its dependencies.
2. Selection of the hydraulic model: Choose a hydraulic model from the available options.
3. Selection of the water quality/chemical reaction model: Define the water quality model based on MSX configuration.
4. Mapping through GIS maps: Visualize the network on the map.
5. Data fetching from Excel Files: Import sensor and sampling data locally.
6. Data fetching from API: Retrieve data using credentials from remote sources.
7. Data manager: Manage and organize input files via local or remote access.
8. Calibrate and validate models: Use real measurements to refine model accuracy.
9. Application of What-if Scenarios through a "Scenario builder" app: Define actions such as injections, initial concentrations, and parameter variations.
10. Scenario simulator: Run the scenario to simulate system response under selected conditions.
11. Model outputs through a "Results explorer" app: Visualize outputs through maps and plots for informed decision-making.

2 Installation Instructions

dbpRisk 2.0 is a QGIS plugin developed to simulate the formation of DBPs in WDNs. QGIS is a free and open-source GIS platform used worldwide for creating, editing, and analyzing spatial data. It supports a wide variety of formats and offers a flexible plugin architecture, making it ideal for developing specialized tools like *dbpRisk 2.0*. By building *dbpRisk 2.0* as a QGIS plugin, we enable users to integrate risk models with geospatial data seamlessly, perform advanced spatial analyses, and produce clear visual outputs - all within a single interface. To install *dbpRisk 2.0*, users first need to **install QGIS from its official website** here.

After launching QGIS, the plugin can be added via the **Plugins > Manage and Install Plugins** menu (**Figure 1**).

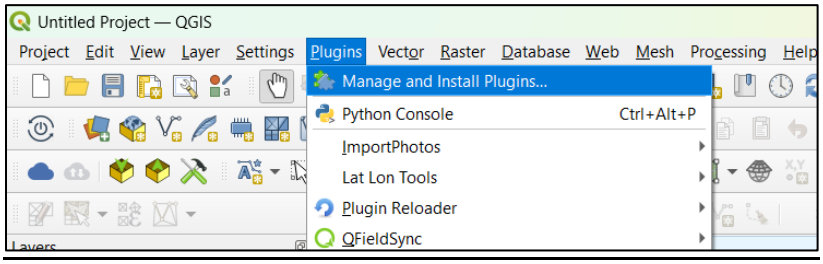


Figure 1: Installing the dbpRisk 2.0 plugin in QGIS

In the plugin manager window, search for *dbpRisk 2.0*, then select it from the list and click **Install Plugin**, as shown in **Figure 2**. Note that, from Plugins, you can install the open-source version of the plugin.

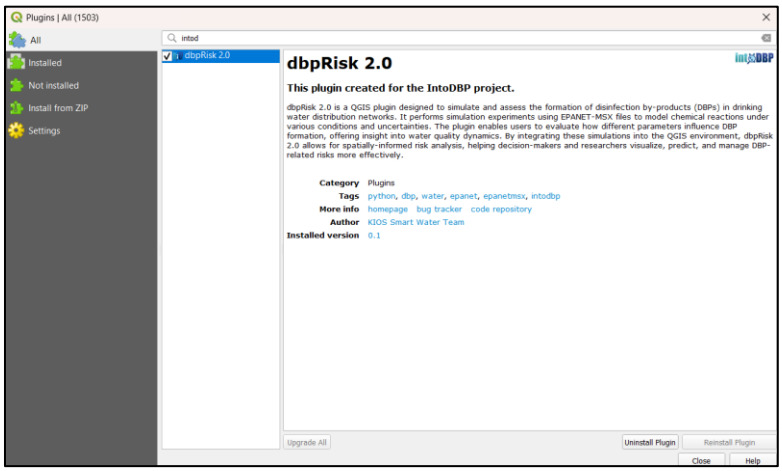


Figure 2: Installation of dbpRisk 2.0 in QGIS using a .ZIP file

Once installed, the plugin will appear in the QGIS interface and can be accessed from the corresponding menu (Plugins) or from a toolbar in the interface (**Figure 3**). Users can move the toolbars to customize the QGIS user interface as desired.

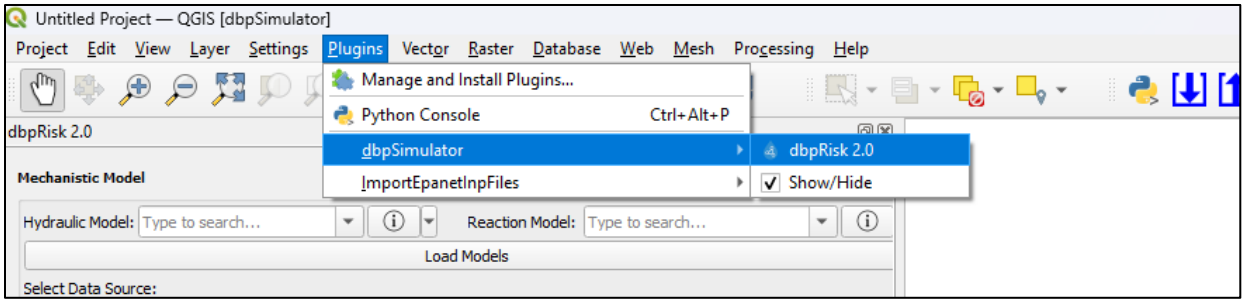


Figure 3: Loading the installed dbpRisk 2.0 plugin in QGIS

2.1 Operating system requirements

Below you can find the operating system requirements for running QGIS and the plugin:

Minimum Requirements

- Operating System: Windows 10 (64-bit), macOS 10.14+, or a modern Linux distribution
- Processor: Dual-core CPU (e.g., Intel Core i3 or AMD equivalent)
- RAM: 8 GB
- Storage: 10 GB available disk space for installation (additional space needed for data)
- Graphics: OpenGL 3.2 compatible graphics card
- Screen Resolution: 1366x768

Recommended Requirements

- Operating System: Windows 11 (64-bit), macOS Monterey or later, Ubuntu 22.04+
- Processor: Quad-core CPU (e.g., Intel Core i5/i7 or AMD Ryzen 5/7)
- RAM: 16 GB or more (especially for large datasets or complex analysis)
- Storage: SSD with at least 10 GB free space
- Graphics: Dedicated GPU with at least 2 GB VRAM (NVIDIA/AMD)
- Screen Resolution: 1920x1080 or higher

3 User interface

The user interface of the *dbpRisk 2.0* application is presented in **Figure 4**. This interface represents a fully integrated simulation environment for modeling and analyzing water quality in a WDN, using QGIS.

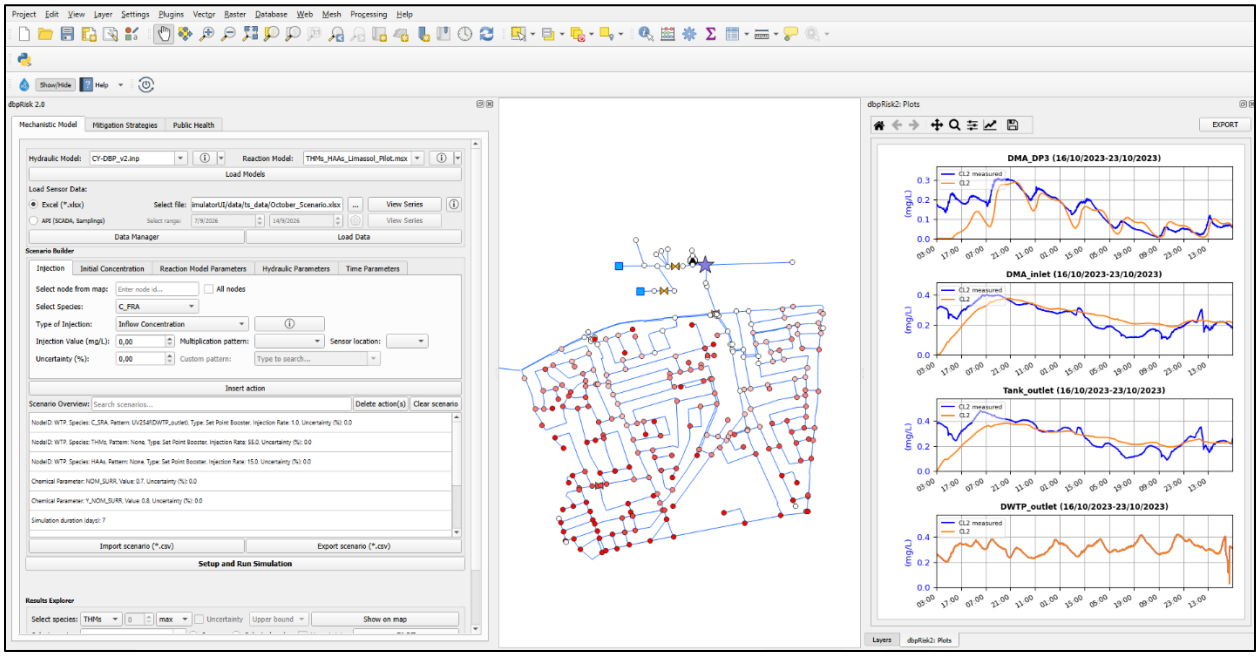


Figure 4: User interface

The menu bar (see **Figure 5**) contains standard QGIS options such as: Project, Edit, View, Layer, Settings, Plugins, Vector, Raster, Database, Mesh, Processing, and Help. These menus provide access to essential GIS functions including map management, layer operations, plugin integration, and data processing.

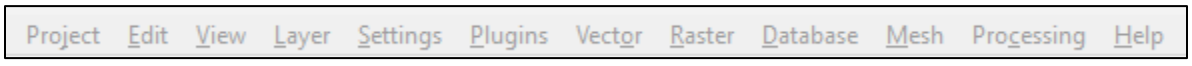


Figure 5: Menu bar

The toolbar (**Figure 6**) provides quick access to common tools and actions, such as navigation tools (zoom, pan, select), layer controls, attribute table viewer, map saving/exporting tools, plugin controls including access to the custom *dbpRisk 2.0*.



Figure 6: Toolbar

The custom *dbpRisk 2.0* plugin is organized into the following sections (**Figure 7**):

- Mechanistic model
  - Scenario builder
  - Setup and Run Simulation
  - Explorer
- Mitigation Strategies
- DBP Exposure

The map view displays the WDN overlaid on a geographic map, with nodes and links clearly visualized. The plot panel presents time series plots of simulated values for selected species at sensor locations and selected nodes.

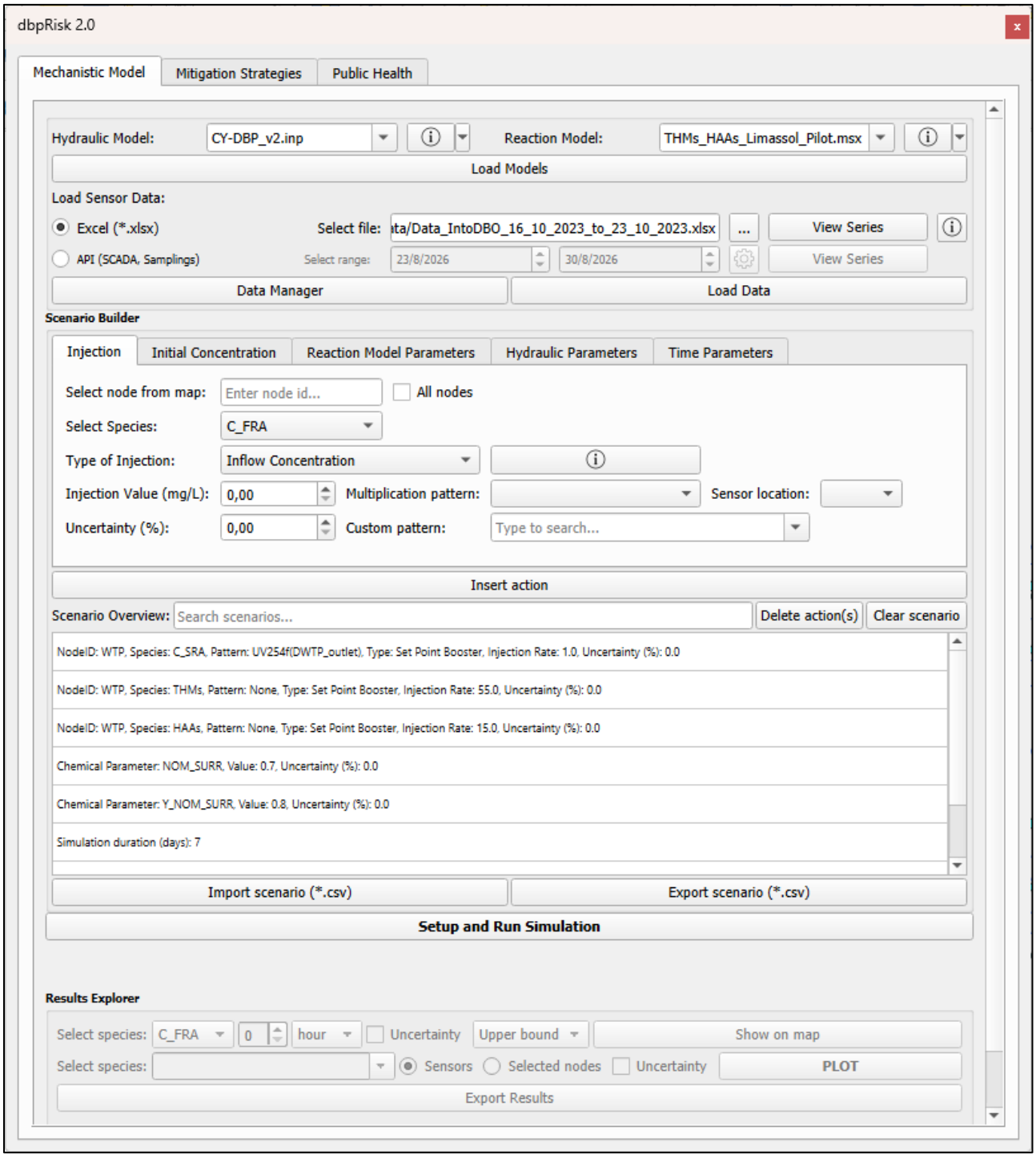


Figure 7: Main interface

3.1 Mechanistic model

The mechanistic model configuration interface, shown in **Figure 8** includes:

- Selection of hydraulic and reaction models
- Button to load models onto the map
- Option to choose data source (excel or API)
- File browser to select excel data file
- Data manager access
- Model calibration button (currently disabled)

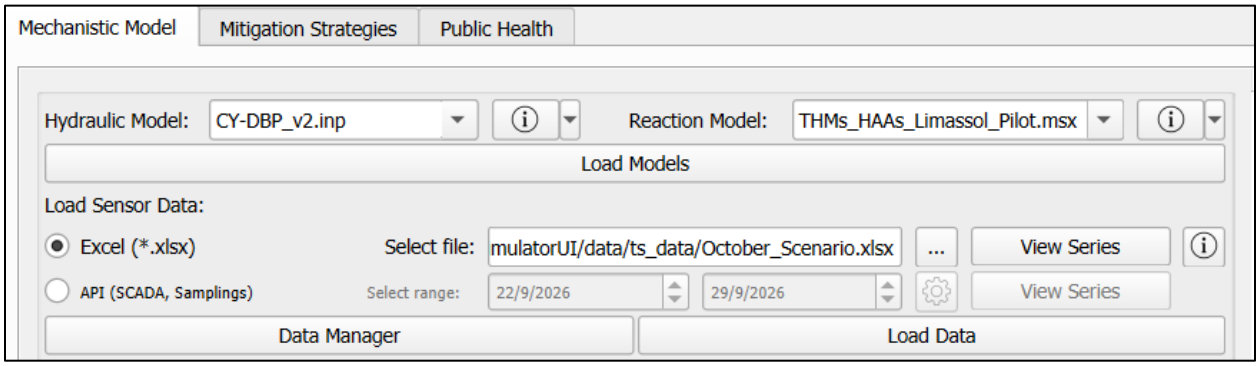


Figure 8: Mechanistic model configuration interface

3.1.1 Select hydraulic model

Initially, the user can select a hydraulic model (EPANET file) from a dropdown list or by selecting the option **IMPORT FILE**, you can import your own EPANET input file (**Figure 9**).

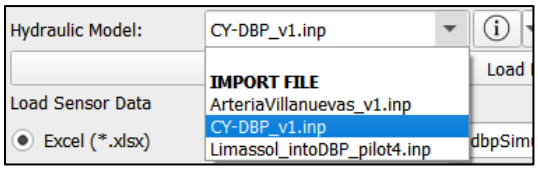


Figure 9: Select hydraulic model

By clicking the corresponding information button (**Figure 10**), detailed information about the selected model can be accessed, as shown in **Figure 11**. Additionally, the user can open the model on the EPANET GUI (**Figure 12**) or in a notepad.

