Final Soft-sensors module for water quality monitoring and performance improvement

Deliverable 4.3

WP4 Soft sensors for water quality monitoring and improved water system performance awareness

Funded by the European Union

GRANT AGREEMENT NUMBER	101082035		
FULL TITLE / ACRONYM	ToDrinQ		
START DATE	01-12-2022	DURATION	48 months
END DATE	30-11-2026		
PROJECT URL	www.todrinq.eu		
WORK PACKAGE No and title	WP4 Soft sensors for water quality monitoring and improved water system performance awareness		
DELIVERABLE TITLE	Deliverable D4.3 Final Soft-sensors module for water quality monitoring and performance improvement		
ACTUAL DATE OF DELIVERY	12-12-2025 (1) & 02-04-2026 (2) & 26-05-2026 (3)		
NATURE	R	DISSEMINATION LEVEL	Public
LEAD BENEFICIARY	National Technical University of Athens		
RESPONSIBLE AUTHOR	Prof. Christos Makropoulos		
CONTRIBUTIONS FROM	Vasiliki Thomopoulou, NTUA Panagiotis Kossieris, NTUA Greg Kyritsakas, TUD Siddharth Seshan, KWR George Bariamis, NTUA Prof. Christos Makropoulos, NTUA Prof. Luuk Rietveld, TUD		

Document history

Version	Description	Author	Organisation short name
V0.1	First draft	Vasiliki Thomopoulou Panagiotis Kossieris Greg Kyritsakas Siddharth Seshan George Bariamis	National Technical University of Athens
V0.2	Review	Panagiotis Kossieris	National Technical University of Athens
V0.3	Review	Alain Magis	CEBEDEAU
V0.4	Draft final 2	Panagiotis Kossieris George Bariamis	National Technical University of Athens
V1.0	Final version	Luuk Rietveld	Delft University of Technology

Quality control

Author	Organisation short name	Role	Date
Panagiotis Kossieris	National Technical University of Athens	Deliverable Leader	7-11-2025 & 11-05-2026
Christos Makropoulos	National Technical University of Athens	Work Package Leader	10-11-2025 & 11-05-2026
Alain Magis	CEBEDEAU	Reviewer 1	21-11-2025
Luuk Rietveld	Delft University of Technology	Scientific project coordinator	09-12-2025
Danitsja van Heusden	Delft University of Technology	Project coordinator	12-12-2025
Panagiotis Kossieris	National Technical University of Athens	Deliverable Leader	14-4-2026
Christos Makropoulos	National Technical University of Athens	Work Package Leader	14-4-2026
Luuk Rietveld	Delft University of Technology	Scientific project coordinator	14-4-2026
Danitsja van Heusden	Delft University of Technology	Project coordinator	20-4-2026 & 26-05-206

Abbreviations

ANN		Artificial Neural Network
AOC		Assimilable Organic Carbon
AWD		Amsterdam Supply Dunes
CT		Disinfectant (Ozone) Exposure (Concentration x contact Time product)
CV	-	Computer Vision
DC	-	Demo Case
DCL	-	Demo Case Leader
DWTP	-	Drinking Water Treatment Plant
EC	-	European Commission
EO	-	Earth Observation
ET	-	Evapotranspiration
EU	-	European Union
EWS	-	Early warning system
FAI	-	Floating Algal index
FAIR	-	Facebook AI Research team
GNDVI	-	Green Normalized Difference Vegetation Index
HEC-HMS		Hydrologic Engineering Center, Hydrologic Modeling System
ISA	-	Intelligent Spectral Analyzer
ISRIC	-	International Soil Reference and Information Centre
LAI	-	Leaf area index
LE	-	Latent Heat flux
LSTM	-	Long Short-Term memory
MAE		Mean Absolute Error
MCI	-	Maximum Chlorophyl Index
ML	-	Machine Learning
MLP	-	Multi layer Perceptron
MSA		Mean Squared Error
MSI	-	Multi Spectral Instrument
NCEP	-	National Centers for Environmental Prediction
NDAI	-	Normalized Difference Algae Index
NCDI	-	Normalized Difference Chlorophyll Index

NDRE - Normalized Difference Red-Edge Index
NDTI - Normalized Difference Turbidity Index
NDVI - Normalized Vegetation Index
NIR - Near InfraRed spectrum
NNLI - Normalized Nutrient Load Index
NOAA GFS - National Oceanic and Atmospheric Administration - Global Forecast System
NSE Nash–Sutcliffe Efficiency
PC - Project Coordinator
PET - Potential Evapotranspiration
PINN Physics Informed Neural Network
PLE - Potential latent heat flux
RAE Relative Absolute Error (RAE).
RECI - Red-Edge Chlorophyll Index
RF - Random Forest classifier
RMSE Root Mean Squared Error
SCADA - Supervisory Control and Data Acquisition
SITS - Satellite Images Time Series
SMA - Soil Moisture Accounting
SWIR - Short-wavelength infrared spectrum
TSS Total suspended solids
VCI - Visual Cyanobacteria Index
WP - Work Package
WPL - Work Package Leader
WQI - Water Quality Index

Short name	Legal Name
TUD	TECHNISCHE UNIVERSITEIT DELFT
RWTH	RHEINISCH-WESTFAELISCHE TECHNISCHE HOCHSCHULE AACHEN
CEB	CENTRE BELGE D'ETUDE ET DE DOCUMENTATION DE L'EAU
NTUA	ETHNICON METSOVION POLYTECHNION
KWR	KWR WATER BV
WTNT	STICHTING WATERNET
HWL	HET WATERLABORATORIUM NV
VEF	VEOLIA EAU - COMPAGNIE GENERALE DES EAUX SOCIETE EN COMMANDITE PAR ACTIONS
VEOCZ	VEOLIA CESKA REPUBLIKA, A.S.
EYDAP	ETAIREIA YDREYSEOS KAI APOCHETEFSEOS PROTEYOYSIS ANONIMI ETAIREIA
OLI	OLISENS TECH
OXY	OXYMEM LIMITED

Short name	Legal Name
ORV	ORVION B.V.
CHEM	CHIMIKI TECHNOLOGIA P. DIMOPOULOU -P.TAZES & SIA OE
WE	WATER EUROPE
ALTIS	ALTIS Groupe SA
BNV	BNOVATE TECHNOLOGIES SA
VEHO	Veolia Holding Ceska Republika AS

Table of contents

List of figures

List of tables

Executive summary

The present deliverable (D4.3) reports progress on Tasks 4.1, 4.2 and 4.3 of the ToDrinQ project, focusing on advancing water quality monitoring and performance improvement through the development of soft sensors, leveraging innovative data fusion techniques, machine learning, and stakeholder collaboration. These tasks were implemented in parallel with key technical developments but also in collaboration with demo cases (DC#1 Amsterdam, DC#2 Athens and DC#3 Val de Bagnes) as expertise, information and data providers.

In total **we identified ten (10) soft sensor use cases** (at catchment and water treatment plant levels of water supply chain) including early detection of nutrient runoff in Yliki Lake (**DC#2**) and bacteriological contamination in Sarayer (**DC#3**), estimation of chlorophyll-a concentrations, prediction of algal blooms, and the Water Quality Index (WQI) at pixel-level resolutions (DC#2). Additionally, we developed soft sensors to optimize water treatment processes by predicting turbidity events at treatment plant inlets (**DC#1**), reducing chemical consumption in coagulation processes, and estimating ozone exposure in ozonation tanks.

The modules provide a suite of powerful, data-driven tools validated across three of the ToDrinQ key demonstration cases: Amsterdam (DC#1), Athens (DC#2) and Val de Bagnes (DC#3). These tools are tailored to specific, high-priority utility challenges:

- **For Catchment & Source Protection (Athens & Val de Bagnes):** A set of seven sensors provides enhanced situational awareness. This includes an Early Warning System (EWS) for nutrient runoff in the Yliki Lake catchment, leveraging Earth Observation (EO) data, complemented by sensors for Chlorophyll-a, pH, Dissolved Oxygen (DO), Algal Bloom Probability, and an integrated Water Quality Index (WQI). For the Alpine springs of Val de Bagnes, an EWS for bacteriological contamination was developed, linking meteorological data to contamination events.
- **For Drinking Water Treatment Plant Optimization (Amsterdam):** Three sensors were developed to improve treatment process efficiency. These include a sensor for turbidity prediction in the coagulation-flocculation process, an EWS for turbidity events at the Drinking Water Treatment Plant (DWTP) inlet, and a predictive model for ozonation (CT) exposure.

The development of soft sensors has made significant progress beyond state-of-art across various applications, highlighting also both successes and areas for improvement as planned.

The findings from these ten sensors provide a robust, evidence-based foundation for integration into the WP7 modular platform under Deliverables 7.2 and 7.3.

Between M24 and M36 all the soft-sensors progressed from beta version (in D.4.1) to completed alpha versions (D4.3). The improvements and changes made, include activities such as

a) Resolution of known issues and bugs identified during beta testing
b) testing of different model architectures and comparisons (such as GRU, LSTM, and MLP, or ARIMA-X against tree-based models)
c) extension of training datasets and validation periods to include more recent operational data
d) replacement of manual neural network architectural selection with automated methods, mainly Byesian Hyperparameter Optimization.
e) Input pipelines were redesigned to use raw satellite bands, instead of indices.

f) Methods for handling imbalanced data, Explainable AI (XAI) tools, and uncertainty quantification were also added.

Components that were conceptualized at the beta version were technically matured during the M24-M36 period including the WQI soft-sensor and the additional development of its components; pH, the DO soft-sensors, as well as the bromate prediction in the ozonation outlet.

The overall progress and technical maturity of the developments are presented in the table below

SoSE #	Name	Delivered in D4.1 (Beta Version)	Delivered in D4.3 (final Version)	Progress made within the M24-M36 period
1	**Early Warning System for Nutrient Runoff**	• Initial U-Net based crop type segmentation models, NNLI formulation, mapping of agricultural practices, HEC-HMS model. • Beta version of the patch reconstruction pipeline (contained stitches artifacts)	Operational early warning system with corrected patch reconstruction approach.	• An issue in the patch reconstruction process that was used to deliver the full segmented map was identified and fixed between M23-M36. • The artifacts presented in the beta version of the crop type map were eliminated. • The logic of the soft-sensor was preserved (U-Net segmentation architectures, loss function setup, NNLI formulation, agricultural practices mapping, and HEC-HMS hydrological modelling) from the beta version. The updates between M24-M36 do not alter the underlying scientific approach and preserve spatial consistency.
2	**Chl-a concentration estimation**	• Beta Chl-a soft-sensor based on spectral indices and feature selection based on correlation. • Manually selected architectures and aggregated performance metrics.	• Final Chl-a soft-sensor. Receives 10 raw bands of Sentinel-2 data as inputs. • Imbalance handling (inverse-frequency weighting with square-root smoothing. Bayesian hyperparameter optimization, controlled experiments between imbalance and baseline models. • Permutation Feature Importance XAI pipeline	• The input set was redesigned to include 10 raw Sentinel-2 bands. • Spectral indices and correlation-based selection were removed. • Inverse frequency weighting with square-root smoothing for imbalance data handling was introduced. • Manual architecture selection for the Neural Network (NN) development was discarded. Bayesian Hyperparameter Optimization (HPO) was introduced. • The residuals were examined and compared between the imbalance handling and benchmark architecture using per-bin error analysis. • Explainable AI (XAI) methods were introduced. Global Model Agnostic XAI methods (Permutation Feature Importance (PFI)) were utilized to explain and validate model predictions based on the underlying physical relationships

SoSE #	Name	Delivered in D4.1 (Beta Version)	Delivered in D4.3 (final Version)	Progress made within the M24-M36 period
3	**pH estimation**	Not part of the beta version	• pH soft-sensor predicting non-optically active variables from 10 Sentinel-2 raw bands. • The imbalance handling model was compared with the baseline model.	• A new component for the EYDAP demo case was introduced (pH estimation). • Designed and trained two deep MLP architectures in a controlled experimental framework (baseline model and imbalance handling model). • The models were optimized using Bayesian HPO. • Conducted per-bin error analysis in addition to aggregate metrics to evaluate the behaviour of the model across the pH values range.
4	**DO estimation**	Not part of the beta version	• DO soft-sensor predicting non-optically active variables. • Deep MLP architecture (derived from Bayesian Hyperparameter Optimization)	• A new component for the EYDAP demo case was introduced (DO concentration estimation). • Designed and trained a deep Multi-Layer Perceptron (MLP) with 11 fully connected layers with the configurations extracted from Bayesian HPO. • The performance and generalization of the model was evaluated
5	**Bloom Occurrence Probability Estimation**	Beta soft-sensor based on Categorical Naive Bayes classifier, using meteorological 2012-2023 meteorological time window.	• Soft-sensor based on GRU encoder-decoder Neural Network, producing 3-day forecasts at daily temporal resolution, using raw NOAA GFS meteorological data as inputs.	• Replaced the Categorical Naive Bayes classifier used in D4.1 with a GRU encoder-decoder architecture. • Multiple architectures were tested such as MLP, and LSTM and compared with the GRU approach. • The input pipeline was redesigned, using sequential NOAA GFS meteorological forecasts. • The MODIS FAI prior probability information is used to provide the spatial conditioning to the model. • The training data window was updated from 2012-2023 to 2020-2025. • The teacher forcing training method was used for GRU model training.
6	**Water Quality Index**	Proposed methodology. The soft-sensor was described conceptually. An MLP-based WQI model was planned. The Input list included Electrical Conductivity.	• The WQI was fully implemented, based on the results of the pH, DO and Chl-a soft-sensors.	• The WQI soft-sensor fully implemented, moving from the conceptual idea to full implementation, building on-top of the three upstream soft-sensors (pH, DO and Chl-a) • The parameter list was refined by removing Electrical Conductivity (EC) • The methodological framework was modified. Instead of training an MLP to directly predict the WQI, the index is estimated as the weighted aggregation of the three soft-sensor outputs.
7	**Early Warning System for Bacteriological Contaminati**	The beta version of the early warning system was conceptually designed, based on the workflow presented in Soft-	• The soft-sensor was redesigned. The final version includes a direct prediction problem, using BactoSense data, ERA-5 reanalysis data, and	• The Early Warning system for bacteriological contamination was redesigned from the beta version. The problem in the final version is framed as the direct prediction of the time-series of bacteriological contamination level, using meteorological variables. • The input pipeline was assembled and pre-processed based on the BactoSense in-situ data (ICC and HNAP),

SoSE #	Name	Delivered in D4.1 (Beta Version)	Delivered in D4.3 (final Version)	Progress made within the M24-M36 period
	on	Sensor #1.	ALTIS meteorological data. ● Cross-correlation feature selection was implemented and a two-stage modelling approach comparing regression and classification was implemented. Gini impurity feature importance was used for explainability	ERA-5 meteorological reanalysis data. ● In-situ meteorological data provided from ALTIS was used to validate the ERA-5 data input. ● Cross-correlation features selection was implemented. Snow Cover, Temperature, Precipitation, Snow Melt, Snow Depth and Runoff were selected as the predictors. ● Two modelling approaches were tested. The first treated the problem as a regression task (3 models were compared ARIMA-X, LSTM, and GRU). The second reframed the problem as a classification task and was trained using Random Forest classifier (in 2-class and 3-class targets), ● Gini impurity feature importance was selected and applied as the XAI method to interpret the classifiers.
8	**Estimation of turbidity of the coagulation-flocculation process**	● Beta version of the soft sensor was developed based on the requirements identified from discussions with the end-users during Task 4.1. ● Beta version soft sensor developed with available data and conducting hyperparameter optimization to train various models.	● Final (alpha version) soft sensor was delivered which was an improved version compared to the beta release, with a higher prediction accuracy (Beta version Test R2 score: 0.42 - 0.47; Alpha version: 0.62-0.65) ● An uncertainty analysis on the alpha version soft sensor was conducted and reported.	● Changes were made to the features used for model training. Namely, two new features were added, days after dredging and the year, based on process insights gained from the end-user. ● New models were trained and the routine for identifying the best combination of hyperparameters was conducted. ● More accurate performing models were identified and reported. ● A detailed uncertainty analysis was conducted for two types of uncertainties, namely - epistemic uncertainty, and aleatoric uncertainty. The results of this analysis were provided in D4.3
9	**Early prediction of turbidity in DWTP inlet.**	The beta version of model was predicting turbidity events as a classification problem with a maximum prediction of up to 3 hours ahead.	The last version is a regression model that predicts actual turbidity measurements up to 6 hours ahead.	• The variable analysis conducted in the beta model included significantly fewer input variables compared to the final model version. • Feature analysis was expanded to investigate multicollinearity among a broader range of available source water quality and environmental variables. • Additional analyses were performed to further improve the model's predictive efficiency and overall performance. • Further correlation analysis was carried out to better understand the relationship between upstream water quality deterioration and turbidity levels at the drinking water treatment plant (DWTP).

SoSE #	Name	Delivered in D4.1 (Beta Version)	Delivered in D4.3 (final Version)	Progress made within the M24-M36 period
10	**Prediction of the ozonation exposure, CT, to improve the ozonation process.**	The beta version is an initial model for predicting ozone exposure (CT). Methodology is set and preliminary refinement was made.	The final version enhances the CT prediction analysis and additionally it predicts the bromate concentration in the ozonation outlet enabling the monitoring and management of the ozone process.	• The beta version primarily provided the conceptual design of the model, whereas the final version focused on refining the framework through a comparative evaluation of multiple machine learning algorithms, including an improved version of a Physics-Informed Neural Network (PINN). • The updated model version predicts bromate concentration at the outlet, a functionality that was not included in the beta version. • The final version includes an interactive dashboard that continuously displays predicted CT values and bromate concentrations, benchmarked against the minimum required thresholds to ensure the efficiency and reliability of the ozonation process.

1 Introduction

The EU's recast Drinking Water Directive advocates a risk assessment and management approach from source to tap, aiming for high drinking water standards. The recast Drinking Water Directive[1] (effective from January 2021) is the European Union's legislation for water intended for human consumption with the overarching goal of protecting human health. The directive covers all types of water used for drinking, cooking, food preparation or other domestic purposes regardless of its origin (distribution network, from tankers, put into bottled or containers, also including spring waters). Key aspects of the revised directive include up-to-date water quality standards; emerging pollutants such as Per- and Polyfluorinated substances (PFAS), and microplastics; a preventive approach to reduce pollution at its source; initiatives to improve water access for vulnerable groups; efforts to promote tap water usage to decrease plastic bottle consumption; harmonization of standards for materials in contact with water; and actions to reduce water leakages and enhance sector transparency.

For the purposes of the Drinking Water Directive, water intended for human consumption must meet the following minimum requirements:

- **Water is free from any micro-organisms and parasites and from any substances** which, in numbers or concentrations, constitute a potential danger to human health.
- Water meets the minimum requirements set out in Parts A (**microbiological parameters**), B (**chemical parameters**) and D (**parameters relevant for the risk assessment of domestic distribution systems**; *Legionella* and *Lead*) of Annex I.
- Member States have taken all other measures necessary to **comply with Articles 5 to 14**, emphasizing a risk-based approach to water safety, covering the entire water supply chain from source to consumer. This includes risk assessments and management of catchment areas, supply systems, and domestic distribution systems, with specific deadlines for implementation.

Annex I also include Part C which catalogues the indicator parameters in water which do not directly affect public health, but they are essential for assessing the performance of production and distribution systems for drinking water. They help in evaluating water quality, identifying any shortcomings in water treatment, and play a crucial role in the quality of their water. Consequently, it's important for Member States to monitor these parameters to ensure the continued safety and reliability of drinking water.

According to Art 13 of the DWD "*Member States shall take all measures necessary to ensure that regular monitoring of the quality of water intended for human consumption is carried out in accordance with this Article and Parts A and B of Annex II, in order to check that the water available to consumers meets the requirements of this Directive and in particular the parametric values set in accordance with Article 5. Samples of water intended for human consumption shall be taken so that they are representative of its quality throughout the year*. This requirement **indicates an ongoing, year-round commitment to monitoring**.

Monitoring programs ensure compliance with drinking water standards safeguarding the water users. Analysis of this data aids legislators and water managers in evaluating the success of current water policies while also supporting trends in drinking water quality parameters, leading to the development of new strategies.

The monitoring process involves systematic **collection and analysis of water samples from various sources where raw or treated water is stored or processed** (including reservoirs/lakes, rivers, aquifers,

[1] DIRECTIVE (EU) 2020/2184 on the quality of water intended for human consumption, recast

water treatment facilities, and water distribution systems). The primary goal is **to detect and measure the presence of chemical** (e.g., dissolved oxygen, nutrients, alkalinity, heavy metals, organic micro-pollutants, other dissolved salts, pH etc.), **biological** (pathogenic bacteria, viruses and protozoa, algae, and waterborne plants), **and radiological substances that could potentially pose health risks to users**.

The methods and frequency of water quality monitoring can vary depending on the type of water resource, remoteness of the sampling locations, access to funds and advanced technological tools and the potential risks associated in case of contamination event.

Hard (physical) sensors, remote sensing, earth observations and automated sampling stations are the set of choices available for this purpose and are increasingly employed alongside established laboratory analyses to provide (near) real-time data feed and early warnings (in case of critical events). This approach enables water operators and national authorities to respond promptly to contamination events safeguarding public health and maintaining public confidence in the water supply with regular reporting on water quality performance indicators.

1.1 Soft sensors for water quality monitoring from source to tap

The global drinking water supply faces significant challenges, such as climate change, emerging pollutants, demographic shifts, and aging infrastructure and workforce. To protect water from source to tap, a proactive approach is needed to enhance the resilience of drinking water systems. Soft (Software) sensors, also known as virtual or inferential sensors, can play a significant role in monitoring drinking water quality. These sensors use algorithms to estimate parameters that are difficult, costly, or impossible to measure in a direct way.

Soft Sensors can be very helpful when employed as real time monitoring systems (Juntunen et al., 2013) are cost effective (Djerioui et al., 2019) since they can infer water quality characteristics from already available data using computational methods/models. They can also support predictive analysis and estimate a wide range of parameters like pH (Bresciani et al., 2019), turbidity (Elhag et al., 2019), bacterial content (Mohammed et al., 2018), and other pollutants (Kapalanga et al., 2021). Soft sensors by their nature can integrate data, information and knowledge from various sources while remaining adaptable (e.g., during changing standards) and accessible especially when they employ earth observation information.

The parameters that the soft sensors are developed to estimate are also called target variables (or predictands). The measured variables that are used to predict the target variable are called input variables (or predictors). The estimated parameters (target-parameters) differ based on the needs, the data availability, and the system characteristics of each case study.

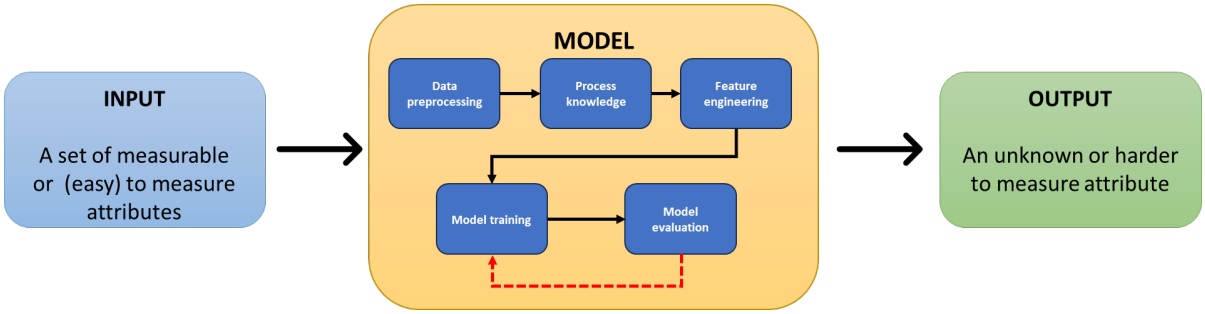


Figure 1: Schematic representation of soft sensor

The use of soft sensors for drinking water quality monitoring can fully cover the water supply chain, from catchment (water availability) to tap (water consumption) when utilized with proper adjustments, adequate data for training and regular fine tuning (Figure 2). It is worth noting that catchment functions along with the distribution networks have the largest spatial coverage enabling them as active test beds for soft sensors development to monitor water quality and enhance potable water distribution.

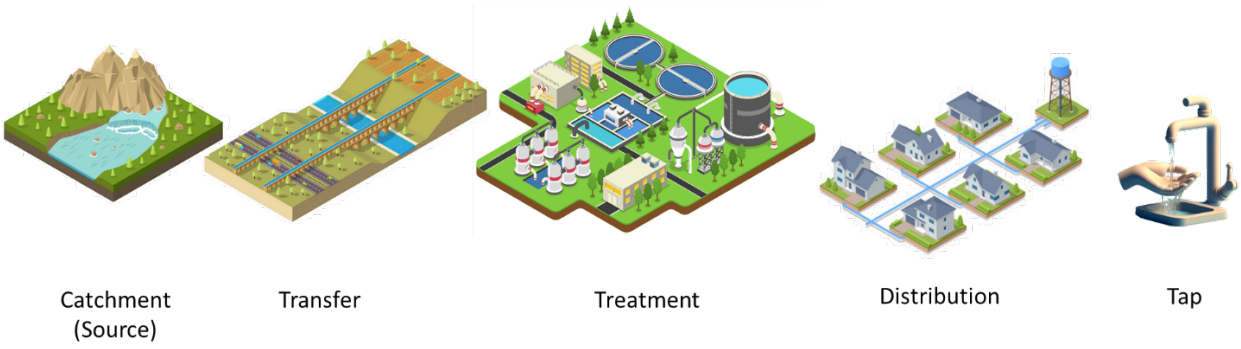


Figure 2: Drinking water supply chain from catchment to tap user.

1.2 Soft sensors at source/catchment level

Catchments play a critical role in moderating water quality and quantity. Catchments via ecosystem services provide (among others) storage, conveyance, natural filtration, removing pollutants and excess nutrients from water before it enters streams, lakes, and large tributaries. Healthy watersheds can contribute to providing reasonably clean raw water. However, apart from natural processes and functions, anthropogenic drivers exert pressures to components of catchments in surface and groundwater bodies impacting water availability and water quality. Anthropogenic activities like agriculture, industrial production, mining industries, urban development and energy production are the major anthropogenic drivers directly affecting water quality conditions and flow regime. The most-commonly predicted target variables on soft sensors and models' development are water flow (discharge) groundwater storage (level measurements), pH, stream temperature, turbidity, and Chl-a concentrations (especially in natural lakes).

Surface water discharge and groundwater level monitoring are important parameters indicating the overall water availability in the boundaries of the catchment. Monitoring both surface water discharge and groundwater levels is crucial for identifying the dynamics of a water, a process that is highly dependent on elaboration of coupled (or not) hydrological and hydrogeological models. Such models can

be regarded as soft sensors since they are estimators of water flow and groundwater storage also in places/subbasins where physical/hard sensors are not established. Surface water discharge monitoring is typically conducted using a variety of methods, including stream gauges, weirs; devices measuring the flow rate of water passing through a particular location. Groundwater level monitoring is typically done using wells, which are drilled to access the aquifer. Water level sensors are placed in the wells to measure the depth of water.

As far as the catchment functions are concerned, studies highlight the interlinked relationship between catchment activities, such as crop cultivation and livestock practices, and water quality (Leip et al., 2015). Agricultural runoff, including sediment, nutrients, and pesticides, can significantly impact the composition of open water bodies (Gensemer et al., 2018).

More specifically, extensive agricultural and livestock activity can cause the occurrence of eutrophic phenomena at the downstream waterbodies. The main components of the fertilizers are nutrients, mainly nitrogen and phosphorus. The same nutrients are also found in animal waste products (faeces and urine). Nitrogen and phosphorus in general are natural parts of the aquatic ecosystem. An increase in their concentration though causes increases in Chl-a concentration and thus (sometimes) floating algae.

Monitoring pH in flowing rivers and in open water bodies is crucial for assessing the acidity of aquatic environments, providing valuable information about water quality and potential environmental impacts. The importance of pH monitoring has been highlighted in many studies, as it influences many chemical and biological processes in aquatic ecosystems. Decreasing pH levels can affect the overall health of aquatic organisms (Gensemer et al., 2018). In open water bodies, continuous monitoring of pH is essential to identify potential sources of contamination and to understand the system's resilience to external factors. Literature reflects the use of both traditional in-situ measurements (Fadillah Rahmat et al., 2016) and automated sensor networks for real-time pH monitoring, contributing to a better understanding of dynamic changes in water chemistry.

Monitoring of stream temperature is crucial as it affects the ecosystem by influencing many chemical and biological processes. Stream temperature is affected by changing precipitation patterns, solar radiation, turbidity, and other factors. When the temperature rises, water can hold less dissolved oxygen, living aquatic organisms die and thus the ecosystem and the water quality are degraded (Ahearn et al., 2004) (Hamid et al., 2020).

Turbidity monitoring in surface water bodies is a critical aspect of environmental research and management, as it provides valuable insights into water quality and ecosystem health. A comprehensive literature review shows a growing interest in research focused on the development of different turbidity monitoring techniques (Silva et al., 2022). Traditional in-situ methods, such as nephelometry and turbidimetry, have been extensively employed for their accuracy, yet limitations persist in terms of spatial coverage and real-time data acquisition. Recent studies highlight the integration of satellite imagery as a promising approach for large-scale turbidity assessment (Hafeez et al., 2019). Satellites equipped with remote sensing instruments, such as multispectral or hyperspectral sensors, offer the capability to capture spatial and temporal variations in water turbidity over extensive areas like a big lake or a dam. Case studies employing satellite data showcase successful applications in monitoring turbidity levels in diverse open water bodies, including lakes, rivers, and coastal zones (Bresciani et al., 2019) (Feng et al., 2020). The use of both in-situ measurements and satellite-based techniques emerges as a good strategy for enhancing understanding of turbidity dynamics in open water ecosystems.

Chlorophyll-a (Chl-a) concentrations monitoring in surface water bodies is a crucial component of assessing aquatic ecosystem health also affecting treatment processes, as it serves as a proxy to quantify algal biomass (He et al., 2022). Chl-a is directly related to photosynthetic activity as it provides insights

into nutrient availability and the potential for eutrophication. In open water settings, Chl-a concentrations offer valuable information about algal blooms and their environmental implications. The integration of satellite imagery has significantly advanced Chl-a monitoring, allowing for a comprehensive view of large water bodies and facilitating the analysis of spatial and temporal variations (Cao et al., 2020). Various studies highlight the effectiveness of remote sensing techniques, such as multispectral and hyperspectral sensors, in estimating Chl-a concentrations (Olivetti et al., 2023). These approaches, combining satellite data with traditional in-situ measurements, present a practical tool for researchers and environmental managers to better understand algal growth dynamics in expansive aquatic areas.

1.3 Soft sensors at conveyance system level

The conveyance system plays a critical role in the overall water supply chain since it connects the water sources (from abstraction/uptake points) with the Drinking Water Treatment Plants (DWTP). Soft sensors, employed for estimating water quality within the conveyance system, play a key role in predicting the quality of water that eventually reaches DWTP prior to any treatment stage. These sensors can also enable notifications to DWTP operators regarding potential extreme events (disrupting quantities of water) or serious contaminants. Furthermore, implementing a soft sensor in the aqueduct can aid in identifying leaks or spills of substances that could compromise water quality (Yang, 2012).

1.4 Soft sensors at Drinking Water Treatment Plant level

In the DWTP facility, soft sensors play a fundamental role in enhancing the efficiency of water purification by estimating water quality before (raw), during and after treatment (potable). They go beyond merely detecting harmful substances and high concentrations of unwanted elements since they have the capacity to optimize e.g., the chemical dosages employed in the treatment process. Soft sensors can also be developed to calculate process parameters that are not possible to measure or observe in water treatment processes, such as the backwash efficiency during membrane treatment processes, and adsorption capacity of an activated carbon installation. Moreover, soft sensors facilitate timely notifications to operators regarding issues such as filter clogging, eliminating the necessity for the installation of additional hardware sensors. Hence soft sensors can assist in providing advanced understanding of treatment processes during operations and be used as tools in conducting preventing and predictive maintenance of physical assets. Although sensors are not explicitly mentioned during drinking water treatment, implicitly they are frequently used. For example, the head loss over a filter can be seen as an indirect sensor for the level of clogging and thus the time for backwashing; the dosing of chlorine or ozone is an indication for the (logarithmic) removal of pathogenic micro-organisms; the run time of an activated carbon filter can be an indication for the breakthrough of organic micro-pollutants. At research/pilot scale level more soft sensors have been developed to determine calcium pH after softening, using influent data process models; to estimate assimilable organic carbon (AOC) concentrations after ozonation by using differential UV/Vis spectra; to estimate the state of a fluidized bed using differential head loss measurements.

2 Identifying Needs and Designing Soft Sensors (relates to T4.1)

The objective is to conduct a comprehensive needs assessment and requirement analysis for water quality monitoring using soft sensors, from the source to tap. Key parameters like pH, turbidity, dissolved oxygen, and conductivity are critical for maintaining drinking water quality. This process involved engagement with the demo cases and determining soft sensor requirements (accuracy, response time, operational conditions etc) to ensure reliable estimation performance.

2.1 Review on soft-sensors applications

2.1.1 Desk study conducted on remote sensing augmentation for water quality, source to tap

The consortium critically reviewed over 150 publications on soft sensor applications, focusing on various stages of water quality improvement. Each partner contributed specific expertise:

- NTUA addressed water quality at catchment and conveyance levels with an emphasis on remote sensing.
- TUD focused on optimizing drinking water treatment processes.
- KWR investigated hybrid modelling for improved treatment performance.

The desk study included a detailed literature review on remote sensing for water quality monitoring. The outcome was a comprehensive catalogue of soft-sensor applications that organized various water quality target variables with predictor variables. This cataloguing considered spatial and temporal resolutions, aiding the integration of remote sensing technologies to provide a continuous, high-resolution monitoring system throughout the drinking water supply chain.

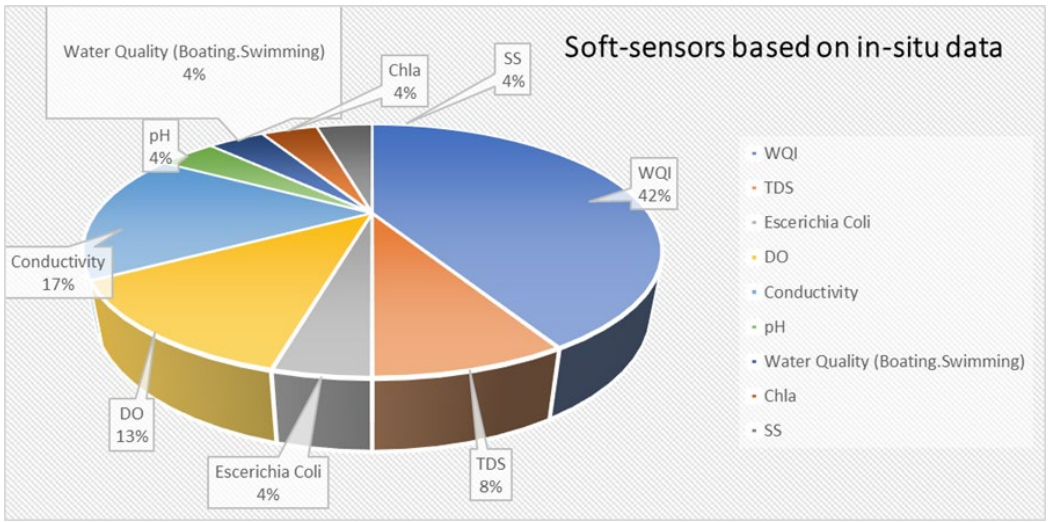


Figure 3: Review of Soft Sensor Use Cases Mapped Based on In-Situ Data Inputs

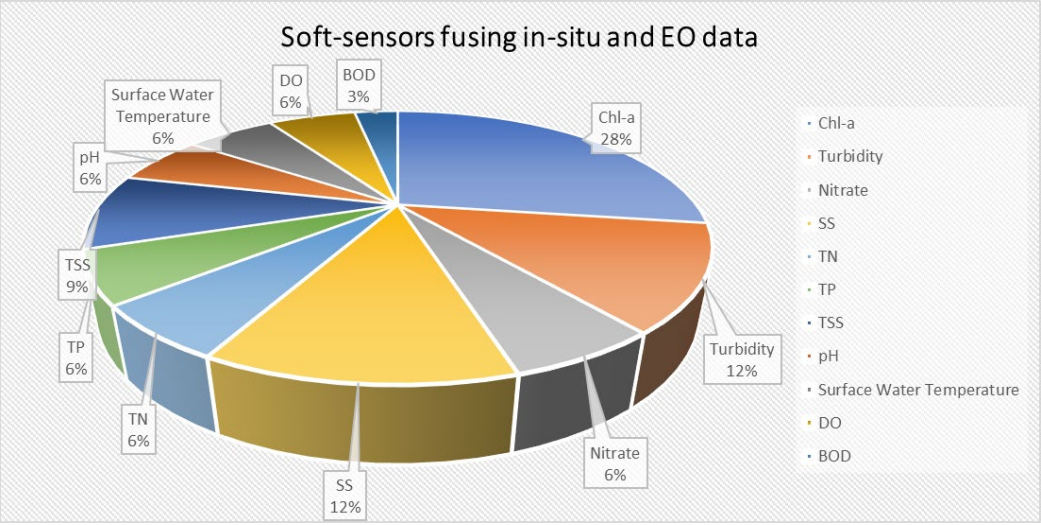


Figure 4: Review of Soft Sensor Use Cases Mapped Based on In-Situ and earth observation (EO) Data Inputs

2.1.2 Overview of identified soft sensor applications

Table 1 provides an overview of the soft sensors use cases identified and next developed in the context of WP4 activities for each DC. The table provides for each DC, the variable that soft-sensor predicts, as well as the temporal and spatial scale of its implementation.

Table 1: Soft sensors use cases for each demo sites.

Soft Sensor No	DC	Soft Sensor Title	Aim (short description)	Hard-to-measure' compound	Variables	Time scale	Spatial Scale
1	EYDAP	Early Warning for nutrient runoff in the Yliki Lake	Provide an early warning system for nutrient runoff for in Yliki lake	Nutrient load in the basin and the nutrient load concentration increase in the lake	Crop type maps, Normalized Nutrient Load Index (NNLI), NOAA GFS precipitation, MODIS evapotranspiration, In-situ discharge, In-situ precipitation, HiHydrosoils-soil properties dataset	Daily	20m (crop type map), 500m (MODIS Evapotranspiration), 27830 meters (NOAA GFS), Point (In-situ precipitation and discharge), 250 m (HiHydrosoils)
2	EYDAP	Chlorophyl-a estimation	Provide Chl-a concentration estimations in frequent time intervals (3-5 days) at the pixel level (10m)	Chl-a (μg/l)	1.Sentinel-2 reflectance data 2. In-situ Chl-a measurements from ISAboat (EYDAP)	3-5 days	10m
3	EYDAP	pH estimation	Provide pH estimations in frequent time intervals (3-5 days) at the pixel level (10m)	pH	1.Sentinel-2 reflectance data 2. In-situ pH measurements from ISAboat (EYDAP)	3-5 days	10m
4	EYDAP	DO estimation	Provide DO concentration estimations in frequent time intervals (3-5 days) at the pixel level (10m)	DO (mg/l)	1.Sentinel-2 reflectance data 2. In-situ DO measurements from ISAboat (EYDAP)	3-5 days	10m
5	EYDAP	Bloom Occurrence Probability	Estimation of the Floating Algal bloom occurrence probability at pixel level	FAI	1.prior bloom occurrence probability at pixel level (through historical MODIS FAI timeseries from 2008 to 2018) 2. Meteorological variable from NOAA GFS (forecast)	Daily	250m

Soft Sensor No	DC	Soft Sensor Title	Aim (short description)	Hard-to-measure' compound	Variables	Time scale	Spatial Scale
6	EYDAP	Water Quality Index (WQI)	Estimation of the Water Quality Index (WQI) in the Yliki lake at pixel level	WQI (a thumbnail of the overall water quality condition)	Soft-sensor 2, 3 and 4 outputs	3-5 days	10m
7	ALTIS	Early warning of Bacteriological contamination in Sarayer	Provide an early warning for bacteriological contamination in Sarayer	Bacteriological contaminants	LULC maps, pasture area, and pasture patterns, discharge data, EO precipitation, Bacteriological load in the basin	Daily	Point
8	WINT	Estimation of turbidity of the coagulation-flocculation process	Enhance operation of water treatment by predicting outcomes, and thereby reduce consumption of chemicals	Organic material and phosphate removal	Turbidity at inlet, temperature, pH, volume, ferric chloride added	hours	not applicable
9	WINT	Early prediction of turbidity in DWTP inlet	Identification of turbidity events in the water entering DWTP	-	Inlet turbidity, Inlet flow, Inlet water temperature, inlet pH, river discharge in 2 different points	Hourly data	not applicable
10	WINT	Prediction of the ozonation exposure (CT) to improve the ozonation process	Soft sensor for the estimation of the ozone exposure in the ozonation tank	Ozone concentration in the ozonation tank	Ozonation inlet: flow, ozone dosage, water temperature, UV Rapid sand filtration outlet: turbidity Ozonation outlet: UV	hourly data - except of daily data in the UV outlet	not applicable

3 Identifying and Processing Data Sources for Soft Sensors

The data sources are of great importance for developing a soft sensor. Data availability determines the type of approach that will be employed for their development (training machine learning (ML) model requires many measurements to have reliable estimations), but also how to approach and monitor the variable of interest (if for example satellite data can be used, the temporal resolution of estimation will be limited by that of the satellite).

Data sources for the development of soft-sensors are of various types as they can either come from in-situ measurements from **hard sensors** (e.g., installed meters), **ex-situ measurements** (measurements that are collected in-situ and analysed in the laboratory), **earth observation data** (from satellites) or even **legacy systems or other databases** that will help identify patterns and train the ML models.

Each data source is characterized by the variable it measures, the frequency with which it measures this variable and the spatial resolution that can be either point or gridded measurement.

3.1 Data from in-situ measurements

A monitoring station is a device employed to sense a physical phenomenon, generating a signal (the output) upon detecting a change in its near environment. An effective sensor exhibits insensitivity to variables other than the one it was designed to measure and must not have any influence on the measured property. In the water quality monitoring in-situ measurements can be installed at the source, at the conveyance system, at the DWTP, at the distribution network and finally at the tap level.

At the catchment level, monitoring stations measure variables related to a wide range of complex phenomena arising from hydrological and physicochemical processes. Measuring only one variable in an open water body in most cases is not adequate for understanding the underlaying processes and its overall condition. Hard sensors at the source level can monitor contamination from foreign sources like nutrients (phosphorus and nitrogen) from the upstream catchment or measure variables related to the characteristics of the eater body itself, like turbidity.

In the conveyance system, sensors can be strategically installed at points of interest to monitor water quality parameters as the water flows from the water source to the DWTP. These sensors measure variables, such as turbidity levels, that can facilitate the treatment process by updating the DWTP operators about the incoming water, while, at the same time, with the installation of several sensors in the conveyance system, problems or faults in the water transport process can be detected.

Sensors are an integral part of the drinking water treatment process. Drinking water treatment is a complex process with many treatment steps in series and with many variations, as the treatment process can be modified depending on the quality of the water that enters the treatment plant. A typical conventional surface water treatment plant, e.g., involves coagulation, flocculation, sedimentation, filtration, and disinfection. In these treatment steps sensors are important for determining the needed dosage of the added chemicals, such as alum and chlorine, to determine the sludge levels in sedimentation tanks, and to Estimate filter clogging for optimizing of the backwash process. Sensors in DWTP are not only used to monitor chemical processes and the filters' state, but they also monitor other water quality related parameters, such as turbidity, UV254 absorbance, temperature, pH, electrical conductivity and dissolved oxygen to assess the overall performance of the DWTP.

Distribution networks could also be monitored to ensure the quality of the water that reaches our taps. The most important sensors for a water distribution network are:

- Temperature sensors, as the water temperature is a principal parameter for physical, biological, and chemical processes in the water distribution network.
- Level sensors, that notify the operator when the water level in the storage tanks are below a certain threshold.
- Water flow sensors provide accurate flow rate and direction information.
- hard sensors in homes, varying from pH, temperature, to electrical conductivity (simple strips or advanced electronic sensors).

Another information source that can be employed for the development of soft sensors is data information from Supervisory Control and Data Acquisition (SCADA) or other legacy systems. SCADA are systems of hardware and software elements for controlling industrial processes such as water purification/treatment processes. It could be described as a distributed computer system that facilitates the monitoring and management of processes.

SCADA systems provide real-time data acquisition and visualization of critical parameters throughout the treatment process, such as water flow rates, pressure levels, chemical dosages and can also automate various control processes, such as adjusting pumps, regulating chemical dosing, and optimizing treatment parameters. These systems are also equipped with alarm and notification systems that alert operators to critical events or deviations from normal operating conditions. Since SCADA systems collect and store historical data, allowing for trend analysis and evaluation of DWTP performance, they can support also comprehensive analyses, allowing the identification of past events associated with changes in critical water quality parameters. Furthermore, data from legacy systems serves as a valuable information source that can be employed for model training purposes.

