Soft-sensors module for water quality monitoring and performance improvement: beta version

Deliverable 4.1

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 D4.1 Soft-sensors module for water quality monitoring and performance improvement: beta version		
ACTUAL DATE OF DELIVERY	29-11-2024		
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 Jasper Imnik, KWR George Bariamis, NTUA Christos Makropoulos, NTUA Luuk Rietveld, TUD		

Document history

Version	Description (Section, page number)	Author	Organisation short name
V0.1	First draft	Vasiliki Thomopoulou, NTUA Panagiotis Kossieris, NTUA Greg Kyritsakas, TUD Jasper Imnik, KWR George Bariamis, NTUA	National Technical University of Athens
V0.2	Review	Jean-François Gérard, CEB	National Technical University of Athens
V0.4	Final draft	Vasiliki Thomopoulou, NTUA Panagiotis Kossieris, NTUA Greg Kyritsakas, TUD Jasper Imnik, KWR George Bariamis, NTUA	National Technical University of Athens
V0.5	Review	Christos Makropoulos, NTUA	National Technical University of Athens
V1.0	Final version	Luuk Rietveld, TUD	Delft University of Technology

Quality control

Author	Organization short name	Role	Date
Christos Makropoulos	National Technical University of Athens	Deliverable Leader	08-11-2024
Christos Makropoulos	National Technical University of Athens	Work Package Leader	08-11-2024
Jean-François Gérard	CEBEDEAU	Reviewer 1	13-11-2024
Luuk Rietveld	Delft University of Technology	Scientific project coordinator	29-11-2024
Danitsja van Heusden	Delft University of Technology	Project coordinator	29-11-2024

Abbreviations

ANN		Artificial Neural Network
AOC		Assimilable Organic Carbon
ASPP SAM -		Atrous Spatial Pyramid Pooling (ASPP)-Spatial Attention Mechanism (SAM)- Unet
AWD		Amsterdam Supply Dunes
BNN	-	Bayesian Neural Networks
CA	-	Consortium Agreement
CNB	-	Categorical Naïve Bayes
CT		Disinfectant (Ozone) Exposure
CV	-	Computer Vision
DC	-	Demo Case
DCL	-	Demo Case Leader
DoA	-	Description of the Action
DWTP	-	Drinking Water Treatment Plant
EAB	-	External Advisory Board
EC	-	European Commission
EO	-	Earth Observation
ET	-	Evapotranspiration
EVI	-	Enhanced Vegetation Index
EU	-	European Union
EWS	-	Early warning system
FAI	-	Floating Algal index
FAIR	-	Facebook AI Research team
GA	-	General Assembly
GNDVI	-	Green Normalized Difference Vegetation Index
HEC-HMS		Hydrologic Engineering Center, Hydrologic Modeling System
KWR	-	KWR WATER BV
LAI	-	Leaf area index
LE	-	Latent Heat flux
LSTM	-	Long Short-Term memory
MCI	-	Maximum Chlorophyl Index
ML	-	Machine Learning

MLP	-	Multi layer Perceptron
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
NTUA	-	National Technical University of Athens
PC	-	Project Coordinator
PET	-	Potential Evapotranspiration
PINN		Physics Informed Neural Network
PLE	-	Potential latent heat flux
RECI	-	Red-Edge Chlorophyll Index
RF	-	Random Forest classifier
SC	-	Steering Committee
SCADA	-	Supervisory Control and Data Acquisition
SITS	-	Satellite Images Time Series
SMA	-	Soil Moisture Accounting
SOMs		Self-organizing Maps
SWIR	-	Short-wavelength infrared spectrum
TUD	-	Delft University of Technology
VCI	-	Visual Cyanobacteria Index
WP	-	Work Package
WPL	-	Work Package Leader
WQI	-	Water Quality Index

Table of contents

List of figures

List of tables

Executive summary

ToDrinQ project intends to support the implementation of the EU's revised Drinking Water Directive (DWD) through the development and testing of a modular ToDrinQ Toolkit. This toolkit will integrate real-time sensing and water quality monitoring technologies, modular treatment processes, and decision-support tools to enhance resilience, risk assessment, and risk management in drinking water systems.

The recast Drinking Water Directive (effective January 2021) introduced a risk-based framework covering the entire water supply chain—from source to tap—to ensure high drinking water standards and protect public health. ToDrinQ toolkit aligns with these priorities by leveraging **innovative technologies and collaboration** and its approach combines **real-time and automated monitoring with advanced analytical tools** to detect and measure chemical, biological, and other contaminants across diverse sources, including reservoirs, rivers, aquifers, and distribution systems. These tools provide near real-time data and **early warnings**, enabling operators to respond swiftly to contamination events, maintain public confidence, and support compliance with water quality standards. Through a co-creative process (WP2) with partner water utilities (Waternet, EYDAP, Veolia France, Veolia Czechia and ALTIS) and a strong consortium of research institutions and technology developers, ToDrinQ drives significant progress beyond the state-of-the-art while addressing the practical needs of the water sector in the face of emerging challenges.