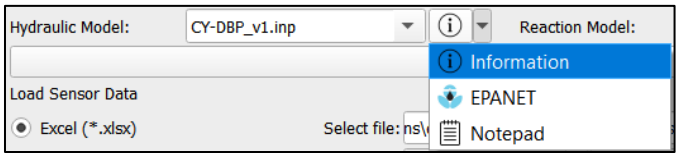


Figure 10: Hydraulic model Information, EPANET, Notepad buttons

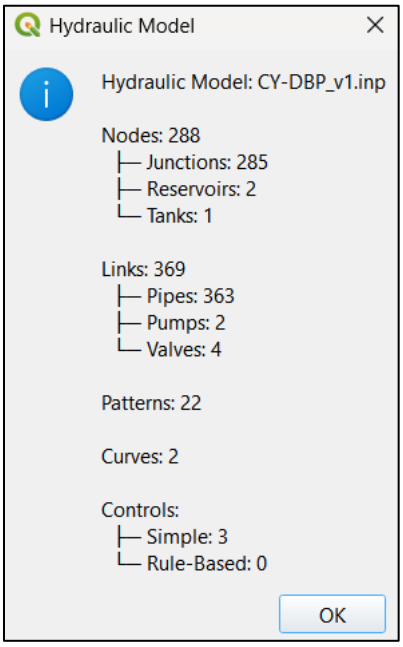


Figure 11: Hydraulic model information

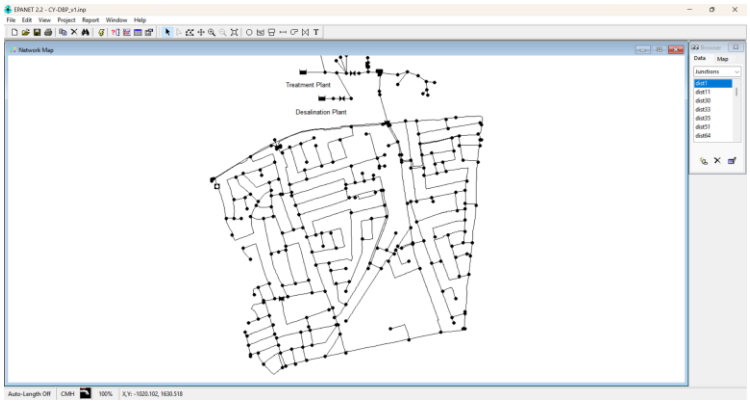


Figure 12: View network in EPANET

3.1.2 Select reaction model

After that, the user can select a reaction model (MSX file) from a dropdown list (**Figure 13**).



Figure 13: Select reaction model

By clicking the corresponding information button (**Figure 14**), the selected MSX file is opened, allowing the user to view the detailed configuration of the reaction model, as shown in **Figure 15**. Users can also prepare their own MSX files by following the same structure as the provided using the option **IMPORT FILE**.

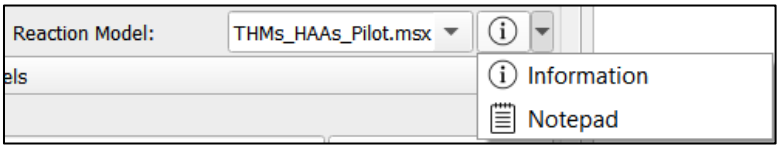


Figure 14: Reaction model Information and Notepad buttons

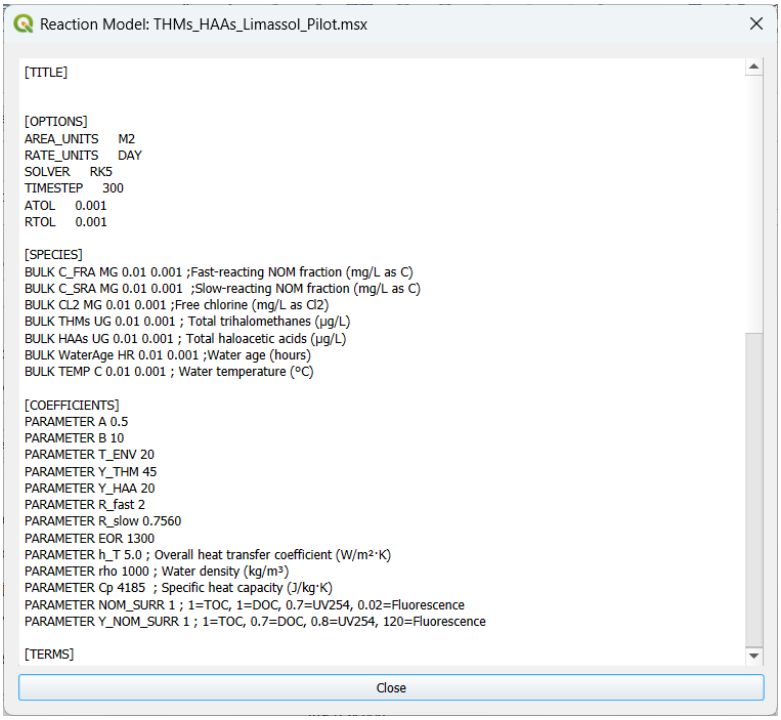


Figure 15: Selected MSX file

3.1.3 Load GIS on map

After selecting the hydraulic and reaction models, the user can load the associated GIS data onto the map by clicking the 'Load Models' button (**Figure 16**).



Figure 16: Load models

The corresponding network elements then appear on the map for visualization and interaction, as shown in **Figure 17**.

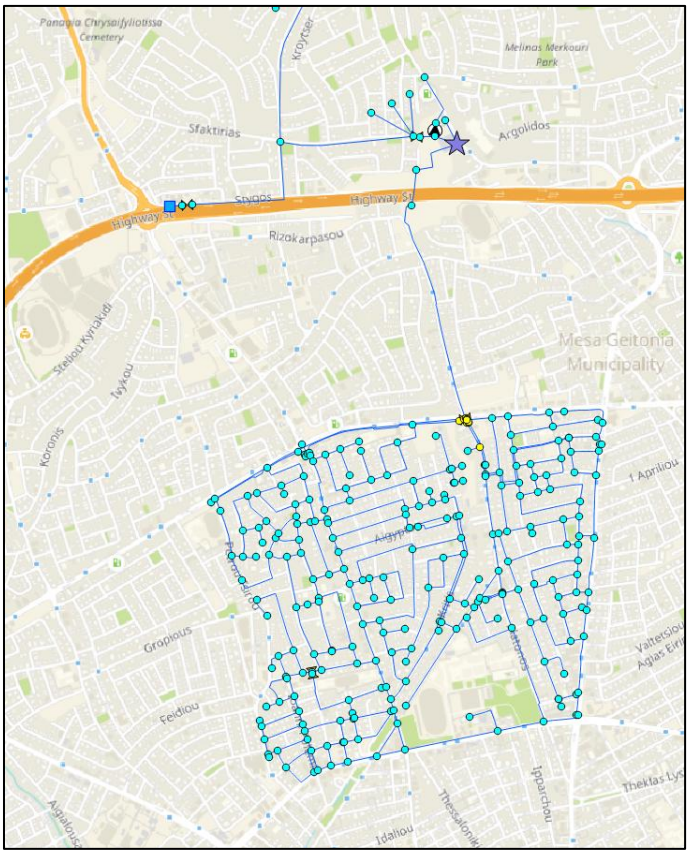


Figure 17: Map view displaying the loaded GIS data

3.1.4 Data loading

The user can either select an Excel file containing sensor data for further processing or retrieve data directly from the API by providing valid user credentials (username and password). When

the API option is selected, the user can also specify the desired period for data retrieval. This functionality is now active (**Figure 18**).

Furthermore, using the 'View Series' option, the user can visualize the loaded data as time series. A pop-up window displays the data and allows the user to filter it by parameter and location. One of the available parameters is the flow through the sensor, which is displayed on the secondary y-axis. This information is important for assessing the validity of the sensor measurements. A flow value of zero indicates that no water is passing through the sensor and, consequently, that the sensor is not operating correctly and the corresponding measurements may be unreliable. As shown in **Figure 19**, the low chlorine values observed on the last day of the dataset are faulty measurements resulting from the absence of water flow through the sensor.

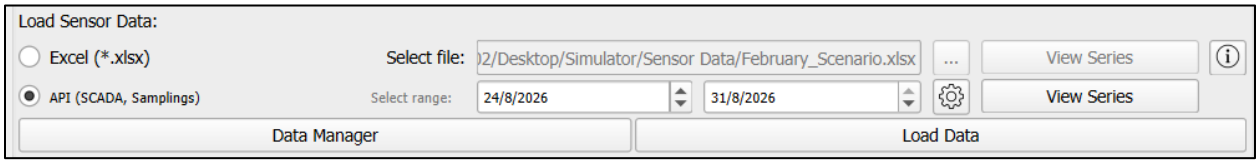


Figure 18: Data loading interface

.

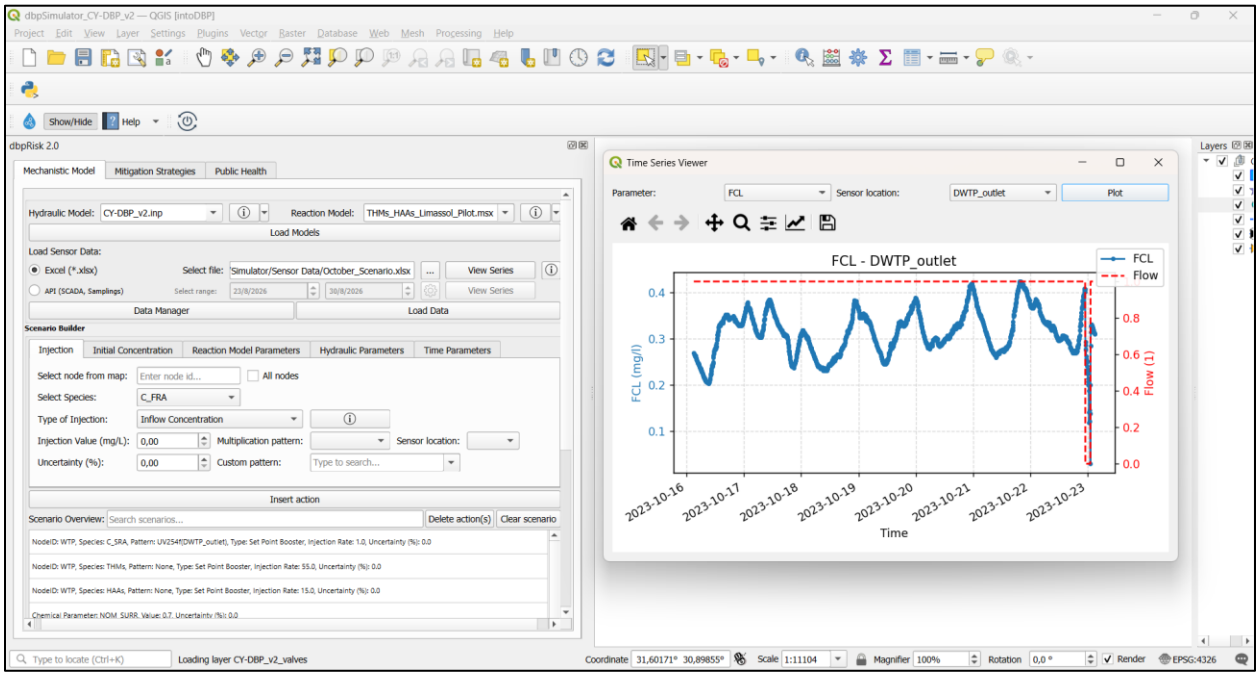


Figure 19: Sensor data loading and visualization interface

3.1.5 Data manager

By clicking the 'Data Manager' button (**Figure 20**) the user can search for specific data files, either by selecting excel files locally or - once available - by providing a URL (e.g., from Zenodo) for remote data access. The latter option is currently not available but is planned for implementation in a future version.

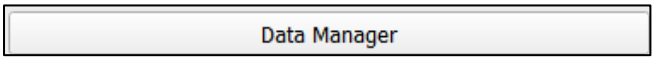


Figure 20: Data manager button

3.1.6 Calibrate and validate models

The hydraulic and water quality models can be calibrated and validated using real-world measurements obtained from field sensors and sampling results, by clicking the 'Calibrate Models' button (**Figure 21**). These data can be imported via excel files or retrieved through APIs for calibrating models. This option is currently not available but is planned for implementation in a future version.

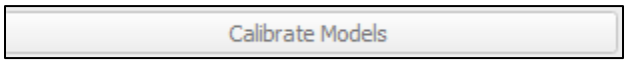


Figure 21: Model calibration button

3.2 Scenario builder (Mechanistic Model)

Figure 22 shows the scenario builder interface, which supports the following actions:

1. Injection
2. Initial Concentration
3. Reaction Model Parameters
4. Hydraulic Parameters
5. Time Parameters

The developed software enables users to simulate what-if scenarios including various conditions, uncertainties and considering climate change to assist in the decision-making process.

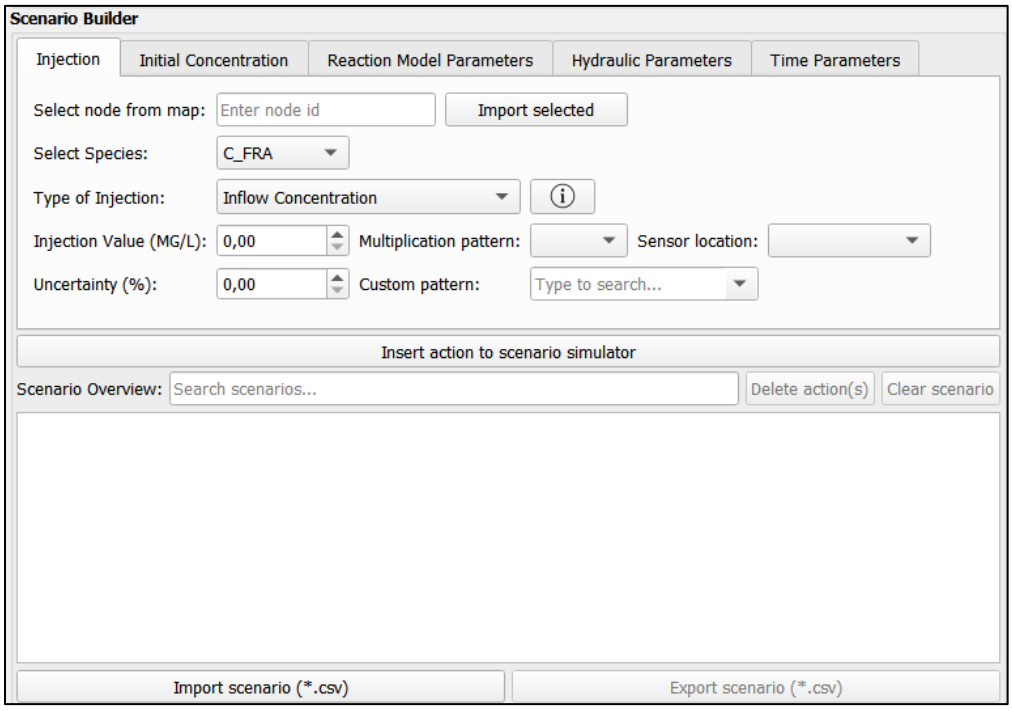


Figure 22: Scenario builder interface

3.2.1 Injection

Injection action includes a node ID, available species (based on the MSX file), the type of injection, the injection value and its uncertainty, as shown in **Figure 23.**

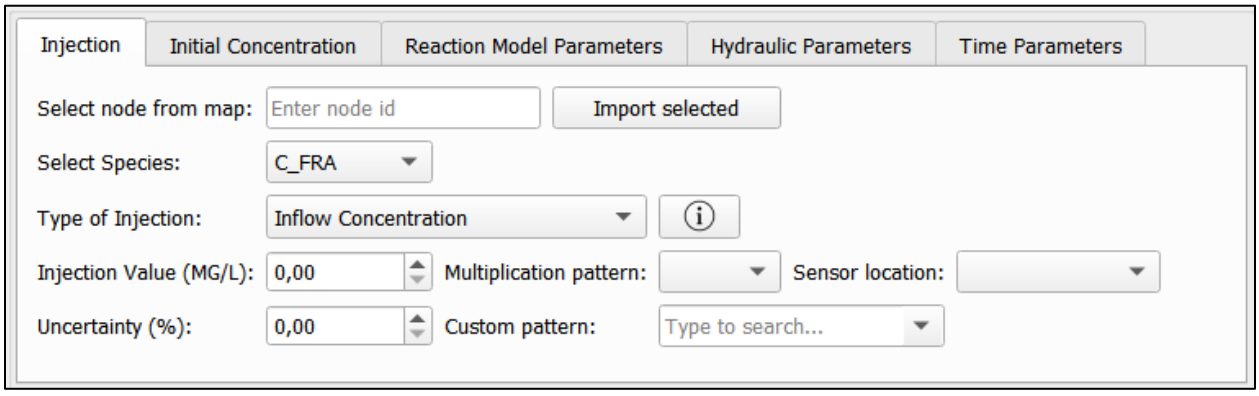


Figure 23: Injection parameters

The user can select a node directly from the map and by clicking 'Import selected' button the corresponding node ID is captured and defined for use in scenario builder.

Supported injection types include inflow concentration, mass inflow booster, set point booster and flow paced booster (**Figure 24**). More details for each injection type can be found in the EPANET manual.

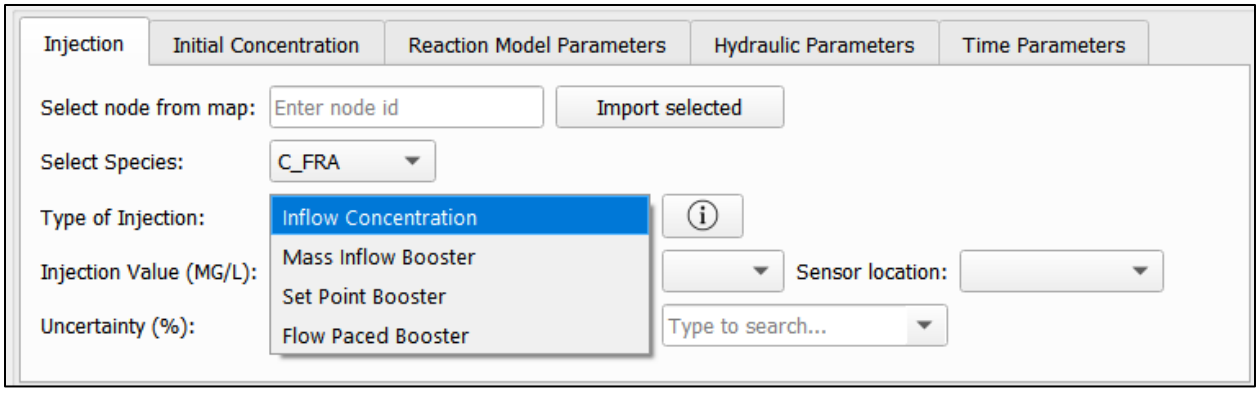


Figure 24: Injection types

3.2.2 Initial concentration

Initial concentration action includes the node ID, available species (based on the MSX file) and the initial concentration value and its uncertainty, as shown in **Figure 25**.