3.2 Data from ex-situ measurements

Ex-situ measurements are those that require the analysis of the water quality related parameter to be done in the laboratory. This requires the storage and transport of the sample taken from the water intake source or from the reservoir to lab facility. Due to the nature of the ex-situ measurements (require laboratory analysis), they are not performed at all the stages from source to tap, also depending on (national) legislation. Data from ex-situ measurements mostly comes from surface/groundwater water bodies the drinking water treatment plants inlets and outlets, the service reservoirs and from randomly selected consumers taps in different parts of the distribution network. All the parameters measured with installed sensors can also be measured in the laboratory, but there are some variables that explicitly require laboratory analysis.

The frequency of ex-situ measurements is usually not high, depending on the purpose of the measurements. It can be once per day up to once per (several) months(s). In addition, it is not always possible to track the conditions under which the samples are taken. These make the use of ex-situ measurements sometimes difficult to be used for the calibration of soft sensors.

3.3 Earth observation data

Earth observation (EO) data is a term that incorporates both data derived from satellite instruments and data from in-situ measurements (from installed ground sensors and airborne platforms such as weather stations[2]). EO data can facilitate the monitoring of natural resources and offer plenty of information for many physical processes like soil water content (soil saturation), precipitation estimations, soil properties (like hydraulic conductivity, porosity etc.) vegetation cover etc.

Satellite imagery refers to images of the Earth's surface captured by satellites orbiting the earth. The instruments installed in these satellites are equipped with sensors designed to detect electromagnetic radiation returning to the sensors after reflecting on the earth's surface. Consequently, satellite imagery not only facilitates the collection of information on the upstream basin, such as crop types and soil moisture conditions, but also provides insights into the quality of the water body itself. This is achieved by translating reflectance values from specific targets into variables with physical meaning. The characteristics of satellite information include temporal resolution (how often the satellite revisits the area of interest), spatial coverage (the extent of each satellite image), and spatial resolution (the size of the smallest item the sensor can detect).

When utilizing satellite imagery and EO data to estimate water quality parameters, two approaches may be adopted. One approach involves simulating and estimating catchment functions and linking them to quality-related parameters of interest in the outlet (the target water body). For instance, one might estimate nutrient or faecal coliform bacteria runoff resulting from agricultural and livestock processes upstream. The second approach focuses on using satellite imagery within the water. This can be accomplished in two ways: utilizing reflectance values to estimate water quality-related indexes or using them to directly estimate the values of the parameters of interest. The latter is accomplished by using the reflectance data or empirical indices and coupling them with in-situ measurements to train data-driven or hybrid models to estimate the target variable through space.

3.4 Available instruments

The available techniques of remote sensing data are divided into two major categories each of which can support in different parts of water quality monitoring: Passive and active instruments. Their main difference is the energy source that is used for the sensing technique.

Passive instruments derive their energy from electromagnetic radiation emitted by the sun. This radiation reaches the Earth's surface, where some of it is absorbed and scattered by various targets. The satellite instrument reads and stores the radiation that returns to the sensor. The underlying principle is that different targets, characterized by distinct materials and textures, absorb and scatter light in unique ways. Consequently, each target on the Earth's surface produces a distinctive spectral profile at the sensor. The variation in reflectance detected by the sensor is influenced not only by the nature of the target itself (such as the disparate spectral signatures of water and buildings) but also by the specific properties and characteristics of the target. For instance, a turbid water body will yield a slightly different spectral signature than a non-turbid one. This implies that passive remote sensing can play a role in the monitoring process by leveraging passive satellite data to estimate optically active variables related to water quality such as Chl-a, turbidity, and floating algae.

Active instruments operate differently in the sensing process, utilizing their own electromagnetic radiation. They function by emitting a radiation beam towards the Earth's surface. Active instruments can

[2] Group on Earth Observations (GEO), GEO at a Glance- link

operate in various parts of the electromagnetic spectrum, with the microwave part being the most common. This preference is due to the greater wavelength of microwaves compared to other parts (such as visible and infrared radiation), which enables them to penetrate clouds effectively. In the context of monitoring water bodies, active remote sensing proves invaluable for extracting the shape of water bodies even on days with extensive cloud coverage. This capability enhances the reliability and ease of preprocessing satellite and in-situ data. An active instrument employs its sensor to measure the angle, polarization, and intensity of the returning beam. Through this process, it can estimate properties related to the target on the Earth's surface. For example, a building will scatter the microwave beam differently than dense foliage, allowing the active instrument to discern distinctive characteristics.

4 Soft Sensors for Athens Demo Case

4.1 Introduction and demo case description

The upstream basin of the Yliki lake, the Boeotikos Kifissos river basin, is characterised by intense agricultural activity, and intensive use of fertilizers. The main constituents of fertilizers are nutrients, specifically nitrogen and phosphorus. Nitrogen and phosphorus are also the major drivers of eutrophication. Eutrophication phenomena can be classified into two categories: natural and cultural. The former takes place very slowly in geological time but is accelerated by anthropogenic activities, which cause the latter, cultural eutrophication. Cultural eutrophication has become more widespread from the 1940s onwards, and it has been characterized as the most widespread water quality problem [1]. Cultural eutrophication is caused by increasing inputs of nutrients, specifically nitrogen and phosphorus, which are found in excrement of livestock, human sewage, and synthetic fertilizers [2].

Figure 5: Athens water supply system operated by EYDAP.

The Boeotikos Kifissos river basin, the study area, originates from springs on Mount Parnassos in the foothills of Fokida. The river flows through the Boeotikos valley and eventually discharges into Lake Yliki. The basin covers almost 2,000 km^2, mainly in the Boeotia prefecture with smaller parts extending to Fthiotida and Fokida. It lies at the southeastern part of the basin, where the Kopaida plain was formed after the same-named lake was drained between 1880 and 1930. To the south, it borders the plains of Vagia and Thebes.

Although the need to drain the lake became evident after Greece's independence, the lake had also been drained in ancient times by the Minyas. In the 14th century BC, they built an underground tunnel to divert water from the lake to the Evian Gulf. The second drainage employed a similar method, but this time the water from Lake Kopaida was directed to Lakes Yliki and Paralimni. Initially, a French company managed the project, but in 1887, the rights were transferred to the British "LAKE CORAIS Co. Ltd." The project drained a total of 250,000 acres, enabling intensive farming of wheat, cotton, pulses, and corn.

Lake Yliki, situated east of Thebes, is the ninth largest lake in Greece, covering more than 19 square kilometers with a perimeter of around 50 kilometers. Before the drainage of Kopaida, the lake's average

depth was about 4 meters, but after the creation of the Kopaida plain, its area nearly doubled, and its depth can now exceed 30 meters. To the east of Lake Yliki is Lake Paralimni, which is fed by Yliki through depressions and karstic formations at the lake's bottom. Yliki is surrounded by Mount Ptoe to the north, Messapio to the east, and Sphingio to the west, while lower hills enclose it to the south. The surrounding mountains and small streams, in addition to the Boeotian Kifissos River, contribute to the lake's water supply. When full, Lake Yliki can hold over 660 million cubic meters of water.

Lake Paralimni is situated on the border between the prefectures of Boeotia and Evia, within the administrative boundaries of the municipalities of Thebes and Chalkidea. It is encircled by the low mountains of Ptoe to the north and Messapio to the south. The lake receives water from the nearby Lake Yliki. Covering an area of approximately 10.97 km², the lake has an elongated shape with a maximum length of about 8 km and a maximum width of 2 km. The average annual natural runoff into the lake is estimated at 3.50 million m³ per year. Together with Lake Yliki and the Kifissos River in Boeotia, Lake Paralimni forms part of the Kifissos system of Boeotia, which is included in the Natura 2000 network.

The Kifissos River basin is predominantly characterized by a Csa climate, which is a Mediterranean climate with dry, very hot summers. In the upper part of the basin, other climate types are also present to a lesser extent, including Csb (Mediterranean climate with dry, warm summers), Dsb (continental climate with dry, warm summers), and Dsc (continental climate with dry, cool summers).

In the Water Region of Eastern Central Greece, the Boeotian Kifissos River and Lake Yliki are considered sensitive receptors. According to current regulations, the discharge of liquid or solid waste into these areas is prohibited. Directive 91/271/EEC requires towns and cities with populations equivalent to more than 10,000 people, located in the catchment areas of sensitive receptors, to establish a sewerage network and wastewater treatment plants by the end of 1998, according to the timetable set by the directive. As a result of these restrictions, wastewater treatment plants have been constructed in several cities, including Lamia, Chalkida, Oinofyta, Thebes, Livadia, Kamena Vourla, Loutra Edipsos, and Skiathos. Many of these plants provide biological treatment with nitrogen and phosphorus removal.

The primary non-point sources of pollution stem from agricultural and livestock activities, particularly free-range farming and extensive use of fertilizers, which contribute significantly to nutrient loading in surface and groundwater. The dominance of the agricultural activity in the area is obvious from the Land-Use-Land-Cover map where cropland covers more than 30% of the total area (see Table 2)

Table 2: Land-Use and Land-Cover table of the study area

LULC Class	Trees	Crops	Shrub & scrub	Built	Grass	Bare Soil	Water	Snow and ice
Percentage (%)	45.90	33.57	15.57	3.24	1.09	0.53	0.05	0.04

Challenges to be addressed

Ensuring sustainable water management within the basin demands close attention to the nutrient load coming from the large tracts of cultivated land. Nutrients like nitrogen and phosphorus, often present in fertilizers, can seep into nearby rivers and downstream water bodies. This leakage risks causing eutrophication, a condition that spurs rapid algae growth, depletes oxygen, and can degrade water quality—complicating water treatment efforts.

Although in-situ measurements of nutrient levels provide accurate data, gathering this information can be both costly and time-consuming. To aid decision-makers, soft-sensors are developed. These sensors blend direct measurements with satellite data, utilizing Earth Observation technologies' broad spatial and temporal reach. This integration improves both the frequency and geographical range of water quality assessments for Lake Yliki, delivering essential insights that align with water management needs.

4.2 Soft Sensor 1 - Early Warning System for Nutrient Runoff

The concept of early warning systems (EWS) is basically a strategy for monitoring risks and providing timely alerts for potential crisis. Early warning systems specifically for water quality have been previously studied. Wuang et al. [3], developed an early warning system for pollution risk assessment in major inland water bodies of China. In this study, publicly available hydrological and water quality data were utilized, while at the same time, web scraping methods were applied for the collection of these data in real-time. The modelling approach for the early warning system was based on a modified Long Short-Term Memory (LSTM) network, which allowed quick evaluation and the water quality related risk assessment. Another study that took place in the Three Gorges Reservoir [4], focused on the forecasting of pollution accidents. The model utilized a two-dimensional water quality model in combination with water quality security standards to predict the spatiotemporal trends in the pollutant levels and the early warnings and forecasts for water quality safety. In [5], Sentinel-2 and Landsat-8 satellite were used to monitor water quality in the Mar Menor coastal lagoon. These satellite data allowed for the detection of critical changes in the water quality parameters, which were used as early indicators to trigger warnings about potential water quality issues, such as eutrophication or harmful algal blooms.

Based on this brief review, the conceptualization of an early warning system can be summarized in the following constituents.

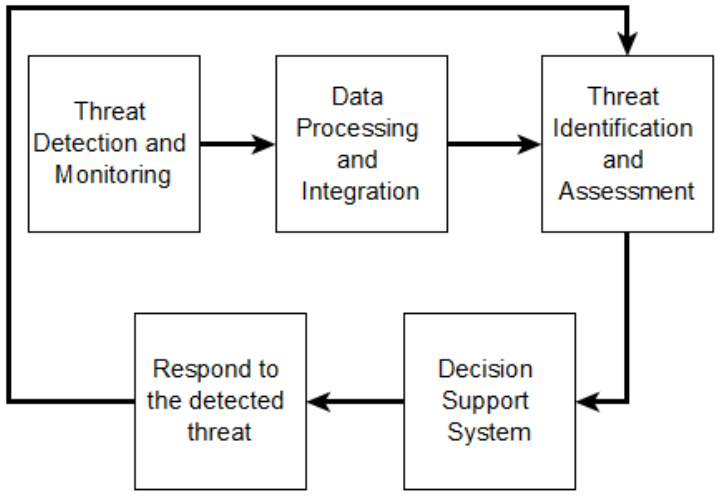


Figure 6: Soft sensor constituents

- **Threat Detection and Monitoring:** Monitoring of the environment for potential threats or hazards. This includes data from in-situ sensors, ex-situ measurements, or other data sources like social media, surveillance systems etc.
- **Data Processing and Integration:** Data transformation into information that can be fed into the early warning system. This may include data preprocessing methods like noise filtering, data normalization and methods for the aggregation of data from different sources.
- **Threat Identification and Assessment:** Includes the assessment of the potential impact of the threat based on the processed data. This assessment includes the utilization algorithms and rule-based systems (e.g., for the threshold determination), the identification of the severity of the threat and the impact assessment.
- **Decision Support System:** Includes the interpretation of the threat assessment and the choice of the appropriate actions. This assessment can be performed by decision making frameworks or can be performed based on expert knowledge.
- **Respond to the detected threat:** Includes the response coordination and action to the emerging threat based on response protocols or systems that carry out the mitigation of the threat.

The first soft sensor concerns the development of an early warning system for nutrient runoff in the Yliki Lake. The aim of this soft sensor is to develop an early warning system regarding eutrophication phenomena in the downstream of basins characterized by intensive agricultural activities, with the utilization of EO data. The proposed approach consists of two main components: the estimation of the nutrient load in the basin and a rainfall-runoff model for the transformation of the precipitation into runoff.

4.2.1 Problem statement and soft-sensor development flow-chart

An effective early warning system for nutrient runoff relies on two key components: estimating nutrient loads in the upstream basin and developing a hydrological model to assess runoff, which will help predict nutrient transport to the water body.

The first component, nutrient load estimation, requires detailed crop mapping and quantification of nutrient inputs. Once crop types are identified, agricultural practices and associated nutrient loads should

be assessed to determine their impact. The second component, a hydrological model for converting rainfall into runoff, must support real-time simulations to be operationally viable. This requires careful selection of datasets to ensure they meet the system's demands for accuracy and responsiveness.

To ensure the system achieves accuracy and efficiency at the operational level, the datasets used for estimating nutrient loads and modelling runoff should have appropriate temporal and spatial scale. For the treatment process to adequately prepare for increased nutrient concentrations in the lake, precipitation forecasts should be integrated to estimate discharge in advance. Simultaneously, nutrient load dynamics must be captured and quantified with sufficient temporal and spatial resolution. In this framework, a one-day-ahead discharge forecast will be simulated using a continuous hydrological modelling scheme. Nutrient loads in the basin will be estimated at a monthly temporal scale, reflecting the fertilization patterns associated with the crops in the area. The crop type mapping is conducted at a spatial resolution of 20 meters, which is sufficient to capture crop types at the plot scale, ensuring the detailed representation of agricultural practices and their contributions to nutrient dynamics.

4.2.2 Review of state-of-art approaches

The proposed model provides a novel methodology and addresses key limitations of the existing approaches regarding nutrient runoff in inland water bodies in all three main aspects of the proposed method.

Starting with the nutrient load estimation which also includes the crop type classification, deep neural networks have been previously applied for this image segmentation task, even the 3D Unet has been previously applied. (Jia et al., 2024), proposed an advanced ASPP-SAM-UNet algorithm integrating spatial attention mechanisms and multi-scale features to improve crop classification accuracy in remote sensing, demonstrated in Bayan County, China. (Ayushi and Buttar, 2024), proposed a fully convolutional encoder-decoder architecture to address the challenges like irregular field shapes, small plot sizes and cloud cover on a relatively small dataset. The novelty of the proposed model for crop type classification lies primarily in its handling of imbalanced datasets, which is a significant challenge in both classification and regression tasks in machine learning (Pereira and Saraiva, 2020; Rout et al., 2018; Spelmen and Porkodi, 2018). Including imbalanced data handling in image segmentation models is crucial for ensuring accurate segmentation of underrepresented classes (such as small but nutrient-intensive crops) and preventing bias towards more dominant crop types. Regarding nutrient load quantification in the basin, the proposed NNLI offers a novel and effective approach to measure nutrient load in the absence of in-situ nutrient measurements from upstream crops. By combining agricultural practices with the crop type classification map, it provides a simple yet robust estimation of nutrient load at the pixel scale, as well as across the entire basin.

At the same time, the hydrological model presented in the early warning system utilizes methodologies that are still in their infancy. More specifically, the utilization of satellite data in hydrological simulation has gained ground recently. Most studies rely on satellite precipitation data to drive hydrological simulations (Bitew et al., 2012; Bitew and Gebremichael, 2011; Zhao et al., 2015), while others use satellite data for evapotranspiration estimates (Herman et al., 2018; Immerzeel and Droogers, 2008; Jiang et al., 2020; Kite and Droogers, 2000). The hydrological model developed for the first soft sensor goes a step further by integrating Earth Observation (EO) data for precipitation, evapotranspiration, and soil properties, using the HiHydrosoils dataset to enhance continuous hydrological simulations.

4.2.3 Data sources and data preprocessing

The early warning system for nutrient runoff is based on the utilization of in-situ and EO data. EO and in-situ data are utilized for the development of each individual development.

Data sources for the crop type classification

In-situ data

The in-situ data were provided by the Greek Payment Authority of Common Agricultural Policy (OPEKEPE) for the year 2023. That is, the data refers to the crop census in the study area for the year 2023. The data were provided in static vector files accompanied by the memo for the code corresponding to each crop type. In the area, there are over 30 classes in an area of about 825 square kilometers. However, more than 93% of the area is covered by the following classes presented in Table 3.

Table 3: Crop types in the study area

Crop Type	Area km²	Percentage (%)
Pasture	182.997	22.18
Cotton	155.362	18.83
Cattle Feed	151.260	18.33
Olive Grooves	77.877	9.44
Cereals	75.427	9.14
Fallow	64.836	7.86
Wheat	43.974	5.33
Corn Irrigated	22.620	2.74

Pasture occupies the largest area at 182.997 km² (22.18%), followed by cotton at 155.362 km² (18.83%) and cattle feed at 151.260 km² (18.33%). Other significant crops include olive groves, wheat, and fallow land, with percentages ranging from 9.44% to 5.33%, while irrigated corn covers the smallest area at 2.74%.

EO Data

The input data for the crop type classification task consists of multi-temporal satellite data acquired from the Sentinel-2 MSI satellite mission. Sentinel-2 captures data in 13 different spectral bands, ranging from visible to near-infrared and shortwave infrared. The mission includes two identical sun-synchronous satellites, Sentinel-2A and Sentinel-2B, both operating at an altitude of 786 km. Sentinel-2A was launched in 2015, followed by Sentinel-2B in 2017. Each satellite has a return time of 10 days at the same viewing angle, but because this is a two-satellite constellation, some areas can be observed two or more times every 10 days, albeit from different viewing angles. Both satellites carry identical instruments—the Multi-Spectral Instrument (MSI, see Figure 7)Figure 7: The MSI on-board the Sentinel-2 mission —which records reflectance across the 13 spectral bands (see Figure 8). The MSI operates using a push-broom (along-track) scanning method, which meets the mission's requirements for a large swath width and ensures high geometrical and spectral accuracy in the measurements.

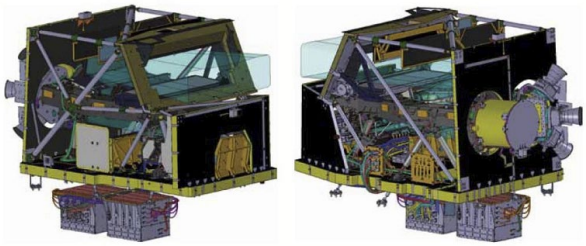


Figure 7: The MSI on-board the Sentinel-2 mission

The spatial resolution of the MSI instrument varies across its spectral bands, capturing data at 10, 20, and 60 meters. Band 1, which records reflectance at a central wavelength between 442.2 and 442.7 nm with a bandwidth of 21 nm, has a spatial resolution of 60 meters. The visible spectrum—Blue (Band 2), Green (Band 3), and Red (Band 4)—is recorded at a 10-meter resolution. The three bands that capture reflectance at the vegetation red edge have a 20-meter resolution, while two bands in the Near InfraRed (NIR) spectrum record at 10 meters and 20 meters, respectively. The water vapor band and the Cirrus Short-wavelength infrared spectrum (SWIR) band also have a 60-meter resolution, whereas the other two SWIR bands have a resolution of 20 meters.

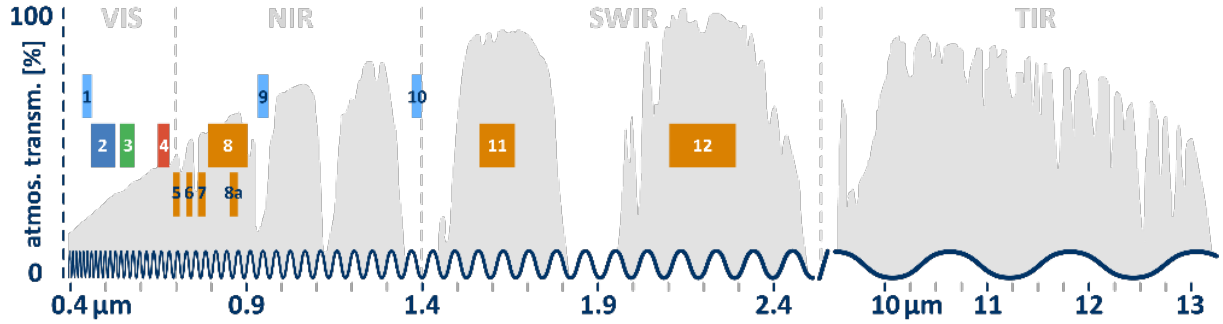


Figure 8: The spectral bands of Sentinel-2 satellite

Data sources for the hydrological model

In-situ precipitation and discharge

The in-situ data used for model development include precipitation records from the station presented in Figure 9, from 2008 to 2017 and from 2019 to 2021. The calibration of the model was performed based on the in-situ precipitation measurements by setting the calibration period from October 1st, 2008 to September 31st, 2017 and the validation period from October 1st 2019 to September 31st 2021. The simulations are at daily time-step and the precipitation was calculated based on the Thiesen method, per subbasin.

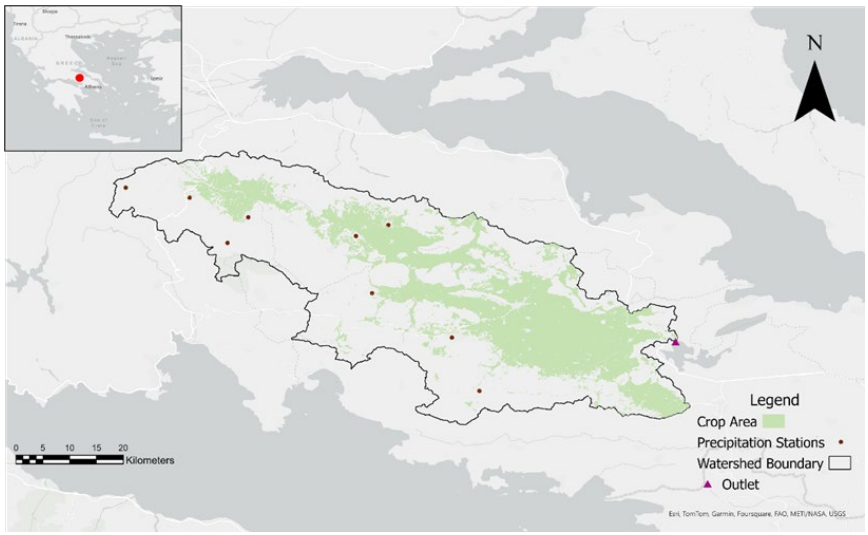


Figure 9: Map of the study area

Earth Observation Datasets

EO datasets are employed to support the calibration of HEC-HMS, and particularly the Soil Moisture Accounting (SMA) loss method, as well as to drive the model with gridded meteorological data input. Particularly, HiHydrosoils dataset [6] are used in the parameter mapping of SMA method, by providing data on soil properties. NOAA GFS precipitation data was used to drive the calibrated model with gridded precipitation forecasts. In the following paragraph each EO dataset employed is discussed and a summary of the datasets used in provided in Table 4.

- HiHydrosoils v2.0 [6] builds upon the ISRIC SoilGrids 250 m (De Sousa et al., 2021), offering an enhanced, high-resolution global dataset for soil hydraulic properties. SoilGrids 250 m produces information combining soil observations from more than 200000 locations (in-situ data), over 400 covariates related to vegetation, climate and geology with machine learning models. It provides information about soil properties such as bulk density, soil organic carbon, soil pH etc., in 6 different (standard) depths. The variables included in this dataset are used as input for deriving soil hydraulic properties for the HiHydrosoils v2.0 dataset. For the conversion of the soil properties into soil hydraulic functions, pedotransfer functions were utilized, while for the calculation of the Hydrologic Soil Group (HSG), the absolute depth to bedrock and the simulated groundwater depth were used as input. The hydraulic properties in the HiHydrosoils dataset are presented in the six different depths for which ISRIC SoilGrids250m provides the soil properties. Layer 1 spans from the surface to a depth of 5 cm, while layer 2 represents the next 10 cm (from 5 to 15 cm). Layer 3 includes the region from 15 cm to 30 cm and layer 4 includes the next 30 cm. Layer 5 spans from 60 to 100 cm deep while the final (sixth) layer starts from the lower 100 cm and expands downwards to the lower 200 cm.
- Evapotranspiration estimations were obtained from NASA MODIS ET dataset, which is an 8-day composite dataset, with a spatial resolution of 500 m. The estimations are based on the Penman-Monteith method [8] and reanalysis of meteorological and vegetation-related remote sensing data to calculate evapotranspiration (ET), potential evapotranspiration (PET), latent heat flux (LE) and potential latent heat flux (PLE). The algorithm used for the extraction of ET/PET considers both the surface energy and vegetation related indices like Leaf Area Index (LAI) and Normalized Vegetation Index (NDVI) for the calculation of vegetation cover. The pixel values for both Evapotranspiration layers (ET and PET) derived by aggregating the values from all eight days within the composite period [9].

- The Global Forecast System (GFS) (Wu et al., 2011) is a weather forecasting model developed by the National Centers for Environmental Prediction (NCEP). GFS integrates global models for the nexus atmosphere, ocean, land/soil, sea, and ice, providing forecasts for various weather-related processes, such as wind, temperature, ozone concentration, and precipitation. It updates every 6 hours, with new data available four times a day, and produces 384-hour weather forecasts on a 28 km grid, with intervals of 1 and 3 hours.

A summary of the EO datasets used for the development of the early warning system for nutrient runoff are provided in Table 4.

Table 4: Summary of the EO datasets used for the development of the Early Warning System

Dataset	Variable	Temporal Resolution	Spatial Resolution	Reference
Sentinel-2 MSI	Surface reflectance data	5 days	10-60 m	Drusch et al., 2012
HiHydrosoils v2.0	Soil hydraulic properties	-	250 m	Simons et al., 2020
MODIS ET/PET	Potential evapotranspiration	8 days	500 m	Running et al., 2021
NOAA GFS	Precipitation forecast	Daily	27830 m (0.25s°)	Clough et al., 2005

In-situ and Satellite Data preprocessing

Before data can be used to develop any components of the early warning system, it must undergo preprocessing to correct discrepancies, extract meaningful information, and format it appropriately for model development.

In-situ data pre-processing

Discharge time series correction: An analysis of the basin outflow data from 1994 to 2022 reveals multiple instances where identical values are recorded more than twice. Notably, this excludes the zero runoff values observed during the summer months, which are clearly due to the natural absence of runoff rather than measurement error. However, in other months, when runoff values remain constant for two or more days—particularly when such values are recorded with a precision of seven decimal places, as seen at the Boeotian Kifissos basin outlet—this consistency cannot be attributed to natural causes.

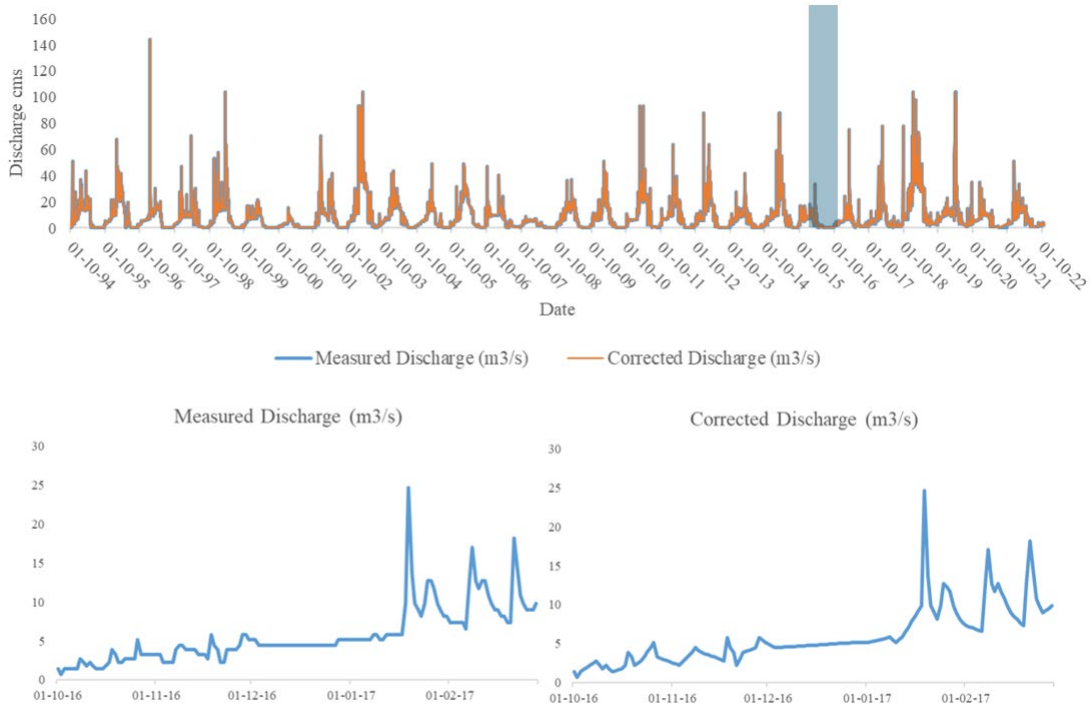


Figure 10: Discharges at the Kifissos river basin outlet

To address this discrepancy, the measurements were adjusted by correcting consecutive identical values recorded on successive days. This was done by removing the repeated values and filling in the time series using linear interpolation.

Satellite data pre-processing

The satellite data from Sentinel-2 is used for the crop type classification and treated as images time series. More specifically, as the reference crop type dataset includes the census data from the year 2023, several satellite image instances have been selected for model development. More specifically, 11 representative instances of the study area were selected based on the quality of the images and specifically the cloud cover percentage. Only instances in which the cloud cover was less than 12% were selected, in order to minimize the effect of the clouds in the reflectance measurements. The time selected for the analysis are provided on Table 5.

Table 5: The instances used in the SITS data cube

Month	Date	Cloud Cover (%)
January	3 January 2023	0.45
February	2 February 2023	3.90
March	14 March 2023	1.72
April	13 April 2023	0.01
June	22 June 2023	10.57
July	2 July 2023	3.90
August	16 August 2023	0.36
September	10 September 2023	7.04
October	10 October 2023	0.19
November	19 November 2023	0.03
December	19 December 2023	0.006

The Satellite Images Time Series (SITS) obtained, are transformed into a Data Cube (DC). A DC is a data structure for multidimensional data, the data within a DC is explained by specific dimensional values. This way, by using multidimensional arrays, spatiotemporal data can be meaningfully represented. The DC produced for the supervised crop type classification task is presented in the Figure 11. The developed DC is a 4D tensor of shape 11∗2694∗4678∗7. The first dimension represents the time instances (i.e., the number of images collected), the second represents the columns while the third represents rows. Finally, the last dimension represents the spectral bands of each time instance.

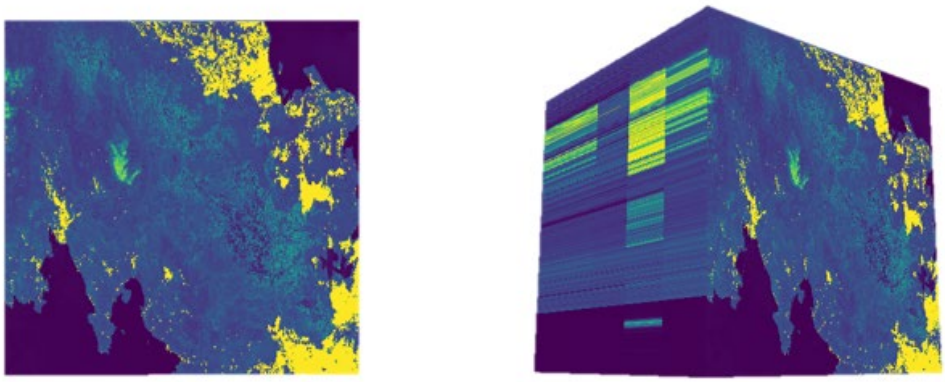


Figure 11: The Sentinel-2 data cube.

4.2.4 Materials and methods

As mentioned above, the development of this soft sensor requires the development of two individual sub-models. The nutrient load estimation and the rainfall to runoff transformation for the transfer of the nutrients from the basin to the lake. The proposed methodology is illustrated in the flowchart presented in Figure 12. The lighter shades represent the data sources, while the slightly darker shades represent the individual components contributing to the development of the early warning system.

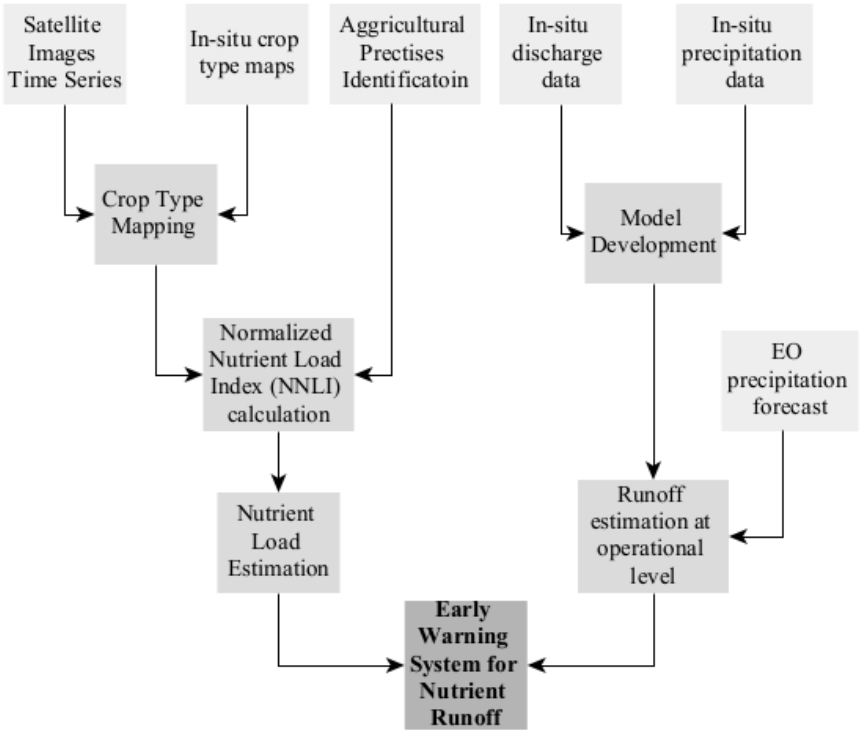


Figure 12: Flowchart of the early warning system.

In-situ crop type data and the Satellite Images Time Series (SITS) are used to develop a supervised computer vision model for the crop type classification, while the in-situ precipitation and discharge are used to calibrate and validate the hydrological model, developed using HEC-HMS [13].

Nutrient Load in the Basin

The first component of the early warning system includes the estimation of the nutrient load in the basin. This component is composed of three other individual components as presented in the following Figure 13.

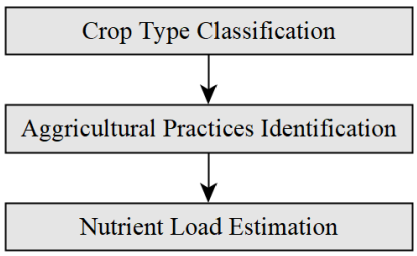


Figure 13: Components of soft sensor on nutrients load

The crop-type classification task leverages in-situ crop maps and satellite data cubes to develop a supervised classification model. Based on existing literature, agricultural practices for the major crops in the study area are then identified. Finally, the nutrient load in the basin is quantified using the proposed Normalized Nutrient Load Index (NNLI).

Crop Type Classification

For the crop type mapping task in the study area, Computer Vision (CV) methods are combined with Earth Observation (EO) data. The crop type classification task is handled as an image segmentation task. For this image segmentation task, three different architectures are compared. The *first* architecture, serving as the baseline model, is based on the basic **Unet architecture** [14] . In the *second* architecture examined, the time dimension is compressed in the encoder part of the network. The *third* architecture includes the baseline model but with attention modules at the skip connections. The selection of the loss function and its parameters is also carefully examined. Given the highly imbalanced reference dataset, it is essential to address this imbalance effectively. To tackle this, the focal loss function is chosen, and various parameters of the function are tested to determine the optimal configuration.

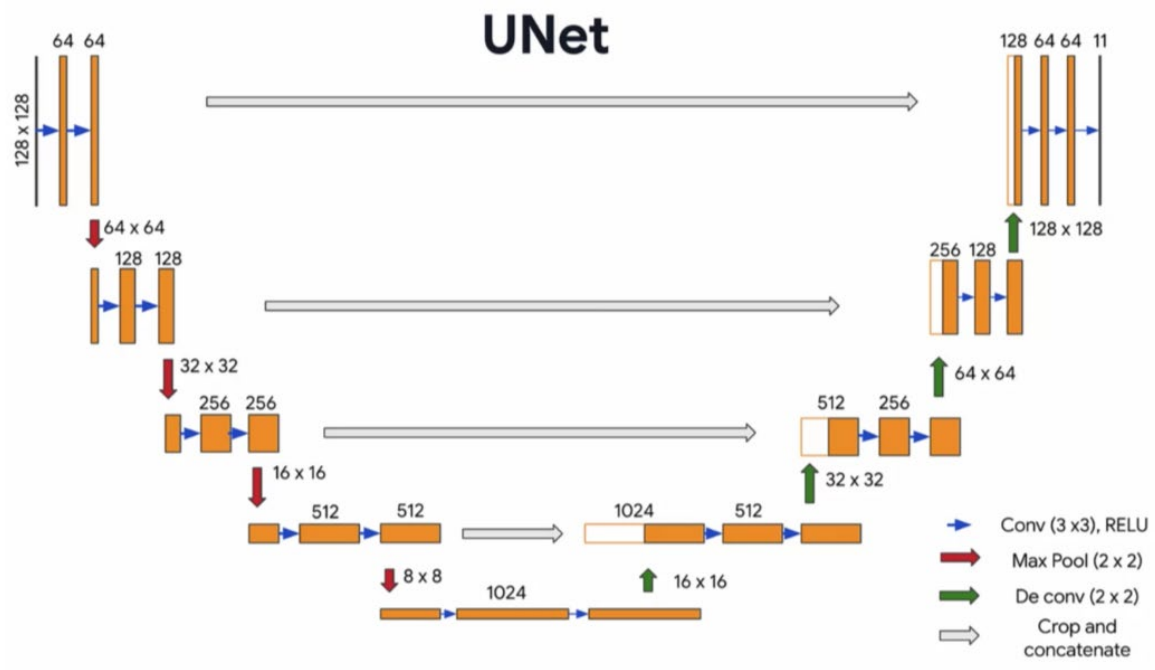


Figure 14: The UNet architecture

Baseline Model

The baseline model follows the typical Unet architecture which consists of consecutive convolutional and MaxPooling layers in the encoder and transpose convolution layers at the decoder, with skip connection from each convolutional block to the corresponding decoder block. The encoder part consists of a series of convolutional layers, followed by non-linear activation functions. MaxPooling operations are utilized in order to progressively reduce the spatial dimensions while increasing the feature representation. The decoder part utilizes transpose convolutional layers, to progressively reconstruct the spatial dimensions of the input image. In this context, the important spatial information extracted by the encoder part, skip-connections are introduced between each layer of the encoder and the decoder part. This way, both the low-level details of the encoder and the high-level abstractions of the decoder are combined, achieving balance between localization and contextual representation.

Baseline Model with Compression of the Time Dimension

The baseline model follows the typical UNet architecture which consists of consecutive convolutional and MaxPooling layers in the encoder and transpose convolution layers at the decoder, with skip connection from each convolutional block to the corresponding decoder block. The encoder part consists of a series of convolutional layers, followed by non-linear activation functions. MaxPooling operations are utilized in order to progressively reduce the spatial dimensions while increasing the feature representation. The decoder part utilizes transpose convolutional layers, to progressively reconstruct the spatial dimensions of the input image. In this context, the important spatial information extracted by the encoder part, skip-connections are introduced between each layer of the encoder and the decoder part. This way, both the low-level details of the encoder and the high-level abstractions of the decoder are combined, achieving balance between localization and contextual representation.

Baseline Model with Attention at the skip connection

The baseline model follows the typical UNet architecture which consists of consecutive convolutional and MaxPooling layers in the encoder and transpose convolution layers at the decoder, with skip connection from each convolutional block to the corresponding decoder block. The encoder part consists of a series of convolutional layers, followed by non-linear activation functions. MaxPooling operations are utilized in order to progressively reduce the spatial dimensions while increasing the feature representation. The decoder part utilizes transpose convolutional layers, to progressively reconstruct the spatial dimensions of the input image. In this context, the important spatial information extracted by the encoder part, skip-connections are introduced between each layer of the encoder and the decoder part. This way, both the low-level details of the encoder and the high-level abstractions of the decoder are combined, achieving balance between localization and contextual representation.

Table 6: Models tested for the supervised crop type classification task

Model	Architecture Characteristics	Total Parameters	Trainable Parameters
1.Baseline	Basic Unet architecture but with 3d convolutions	22,673,860	22,667,972
2.Baseline with Time Compression	Baseline model with compression of the time dimension at the encoder	23,185,988	23,180,100
3.Baseline with Attention modules at the skip connections	Baseline model with attention mechanisms at the skip connections	22,851,304	22,844,456

Class imbalance and the loss function

Class imbalance is the situation in which, in a classification task, the instances of each class of the dataset are not equally represented. In fact, class imbalance is a situation that can occur in many real-world scenarios, like the prediction of extreme precipitation events. In these cases, the rare (extreme) events are of great interest but are very rare and are overshadowed by the most common instances. In case class imbalance is not handled correctly, it can lead to biased models, not effective in predicting the minority classes. One of the major issues when dealing with class imbalance in imbalanced datasets is related to the evaluation metrics utilized to evaluate the performance of the model. Simpler metrics can be misleading. For examples, in a multiclass classification problem with imbalanced data with 5 classes and 100 instances, in the case that class one is represented by 85 instances, and the rest 15 are shared by the other false classes, if the model predicts correctly the “easy-to-classify” class, and messes up with the rest 4 classes, the model accuracy would still be 85%, and on the same time the model would be useless because it cannot predict the “hard-to-classify” classes.

The are various ways to deal with class imbalance like resampling the training dataset, assign class weights and choosing the proper loss function. When the resampling of the training dataset is chosen to address class imbalance, it can be performed through either random oversampling of the minority class or random under-sampling of the majority class. In the first case, samples from the majority class are randomly eliminated until the classes are balanced. In the second case, which is the reverse of the first, the number of samples in the minority class is increased by randomly replicating existing samples. While oversampling the minority class avoids information loss, it also increases the risk of overfitting. On the other hand, the class weighting technique assigns varying weights to each class in the training data, penalizing the model more heavily for misclassifying harder-to-predict classes. However, one drawback is that it might be ineffective in cases of extreme class imbalance, as the model could still favour the majority class. This method involves modifying the loss function.

Another approach for addressing imbalanced data is to select an appropriate loss function. Focal Loss is specifically designed to manage class imbalance in classification tasks. Introduced by the Facebook AI Research team (FAIR) (Lin et al., 2017), Focal Loss serves as an alternative to cross-entropy (Figure 15) by giving greater weight to difficult or frequently misclassified examples while reducing the weight of easy examples. This is accomplished through a modulating factor that adjusts each sample's impact on the overall loss.

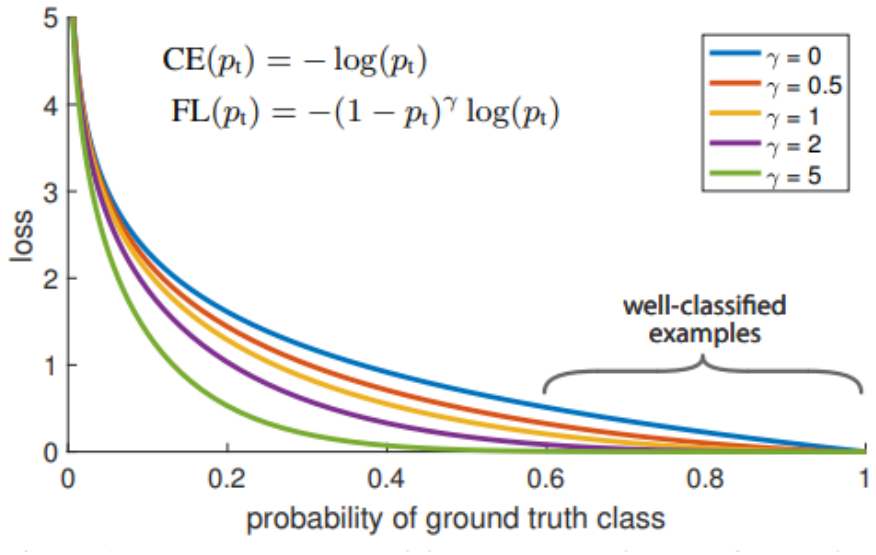


Figure 15: Focal Loss and Cross Entropy functions

$$Cross - EntropyLoss = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C} y_{n,i}\log\left(y'_{n,i}\right)$$

$$Focal\ Loss = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C}(1 - y''_{n,i})^{\gamma}\, y_{n,i}\left(y'_{n,i}\right)$$

From the supervised classification task, the focal loss function is used to tackle the class imbalance problem. In addition to that, and as the dataset is highly imbalanced, the focal loss function is combined with weighted sampling for the model training.