The Deliverable D4.1 reports progress on Tasks 4.1, 4.2 and 4.3 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. The findings provided a comprehensive framework for designing and developing soft sensors (Task 4.1), integrating in-situ, online/offline and open access earth observation data (Task 4.2) to enhance water quality monitoring and treatment across these three demo cases.

In total **we identified eight (8) soft sensor use cases** (at catchment and 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 development of soft sensors has made significant progress across various applications, demonstrating both successes and areas for improvement as planned. The early warning system for nutrient runoff successfully integrates crop classification models with hydrological tools, offering stakeholders actionable insights into nutrient load locations and timing. Similarly, the chlorophyll-a concentration model shows **promising predictive capabilities**, although its robustness needs further testing in diverse environments. The bloom occurrence probability sensor effectively combines meteorological forecasts and basin-level data but requires advanced modeling techniques, like Bayesian Neural Networks, to enhance its accuracy. Other models, such as the Water Quality Index (WQI) and bacteriological contamination early warning systems, are progressing toward practical implementation. The WQI sensor holds promise for high-resolution water quality monitoring at the pixel level, while the bacteriological contamination model provides timely notifications by linking pollutant sources to groundwater sources to rainfall/snow patterns in Alpine regions. The turbidity prediction

model for coagulation-flocculation processes has achieved better performance than traditional approaches by combining physics-informed and deep learning methods. Similarly, the turbidity prediction sensor for drinking water treatment plant (DWTP) intake shows reliable results but requires additional data, ensemble modeling, and incorporation of rainfall impacts to optimize performance. Lastly, the ozone exposure prediction sensor demonstrates strong outcomes using Random Forest models, with plans to refine Physics-Informed Neural Networks and expand its scope to predict bromate and assimilable organic carbon (AOC).

Future steps towards the finalization of the tasks for all soft sensors include enhancing models robustness and explainability based on xAI principles, linking with hard sensors data inputs (WP3) and testing in broader scenarios. These developments when integrated in the ToDrinQ platform (WP7) aim to improve predictive accuracy, operational efficiency, and stakeholder trust, paving the way for these technologies to become vital tools in modern water quality monitoring, management and regulatory compliance.

1. Introduction

The EU's recast Drinking Water Directive (DWD) 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 DWD, 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[2].

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, water treatment facilities, and water distribution systems). The primary goal is **to detect and measure**

[1] DIRECTIVE (EU) 2020/2184 on the quality of water intended for human consumption, recast - link
[2] Why monitor water quality? - link

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.

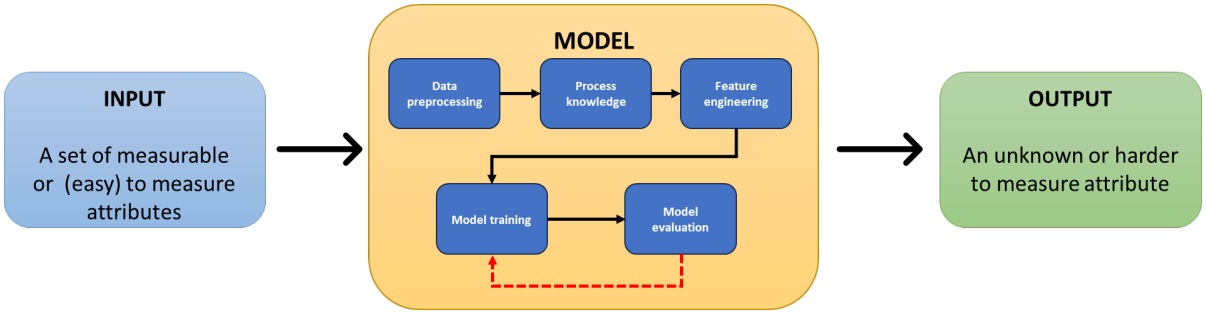


Figure 1: Schematic representation of soft sensor

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.

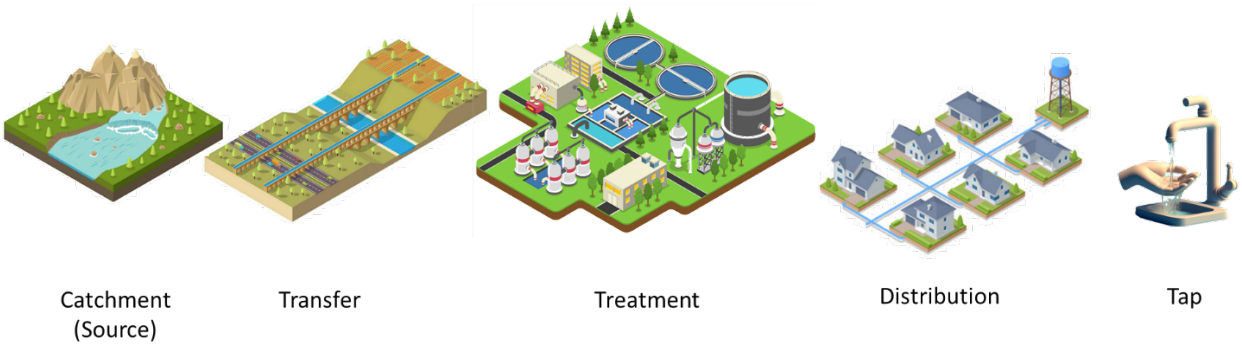


Figure 2: Drinking water supply chain from catchment to tap user

1.2 Soft sensors at source/catchment level

Catchments play a critical role in modulating 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 relatively 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 storage (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 (transfer) 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 DWTP 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.