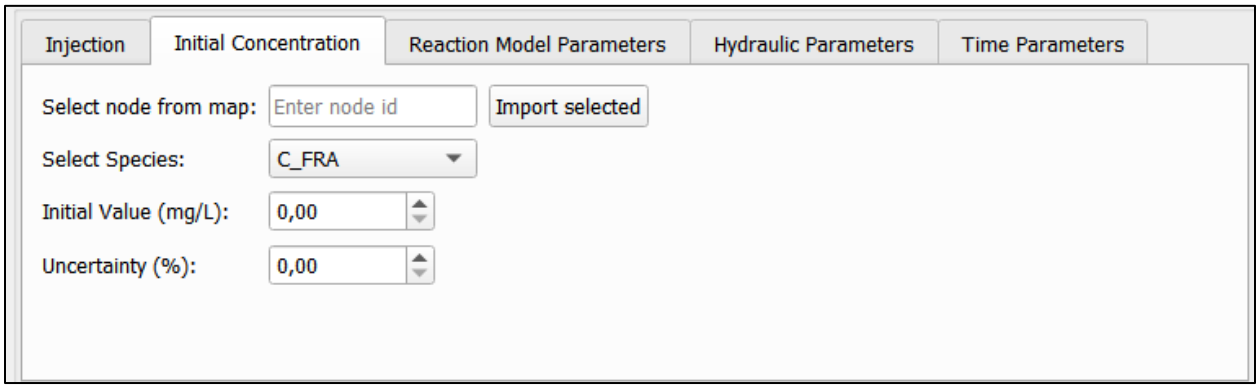


Figure 25: Initial concentration parameters

The user can select a node directly from the map and by clicking 'Import selected' button the corresponding node ID is captured and defined for use in scenario builder.

3.2.3 Reaction model parameters

Reaction model parameters' actions include parameters based on the MSX file, their values and uncertainty, as shown in **Figure 26**.

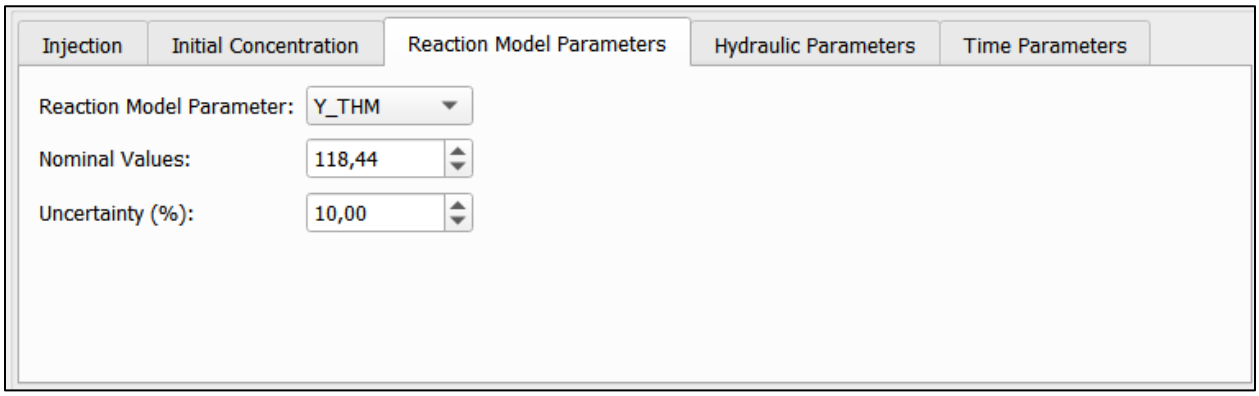


Figure 26: Reaction model parameters

3.2.4 Hydraulic parameters

Hydraulic parameters scenarios include demand uncertainty, as shown in **Figure 27**. A number of Monte Carlo simulations can be also defined by the user.

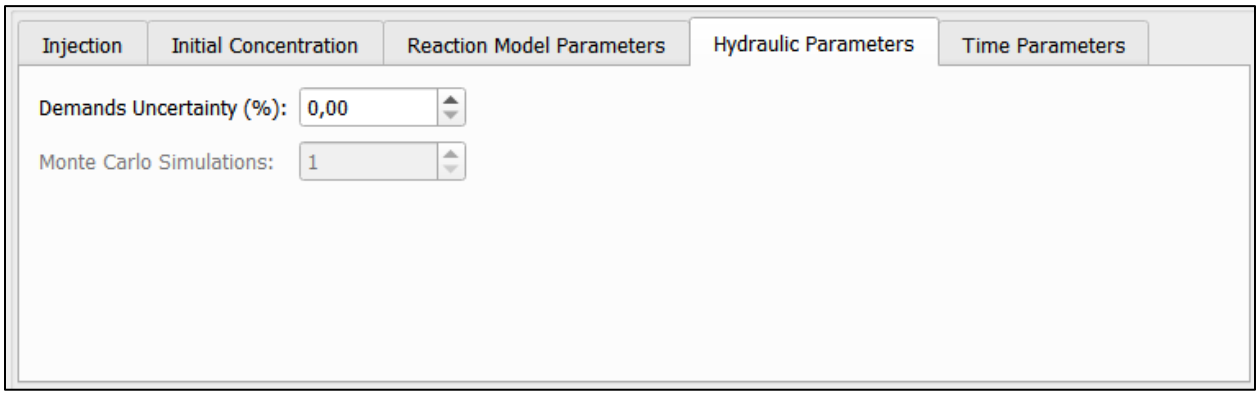


Figure 27: Hydraulic parameters

3.2.5 Time parameters

In this tab, users can set the simulation duration (e.g. 7 days) (**Figure 29**).

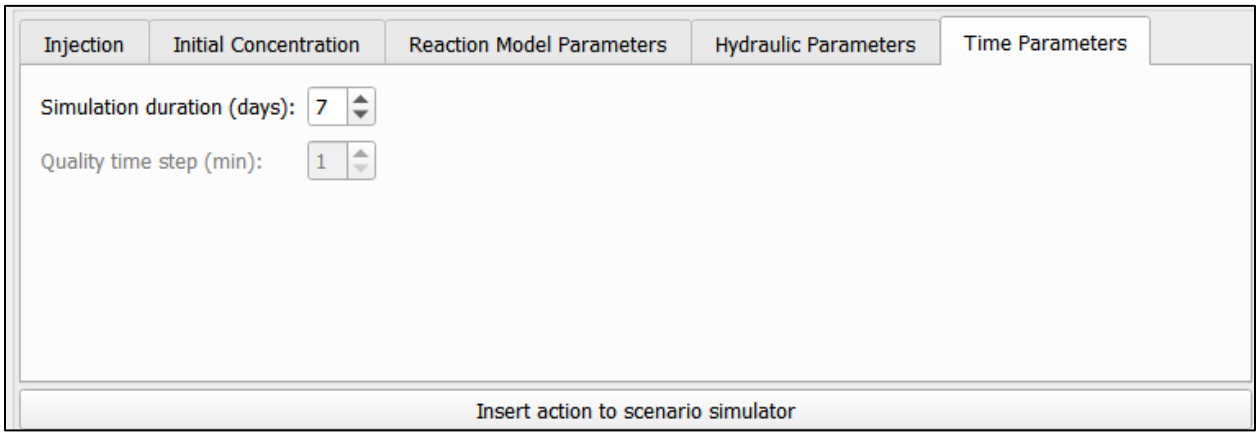


Figure 28: Time parameters

3.2.6 Insert action into the scenario simulator

The user can select the desired parameters of each type to build an action and then insert the desired action to the scenario simulator box, by clicking the 'Insert action to scenario simulator' button (**Figure 29**).

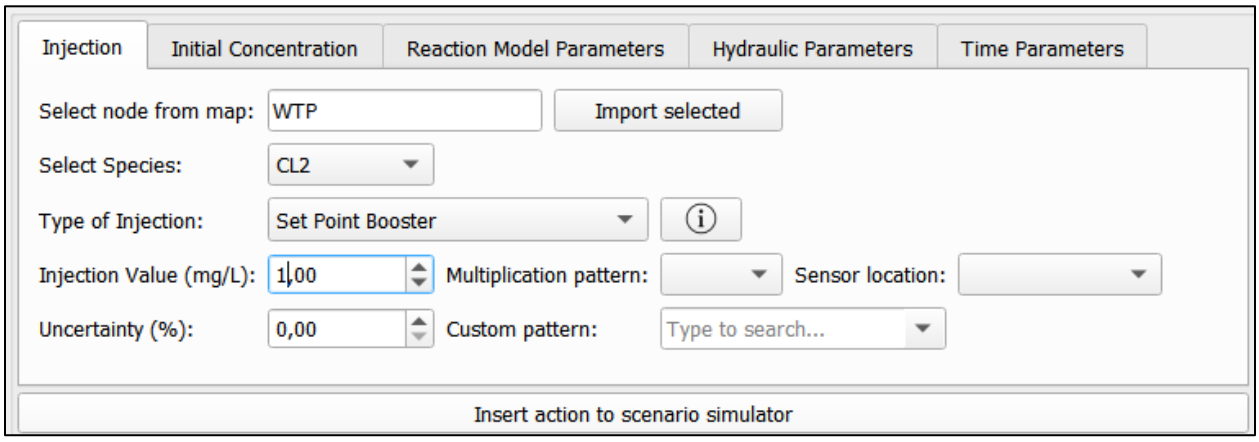


Figure 29: Insert action into scenario simulator

Figure 30 shows the scenario overview which provides a list of all configured actions. Users can manage actions using the 'Delete action' or 'Clear scenario' buttons, import/export scenarios, and search for any word to display actions containing the searched word.

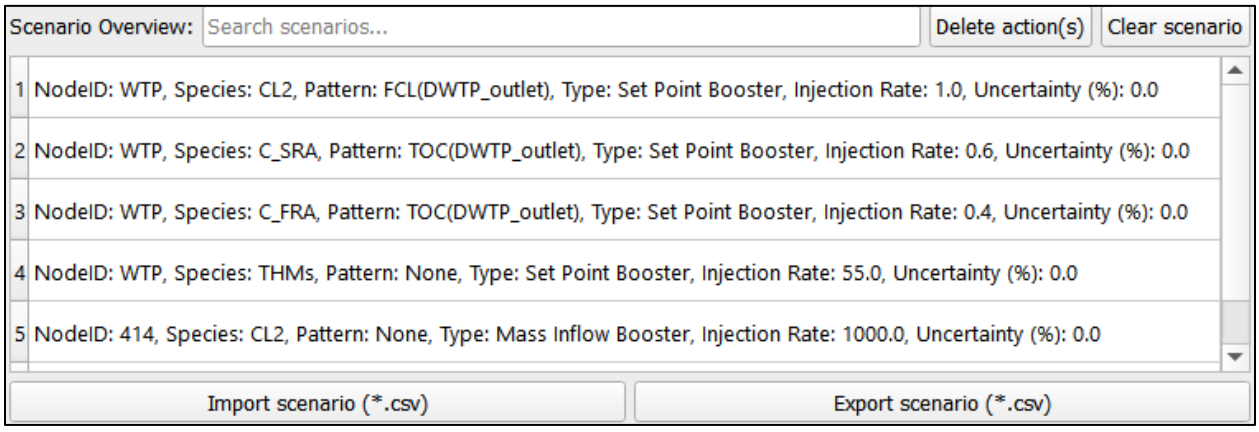


Figure 30: Scenario simulator interface

The user has the flexibility to manage the configured scenario by removing individual actions or clearing all actions entirely. A specific action can be deleted by clicking the 'Delete action' button (**Figure 31**), while the 'Clear scenario' button removes all actions from the scenario simulator (**Figure 32**). A confirmation warning is displayed before any deletion is finalized, ensuring that the user explicitly confirms the removal of actions.

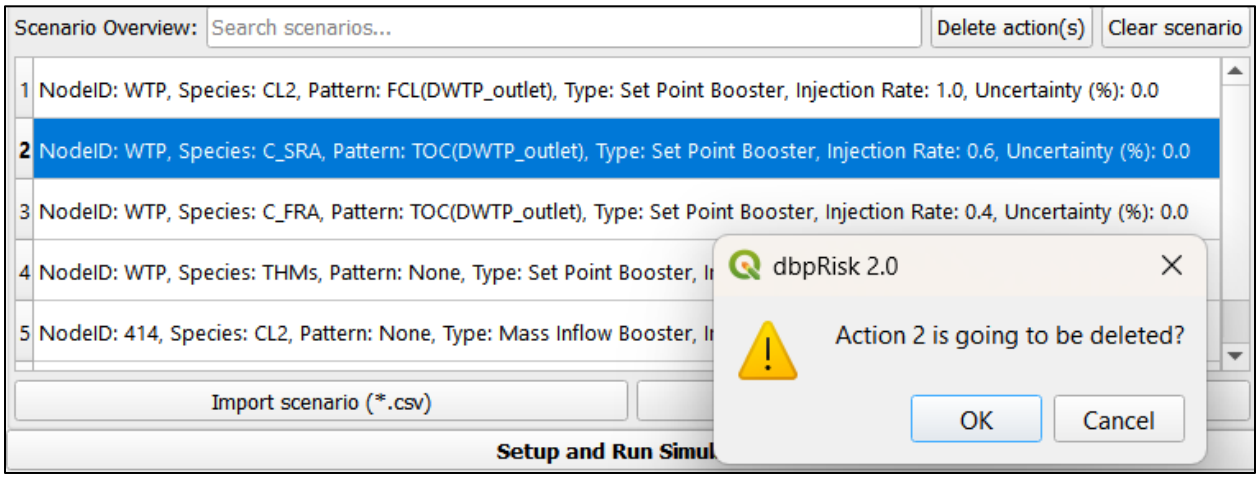


Figure 31: Delete action

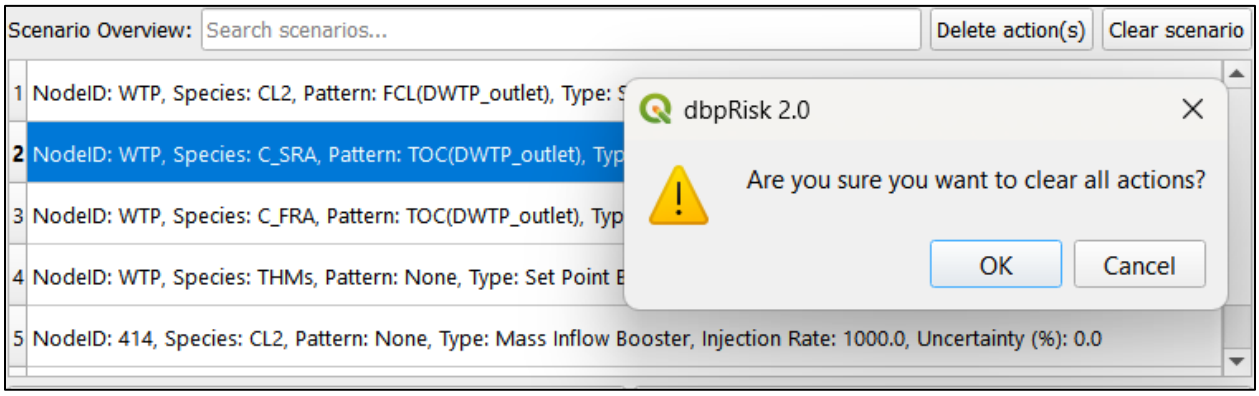


Figure 32: Clear scenario

Alternatively, a predefined scenario can be imported directly from a CSV file by clicking the 'Import scenario' button (**Figure 33**), enabling faster setup and reuse of previously configured scenarios.

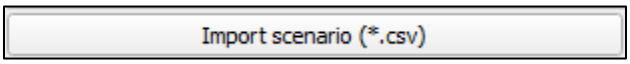


Figure 33: Import scenario button

The user can also export the configured scenario as a CSV file by clicking the 'Export scenario' button (**Figure 34**). This feature supports scenario sharing and future reuse.

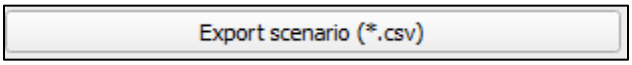


Figure 34: Export scenario button

3.3 Scenario simulator (Mechanistic Model)

In this section, once all desired actions have been added and retained in the scenario simulator box, the user can simulate the what-if scenario to support decision-making by clicking the 'Setup and Run Simulation' button (**Figure 35**). In the time parameters, the user can change the simulation duration and run it again without needing to add it in the scenario overview.



Figure 35: Setup and Run Simulation button

3.4 Results explorer

Figure 36 shows the results explorer interface, which allows users to:

- Select species
- Choose a specific hour (the range depends on the defined simulation duration) to view data for a particular simulation time
- Click the "Show on map" button to display the corresponding values on the map
- Select one or two species to plot at either sensor locations or selected nodes
- Use the "PLOT" button to generate visualizations

The "Include Samplings (API)" checkbox is currently disabled but is planned for future implementation.

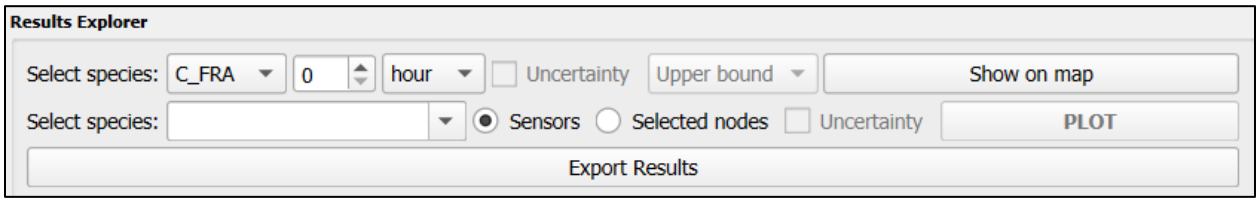


Figure 36: Results explorer interface

The user can select a species from a dropdown list (**Figure 37**), choose the hour option (**Figure 38**) and specify a particular hour from a dropdown list.