Agricultural Practices Identification and Nutrient Load Estimation

To incorporate the nutrient load on the basin as input to the model, the load should be quantified. For that, the Normalized Nutrient Load Index (NNLI) is proposed in this section. The estimation of NNLI is based on the crop type and the agricultural practices identified in the region of interest. A flowchart of the proposed NNLI is provided in the following figure.

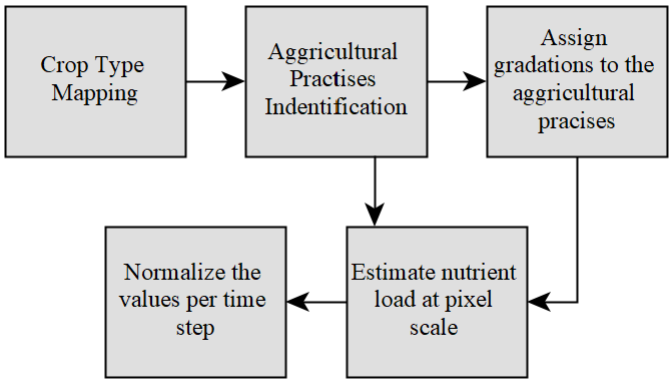


Figure 16: Flowchart for the estimation of the proposed NNLI

Based on the above figure the estimation of the NNLI can be broken down into five steps.

The *first step* involves mapping the types of crops in the area. Identifying these crop types is essential for estimating nutrient loads and determining fertilization periods, which make up the *second step*. For most crops, fertilizer is typically applied during two periods: the base dressing and the top dressing. The base dressing occurs during the sowing season and usually requires more fertilizer, which is applied directly to the soil. The top dressing, applied later in the growing season, requires less fertilizer and is applied to the surface of the crops.

In the *third step*, agricultural practices are assigned gradations: base dressing periods are labelled as 1, top dressing periods as 0.5, and areas with no fertilization as 0. These labels are applied to each pixel based on its identified agricultural practices for each time step, allowing the NNLI to account for variations in fertilizer use.

For each time step, the nutrient load in the study area is calculated as the weighted sum of the nutrient-loaded pixels (*step 4*). After estimating the nutrient load for each time step, a normalization equation is applied (*step 5*). The NNLI, calculated for each time step, provides an overview of the nutrient load across the entire study area.

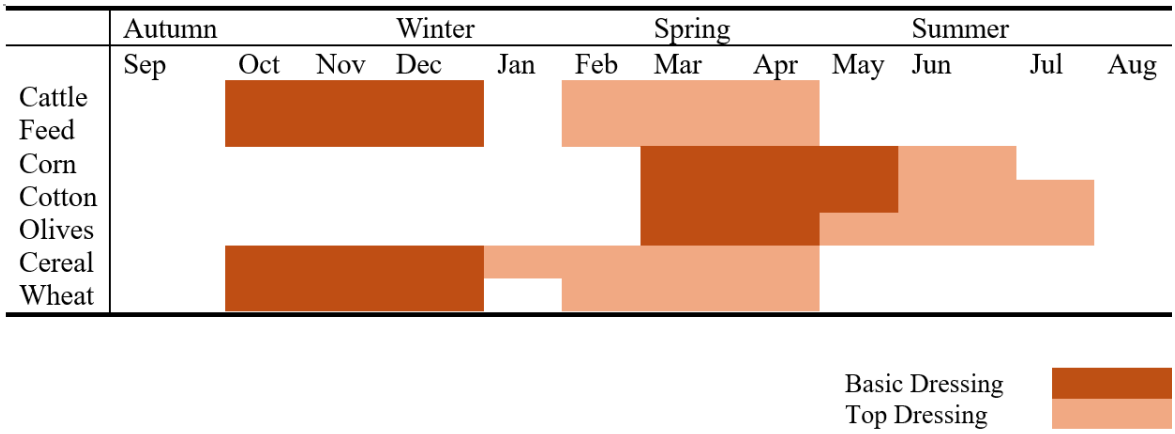
A general overview of common agricultural practices for each crop type as derived from the related literature for each crop type is presented in the following paragraphs.

- Cotton is planted usually from March to May, as it requires a soil temperature around 15°C. It requires a significant amount of nitrogen, phosphorus and potassium to be applied at various stages. Nitrogen especially is split into multiple applications, from planting, mid-season and before flowering. It is harvested from late summer to fall.
- Olive Trees are perennial and are not planted annually but new trees are planted in early spring. They require regular fertilization with nitrogen and phosphorus which is done in early spring, sometime during the growing season and after harvest.

- Wheat, in particular winter wheat, which is found in the stud area, is sown in the fall between September to November. Before the sowing season phosphorus is applied, while nitrogen is usually applied at planting and again at tillering.
- Corn is typically planted in spring, from April to June as it requires a soil temperature around 10°C. Corn is characterized as a heavy feeder, and it needs significant amounts of nitrogen phosphorus and potassium. Nutrients are applied in multiple stages from planting to the six-leaf stage and before tasting. It is harvested from August to October.

Based on the above, the Table 7 summarizes the agricultural practices in the Boeotikos Kifissos river basin.

Table 7: Agricultural Practices and fertilizer application periods



Hydrological Modelling

Model description and set-up

The data used for the hydrological model development includes both in-situ measurements and EO data. Both the in-situ and the EO data are incorporated into the HEC-HMS model. The software includes a variety of sub-models for infiltration, rainfall-runoff transformation, baseflow estimation, hydrological routing, and modules that enable continuous simulation, such as the Soil Moisture Accounting (SMA) loss method [15].

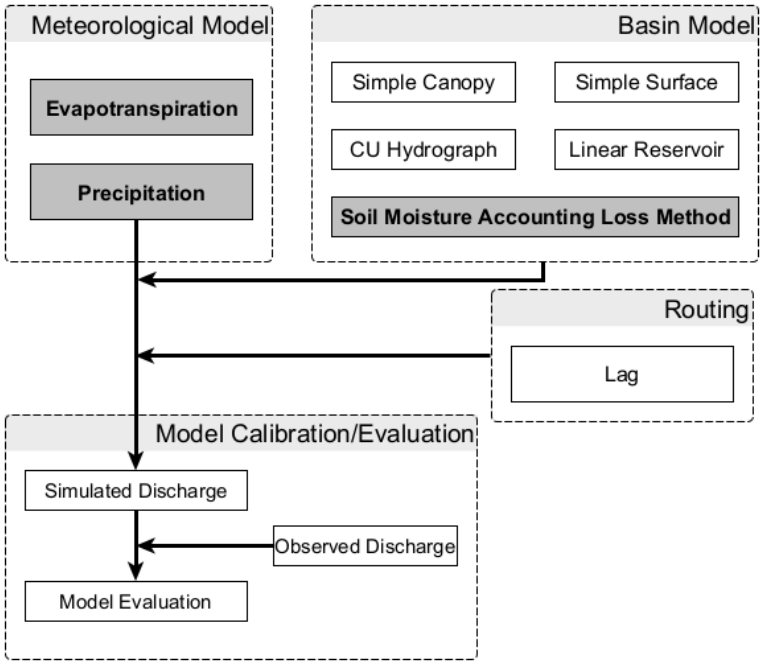


Figure 17: Flowchart of model description

The model is composed of two basic sub-modules, the basin model and the meteorological model. The meteorological model is simpler to parameterize and requires the main drivers of the hydrological model like the precipitation and the evapotranspiration. In this set-up, we have selected the monthly average potential evapotranspiration method, with values calculated from the MODIS mission observation. For the precipitation, both in-situ and EO precipitation data have been utilized.

With regards to the basin model, this includes the following sub-models: canopy storage, surface storage, a model to transform excess precipitation to direct runoff, a model that simulates water losses in the soil and baseflow. In detail:

- Canopy storage is anticipated to vary based on the type of vegetation species and the density of vegetation cover. However, the connection between canopy storage and these factors is not well understood yet [16], and due to this the estimation of maximum canopy storage was obtained in the calibration procedure. The crop coefficient was set to 1 and evapotranspiration to "Wet and Dry periods". For water uptake, we used the "Simple Canopy" method, assuming that water is drawn from the soil at the potential evapotranspiration rate.
- The estimation of surface depression storage values relies on the initial estimates provided by Bennet and Peters [15] in combination with the slope percentage value from the Soil Survey Manual (NRC Soil Science Division, 2017.).
- To convert excess rainfall into direct runoff we employed the Clark Unit Hydrograph method [18]. This method requires the calculation of both the time of concentration and a storage coefficient, thus accounting for attenuation and diffusion processes. This approach is realistic since the lower part of the river course is characterized by almost negligible slopes.
- To simulate water loss in the soil, special focus was given to the proper identification of parameters of HEC-HMS associated with the simulation of water movement within the soil. To enable continuous simulation, we employed the SMA loss method, which divides the soil into five layers: the canopy interception layer, surface depression layer, soil profile storage layer and two groundwater layers, termed GW1 and GW2, respectively. The method has seven parameters. Starting with maximum

infiltration, it sets the upper bound on infiltration from surface storage into the soil. Next, soil percolation defines the upper bound on percolation from the soil storage layer into the upper groundwater layer. Finally, percolation rates in GW1 and GW2 set the upper bound on the percolation from the upper groundwater layer to the lower groundwater layer, and the upper bound on deep percolation, respectively. In addition to the above-mentioned parameters the SMA method requires the specification of initial wetness conditions for both the soil and the groundwater layers. EO data were employed to support the calibration of all SMA parameters and the proper identification of initial conditions.

- Finally, the linear reservoir approach was chosen to model the baseflow recession after precipitation events. This is directly related to the loss method (the infiltration calculated from the loss method is the inflow to the linear reservoir). The GW1 and GW2 coefficients for the linear reservoir correspond to the GW1 and GW2 coefficients of the SMA loss method. The GW1 and GW2 fractions in the linear reservoir determine how the water from the loss method is divided into the groundwater layers. The attenuation during the routing in each one of the groundwater layers is accounted for by the number of reservoirs/layers, increasing with increasing number of reservoirs.

Coupling HEC-HMS model with EO soil properties data

In this work, we take advantage of the information provided by the HiHydrosoils dataset to support the calibration of HEC-HMS model, and specifically the parameters of SMA loss method associated with the simulation of the downward water movement in the soil and towards the outlet. This dataset provides information with respect to most SMA method variables, while its gridded data availability allows accounting for the spatial heterogeneity of the basin's soil hydraulic properties. To evaluate the added value derived from such an approach, we contrast with another model setup, which uses *default* parameter estimation procedures proposed in the literature. The two approaches are detailed in the following sections.

HiHydrosoils dataset provides soil-related parameters in six standard depths. However, these depths do not necessarily represent either the total soil depth of an area (HiHydrosoils provide information up to the first 200 cm, while soils can be deeper or shallower) or the variation of the soil properties as we move to deeper layers. For example, the average soil conductivity profile from 15 to 30 cm, the third layer of the dataset, is probably not the same from 15 to 20 cm. Thus, reasonable assumptions should be made with respect to the selection of soil layers and parameters, also in accordance with the structure of the SMA model. For example, the Tension Zone layer will be considered as a part of the soil storage layer (as described in the SMA model) and the tension storage will be estimated for the first 15 cm of the soil.

To map the parameters provided by HiHydrosoils into the domain of SMA method, we developed the methodology illustrated in Figure 18**.** Starting with the maximum infiltration parameter, the parameter values of each subbasin were derived from the saturated hydraulic conductivity layer within the upper 15 cm of soil [19]. Soil percolation represents the speed at which water migrates to the initial groundwater layer and it was considered equal to the saturated hydraulic conductivity in the 15 - 100 cm range. When addressing percolation in two groundwater layers (termed GW1 and GW2), we refer to the saturated hydraulic conductivity for the deepest layer accessible in the HiHydrosoils dataset (100 - 200 cm). Soil storage values were estimated by multiplying the soil depth by the weighted soil porosity, and tension storage was calculated by computing the weighted field capacity times the average soil depth in each subbasin. The GW1 and GW2 parameters were estimated based on the storage coefficient derived by the Clark unit hydrograph method. To introduce diversity in parameter values among subbasins in accordance with their spatial heterogeneity, the initial calculations for GW1 and GW2 storages were made in proportion to their soil storage. For the reasons mentioned above, these parameters underwent further refinement during the calibration procedure.

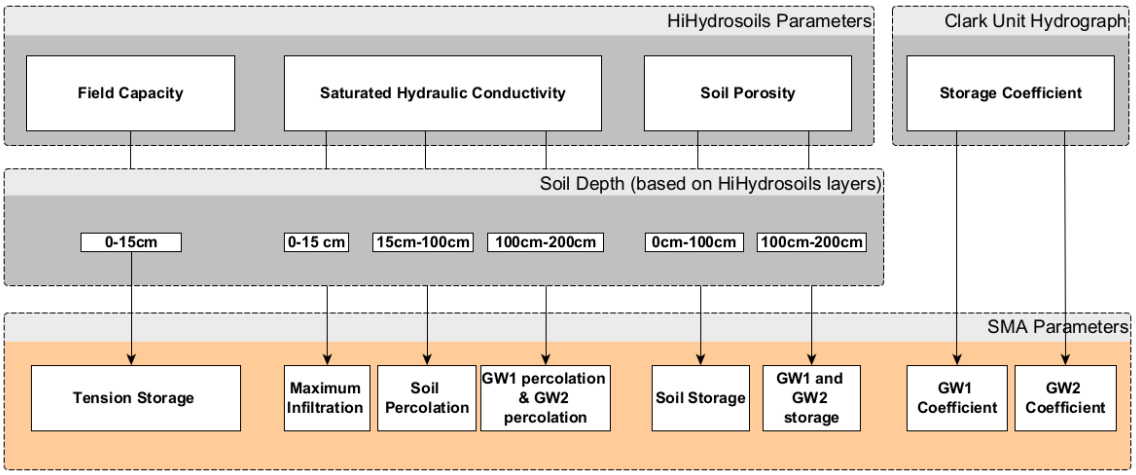


Figure 18: Parameters involved in the hydrological model

4.2.5 Results

Results of the supervised crop type mapping

For the evaluation of the models tested different metrics suited for multiclass classification were utilized. For the training of the neural networks the loss function selected is the Focal Loss function, while the evaluation metric used during training is the accuracy metric of the classification. Because relying only on the loss function and a single performance metric can be misleading, especially when dealing with imbalanced datasets.

Accuracy: Accuracy basically measures the overall correctness of the model without distinguishing between different types of errors like false positives or false negatives. If the dataset is imbalanced (e.g., 95% of the samples belong to one class), a model could achieve high accuracy by simply predicting the majority class all the time, even if it completely ignores the minority class.

$$Accuracy = \frac{Correct\ Predictions}{All\ Predictions} \quad \text{Eq. 3}$$

Precision: Precision measures the model's ability to identify instances of a particular class correctly.

$$Precision(ClassA) = \frac{TPa}{TPa + FPa} \quad \text{Eq. 4}$$

$$Precision\ (Macro - averaged) = \frac{Precision\ (Class\ A) + \cdots + Precision\ (Class\ N)}{N} \quad \text{Eq. 5}$$

F1 score: F1 score is the harmonic mean of precision and recall. It's a balanced metric that is particularly useful when you want to balance the trade-off between precision and recall, especially in cases where you care about both types of errors equally.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{Eq. 6}$$

Cohen's Kappa coefficient: Cohen's Kappa measures the agreement between the model's predictions and the actual labels while accounting for the possibility of chance agreement. This metric is particularly useful when you want to know if the model is genuinely performing well or if the performance is inflated due to the dataset's distribution.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad \text{Eq. 7}$$

where, p_o is the observed agreement, defined as the number of agreements divided by the total number of items, and p_e is the expected agreement by chance, calculated as the sum of the product of the marginal probabilities for each class, given respectively by:S

$$p_o = \frac{Number\ of\ agreements}{Total\ number\ of\ items} \quad \text{Eq. 8}$$

$$p_e = \sum_{i=1}^{C} p_{i,1} * p_{i,2} \quad \text{Eq. 9}$$

This combination of metrics will give an in-depth view regarding general accuracy, which shows the competence of the model to correctly identify minority classes while balancing false positives and false negatives.

Baseline Model: Regarding the results of the baseline model, the network seems to perform consistently across all the metrics presented. Regarding the accuracy, gamma = 3, provides the best accuracy, meaning highest number of correct classifications. The same is true for the F1 score and the Kappa coefficient. On the other hand, precision improves as the gamma values increase, gamma = 4 provides the best precision. Generally, gamma=3, provides the best balance across the metrics (Figure 19).

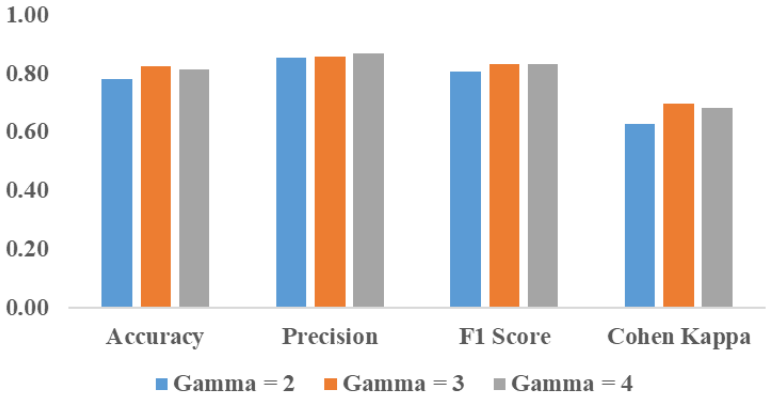


Figure 19: Baseline U-Net results

Compression of the time dimension: The second model, in which the time dimension is compressed, shows the same behaviour as the baseline model, but with slightly better performance metrics. Gamma = 3 has the best overall performance, while the model with gamma value equal to 4 is a close second, as it achieves better precision (Figure 20).

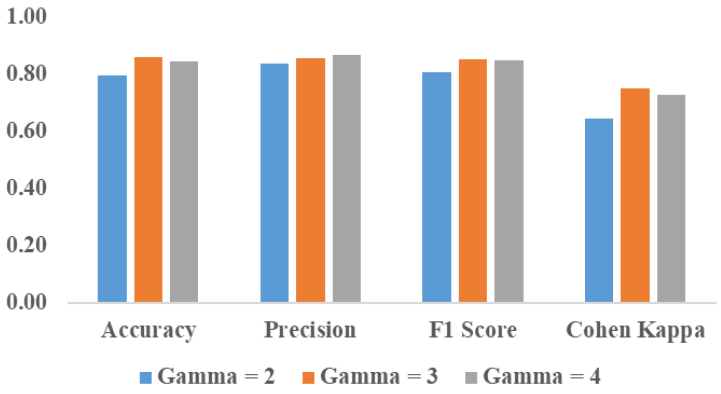


Figure 20: Time Compression U-Net Results

Attention at the skip connections

The third model, even though it mirrors some of the patterns observed in the results of the two previous models, achieves the highest precision for gamma = 2. Gamma = 4 is again close to the metrics of the model with gamma value equal to 3 but it doesn't outperform the latter (Figure 21).

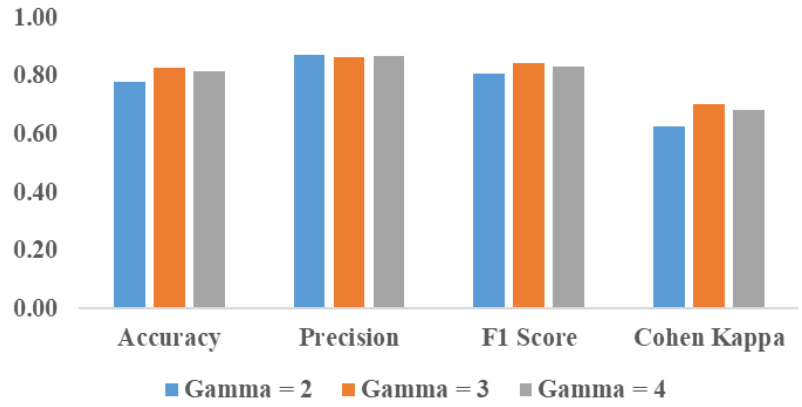


Figure 21: U-Net with attention at the skip-connections results

Comparison of the best performing model of each architecture:

Generally, for all three of the models, the gamma value equal to 3, seems to achieve the best results with respect to the evaluation metrics. The model that compresses the time dimension has the best accuracy and is followed by the benchmark model. In terms of precision, the model with the attention modules achieves the highest scores, meaning it is the most effective in reducing the false positive. In terms of F1 score, the second model performs better than the other two models, while the same is also true for the Kappa coefficient.

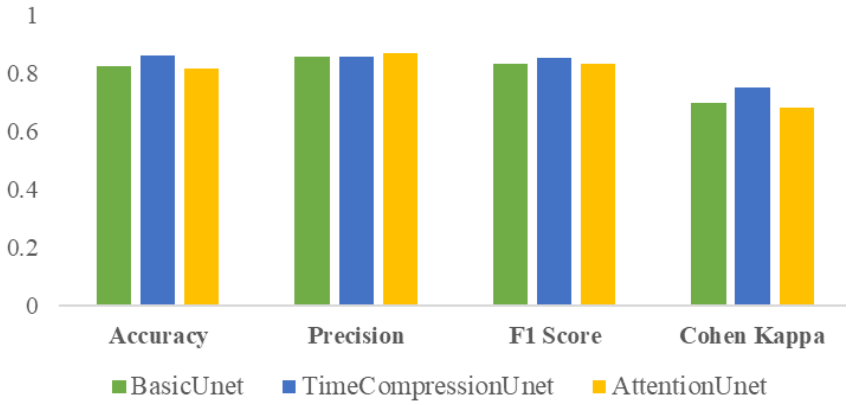


Figure 22: Comparison of the best performing models of each tested architecture

In conclusion, the benchmark model has a solid performance across all metrics, and the attention model performs the highest in terms of precision, but the model that compresses the time dimension seems to be the best performing model overall. Thus, the segmented map of the model that compresses the time dimension and for gamma = 3, is presented in the following figure (see Figure 22).

The three sets of metrics for training, validation, and test data highlight that class imbalance has a tangible effect on how well the model performs. This issue is especially visible in classes with fewer samples, like those for fallow fields and cereals, where we see a dip in precision, recall, and F1 scores. This likely points to the difficulty the model faces in learning consistent, general features when there's limited data to work with. While standard class-balancing techniques were applied, their results were mixed, leaving room for improvement in these smaller classes.

For the final version, the reconstruction logic used to stitch the model's output patches into the full segmented crop-type map was corrected, so that the predicted patches are now reassembled into a spatially coherent map without the edge artifacts present in the beta version. The updated crop classification map is shown in Figure 23.

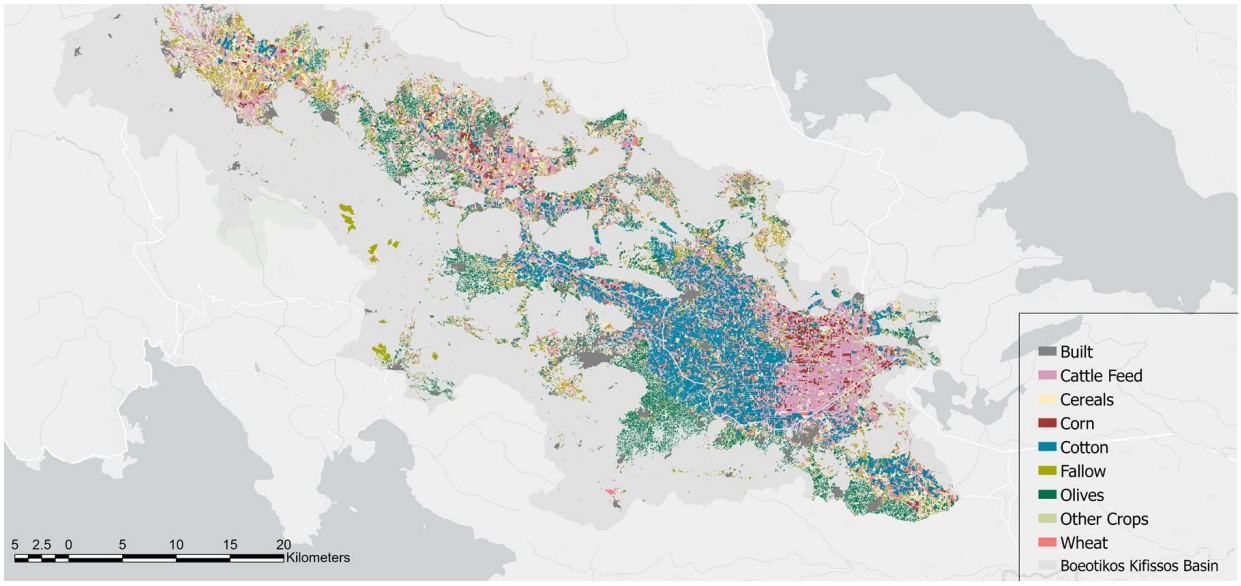


Figure 23: Segmented map of the crop types in study area

Results of the nutrient load estimation at the basin

The application of the proposed NNLI for nutrient load estimation in the basin provides a standardized and understandable method for quantifying nutrient levels. This approach enables a comprehensive assessment of nutrient loading on a monthly timescale, highlighting periods when the basin experiences the highest nutrient concentrations. The assessment can be conducted for the entire basin by computing the overall NNLI, but it can also identify the most nutrient-loaded areas, as the maps provide pixel-level information. The monthly NNLI maps presented below (see Figure 24) reflect the updates outputs produced for the final version of the early warning soft-sensor, regenerated after the correction of the patch reconstruction logic. The underlying NNLI values and the monthly aggregation framework are unchanged from the beta version. Only the spatial representation of the results has been corrected, producing maps that are more spatially coherent than those originally shown, without altering the numerical outputs of the nutrient load estimation.

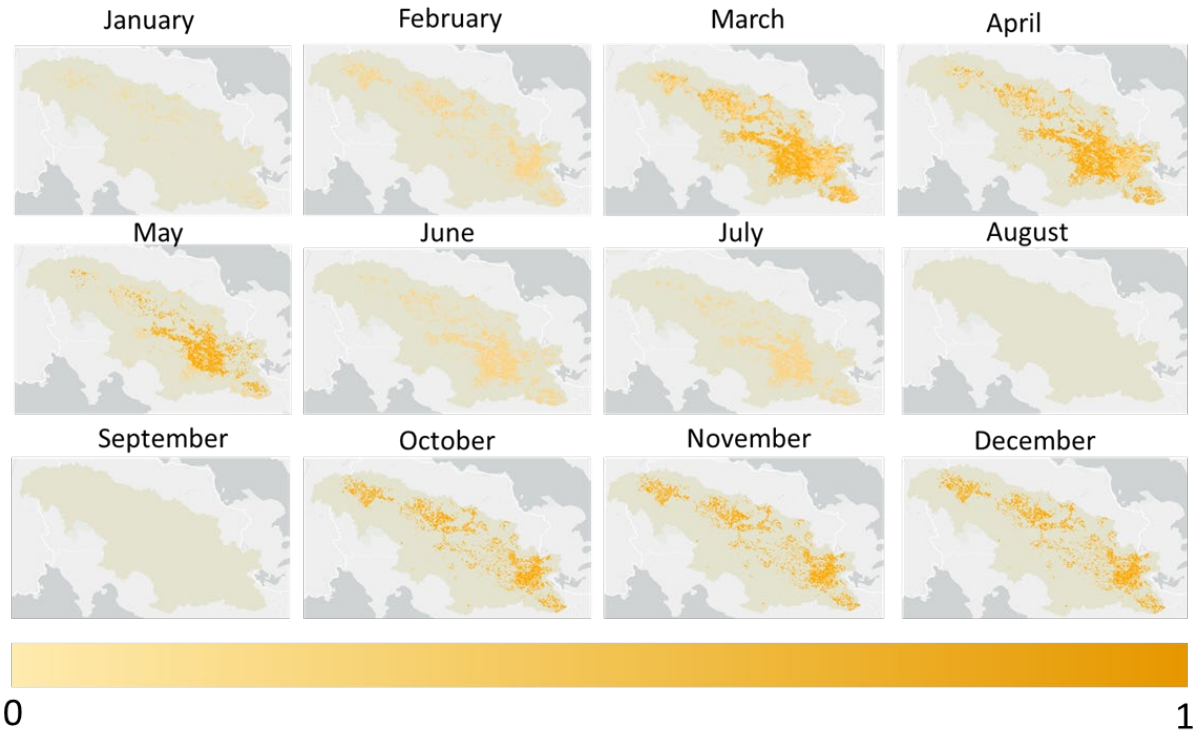


Figure 24: Monthly NNLI at the basin

Hydrological Modelling

The results of the EO-based calibration and validation are presented in the following paragraph. The model was first calibrated using in-situ precipitation and EO to simulate the soil properties (HiHydrosoils dataset) and the evapotranspiration losses (MODIS evapotranspiration). The NOAA forecast precipitation is used on the in-situ calibrated model for the validation period and the results for both simulations (in-situ and NOAA forecast precipitation) are evaluated based on their Nash–Sutcliffe efficiency (NSE) value.

Table 8: Evaluation criteria of the hydrological model

	NSE Calibration	NSE Validation	NSE validation (NOAA forecast)
HiHydrosoils-based model	0.606	0.569	0.389

Based on the results presented in the Table 8, the NSE calibration value indicates a moderately strong model performance for the calibration period, meaning the model captures more than 60% of the variance in the observed data. The validation for the in-situ precipitation indicates that the model performs reasonably well in the validation period too, with the small drop being attributed to the input of the new data. At the same time, the model's performance when using the NOAA forecast precipitation drops noticeably when compared to the in-situ validation values using the in-situ precipitation. This drop can be attributed to forecast inaccuracies and to the fact that the model was calibrated using in-situ precipitation from point stations and not gridded precipitation forecasts. Nevertheless, the model still performs better than random predictions.

4.2.6 Conclusion and next steps

Stepping from the beta version of this soft-sensor in D4.1 towards the current implementation, a number of improvements were made to enhance the early warning system. An issue in the patch reconstruction logic used to assemble the full segmented map from the model's output patches was identified and corrected. The bug that was identified and fixed between M24 and M36 was located exclusively in the post-processing patch reconstruction pipeline (the step that stitches the U-Net's output patches back into a full spatial map) and not in the hydrological model itself, nor in the U-Net's learning process. First, in the beta version, the reconstruction of the full crop-type map from the model's output patches was implemented using a simple overwrite strategy in which each predicted patch was placed directly onto the output canvas, overwriting any previously written values in the regions where consecutive patches overlapped. Second, in the beta version, a geotransform mismatch between the patch reconstruction output and the source image caused the predicted crop-type map to cover only a compressed central strip of the Boeotikos Kifissos Basin, leaving large portions of the cultivated area unclassified. In the final version, after propagating the source image geotransform directly to the output and correcting the patch averaging logic, the classification covers the full basin extent. The U-Net was trained on individual patches, and its patch-level predictive accuracy was unaffected by how those patches were later assembled.

The reconstruction process now stitches the predicted patches back together into a spatially coherent crop-type map. This way the artifacts that were present in the beta version's reconstruction output were eliminated. Following this correction, the full crop classification maps were The NNLI maps were revised, propagating the patch reconstruction fix. The remaining methodological components of the soft-sensor (the U-Net-based segmentation architectures, the loss function setup, the NNLI formulation, the agricultural practices mapping, and the HEC-HMS hydrological modelling framework) were preserved from the beta version as they had already been validated. The effect of these above-mentioned updates is more spatially consistent crop-type maps, that do not alter the underlying scientific approach of the early warning system.

Overall, the final version of the early warning system integrates a set of models that together describe the factors contributing to eutrophication in the lake. On the nutrient load side, a supervised crop-type classification model is combined with agricultural practices to quantify the nutrient load in the basin, providing stakeholders with valuable insights into when and where nutrient loads concentrate. In parallel, a hydrological model driven by EO data and precipitation forecasts serves as an essential tool for discharge estimation at the operational level. Combining these two components enables stakeholders to anticipate when nutrient runoff will occur and to prepare their treatment processes accordingly.

Further extension such as the mapping of crop types across multiple historical years to examine how shifts in crop patterns influence nutrient loads downstream, or refining the temporal resolution of nutrient load estimates from monthly to bi-weekly intervals, are recommended as directions for future research beyond the current deliverable.

4.2.7 Replicability potential

The replicability of the proposed early warning system relies on the fact that EO data is present in all aspects of the methodology. This methodology can be applied regardless of the study area or the basin characteristics, if certain in-situ data are available (e.g., in-situ crop type maps and discharge data).

4.3 Soft Sensor 2 - Chl-a concentration estimation

4.3.1 Problem statement and soft-sensor development flow-chart

Measuring Chl-a concentration is crucial, as it serves as an indicator for the among of photosynthetic plankton present in the lake. The amount of Chl-a in a collected water sample is used as a measure of the concentration of suspended phytoplankton. There are many ways to measure chlorophyll including spectrophotometry, chromatography and fluorometry. Spectrophotometry includes the collection of a large water sample, which then undergoes the filtration process, followed by the mechanical rapturing of the collected cells, and finally the extraction of Chl-a. This extract is then either analysed by the spectrophotometric method, or the chromatography method. The fluorometric method, also requires the same steps as the spectrophotometric method and then a fluorometer is used to measure the discrete molecular Chl-a fluorescence.

These monitoring methods (in-situ sampling followed by ex-situ analysis) have significant disadvantages. They are time consuming, costly and labour intensive as they require experienced and efficient analysis to generate accurate results. At the same time, they are not able to provide continuous monitoring as the collection of samples at reasonable time intervals would be extremely costly and time-consuming. A remedy to the above challenges of the efficient Chl-a concentration monitoring can be provided by the utilization of EO data. EO methods can produce higher volumes of data than conventional monitoring as it offers the potential for more spatially and temporally dense data collection to support estimates when used to augment in situ measures.

The estimation of Chl-a concentration is feasible using EO data as the Chl-a is a photosynthetic pigment found in phytoplankton which absorbs light in the blue and read part of the spectrum and reflects light in the green part of the spectrum. This absorption-reflection pattern creates a “spectral signature”. These absorption and reflection properties of Chl-a pigments are distinct and cause measurable variations in watercolour. Optical sensors on-board satellite missions are able to detect the spectral signatures associated with Chl-a in water bodies. In addition to the measuring of the reflectance, satellite imagery also provides frequent revisit time which makes it valuable for the timely and accurate estimation of Chl-a concentration without the need for extensive in-situ sampling.

4.3.2 Review of state-of-art approaches

In recent years, many studies that utilize remote sensing data for Chl-a concentration estimation have been proposed. (Barraza-Moraga et al., 2022), evaluated the use of Sentinel-2 MSI data for estimation

Chl-a concentration in a lake in south-central Chile using multiple linear regression. Nas et al. (2009) utilized data from Terra ASTER and in-situ Chl-a measurements to map the spatial distribution of Chl-a in the Lake Beyehir in Turkey, while (Moses et al., 2009), calibrated and validated a three-band and tow-band model using MERIS reflectance in the red and near-infrared spectral regions to estimation Chl-a. While Chl-a estimation using remote sensing has been extensively applied and various modelling approaches tested, robust methodologies for Chl-a estimation in inland water bodies remain in their early stages despite the increasing availability of remote sensing data and computational resources.

The methodological approach followed in this study aligns with state-of-the-art for environmental remote sensing, using remote sensing and deep learning to capture non-linear relationships. Beyond that, however, the development of the Chl-a soft-sensor showed that, in cases of highly imbalanced data, simply adopting a state-of-the-art model utilizing EO data as inputs is far from sufficient. Integrating imbalanced data handling helps ensure the generalization of the model while at the same time the application of explainable methods helps not only with interpretation of the model's predictions but also serves as an essential tool for scientific validation and feature selection, ultimately moving the field from using black-box models towards developing models that are reliable and interpretable.

4.3.3 Data sources and data preprocessing

For the development of the Chl-a concentration estimation soft-sensor, in-situ measurements and EO data are combined.

Data inputs

In-situ measurements: In-situ Chl-a measurements are obtained from the ISA Unmanned Surface VehicleBoat provided by EYDAP. The boat is equipped with the basic set of sensors for monitoring Dissolved Oxygen (DO), temperature, pH and electrical conductivity (EC), Chlorophyll-a (Chl-a), and is also equipped with the Intelligent Spectral Analyzer (ISA) which is a spectrophotometer that can identify parameters of interest such as total nitrogen (TN), orthophosphate (-PO_4) and nitrate (NO_3).

The sampling campaigns were conducted in various times from 2019 to 2022 during the INCATCH project (Katsouras et al., 2021). The dates of the in-situ sampling are presented on the following Table 9.

Table 9: Sampling Campaigns for water quality measurements in Yliki lake

2019	2020	2021	2022
February 11th	May 28th	May 24th	November 3rd
February 19th	August 5th		
March 6th	September 24th		
March 19th			
April 2nd			
April 22nd			
May 3rd			
May 29th			
July 24th			
September 25th			

EO data: The EO data used for the soft sensor development include data from the Sentinel-2 mission. Sentinel-2 provides high resolution, multispectral optical imagery. Sentinel-2A was launched in 2015, while Sentinel-2B was launched in 2017. Both satellites carry a Multispectral Instrument (MSI) that capture data across 13 spectral bands.

Preprocessing

The preprocessing methods applied in the development of this soft sensor include both the individual preprocessing of each data source and the combined processing of data sources through temporal and spatial alignment of in situ and Earth Observation (EO) data.

In-situ data pre-processing

Starting with the in situ Chl-a measurements, each sampling campaign was mapped to represent the different routes taken by the boat across the lake. Outlier removal was then performed based on expert judgment and a literature review of accepted Chl-a concentration values for oligotrophic-mesotrophic lakes, such as Lake Yliki. During some campaigns, recorded values occasionally reached as high as 200 µg/L. Expert knowledge indicates that Chl-a concentrations below 2.0 µg/L are categorized as "low," while concentrations between 2.0 µg/L and 8.0 µg/L are considered "intermediate." The upper accepted value for each campaign was determined based on the recommendations provided in the campaign experts' report.

Satellite Data Preprocessing

When acquiring data from remote sensing imagery, several factors can influence image quality. These include sensor characteristics, atmospheric and weather conditions, sun-glint effects, and potential sensor faults. To obtain more reliable information about the target, it is essential to preprocess the satellite imagery. This involves selecting satellite images that have passed quality control, ensuring only

those with approved quality are used. For the Sentinel-2 data used in the estimation of Chl-a concentration with a soft-sensor, preprocessing steps included filtering images based on cloud cover percentage. Only images with less than 15% cloud cover were selected, and a cloud mask was applied. Pixels containing cirrus or cumulus clouds were removed from all images used.

Water body extraction

Extracting the shape of a water body is essential when using satellite data for estimating water quality parameters, as it serves several crucial purposes in the data processing workflow. Dynamically extracting the shape of the waterbody allows excluding land in the coastal regions which would introduce noise and inaccuracies in the water quality parameters. Additionally, defining the precise shape of the waterbody helps avoid edge pixels which might capture both land and water. Finally, because the water level in the Yliki lake presents high variability dew to the underground water losses, the area and shape of the water body fluctuate, potentially exposing or covering new areas. Defining the current water boundary helps accurately assess the area and volume of water at a specific time.

Temporal and Spatial alignment

Temporal alignment involves synchronizing the timing of the satellite image with the date of in-situ sampling. To achieve this, a 3-day time window before and after the in-situ measurement is used to select the satellite pass closest to the sampling date. This alignment is crucial because water quality indicators, such as Chl-a levels, exhibit significant variability. According to the literature, a time window of up to 7 days is recommended to ensure meaningful alignment between in-situ and satellite data, thereby enhancing the accuracy and relevance of the model development.

Spatial alignment refers to ensuring that satellite data matches exactly with the locations of the in-situ measurements. This is important because water quality can differ a lot within a single body of water due to currents, varying depths, or localized pollution sources. For that, after the in-situ campaigns were mapped, for each instance, the value of the corresponding pixel of the Sentinel-2 mission was extracted.

4.3.4 Materials and methods

For the estimation of Chl-a concentration in the Yliki lake, in-situ measurements are combined with satellite observations for the timely and accurate estimation of Chl-a in the lake. The developed flowchart is provided in Figure 25.

Feature Extraction

For the development of the Chl-a soft-sensor the reflectance values recorded by the MSI sensor on-board the Sentinel-2 satellite. More specifically, starting with the visible light bands, the 10 bands used for this development include:

1. B2 (Blue, 490 nm) due to the strong absorption peak that Chl-a presents in the blue part of the spectrum.
2. B3 (Green, 560 nm) as Chl-a reflects green water, which is why waterbodies rich in phytoplankton appear green.
3. B4 (Red, 665 nm) which is one of the most important bands both for the inland and the coastal Chl-a estimation application, because of the second absorption peak that Chl-a presents in this part of the electromagnetic spectrum.

In addition to the visible part of the spectrum, the red-edge and near-infrared bands are able to capture key reflectance features caused by Chl-a.

4. B5 (Red Edge 1, 705 nm) follows the absorption in the red part of the spectrum. The height of the absorption peak captured by B5 is directly related to Chl-a concentration.
5. B6 (Red Edge 2, 740 nm) and B7 (Red Edge 3, 783 nm) further define the shape of the reflectance peak that starts and B4 and B5. These bands are commonly used as ratios (B7/B6) or in combination with B4 and B5, in many common algorithms such as the Normalized Difference Chlorophyll Index (NDCI).
6. B8 (NIR, 842nm) and B8A (Narrow NIR, 865 nm) provide a baseline dark reference and are used to separate land from water, as well as help in the detection of suspended sediments.

Finally, the bands representing the Near- and Short-wave Infrared bands of the spectrum, these include B11 (SWIR 1, 1610nm) and B12 (SWIR 2, 2190nm). These two bands are often used for sun glint correction algorithms which is described as the removal of the specular reflection of sun light from the water surface, which can overwhelm the water-leaving radiance signal. These algorithms are based on the assumption that water completely absorbs all light in the Near-Infrared (NIR) or Short-Wave Infrared (SWIR) spectrum, meaning any signal detected in these bands is assumed to be glint or atmospheric noise. This assumption though is frequently violated in optically complex water bodies. In highly turbid waters, shallow areas and especially in areas with visible algae scum NIR and SWIR parts of the spectrum reflect light. For the above-mentioned reasons traditional deglinting pre-processing is skipped and B11 and N12 are used as inputs for this development as the target variables (Chl-a) is related to variables that may cause high reflectance values in these two bands. The remaining bands (B1, B9, B10) are excluded entirely because they contain no surface information.

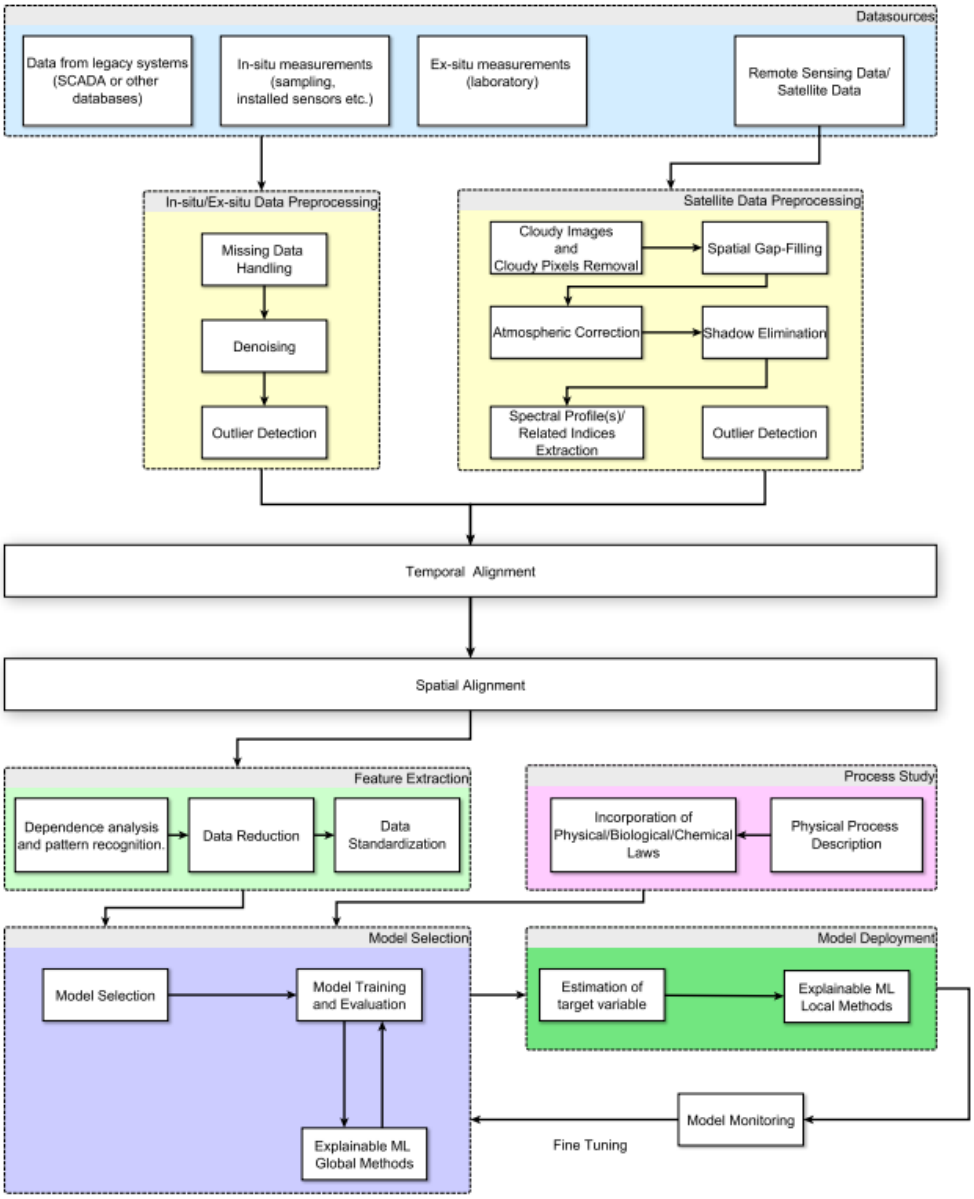


Figure 25: Chl-a estimation flowchart

Model Development

Imbalanced data handling

The Chl-a soft-sensor development involved several steps prior to final model selection. As shown in the figure, the dataset is heavily skewed toward values between 2-4 μg/l, while higher values (>8μg/l) are significantly underrepresented (see Figure 26 and Table 10)**.** This imbalance can cause the model to overfit to the most common values and fail to generalize well to the less frequent, higher concentrations. For that reason, imbalance data handling methods for regressions were tested. Two models were thus developed, one that didn't take into account the imbalance of Chl-a and one that did.