1.5 Soft sensors at water distribution level

Soft sensors at the water distribution network level can play an important role in ensuring the safe delivery of potable water to end-users. In particular, the development of soft sensors for predicting potential leakages and pinpointing their locations is valuable (saving both water and energy). In addition,

the utilization of hard sensors installed in the distribution network further enhances the capability to estimate water quality parameters that may not be directly measured, such as turbidity (Meyers et al., 2016). A common soft sensor for recontamination and/regrowth in the distribution network is the (consumption of) residual chlorine (when applied). Further, the average concentration of AOC after treatment can give an indication for regrowth potential in distribution networks in not chlorinated systems and the average saturation index after treatment of the corrosion/scaling potential in the distribution system. By leveraging these technological advancements, we not only optimize the efficiency of the distribution process but also contribute to the overarching goal of providing consistently safe and high-quality water to the end-users.

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 involves stakeholder engagement and determining sensor requirements (accuracy, response time, operational conditions etc) to ensure reliable field performance.

2.1 Review on soft-sensors applications

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 monitoring and treatment improvements. with each partner contributing specialized expertise. The National Technical University of Athens (NTUA) concentrated on assessing water quality at both catchment and conveyance levels, leveraging advanced remote sensing techniques for enhanced monitoring and analysis. The Technical University of Dresden (TUD) prioritized the optimization of drinking water treatment processes, aiming to improve efficiency and reliability. Meanwhile, KWR Water Research Institute explored hybrid modelling methodologies to enhance treatment performance, combining traditional and data-driven techniques for more robust and adaptive solutions. The outcome was a comprehensive catalogue of soft-sensor applications that organized various water quality target variables with predictor variables. This cataloguing considered various spatial and temporal resolutions, aiding the integration of remote sensing technologies with in-situ to provide a continuous, high-resolution monitoring system throughout the drinking water supply chain. Figures 3 and 4 illustrate the distribution and application of target variables for soft sensors. Figure 3 provides an indicative representation of the distribution of target variables when soft sensors rely solely on in-situ data inputs. In contrast, Figure 4 highlights the target variables in use cases where soft sensors integrate in-situ data with Earth Observation (EO) data. Notably, chlorophyll-a, turbidity, and suspended sediments (SS) emerge as the primary target variables, as they are optically active and particularly suited for such hybrid approaches.