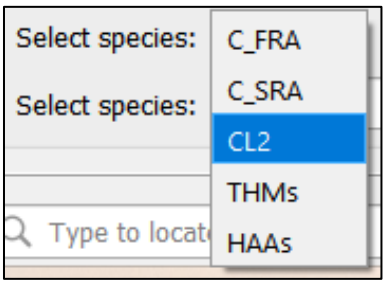


Figure 37: Select species

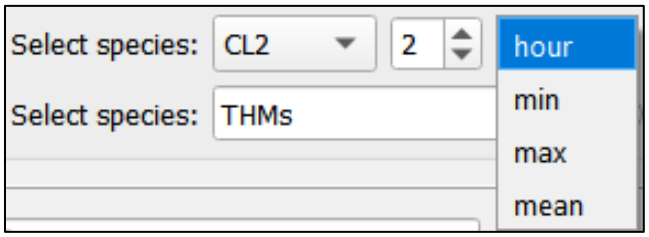


Figure 38: Choose hour option

By clicking the 'Show on map' button (**Figure 39**), the corresponding simulated values can be visualized on the map for that specific time, as shown in **Figure 40**. When hovering the mouse over a junction, the species name and its simulated value are displayed.



Figure 39: Show on map button

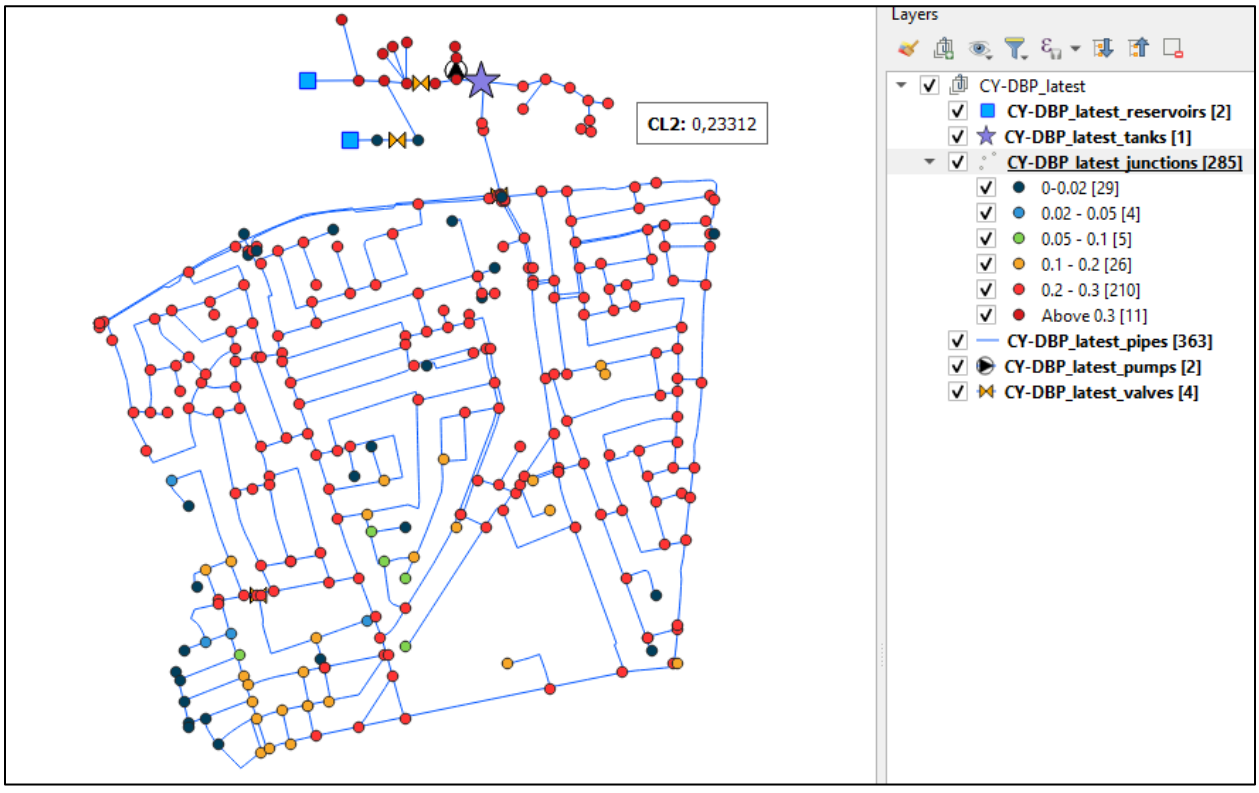


Figure 40: Simulated values on map

Additionally, the platform allows the display of the minimum, maximum and mean values of these species over the simulation period, by selecting the appropriate option from the dropdown list (**Figure 41**).

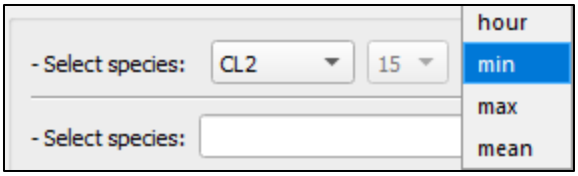


Figure 41: Show minimum values

The user can select one or two species (**Figure 42**), select Sensors and by clicking the 'PLOT' button (**Figure 43**) the simulated species at sensor locations can be visualized, as shown in **Figure 44**.

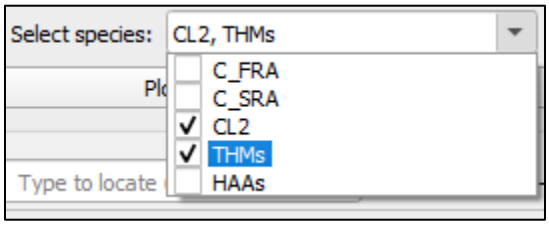


Figure 42: Select species

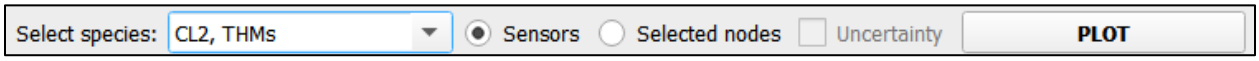


Figure 43: Plot species at sensor locations

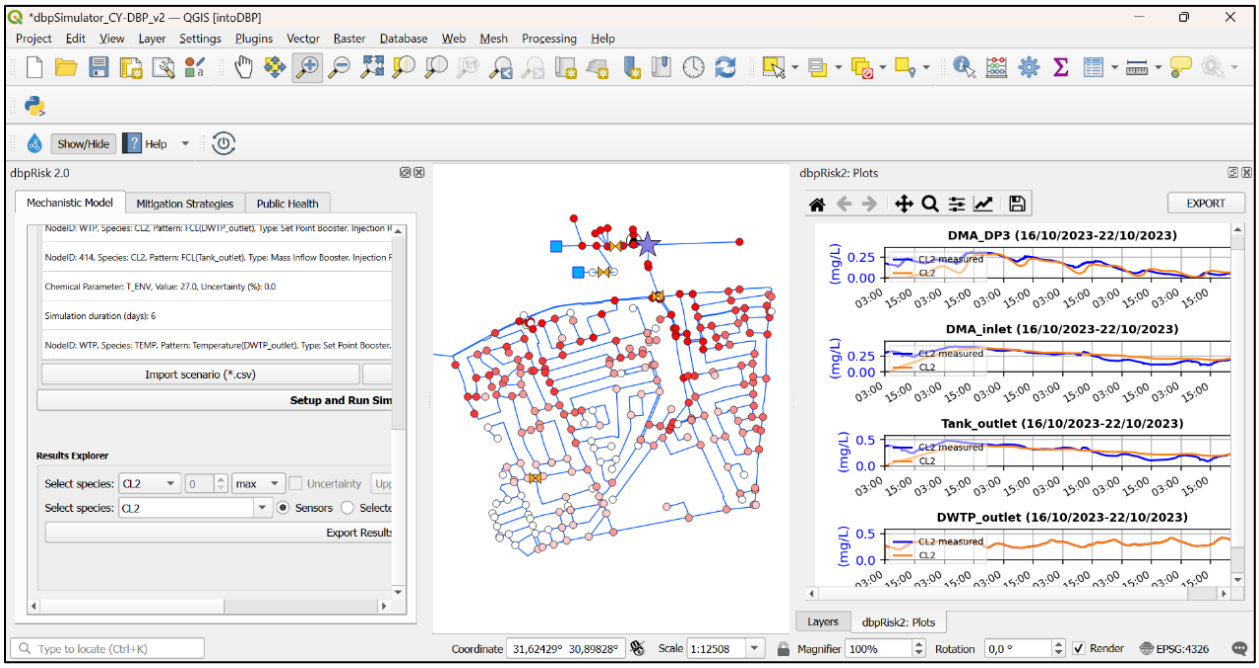


Figure 44: Plots of species at sensor locations

The user can also select nodes directly from the map, choose the "Selected nodes" option and by clicking the 'PLOT' button (**Figure 45**) the simulated species at these nodes can be visualized, as shown in **Figure 46**.



Figure 45: Plot species at selected nodes

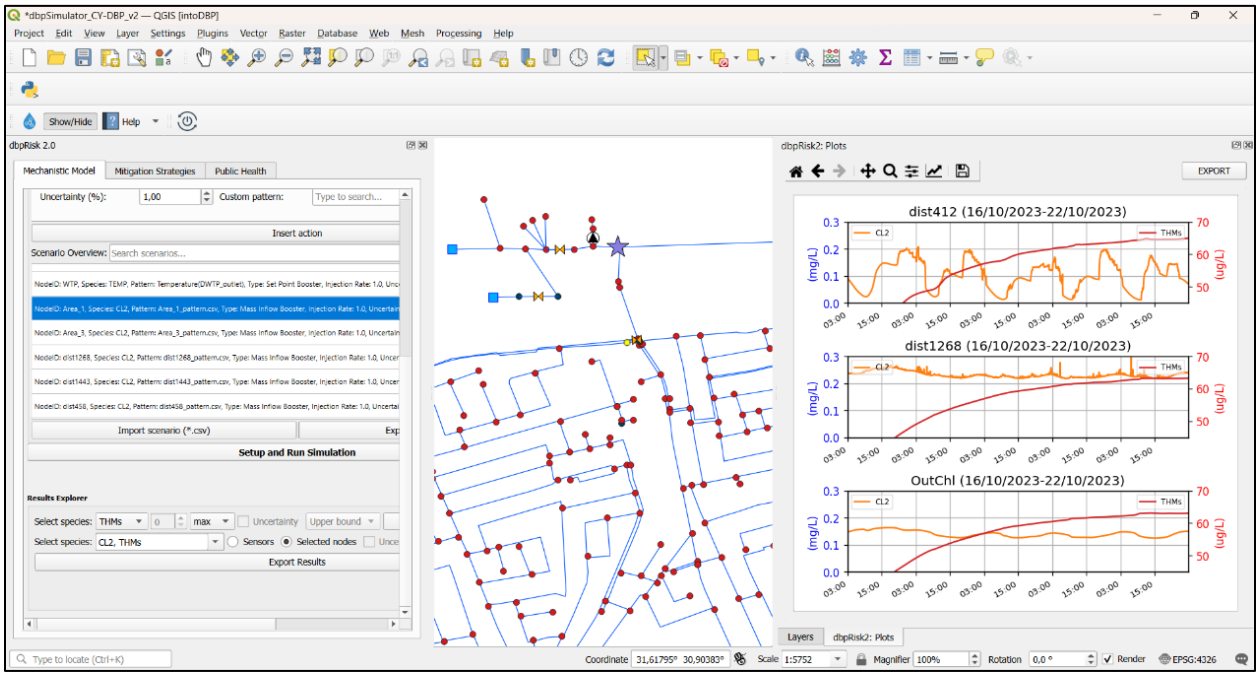


Figure 46: Plots of species at selected nodes

3.5 Mitigation Strategies

The proposed mitigation strategy consists of two sequential optimization steps: **(i) optimal placement of booster chlorination stations and (ii) optimization of the time-varying chlorine injection schedules** at the selected booster locations. The two optimization problems exploit the principle of superposition to represent the relationship between chlorine injections and network-wide disinfectant residuals, allowing the optimization to be formulated and solved efficiently using mathematical programming techniques.

The **first optimization problem** determines the optimal locations for booster chlorination stations. The algorithm identifies a set of locations that provides effective control of disinfectant residuals throughout the network while considering the maximum number of boosters that can be installed. The **second optimization problem** determines the chlorine injection pattern at the selected booster locations over time. The resulting schedules aim to maintain disinfectant concentrations within the specified operating limits while minimizing the formation of undesirable disinfection by-products such as THMs.

The optimization is formulated as a **multi-objective problem**, allowing competing objectives to be considered simultaneously rather than combining them into a single performance measure. The method therefore generates a set of Pareto-optimal solutions, representing different trade-offs between the objectives. These solutions allow the user to select the operational strategy that best satisfies the required water-quality and operational constraints.

To run the optimization algorithm, the user must provide the following inputs (**Figure 47**):

- **Hydraulic model:** the water distribution network model on which the optimization is performed.
- **Lower chlorine bound:** the minimum allowable disinfectant residual concentration.
- **Upper chlorine bound:** the maximum allowable disinfectant residual concentration.
- **Maximum number of boosters:** the maximum number of booster chlorination stations that may be selected by the optimization algorithm.

Based on these inputs, the algorithm identifies the **optimal booster locations** and subsequently determines the **optimal disinfection/injection schedule** for the selected boosters, while satisfying the specified chlorine concentration limits.

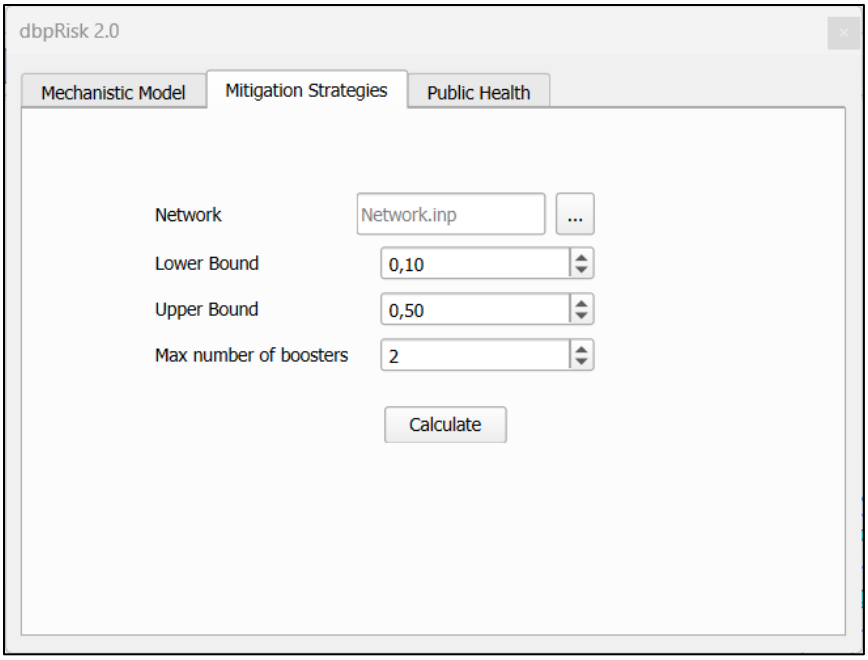


Figure 47: Inputs required for the optimization algorithm to control DBP formation

3.6 DBP Exposure

We assess the DBP exposure utilizing the ADD - average daily potential dose (mg kg^{-1} day^{-1}),

$$ADD = \frac{C * IR * EF * ED}{BW * AT}$$

where C the pollutant concentration (mg/L), IR the injection rate of daily drinking water (L day^{-1}), EF the exposure frequency (days $year^{-1}$), ED the exposure duration (year), BW the body weight (kg) and AT the average time (days).

Monte Carlo simulations can be employed to address uncertainties including ingestion rate, body weight, and exposure duration.

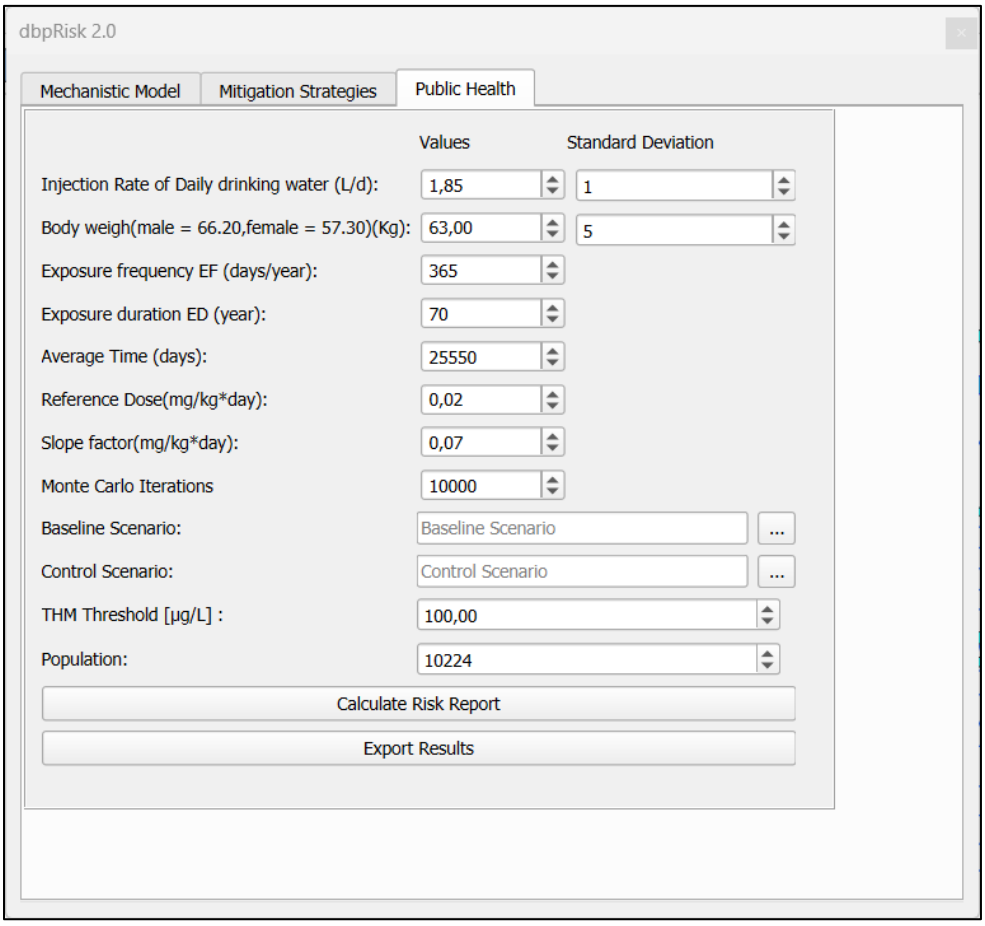


Figure 48: Inputs required for the DBP exposure assessment

4 Use case

This section guides the user through a complete example of how to use dbpRisk v2.0, for two scenarios. A baseline scenario to define the current situation of a case study and the corresponding optimization scenario.