The model that addressed the skewed data distribution was inverse-frequency weighting. This approach discretized the target variable (Chl-a) into fixed width bins (2μg/l) and counts the number in each bin. Afterwards, a weight proportional to the inverse frequency was calculated. These weights were then reduced, using the square root to prevent the extreme influence of the rate samples. Finally, each

individual sample in the dataset was assigned the resulting weight corresponding to its bin. This way the model is able to place greater emphasis on fitting the under-represented Chl-a values.

Model Hyperparameters and Architecture

The model parameters for each Chl-a prediction model were selected using hyperparameter tuning methods. Hyperparameter tuning is defined as the process of finding the optimal set of model configurations (e.g., number of layers, number of neurons, learning rate etc.) that yields the best performance. The ranges and values of each of these hyperparameters that define where the tuning algorithm is allowed to explore define the search space. For the Chl-a models a Bayesian optimization tuner was employed to find the best model configuration by minimizing the validation loss over a total of 30 trials. The search space of the models architecture including the tuning of the hidden layers (2 to 25) and for each of those layers the number of units (128 to 512 in increments of 32). The tuner also configured the optimization algorithm choosing between Adam and SGD**.** The learning rate and the momentum (in case of SGD) were also sampled. The ReLU activation function was used for all hidden layers as well as the final output layer.

Explainability

Explainability refers to a set of methods used to understand the decision-making process of complex models. Large neural network and machine learning models are often referred to as "black-boxes" since their internal processes are difficult to interpret. In particular, in the case of prediction of a water quality parameter like Chl-a, explainable methods are essential for scientific validation, and they are necessary for scientific validation since they rely on remote sensing and physics-based principles.

There are several explainable methods that interpret these black-box models. A global method is the Permutation importance that defines how much each input matters by estimating the decrease in model performance when the values of the input are randomly shuffled. For the Chl-a soft sensor, the application of the Permutation Importance XAI method served a dual purpose, since beyond validating the model's learned patterns as mentioned above, the XAI results were directly used for feature selection. In particular, two more focused models were developed based on this analysis, one with and one without imbalance handling, using only the top 5 most important inputs as derived from the XAI analysis, a reduction from the 10 inputs used in the original models.

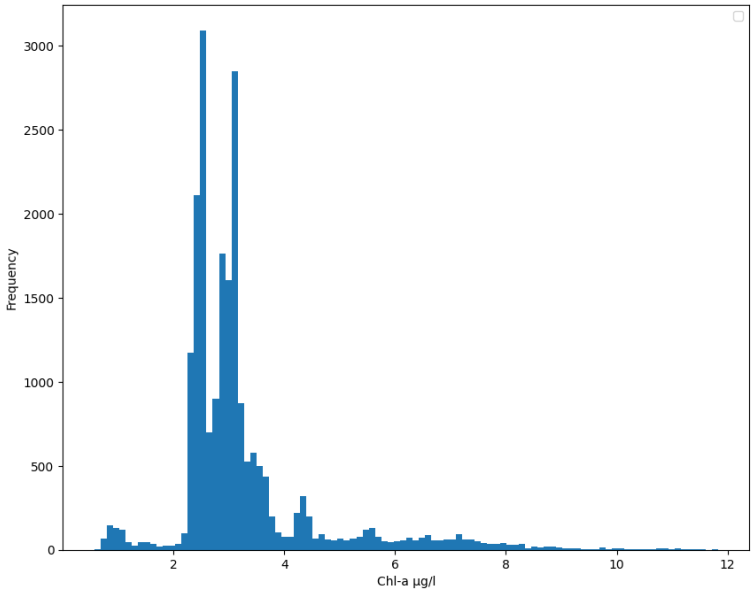


Figure 26: Histogram of Chl-a concentration

Table 10: Number of samples per Chl-a concentration range

Category	Chl-a concentration range	Number of Samples
Low	0-2 μg /l	954
Medium	2-8 μg /l	18,997
High	>8 μg /l	352

To address this issue, the weighted sampling method was employed. This approach assigns higher weights to the minority classes to oversample them and lower weights to the majority class to under sample it, ensuring a more balanced representation during model training. The classes were defined based on Chl-a concentrations, which were divided into seven histogram bins. The imbalance model, which performed better at predicting high concentration Chl-a values, degraded Band 5 (from 1st to 4th), and instead learned that Band 3 is the most critical feature, with importance almost double that of the next feature (Band 4).

4.3.5 Results

The results of the benchmark models (the models that use 10 band inputs) indicate a relatively strong performance in terms of prediction accuracy. The benchmark model that didn't account for the data imbalance (named "Basic" from now on) achieved R2 0.862 during training and R2 0.849 during testing. The benchmark model that did account for the imbalance in the dataset (named "Imbalance" from now on) achieved MAE 0.257, MSE 0.286 and R2 0.882 during training and MAE 0.232, MSE 0.220 and R2 0.848 during testing. While the overall metrics do not seem vastly in favour of the imbalanced-handling model (see Table 11) its improvement and superiority are much more obvious when the per bin error are inspected. The results of the per-bin error are presented in the Figure 27.

Table 11: Chl-a model results

	Training		Testing	
	Basic	Imbalance	Basic	Imbalance
MSE	0.259	0.22	0.284	0.286
MAE	0.258	0.232	0.268	0.257
R2	0.862	0.882	0.849	0.848

By analysing the per-bin error it is obvious that the basic model achieves a low average error by being accurate at the "most-common" bins, especially at the bin 2-4 μg/l, which mask its inability to generalize at the rare, high-values. On the other hand, the imbalance-handling model is able to trade a small loss in accuracy in common bins, especially 4-6 μg/l, for a big improvement in the rare, high values.

In more detail, the common bins in both training and testing both for the basic and imbalance models show similar MAE values, indicating that the models are equally effective in predicting the common Chl-a values. In contrast, for the 8-10 μg/l bin, the imbalance model achieves less than half the MAE that the basic model achieves, while in the 10-12 μg/l bin, the basic model fails, with testing MAE of 3.689. The imbalance model has its highest error in this bin but performs significantly better with testing MAE of 2.056.

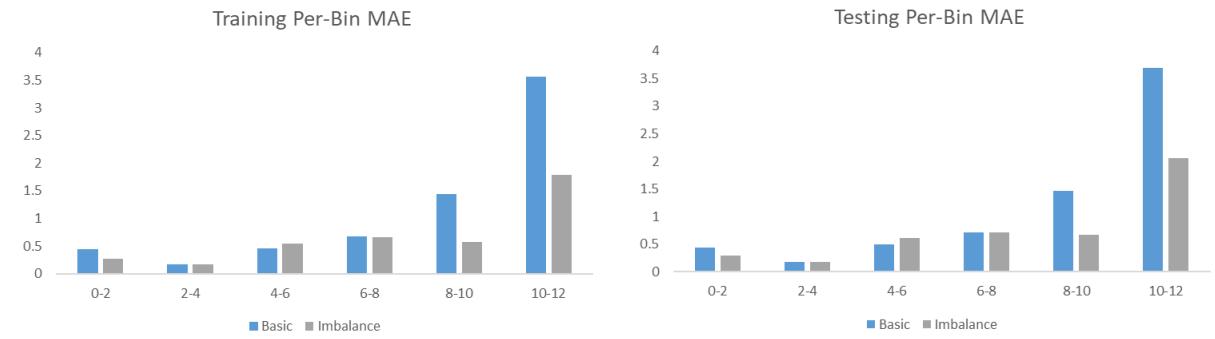


Figure 27: Per-bin Chl-a MAE

Explainability Results

The reults of the permutationfeature importance method for the basic and imbalance model are presented in Figure 28. First, both models rank as more importand the same 5 features, though the basic model ranks in different order B2 and B4 in training and testing suggesting that thes 5 bands contain the majority of the necessary signal for Chl-a prediction. Also, both on the training and testing predictions, the importance values are nearly idenical for every feature, indicating the stability of both the basic and the imbalance model and the relibaility of the results.

When examining the XAI results in more detail though, the order of the Permutation Importance results are different. The basic model, which fails to correctly predict the higher Chl-a values, learns to rely heavily on Band 5 (PFI = 1.76). The importance of Band 5 is more than 60% higher than the next feature (Band 3), suggesting that Band 5 is identified by the model as an adequate predictor of the most common Chl-a values (2-4 μg/l), but at the same time, a poor input for the higher Chl-a values.

The results of the explainability analysis align with the established remote sensing literature. The basic model which predicted adequatelly the common Chl-a consentration, indentified Band 5 as the most importand feature, which is consistent with the literature, as it is one of the most widely used bands for predicting Chl-a consentrations. The same is true for the imbalance model also, as Band 3 is the feature that represents the green reflectance peak, which is the fundamental definition of how Chl-a is identified from space. The imbalance model is forced to generalize across the entire Chl-a range, and thus learned the universal rules which states that more chlorophyll equals more green light reflected.

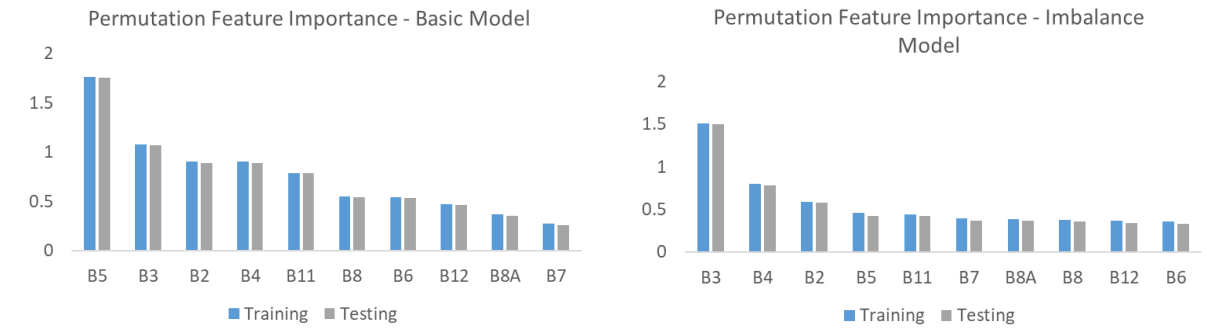


Figure 28: Permutation Feature Importance Results

Explainability for feature selection

The final two models developed used the results of the permutation importance method. The inputs in these two models were the same, as the permutation importance analysis for both the basic and the imbalance model identified Band 2, Band 3, Band 4, Band 5, and Band 11 as the most important bands, even though in a different order.

Table 12: XAI Models Results

	Training		Testing	
	Basic	Imbalance	Basic	Imbalance
MSE	0.315	0.422	0.372	0.489
MAE	0.289	0.332	0.31	0.36
R2	0.832	0.775	0.802	0.74

Based on the results of the 5-input models, presented in Table 12 the basic model which didn't account for the data imbalace seems to outperform the model that did. However, this image changes when inspecting the per-bin errors. This analysis again makes it apparent that the basic model's strong overall score is driven by its high accuracy on common values. Although the imbalance model shows somewhat poorer overall performance, it maintains its superior performance on the under-represented, high-value bins, an advantage masked by the averaged metrics (see Figure 29).

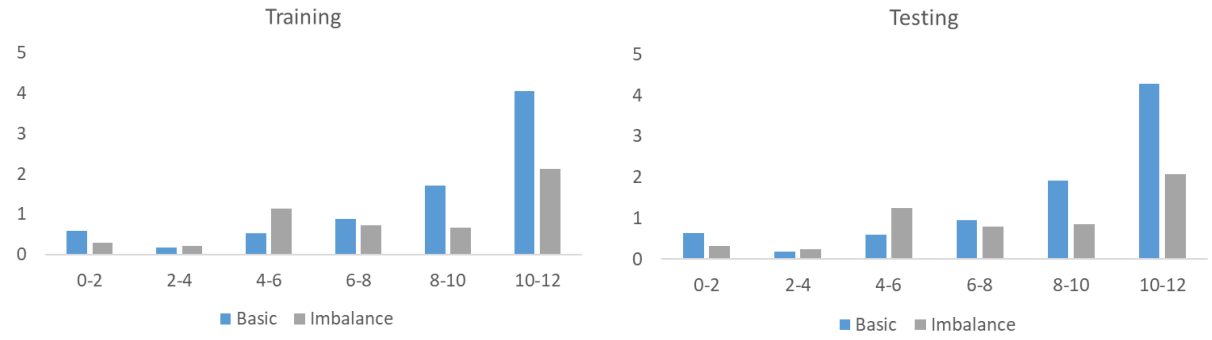


Figure 29: Per-bin MAE - XAI models

4.3.6 Conclusions

The final version of the Chl-a soft sensor was redesigned to improve its operational performance and interpretability. The input feature set was rebuilt around 10 raw Sentinel-2 bands, replacing the spectral indices and correlation-based selection used in the beta version. Imbalance handling was formalized through inverse-frequency weighting with square-root smoothing, and a controlled comparison framework was set up in which models with and without imbalance handling were trained in parallel. Manual architecture selection was replaced by Bayesian hyperparameter optimization, while per-bin error analysis and a complete Permutation Importance XAI pipeline were introduced.

Overall, the development of the of the Chl-a soft-sensor demonstrated that the standard aggregated performance metrics are insufficient and misleading for developing results from imbalanced environmental data. The models that didn't account for the imbalance in the Chl-a datasets achieved high average scores, but in both cases, deeper investigations revealed that it mostly predicted the most common Chl-a values (2-4 μg/l) and failed to generalize to the higher values. Thus, the implementation of the inverse-frequency weighting was proven to be essential, as it forced the model to learn the

underrepresented data in both cases. In addition to that, the explainability methods provided some key benefits as it not only validated that the models have learned the relationship that are backed up with the remote sensing literature but also served as a powerful tool for feature selection and model simplification. The analysis identified a clear set of inputs (B2, B3, B4, B5 and B11) that contained the primary predictive signal which allowed for the development of simpler but almost adequately powerful models. In Figure 30 some indicative maps of the are presented.

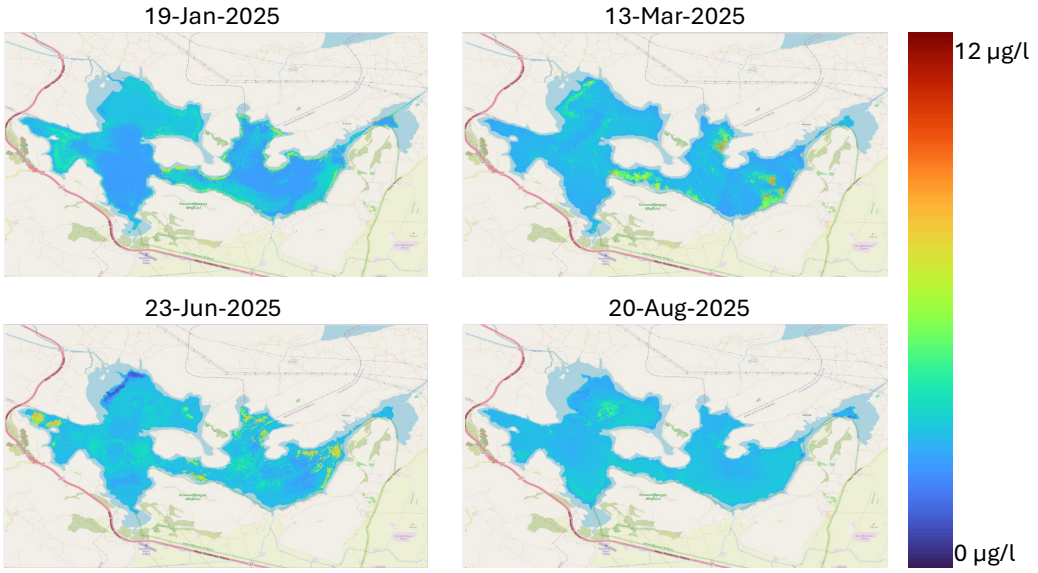


Figure 30: Indicative Chl-a soft-senor maps

Quantitatively, the beta version relied on spectral indices and correlation based feature selection with manually chosen architectures and achieved a training MAE of 0.295 µg/l, a testing MAE of 0.348 µg/l, and an R^2 of 0.871 on the test set. The final version that introduced a controlled two-model framework achieved a testing MAE of 0.268 µg/l and R^2 = 0.849, while the imbalance-handling model achieves a testing MAE of 0.257 µg/l and R^2 = 0.848, representing a reduction in testing MAE of approximately 23% and 26% respectively compared to the beta. Additional methodological advances include replacement of spectral indices with 10 raw Sentinel-2 bands as inputs, introduction of Bayesian hyperparameter optimization replacing manual architecture selection, and a full Permutation Feature Importance XAI pipeline for model validation and feature selection.

Further extensions such as cross-region validation on lakes outside the EYDAP demo case to assess geographic transferability, or the incorporation of convolutional layers to exploit spatial context in the satellite imagery are recommended as directions for future research beyond the current deliverable.

4.3.7 Replicability potential

From the perspective of replicability, the proposed workflow is replicable, starting from the steps of preprocessing to hyperparameter tuning and model training. This process can be repeated for any water body with different characteristics, and even the common and rare values are different from one another, ensuring the robustness of the workflow from beginning to end. This covers not only the development of the model but also the analysis and evaluation of the model's learned patterns through the application of explainability methods.

4.4 Soft Sensor 3 & 4 – pH and DO estimation

4.4.1 Problem statement and soft-sensor development flow-chart

The pH and DO soft-sensor were developed as new components of the EYDAP demo case, extending the soft-sensor monitoring beyond Chl-a to cover two additional water quality parameters that are critical both for ecosystem monitoring and drinking water treatment operations. The inclusion of these two variables also serves as the foundation for the subsequent development of the WQI soft-sensor, which combines the outputs of the Chl-a, pH, and DO models into a single operationally meaningful water quality score.

Dissolved Oxygen (DO) is defined as the amount of oxygen dissolved in the water. DO is measured in aquatic ecosystems because its depletion is one of the most destructive consequences of eutrophication and algal blooms. When algal blooms die, the bacteria decompose and the dead organic matter consumes big amounts of DO. This may lead to hypoxic conditions and “dead zones” These dead zones are uninhabitable for fish and result in mass fish deaths and thus a loss in biodiversity. In terms of the effect DO has in the treatment process, the dissolved metals in low DO concentrations may reduce the treatment efficiency and even lead to discoloured water. Moreover, pH, is the measure of water alkalinity or acidity and also affects aquatic organisms. Effectively monitoring of pH values is important not only because most aquatic species have a relatively narrow optima range, but also because it directly impacts some treatment stages, such as the coagulation and chlorine disinfection process, which is less effective at high pH values. This may lead the operators to use higher chlorine doses to achieve the same level of pathogen removal.

4.4.2 State of the art

The novelty of this work is related to the fact that the two soft-sensors developed predict two non-optically active variables using satellite data. This was achieved thanks to the capability of deep neural networks to overcome the lack of direct optical signal from parameters such as pH and DO by analysing the spectral information and identifying complex, non-linear relationships to estimate these parameters. Additionally, another important contribution of these developments is the incorporation of imbalanced data handling to the prediction process, as standard deep learning models might successfully predict common water conditions from satellite images, they may fail to predict outliers or uncommon events, which are also the most operationally significant.

4.4.3 Data sources and data preprocessing

The data sources and preprocessing steps employed for the development of these two soft-sensors (pH and DO) are the same as those utilized for the Chl-a soft-sensor and are extensively mentioned in Section 15.3. In sort, the input for the Soft-Sensors are the 10 Sentinel-2 bands, i.e., B2, B3, B4, ,B5, B6, B7, B8, B8A, B11, B12. The target values (pH and DO measurements) were sourced from EYDAP’s INCATCH Boat for the dates specified in Section 4.3. Similarly, the temporal and spatial alignment for these two soft-sensors was performed using the same methodology previously detailed for the Chl-a soft-sensor. The histograms of the two target datasets (pH and DO) are presented in Figure 31. Both datasets show significant distributional imbalances.

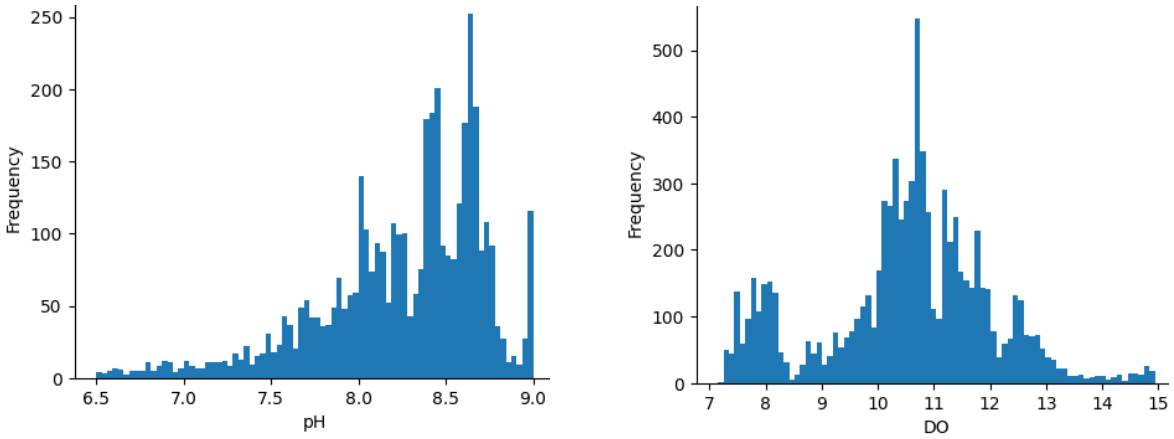


Figure 31: Histograms of pH and DO

The DO dataset is bimodal with two distinct clusters of DO concentrations, and even if the data between these two clusters is sparse a predictive model would have clear separate groups to learn from. In contrast, the pH dataset exhibits a left skew and a dense concentration of samples between 8 and 9 and a long tail of values below 8, and thus a lack of samples for the neutral or slightly acidic states.

4.4.4 Materials and methods

pH soft-sensor

For the development of the pH soft sensor, two different model architectures were tested, one that accounted for the data imbalanced described above, and one that didn't. The architecture of the model that didn't account for the imbalance in the data (Basic pH model) is a deep Multilayer Perceptron (MLP) whose specific architecture was derived from a hyperparameter tuning process to find an optimal configuration. The model consists of five fully connected layers (dense layers) and the final output layer. The model processes the input data from the Sentinel-2 bands (10 input features) through a deep stack of hidden layers (with 320, 480, 480, 288 and 480 respectively) and the output is compressed into a single neuron (output layer).

Similarly, the architecture of the model that did account for the imbalance in the data, also emerged from the hyperparameter optimization process. It's a deep Multilayer Perceptron, that consists of thirteen sequential dense layers. The network structure includes different sizes of hidden layers , from 256 to 416 neurons. The methods used to handle the imbalance in the data is inverse-frequency weighting, as described in the development of the Chl-a soft-sensor.

DO soft-sensor

The DO model architecture is also identified through the hyperparameter tuning process and is significantly deeper than the pH model. More specifically, this sequential neural network is composed of eleven fully connected layers. The model processes the 10 input features from the Sentinel-2 bands through a series of hidden representations (e.g., layers with 192, 64, 512 and 320 neurons) and outputs a single prediction (single neuron output layer) of the DO concentration prediction.

4.4.5 Results

Both soft-sensors achieved relatively good results, even though both pH and DO are non-optically active variables. In the following paragraphs the results of these two models are presented in more detail.

pH soft-sensor

During the pH soft-sensor development two models were compared, one that didn't account for the imbalance in the data, and one that did. The model that did account for the data imbalance, achieved a better fit on the training data it wasn't able to generalize as well as the basic model on the unseen test data. The mode that didn't account for the data imbalance achieved both higher R2 and lower MSE scores. These results suggest that the attempt to correct the imbalance in the data for the pH values have led to overfitting.

Table 13: Performance metrics of pH soft sensor

	Basic pH		Imbalance pH	
	Training	**Testing**	**Training**	**Testing**
R2	0.725	0.67	0.757	0.62
MSE	0.06	0.078	0.053	0.09
MAE	0.146	0.172	0.127	0.17

However, when examining the per-bin MAE (see Figure 32), some insights are revealed, and that is that the model that handled the imbalance was highly successful in its goal to improve the predictions in the underrepresented regions.

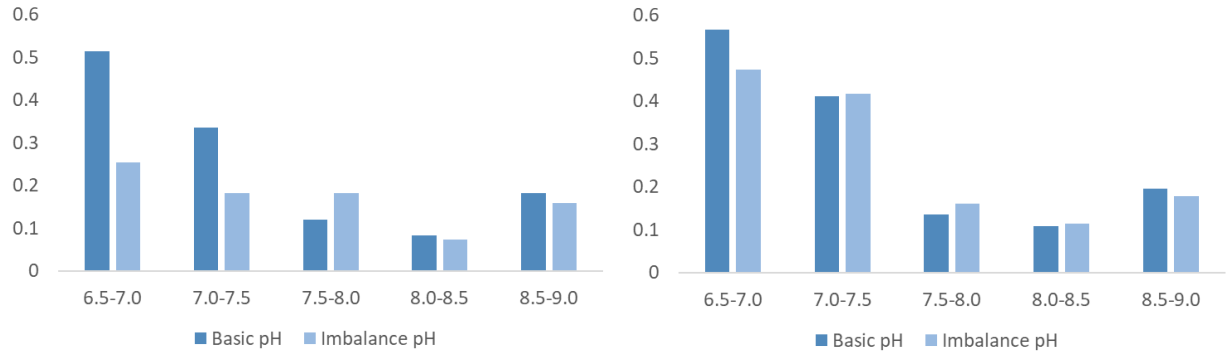


Figure 32: Per-bin pH MAE

On the testing set specifically, it achieved a lower MAE in the rarest bin (6.5-7.0). This means that the imbalance technique made the model learn from the tail in the data. Still, a trade-off appears, as its focus on the underrepresented data has decreased the accuracy in the most common bins, were the basic model performed better. Some indicative maps of the pH soft-sensor results are provided in Figure 33.

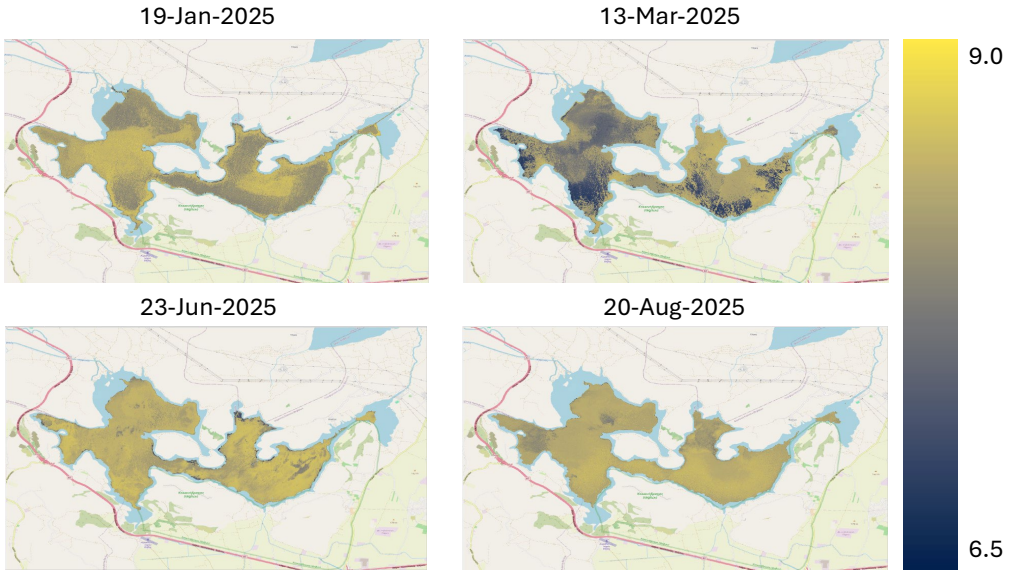


Figure 33: Indicative pH soft-sensor maps

DO soft-sensor

The results for the DO prediction model shows very good performance and strong generalization. The model achieved high R2 during training (0.914) and maintained an almost identical score during testing (0.899). The model evaluation results are presented in Table 14.

Table 14: DO model results

	Training	Testing
R2	0.914	0.899
MAE	0.242	0.272
MSE	0.192	0.232

This shows that the model was able to successfully explain 90% of the variance in the unseen data. In addition to that, the very small difference between the training and the testing error metrics shows that the model is robust and doesn't present overfitting behaviour. Some results of the DO soft-sensor are provided below (see Figure 34).

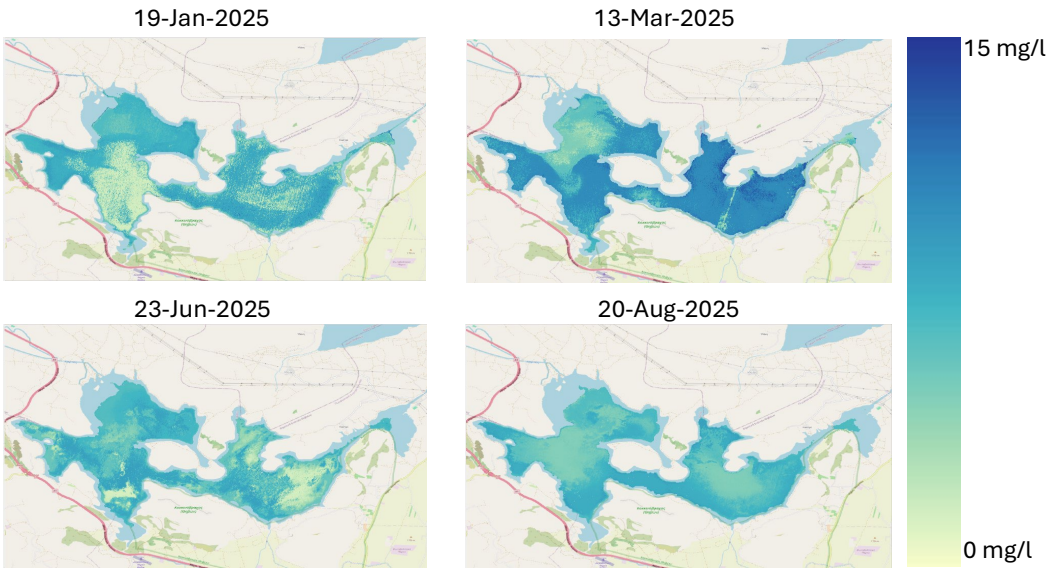


Figure 34: Indicative DO soft-sensor maps

4.4.6 Conclusions

Stepping from the beta version of this deliverable (D4.1) towards the current implementation, the pH and DO soft-sensors were introduced as entirely new developments and were not part of the beta version. Reusing the established, validated framework from the redesigned Chl-a sensor, the final versions of the pH and DO soft-sensor were built.

The development of the pH soft-sensor showed although typical modelling techniques can produce excellent performance, a more thorough analysis is necessary for the effective development of operationally dependable tools when dealing with imbalanced data. The main challenge had to do with the left skew of the pH data, in which low measurements were uncommon. In this case, the imbalance model was much more accurate in predicting the underrepresented values, even though the basic model has superior overall metrics. Because these low values represent important events, a model that is able to accurately predict these low values will be more robust in the operational context.

As this sensor was entirely new in D4.3 (not present in the beta), the progression is from 0% delivery to a fully trained and evaluated model. The final basic pH model achieves $R^2 = 0.670$ on the test set (MAE = 0.172, MSE = 0.078), while the imbalance-handling model achieves $R^2 = 0.620$ (MAE = 0.170, MSE = 0.090).

The imbalance model shows superior performance in the underrepresented low-pH range (6.5-7.0 bin), which corresponds to the operationally critical acid events.

For the DO soft-sensor, also a new delivery in D4.3, the final model (11-layer deep MLP derived from Bayesian HPO) achieves $R^2 = 0.899$ on the test set (MAE = 0.272, MSE = 0.232), with a training R^2 of 0.914, demonstrating strong generalization and negligible overfitting. This represents a full delivery from the conceptual stage in the beta version.

4.4.7 Replicability potential

The replicability potential of the pH and Dissolved Oxygen (DO) soft sensors is high, as they rely primarily on globally available, open-access Sentinel-2 Earth Observation data. The developed methodological framework, which combines deep neural networks with advanced imbalance handling techniques, serves as a transferable blueprint that can be adapted to other water bodies facing similar monitoring challenges. However, since pH and DO are non-optically active variables inferred through complex non-linear correlations, successful replication in new locations requires adequate local in-situ data to retrain the models and capture the specific environmental characteristics of the target site.

4.5 Soft Sensor 5 - Bloom Occurrence Probability Estimation, Floating Algal Index

4.5.1 Problem statement and soft-sensor development flow-chart

Monitoring algal blooms in eutrophic water bodies, like the Yliki lake, is crucial because it helps manage water quality, protect aquatic ecosystems, and safeguard public health. Because of the importance of floating algae for the environment in general and for water quality in particular, the spatiotemporal variability of floating algae has been extensively studied. Generally, algal blooms appear suddenly (D'Silva et al., 2012). Although nutrient overload is one of the main drivers of algal blooms, which can somehow be monitored and the nutrient load in the water bodies can be measured or estimated, there are other drivers which make the appearances of bloom events sudden. Some of these drivers are related to meteorological conditions (Zhang et al., 2016). Warm temperature, radiation, and calm weather conditions can accelerate algae production, while at the same time certain water currents and winds can concentrate algae in specific areas, making blooms appear more sudden and intense (Tian et al., 2018).

Satellite data have been previously utilized for the estimation of floating algae in inland water bodies. Traditional ocean colour algorithms have been extensively developed and utilized to quantify the water surface features in satellite imagery (Qi et al., 2020). The chlorophyll fluorescence product from MODIS sensor is a Level 2 product which contains parameters used to describe the ocean chlorophyll fluorescence properties and uses the sea-surface leaving radiance at the fluorescence wavelength of 683 nm as a relative measure of the fluorescence line height ("MODIS Chlorophyll Fluorescence (MOD 20)," n.d.). In 2006, Gower (Gower et al., 2006) first used MERIS and MODIS data to map the extensive surface slicks in the Gulf of Mexico. Moreover, the utilization of the Maximum Chlorophyl Index (MCI) algorithm and MERIS data provided the mapping and assessment of algae distributions in the same area (Gower and King, 2008). In 2009, Shi and Wang (Shi and Wang, 2009) analysed the development (and vanishing) of the 2008 Yellow Sea green tide (Zheng et al., 2022), by utilizing MODIS data and the Normalized Difference Algae Index (NDAI). On the same year Hu (Hu, 2009a) proposed the Floating Algal Index (FAI)

for the monitoring of green microalgae blooms (GMB). FAI has been reported to provide better and more stable results in comparison to other indices as it can detect various types of macroalgae like green and brown macroalgae and is also less sensitive to the decreased absorption of water in the NIR region (Oyama et al., 2015a), (Mu et al., 2021a). In the following chapter, the FAI index will be utilized as an estimation of floating algae in the study area.

4.5.2 Review of state-of-art approaches

The prediction of algal blooms has been approached through various methods. Luo et al., (2023), proposed the Vegetation and Bloom (VBI) algorithm to distinguish aquatic vegetation and algal bloom by utilizing a three-step classification process based on Landsat images. A similar approach was proposed by (Oyama et al., 2015b). The proposed Visual Cyanobacteria Index (VCI) classified cyanobacterial bloom levels into six categories based on field-measured threshold. At the same time, the FAI index, proposed by (Hu, 2009b) and originally developed for ocean environments offers robust detection of floating algae by leveraging its spectral sensitivity to environmental variations. Lastly, a Bayesian approach for bloom probability prediction has been previously proposed (Mu et al., 2021b)which integrated MODIS-derived FAI and meteorological data. In contrast, the proposed methodology integrates a broader range of environmental variables and meteorological forecasts from NOAA GFS, with satellite-derived FAI. These additional variables enhance the model's capacity to capture the complex drivers of algal blooms while the inclusion of forecasts enables the short-term prediction at operational level.

4.5.3 Data sources and data preprocessing

This section presents the data used for estimating the probability of floating algae occurrence. The dataset includes historical records of floating algae from the MODIS satellite, which are used to calculate the prior probability of algae presence in each pixel. Meteorological and discharge data (the results of the hydrological model) are also incorporated, along with an estimate of nutrient load in the basin, for which a Normalized Nutrient Load Index (NNLI) is proposed.

MODIS REFLECTANCE DATA

The model developed in this section is a predictive model which requires a significant amount of data for training, especially for the confident estimation of the prior probability. The meteorological conditions, including cloud coverage make the acquisition of satellite images with sufficient resolution difficult, and therefore the incorporation of high temporal resolution satellite images was necessary to obtain sufficient data for the model development.

The Moderate Resolution Imaging Spectroradiometer (MODIS) sensor, aboard NASA's Terra and Aqua satellites, has been operational since its launch in 1999. MODIS provides a daily revisit frequency and captures data across 36 spectral bands, with wavelengths ranging from 0.4µm to 14.4µm. The spatial resolution of these observations varies between 250m and 1km. MODIS data have been widely applied in numerous fields, including monitoring vegetation health via vegetation indices (Kirana et al., 2020; Kloos et al., 2021; Tripathi et al., 2014), tracking land cover changes (Lunetta et al., 2022; Usman et al., 2015; Yin et al., 2014; Zhan et al., 2002), assessing fluctuations in water levels of large inland water bodies (Khandelwal et al., 2017; Ling et al., 2020; Ovakoglou et al., 2016), and identifying wildfire outbreaks (Balch et al., 2020; Kaufman et al., 1998, n.d.; Raffuse et al., 2013).

In this study, MODIS data spanning from 2012 to 2022 were collected. After excluding images affected by clouds and thick aerosols, a total of 3,580 MODIS images were processed. The spatial resolution of the data was set at 250m following resampling. Lake Yliki encompasses 464 pixels of this 250m resolution in total. For each pixel, and for each instance the FAI was calculated and afterwards the bloom or non-bloom state was classified based on the FAI values. The index is calculated as the difference between the reflectance at 859 nm and the linear baseline between the red band (645 nm) and the shortwave infrared bands (1240 nm).

$FAI = R_{rc,NIR} - R'_{rc,SWIR}$ where Eq. 18

$R'_{rc,SWIR} = R_{rc,RED} + (R_{rc,SWIR} - R_{rc,RED}) * (\lambda_{NIR} - \lambda_{RED})/(\lambda_{NIR} + \lambda_{RED})$ Eq. 19

When the pixel's FAI value was greater than or equal to 0, it was classified as 1, indicating the presence of a bloom. Conversely, when FAI was below 0, the pixel was assigned a class variable of 0, indicating the absence of a bloom. This threshold was determined based on existing literature, where FAI = 0 has shown consistent results with visual interpretation of bloom presence.

Meteorological Data

Meteorological data used for estimating the probability of algal bloom occurrences were sourced from the NOAA GFS. This data includes daily forecasts of wind speed, temperature, relative humidity, total precipitation, and radiation, covering the period from 2020 to 2025 (see Table 15). These measurements represent forecasted values on a daily time scale.

Table 15: Variables used for the bloom occurrence probability estimation

Dataset/Model	Variable
NOAA GFS	Meteorological Variable Forecasts – wind speed, temperature, relative humidity precipitation and radiation
MODIS Reflectance Data	Floating Algal Index calculation

4.5.4 Materials and methods

The objective of this section is to develop a predictive model for algal bloom occurrence in Lake Yliki. The developed model utilizes an encoder-decoder architecture, which is a sequence-to-sequence architecture, suitable for time-series forecasting. The first component of the model, the encoder is built with Stacked Gated Recurrent Unit (GRU) Layers. GRUs are designed to handle sequential data, while stacking these layers allows the model to identify patterns and features of the inputs. The encoder receives the sequential meteorological data as inputs and compresses it (context vector) in order to represent the summary of the relevant past information.

The encoder part of the model receives two inputs. The meteorological forecast time series for 4 days and the per-pixel monthly prior probability of bloom occurrence per pixel. The meteorological input is important as algal blooms are heavily dependent on environmental conditions. The prior probability of bloom occurrence is derived from the FAI time series, as described in the previous paragraph. After calculating the probability, for each pixel corresponds 12 prior probabilities, one for each month, of bloom occurrence. The prior probability input is static and is used to show the model which areas are more prone to blooms.

The context vector derived from the encoder is then passed to the decoder. The decoder receives this summary and generates the output sequence, which is basically the corrected probability of bloom occurrence of each horizon. The model generates predictions for 3 horizons in total at the daily time step. After the first horizon (1 day ahead) is predicted, the result is used as input for the prediction of the next horizons. In order to train this model effectively, teacher forcing is used. By using this method, the true previous target (the actual observed bloom status from the previous day) is used during training and fed into the decoder to help predict the next step. Using this method helps stabilize the learning process while at the same time helps the model converge faster by preventing the model from compounding its own errors. This approach is not used at inference or evaluation, instead the model operates

autoregressively, feeding its own last prediction back into itself to "roll forward" and generate the bloom probabilities for the full 3-day forecast horizon.

4.5.5 Results

This GRU encoder-decoder model had a strong performance for predicting algal bloom occurrence probability. More specifically, for the 1-day forecast (+0d) the model achieved Area Under the Curve (AUC) metric of 0.765 and Area Under the Precision-Recall Curve (AUPRC) of 0.715, with a relatively balanced F1 score of 0.670. There was a drop in the performance for the 2-day forecast, but this was expected from the autoregressive inference method in which prediction error may compound. More specifically, the AUC metric falls to 0.721 and the AUPRC drops to 0.618. Regarding the three day forecast (+2d), the metrics improve and F1 score of 0.663 is achieved. The model results are presented in detail in the table below (see Table 16).

Table 16: Per-horizon soft-sensor results

Horizon	AUC	AUPRC	Precision	Recall	F1 score
+0d	0.765	0.715	0.591	0.772	0.670
+1d	0.721	0.618	0.535	0.861	0.660
+2d	0.725	0.622	0.541	0.856	0.663

Another interesting finding is the models tendency to prioritize Recall over Precision in all three horizons. This means that in general, the model is able to identify most of the actual bloom events, as it presents low false-negative rates, but this comes at the cost of the higher false-positive rate were the model is more likely to identify a non-bloom state of a pixel as a potential bloom.

4.5.6 Conclusions

The Bloom Occurrence Probability soft-sensor was redesigned between the beta version in D4.1 and the final version presented here. The Categorical Naïve Bayes classifier used in the beta version, was replaced by a GRU encoder-decoder neural network, a sequence-to-sequence deep learning architecture designed for time-series forecasting. Additionally, the input pipeline was simplified. The model now receives as inputs raw sequential meteorological data directly, with the prior probability derived from the MODIS FAI time series repurposed as a static per-pixel, per-month input giving the encoder a spatial conditioning signal. The meteorological data window was also updated from 2012–2023 to 2020–2025 to align with the operational forecast horizon.

In conclusion, the developed soft sensor proved capable of predicting the probability of algal bloom occurrence over a three-day horizon. The model achieved good performance by utilizing daily meteorological forecasts in combination with monthly bloom event probabilities. These inputs are essential for ensuring the model's accuracy, as bloom events are influenced not only by meteorological conditions but also by the specific location within the water body (e.g., nearshore or downwind areas). The performance metrics were good for the 1-day forecasts (+0d), and even with the expected drop dew to the autoregressive prediction approach, remained relatively stable for the next two days of forecast horizons. The most critical operational feature of the soft-sensor is its clear tendency to maximize Recall and this way making sure that the majority of the true bloom events are identified. This trade-off between Recall and Precision even though reveals a slight weakness of the model to identify non-bloom events is

helpful at the operational level because the cost of not being able to predict a true bloom event is higher than the cost of a false alarm.