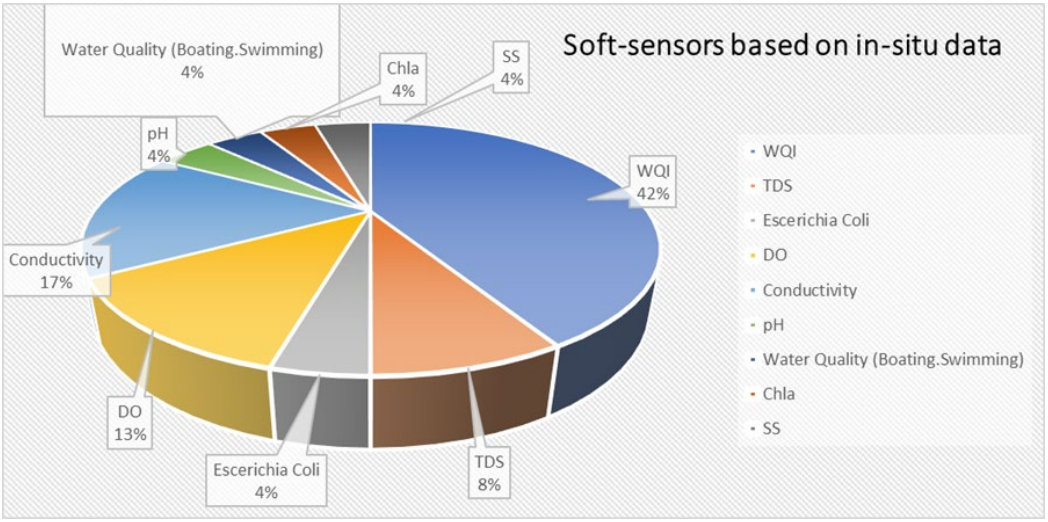


Figure 3: Review of Soft Sensor Use Cases Mapped Based on In-Situ Data Inputs

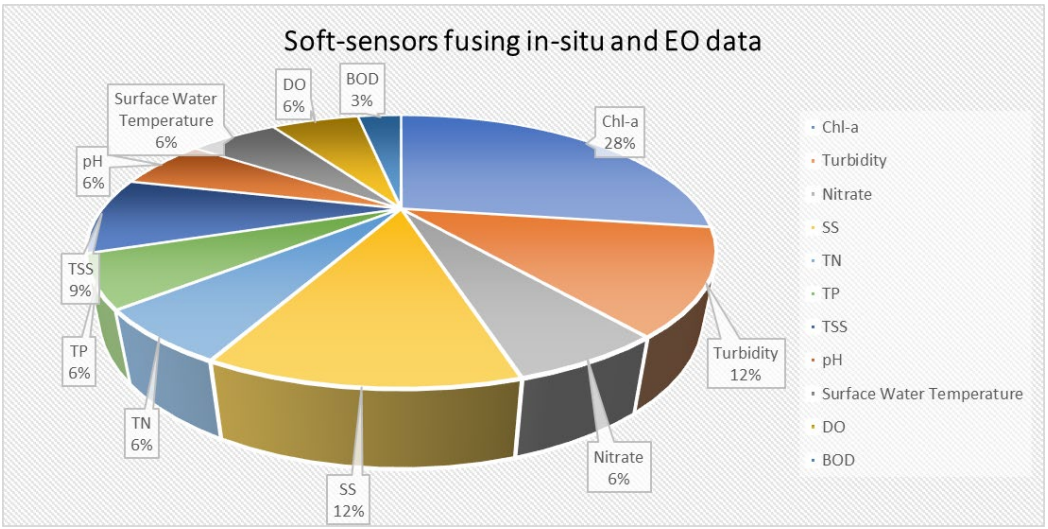


Figure 4: Review of Soft Sensor Use Cases Mapped Based on In-Situ and earth observation (EO) Data Inputs

Overview of soft sensor applications identified

The identified soft sensor use cases are listed in Table 1 below, organized by their location within the water supply chain

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 concentration 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, and remote sensing indices 2. In-situ Chl-a measurements from ISA boat (EYDAP)	3-5 days	10m
3	EYDAP	Bloom Occurrence Probability Estimation Floating Algal Index (FAI)	Estimation of the Floating Algal bloom occurrence probability at pixel level	FAI	1. Prior bloom occurrence probability at pixel level MODIS FAI timeseries from 2012 to 2022) 2. Normalized Nutrient Load at the upstream basin 3. Meteorological variable from NOAA GFS (forecast) 4. Discharge data (from the hydrological model developed in the early warning system)	Daily	Point (Discharge) & 250m (MODIS) & 27830m (Meteorological variables)
4	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)	1.Sentinel-2 reflectance data, and remote sensing indices 2. In-situ Chl-a, turbidity, pH, and DO measurements from ISA boat (EYDAP)	3-5 days	10m
5	ALTIS	Early warning System of Bacteriological	Provide an early warning for bacteriological	Bacteriological contaminants	LULC maps, pasture area, and pasture patterns, discharge data, EO precipitation, Bacteriological load in the basin	Daily	To be investigated

Soft Sensor No	DC	Soft Sensor Title	Aim (short description)	Hard-to-measure' compound	Variables	Time scale	Spatial Scale
		contamination in Sarayer	contamination in Sarayer				
6	WTNT	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
7	WTNT	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
8	WTNT	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 (relates to T4.2)

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 (hard sensors)

A hard sensor 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 drinking water treatment plant (DWTP), at the distribution network and finally at the tap level.

At the catchment level, hard sensors 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, hard 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.

Hard 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 hard 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. Hard 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 data source that can be employed for the development of soft sensors is 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[3]). 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 technologies (passive/active instruments)

The available technologies 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

[3] Group on Earth Observations (GEO), GEO at a Glance, last accessed: 15, November 2023- 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/Demo case description

The basin on the upstream of Yliki Lake, the Boeotikos Kifissos river basin, is characterized by intense agricultural activity, and extensive 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 (Schindler, 2012). 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 (Schindler and Vallentyne, 2008).

Figure 5: Athens demo case area

The study area of Boeotikos Kifissos river basin 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 in 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: Cover classes with the highest estimated probability for the years 2016–2023, based on the Dynamic World V1 Dataset classification in 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 for nutrient runoff in the Yliki Lake

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. Guang et al. (Wang et al., 2022), 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 water quality related risk assessment. Another study that took place in the Three Gorges Reservoir (Ding et al., 2017), 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 (Caballero et al., 2022), 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.