First, select a hydraulic model from the dropdown list - e.g., CY-DBP_v1.inp (**Figure 49**).

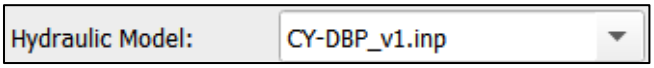


Figure 49: Select hydraulic model

Then, select a reaction model from the dropdown list - e.g., THMs_HAAs_Pilot.msx (**Figure 50**).

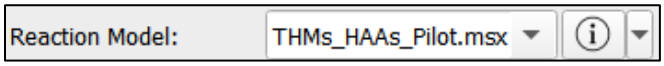


Figure 50: Select reaction model

After selecting the hydraulic and reaction models, load the associated GIS data onto the map by clicking the 'Load Models' button (**Figure 51**).



Figure 51: Load models

The corresponding network elements then appear on the map for visualization and interaction, as shown in **Figure 52**.

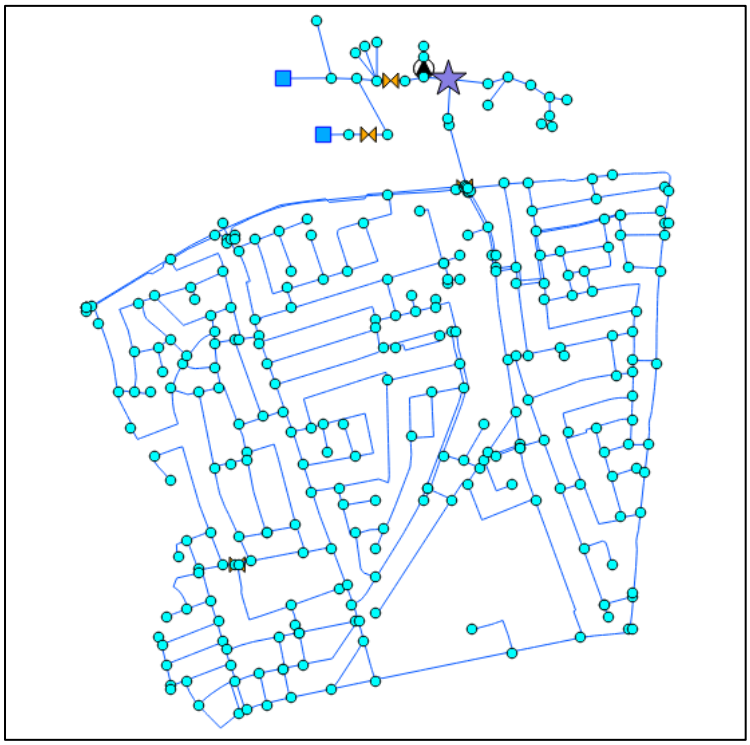


Figure 52: Map view displaying the loaded GIS data

Select an excel file containing sensor data - e.g., sensors_data.xlsx (**Figure 53**).

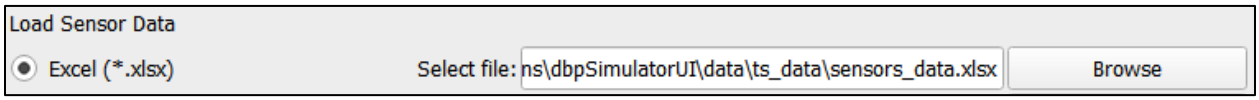


Figure 53: Load sensor data

Subsequently, configure the required parameters for each selected action type. Once configured, insert the defined action into the Scenario Overview by clicking the "Insert Action to Scenario Simulator" button.

4.1 Baseline scenario:

As shown in **Table 1**, disinfection is based on continuous chlorine injection at the WTP, with a chlorine concentration pattern retrieved from the sensor at the DWTP outlet. Additional rechlorination is represented by chlorine injection at Node 414, located upstream of the storage tank, thereby representing the existing tank rechlorination practice. The chlorine pattern retrieved from the Tank outlet sensor is used to calibrate the rechlorination process. NOM is represented

using UV sensor measurements, which are used as the surrogate measurement for NOM in the Digital Twin. The MSX chemical parameters, including NOM_SURR = 0.7 and Y_NOM_SURR = 0.8, are calibration parameters that allow the EPANET-MSX model to simulate NOM-related reactions and DBP formation using UV measurements as the NOM representative. Temperature is represented by using the corresponding sensor measurements, while the environmental temperature parameter T_ENV = 27°C is used to define the outside environmental conditions. DBPs initial concentrations at the WTP outlet are defined for THMs and HAAs.

Table 1: Baseline scenario disinfection strategy

Baseline scenario	**Parameter**
NodeID: WTP, Species: CL2, Pattern: FCL(DWTP_outlet), Type: Set Point Booster, Injection Rate: 1.0, Uncertainty (%): 0.0	Chlorine
NodeID: 414, Species: CL2, Pattern: FCL(Tank_outlet), Type: Mass Inflow Booster, Injection Rate: 10000.0, Uncertainty (%): 0.0	
NodeID: WTP, Species: C_SRA, Pattern: UV254f(DWTP_outlet), Type: Set Point Booster, Injection Rate: 1.0, Uncertainty (%): 0.0	NOM
Chemical Parameter: NOM_SURR, Value: 0.7, Uncertainty (%): 0.0	
Chemical Parameter: Y_NOM_SURR, Value: 0.8, Uncertainty (%): 0.0	
NodeID: WTP, Species: TEMP, Pattern: Temperature (DWTP_outlet), Type: Set Point Booster, Injection Rate: 1.0, Uncertainty (%): 0.0	Temperature
Chemical Parameter: T_ENV, Value: 27.0, Uncertainty (%): 0.0	
NodeID: WTP, Species: THMs, Pattern: None, Type: Set Point Booster, Injection Rate: 55.0, Uncertainty (%): 0.0	DBPs
NodeID: WTP, Species: HAAs, Pattern: None, Type: Set Point Booster, Injection Rate: 15.0, Uncertainty (%): 0.0	

Simulation duration (days): 7	Simulation Duration

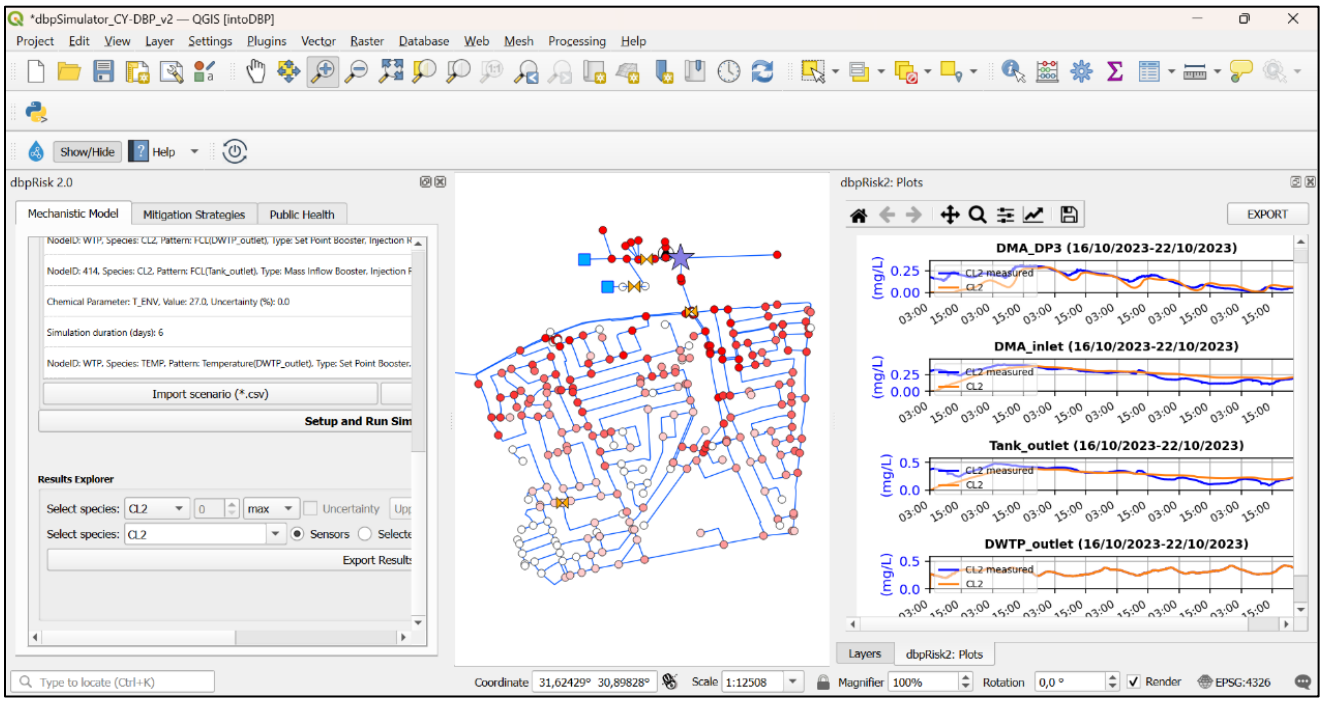


Figure 54: Simulated maximum chlorine concentration values plotted on the map along with the corresponding time series at sensor locations compared to the field sensor data (baseline scenario)

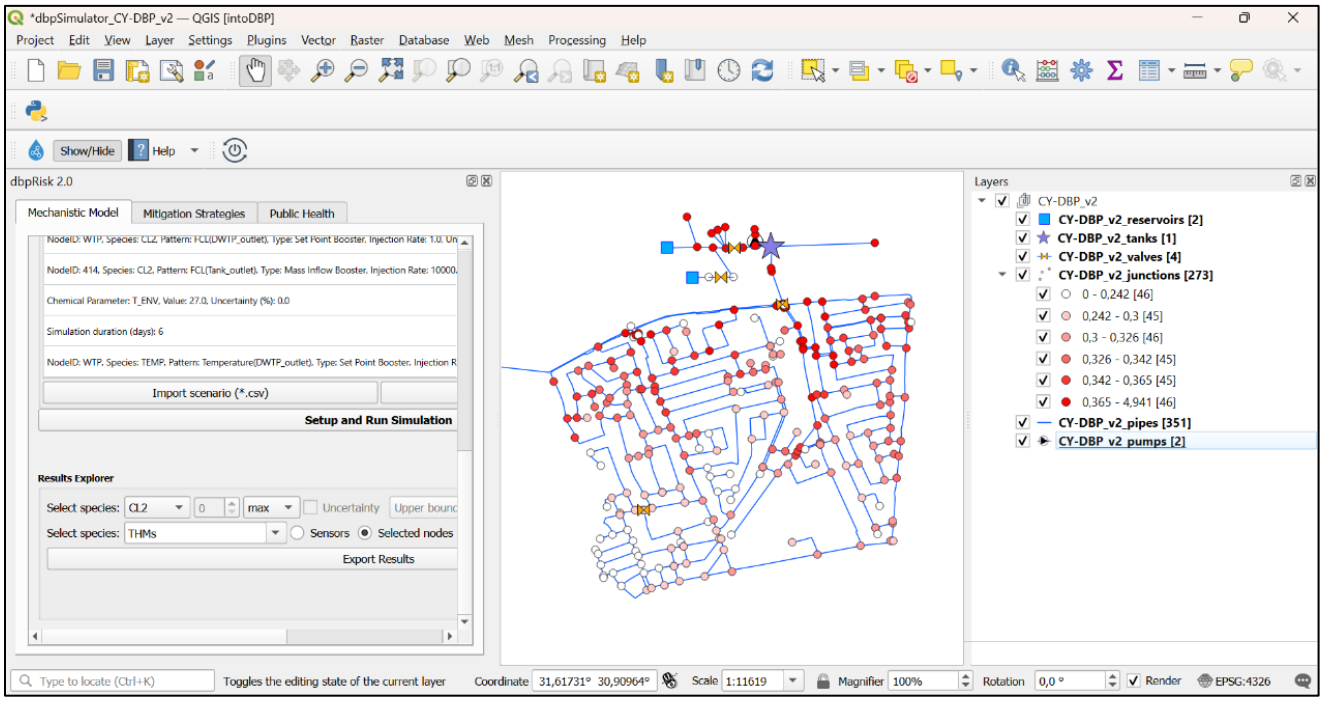


Figure 55: Simulated maximum chlorine concentration values plotted on the map with the corresponding legend (baseline scenario)

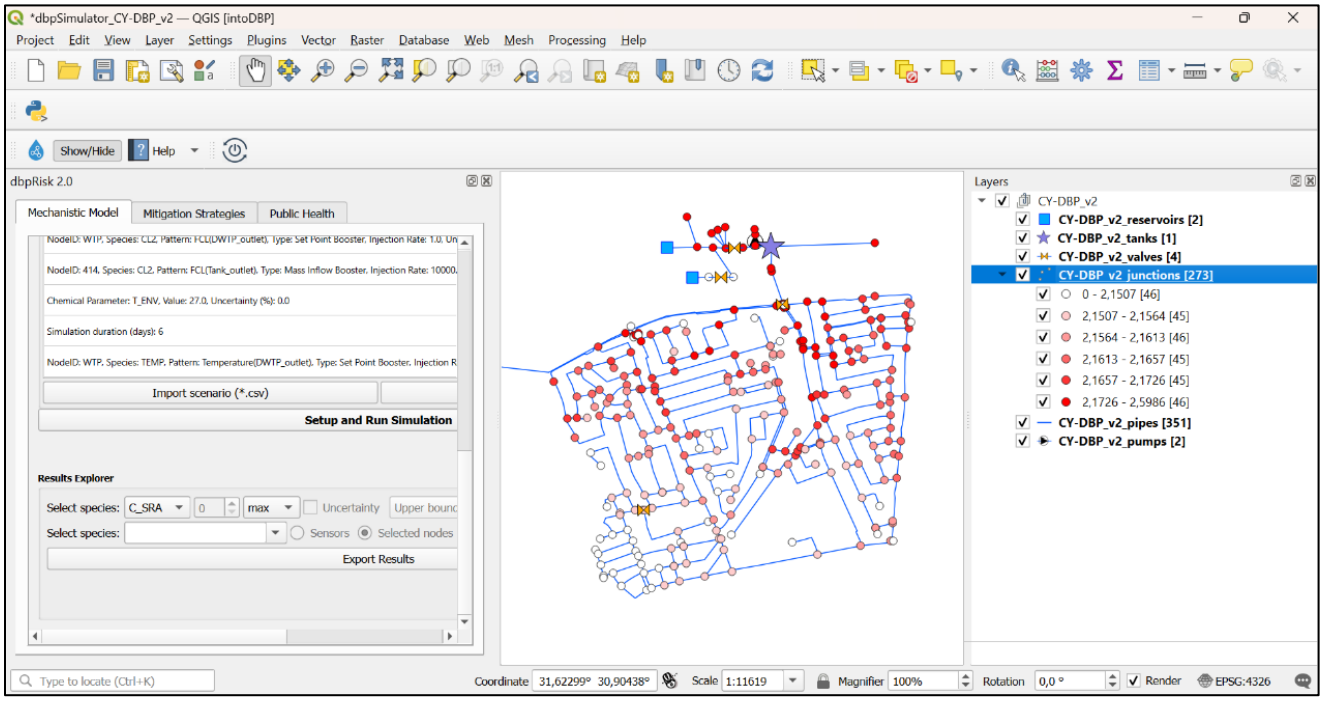


Figure 56: Simulated maximum NOM (UV) concentration values plotted on the map with the corresponding legend (baseline scenario)

Figure 54 and **Figure 55** present the simulated maximum chlorine concentrations across the network together with simulated and measured chlorine time series at the available sensor locations. The spatial distribution shows higher chlorine concentrations in the upstream/northern part of the network (>0.3 mg/L), while lower residuals occur toward the downstream distribution areas (<0.3 mg/L). The sensor time series show that the model reproduces the main temporal variations observed in the field measurements.