The beta version (D4.1) used a Categorical Naïve Bayes (CNB) classifier with an overall accuracy of 0.552. The CNB model showed a precision of 0.51 and recall of 0.73 for the non-bloom class (class 0), and precision of 0.65 and recall of only 0.42 for the bloom class (class 1), yielding F1-scores of 0.60 and 0.51 respectively, indicating a model that struggled to detect actual bloom events. The final version replaces the CNB entirely with a GRU encoder-decoder neural network, a sequence-to-sequence deep learning architecture that additionally introduces a 3-day forecast horizon is not present in the beta. For the 1-day-ahead forecast, the final model achieves AUC = 0.765, AUPRC = 0.715, Precision = 0.591, Recall = 0.772, and F1 = 0.670. Compared to the beta, recall for bloom events improves from 0.42 to 0.772 (an increase of >80%) and the overall F1 for the bloom class rises from 0.51 to 0.670. Performance remains stable across the 3-day autoregressive horizon (AUC = 0.725, F1 = 0.663 at +2d), a forecasting capability absent in the beta version.

4.5.7 Replicability potential

The replicability of the Bloom Occurrence Probability soft sensor is exceptionally high, as it relies exclusively on globally accessible, standardized datasets: MODIS satellite imagery for the Floating Algal Index and NOAA GFS for meteorological forecasts. The developed GRU encoder-decoder architecture provides a transferable framework that can be deployed to other eutrophic water bodies simply by defining the new spatial boundaries, eliminating the need for expensive in-situ hardware. However, while the data sources are global, successful upscaling requires retraining the model with local historical data to ensure it captures the specific environmental drivers and bloom dynamics unique to the new target location.

4.6 Soft Sensor 6 – Water Quality Index

4.6.1 Problem statement

For the assessment of eutrophic water bodies, like the Yliki lake, the providing a comprehensive overview of water quality is essential as eutrophic water bodies need careful monitoring to prevent further degradation and manage the risks associated with eutrophication like hypoxia. The Water Quality Index (WQI) is one of the most used tools to describe water quality. It based on physical chemical and biological factors, combined into a single value. By combining the water quality parameters into a single value, the end users are able to describe the water quality state by only using a single value, basically serving as a thumbnail for the overall water quality. By incorporating different water quality parameters into a single thumbnail, more effective management can be achieved to mitigate the adverse effects of eutrophication.

4.6.2 Review of state-of-art approaches

The proposed methodology, which develops a water quality index (WQI) by training separate models for each water quality parameter using EO data as input, offers a novel approach compared to existing methods proposed. Najafzadeh et al. (2021), estimated the WQI in the Karun River while in (Najafzadeh and Basirian, 2023) he studies the WQI at the Hudson River. In these studies, various data-driven models have been tested, and satellite data have been used as input while in-situ measurements were used for

the model training. These methods focus on predicting WQI values directly. The utilization of satellite images and in-situ data has been proposed in previous studies and provides an important tool for the monitoring of the overall water quality of the water body as well as the spatial distribution of the WQI.

4.6.3 Data sources and data preprocessing

For the development of the WQI soft-sensor, in-situ measurements and EO data are combined, the same way as in the Chl-a estimation soft-sensor.

Data inputs

In-situ data: For this soft-soft sensor the in-situ measurements of the ISA EYDAP Boat are utilized. Specifically, the measurements of Chl-a, DO, and pH are used. The campaigns utilized for model development are the same as those for the development of the Chl-a soft sensor and are presented in Section 4.3.

EO-data: For the development of this soft sensor multispectral data are utilized. The multispectral data are those of the Sentinel-2 mission, the same way as for the development of the Chl-a, pH and DO soft sensor.

Preprocessing

For the WQI soft-sensor development, the pre-processing steps required are described in the previous sections (see section 4.3.4 and 4.4.4), including the development of the three models and the generation of the required input maps of Chl-a, DO and pH.

4.6.4 Materials and Methods

For model development, as previously mentioned, the workflow closely mirrors the process used in developing the Chl-a soft sensor. The flowchart is presented in Figure 35. In the “Selection of Water Quality Parameters” section of the proposed flowchart, the chosen parameters are Chl-a, DO and pH. The parameter selection is the initial step of the WQI process, presents significant variations between different WQI models, and is based on the characteristics of the water body and the measured variables. For the Yliki lake, the variables were selected based on the measured variables and the ability to model each of the selected variables through remote sensing data, based on the existing literature and their environmental significance.

In the two previous sections (section 15.3 and 15.4) the process of development and the results of the Chl-a, DO and pH soft-sensors are described. These three developments provide information about these three water quality variables for the whole surface of the lake. The results will be used as input to the WQI estimation model.

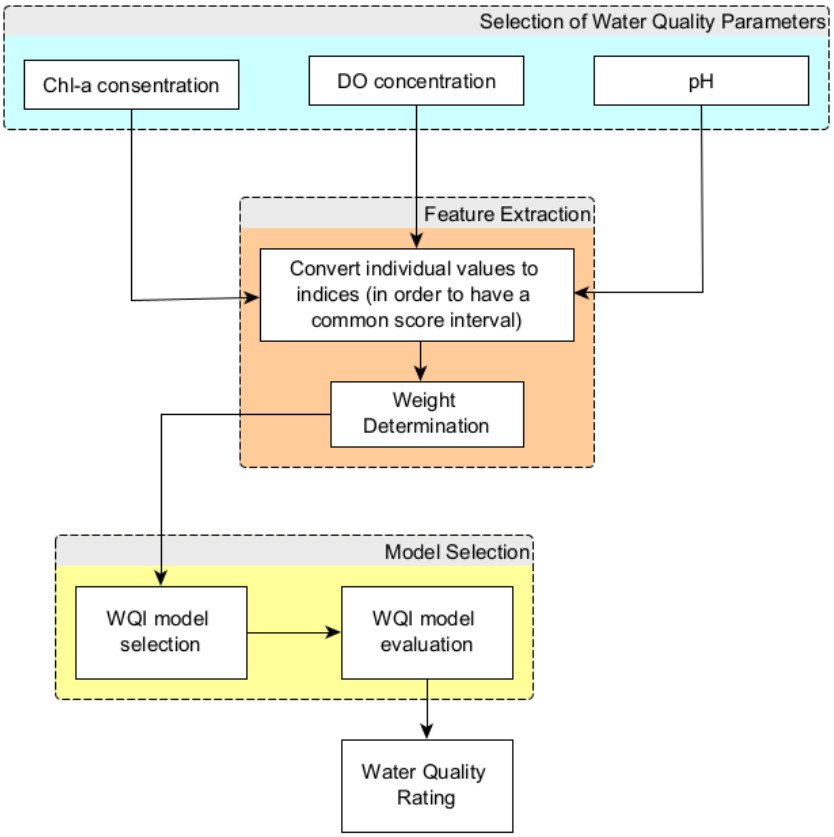


Figure 35: WQI estimation flowchart

After the estimation of these three water quality variables is complete through the three soft-sensors, the individual values of each pixel is converted into indices. This step is necessary because the extraction of the index values requires all variables to share a common scoring interval. The sub-indexing process typically employs linear interpolation functions or rating curve functions.

After converting the individual values to indices, the next step is weight determination. The weights for each variable are assigned based on their relative significance in the assessment process. Most WQI models use unequal weighting techniques, where the sum of all parameter weights equals 1. Following this, the WQI is computed by combining the individual parameter sub-indices, adjusted by their respective weights. A rating scale is then applied to classify water quality based on the overall index score. Through the aggregation function (Step 4), parameter weights can significantly influence the final index value. Therefore, WQI model robustness is best achieved by employing an unequal parameter weighting system and carefully assigning the most appropriate weights.

For the soft-sensor development for the Athens demo case the WQI is formulated as a weighted average of the three variables. Each variable is first converted into the standardized sub-index with values ranging from 0 (e.g., Chl-a > 8.0) to 100 representing the best conditions (e.g., pH > 7.7).

The final weights for the WQI were assigned at 0.05 for DO, 0.35 for pH and 0.6 for Chl-a because Chl-a is the main driver of eutrophication and thus received the highest weight, while DO was assigned the lowest weight due to its comparatively lower variability in the dataset.

4.6.5 Results

The WQI soft-sensor is basically a synthetic index and no ground truth data for the WQI are available in order to have a conventional performance evaluation. Thus, in this section, the results consist of the application of the WQI formula described above, to the predicted outputs of the three soft-sensors described in the previous sections. The output is a score for water quality over time and for each pixel covering the surface of Yliki lake. Some indicative results of the maps are presented in Figure 36.

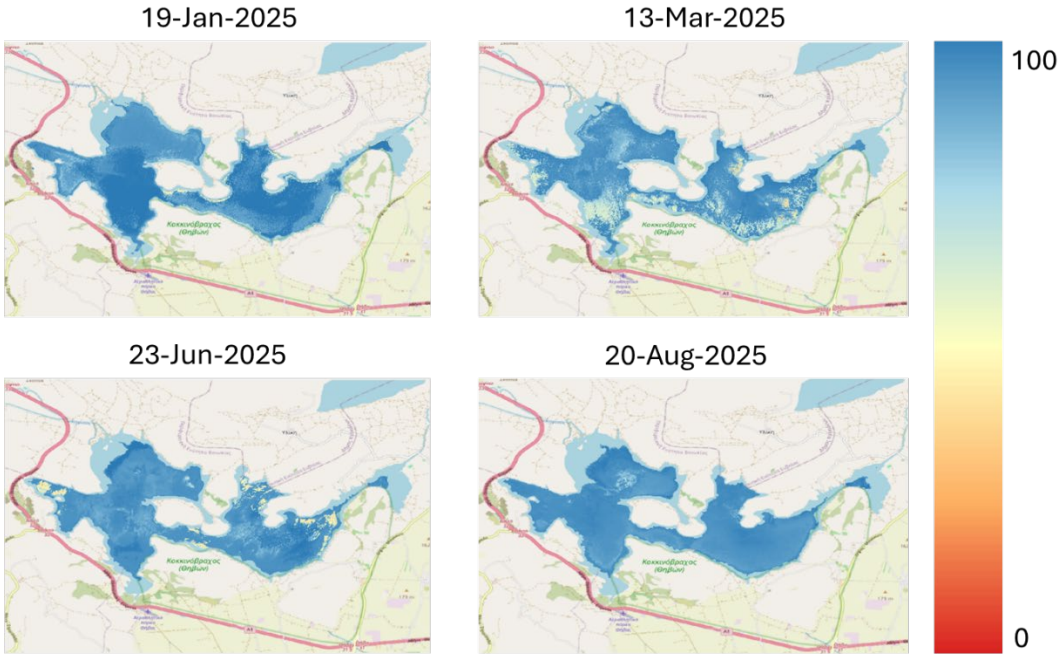


Figure 36: Indicative WQI soft-sensor maps

4.6.6 Conclusion and next steps

The WQI soft-sensor has changed from a planned development in the beta version (D4.1) to a fully implemented tool in the current deliverable. In D4.1 the WQI soft-sensor was described only at the level of a proposed methodology. For the current implementation, the soft-sensor was built on top of the three upstream soft-sensors developed in this deliverable, Chl-a, pH and DO, that provide pixel level input maps that the WQI requires. The parameter list was also refined as electrical conductivity was removed and the final index is computed from the three parameters linked to eutrophication and treatment operations. The methodological framework was simplified and made more transparent compared to what had been described in the beta version. Rather than training a dedicated Multi-Layer Perceptron for the WQI itself, as originally proposed, the final soft-sensor implements the WQI as a deterministic weighted aggregation of the upstream model outputs. This design choice makes the soft-sensor fully interpretable and easier to adapt to other water bodies by adjusting weights and sub-index rating curves to local conditions.

In terms of progress from beta to final, the beta version (D4.1) described the WQI only as a proposed methodology with no implementation. The final version delivers a fully operational weighted aggregation pipeline combining the outputs of the Chl-a ($R^2 = 0.848$), pH ($R^2 = 0.620$) and DO ($R^2 = 0.899$) soft-sensors with weights of 0.60, 0.35 and 0.05 respectively. While no independent ground-truth WQI exists, the upstream model performance metrics serve as the quantified basis for evaluating the composite index quality.

4.6.7 Replicability potential

The replicability of the Water Quality Index (WQI) soft sensor is inherently linked to the transferability of its underlying input models (Chlorophyll-a, pH, and DO), making it a highly modular and adaptable tool. The proposed methodological framework, which converts these individual parameters into standardized sub-indices and aggregates them via a weighted scoring system, is universally applicable to any water body regardless of location. However, successful replication requires adjusting the specific weighting factors and sub-index rating curves to align with local regulatory standards, environmental priorities, and the specific water quality characteristics of the new target ecosystem.

5 Soft Sensors for Val de Bagnes demo case

5.1 Introduction and Demo Case description

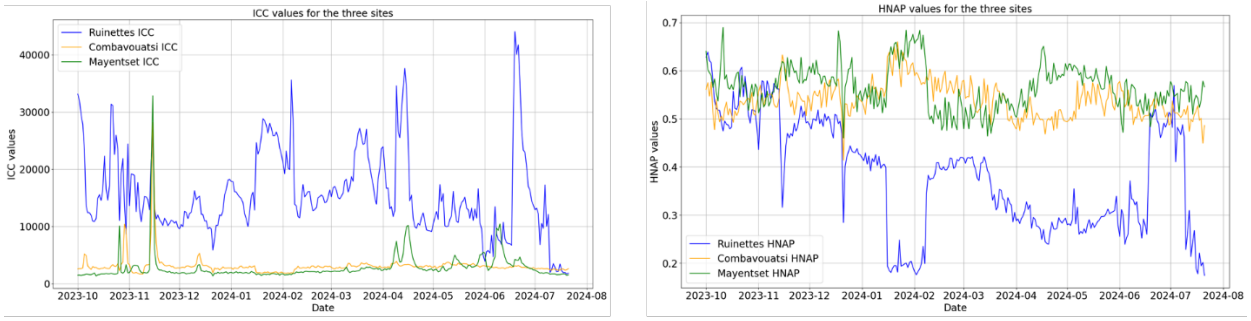
The Mayentzet demonstration case known for heavy snowfall in winter, is susceptible to bacterial contamination, especially during the snowmelt period. Combined with the absence of remotely connected sensors in the region, the challenge of monitoring water quality in real-time becomes even more pronounced. Given that this reservoir primarily serves the residents of Verbier, the crucial importance of effective water quality surveillance cannot be underestimated. Therefore, this demonstration case assumes critical importance in addressing the urgent need for enhanced real-time monitoring capabilities in remote mountainous regions. The challenge of this demonstration case was to improve real-time water quality monitoring.

This reservoir, primarily serving the residents of Verbier, is fed by three different sources: Mayentzet, Combavouatsi, and Ruinettε. The Mayentzet and Combavouatsi sources undergo no treatment before reaching the Mayentzet reservoir and these are the locations of interest of soft sensors development in the DC of Val de Bagnes. In comparison, the Ruinette source receives chlorination treatment before joining the Mayentzet reservoir. This decision was made while the first two sources are of excellent quality, so it is important to sustain drinking water in its natural state. Conversely, the Ruinette source is more unstable. However, the excellent quality of the water does not eliminate the risk of bacterial contamination, and since the Mayentzet reservoir supplies the Verbier area and many residents, it is Altis' duty to protect its population. Therefore, the primary challenge of this demonstration case is to enhance real-time water quality monitoring capability, a critical need for providing quality water and ensuring water safety in the event of contamination.

5.2 Soft Sensor 7 - Early Warning System for Bacteriological Contamination

5.2.1 Problem statement and soft-sensor development flow-chart

This development addresses the challenge of monitoring and predicting bacteriological contamination the water sources supplying the Mayentset reservoir that primarily serves Verbier. The reservoir is fed by three water sources, namely Mayentset, Ruinettes and Combavouatsi. Regarding the water quality of each water source, Combavouatsi and Mayentset sources deliver very good water quality. In contrast, the Ruinettes source is more unstable, with occasional bacteriological contamination events as presented in Figure 37. The goal of this project is to develop a soft sensor that proactively monitors and predicts contamination levels. Specifically, this work will examine the feasibility of using meteorological variables to predict bacteriological contamination events in untreated water.



Daily ICC Measurements

Daily HNAP Measurements

Figure 37: Daily BactoSense measurements

5.2.2 Review on state-of-art

The state-of-the-art contribution of this work lies in demonstrating that while meteorological variables alone are insufficient for regression-based prediction of specific bacteriological counts, they are highly effective within a classification framework. By shifting to a risk-based approach, the developed Random Forest model successfully links open-access meteorological data (ERA-5) to contamination events (HNAP), achieving high predictive accuracy (0.88). This establishes a novel, cost-effective pathway for monitoring remote Alpine sources where continuous biological sensing is often technically or financially unfeasible.

5.2.3 Data sources and data preprocessing

Data Sources

In order to develop this predictive model, two categories of data were collected: water quality measurements of Intact Cell Count (ICC) and High Nucleic Acis Percentage (HNAP) from the BactoSense sensor and reanalysis meteorological variables from the ERA-5 dataset. In addition, ALTIS provided measurements of three variables, namely precipitation snowfall and temperature used for the validation of the ERA-5 reanalysis variables. The available data sources include both in-situ and reanalysis data and are presented in the following table 17:

Table 17: Datasets for the early warning system for the Val de Bagnes demo case

	Variable	Time Scale
1	BactoSense Measurements of ICC and HNAP	6-hour sampling interval
2	Precipitation, Snowfall and Temperature (in-situ) provided from ALTIS	Hourly
3	ERA-5 Meteorological variables	Hourly

Preprocessing

Before being used as input to the models the ERA-5 meteorological variables were validated using the in-situ ALTIS meteorological measurements. This validation showed that the ERA-5 data are reliable and showed a strong correlation with the in-situ meteorological data (e.g., R = 0.96 for Temperature).

Regarding the variable selection, from the available ERA-5 variables some of the most important features were selected including Snow Cover, Temperature, Precipitation, Snow Melt, Snow Depth and Runoff.

5.2.4 Materials and methods

This development used a two-stage methodological approach. In the first stage the problem was addressed as a regression problem while on the second stage the bacteriological contamination prediction problem was addressed as classification of the level of the bacteriological contamination.

For the regression approach, the goal was to predict the continuous value of ICC and HNAP for which three different models were developed and compared:

- ARIMA-X, serving as the benchmark model. For this approach, the traditional time series forecasting model was used as the baseline to capture linear trends and relationships with the meteorological variables.
- Long Short-Term Memory (LSTM) Neural Network which is a type of neural network (RNN) that is designed to handle complex non-linear and long-term dependencies in time series data
- Gated Recurrent Unit (GRU), which is a more computationally efficient RNN alternative to LSTM also used for modelling long-term dependencies.

In contrast, for the classification approach the problem was defined as the prediction of the category of the bacteriological contamination in the reservoir (e.g., low or high) and not as the prediction of the exact continuous value of the contaminants. For this approach, the methodological approach included the variable selection process, the mode selection and finally the feature importance using XAI methods.

- For the feature selection process, the cross-correlation analysis was employed in order to select the most predictive variables and their time lags. This analysis used a 30-day maximum lag, and only variables with cross-correlation value greater than 0.15 at any lag were included.
- The Random Forest Classifier was chosen as a powerful ensemble learning algorithm that builds multiple decision trees and uses the majority vote to determine the final class.
- The Gini impurity importance was used as the explainable method for the evaluation of feature importance by measuring how much each feature helps to create more accurate classification when splitting the data.

Two separate classification experiments were conducted: a 2-Class classification and a 3-Class classification experiment.

For the 2-class classification the target variables were split into two classes: Class 0 and Class 1 based on the median values of the two target variables ICC and HNAP:

- ICC threshold: 14150.0
- HNAP threshold: 0.397 (39.7 %)

For the 3-class classification the data were split into three classes: Class 0, Class 1 and Class 2 based on the median and 85th percentile of the target variables:

- Median (1st threshold): 14150.0 for ICC and 39.7 for HNAP
- 85th percentile (2nd threshold): 23914.75 for ICC and 0.507 (50.7) for HNAP

5.2.5 Results

Regarding the regression models, they were very unstable. The attempt to predict the exact continuous values of ICC and HNAP in water using only meteorological variables failed. The model examined were not able to predict the sudden spies of the contamination events (e.g., see Figure 38 that represent the results of the ARIMA-X model for the regression task).

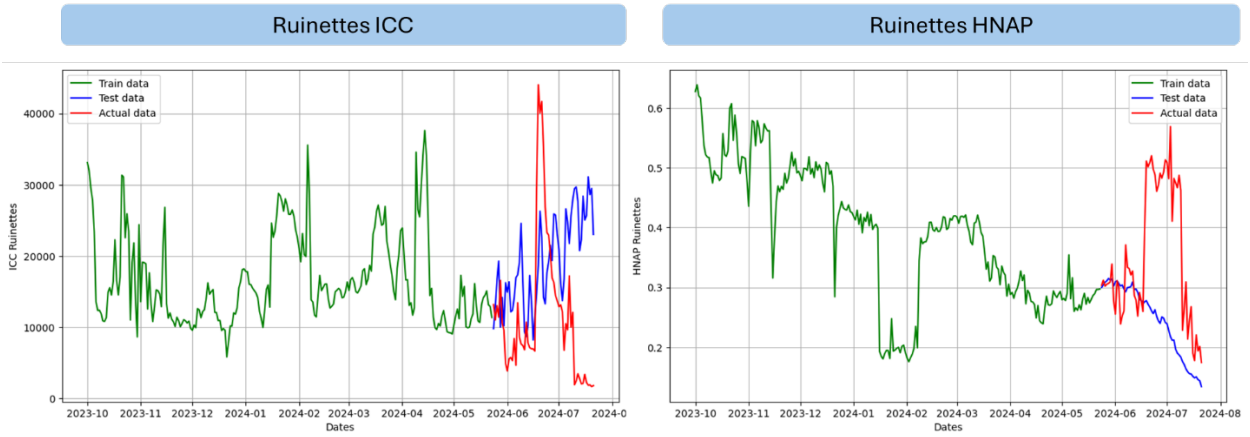
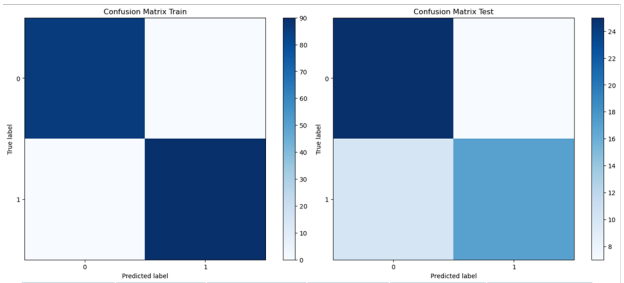


Figure 38: ARIMA-X time-series results for ICC and HNAP

On the other hand, the classification models were successful, and the RF Classifier provided valuable results. More specifically, the 2-class classification model for HNAP was most successful and achieved accuracy of 0.88 on the testing data. Additionally, the high F1 scores, (see Figure 39) indicate that the model is reliable for predicting low and high levels of bacteriological contamination. Regarding the 2-class classification model for the ISS classification it also performs adequately well and achieved 0.712 accuracy on the testing data.

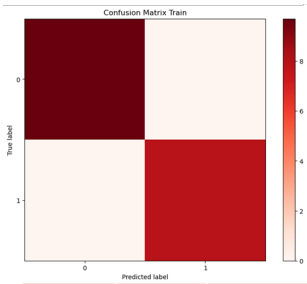
Ruinettes **ICC** Training



	Class	Precision	Recall	F1-Score	Accuracy
Training	**0**	1	1	1	1
	1	1	1	1	
Testing	**0**	0.71	0.78	0.75	0.712
	1	0.71	0.63	0.67	

Ruinettes **ICC** Testing

Ruinettes **HNAP** Training



	Class	Precision	Recall	F1-Score	Accuracy
Training	**0**	1	1	1	1
	1	1	1	1	
Testing	**0**	0.86	0.83	0.84	0.88
	1	0.89	0.92	0.90	

Ruinettes **HNAP** Testing

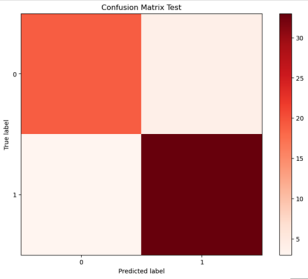
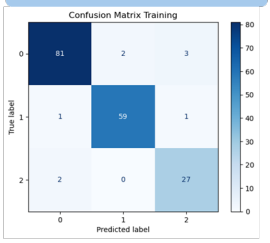


Figure 39: 2-class classification results for ICC and HNAP

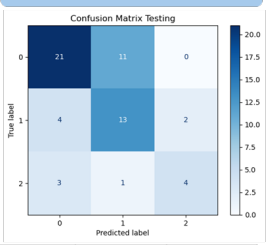
When it comes to the three-class classification problem, a performance drop can be identified, as there was a drop in accuracy in comparison to the 2-class classification problem (HNAP: 0.746 and ICC: 0.644). This indicates that the models are able to predict the shift from the median values, they struggle to differentiate between high and very high events (Figure 40).

Ruinettes **ICC** Training

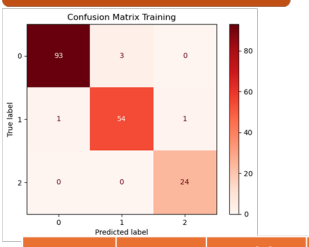


	Class	Precision	Recall	F1-Score	Accuracy
Training	0	0.96	0.94	0.95	0.949
	1	0.97	0.97	0.97	
	2	0.87	0.93	0.90	
Testing	0	0.75	0.66	0.70	0.644
	1	0.52	0.68	0.59	
	2	0.67	0.50	0.57	

Ruinettes **ICC** Testing



Ruinettes **HNAP** Training



	Class	Precision	Recall	F1-Score	Accuracy
Training	0	0.99	0.97	0.98	0.97
	1	0.95	0.96	0.96	
	2	0.96	1	0.98	
Testing	0	0.83	0.83	0.83	0.746
	1	0.70	0.73	0.72	
	2	0.67	0.60	0.63	

Ruinettes **HNAP** Testing

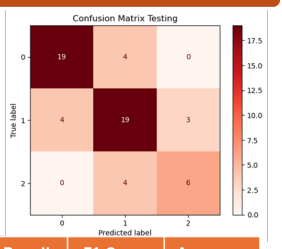
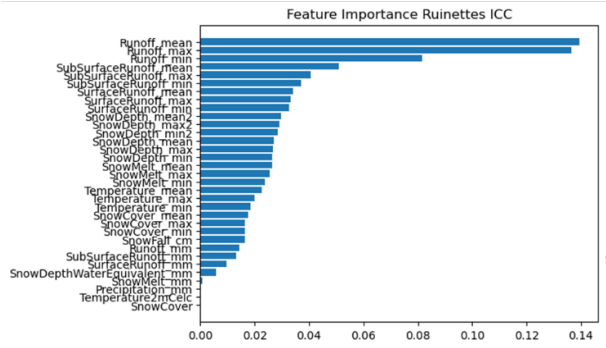


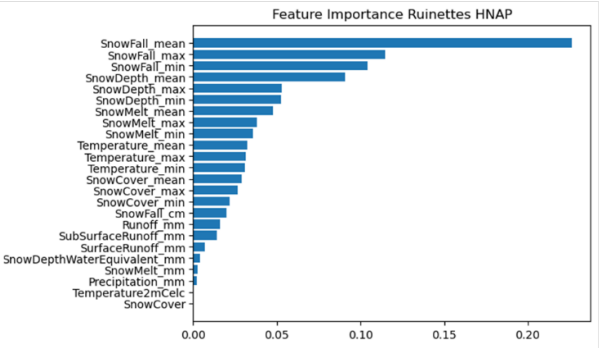
Figure 40: 3-class classification results for ICC and HNAP

Finally, regarding the Gini Importance, the Gini impurity measured the runoff was the key driver for the ICC prediction both for the two class and three class problem and snow was identified as the key driver for HNAP predictions. This shows that the contamination events in terms of ICC are driven by the transport of water though and over the ground, while the activity level of the bacteria is more closely linked to the snow dynamics than to simple water runoff (see Figure 41).

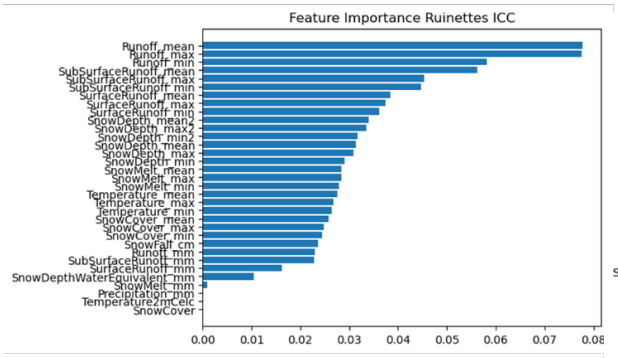
Ruinettes **ICC** Feature Importance



Ruinettes **HNAP** Feature Importance



Ruinettes **ICC** Feature Importance



Ruinettes **HNAP** Feature Importance

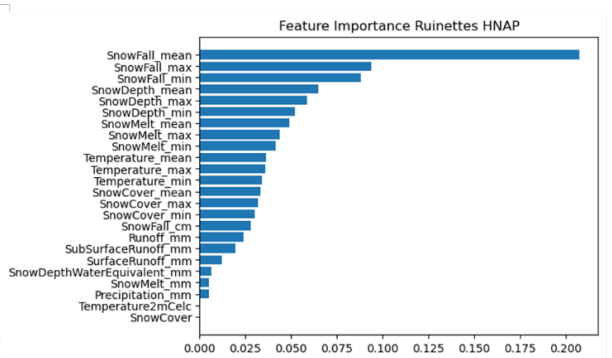


Figure 41: Explainability results for HNAP and ICC classifiers

5.2.6 Conclusions

From the beta version towards the final version the early warning system for bacteriological contamination was redesigned. In the current deliverable, the problem has been reframed as the direct prediction of bacteriological contamination from meteorological variables and was treated as a time-series prediction problem. The input data for this approach have been gathered and processed, with the final version relying on BactoSense in-situ measurements of ICC and HNAP, ERA-5 reanalysis meteorological variables, and ALTIS-provided in-situ meteorological measurements used to validate the ERA-5 data. A cross-correlation-based feature selection step was introduced, with a 30-day maximum lag and a 0.15 cross-correlation threshold, identifying Snow Cover, Temperature, Precipitation, Snow Melt, Snow Depth, and Runoff as the operational predictors.

In the first stage of the final version development, the problem was treated as a regression task, and three models were compared (ARIMA-X, LSTM and GRU). All three failed to predict the sudden contamination spikes, and this failure is reported as an explicit negative finding. In the second stage, the problem was reframed as a classification task, where the goal became predicting the level of bacteriological contamination (low, high, or optionally very high) rather than its exact continuous value. A Random Forest classifier was trained for both a 2-class and a 3-class setup, with thresholds defined from the median and 85th percentile of the target variables, and Gini impurity feature importance was used to explain the models' decisions.

In quantified terms, the beta version conceptually designed the EWS based on the SS1 workflow, without a trained model or reported performance metrics. The final version delivers two trained classifiers with the following results: the 2-class HNAP Random Forest classifier which achieves accuracy = 0.88 on the test set with balanced F1 scores across both classes and the 2-class ICC classifier which achieves accuracy = 0.712. For the 3-class setup, performance drops as expected (HNAP: 0.746, ICC: 0.644), reflecting the inherent difficulty of distinguishing high from very-high contamination events. The regression stage (ARIMA-X, LSTM, GRU) was tested and reported as negative, with all three models failing to predict contamination spikes.

In conclusion, the classification framework was far more effective and reliable than regression for predicting bacteriological contamination events in the Verbier system using only meteorological variables, and the 2-class HNAP model stands out as the strongest operational candidate for providing early warning of potential contamination events. The work also demonstrates that for remote sources in which continuous biological sensing is technically difficult, meteorological data combined with a classification-based risk framework can provide a cost-effective monitoring tool.

5.2.7 Replicability potential

The methodology described provided a state-of-the-art approach that identified the limitation of the regression methods for predicting bacteriological contamination using meteorological variables. The selection of the classification framework was proven to be more effective. In addition, the study is highly replicable as the data sources, sampling periods and models tested are clearly defined. The same is true regarding the processing logic and the threshold definition.

The replicability potential of this Early Warning System is high, as it leverages the globally available ERA-5 reanalysis dataset, eliminating the absolute dependency on dense, expensive local hardware networks. The proposed workflow, which validates reanalysis data against sparse in-situ checks and defines site-specific risk thresholds, acts as a transferable blueprint for other utilities managing karst or mountain spring sources. However, successful upscaling requires retraining the classification model with local

historical data to capture the specific lag times and hydro-geological responses unique to each catchment's snowmelt and runoff dynamics.

6 Soft sensors for Amsterdam demo case

6.1 Introduction and demo case description

Waternet supplies more than 90 million m^3 of drinking water annually to consumers in the Amsterdam area. The water is produced at two different DWTPs, Leiduin and Weesperkarspel.

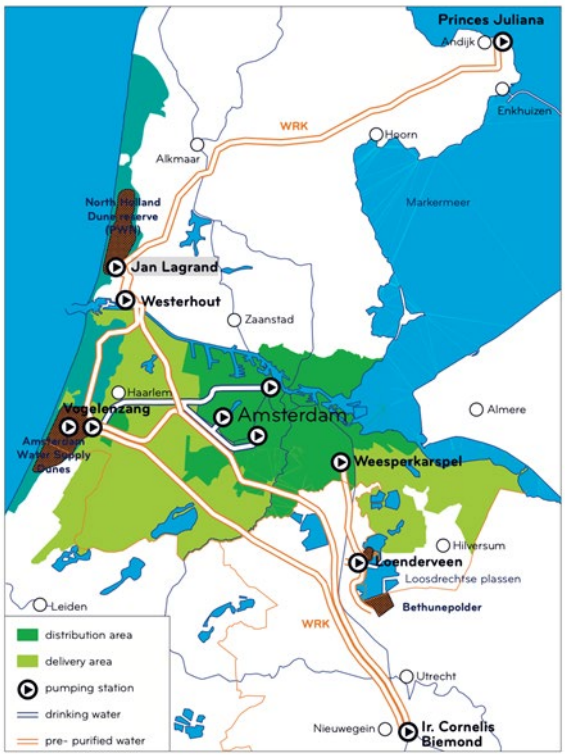


Figure 42: The Amsterdam demo case

Leiduin is the main DWTP as it produces approximately 70% of the water that feeds the Amsterdam area. The main source of the drinking water produced in the Leiduin plant is river water from the Lek Canal, supplemented by natural dune water. The treatment process consists of the pretreatment phase that takes place in Nieuwegein and the main treatment that takes place in Leiduin. The pretreatment consists of 2 stages, coagulation and rapid sand filtration and then the pre-treated water is transported to the Amsterdam Water supply Dunes (AWD) through 3 pipelines of 210 km length using 8 pumps. The main treatment process starts in the AWD with the infiltration of pretreated water. Thereafter, the following stages are rapid sand filtration, ozonation, that is used for both oxidation and disinfection, softening of the water, carbon filtration and slow sand filtration. The treated water is then stored in two different service reservoirs (storage tanks). Finally, the water is distributed in the Amsterdam area using pumps and 3 large pipelines. A schematic of this plant is presented in the following Figure 43.

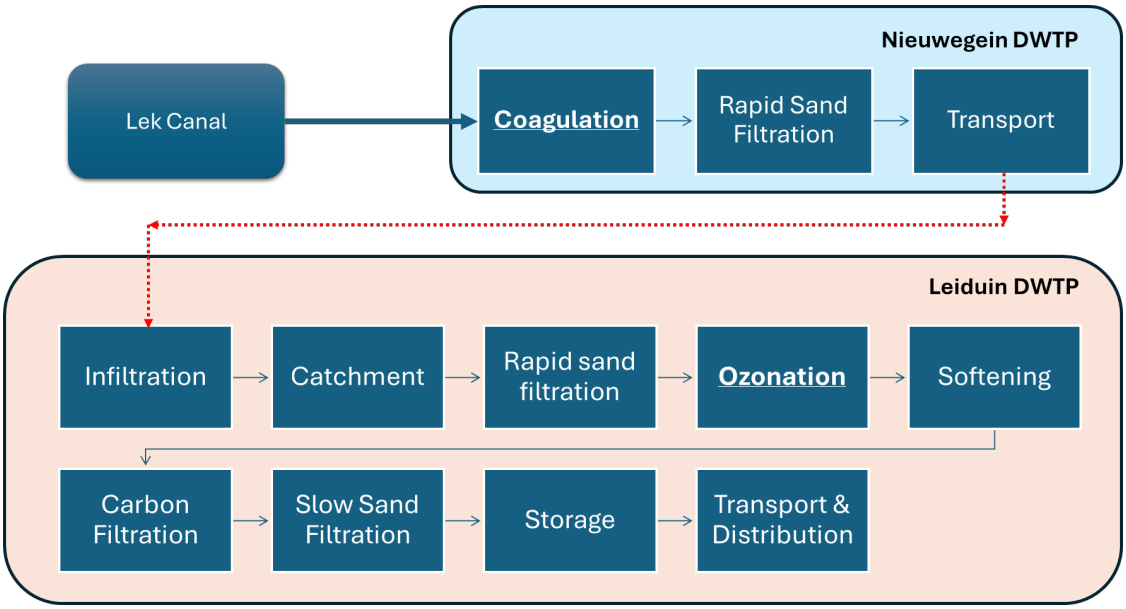


Figure 43: schematic of the Leiduin drinking water treatment plant

With this work, we aim to improve the treatment process of the Leiduin reservoir with the generation of 3 different soft sensors. More specifically the three soft sensors are as follows:

- **Soft Sensor 8 - Estimation of turbidity of the coagulation – flocculation process.**

  With this soft sensor the aim is to accurately predict the turbidity in the outlet of the coagulation – flocculation process 6 hours ahead, to inform Waternet's operational staff and aid them to adjust the coagulant dosage.

- **Soft Sensor 9 – Early prediction of turbidity in the DWTP inlet.**

  This soft sensor aims to predict water quality events that could influence the treatment process by predicting the turbidity in the inlet with some time in advance. Thus, with this tool, the operators will be informed with significant time in advance to prepare the DWTP and guarantee its continuous operation.

- **Soft Sensor 10 – Prediction of the ozonation exposure (CT) to improve the ozonation process.**

  This soft sensor aims to provide a daily estimation of the ozone exposure and consequently provide information about the efficiency of ozonation process. By providing daily ozone exposure estimates, Waternet's process engineers can make more precise adjustments to the ozone dosage, effectively responding to any fluctuation in bacteriological risks and optimizing the energy consumption of the ozone generators.

6.2 Soft Sensor 8 - Estimation of turbidity of the coagulation-flocculation process

6.2.1 Problem statement and soft-sensor development flow-chart

Waternet's treatment location in Nieuwegein, pre-treatment of the raw water from the Lek Canal is conducted. One of the treatment steps is the coagulation-flocculation treatment process, which relies on the dosing of the chemical ferric chloride to remove contaminants like organic carbon and phosphates. The control and operation of this treatment step such as the amount of ferric chloride dosed at a given time is based on measurement data and expert knowledge of the operators and process engineers. To support the decision making by these employees and minimize the use of ferric chloride, a soft sensor was developed that can help interpret the performance of the coagulation-flocculation process and give advice regarding its operation.

For this soft sensor, a model was built to predict the turbidity at the outlet of the coagulation-flocculation process, based on readily available measured input parameters. This model allows operators to have predictive insights on this important effluent parameter, thereby giving them a good indication of process performance in future scenarios. The turbidity at effluent is a parameter that operators use to steer the coagulation-flocculation treatment: it is a measure of the organic and particulate material in the water, which are removed by the coagulation-flocculation process. The ferric chloride reduces this turbidity but adding too much is both costly and unnecessary for reducing the turbidity below a threshold value set by national guidelines. As such, having a good prediction of the effluent turbidity will help enhance the operation of this treatment, reducing the need for ferric chloride, lowering both environmental impact and overall cost. The type of model that we develop is a hybrid physics + deep learning-based model, a state-of-the-art model type, that combines physicochemical knowledge of the treatment method with deep learning methods. We have found that this hybrid model provides better predictions than deep-learning or physics-based models alone.

This soft sensor was largely developed in the previous iteration of this work package and reported in Deliverable 4.1. In the subsequent iteration, improvements were made to the soft sensor to increase its accuracy. In addition, an uncertainty analysis was conducted and quantified to support the assessment of the predictions made by the different modules of the hybrid soft sensor.

6.2.2 Data sources and data preprocessing

The dataset used for training and testing the models, were provided by Waternet. The different types of sensor data that are used by the soft sensor are summarized in Table 18.

The first seven signals in Table 18 are model input or *feature* parameters, and the last one is the model output or *target* parameter. In this iteration, two additional features were added (days after dredging and the year) to the previous list and one was removed (flowrate). Turbidity, temperature and pH are measured by hard sensors at the inlet or outlet, and the amount of ferric chloride ($FeCl_3$) added for treatment, the time since last dredging event and the lanes used are actively managed. The time since last dredging event. The lanes used refers to which separate reactor paths are used: in Nieuwegein there are three reactor lanes, which can be closed or opened. Multiple lanes can be used at once. In the data as provided by Waternet, Lanes Used refers to which of the three lanes are opened for that specific data point.

The data preprocessing method was chosen to mimic data preprocessing as performed by Waternet internally, for two reasons: (i) their data scientists had been working for some time with this data, and as such had gained good insight; and (ii) one other interest was to investigate whether the novel model type that we develop in this project outperforms more traditional models. Therefore, the soft sensor data was pre-processed in the same way as measurement data was pre-processed by Waternet, to the extent that this was possible or meaningful. The preprocessing was done in order as follows:

Table 18: Data as used or predicted by the model.

Data	Data type	Range	Units	Notes	Feature or target
Turbidity (influent)	Floating point number	4-60	FTU	Turbidity at inlet.	Feature
Temperature (influent)	Floating point number	3-25	Degrees Celsius	Temperature at inlet.	Feature
pH (influent)	Floating point number	7.7-8.7	-	pH at inlet.	Feature
Ferric chloride ($FeCl_3$) added	Floating point number	0.2-6.5	mg Fe L^{-1}	Amount of ferric chloride added.	Feature
Days since dredging	Floating point number	0-956	days	Time since last dredging event.	Feature
Lanes used	Set of Booleans	[0-1, 0-1, 0-1]	-	Which of the three lanes are used at this time.	Feature
Year	Integer	2021-2024	-	Year in which measurement occurs	Feature
Turbidity (effluent)	Floating point number	1.5-15	FTU	Turbidity at outlet. Steering parameter for operators.	Target

- Outlier removal: when data was above or below a threshold, remove data points. Thresholds were kept the same as thresholds set by Waternet operators.
- Non-representative data removal: remove periods of data that are not representative of normal operating conditions, as determined by Waternet operators.
- Data scaling: all floating point feature parameters were normalized, *i.e.* centered around 0, magnitude of the data was set equal to their standard deviation.
- Data resampling: one data point measured per six hours (i.e., all data points at 0:00, 06:00, 12:00 and 18:00 hours) was included in the data set, other points discarded.
- Data was split into a training and testing set: three consecutive weeks' worth of data becomes part of the training set, the fourth week becomes part of the testing set.

The target parameter was coupled with the seven feature parameters as follows: as the water takes about 6 hours to pass through the reactor, the feature parameters at time t were coupled with the target parameters at time $t + 6h$. This was also in line with the preprocessing as performed by Waternet. It should be noted that the raw data also contained time stamp information, but this was not included in the model data set, as the intention was to build a time-independent model.

6.2.3 Materials and methods

In this work, a physics-informed neural network was developed: a hybrid deep-learning + physics-based model that incorporates physics or chemistry-based knowledge in its geometry. In this case, we develop a multi-branch type neural network model, where one branch represents physicochemical knowledge and the other branch represents a deep learning network.

This is a novel type of model that has only seen limited development in academic and practical contexts (Rai and Sahu, 2020; Seyyedi et al., 2023). They are an applied type of model to the more generic neural network models, where the known physics are incorporated in the model directly, and can take advantage of both the versatility and data-driven knowledge of machine learning models, while also drawing from the known broad physics-based knowledge. This combination is a recent step and allows for the versatile development of models for specific purposes (Seyyedi et al., 2023)

In earlier work by KWR, a physics-based model was developed, which was used as a basis for the physics part of the model. This model is explained in detail in a report by KWR,3 and we will summarize this model only briefly here. Coagulation-flocculation is a technique where ferric chloride is added to water, reacts and forms particles, which then find each other, aggregate together and grow into clusters or flocs. Contaminants then adsorb to these large flocs, and the weight of these flocs makes them sediment onto the bottom, allowing easy removal. The physics-based model is based on the DLVO theory,4 which describes the probability that two particles that find each other, will stick together. This sticking, or aggregation, is an important part of the coagulation-flocculation process. The DLVO-based model takes several parameters, such as the temperature, salt concentration and electric charge on the particles (given as the *zeta potential,5* a measurable parameter) and calculates this probability. The exact formulas and model details are beyond the scope of this report but can be found described in the earlier report[3].

This physics-based model was incorporated in a deep learning based neural network model. The goal of this overall model is to take the six feature parameters and predict one target parameter. This overall model was developed as graphically depicted in Figure 44.