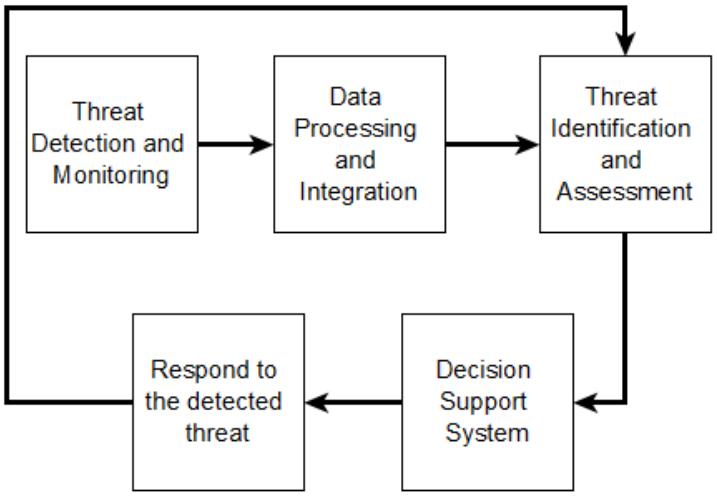


Figure 6: Soft sensor constituents

- 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

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 modeling 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 10 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 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 was 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) —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.

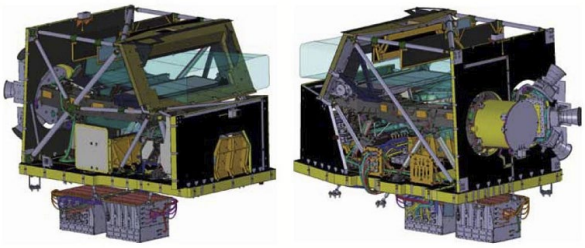


Figure 7: The MSI on-board the Sentinel-2 mission

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 SWIR band also have a 60-meter resolution, whereas the other two SWIR bands have a resolution of 20 meters.

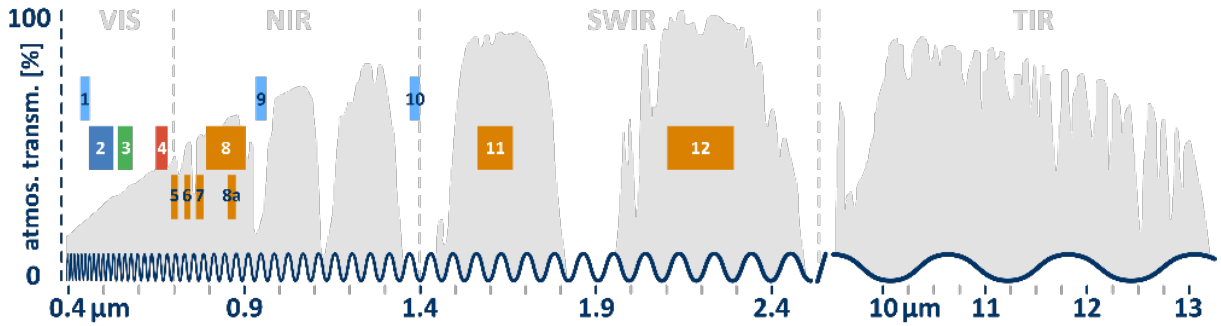


Figure 8: The spectral bands of Sentinel-2 satellite

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 2091 to September 31st 2021. The simulations are at daily time-step and the precipitation was calculated based on the Thiesen method, per subbasin.

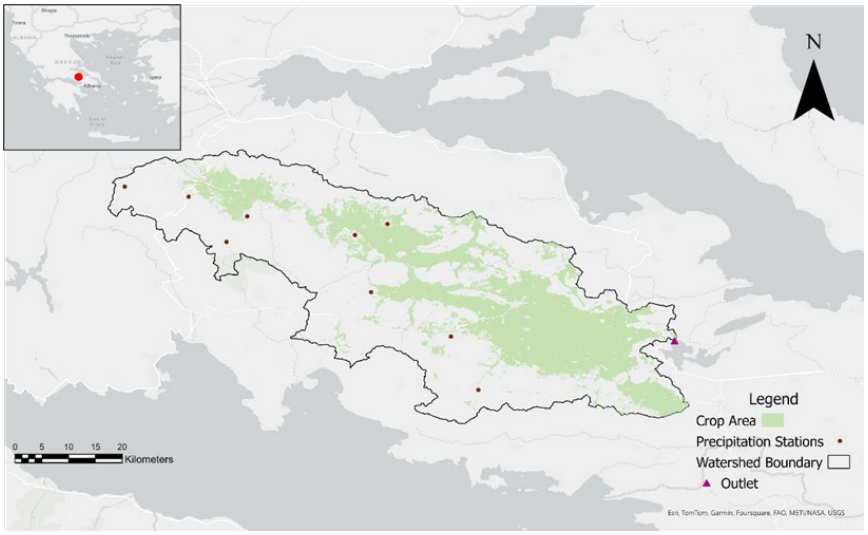


Figure 9: Map of the Kifissos basin area upstream of Yliki Lake

Earth Observation Datasets

EO datasets are employed to support the calibration of HEC-HMS (Feldman, 2016), and particularly the Soil Moisture Accounting (SMA) (Bennett and Peters, 2004a) loss method, as well as to drive the model with gridded meteorological data input. Particularly, HiHydrosoils dataset (Simons et al., 2020) are used in the parameter mapping of SMA method, by providing data on soil properties. NOAA GFS precipitation (Clough et al., 2005) 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 (Simons et al., 2020) 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 (Running et al., 2017), which is an 8-day composite dataset, with a spatial resolution of 500 m. The estimations are based on the Penman-Monteith method (Pereira, 1998) 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 (Running et al., 2017).