Figure 56 presents the simulated maximum NOM concentration represented by the UV surrogate. The values are concentrated within a relatively narrow range of approximately **2.15–2.17 UV units**, with only small spatial variations across the network. This indicates that the NOM input remains comparatively uniform throughout the case-study area

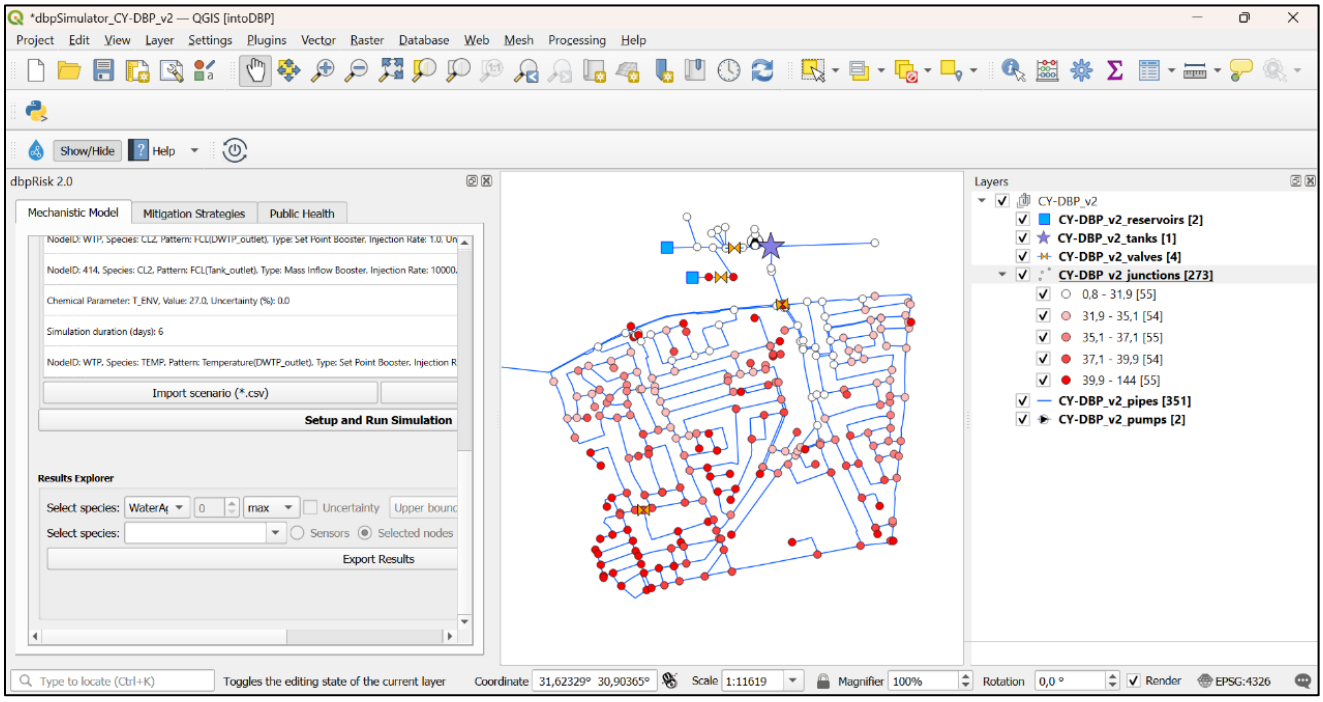


Figure 57: Simulated maximum water age values plotted on the map with the corresponding legend (baseline scenario)

Figure 57 shows maximum water ages ranging approximately from 31 to 40 h. Higher water ages are mainly observed in the downstream parts of the distribution network, indicating longer residence times in these areas and consequently greater potential for DBP formation.

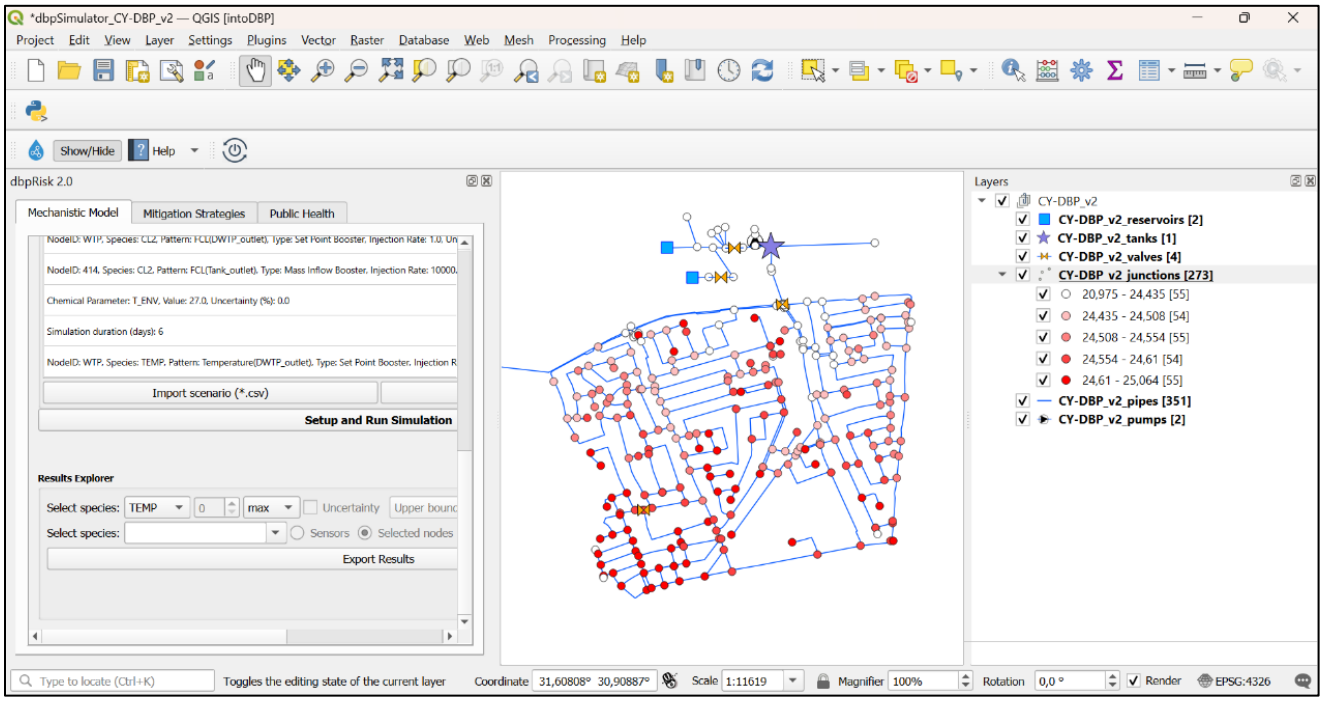


Figure 58: Simulated maximum temperature concentration values plotted on the map with the corresponding legend (baseline scenario)

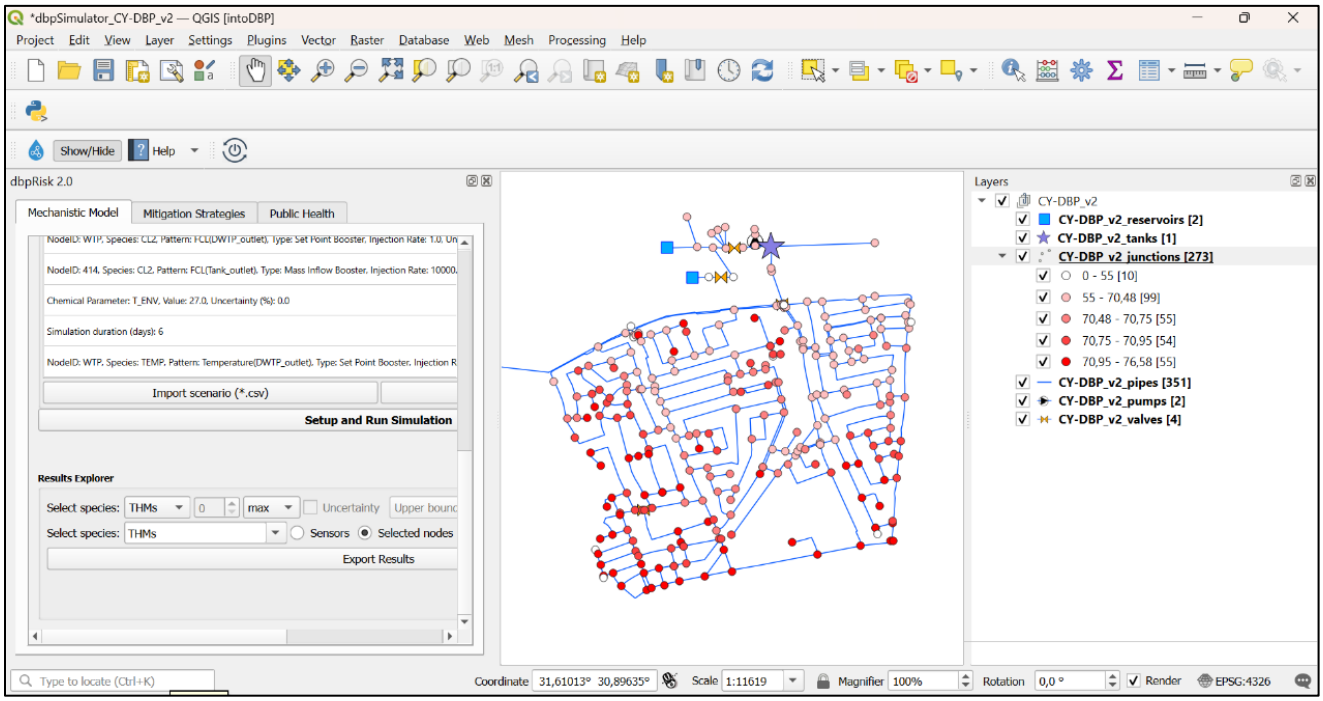


Figure 59: Simulated maximum THM concentration values plotted on the map with the corresponding legend (baseline scenario)

Figure 58 shows maximum simulated temperatures between approximately 21 and 25 °C. Higher temperatures are predominantly distributed across the main downstream network, while lower values occur in the upstream areas. The temperature field provides an important environmental condition for the simulated chlorine decay and DBP formation processes.

Figure 59 presents maximum THM concentrations ranging from approximately 55 to 77 µg/L. Most nodes fall within the 55–71 µg/L range, while the highest concentrations (71–77 µg/L) are concentrated mainly in downstream sections of the network. The spatial increase in THM relative to the WTP initial value of 55 µg/L indicates additional formation within the distribution system.

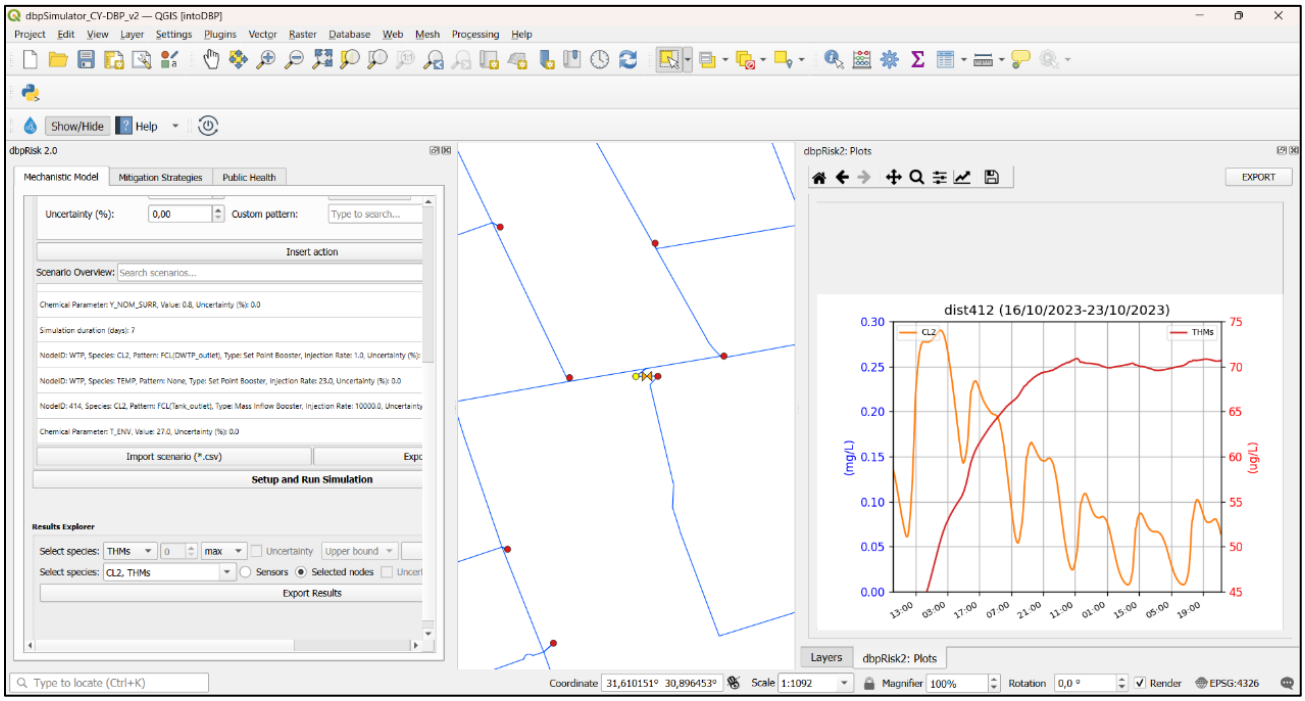


Figure 60: Chlorine and THM concentration time series for a specific node within the distribution zone (baseline scenario)

Figure 60 illustrates the temporal evolution of chlorine and THM at node dist412. Chlorine varies approximately between 0.01 and 0.30 mg/L, with repeated daily fluctuations, while THM increases from approximately 45 μg/L to around 72–73 μg/L over the simulation period.

Figure 61 highlights the network nodes exceeding the defined THM threshold of 70 μg/L. A substantial portion of the affected nodes is concentrated in the central and downstream areas of the distribution network. The assessment identifies an affected population of approximately 6,757 people out of the total population of 10,224 (**Figure 62**), demonstrating the spatial extent of the baseline THM risk. **Figure 62** presents the Monte Carlo frequency distribution of THM average daily dose (ADD). The distribution has a mean ADD of approximately 2.10×10^{-3} mg/kg-day.

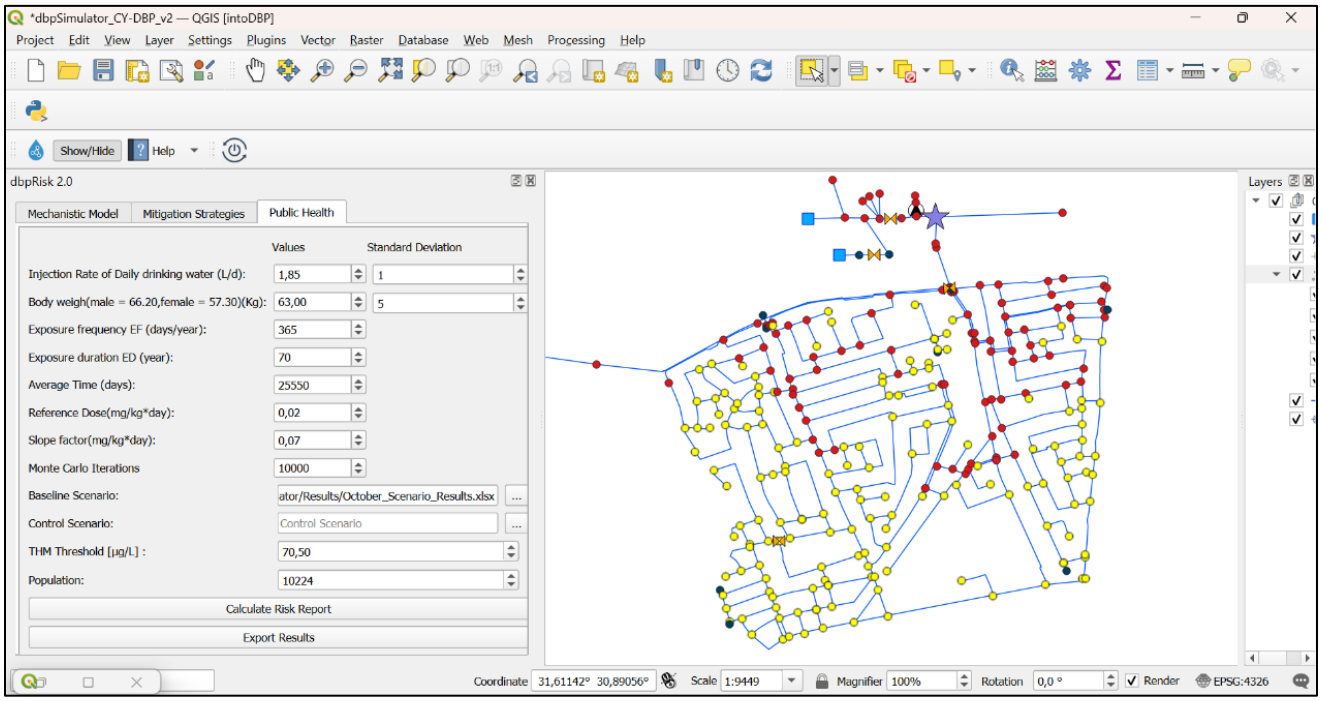


Figure 61: DBP Exposure dashboard with nodes above the defined THM threshold level highlighted

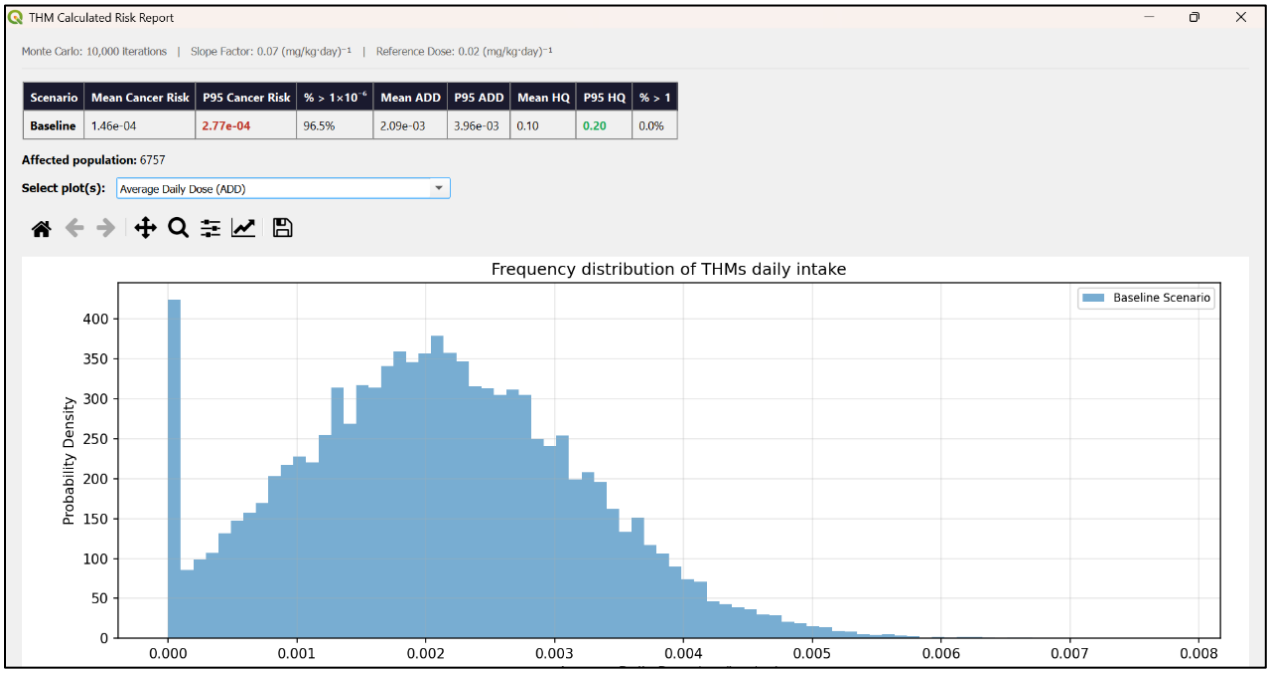


Figure 62: DBP Exposure dashboard indicating the THM Average Daily Dose (mg kg-1 day-1) frequency distribution under (baseline scenario)

4.2 Optimization scenario

The optimization strategy is defined by utilizing the algorithms developed and presented in the main deliverable and integrated into the software. Using the corresponding interface (**Figure 47**), the network is loaded, the targeted network chlorine levels are defined (0.10–0.50 mg/L) along with the available number of boosters (m = 1 – 3 booster stations).