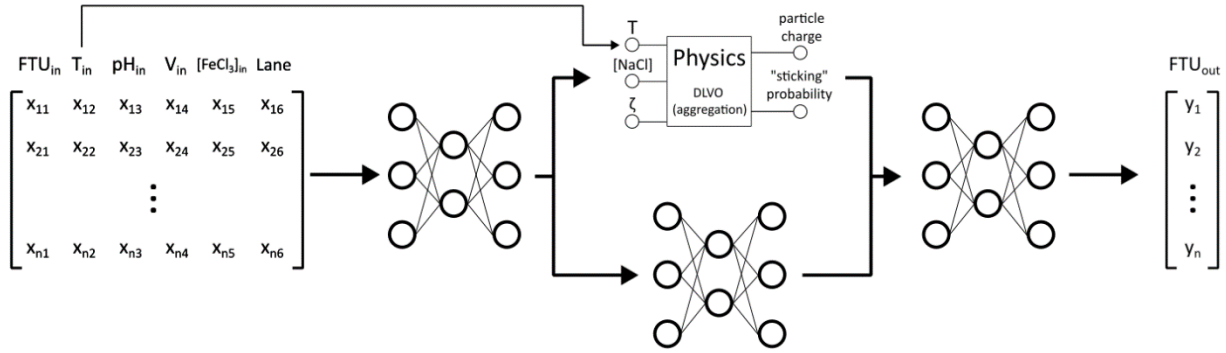


Figure 44: Geometry of the model.

The data enters into a neural network, which pipes its outputs into two branches. The first branch is the physics model, the second branch is another neural network. The outputs of both branches are piped into a final neural network, and its outputs are the predictions of the model.

3 J. N. Immink et al., *Softsensor flocculatie*, BTO report KWR. January 2023.
4 Derjaguin, B. V. *et al., The Derjaguin—Landau—Verwey—Overbeek (DLVO) Theory of Stability of Lyophobic Colloids*, Surface Forces, 1987.
5 Zasoski, R. J., *Zeta potential*, Encyclopedia of Soil Science, 2016

The input data (Figure 44, left) is first piped into a neural network subsection. This subsection performs its calculations and generates outputs; the output of this subsection is then piped into two model branches. The first branch is a DLVO-based physics model, the other is another neural network subsection. The physics model branch takes three input parameters. The first is temperature, which is taken directly from the input data. The other two are taken from the outputs of the first neural network and mapped onto the physical parameters that are necessary for this physics-based model. These parameters are then used to predict a particle charge and a "sticking" probability or aggregation probability.

In the other branch, a second neural network subsection is placed. Its inputs are not mapped as with the other branch, and its inputs and outputs are transformed directly as a normal neural network would. Finally, a last neural network subsection takes in the outputs from both branches, the two physics-based predictions and the neural network-based predictions. It generates a single output parameter, which is then mapped to be the turbidity at the outlet, or the target parameter.

This geometry was chosen to be able to translate the input parameters to output parameters as best as possible and take advantage of the strengths of both neural networks and physics-based models. The first neural network translates and prepares the input data to the input parameters of the physics model: while parameters such as iron chloride concentration and pH are related to the particle charge needed by the physics model, they are not directly translatable without more information. To generate input data for the physics model, the data should be transformed as best as possible. From a physicochemical point of view, this is not possible, but from an experience (or data) point of view, a reasonable guess could be made based on the knowledge we have at that point. This first neural network tries to achieve this as well as possible: guessing the physics model input parameters, given the information that the data contains. The third neural network has a similar purpose, in that it translates the particle charge and sticking probability to the turbidity, given the information that the model has from the physics-based and neural-network-based predictions at that point.

Finally, the middle neural network subsection is intended to not force all the information that is in the data through the physics model, as the physics model captures the aggregation phenomena but no other phenomena that also affect the turbidity. To make sure the model can also capture these types of phenomena effectively, the model pipes the information through a separate branch apart from the physics model that captures only part of the phenomena in the system.

6.2.4 Results

In order to achieve the best possible model, hyperparameters that govern specifics about the neural network subsections were explored to find the optimal values. These hyperparameters are for example the rate at which the model learns or the size and shape of the neural network subsections. However, it is beyond the scope of this document to discuss details about each individual hyperparameter, more detailed reading can be found elsewhere. The hyperparameter optimization was done using a random search algorithm: hyperparameters are randomly chosen within a range, and the performance of that set of hyperparameters was determined. Repeating this a number of times tells us which hyperparameter ranges yield to optimal models. Hyperparameters can be found in Table 18.

For the purpose of finding the optimal hyperparameters, 100 models were trained using 80% of the training data set, using a broad range for each parameter. After the optimal parameter ranges were identified, 100 more models were created with a second, narrower hyperparameter range, and the optimal value was chosen for each hyperparameter using this information. The results can be found in Table 18.

In the current iteration, with the changes made to the features used for model training, new models were trained and a search was conducted for the hyperparameters. This led to different optimised values than the ones that were reported in the previous deliverables. The results can be found in Table 19.

In order to investigate the model geometry and investigate the addition of the multi branch model, we test three different model geometries: a model where all information was piped through the physics model (physics-piped-model), where the bottom branch in the model (Figure 45) was turned off; a purely neural network-based model, where the physics-based branch (top branch) was turned off; and the complete model as seen in Figure 38 without alterations (multi-branch-model).

Table 19: Hyperparameters and their optimized values that were varied for the three types of models.

Hyperparameter	Optimized value (Physics-piped model)	Optimized value (Purely neural-network model)	Optimized value (Multi-branch model)
Epochs	70	15	50
Learning rate	$1 \cdot 10^{-3}$	$2.5 \cdot 10^{-3}$	$6 \cdot 10^{-4}$
Neurons per hidden layer, first neural network subsection	110	60	85
Number of hidden layers, first neural network subsection	110	10	10
Neurons per hidden layer, second neural network subsection	N/A	60	85
Number of hidden layers, second neural network subsection	N/A	10	10
Neurons per hidden layer, last neural network subsection	90	40	65
Number of hidden layers, last neural network subsection	2	2	3

The performances of the final models can be found in Table 20. We calculate the performance by the coefficient of determination, or R^2:

$$R^2 = 1 - \frac{\sum_i^N (y_i - \hat{y}_i)^2}{\sum_i^N (y_i - \bar{y})^2} \qquad \text{Eq. 25}$$

where $\hat{y}_i$ and y_i are the predicted and measured target value at data point i respectively, N is the number of data points considered for this calculation, and $\bar{y}$ is the average value of observed data targets calculated over N data points. Example predictions can be found in Figure 45. In Table 19, the improvements achieved in this iteration with the new model trainings conducted with refined input features and a new hyperparameter search are shown. The performances on unseen data (testing set) from the best model in iteration 2 are between 0.62-0.65. It is notable that the multi-branch model has the highest performance together with the neural network model, even though the difference are relatively small. This aligns with the expectation that, if a model has access to the physics-based knowledge and machine-learning based data interpretation, it can interpret the data best. At the same time, the outputs of the physics model can be extracted and interpreted for their physical meaning. While outside of the scope of the current project, further work could also investigate whether too low or too high predictions are more desirable and tuning a model to that finding.

Table 20: Performances of the final model. For reference: the R^2

	R^2 training set (iteration 1)	R^2 training set (iteration 2)	R^2 testing set (iteration 1)	R^2 testing set (iteration 2)
Physics-piped model	0.72	0.89	0.42	0.62
Purely neural-network model	0.77	0.86	0.44	0.64
Multi-branch model	0.74	0.92	0.47	0.65

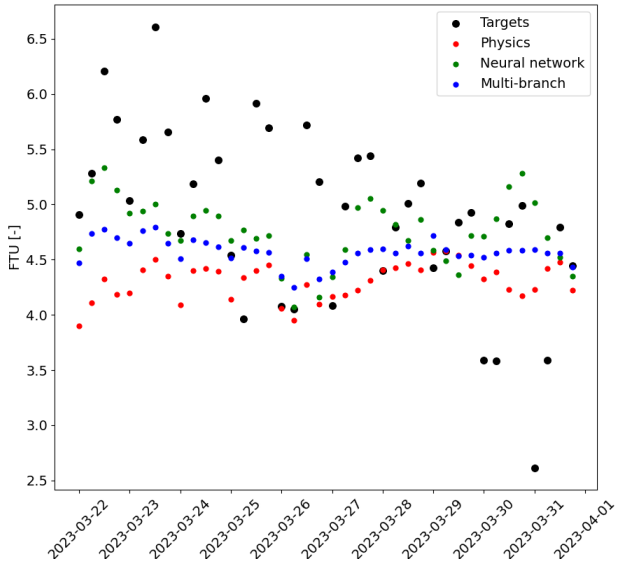


Figure 45: Example data points reflecting the original data targets (effluent turbidity) and the predictions by the three models.

6.2.5 Uncertainty analysis of the hybrid model

The three types of models were assessed for two types of uncertainties. First, epistemic uncertainty – the uncertainty arising from randomness in the model training. Every model will have slightly different parameters and therefore slightly different outputs. The uncertainty between different models trained on the same data and same hyperparameters is called epistemic uncertainty. Second, aleatoric uncertainty – uncertainty arising from uncertainty in the data which results in uncertainty in the output.

Uncertainty caused by model training (Epistemic uncertainty)

In 20, we show the uncertainty from the model for each of the three model types. The model predictions are calculated as described in the report. These uncertainties are calculated as follows. First, a training and a testing set is created from the Waternet dataset, using the same method and preprocessing as described in section 6.2 of the report. Then a set of hyperparameters is chosen for which the epistemic uncertainty is calculated. In this case we choose the best performing hyperparameter set, to get the uncertainty for the best performing model. A number of models is created through randomly initializing the weights and biases of the neural network modules (using the Glorot initialization routine. The model is then trained using the training data set and its performance is quantified using the testing set. The calculated performance metrics are the mean squared error (MSE), total squared error (TSE) and the coefficient of determination (R^2). These metrics are explained in section 6.2 of the report. For these performance metrics and for the prediction values over the full testing set, the standard deviations are calculated. For calculating the uncertainty of the best performing set of hyperparameters (see section 6.2 of the report), 20 models are created using that hyperparameter set.

Table 21: Performance parameters, average predictions and uncertainties for the best performing model of each model type.

Performance parameter	Multi-branch	Neural network-model	Physics-based model
Mean squared error (lowest)	0.989	1.048	1.094
Total squared error (lowest)	3824	4052	4225
Coefficient of determination (highest)	0.650	0.639	0.617
Prediction mean Uncertainties	5.177	5.119	5.243
Standard deviation of MSE	0.129	0.152	0.675
Standard deviation of TSE	499	588	2607
Standard deviation of R2	0.040	0.035	0.216
Standard deviation of prediction	0.397	0.411	0.657

It is clear that there are minor differences between the best performing models of the three different model types, with the multi-branch network performing slightly better than the other two. Whereas the model training has a significant effect on the exact model that is being trained, the standard deviation of the predictions with the testing set are about 8.0% of the magnitude of the target parameter for the neural-network and multi-branch based models. It is notable that the standard deviation for the physics-based model is much higher. This can be attributed to the model training: the physics layer relies on acceptable in- and output values which are physically sound. If values lie outside of physical reality, the loss function drops to drive training towards realistic numbers. It was seen that during training, the physics model can occasionally get to very high prediction error due to these non-physical numbers.

Uncertainty caused by variance in data (Aleatoric uncertainty)

We also investigate the uncertainty arising from uncertainty in the data. We take the best performing model which is also being used for further development. We test the aleatoric uncertainty in two ways:

- The variance when a random number is either added or subtracted from the original number in a particular input parameter, randomly over the entire data set. We add a random number to one parameter in the testing set, randomly generated around 0 with a standard deviation of 1. The number added is different for each same in the testing set. This is repeated 1000 times for the entire set and the mean, and standard deviation in predictions are found in Table 22.
- We also calculate the variance for the same parameters when they are varied with a certain percentage. This was done by either adding an absolute number to the parameter to be varied or by multiplication with a factor, again different numbers for each sample in the testing set. The factor was 1 with a standard deviation of 0.1. This is repeated 1000 times for the entire set and the mean, and standard deviation in predictions are found in Table 23.

Table 22: Standard deviations when a random number is added to the depicted input parameter. The random number is 0 with a standard deviation of 1. The number added is different for each sample in the data set.

Input parameter	Mean input magnitude	Standard deviation output (multi-branch model)	Standard deviation output (neural network)	Standard deviation output (physics-based)
Turbidity at inlet	16.01 (FTU)	1.494	1.606	1.672
Ferric chloride dosed	2.65 (mg/L)	2.187	2.464	2.646
Temperature at inlet	11.5 (oC)	1.508	1.615	1.684
pH at inlet	7.98 (-)	3.016	2.703	2.388
Time since last dredging	291 (days)	1.496	1.611	1.675

Table 23: Standard deviations when the data for a specific input parameter is multiplied by a random factor. The random factor is 1 with a standard deviation of 0.1. The factor is different for each sample in the data set.

Input parameter	Mean input magnitude	Standard deviation output (multi-branch model)	Standard deviation output (neural network)	Standard deviation output (physics-based)
Turbidity at inlet	16.01 (FTU)	1.493	1.606	1.672
Ferric chloride dosed	2.65 (mg/L)	1.525	1.673	1.719
Temperature at inlet	11.5 (oC)	1.503	1.614	1.713
pH at inlet	7.98 (-)	2.668	2.572	2.488
Time since last dredging	291 (days)	1.509	1.620	1.752

Some parameters have more effect on prediction than others do, with ferric chloride dosed and pH at inlet being strongly affecting parameters. Other parameters do not affect uncertainty much, and it seems like here an inherent uncertainty is found, which a single-parameter uncertainty has to overcome in order to affect overall uncertainty.

Another interesting point is that the inherent uncertainty of the multi-branch model is lower than the uncertainties of the other two models. This might be due to the multi-path nature of the model, allowing for several options for the data to pass.

6.2.6 Conclusion and next steps

In this work we report on the development of a novel physics-informed neural network, developed for the operation of the coagulation-flocculation water treatment technique as operated by Waternet. The model is intended to help operators make informed decisions regarding chemical dosing, optimizing environmental impact and minimizing costs. We have developed a hybrid physics-based + deep-learning based model, which takes advantage of the physicochemical knowledge of this water treatment technique, but also the data-based modelling that deep learning models can provide. We show that the model can predict turbidity at the outlet of the water treatment relatively well and achieve better performances than traditional machine learning models and achieve predictions which a purely physicochemical model would not be able to make.

In the next steps of this project, we intend to optimize the models further with several steps. First, we will feed a new version of the model with a broader range of data, which will include a broader window of data points, reflecting a different type of operation and experiences with the operation. This will make the model more robust for unseen circumstances. Furthermore, a broader hyperparameter optimization regime will improve the model further. Finally, we intend to investigate the degree to which each branch in the model contributes: by diminishing or "turning off" the contribution of each of the two branches in the model, we can investigate to what extent the physics model and the deep-learning models contribute to the final predictions and help tune the model to achieve even better performances.

6.3 Soft Sensor 9 - Early prediction of turbidity in DWTP inlet

6.3.1 Problem statement and soft-sensor development flow-chart

DWTPs constitute critical components within drinking water systems, ensuring that water provided to consumers is devoid of harmful microorganisms and hazardous substances. In Amsterdam, more than 1,4 million customers use tap water from the Waternet, the water-cycle company of Amsterdam and surrounding areas, with the average daily water consumption of 141 liters per person in 2021. Approximately two thirds of Amsterdam's tap water originate from Lek canal, a canal situated east of Nieuwegein, where the water is pumped into large ponds and pre-purified there with coagulation and sedimentation process.

Coagulation is a fundamental step in ensuring the quality and safety of drinking water, as it effectively removes pathogens, and improves the overall clarity of the water. Optimizing dosing and initial water quality conditions are crucial for guaranteeing efficient coagulation with many factors influencing the process such as the presence of total suspended solids (TSS), temperature of the water and pH [49]. Direct measurement of water-quality indicator parameters, such as coliform and *e-coli* bacteria or TSS, can be challenging, often requiring laboratory experiments with delays ranging from days to weeks. Therefore, water utilities use turbidity levels, a physical property of the fluids a parameter that is easy to measure with hard sensors, as a water quality indicator that reflects the presence of TSS in the raw water [50,51]. High turbidity in the raw water imposes higher risk in the quality of the drinking water, therefore, the aim of the coagulation process to reduce turbidity levels in raw water and consequently fully treat the water in the following processes [51].

Turbidity is highly fluctuated because of continuous changes in the river water flow, the sudden rainfall events and the human-made activities [50]. These conditions pose significant challenges to the task of achieving and maintaining optimal conditions for the coagulation process [50,51]. Traditional methods for identifying the optimal coagulant dosage, like jar testing, are not only costly and time-consuming but also yield delayed results, often failing to response to the rapid fluctuations in source water quality [51]. This delay impedes the timely implementation of corrective measures essential for ensuring the continued efficacy of treatment processes. Consequently, timely implementation of corrective measures for water quality events becomes a considerable challenge. One possible approach to address this challenge is the development of source water quality prediction methodologies. Such methodologies would provide sufficient response time for the DWTP operators to anticipate and prepare for changes of the water quality conditions.

Data-driven approaches, particularly machine learning (ML), have become essential tools in engineering due to their capacity to address complex nonlinear challenges. ML models such as artificial neural

networks (ANNs) or decision trees have been extensively employed in water quality prediction [52]. Numerous studies have estimated various water quality parameters across the supply chain, from source to tap. For example, Ortiz Lopez et al. [53] used rainfall and discharge data as inputs for two ML algorithms—support vector machines and ANNs—while Lu and Ma [54] applied a hybrid approach combining random forest with XGBoost to predict turbidity at the inlet of DWTPs.

Building upon these developments, this work aims to further contribute to the field of data-driven applications for improving DWTP efficiency by developing a ML algorithm-based model for the for short-term turbidity prediction in the Nieuwegein DWTP. By providing early turbidity events forecasting, this work aims to support Waternet's process engineers in enhancing the coagulation process management strategies. The developed methodology in firstly identifies the key factors that influence turbidity in the Lekanaal, then identifies the combination of variables (features) that yield the most effective turbidity predictions and, finally, investigates how far in advance can turbidity levels be reliably predicted.

6.3.2 Data sources and data preprocessing

Data Collection

The Nieuwegein DWTP takes water from the Lekkanaal. Lekkanaal is a canal that connects the River Lek to the Amsterdam-Rhine Canal at Nieuwegein. Located within the Rhine–Meuse–Scheldt delta, this area has a complex water network interconnected by rivers and canals. However, the flow direction in Lekkanaal can be bidirectional. The source of the water can be traced back to the River Rhine, which enters the Netherlands at Lobith. There are multiple stations located across the river that measure the river discharge flow and some water quality parameters with a certain frequency – 10 minutes frequency sensor measurements for the river discharge, daily sensor measurements for conductivity and weekly sampling for water quality parameters. The main station that monitors the water that enters in the Netherlands is located in Lobith. Other stations that are located upstream of the DWTP are the Hagestein boven river station and the Nieuwegein station in the Lekkanaal. These stations are managed by the Rijkswaterstraat (Public Works and Water Management) and RIWA (Association of River Water companies) and their data are publicly available. In the DWTP intake, Waternet monitors 4 different parameters through their SCADA systems, water flow, water temperature, pH of the raw water and turbidity of the raw water with a 5-minute frequency. For this part of this work, we concentrated only in the sensor data with high frequency. Overall, the different data variables, the data availability period, the location that the variables are measured and the data source are presented. Finally, to further explore

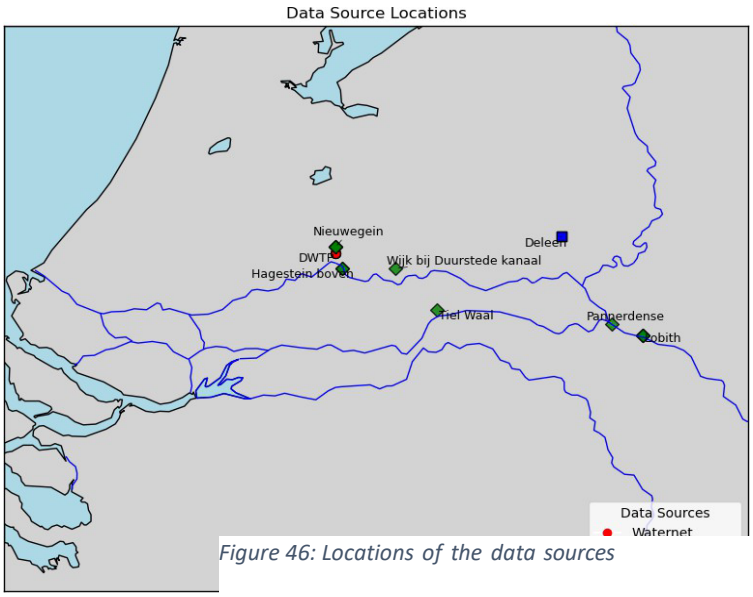


Figure 46: Locations of the data sources

the influence of the weather on the canal turbidity, rainfall, pressure and temperature data collected from Deleen weather station, the closest weather station to the DWTP. The data from this source belong to The Royal Netherlands Meteorological Institute (KNMI) and is publicly available. The locations of the data sources are presented in the following Figure 46 and the overview of the available data in the following Table 24.

Table 24: Datasets used for the soft sensor development

Data source	Parameter	Name	Location	Lon	Lat	Temporal resolution
Waternet	Turbidity	sensor_turbidity	Nieuwegein	5.113	52.022	5M
Waternet	Discharge	sensor_dis	Nieuwegein	5.113	52.022	5M
Waternet	Water temperature	sensor_temperature	Nieuwegein	5.113	52.022	5M
Waternet	pH	sensor_pH	Nieuwegein	5.113	52.022	5M
Rijkswaterstaat	Discharge	Lob_dis	Lobith	6.145	51.846	10M
Rijkswaterstaat	Discharge	Hag_dis	Hagestein boven	5.136	51.990	10M
Rijkswaterstaat	Discharge	Nieu_dis	Nieuwegein	5.113	52.034	10M
Rijkswaterstaat	Discharge	Pan_dis	Pannerdense	6.043	51.870	10M
Rijkswaterstaat	Discharge	Tiel_dis	Tiel Waal	5.456	51.901	10M
Rijkswaterstaat	Water level	Nieu_wl	Nieuwegein	5.113	52.034	10M
Rijkswaterstaat	Water level	Wijk_wl	Wijk bij Duurstede kanaal	5.316	51.989	10M
Rijkswaterstaat	Water temperature	Lob_wt	Lobith	6.145	51.846	10M
Rijkswaterstaat	Water temperature	Hag_wt	Hagestein boven	5.136	51.990	10M
Rijkswaterstaat	Water temperature	Nieu_wt	Nieuwegein	5.113	52.034	10M
Rijkswaterstaat	EC	Nieu_EC	Nieuwegein	5.113	52.034	10M
Rijkswaterstaat	EC	Lob_EC	Lobith	6.145	51.846	1H
Rijkswaterstaat	pH	Lob_pH	Lobith	6.145	51.846	1H
Rijkswaterstaat	Turbidity	Lob_turbidity	Lobith	6.145	51.846	1H
KNMI	Pressure	P	Deleen	5.873	52.056	1H
KNMI	Rainfall	RH	Deleen	5.873	52.056	1H
KNMI	Temperature	T	Deleen	5.873	52.056	1H

Sensor Data Preprocessing

Due to sensor sensitivity and potential fouling from material buildup, as well as periodic recalibration requirements, the raw sensor data often contained inaccuracies. To ensure reliable data quality, we implemented a five-step preprocessing approach as proposed by Gleason et al [54]:

- **Timestamp Errors and Missing Data Replacement**: Timestamp errors and missing data points were identified and replaced using data interpolation, as there were no extended periods of

missing data or large timestamp discrepancies. Interpolation was chosen because the errors and gaps were minimal and non-consecutive, making it suitable for filling short gaps in a time series.

- **Single-Point Outliers Replacement**: Outliers were identified as values significantly deviating from surrounding data points. To detect these, we calculated the z-score (i.e., the difference between the point and the dataset's mean, divided by the standard deviation). Any data point with a z-score exceeding a threshold of 100 was flagged as an outlier and replaced via interpolation. In this step a large number of turbidity data were identified which is probably related to the sensitivity of the turbidity sensors in material accumulated in their optical lens.
- **Threshold-Based Replacement**: Waternet has established minimum and maximum acceptable thresholds for several parameters (e.g., temperature, flow, pH). For instance, acceptable water temperatures range between 3°C and 27°C. Data points outside these thresholds were deemed invalid and were replaced with interpolated values to maintain continuity.
- **Flat-lining data**: Flat-lining data occurs when sensors are returning the same value repeatedly. To tackle this issue, the total period that the sensor repeats the same value is calculated. In this work 3 thresholds were set, an 8-hour threshold for the flow and discharge data, a 4 hour threshold for the turbidity, and a threshold of 12 hours threshold for pH and temperature. Overall, 8 periods of flat-lining temperature data (5 days of data), 270 periods of flat-lining turbidity data (7 days) and 3 days of pH data were identified. The data were replaced with the weekly medians when it was possible but in the periods where the flat-lining was more than a day, these data were removed from the final dataset.
- **Drift Correction**: Some sensors, particularly turbidity sensors, are prone to drift over time. To correct this, we calculated a four-week rolling mean. Successive changes in this weekly average were flagged as drift, and any data showing sustained shifts were corrected using asymmetric least squares regression.

After completing the data cleaning, the pre-processed data were aggregated from 5-minute and 10-minute frequency to hourly frequency and merged into a unique dataset using the date (day and hour of measurement.

6.3.3 Materials and methods

Runoff Time Analysis

When applying ML to predict turbidity, understanding water runoff time from upstream monitoring locations to the DWTP is crucial. This travel time, referred to as the time lag, determines how upstream measurements—such as turbidity and other water quality parameters—correspond to conditions downstream. If the time series are misaligned due to an unknown lag, the ML model may suffer from reduced accuracy by relying on data that does not properly reflect the timing of events at the DWTP.

To estimate this time lag, the peak-matching method was used for discharge data when distinct peaks were present. This method compares discharge time series from upstream and downstream monitoring stations, identifying corresponding peaks—significant increases in discharge—in both datasets. The time difference between these peaks provides an estimate of runoff time.

However, the Lekkanaal and Amsterdam-Rhine Canal exhibit different flow characteristics due to hydraulic regulation. Water in these canals can flow in both directions, and discharge peaks are often unclear. In such cases, correlation analysis with time lag was applied instead. Specifically, Spearman correlation was calculated between turbidity and other water quality time series with time lags ranging from 0 to 15 days in 1-hour increments. Spearman correlation is preferred over the commonly used

Pearson correlation because it is rank-based and better suited for detecting non-linear relationships. The Spearman correlation coefficient (ρ) is calculated as:

$$\rho = 1 - \frac{6\sum_i^n d_i^2}{n(n^2-1)}$$

where $d_i = r(X_i) - r(Y_i)$ is the difference between the ranks of each pair of observations, n is the number of data points, and $r(X_i)$ and $r(Y_i)$ are the ranks of X_i and Y_i in their respective datasets.

The estimated travel time corresponds to the time lag at which the correlation reaches its peak. Once identified, the upstream parameter time series were shifted by this optimal lag before being fed into the ML model. This alignment step is crucial to ensure that the model learns from temporally relevant input features, thereby improving its ability to capture causal relationships and enhancing overall predictive performance.

Correlation identification

The correlation identification included the use of self-organising maps (SOMs) [56], and correlation analysis using the Spearman cross correlations with a time lag. Self-organising Maps is an unsupervised neural network clustering methodology that visualises multidimensional datasets in 2-D plots where each plot represents a different variable. The visualisation plots can effectively reveal and communicate hidden non-linear complex correlations between multiple variables, even when part of the dataset is missing or incomplete. This visual and qualitative correlation is achieved as SOM locates the correlated clusters of all the variables in the same area of their own map. Thus, a visualization of the correlations is achievable.

The methodology followed in this work is as follows:

- Spearman correlation obtained from runoff time analysis is used to evaluate the relationship between turbidity and the other available variables.
- Every variable with Spearman corelation higher than 0.3 or lower than -0.3 was considered high enough to be selected as a potential candidate
- The selected variables were then further analysed using SOMs to explore further multi-correlations

Turbidity predictive model using data-driven algorithms

In this work 3 data driven models were implemented, **ARIMA** (AutoRegressive Integrated Moving Average) a classical statistical model for univariate time series forecasting [57] ; **Random Forest** (RF), a non-linear, multivariate model based on decision tree ensembles [58] ; and **Long Short-Term Memory** (LSTM), a deep learning model designed for sequential data [59]. These models were trained and evaluated using consistent data splits and forecast horizons (1 hour, 3 hours, and 6 hours ahead). While ARIMA provides a baseline from traditional statistical modeling, RF introduces a non-linear multivariate benchmark, and LSTM represents a more advanced approach capable of capturing complex temporal dependencies and feature interactions in dynamic water quality conditions.

ARIMA models

ARIMA models the future values of a variable as a linear combination of its own past values (autoregressive part), past forecast errors (moving average part), and differencing to achieve stationarity (integrated part). The model is specified as ARIMA(p, d, q), where:

- p is the number of autoregressive terms
- d is the number of times the data is differenced to make it stationary
- q is the number of moving average terms

The ARIMA model is applied to the univariate turbidity time series. The entire dataset is split chronologically, with the first 85% used for model training and the remaining 15% reserved for testing. The training process involves the following steps:

Stationarity Check: The augmented Dickey-Fuller (ADF) test is used to assess whether the series is stationary. If necessary, differencing is applied to remove trends.

Hyperparameter Selection: The optimal values for p, d, and q are chosen using a grid search guided by the Bayesian Information Criterion that, compared to the Akaike Information Criterion, penalizes model complexity more strongly. This ensures a more parsimonious model, especially important for large datasets like this work.

Model Fitting and Forecasting: The final ARIMA model is fit to the training data. To evaluate the ARIMA model at different forecast horizons (e.g., 1-step, 3-step, 6-step ahead), a fast rolling forecast method is used. For each time step t in the test set, the trained model generates a multi-step forecast using get_forecast(steps=h), where h is the desired forecast horizon. The last forecast in the multi-step forecast sequence is extracted for com- parison with the actual observed value. The model is then updated incrementally using model_fitted.append([new_observation], refit=False) to incorporate the next true observa-tion, avoiding costly retraining. This process is repeated iteratively, simulating how the model would perform in a real-time deployment setting.

Random Forrest

Random Forest is an ensemble learning method based on decision trees. By constructing multiple trees and aggregating their predictions, RF enhances prediction accuracy and robustness. It effectively captures nonlinear relationships and can handle high-dimensional feature spaces.

Before training, all continuous variables were normalized to the range [0, 1] using MinMaxScaler for consistency across modeling pipelines. The scaler was fitted on the training data only, and the same transformation was applied to the test set using stored parameters to avoid in- formation leakage. Additionally, the target variable, sensor_turbidity, was log-transformed using np.log1p to assess its impact on RF model performance, potentially reducing skew- ness, stabilizing variance, and enhancing the model's ability to capture non-linear patterns.

Like ARIMA model training, the Random Forest model is trained on the first 85% of the data, while the remaining 15% is reserved for testing. K-fold cross-validation is used to assess model performance by splitting the training data into multiple subsets, ensuring reliable hyperparameter tuning and reducing overfitting.

The hyperparameter search space includes:

- Target variable: sensor_turbidity or sensor_turbidity_log
- Number of trees (n_estimators): 100 to 1000
- Maximum tree depth (max_depth): 1 to 10
- Minimum number of samples required to split a node (min_samples_split): 2 to 20
- Minimum number of samples required at a leaf node (min_samples_leaf): 1 to 20
- Number of features considered for splitting (max_features): from $\sqrt{availalbe\ features}$ to $\frac{\text{available } features}{3}$

LSTM

In this work, the Long Short-Term memory (LSTM) model employed a sequence-to-one structure. Stacked LSTM layers process the input sequence, and the final hidden state is passed to a fully connected layer to produce a single turbidity forecast.

The same preprocessing steps described for the RF model were applied here, including feature scaling with MinMaxScaler, log-transformation of the target variable sensor_turbidity, and TimeSeriesSplit with 5 folds for k-fold cross-validation. These steps ensured fair comparison across models and preserved the temporal characteristics of the data.

Unlike Random Forest, which use flat feature inputs, LSTM requires structured sequences of data. A sliding window approach was applied, where each input sample consists of 24 consecutive hourly records of all selected features. The corresponding output label is the turbidity value at the desired future time step (e.g., 1-hour, 3-hour, 6-hour ahead). Models were trained using the Adam optimizer and MSE loss. Batch size was set to 16, and training was conducted for up to 100 epochs with early stopping based on validation loss.

Hyperparameter tuning was conducted using the Optuna framework, applied separately for each forecast horizon. The optimization process targeted the best validation performance across cross-validation folds, with mean MSE as the objective metric to minimize. The hyperparameter search space includes:

- Target variable: sensor_turbidity or sensor_turbidity_log
- Number of LSTM layers: 1 to 2
- Hidden size: 16 to 256
- Learning rate: 10^{-5} to 10^{-1}

6.3.4 Evaluation metrics

Six metrics were used to assess predictive accuracy: Nash–Sutcliffe Efficiency (NSE), NSE on the log-transformed target (log NSE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Relative Absolute Error (RAE).

- NSE: Measures the model's predictive skill compared to the mean of observed data, with values closer to 1 indicating better performance.
- Log NSE: Applies NSE to the log-transformed target variable, emphasizing performance on low-value events; higher values indicate better fit.
- MSE: Calculates the average squared difference between predicted and actual values, penalizing larger errors more heavily; lower values indicate better accuracy.
- RMSE: The square root of MSE, providing error magnitude in the same units as the target; lower values reflect higher accuracy.
- MAE: Computes the average absolute difference between predicted and actual values; lower values indicate better performance.
- RAE: Computes the mean of absolute errors divided by their corresponding absolute observed values, expressed as a percentage; lower values indicate better accuracy.

6.3.5 Feature attribution analysis

To enhance model performance and interpretability, this study applied both greedy feature selection and post hoc interpretation techniques tailored to the respective model types.

To evaluate the predictive contribution of correlated input variables, a greedy feature se- lection approach was applied using sensor_turbidity as the baseline. Candidate features identified through prior correlation analysis were added one at a time, forming new input combinations. For each combination, a ML model was trained and validated to determine whether the added feature improved predictive performance. This stepwise process allowed for a systematic evaluation of the marginal utility of each feature in enhancing turbidity forecast accuracy.

Following the identification of optimal feature sets, interpretability tools were applied to quantify and visualize the influence of each input feature. For the Random Forest model, two interpretability methods were used:

- Built-in feature importance scores were extracted, which reflect the average contribution of each variable in reducing prediction error across the ensemble.
- SHAP (SHapley Additive exPlanations) was used to provide more nuanced local and global attributions. Summary and beeswarm plots were generated using TreeExplainer to visualize how each feature influenced predictions, including the direction and magnitude of its effect.

Interpreting the LSTM model required gradient-based techniques. The Captum library was used to apply Integrated Gradients, which attributes importance scores by integrating gradients along a baseline-to-input path.

6.3.6 Results

Data Quality assessment

Figure 47 shows overall quality of the original data collected from Waternet, Rijkswaterstaat, and KNMI.

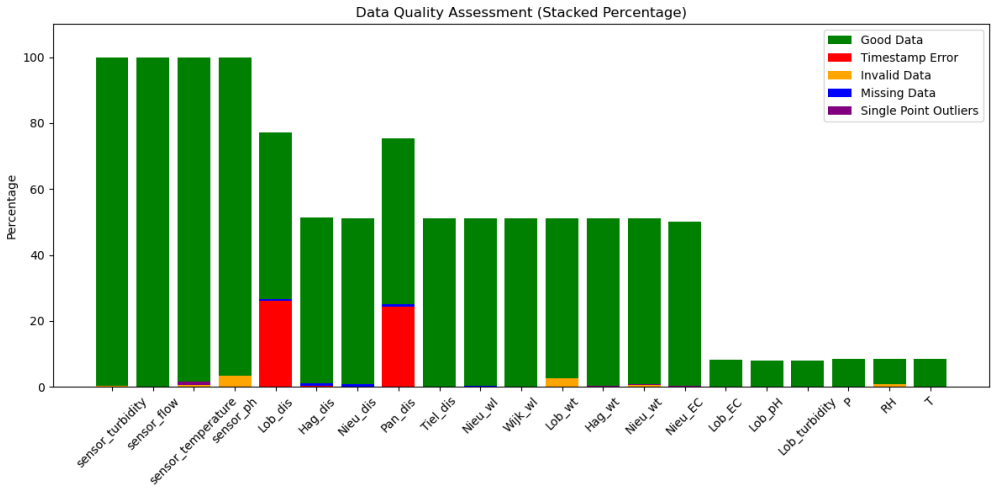


Figure 47: Data quality overview

The aforementioned in the methods section part methodology was applied in all the sensors where invalid or missing data were identified. An example of the data preprocessing application for the identification of the single point outliers in Lobith's temperature sensor is presented in the following Figure 48.

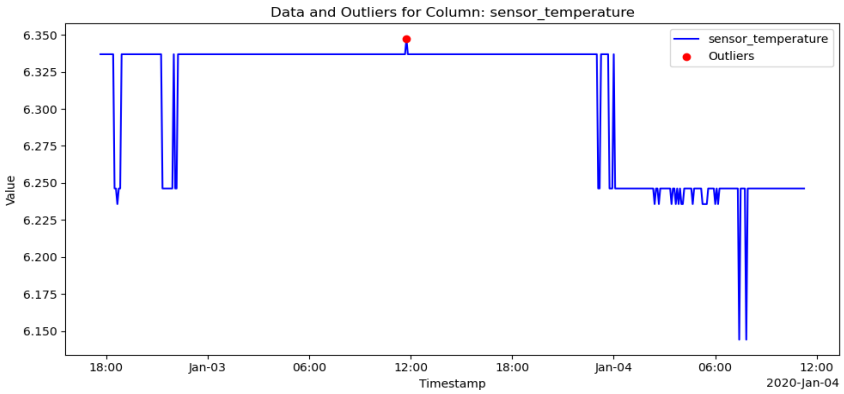


Figure 48: Single point outliners of Lob_EC

The discrete statistics of the processed data is presented in the following Table 25.

Table 25: Data overview

Parameter	Unit	Mean	Std	Min	Median	Max
sensor_turb	FTU	16.06	7.73	3.65	14.66	167.33
sensor_flow	m^3/h	12874.56	19144.44	0.00	12923.84	174764.31
sensor_ten	°C	3.35	6.41	0.00	2.31	38.43
sensor_ph		4.97	0.34	3.75	5.02	10.58
Lob_dis	m^3/s	2124.90	1210.32	704.95	1692.51	7464.83
Hag_dis	m^3/s	219.46	400.56	0.00	29.52	1380.26
Nieu_dis	m^3/s	0.00	0.00	-52.72	0.00	0.00
Pan_dis	m^3/s	1526.30	758.43	737.31	1332.03	4968.93
Tiel_dis	m^3/s	1494.37	758.42	541.57	1259.33	5083.54
Nieu_wl	cm	-39.33	13.30	-174.56	-39.22	66.49
Wijk_wl	cm	-39.38	4.77	-120.76	-39.32	15.92
Lob_wt	°C	9.70	6.00	-1.61	9.22	26.23
Hag_wt	°C	10.03	6.19	-2.56	9.33	27.67
Nieu_wt	°C	12.13	7.71	-2.86	12.00	26.17
Nieu_EC	S/m	4.60E+07	9.32E+06	2.10E+07	4.37E+07	5.69E+07
Lob_EC	S/m	5.50E+07	7.72E+06	0.00	5.57E+07	5.57E+07
Lob_pH		7.98	0.17	7.26	7.94	8.94
Lob_turbidity	FTU	21.47	17.25	2.25	16.42	514.50
P	Pa	10154.60	7.00	10103.96	10151.04	10204.77
RH	%	91.47	7.05	47.86	92.93	235.00
T	°C	10.12	7.72	-13.10	9.96	82.67
RH_daily_a	mm	31.52	63.60	0.00	3.00	1141.50

Runoff Time Analysis

Runoff time analysis was conducted to determine the time lag between collected time series with the target variable sensor_turbidity, which will ensure temporally relevant features are used for ML model training.

First, the peak match method was applied to identify the runoff time between discharge stations. The following Figure 49 shows the peaks of Lob_dis and Hag_dis, where 13 common peaks were identified. The average travel time between these peaks was calculated to be around 28 hours.

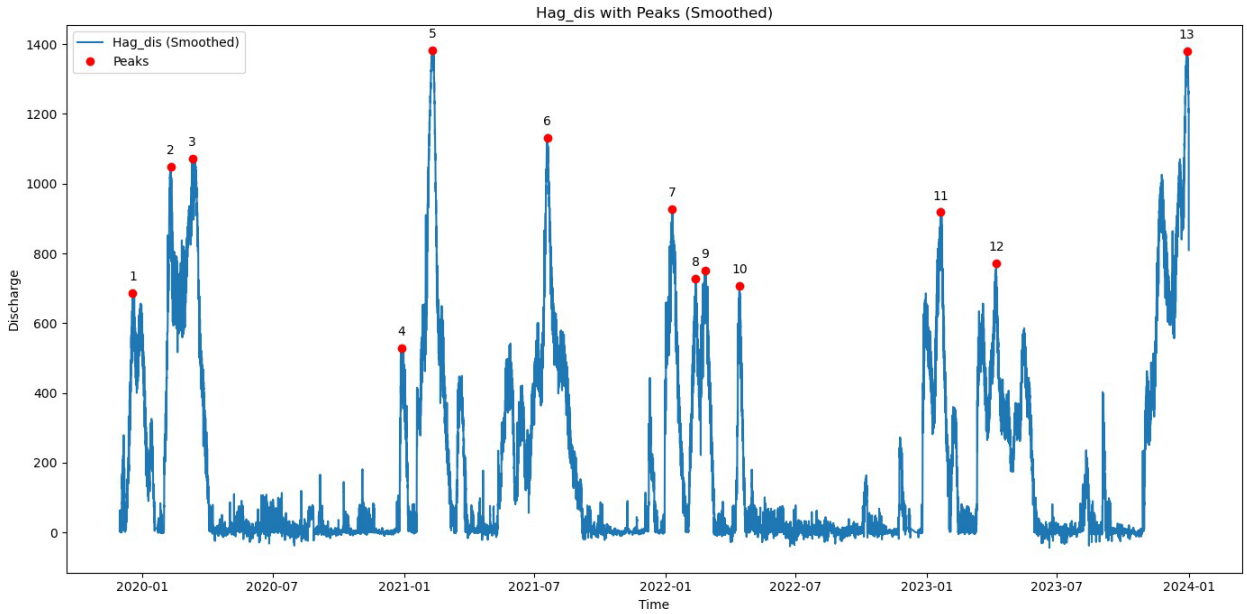


Figure 49: Peaks of Lob_dis and Hag_dis

Overall, the calculated travel distance between the discharge stations, starting from Lobith and ending to Hagestein is presented in the following Table 26.

Table 26: Average traveling time from peak match method

Time series	Travel time
Lob_dis	0
Pan_dis	1:19:10
Tiel_dis	16:45:00
Hag_dis	28:13:38

The following Figure 50 presents discharge data from Nieuwegein, located at the downstream end of the Lekkanaal. Unlike river discharge time series, the canal flow exhibits bidirectional behaviour and lacks distinct peak patterns. As a result, the peak-matching method, which depends on clear maxima to estimate time lags, was not applicable. Instead, a correlation-based approach was adopted. Spearman correlation was calculated between each input time series and the target variable sensor_turbidity over time lags ranging from 0 to 15 days.

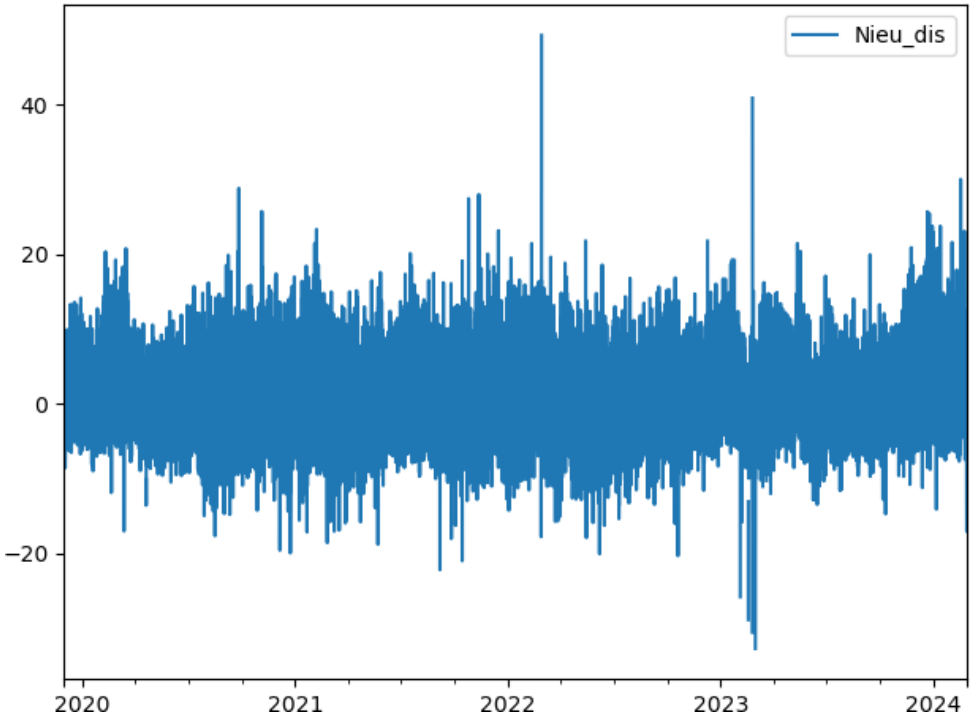


Figure 50: Nieu_dis

In the following figure 51 the result of the Spearman correlation between sensor_turbidity and discharge time series are presented. Distinct correlation peaks can be observed from discharge measurements along the river. The time lag difference between the correlation peaks of Lob_dis and Pan_dis is 2 hours, which aligns with the travel time obtained from the peak match method, as well as for Tiel_dis (17 hours).