- The Global Forecast System (GFS) (Clough et al., 2005) 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.

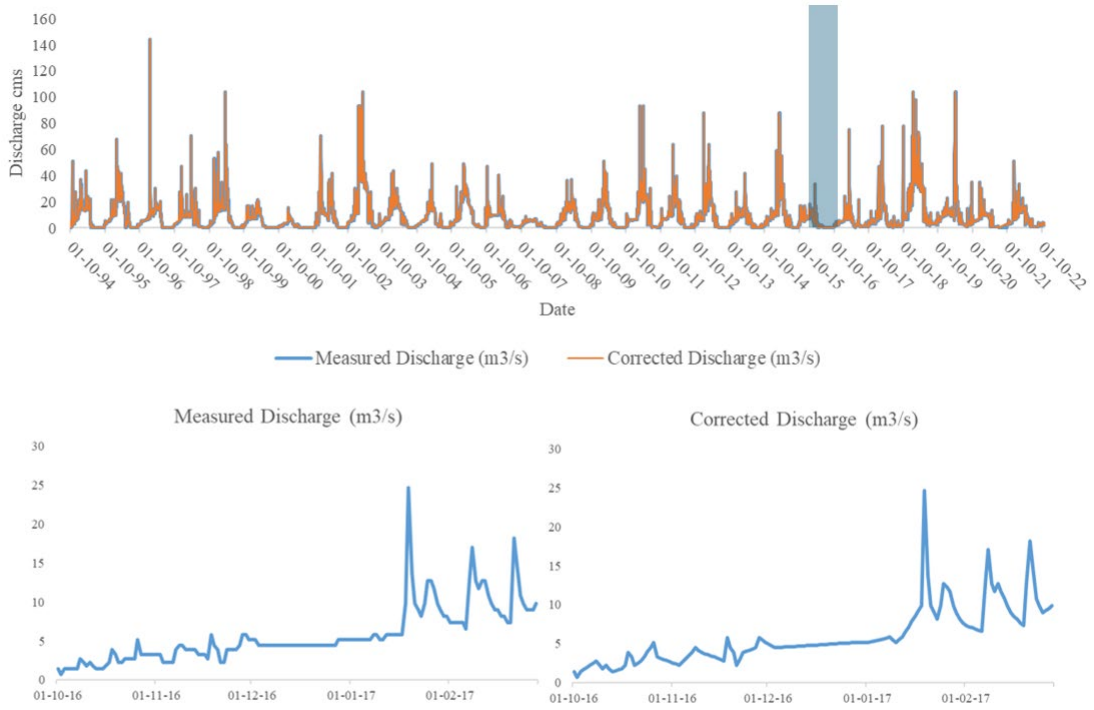


Figure 10: Discharges at the Kifissos basin outlet

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 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 instances selected for the analysis are provided at 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

Month	Date	Cloud Cover (%)
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.