The optimized strategy maintains the WTP chlorine injection while removing the rechlorination at Node 414 upstream of the tank. Instead, chlorine is supplied through five distributed booster locations, with Area_1 and Area_3 serving the Zone 2 Transport Network (Primary Transmission System), while dist1268, dist1443, and dist458 serve the DMA 230 downstream distribution area (**Figure 63**). All booster stations use time-varying mass-based chlorine injection patterns (**Figure 64 and Figure 65**), allowing the chlorine dose to be spatially and temporally distributed according to the optimization results. Area_1 generally operates at higher doses, reaching approximately 1,100 mg, whereas Area_3 reaches approximately 600 mg. Both boosters exhibit a clear repeating diurnal pattern, with higher daytime dosing and lower dosing periods during the night. Dist458 contributes quite low mass over most of the period, while dist1443 operates at approximately 5–20 mg and dist1268, which is located at the entrance of the DMA, provides the largest contribution, generally around 10–50 mg with several short-duration peaks. This configuration aims to maintain chlorine residuals within the target operating range while avoiding excessive chlorine addition and limiting subsequent DBP formation (**Figure 66**). Most residuals remain approximately within the 0.10–0.50 mg/L target operating range, with the majority of node trajectories concentrated below approximately 0.40 mg/L. Short-duration peaks above the target range occur at locations close to booster injection points reflect localized effects of the mass-based injection schedules. The proposed new booster injection patterns are extracted into .csv files which can be then used to run the control scenario as show in

Table 2.

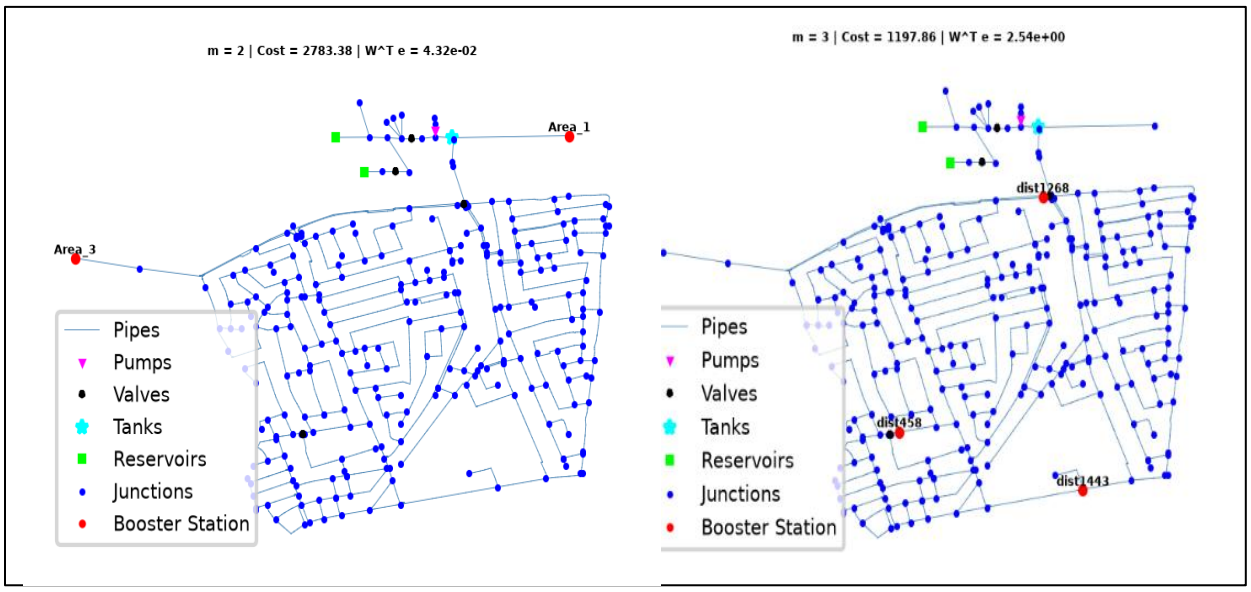


Figure 63: DBP formation control strategies with optimal booster placement across Zone 2 (left) and DMA 230 (right)

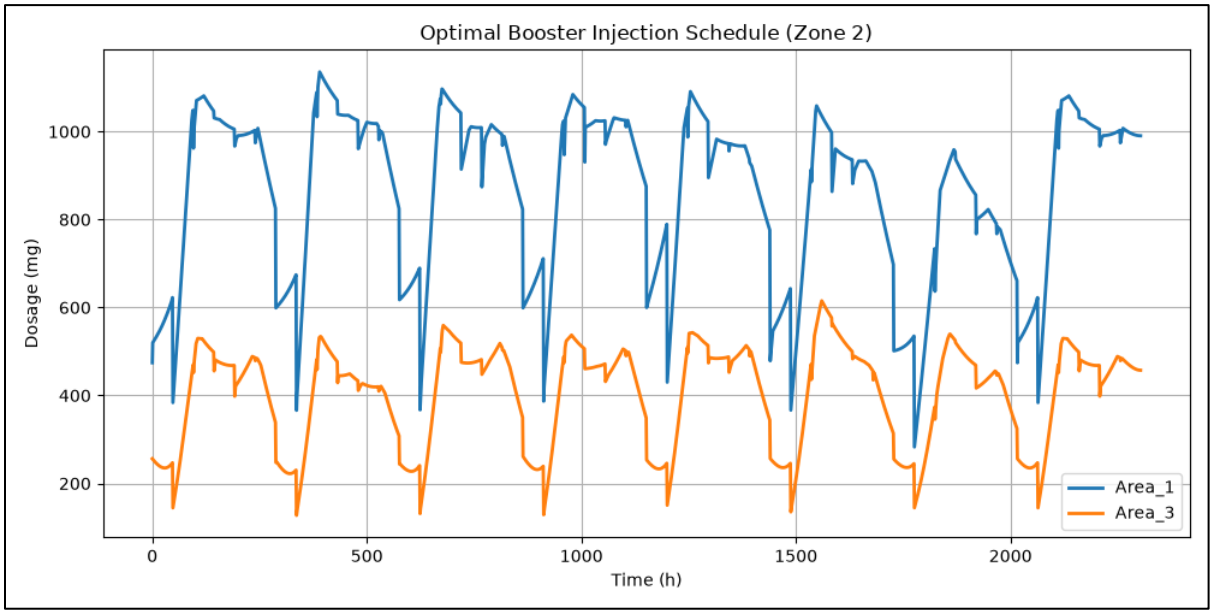


Figure 64: Diurnal dosing pattern for Zone 2 chlorine boosters

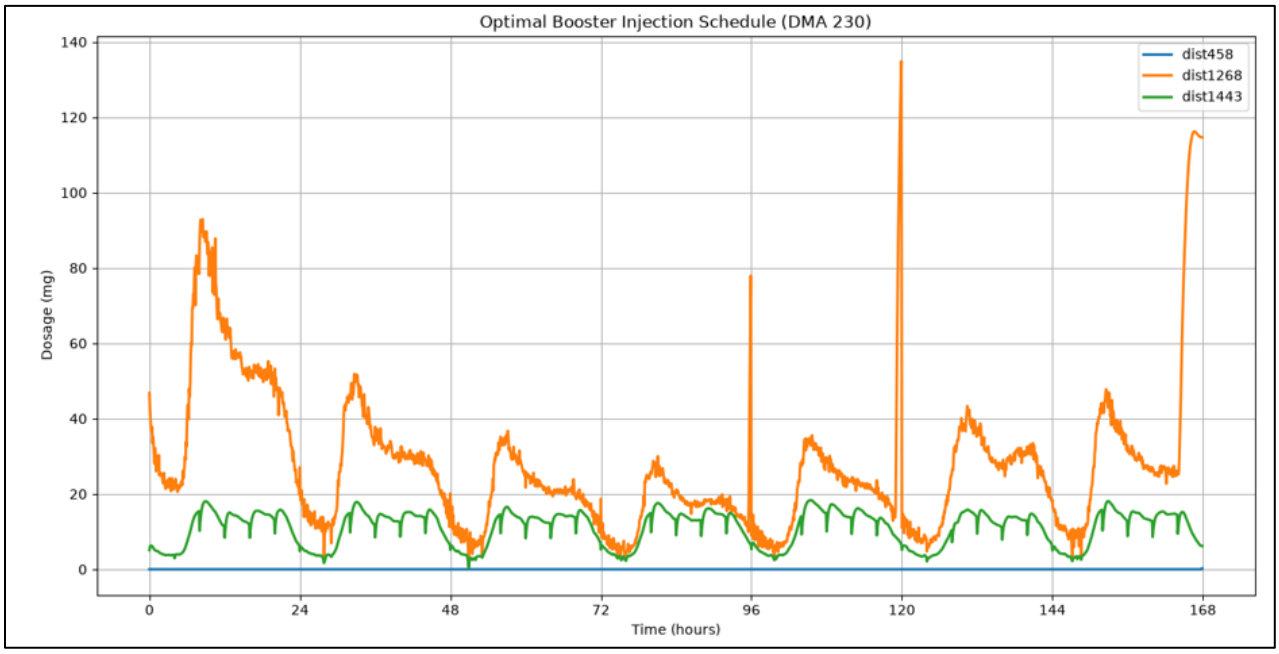


Figure 65: Diurnal dosing pattern for DMA 230 chlorine boosters

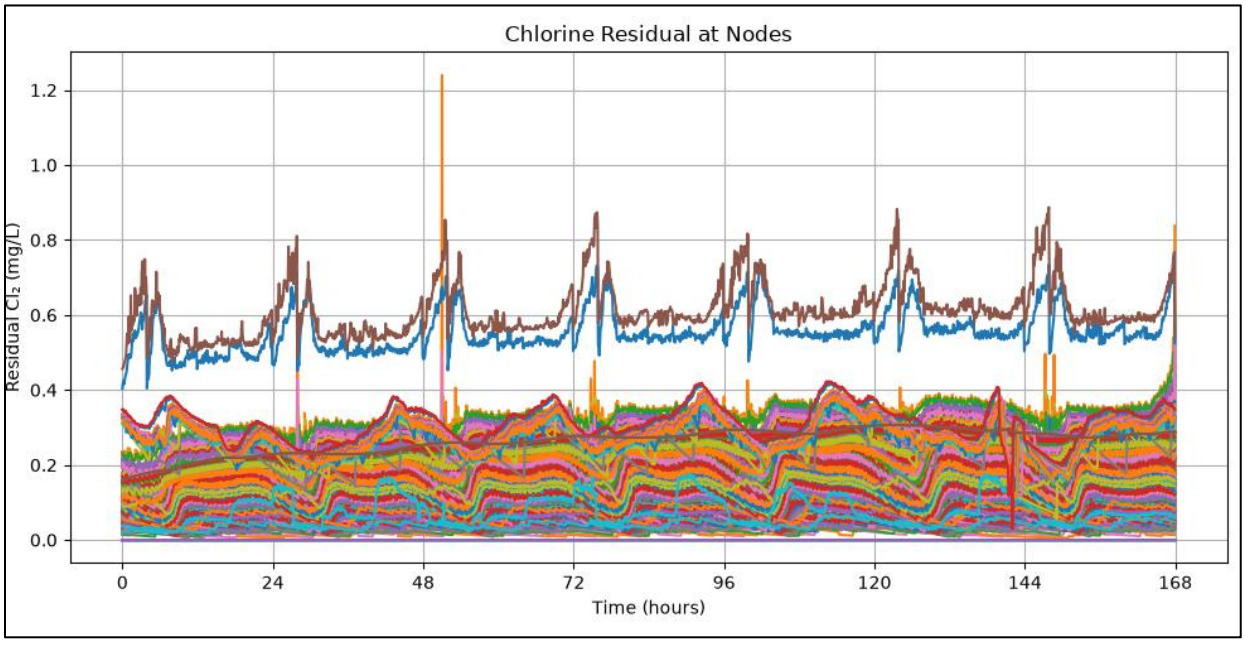


Figure 66: Chlorine residuals at all nodes following the application of the mitigation strategy

Table 2: Control scenario disinfection strategy

Control scenario	Parameter
NodeID: WTP, Species: CL2, Pattern: FCL(DWTP_outlet), Type: Set Point Booster, Injection Rate: 1.0, Uncertainty (%): 0.0	Chlorination
NodeID: Area_1, Species: CL2, Pattern: Area_1_pattern.csv, Type: Mass Inflow Booster, Injection Rate: 1.0, Uncertainty (%): 0.0	Booster Chlorination
NodeID: Area_3, Species: CL2, Pattern: Area_3_pattern.csv, Type: Mass Inflow Booster, Injection Rate: 1.0, Uncertainty (%): 0.0	
NodeID: dist1268, Species: CL2, Pattern: dist1268_pattern.csv, Type: Mass Inflow Booster, Injection Rate: 1.0, Uncertainty (%): 0.0	
NodeID: dist1443, Species: CL2, Pattern: dist1443_pattern.csv, Type: Mass Inflow Booster, Injection Rate: 1.0, Uncertainty (%): 0.0	
NodeID: dist458, Species: CL2, Pattern: dist458_pattern.csv, Type: Mass Inflow Booster, Injection Rate: 1.0, Uncertainty (%): 0.0	
NodeID: WTP, Species: C_SRA, Pattern: UV254f(DWTP_outlet), Type: Set Point Booster, Injection Rate: 1.0, Uncertainty (%): 0.0	NOM
Chemical Parameter: NOM_SURR, Value: 0.7, Uncertainty (%): 0.0	
Chemical Parameter: Y_NOM_SURR, Value: 0.8, Uncertainty (%): 0.0	
NodeID: WTP, Species: TEMP, Pattern: Temperature (DWTP_outlet), Type: Set Point Booster, Injection Rate: 1.0, Uncertainty (%): 0.0	Temperature
Chemical Parameter: T_ENV, Value: 27.0, Uncertainty (%): 0.0	
NodeID: WTP, Species: THMs, Pattern: None, Type: Set Point Booster, Injection Rate: 55.0, Uncertainty (%): 0.0	DBPs

NodeID: WTP, Species: HAAs, Pattern: None, Type: Set Point Booster, Injection Rate: 15.0, Uncertainty (%): 0.0	
Simulation duration (days): 7	Simulation Duration

Figure 67 shows the spatial distribution of maximum chlorine following optimization. Compared with the baseline scenario, where the spatial distribution exhibits higher chlorine concentrations in the upstream/northern parts of the network (>0.3 mg/L) and lower residuals in the downstream distribution areas (<0.3 mg/L), the optimization approach incorporating boosters significantly reduces this spatial variability. The optimized scenario maintains more uniform chlorine residuals throughout the network, keeping concentrations closer to the desired level across all distribution areas.

Figure 68 shows a reduction in the spatial extent of high THM concentrations compared with the baseline. The maximum THM values are approximately 65-70 μg/L, compared with the baseline maximum of 75 μg/L. The higher THM classes above approximately 70μg/L are effectively eliminated, indicating a reduction in the highest DBP concentrations following optimization.