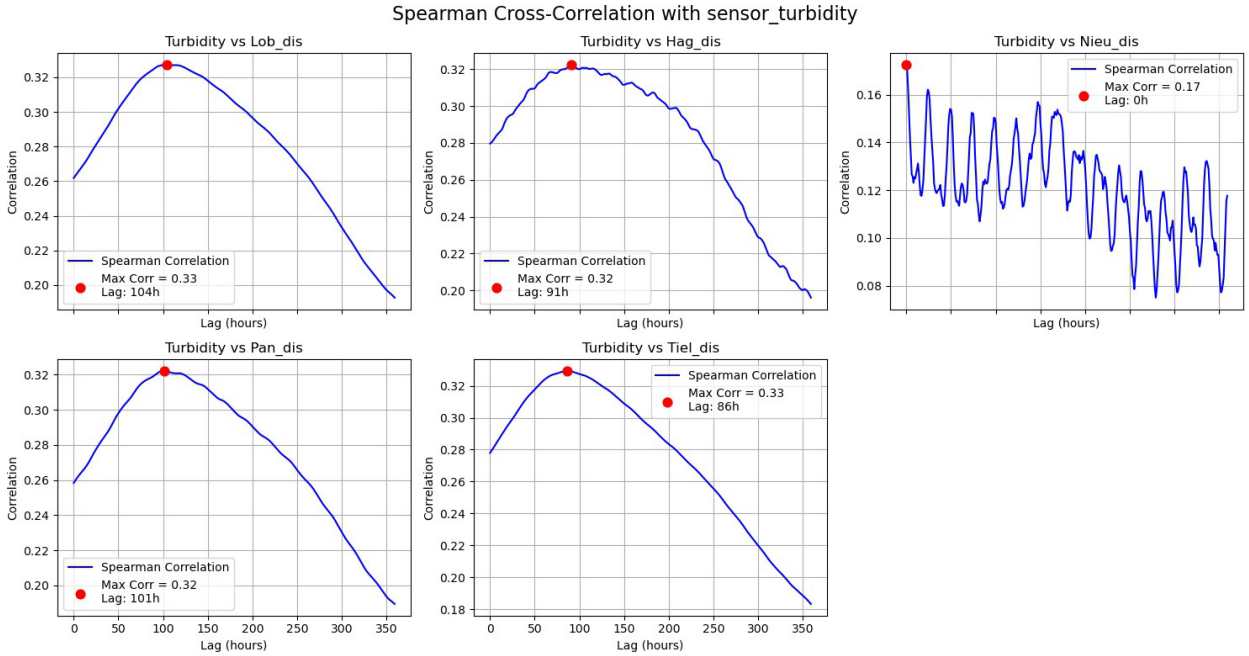


Figure 51: Spearman Cross_Correlation between sensor_turbidity and discharge

The time lag between the correlation peaks of Lob_dis and Hag_dis is 13 hours—shorter than the travel time estimated using the peak-matching method. This discrepancy arises because peak matching focuses on flood periods, when weirs and floodplains increase the runoff time. In contrast, correlation analysis accounts for all flow conditions, capturing the generally faster and smoother water movement under normal, unobstructed conditions.

Correlation Identification

The following table provides an overview of the Spearman correlation results. The upstream discharge data show the strongest positive correlations with sensor_turbidity, while temperature-related data and EC values in Nieuwegein exhibit the strongest negative correlations. The time lags indicated in the table will be used to align the sensor_turbidity time series with the other parameters. Among collected time series, 11 of them were identified with maximum Spearman correlation coefficients exceeding 0.3 or falling below -0.3. These belong to four categories of parameters: upstream discharge and turbidity showed strong positive correlations with sensor_turbidity, while EC and water temperature exhibited strong negative correlations. The strongest positive correlations with sensor_turbidity were observed for upstream discharge and turbidity time series. Increased discharge from upstream areas, especially during storm events or snowmelt, typically mobilizes sediments, organic matter, and other suspended solids from the riverbed and catchment surfaces, leading to elevated turbidity downstream. These suspended materials are transported with the flow, causing spikes in turbidity levels at downstream monitoring points such as the DWTP intake. In contrast, water temperature and EC measured near the DWTP showed strong negative correlations with sensor turbidity. For EC, this inverse relationship is often explained by the dilution effect: high turbidity events usually occur during periods of increased discharge, which introduce large volumes of sediment-rich but ion-poor runoff into the system. As a result, the concentration of dissolved ions—and thus EC—tends to decrease when turbidity rises [60]. Water temperature also tends to be lower during high-discharge events, where rainfall or snowmelt contributes cooler water to the system. Moreover, warmer temperatures generally promote particle settling due to decreased water viscosity, leading to lower turbidity under calm, stable conditions [61]. Therefore, both EC and temperature exhibit inverse relationships with turbidity and provide valuable complementary information for turbidity prediction.

Table 27: Average traveling time from peak match method

Parameter	Max correlation	Time lag (h)
sensor_turbidity	1	0
Lob_turbidity	0.3391	118
Tiel_dis	0.3292	86
Lob_dis	0.3272	104
Hag_dis	0.3223	91
Pan_dis	0.322	101
Nieu_dis	0.1726	0
sensor_ph	0.1006	0
RH_daily_accum	0.0494	156
RH	0.0328	120
Nieu_wl	-0.0718	52
Wijk_wl	-0.074	52
Lob_EC	-0.1127	53
P	-0.1269	339
Lob_pH	-0.1639	152

Parameter	Max correlation	Time lag (h)
sensor_flow	-0.2285	0
Nieu_EC	-0.3293	0
T	-0.3896	36
sensor_wt	-0.4252	0
Nieu_wt	-0.4256	9
Hag_wt	-0.4292	8
Lob_wt	-0.4343	8

Based on the Spearman correlation analysis presented in the previous section, feature selection was carried out to identify candidate variables for ML model training. Parameters that exhibited clear positive or negative correlations with the target variable sensor_turbidity were first shortlisted as potential predictors. These selected features were then used to generate SOMs to explore non-linear relationships and potential feature redundancy. This two-step process ensured that only relevant and informative variables were considered in subsequent modeling efforts.

These 11 timeseries and sensor_turbidty were used to generate SOMs, with the results shown in the following figure. The first row of SOMs presents senor_turbidity and its positively correlated variables. High turbidity events at the DWTP (bottom-right corner of the SOM) are consistently associated with periods of elevated discharge in the Lower Rhine catchment. This pattern is also reflected in the turbidity measurements at Lobith, suggesting that increased upstream discharge enhances the transport of suspended matter into the downstream reaches and subsequently affects raw water quality at DWTP.

The second row shows variables that are negatively correlated sensor_turbidity. In contrast to the first row, high turbidity events at the DWTP are associated with low air and water temperatures. Notably, the SOMs of Nieu_EC exhibit an opposite pattern compared to the discharge-related SOMs. This finding is consistent with the diluting effect of increased river discharge on ionic concentrations. Elevated discharge reduces the EC of the river water by diluting solute concentrations, while simultaneously increasing turbidity through enhanced mobilization and transport of suspended solids.

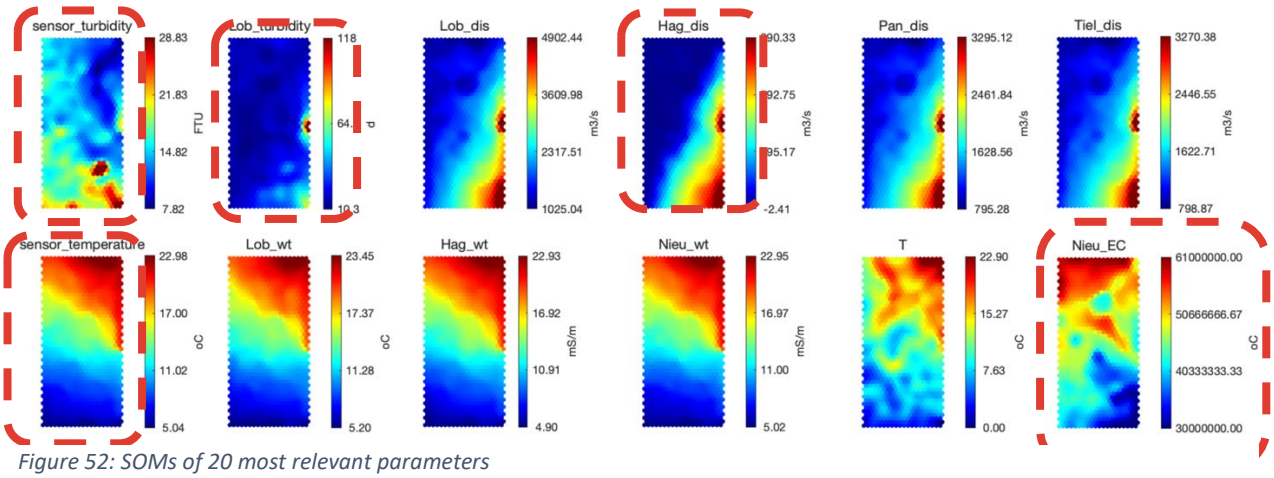


Figure 52: SOMs of 20 most relevant parameters

In addition to the numerical features, two categorical attributes, flow_direction and month, were included in the SOM analysis to provide contextual insight into the observed turbidity patterns. The flow_direction attribute represents the direction of water movement in the Lekkanaal: a value of 1 indicates flow from the Lek River toward the Amsterdam-Rhine Canal (northward), while -1 indicates the

reverse(southward). Incorporating flow_direction helps to identify the primary source of turbidity in Lekkanaal since the flow there is bidirectional. Similarly, the month attribute adds seasonal context, helping to identify temporal patterns in turbidity related to climatic and hydrological cycles.

As shown in the following figure, high turbidity events are predominantly associated with a positive flow direction, indicating that water flows from the Lek River toward the Amsterdam-Rhine Canal. This suggests that the Lek River is a major source of turbidity during these events. This observation aligns with the earlier correlation analysis, where upstream discharge showed a strong positive correlation with sensor turbidity. The Lek River generally has a higher discharge than the Amsterdam-Rhine Canal and therefore a greater capacity to transport suspended solids. The categorical SOM for the Month attribute shows that high turbidity events occur most frequently between December and February, corresponding to the winter season. This seasonal pattern supports the previous finding of a negative correlation between turbidity and water temperature, as colder months are associated with lower temperatures.

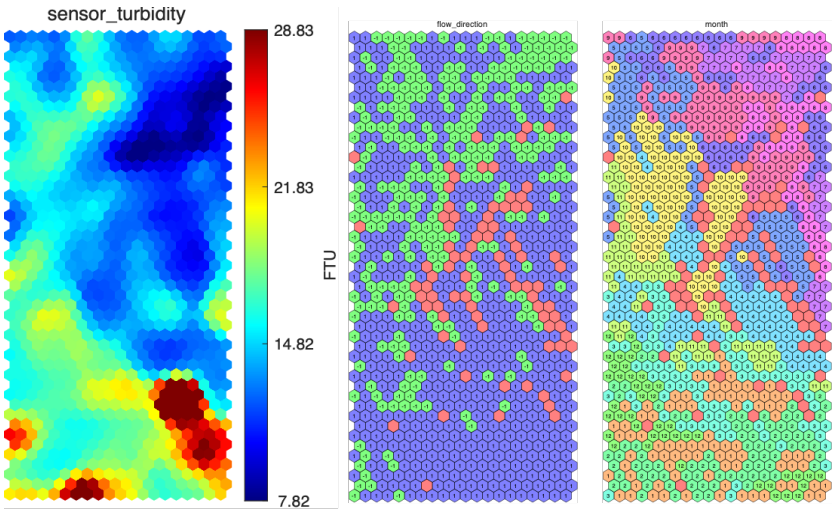


Figure 53: Categorical SOMs

The SOM results showed that the 11 timeseries selected from correlation analysis were relevant for high turbidity events, together with two categorical attributes. However, SOMs derived from parameters of the same kind (e.g., discharge and temperature) were highly like one another, suggesting redundancy within these groups. To avoid overlapping information and improve interpretability, Hag_dis and sensor_temperature were chosen as representative variables from their respective groups, based on their proximity to the DWTP location.

In contrast, the categorical attribute month exhibited minimal variability over the short time horizon considered in this study (24 hours of input to predict the next hour). Since such long-term seasonal effects did not meaningfully contribute to short-term turbidity dynamics, this feature was excluded. Consequently, Hag_dis, sensor_temperature, Nieu_EC, Lob_turbidity, and Flow_direction were selected as candidate features for model training, in addition to sensor_turbidity itself.

Predictive model comparison

This section compares the forecasting performance of ARIMA, RF and LSTM—across different forecast horizons. To ensure a fair comparison, all models were trained using only sensor_turbidity as the input feature. This aligns with the ARIMA model's univariate nature and isolates the models' ability to learn temporal patterns from turbidity measurements alone. The following figure shows the schematic representation of the model setup.

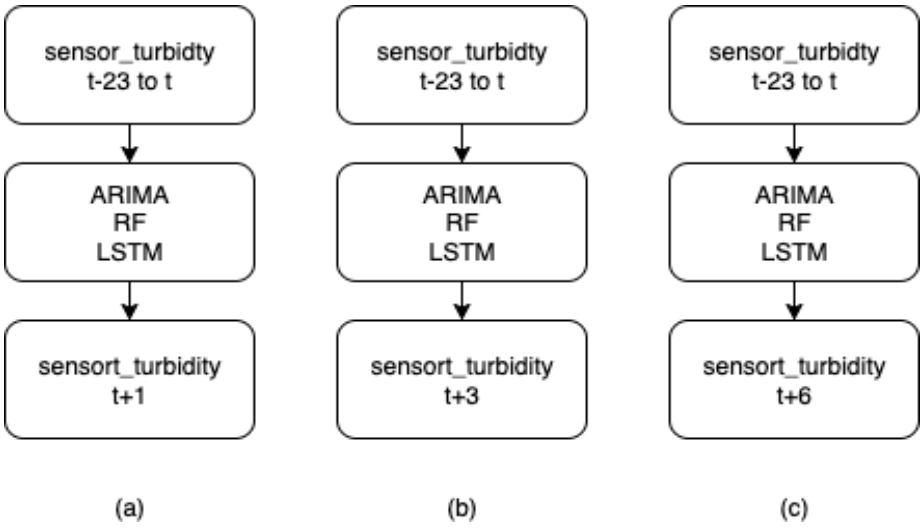


Figure 54: Comparison of model performance across three prediction horizons (t+1, t+3, t+6)

1-hour Ahead Forecast

Figure 55 illustrates the data split used in the 1-hour ahead forecast experiment. Each model was trained on the first 85% of the time series, spanning from January 1, 2020, to July 22, 2023. The remaining 15%, from July 22, 2023, to February 27, 2024, was held out as the test set for evaluating predictive performance.

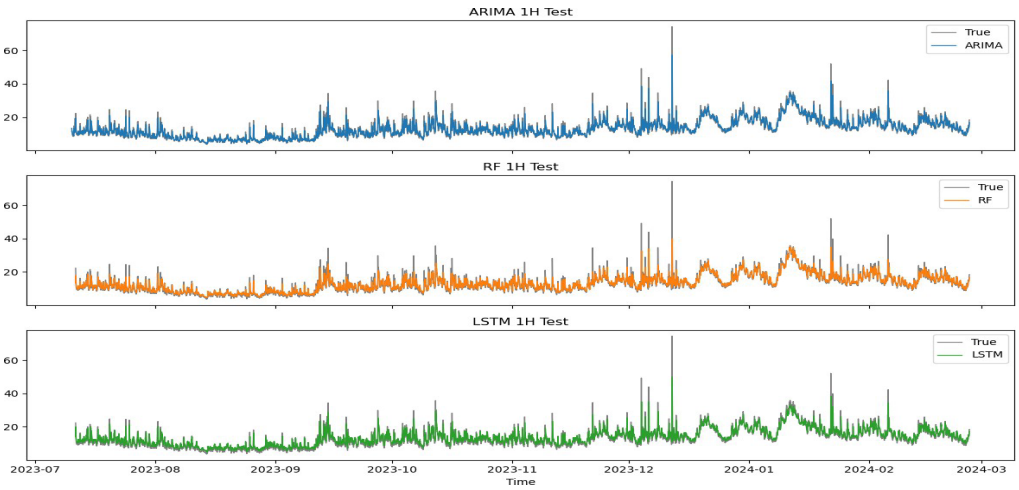


Figure 55: Predictions of three model on test period (Forecast window = 1 hour)

Table 28: Forecast results of three models (Forecast window = 1 hour)

Model	NSE	log NSE	MSE	RMSE	MAE	Relative abs error
ARIMA	0.80	**0.87**	5.60	2.37	**1.25**	**9.19%**
RF	**0.81**	**0.87**	**5.34**	**2.31**	1.30	10.15%
LSTM	0.80	0.85	5.71	2.39	1.49	12.37%

The above table summarizes the performance of the three models on the test set for 1-Hour forecasting. All models achieve superior fit (NSE > 0.8). LSTM performed slightly worse in terms of log NSE (0.85) compared to the other two models (0.87), indicating LSTM performed worse in low-magnitude turbidity events. The RF model also outperformed others in terms of MSE (5.34) and RMSE (2.31), reflecting better handling of large errors. Conversely, ARIMA achieved the lowest MAE (1.25) and relative absolute error (9.19%), suggesting it performs well on typical values. LSTM's highest MSE (5.71) and MAE (1.49) indicate reduced accuracy, underperforming the other two models in terms of typical values and outliers.

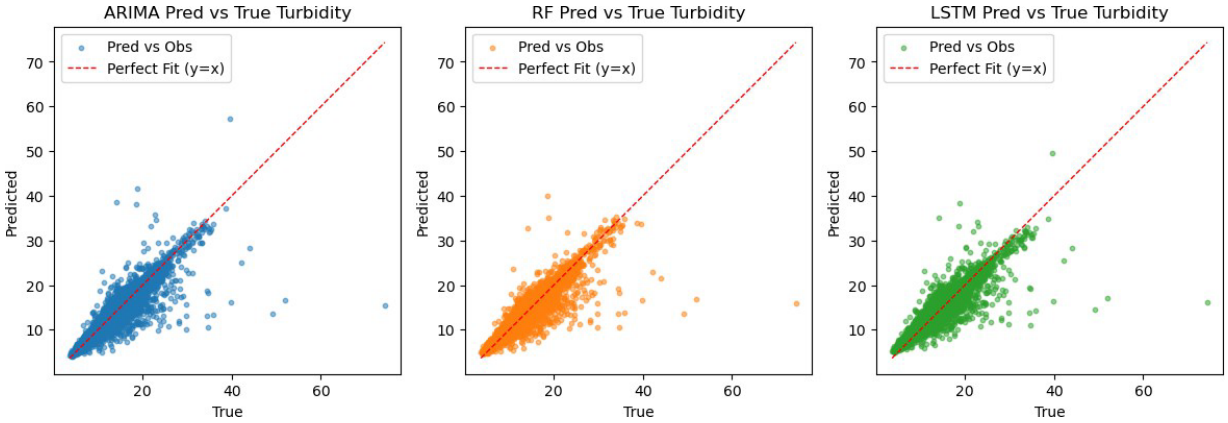


Figure 56: Scatterplot of predictions versus true values of three models (Forecast window = 1 hour)

The above figure 56 shows error scatterplots. Most errors cluster along the zero-error line, indicating good overall fit for typical turbidity values. However, all three models consistently under- estimate peak events, as shown by low prediction at high true turbidity values in the lower part of the plots. ARIMA exhibits the most symmetrical error pattern, with relatively same amount of overestimation (positive errors) and underestimation (negative errors). In contrast, RF shows less pronounced overestimation with fewer large positive errors, contributing to its lowest MSE and RMSE. LSTM displays distinct overestimation for low-turbidity events (e.g., August 2023), aligning with its lowest log NSE.

Figure 56 shows the density of errors from three models. All three curves exhibit positive skewness with longer tails on the left, indicating a tendency to overpredict during non-peak periods and occasional underprediction during peak events. Among the models, ARIMA demonstrates the least bias, with a mean error of -0.0485, aligning with its lowest MAE. In contrast, LSTM exhibits the highest bias (mean error of 0.5448), consistent with its highest MAE, suggesting systematic overprediction. RF, with a mean error of 0.1785 and the lowest standard deviation (2.3037), supports its stability across peak and non-peak events.

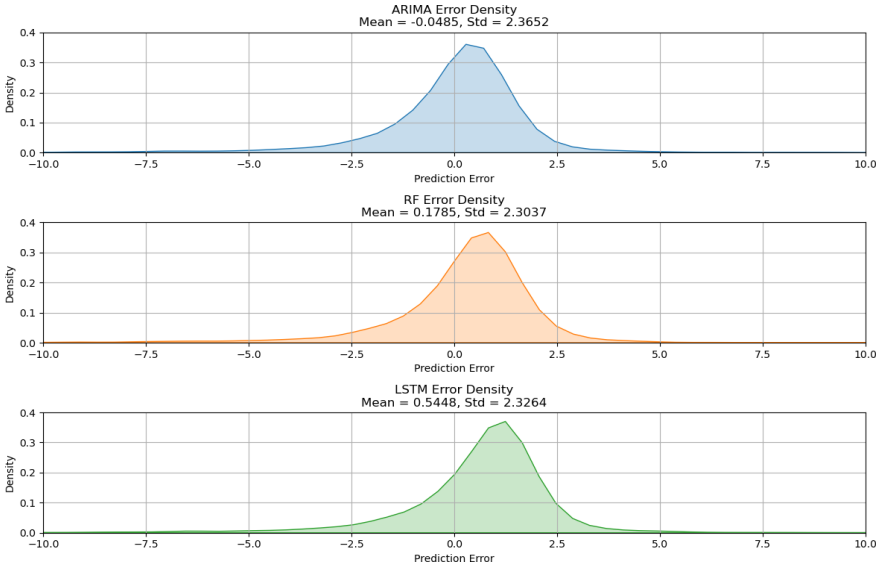


Figure 57: Error density of three models (Forecast window = 1 hour)

All three models deliver robust 1-hour-ahead turbidity predictions but consistently underestimate peak events. LSTM underperforms, with pronounced overprediction in low-turbidity periods. ARIMA provides the least biased predictions, ideal for non-peak forecasting. RF minimizes overestimation with the lowest RMSE.

3-hours ahead Forecasting

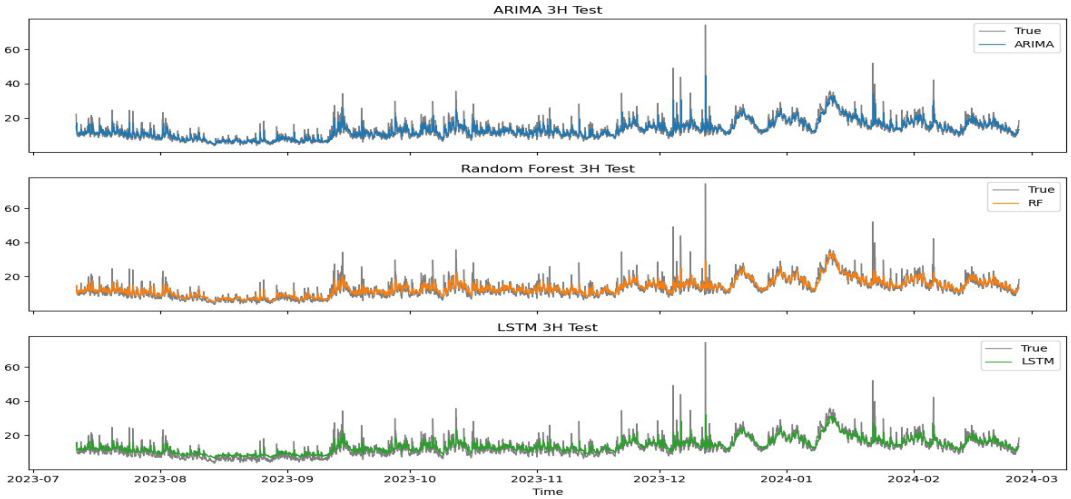


Figure 58: Predictions of three model on test period (Forecast window = 3 hour

Table 29: Forecast results of three models (Forecast window = 3 hour)

Model	NSE	log NSE	MSE	RMSE	MAE	Relative abs error
ARIMA	0.68	**0.76**	9.13	3.02	**1.84**	**14.09%**
RF	**0.70**	**0.76**	**8.46**	**2.91**	1.90	15.45%
LSTM	0.68	0.69	9.22	3.04	2.10	18.58%

In the above table the 3-hour forecast prediction performance is summarized. Compared to 1-hour forecasts, all models show increased underfitting, with NSE dropping from 0.8 to around 0.69. RF maintains the highest NSE (0.70) and lowest MSE (8.46) and RMSE (2.91), while ARIMA achieves the lowest MAE (1.84) and relative absolute error (14.1%). LSTM's log NSE (0.69 vs. 0.76) indicates worsening performance for low-turbidity events, with the highest MSE (9.22) and MAE (2.10). Overall, 3-hour predictions mirror 1-hour trends but with reduced accuracy. RF remains best for dynamic conditions, ARIMA for non-peak events, and LSTM continues to underperform.

6-hours ahead Forecasting

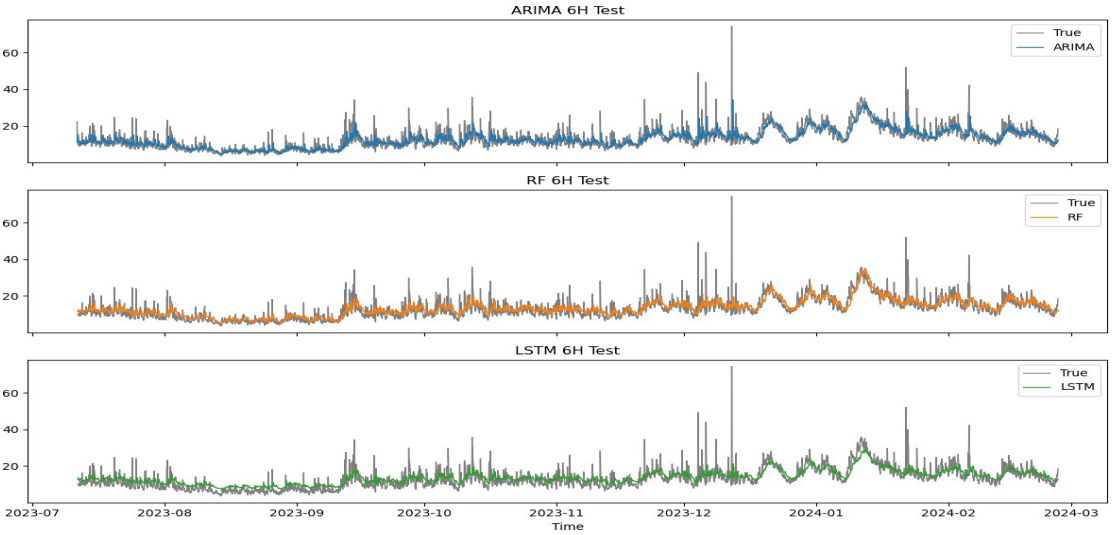


Figure 59: Predictions of three model on test period (Forecast window = 6)

Table 30: Forecast results of three models (Forecast window = 6 hour)

Model	NSE	log NSE	MSE	RMSE	MAE	Relative abs error
ARIMA	0.62	0.69	10.74	3.28	2.17	**16.74%**
RF	**0.66**	**0.70**	**9.76**	**3.12**	**2.16**	17.71%
LSTM	0.62	0.61	10.99	3.31	2.46	22.09%

Table 30 shows a further decline in forecast accuracy compared to 3-hours results. RF maintains the best overall performance with the highest NSE (0.66) and lowest MSE (9.76) and MAE (2.16). ARIMA retains the lowest relative absolute error (16.7%), while LSTM's performance deteriorates further, with the lowest log NSE (0.61) and highest errors.

Overall, 6-hour forecasts demonstrate the expected decline in accuracy, with RF's relative stability reaffirming its suitability for longer-term operational forecasting. Across all forecast horizons (1-, 3-, and 6-hours), ARIMA, RF, and LSTM consistently underestimate peak turbidity events, and accuracy diminishes as the lead time increases. RF consistently achieves the lowest MSE and RMSE, increasingly outperforming the other models at longer horizons. ARIMA performs best for short-term, non-peak conditions, delivering minimal bias and the lowest MAE. Meanwhile, LSTM underperforms overall, particularly struggling with low-turbidity events and rare extremes, exhibiting the highest bias and error variability.

Feature attribution analysis

The sensor_turbidity time series has been the sole input for all three models so far. To optimize model performance, greedy forward feature selection and feature importance analysis were employed to evaluate the impact of additional input features. The objective was to assess whether incorporating highly correlated features, identified earlier, improves prediction accuracy.

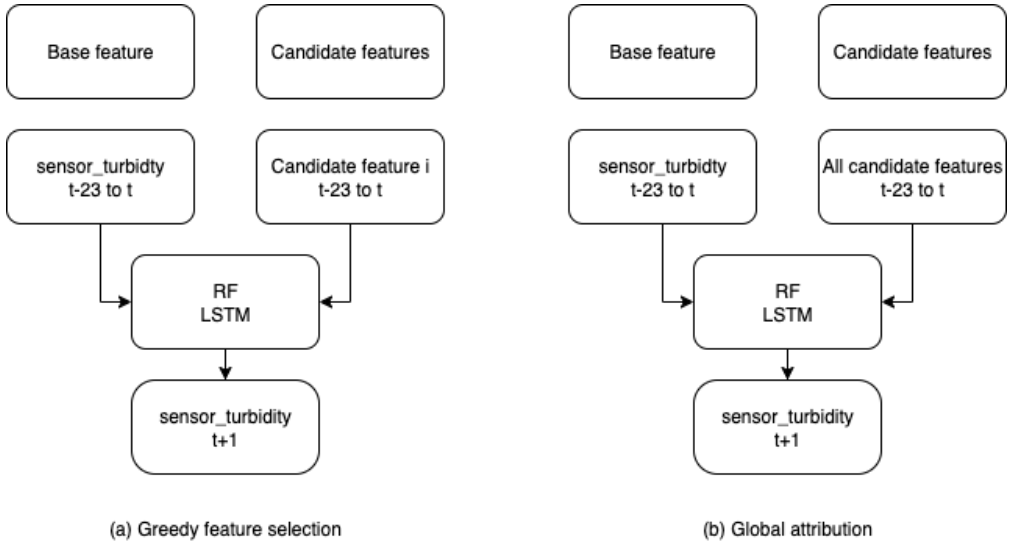


Figure 60:Feature attribution analysis experiments

Figure 60 illustrates the two experimental setups for feature attribution analysis.
A) **Greedy feature selection.** Each candidate feature is individually added to a base feature and used to train a separate model. The change in performance relative to the base feature quantifies the marginal contribution of that candidate. This approach captures the isolated effect of each feature.

B) **Global attribution.** All candidate features are combined with the base feature to train a single model. Feature importance is then computed for all features simultaneously using SHAP.

Figure 61 presents the key performance metrics from the RF model's greedy forward feature selection. The base model, using only sensor_turbidity, outperforms all other feature combinations across all metrics, suggesting that additional features introduce noise and reduce RF model accuracy.

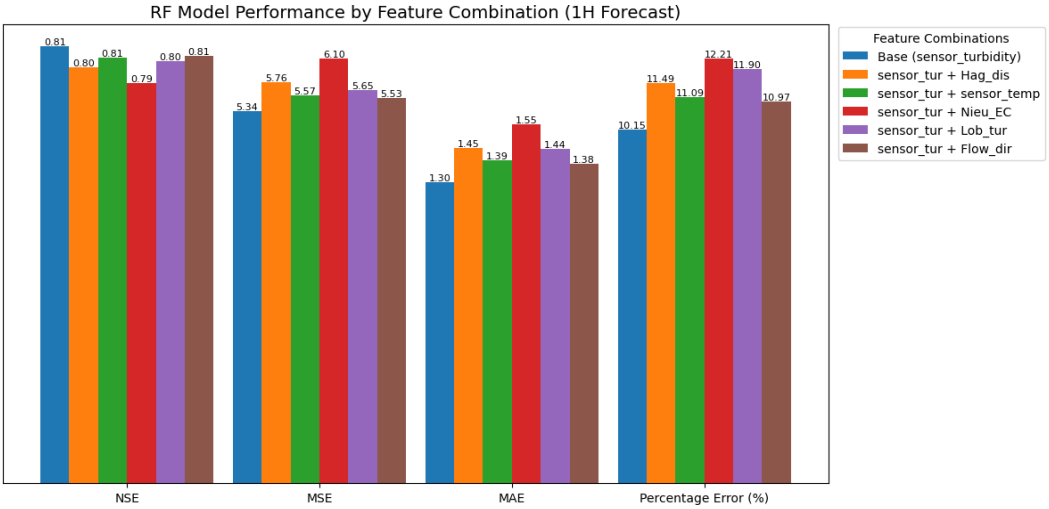


Figure 61: RF model greedy feature selection results (Forecast window = 1)

All five additional features are used to train a new RF model for important analysis. The below figure displays the top 10 most important features for the RF model, all of which are sensor_turbidity time lags. The importance of sensor_turbidity t-1 to t-3 are substantially higher than others, aligning with the ARIMA BIC test, which selected ARIMA(3,0,1) as the optimal model, indicating that the three most recent data points are critical for 1-hour-ahead predictions.

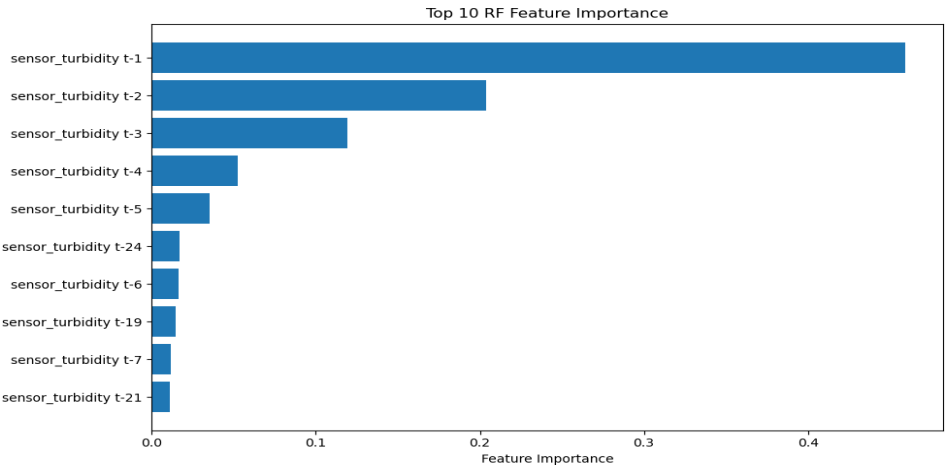


Figure 62: RF feature importance

Figure 62 presents a SHAP summary plot, highlighting the top 10 features based on their mean absolute SHAP values. The results reveal that sensor_turbidity time lags, particularly t-1, t-2, and t-3, dominate the feature importance rankings. This indicates that recent turbidity measurements are the primary drivers of accurate predictions, where high values of sensor_turbidity t-1 to t-3 (red dots on the right) strongly increase predicted turbidity and low values (blue dots on the left) decrease it.

The SHAP analysis reinforces the finding that sensor_turbidity alone is sufficient for robust RF predictions, as additional features introduce noise without substantial benefits. This insight guided the decision to retain the base model for operational forecasting, balancing accuracy and model simplicity.

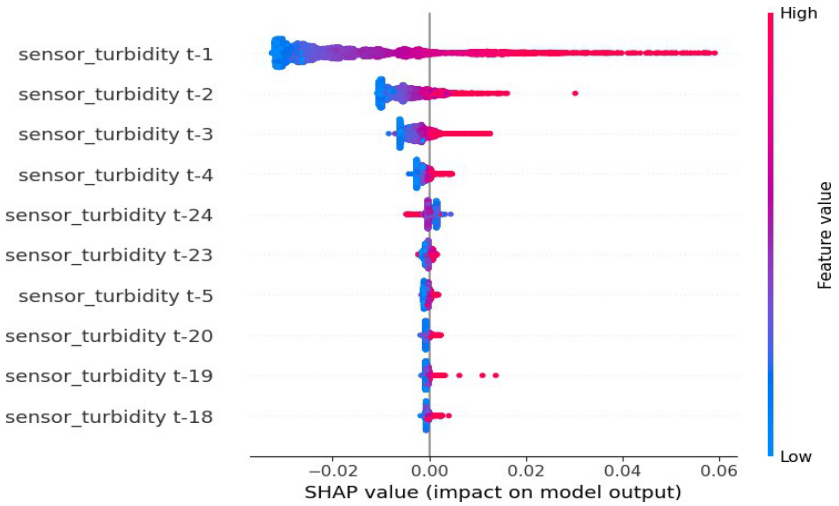


Figure 63: RF SHAP value

In summary, the inclusion of additional features did not lead to improved performance for either the Random Forest or LSTM models in 1-hour ahead turbidity forecasting. Instead, the most recent three turbidity values consistently emerged as the most informative inputs. This suggests that, despite strong correlations identified during feature selection, the added variables failed to provide meaningful predictive insight. Consequently, short-term turbidity forecasting in this context appears to be more of a statistical pattern recognition task than one governed by underlying physical processes. This also explains why simpler models such as ARIMA and RF outperformed the more complex LSTM model, where no significant temporal lag or multivariate interaction to exploit.

6.3.7 Conclusion and next steps

A methodology was developed for understanding the factors that influence high turbidity levels in the drinking water treatment plant (DWTP) intake, and short-term forecasting turbidity with up to certain hours ahead. The key conclusions of this work are as follows:

- Turbidity in the Lekkanaal DWTP is influenced by a combination of physicochemical parameters. High upstream discharge and high turbidity in the river influence the turbidity in the DWTP inlet.
- High-flow events carry significant number of suspended solids in the river which later appear in the DWTP inlet
- Electrical conductivity and water temperature have a reverse corelation with the sensor_turbidity, a finding that indicates most of the events mainly occur in the winter period where rainfall and snow are frequent.
- The seasonal pattern of high turbidity aligns also with the increased discharge of water in the river which occurs mainly in the colder months of the year and promotes the mobilization and transport of particulate materials.
- Feature selection analyses indicated that demonstrated that a univariate model relying solely on sensor_turbidity outperforms multivariate models. This finding highlights the importance of model simplicity, as additional correlated features failed to enhance forecasting performance.
- All models exhibited robust performance for 1-hour-ahead predictions, with NSE exceeding 0.8, however this performance was dropping when the predicting horizon was increased to 3-hour or 6-hour ahead. Nevertheless, RF managed had minimal performance degradation producing a robust model capable of understanding sudden change in the turbidity levels.

- The main limitation of the model is its inability to predict high peak events primarily due to the limited number of such events available for training.

6.4 Soft Sensor 10 - Prediction of the ozonation exposure, CT, to improve the ozonation process.

6.4.1 Problem statement and soft-sensor development flow-chart

Ozonation is widely employed in DWTPs for disinfection, oxidation of micropollutants, and the formation of biodegradable organic matter, which is subsequently removed in activated carbon filtration. In this process, ozone gas is produced by ozone generators and dissolved into water within an ozonation tank, typically consisting of multiple chambers. The efficacy of ozonation is influenced by several factors, with ozone dosage being the most critical. Other factors, such as water temperature, the presence of natural organic matter, and ultraviolet exposure, also play important roles. However, the complex and multi-dimensional interactions between these factors make real-time modeling of the ozonation process challenging, often requiring continuous recalibration of kinetic parameters.

Due to these complexities, water utilities frequently rely on empirical methods to adjust ozone dosage, which can lead to insufficient disinfection or oxidation during critical periods, or unnecessary overproduction of ozone—resulting in higher energy consumption and operational costs. This inefficiency not only impacts the environment but also raises the overall operational costs of the DWTP.

At the Leiduin DWTP, operated by Waternet, process engineers follow a semi-empirical approach to control ozone dosage. Weekly water samples are collected from different chambers of the ozonation process, and disinfectant (ozone) exposure (CT) is calculated by multiplying the ozone concentration in each chamber by the time required for water to pass between chambers. The measured CT values are then compared with the required CT levels for inactivating pathogens such as E. coli, Giardia, and viruses, as outlined in the US EPA guidelines [62] . Adjustments to the ozone dosage are made based on the highest required CT. While this approach is reasonable for periodic control, it is reliant on discrete water samples and therefore cannot account for bacteriological activity between sampling periods. As a result, the system may fail to detect sudden water quality deterioration, or it may lead to overproduction of ozone during periods of low bacteriological activity, unnecessarily increasing energy consumption.

To address these challenges, this study proposes the development of a soft sensor based on data-driven models. The soft sensor aims to provide daily predictions of CT values, using historical process data such as water temperature, flow rate, and ozone dosage. Six different machine learning techniques and combination of those have been tested and compared: random forest (RF), gradient boosting (GB), eXtreme gradient boosting (XGB), Adaptive boosting (AB), light gradient boosting (LGB), and a Physics-Informed Neural Network (PINN) [63,64,65,66,67]. The PINN integrates a data-driven approach with an ordinary differential equation (ODE) that describes the kinetics of ozone consumption in the ozonation tank. This hybrid approach leverages both the predictive power of machine learning and the reliability of physical laws governing ozone decomposition.

By adopting this soft sensor model, Waternet's process engineers will have the capability to adjust ozone dosage on a daily basis, enabling more precise control over the ozonation process. This approach not only improves the efficiency of pathogen inactivation but also reduces energy consumption, operational costs, and environmental impact. Ultimately, this work provides a framework for real-time process optimization, ensuring that high-quality drinking water is consistently delivered to consumers.

By adopting these soft sensor models, Waternet's process engineers will have the capability to adjust ozone dosage on a daily basis, enabling more precise control over the ozonation process. Furthermore, they can evaluate the impact of increased ozone dosages on bromate formation and, therefore, ensure that the quality of the produced drinking water is of high standards. This approach not only improves the efficiency of pathogen inactivation but also reduces energy consumption, operational costs, and environmental impact. Ultimately, this work provides a framework for real-time process optimization, ensuring that high-quality drinking water is consistently delivered to consumers.

6.4.2 Data sources and data preprocessing

Data Collection

The ozonation tank at Leiduin DWTP comprises four separate lanes, each supplied with ozone from a dedicated generator. Each lane contains five chambers, facilitating staged ozone dissolution and distribution. This study utilized two primary data sources: **sensor data** from DWTP's SCADA system and **grab sample data** collected by Waternet at key points in the treatment process.

The sensor data, recorded hourly, were collected from critical locations, including the rapid sand filtration outlet, the ozone generators, and both the inlet and outlet of the ozonation tank. Meanwhile, grab samples were obtained weekly or bi-weekly from the rapid sand filtration outlet and the ozonation outlet, providing a snapshot of water quality parameters at these stages. The study covers a 4-year period from January 2020 to December 2024, offering a robust dataset for analysis. A summary of the specific parameters used in this work is presented in the following Table 31.

Table 31: Datasets used for the soft sensor development

Sensor data stored in the SCADA (units)	Grab samples water quality data (units)
- Turbidity in the Rapid Sand Filtration Outlet (FTU)	-Ozone concentration in compartments 1-5 - Lane 1 (mg/l)
- Water Temperature (°C) in the ozonation inlet	-Ozone concentration in compartments 1-5 - Lane 2 (mg/l)
- UV in the ozonation inlet (254nm)	-Ozone concentration in compartments 1-5 - Lane 3 (mg/l)
-UV in the ozonation outlet(254nm)	-Ozone concentration in compartments 1-5 - Lane 4 (mg/l)
- Ozone generation rate - Line 1 (g/m^3)	-Coliform bacteria in the Rapid sand Filtration outlet (col/l)
- Ozone generation rate - Line 2 (g/m^3)	-Coliform bacteria in the Ozonation outlet (col/l)
- Ozone generation rate - Line 3 (g/m^3)	-Bromide in the ozonation inlet (µg/l)
- Ozone generation rate - Line 4 (g/m^3)	-Bromate in the ozonation inlet (µg/l)
-Flow rate – Lane 1 (m^3/h)	-Bromate in the ozonation outlet (µg/l)
-Flow rate – Lane 2 (m^3/h)	- UV in the ozonation inlet (254nm)
-Flow rate – Lane 3 (m^3/h)	- UV in the ozonation outlet (254nm)
-Flow rate – Lane 4 (m^3/h)	
-Ozone gas flow rate – Lane 1 (Nm^3/h)	
-Ozone gas flow rate – Lane 2 (Nm^3/h)	
- Ozone gas flow rate – Lane 3 (Nm^3/h)	
- Ozone gas flow rate – Lane 4 (Nm^3/h)	

Sensor Data Preprocessing

Due to sensor sensitivity and potential fouling from material buildup, as well as periodic recalibration requirements, the raw sensor data often contained inaccuracies. To ensure reliable data quality, we implemented a four-step preprocessing approach:

1. **Timestamp Errors and Missing Data Replacement**: Timestamp errors and missing data points were identified and replaced using data interpolation, as there were no extended periods of missing data or large timestamp discrepancies. Interpolation was chosen because the errors and gaps were minimal and non-consecutive, making it suitable for filling short gaps in a time series.
2. **Single-Point Outliers Replacement**: Outliers were identified as values significantly deviating from surrounding data points. To detect these, we calculated the z-score (i.e., the difference between the point and the dataset's mean, divided by the standard deviation). Any data point with a z-score exceeding a threshold of 100 was flagged as an outlier and replaced via interpolation.
3. **Threshold-Based Replacement**: Waternet has established minimum and maximum acceptable thresholds for several parameters (e.g., temperature, flow, pH). For instance, acceptable water temperatures range between 3°C and 27°C. Data points outside these thresholds were deemed invalid and were replaced with interpolated values to maintain continuity.
4. **Drift Correction**: Some sensors, particularly turbidity sensors, are prone to drift over time. To correct for this, we calculated a four-week rolling mean. Successive changes in this weekly average were flagged as drift, and any data showing sustained shifts were corrected using asymmetric least squares regression.