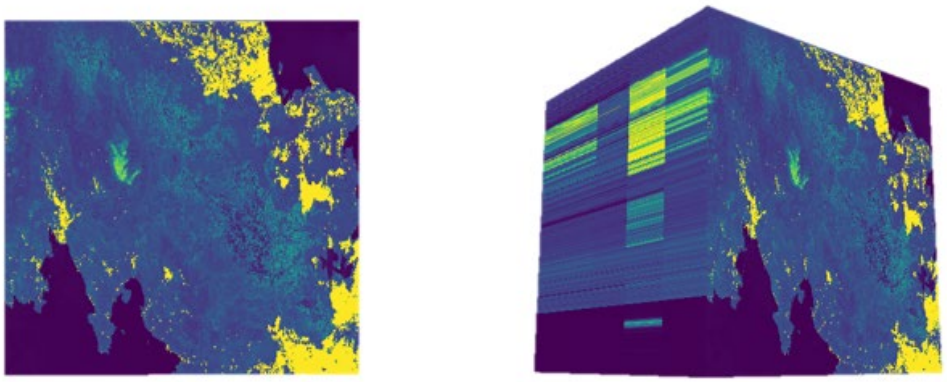


Figure 11: The Sentinel-2 data cube

4.2.3 Material 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.

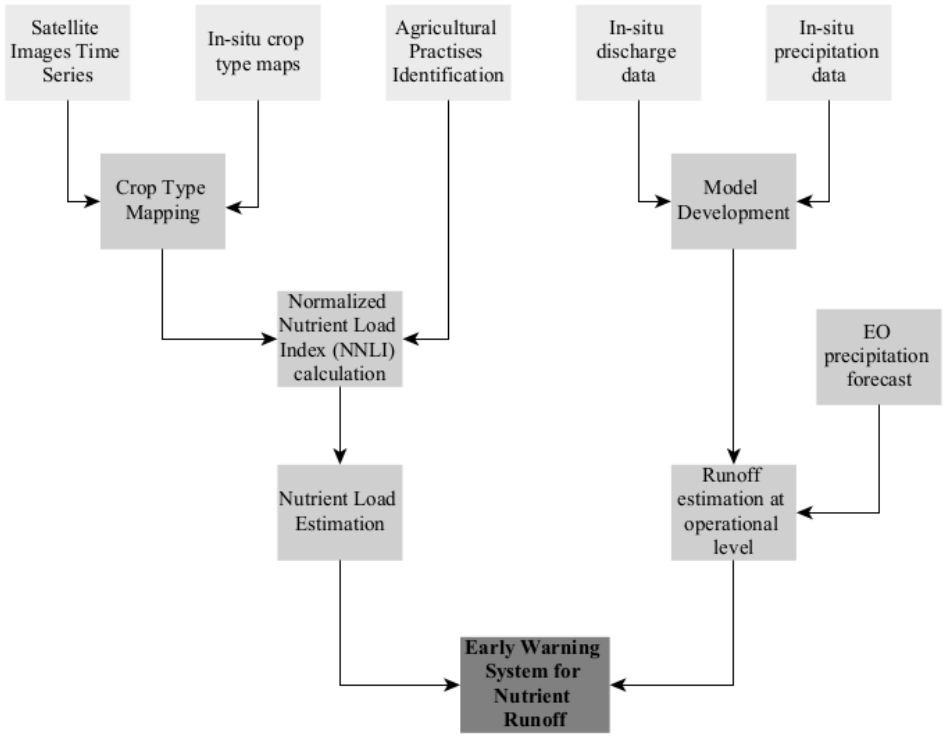


Figure 12: Flowchart of the early warning system

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 ("Hydrologic Modeling System HEC-HMS Technical Reference Manual CPD-74B," 2000).

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:

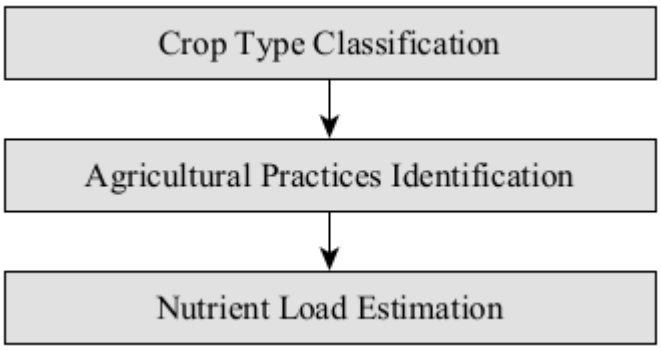


Figure 13: Components of soft sensor on nutrients load

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 (Weng and Zhu, 2015) . 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.