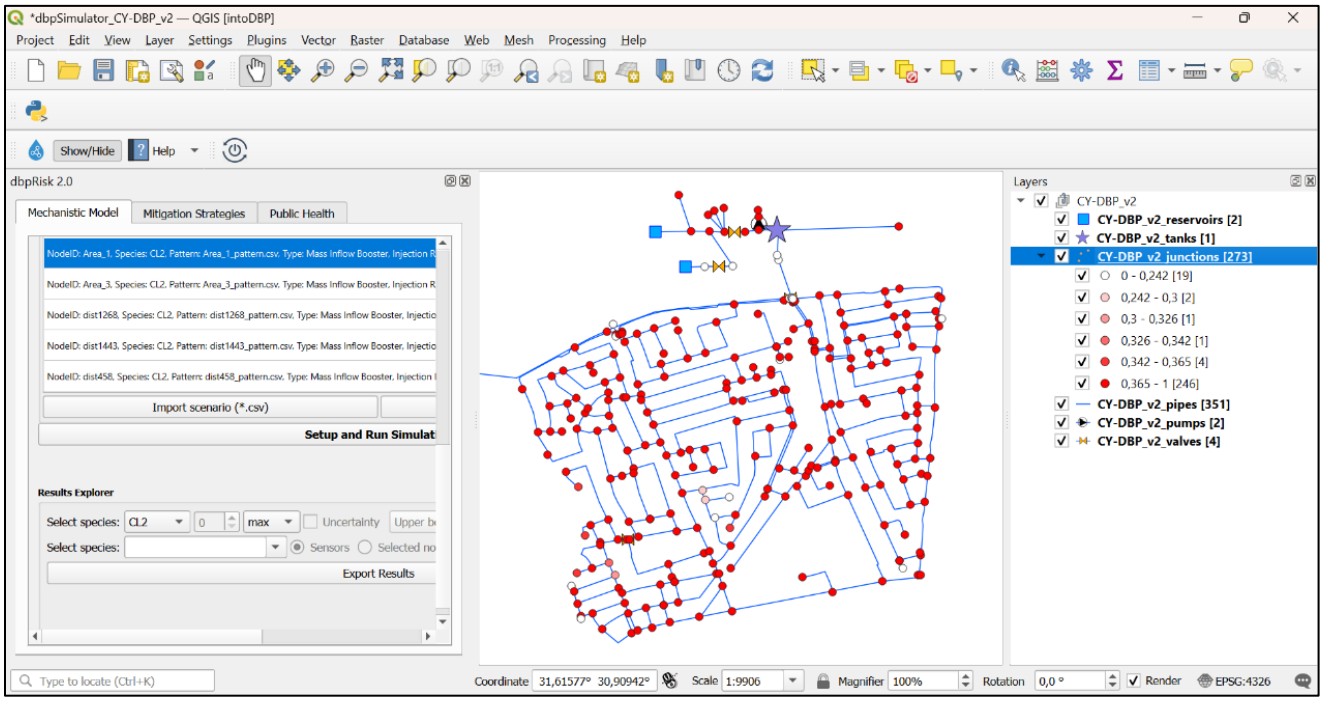


Figure 67: Simulated maximum chlorine concentration values plotted on the map with the corresponding legend (control scenario)

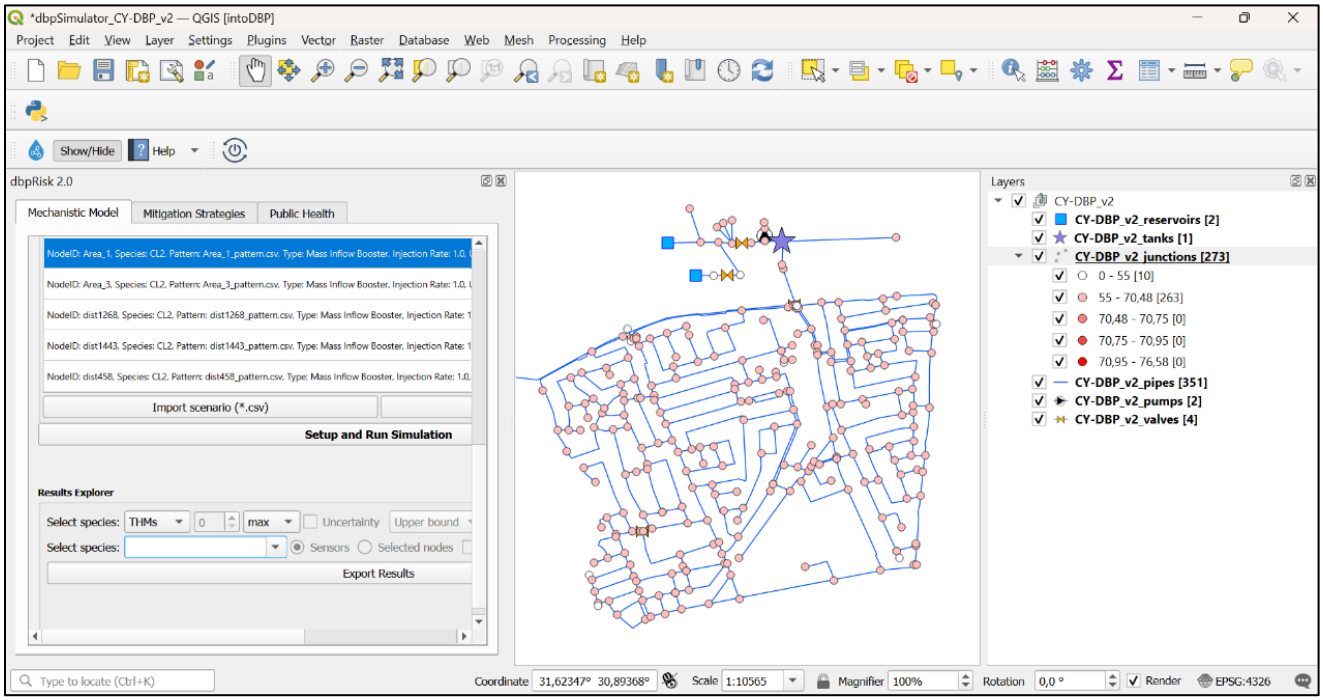


Figure 68: Simulated maximum THM concentration values plotted on the map with the corresponding legend (control scenario)

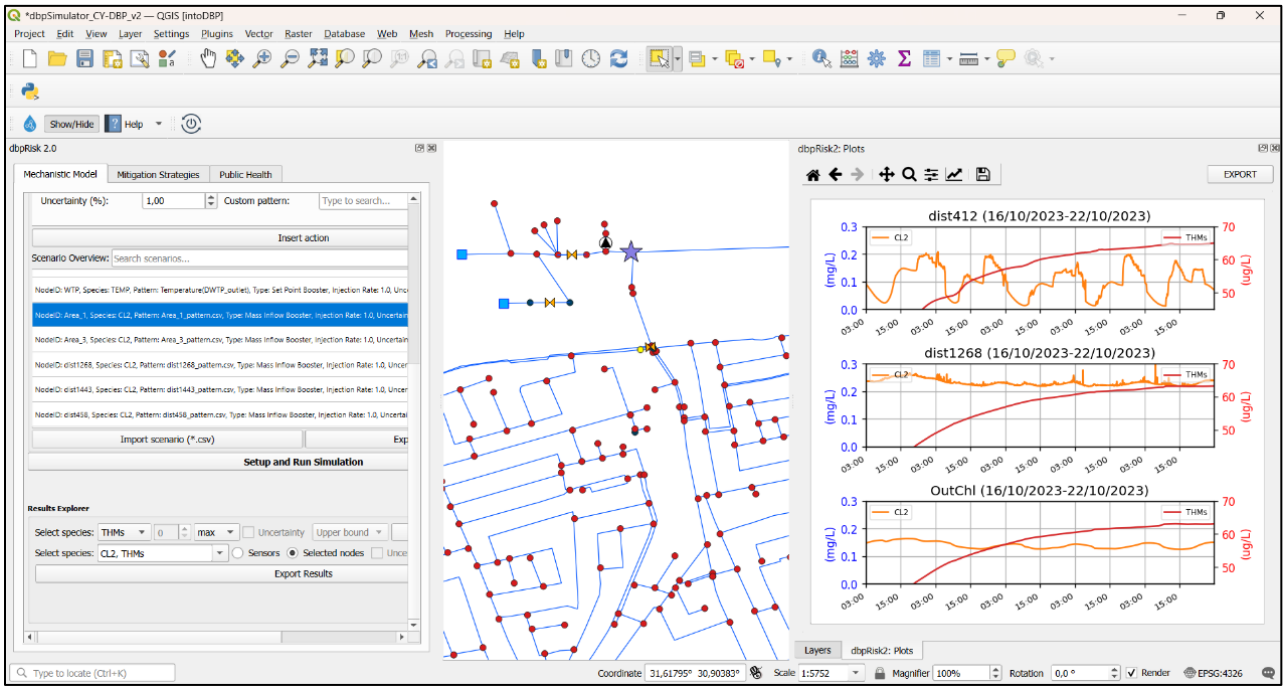


Figure 69: Chlorine and THM concentration time series for specific nodes within the distribution zone (control scenario)

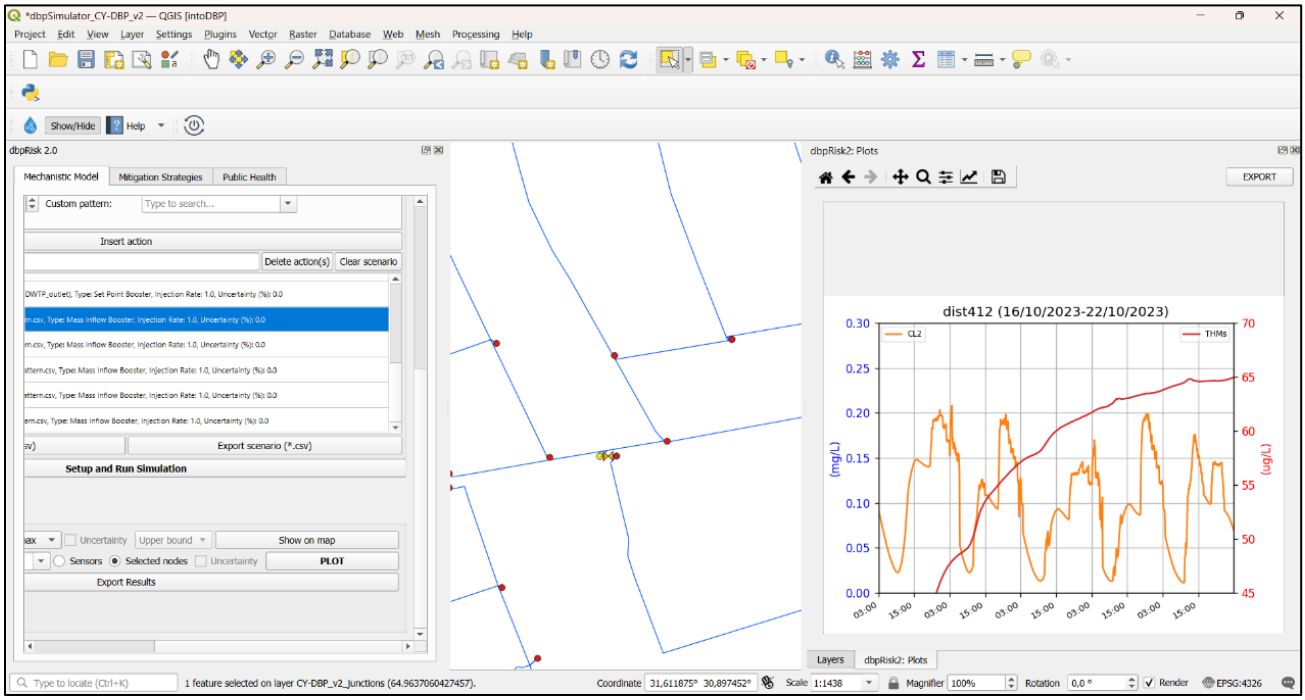


Figure 70: Chlorine and THM concentration time series for a specific node within the distribution zone (control scenario)

Figure 69 compares chlorine and THM time series at dist412, dist1268, and the tank outlet location. Chlorine concentrations generally remain within approximately **0.1–0.3 mg/L**, although higher short-term values occur at locations directly influenced by booster operation. THM concentrations progressively increase during the simulation, reaching approximately **60–65 µg/L** at the illustrated locations. The results demonstrate that the optimized dosing maintains chlorine while limiting the downstream accumulation of THMs. **Figure 70** shows that chlorine at dist412 varies approximately between **0.05 and 0.21 mg/L**, following a repeated daily pattern associated with the optimized booster operation. THM concentrations reach **65 µg/L** but remain below the highest values observed under the baseline scenario. This demonstrates the ability of the optimized strategy to maintain disinfectant residuals while reducing subsequent THM formation.

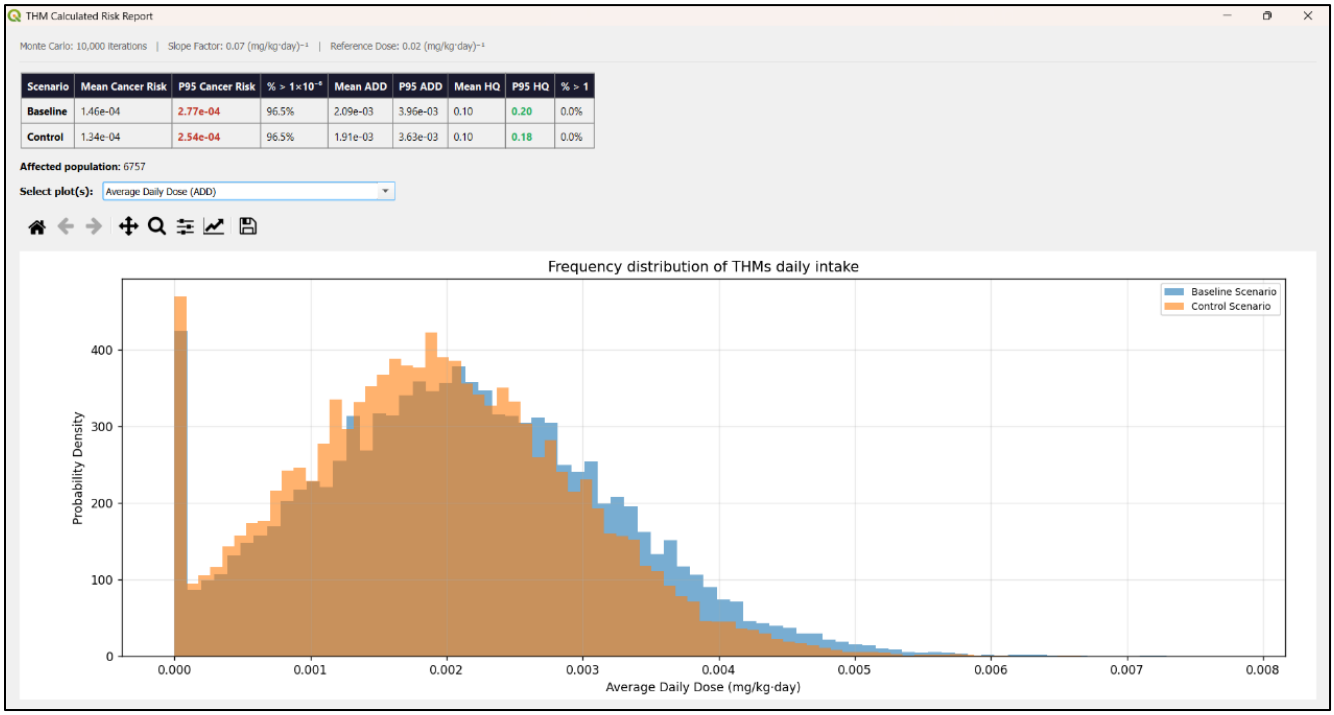


Figure 71: DBP Exposure dashboard indicating the THM Average Daily Dose (mg kg-1 day-1) frequency distribution under (baseline and control scenario)

Figure 71 compares the Monte Carlo THM ADD distributions for the baseline and control scenarios. The control distribution is shifted toward lower ADD values, with the mean decreasing from approximately 2.10×10^{-3} to 1.91×10^{-3} mg/kg-day.

Overall result

Taken together, the figures show that the optimization **redistributes chlorine dosing spatially and temporally and lowers the maximum THM concentration from approximately 70-75 to 65-70 µg/L**. The improvement is also reflected in the DBP exposure assessment, with reductions in ADD. The optimization strategy therefore provides improved control of DBP formation while maintaining chlorine residuals within the defined target.

4.3 Scenarios with uncertainties:

The proposed tool provides the ability to consider uncertainties in all model parameters. Currently, only one parameter with uncertainty can be considered at a time, allowing the user to assess the model's sensitivity to that specific parameter. Below is an example of a scenario involving parameter uncertainty of 50% for the THM yield coefficient, along with the corresponding results. For each simulation, the maximum and minimum bounds for each species are plotted, reflecting the range of variation resulting from the considered parameter uncertainty.

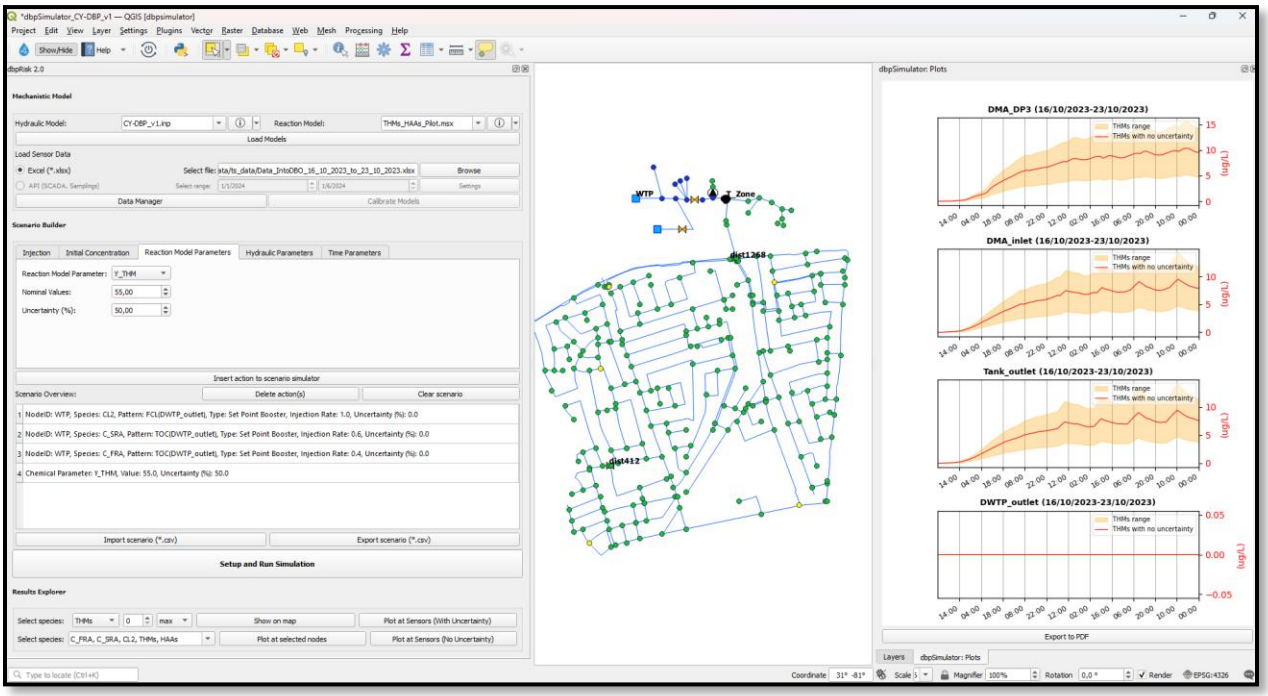


Figure 72: THMs concentration timeseries with uncertainty bounds concerning the THM Yield coefficient