After completing the data cleaning, we calculated the ozone dosage per lane at an hourly frequency. This calculation was based on the ozone generation rate, ozone gas flow rate, and water flow rate. Subsequently, daily averages for each parameter were computed to create a new dataset at a daily resolution.

Grab Samples Data Preprocessing

The grab sample dataset contains measurements collected at irregular intervals. The raw format generally includes three columns: collection date, parameter type, and measured value. The preprocessing involved reformatting this dataset to ensure consistency, so each row represented a distinct date, with each column containing values of specific water quality parameters in a specific treatment process outlet (e.g., UV in ozonation inlet). Regarding the actual ozone exposure (CT), this was calculated as the sum the ozone measured at each chamber times the time that the water required to travel to this chamber from the previous one as the below formula describes:

$$CT = \sum_{i=0}^{n} c_{O_{3,i}*(t_i - t_{i-1})} \qquad \text{Eq. 29}$$

where $c_{O3,i}$ is the ozone concentration at chamber I (mg/l), *n* the total number of chambers and *t* time in minutes.

Combining Datasets

The final preprocessing step involved merging the three datasets using the sample date as the common key. This ensured that all processed sensor and grab sample data were aligned to enable comprehensive analysis in the modelling stages.

6.4.3 Materials and methods

Machine learning algorithms selection

For the development of the soft sensors different data-driven approaches were employed and evaluated. More specifically, the models used were Random Forest (RF), gradient boosting (GB), extreme gradient boosting (XGB), light gradient boosting (LB), adaptive boosting (AB) and a Physics-Informed Neural Network (PINN).

Random Forrest

The Random Forest algorithm was selected due to its robustness in handling nonlinear relationships between input features and the target variable, as well as its ability to handle high-dimensional data. The RF model was trained to predict CT values based on selected input features, such as water temperature, flow rate, ozone dosage.

Gradient Boosting

Gradient Boosting is an ensemble technique that builds models sequentially, with each new model attempting to correct the errors of the combined ensemble so far. It minimizes a chosen loss function via gradient descent, making it a flexible and powerful framework for both regression and classification tasks.

XGboost

XGBoost is an enhanced and scalable version of gradient boosting that incorporates regularization, parallel processing, and efficient handling of sparse data. It's designed for speed and performance, making it one of the most widely adopted machine learning algorithms in practice.

Adaptive boosting

AdaBoost is one of the first boosting algorithms, which combines multiple weak learners, typically shallow decision trees, into a strong classifier. It iteratively adjusts the weights of misclassified samples so that subsequent models focus more on difficult cases, improving accuracy and reducing bias.

Light boosting

LightGBM is a gradient boosting framework by Microsoft that uses a histogram-based approach and leaf-wise tree growth to improve training efficiency and reduce memory usage. It is optimized for large datasets and high-dimensional data while maintaining high accuracy.

For all these models that belong in the ensemble decision trees machine learning category the hyperparameter optimization was implemented using the grid search optimising the number of trees, maximum depth, and minimum samples per leaf. The optimal configuration was selected based on the model's performance in cross-validation in the validation dataset, using Mean Squared Error (MSE) as the evaluation metric.

Physics informed Neural Network

Two Physics-Informed Neural Network (PINN) were developed to integrate physical knowledge of the ozonation process with the data-driven neural network approach, the first one was for the development of the CT soft sensor and the second for the development of the bromate soft sensor.

- **CT soft sensor**

This PINN embeds an ordinary differential equation (ODE) to capture the dynamics of ozone concentration changes within the ozonation tank. The governing ODE, representing ozone consumption, is defined as follows:

$$\frac{d_{C_{o_3}}}{dt} = -k_{O_3} c_{o_3}$$

where K_{o3} is the first order kinetic of ozone decomposition (min^{-1}) and C_{o3} the concentration of ozone (mg/l).

- **Bromate in the ozonation outlet soft sensor**

This PINN embeds the von Gunten & Hoigne model [57] to capture the dynamics of ozone concentration changes within the ozonation tank. The governing ODE, representing ozone consumption, is defined as follows:

$$C_{BrO_3} = C_{BrO_3,ini} + k_{BrO_3} CT$$

$$C_{BrO_3,ini} = F_{BrO_3,ini} * C_{O_{3dos}} + C_{BrO_3,in}$$

where C_{BrO3} is the concentration of bromate in the ozonation outlet (µg/l), $C_{BrO3,ini}$, the initial bromate formation (µg/l), k_{BrO3}, the bromate formation rate constant (µg BrO_3/l /(mg O_3/l * min), $F_{BrO3,ini}$, the constant of initial bromate (µg BrO_3/l /(mg O_3/l), $C_{BrO3,in}$ the influent bromate.

The ODE for the O_3 estimation was used in the work of [58] to describe the ozone decomposition during the ozonation process in two steps the rapid ozone consumption step and the rather slow decay step. The von Gunten & Hoigne model is an empirical regression model for the estimation of the bromate in the inlet and the retention time in the ozonation reactor. In our case study, all the bromate samples collected in inlet of the process had 0 or very minimal concentration of bromate (maximum measured $C_{BrO3,in}$ = 0.002 µg/l), therefore it was assumed that the bromate concentration in the ozonation inlet is 0.

1. **Model Structure**: The PINN incorporates both measured data and the governing formula. A feedforward neural network serves as the data-driven component, while the formula regularizes the model output by penalizing deviations from the ozone decomposition dynamics.
2. **Loss Function and Training**: The PINN was trained using a combined loss function that includes both a data loss component (MSE between predicted and observed ozone concentrations) and a physics-based loss component (formula loss component). The latter penalizes deviations from the foruma by computing the residuals from the differential equation. The model was trained to minimize the total loss function, by treating both the data loss and the formula loss function equally.
3. **K constants estimation:** The constants estimations were calculated using the least square method for calibration in the training dataset.

Ozone Exposure Soft Sensor Development

Setting the input features and outputs of the CT soft sensor

The aim of this soft sensor is to estimate the CT at the specific day d and the specific hour h that the available sensors provide with measurements. As the ozone concentration in each chamber was measured at a specific hour h once every two weeks, for the validation of this model, only the data of the sampling date and hour as well as a 1-hour time lag were used as inputs. An additional input variable was introduced, the residence time in the ozonation reactor. This is calculated by multiplying the flow with the volume inside the reactor. Regarding the outputs, for the ensemble decision tree models the calculated CT at the day d and hour h was set as output and for the PINN model the outputs were the ozone concentrations at each compartment, and the final CT is calculated using the CT formula described above.

Table 32: The inputs and the outputs of the soft sensor.

Input Variables	Source	Output data
- Hourly ozone dosage (mg/l) at the day d for the hours h and h-1	Sensor Data from SCADA	CT at day d and hour h (ensemble decision trees models)
-Hourly Turbidity measurements at the day d for the hours h and h-1	Sensor Data from SCADA	O_3 at each chamber (PINN model)
- Hourly water flow (m^3/h) at the day d the day d for the hours h and h-1	Sensor Data from SCADA	
-Hourly temperature at the day d for the hours h and h-1	Sensor Data from SCADA	
-ozonation lane (chamber 1-4)	SCADA information	
-month of the year (1-12)	SCADA information	
-hour of the measurement (1-23)	SCADA information	
-residence time (h)	Calculated using flow and volume of the ozonation tank	

Grouping the input variables

The total number of available input variables were 12. For this investigation, three groups of input variables were selected. The first one included all 12 variables; the second group excluded the h-1 time lagged sensor data (8 input variables) and the third group is using the second group variables but excludes the residence time and the month of the year.

Table 33: The three group of inputs used for the CT soft sensor.

ID	Input Variables	Number of variables
CT.I.V.1	All	12
CT.I.V.2	Turbidity, ozone dosage, water flow, temperature at day d and hour h, chamber, month, hour, residence time	8
CT.I.V.3	Turbidity, ozone dosage, water flow, temperature at day d and hour h, chamber, hour,	6

Ensemble approaches

To further investigate the potential for improved predictive performance, an ensemble approach will be employed, combining the outputs of the best-performing models. Three ensemble techniques will be considered:

- **weighted averaging**: model predictions are combined using weights proportional to their individual performance during the initial testing period
- **Optimization weight**: The contribution of each model is systematically adjusted to minimize the overall prediction error
- **Stacking**: Stacking is the approach where a meta-model is generated to learn the most effective way to combine the initial model outputs.

By leveraging the complementary strengths of different models, these ensemble methods aim to enhance both robustness and accuracy of the CT soft sensor estimations.

Bromate in the DWTP outlet Soft sensor development

Setting the input features and outputs of the Bromate soft sensor

The objective of this soft sensor is to predict bromate concentration in the DWTP outlet. For its development, the same dataset used for the CT soft sensor is employed, with the addition of the CT estimates provided by the final trained CT soft sensor as an input variable. The total number of variables used in the bromate soft sensor is summarized in the following table 34.

Table 34: Input variables used in the development of bromate soft sensor

Input Variables	Source	Output data
- Hourly ozone dosage (mg/l) at the day d for the hours h and h-1	Sensor Data from SCADA	Bromate (BrO_3) in the ozonation/DWTP outlet
-Hourly Turbidity measurements (FTU) at the day d for the hours h and h-1	Sensor Data from SCADA	
- Hourly water flow (m^3/h) at the day d the day d for the hours h and h-1	Sensor Data from SCADA	
-Hourly temperature at the day d for the hours h and h-1	Sensor Data from SCADA	
-Ozone exposure (CT – mgO_3 * min/l) at the day d for the hour h	Provided by the CT soft sensor	
-ozonation lane (chamber 1-4)	SCADA information	
-month of the year (1-12)		
-hour of the measurement (1-23)	Calculated using flow and volume of the ozonation tank	
-residence time (h)		

Grouping the input variables

The total number of available input variables were 13. For this investigation, two groups of input variables were selected. The first one included all 13 variables; the second group excluded the h-1-time lagged, the residence time and the month of the year.

Table 35: The two group of inputs used for the bromate in the DWTP outlet soft sensor.

ID	Input Variables	Number of variables
Br.I.V.1	All	13
Br.I.V.2	Turbidity, ozone dosage, water flow, temperature at day d and hour h, chamber, hour,	7

Performance evaluation

Each model was evaluated on its ability to predict ozone concentration and CT values using 3 different metrics, mean square errors (MSE), root mean square errors (RMSE) and the coefficient of determination (R^2). The formulas for these performance metrics are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(pred_i - obs_i)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(pred_i - obs_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(pred_i - obs_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(pred_i - obs_i)^2}{\sum_{i=1}^{n}(pred_i - obs_{mean})^2}$$

where $pred_i$ is the predicted value for observation i, obs_i, the actual value for observation I, n, the number of samples, and obs_{mean} the mean of the observed values.

6.4.4 Results

Ozone Exposure Soft Sensor

The model was trained using 80% of the dataset and then it was tested in the remaining 20% which remained unseen. All algorithms were tested to understand which one performs better for this dataset. All models were initially trained using the CT.I.V.1 variables. Following these results, the best models were then used using the CT.I.V.2 and the CT.I.V.3 variables. The PINN model was only trained using the CT.I.V.3 variables. All the results are presented in the following Table 36.

Table 36: Performance metrics of the model in the testing dataset for the CT estimation soft sensor.

Model	Variables group	MAE	MSE	RMSE	R^2
RF	CT.I.V.1	0.223	0.111	0.322	0.49
GB	CT.I.V.1	0.22	0.111	0.333	0.508
XGB	CT.I.V.1	0.21	0.10	0.323	0.537
LGB	CT.I.V.1	0.255	0.116	0.34	0.48
AB	CT.I.V.1	0.228	0.122	0.348	0.46
RF	CT.I.V.2	0.22	0.089	0.297	0.61
GB	CT.I.V.2	0.21	0.072	0.268	0.682
XGB	**CT.I.V.2**	**0.19**	**0.06**	**0.245**	**0.724**
RF	**CT.I.V.3**	**0.19**	**0.083**	**0.288**	**0.65**
GB	**CT.I.V.3**	**0.21**	**0.07**	**0.262**	**0.696**
XGB	**CT.I.V.3**	**0.19**	**0.06**	**0.252**	**0.718**
PINN	CT.I.V.3	0.34	0.11	0.331	0.42

The best performing model is the XGB model using the CT.I.V.2 variables with the same model using the CT.I.V.3 being closely in the second place. The performance of the PINN was suboptimal, likely due to the low values and limited variance of ozone concentrations in the dataset. The 4 best performing models were then used in the ensemble approach and the results are presented in the following table 37.

Table 37: Performance metrics of the ensemble models for the CT estimation soft sensor

Model	MAE	MSE	RMSE	R^2
Weighted	**0.19**	**0.05**	**0.239**	**0.745**
Optimised	0.2	0.051	0.242	0.738
Stacking	0.21	0.08	0.283	0.71

Two of the three ensemble approaches enhanced the performance of the soft sensor. Among them, the weighted averaging approach achieved the best results and was therefore selected for the development of the CT soft sensor. The observed versus predicted plot shown below illustrates that this approach accurately predicts both low and high CT values, demonstrating the model's robustness and applicability under a wide range of operating conditions.

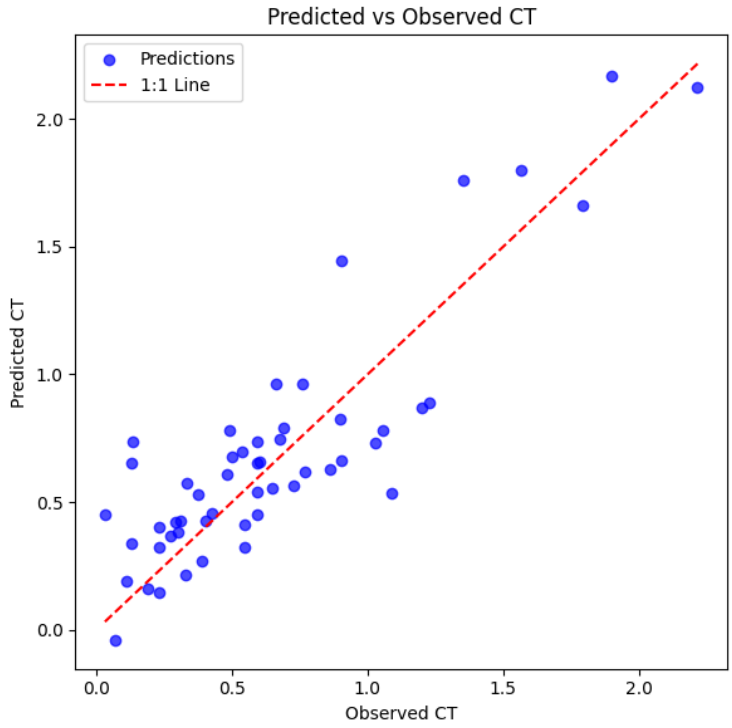


Figure 64: Predicted vs Observed ozone concentration values using the weighted average ensemble for 4 different models.

Bromate concentration in DWTP outlet Soft Sensor

The model was trained using 80% of the dataset and then it was tested in the remaining 20% which remained unseen. All algorithms were tested to understand which one performs better for this dataset. All models were initially trained using the Br.I.V.1 variables. Following these results, the best models were then used using the Br.I.V.2 variables. The PINN model was only trained using the Br.I.V.2 variables. All the results are presented in the following Table 38.

Table 38: Performance metrics of the model in the testing dataset for the Bromate concentration soft sensor.

Model	Variables group	MAE	MSE	RMSE	R^2
RF	Br.I.V.1	0.215	0.098	0.313	0.808
GB	Br.I.V.1	0.128	0.049	0.222	0.902
XGB	Br.I.V.1	0.116	0.038	0.194	0.925
LGB	Br.I.V.1	0.144	0.035	0.188	0.93
AB	**Br.I.V.1**	**0.07**	**0.02**	**0.148**	**0.956**
XGB	Br.I.V.2	0.166	0.058	0.242	0.885
LGB	Br.I.V.2	0.196	0.075	0.275	0.852
AB	CT.I.V.2	0.15	0.079	0282	0.845
PINN	**CT.I.V.3**	**0.34**	**0.11**	**0.331**	**0.42**

The best performing model is the AdaBoost model when all the parameters are used. The observed versus predicted values plot of this model is presented in the below figure 65.

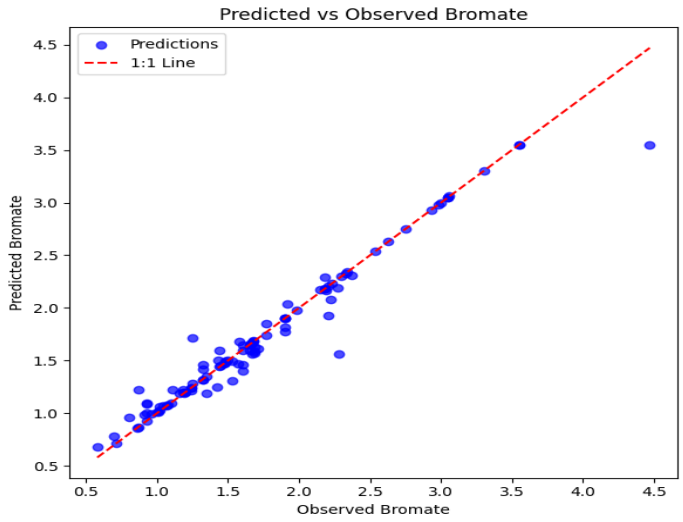


Figure 65: Predicted vs Observed CT values using RF model with 5 input variables

Plotting soft sensor outputs over time

A further analysis of the soft sensor outputs was performed by comparing their hourly predictions over randomly selected chronological periods. Specifically, the CT soft sensor results were compared with the CT values estimated using the EPA-recommended methodology to assess whether the applied ozone dosage was sufficient for bacterial inactivation or if over- or under-dosing occurred. The comparison, illustrated in the following figure, shows that during this particular period, the EPA-suggested CT values are considerably lower than those predicted by the soft sensor, indicating a potential ozone overdose within this timeframe.

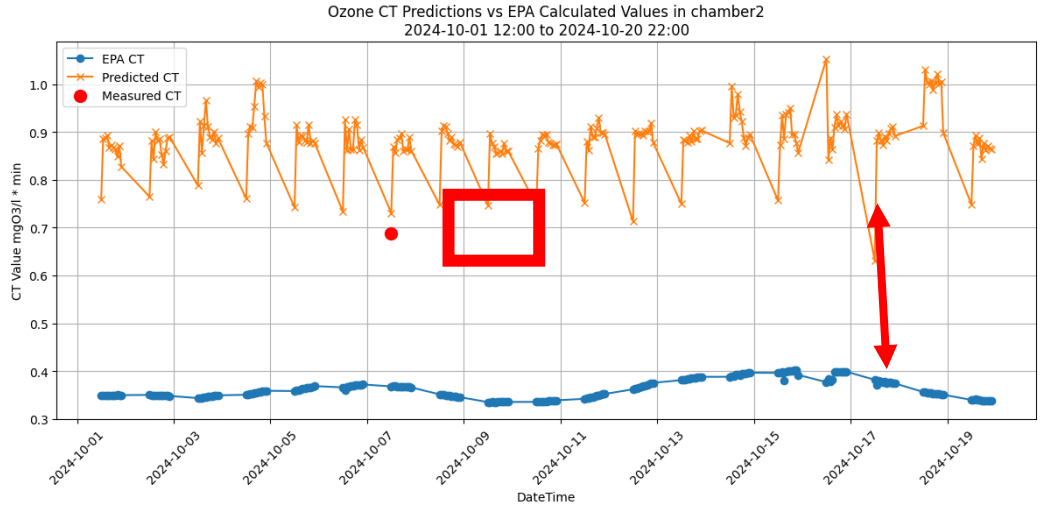


Figure 66 : CT soft sensor estimations, actual measured CT and CT suggestions by EPA

For the bromate soft sensor, a similar analysis was conducted (see figure below). In this case, the plot illustrates the variation of bromate concentration over time and its proximity to the water utility limit of 5 µg/L. This visualization allows for evaluating whether the predicted bromate levels remain within acceptable limits and for identifying potential periods of elevated bromate formation risk.

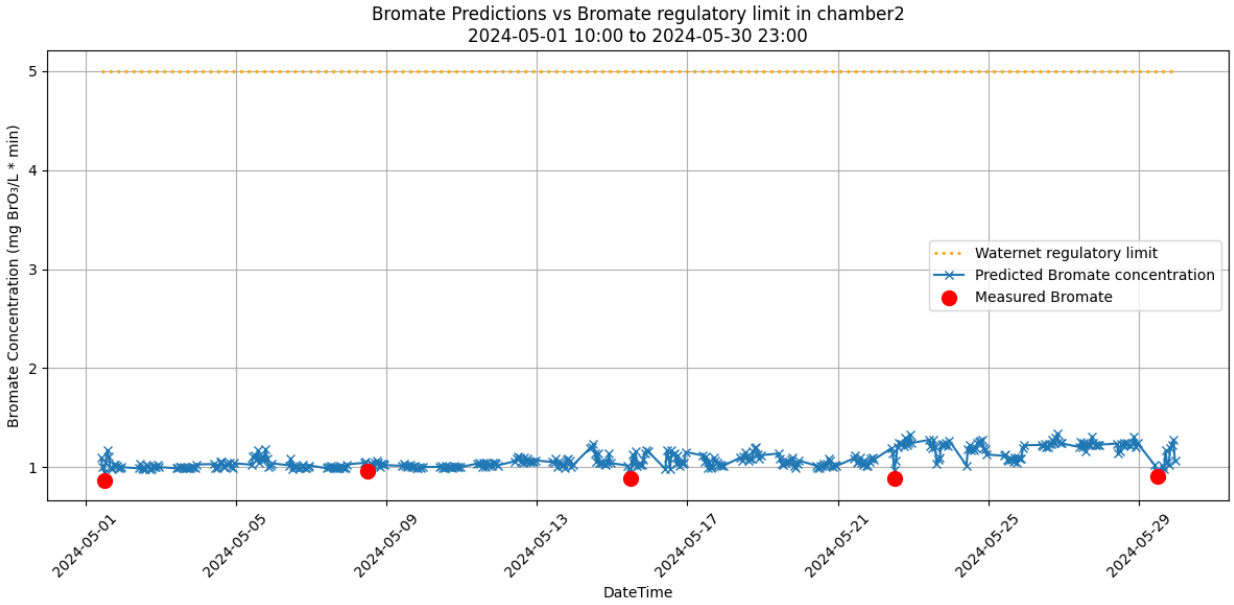


Figure 67: Bromate soft sensor estimations and actual measured Bromate over time

6.4.5 Conclusion and next steps

The development and evaluation of the CT and bromate soft sensors demonstrate the potential of data-driven and hybrid modeling approaches to enhance process monitoring and control in ozonation systems. By integrating historical process data with physics-based insights, the proposed soft sensors can support Waternet's process engineers in achieving more stable and efficient ozone dosing, while minimizing risks associated with bromate formation. The key conclusions from this work are as follows:

- **Enhanced predictive performance:** Among all models tested, ensemble-based approaches—particularly the weighted averaging technique—provided the most accurate CT estimations, effectively capturing both low and high CT ranges.
- **Bromate prediction capability:** The bromate soft sensor, using the CT estimates as an additional input, achieved high accuracy with the AdaBoost model, showing strong potential for predicting bromate concentrations near regulatory limits.
- **Operational insights:** Comparative analysis with EPA-based CT calculations revealed that the soft sensor can identify periods of potential ozone overdosing, enabling more informed operational adjustments.
- **Process optimization potential:** The combined use of CT and bromate soft sensors offers a practical framework for daily ozonation control, improving disinfection efficiency while reducing energy use and mitigating bromate formation risks.

7 Conclusions

This deliverable reports the development of the soft-sensor modules for water quality monitoring and performance improvement, fulfilling successfully the objectives of Tasks 4.1, 4.2, and 4.3, as defined in the Grant Agreement of the project. The work has transitioned from initial needs assessment and data identification to the final design, training, and validation of ten distinct soft sensors across three of the project's key demonstration cases.

The final module provides a suite of powerful, data-driven tools tailored to specific, high-priority challenges faced by water utilities:

Catchment and Source Protection (Athens & Val de Bagnes)

A set of six sensors was developed to provide enhanced situational awareness in complex water bodies. This includes an Early Warning System (EWS) for nutrient runoff in the Yliki Lake catchment, leveraging Earth Observation (EO) data to model crop types and hydrological processes. This EWS is complemented by sensors for monitoring Chlorophyll-a, pH, Dissolved Oxygen (DO), Algal Bloom Probability, and an integrated Water Quality Index (WQI). For the Alpine springs of Val de Bagnes, an EWS for bacteriological contamination was successfully developed, linking meteorological data to contamination events.

Drinking Water Treatment Plant Optimization (Amsterdam)

Three sensors were developed to improve the efficiency and reliability of specific treatment processes. These include a sensor for turbidity prediction in the coagulation-flocculation process, an EWS for turbidity events at the DWTP inlet and a predictive model for ozonation (CT) exposure.

From a methodological point of view, this work yielded several key insights that will inform the project's next phases. The development process confirmed that machine and deep learning models (like RF, LSTMs, and UNet) can effectively simulate the complex, non-linear relationships between EO data, in-situ sensor readings, and real-world water quality outcomes.

Critically, this work highlighted that data imbalance is a primary challenge in environmental modelling. Standard models often fail to predict rare but critical events, such as high Chl-a or low pH values. The implementation of techniques like inverse-frequency weighting proved essential for forcing the models to learn from these underrepresented data points and improve their operational robustness. Furthermore, the use of Explainable AI (XAI) methods (e.g., Gini Importance, SHAP, and PFI) was vital not only for "opening the black box" and validating that models were learning "the right things" but also as a practical tool for feature selection, helping to create simpler, more efficient models.

In summary, the soft-sensor module is complete. The findings from these ten sensors provide a robust, evidence-based foundation for the next stages of the ToDrinQ project: integration into the WP7 modular platform and supporting the WP8 exploitation and dissemination strategy.

8 Upscaling and European added value

8.1 Upscaling and replicability

The ten soft sensors developed in this work package (WP4) represent a significant and replicable advancement for water utilities. Their potential for upscaling is not limited to the three demonstration cases but extends to water systems across Europe, based on the following principles:

- **Methodological Replicability:** While a data-driven model trained for a particular case (e.g., Yliki Lake (DC#2)) cannot be directly applied to a different lake, the methodology used to build it is highly replicable. This deliverable provides a validated workflow (from data processing and model selection to imbalance handling and XAI validation) that utilities can adapt to their local data.
- **Modularity:** The sensors are not a single, monolithic systems. A utility concerned with agricultural runoff can replicate the Nutrient EWS, while another focused on process optimization can adopt the Ozonation (CT) sensor. This modularity lowers the barrier to adoption.
- **Use of Open EO Data:** Sensors developed for catchment monitoring (e.g., in Athens and Val de Bagnes) rely heavily on publicly available Earth Observation data, such as from the Copernicus Sentinel programme. This makes them highly scalable, as the primary data source is free, standardized, and provides pan-European coverage.
- **Interoperability:** The soft sensors are designed as software modules intended for integration with existing utility systems (like SCADA). Their integration into the ToDrinQ platform (under WP7), which is being built on open standards (FIWARE), ensures that they are interoperable and can be deployed in a standardized, future-proof "smart water" ecosystem.

8.2 European added value

The soft sensors identified and developed within the three tasks represent a significant step forward in the monitoring and management of drinking water systems. They bring innovative, data-driven solutions to challenges in water quality assessment at catchment scale, treatment optimization, and contamination prevention in pristine Alpine regions. By addressing critical parameters such as nutrient runoff, turbidity, and bacteriological contamination, the sensors enhance real-time decision-making capabilities, improving both operational efficiency and the reliability of water supply systems.

Beyond their immediate usability in the 3 demo sites of Amsterdam, Athens and Val de Bagnes, these soft sensors exhibit robust potential for replication and upscaling across diverse settings. Their modular design allows for **targeted deployment of specific functionalities**, such as turbidity prediction or nutrient load monitoring in lake water bodies, depending on the needs monitoring of the utilities and the linked DWTPs in the whole supply chain. The integration of earth observation data and advanced machine learning methods ensures adaptability to varied environmental and operational conditions. This versatility makes the soft sensors relevant not only to the case-specific contexts of the ToDrinQ project but also to a broad range of European and global water systems.

The compatibility of these tools with existing SCADA or other information dashboard/systems further enhances their feasibility for widespread adoption. **Minimal infrastructure changes are required, allowing utilities to seamlessly integrate the sensors into their operations**. Pilot implementations in Athens, Amsterdam and Val de Bagnes, are demonstrating the efficacy of these technologies in demanding settings. Moreover, by leveraging standardized frameworks like FIWARE, the sensors are well-positioned for deployment in future smart water treatment initiatives, ensuring interoperability and encouraging further innovation across the sector.

References

[1] D. W. Schindler, "The dilemma of controlling cultural eutrophication of lakes," *Proceedings of the Royal Society B: Biological Sciences*, vol. 279, no. 1746, pp. 4322–4333, 2012, doi: 10.1098/RSPB.2012.1032.

[2] D. W. Schindler and J. R. Vallentyne, "The Algal Bowl," *The Algal Bowl*, Dec. 2008, doi: 10.1515/9781772126341/HTML.

[3] Y. ; ; Wang *et al.*, "Water-Quality Assessment and Pollution-Risk Early-Warning System Based on Web Crawler Technology and LSTM," *International Journal of Environmental Research and Public Health 2022, Vol. 19, Page 11818*, vol. 19, no. 18, p. 11818, Sep. 2022, doi: 10.3390/IJERPH191811818.

[4] X. Ding, J. Zhang, G. Jiang, and S. Zhang, "Early Warning and Forecasting System of Water Quality Safety for Drinking Water Source Areas in Three Gorges Reservoir Area, China," *Water 2017, Vol. 9, Page 465*, vol. 9, no. 7, p. 465, Jun. 2017, doi: 10.3390/W9070465.

[5] I. Caballero, M. Roca, J. Santos-echeandía, P. Bernárdez, and G. Navarro, "Use of the Sentinel-2 and Landsat-8 Satellites for Water Quality Monitoring: An Early Warning Tool in the Mar Menor Coastal Lagoon," *Remote Sens (Basel)*, vol. 14, no. 12, p. 2744, Jun. 2022, doi: 10.3390/RS14122744/S1.

[6] Simons, G.W.H., R. Koster, and P. Droogers, "HiHydroSoil v2.0 - A high resolution soil map of global hydraulic properties.," 2020.

[7] L. M. De Sousa *et al.*, "SoilGrids 2.0: producing quality-assessed soil information for the globe", doi: 10.5194/soil-2020-65.

[8] L. S. Pereira, "Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56," 1998. [Online]. Available: https://www.researchgate.net/publication/235704197

[9] S. W. Running, Q. Mu, M. Zhao, and A. Moreno, "User's Guide MODIS Global Terrestrial Evapotranspiration (ET) Product (NASA MOD16A2/A3) NASA Earth Observing System MODIS Land Algorithm," 2017.

[10] M. Drusch *et al.*, "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services," *Remote Sens Environ*, vol. 120, pp. 25–36, May 2012, doi: 10.1016/J.RSE.2011.11.026.

[11] S. Running, Q. Mu, and M. Zhao, "MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V061 [Data set]," 2021.

[12] S. A. Clough *et al.*, "Atmospheric radiative transfer modeling: a summary of the AER codes," *J Quant Spectrosc Radiat Transf*, vol. 91, no. 2, pp. 233–244, Mar. 2005, doi: 10.1016/J.JQSRT.2004.05.058.

[13] "Hydrologic Modeling System HEC-HMS Technical Reference Manual CPD-74B," 2000.

[14] W. Weng and X. Zhu, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *IEEE Access*, vol. 9, pp. 16591–16603, May 2015, doi: 10.1109/ACCESS.2021.3053408.

[15] T. H. Bennett and J. C. Peters, "Continuous Soil Moisture Accounting in the Hydrologic Engineering Center Hydrologic Modeling System (HEC-HMS)," *Joint Conference on Water Resource Engineering and Water Resources Planning and Management 2000: Building Partnerships*, vol. 104, pp. 1–10, 2004, doi: 10.1061/40517(2000)149.

[16] C. Véliz-Chávez, C. A. Mastachi-Loza, E. González-Sosa, R. Becerril-Piña, and N. M. Ramos-Salinas, "Canopy Storage Implications on Interception Loss Modeling," *Am J Plant Sci*, vol. 05, no. 20, pp. 3032–3048, Sep. 2014, doi: 10.4236/AJPS.2014.520320.

[17] U. NRC Soil Science Division, "Soil Survey Manual 2017; Chapter 4".

[18] C. O. Clark, "Storage and the Unit Hydrograph," *Transactions of the American Society of Civil Engineers*, vol. 110, no. 1, pp. 1419–1446, Jan. 1945, doi: 10.1061/TACEAT.0005800.

[19] M. B. Kirkham, "Infiltration," *Principles of Soil and Plant Water Relations*, pp. 201–227, Jan. 2014, doi: 10.1016/B978-0-12-420022-7.00013-6.

[20] W. G. Buma and S. Il Lee, "Evaluation of Sentinel-2 and Landsat 8 Images for Estimating Chlorophyll-a Concentrations in Lake Chad, Africa," *Remote Sensing 2020, Vol. 12, Page 2437*, vol. 12, no. 15, p. 2437, Jul. 2020, doi: 10.3390/RS12152437.

[21] R. Johansen *et al.*, "Evaluating the portability of satellite derived chlorophyll-a algorithms for temperate inland lakes using airborne hyperspectral imagery and dense surface observations," *Harmful Algae*, vol. 76, pp. 35–46, Jun. 2018, doi: 10.1016/J.HAL.2018.05.001.

[22] I. Ogashawara *et al.*, "The Use of Sentinel-2 for Chlorophyll-a Spatial Dynamics Assessment: A Comparative Study on Different Lakes in Northern Germany," *Remote Sensing 2021, Vol. 13, Page 1542*, vol. 13, no. 8, p. 1542, Apr. 2021, doi: 10.3390/RS13081542.

[23] M. S. D'Silva, A. C. Anil, R. K. Naik, and P. M. D'Costa, "Algal blooms: A perspective from the coasts of India," *Natural Hazards*, vol. 63, no. 2, pp. 1225–1253, Sep. 2012, doi: 10.1007/S11069-012-0190-9/FIGURES/3.

[24] Y. Zhang *et al.*, "Meteorological and hydrological conditions driving the formation and disappearance of black blooms, an ecological disaster phenomena of eutrophication and algal blooms," *Science of The Total Environment*, vol. 569–570, pp. 1517–1529, Nov. 2016, doi: 10.1016/J.SCITOTENV.2016.06.244.

[25] R. Tian, J. Chen, X. Sun, D. Li, C. Liu, and H. Weng, "Algae explosive growth mechanism enabling weather-like forecast of harmful algal blooms," *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–7, Jul. 2018, doi: 10.1038/s41598-018-28104-7.

[26] L. Qi *et al.*, "In search of floating algae and other organisms in global oceans and lakes," *Remote Sens Environ*, vol. 239, p. 111659, Mar. 2020, doi: 10.1016/J.RSE.2020.111659.

[27] "MODIS Chlorophyll Fluorescence (MOD 20)", Accessed: Sep. 25, 2024. [Online]. Available: http://modis-ocean.gsfc.nasa.gov/

[28] J. Gower, C. Hu, G. Borstad, and S. King, "Ocean color satellites show extensive lines of floating sargassum in the gulf of Mexico," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 12, pp. 3619–3625, Dec. 2006, doi: 10.1109/TGRS.2006.882258.

[29] J. Gower and S. King, "Satellite Images Show the Movement of Floating Sargassum in the Gulf of Mexico and Atlantic Ocean," *Nature Precedings 2008*, pp. 1–1, May 2008, doi: 10.1038/npre.2008.1894.1.

[30] W. Shi and M. Wang, "Green macroalgae blooms in the Yellow Sea during the spring and summer of 2008," *J Geophys Res Oceans*, vol. 114, no. C12, Dec. 2009, doi: 10.1029/2009JC005513.

[31] L. Zheng *et al.*, "What causes the great green tide disaster in the South Yellow Sea of China in 2021?," *Ecol Indic*, vol. 140, p. 108988, Jul. 2022, doi: 10.1016/J.ECOLIND.2022.108988.

[32] C. Hu, “A novel ocean color index to detect floating algae in the global oceans,” *Remote Sens Environ*, vol. 113, no. 10, pp. 2118–2129, Oct. 2009, doi: 10.1016/J.RSE.2009.05.012.

[33] Y. Oyama, T. Fukushima, B. Matsushita, H. Matsuzaki, K. Kamiya, and H. Kobinata, “Monitoring levels of cyanobacterial blooms using the visual cyanobacteria index (VCI) and floating algae index (FAI),” *International Journal of Applied Earth Observation and Geoinformation*, vol. 38, pp. 335–348, Jun. 2015, doi: 10.1016/J.JAG.2015.02.002.

[34] M. Mu *et al.*, “Prediction of algal bloom occurrence based on the naive Bayesian model considering satellite image pixel differences,” *Ecol Indic*, vol. 124, p. 107416, May 2021, doi: 10.1016/J.ECOLIND.2021.107416.

[35] R. Tripathi, R. Narayan Sahoo, V. Kumar Sehgal, and P. Misra Sahoo, “Developing Vegetation Health Index from biophysical variables derived using MODIS satellite data in the Trans-Gangetic plains of India,” 2014, doi: 10.9755/ejfa.v25i5.11580.

[36] A. P. Kirana, R. Ariyanto, A. R. T. H. Ririd, and E. L. Amalia, “Agricultural drought monitoring based on vegetation health index in East Java Indonesia using MODIS Satellite Data,” *IOP Conf Ser Mater Sci Eng*, vol. 732, no. 1, p. 012063, Jan. 2020, doi: 10.1088/1757-899X/732/1/012063.

[37] S. Kloos, Y. Yuan, M. Castelli, and A. Menzel, “Agricultural drought detection with modis based vegetation health indices in southeast germany,” *Remote Sens (Basel)*, vol. 13, no. 19, p. 3907, Oct. 2021, doi: 10.3390/RS13193907/S1.

[38] M. Usman, R. Liedl, M. A. Shahid, and A. Abbas, “Land use/land cover classification and its change detection using multi-temporal MODIS NDVI data,” *Journal of Geographical Sciences*, vol. 25, no. 12, pp. 1479–1506, Dec. 2015, doi: 10.1007/S11442-015-1247-Y/METRICS.

[39] R. S. Lunetta, J. F. Knight, J. Ediriwickrema, J. G. Lyon, and L. D. Worthy, “Land-Cover Change Detection Using Multi-Temporal MODIS NDVI Data,” *Geospatial Information Handbook for Water Resources and Watershed Management*, pp. 65–88, Dec. 2022, doi: 10.1201/9781003175025-5.

[40] H. Yin, D. Pflugmacher, R. E. Kennedy, D. Sulla-Menashe, and P. Hostert, “Mapping annual land use and land cover changes using MODIS time series,” *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 7, no. 8, pp. 3421–3427, Aug. 2014, doi: 10.1109/JSTARS.2014.2348411.

[41] X. Zhan *et al.*, “Detection of land cover changes using MODIS 250 m data,” *Remote Sens Environ*, vol. 83, no. 1–2, pp. 336–350, Nov. 2002, doi: 10.1016/S0034-4257(02)00081-0.

[42] A. Khandelwal, A. Karpatne, M. E. Marlier, J. Kim, D. P. Lettenmaier, and V. Kumar, “An approach for global monitoring of surface water extent variations in reservoirs using MODIS data,” *Remote Sens Environ*, vol. 202, pp. 113–128, Dec. 2017, doi: 10.1016/J.RSE.2017.05.039.

[43] F. Ling *et al.*, “Monitoring surface water area variations of reservoirs using daily MODIS images by exploring sub-pixel information,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 168, pp. 141–152, Oct. 2020, doi: 10.1016/J.ISPRSJPRS.2020.08.008.

[44] G. Ovakoglou, T. K. Alexandridis, T. L. Crisman, C. Skoulikaris, and G. S. Vergos, “Use of MODIS satellite images for detailed lake morphometry: Application to basins with large water level fluctuations,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 51, pp. 37–46, Sep. 2016, doi: 10.1016/J.JAG.2016.04.007.

[45] J. K. Balch *et al.*, “FIRED (Fire Events Delineation): An Open, Flexible Algorithm and Database of US Fire Events Derived from the MODIS Burned Area Product (2001–2019),” *Remote Sensing 2020, Vol. 12, Page 3498*, vol. 12, no. 21, p. 3498, Oct. 2020, doi: 10.3390/RS12213498.

[46] Y. J. Kaufman *et al.*, "Kaufman et al., Remote Sensing of Fires from EOS-MODIS Monitoring Global Fires from EOS-MODIS".

[47] S. M. Raffuse *et al.*, "High-resolution MODIS aerosol retrieval during wildfire events in California for use in exposure assessment," *Journal of Geophysical Research: Atmospheres*, vol. 118, no. 19, pp. 11,242-11,255, Oct. 2013, doi: 10.1002/JGRD.50862.

[48] Y. J. Kaufman *et al.*, "Potential global fire monitoring from EOS-MODIS," *Journal of Geophysical Research: Atmospheres*, vol. 103, no. D24, pp. 32215–32238, Dec. 1998, doi: 10.1029/98JD01644.

[49] Sheng, D. P. W., Bilad, M. R. & Shamsuddin, N. (2023). Assessment and optimization of coagulation process in water treatment plant: a review. Asean journal of science and engineering 3, 79–100

50] A review of methods and instruments to monitor turbidity and suspended sediment concentration. Journal of water process engineering 64, 105624 (2024)

[51] Ratnaweera, H. & Fettig, J. (2015).State of the art of online monitoring and control of the coagulation process. Water 7, 6574–6597

[52] Li, L., Rong, S., Wang, R. & Yu, S. (2021).Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: a review. Chemical engineering journal 405, 126673

[53] Ortiz-Lopez, C., Torres, A., Bouchard, C., & Rodriguez, M. (2023). *A methodology for integrating time-lagged rainfall and river flow data into machine learning models to improve prediction of quality parameters of raw water supplying a* treatment *plant*. *Journal of Hydroinformatics, 25*(6), 2406-2426. https://doi.org/10.2166/hydro.2023.122

[54] Lu, H., & Ma, X. (2020). *Hybrid decision tree-based machine learning models for short-term water quality prediction*. *Chemosphere, 249*, 126169. https://doi.org/10.1016/j.chemosphere.2020.126169

[55] Gleeson, K., Husband, S., Gaffney, J. & Boxall, J. (2023).A data quality assessment framework for drinking water distribution system water quality time series datasets. Aqua—water infrastructure, ecosystems and society 72, 329–347

[56] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*(1), 59–69. https://doi.org/10.1007/BF00337288

[57] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

[58] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[59] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[60] Ascott, M., Lapworth, D., Gooddy, D., Sage, R. & Karapanos, I. (2016). Impacts of extreme flooding on riverbank filtration water quality. Science of the total environment 554, 89–101

[61] Li, Y., Xu, Z., Zhan, X. & Zhang, T. (2024). Summary of experiments and influencing factors of sediment settling velocity in still water. Water 16, 938

[62] USEPA 1989 Guidance Manual for Compliance with the Filtration and Disinfection Requirements for Public Water Systems Using Surface Water Supplies. USEPA, Washington, DC

[63] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https:// doi.org/ 10.1023/ A:1010933404324

[64] Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[65] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

[66] Freund, Y., & Schapire, R. E. (1997). *A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

[67] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T.-Y. (2017). *LightGBM: A highly efficient gradient boosting decision tree.* In *Advances in Neural Information Processing Systems* (Vol. 30).

[68] Karniadakis, G.E., Kevrekidis, I.G., Lu, L. *et al.* Physics-informed machine learning. *Nat Rev Phys* **3**, 422–440 (2021). https://doi.org/10.1038/s42254-021-00314-5

[69] European Parliament & Council. (2020). *Directive (EU) 2020/2184 on the quality of water intended for human consumption* (recast) (OJ L 435, pp. 1–62). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32020L2184 EUR-Lex+1

[70] van der Helm, W. C., Rietveld, L. C., Baars, E. T., Smeets, P. W. M. H., van Dijk, J. C.; Modeling disinfection and by-product formation during the initial and the second phase of natural water ozonation in a pilot-scale plug flow reactor. *Journal of Water Supply: Research and Technology-Aqua* 1 September 2008; 57 (6): 435–449. doi: https://doi.org/10.2166/aqua.2008.089