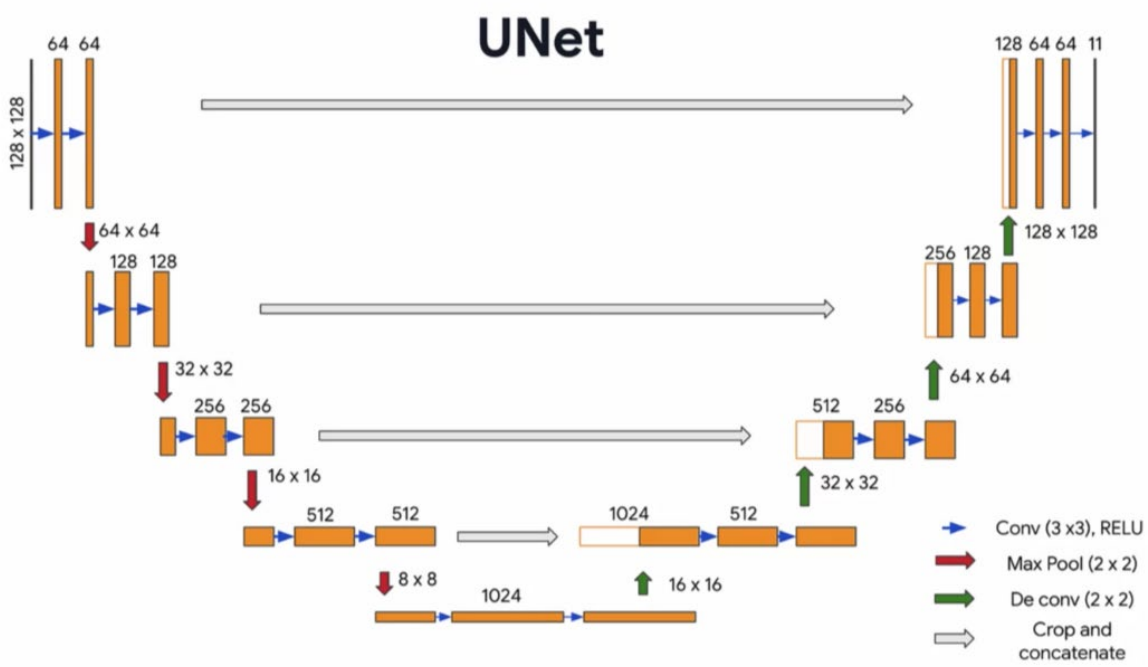


Figure 14: The Unet architecture

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

Within the baseline model, the neural network maintains an uncompressed time dimension within both its encoder and decoder stages. At the bottleneck, the time dimension is not compressed, while the other dimensions of the data cube are reduced. Thus, the second architecture provides the compression of the time dimension at the encoder phase to inspect the importance of information extraction in 4D arrays. Compressing the time dimension in an encoder-decoder architecture such as a U-Net allows several actions to be taken. It first forces the network to shrink and focus only on the most critical features along the temporal axis, hence allowing effective feature extraction. The selective focus, in turn, enhances the generalization capability of the network in recognizing important patterns over time. Secondly, reducing the time dimension helps in dimensionality reduction that decreases the computational complexity of the network. Thus, since the network needs to process fewer time steps, training and inference times become much faster. In models that include a bottleneck, as in autoencoders or U-Nets, there is efficient data representation by compressing the time dimension down to lower dimensions. This allows the network to learn representations in a way that is more compact and informative; this may be highly useful for some anomaly detection tasks or data compression tasks. It compresses the time dimension so that such a network will have improved information flow by focusing on the essential temporal relationships while discarding irrelevant information.

Baseline Model with Attention at the skip connection

In the third model tested, attention modules are applied to the skip connection of the basic Unet architecture of the baseline model to enhance the network's ability to focus on relevant features. In the basic Unet, not all features extracted at different levels contribute equally to the final segmentation output. By applying the attention mechanism, important spatial regions and channel-wise features are emphasized and object boundaries are expected to be more accurately delineated. Additionally, the attentions modules are expected to better capture the long range dependencies of the input data, and thus maintain consistency across different levels of the feature hierarchy. The summary of the model architecture is provided in Table 6.

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 7) 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.

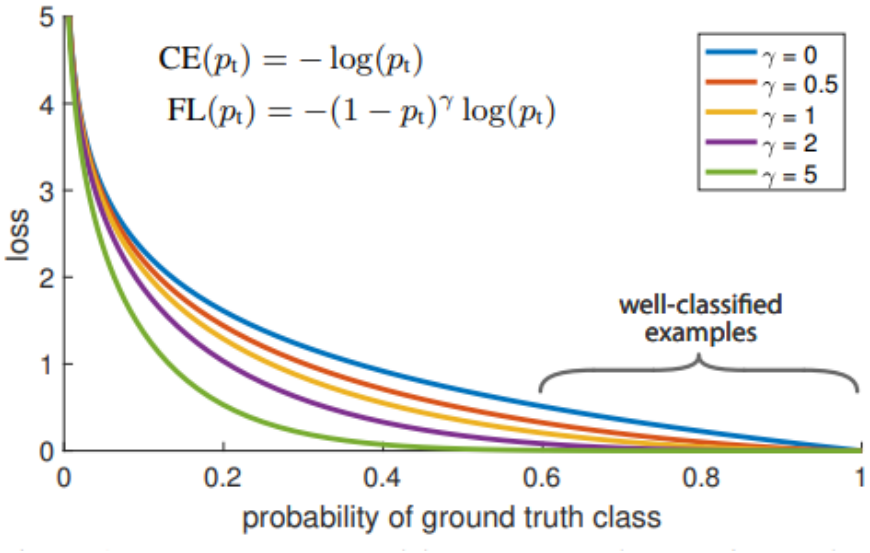


Figure 15: Focal Loss and Cross Entropy functions

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

$$Focal\ Loss = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C}(1 - y''_{n,i})^{\gamma}\, y_{n,i}\left(y'_{n,i}\right) \quad \text{Eq. 2}$$

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.

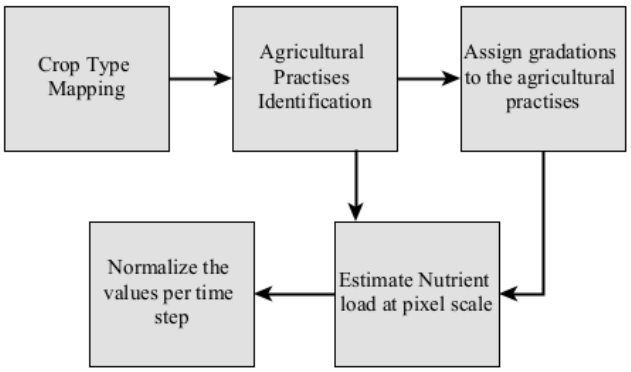


Figure 16: Flowchart for the estimation of the proposed NNLI

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.

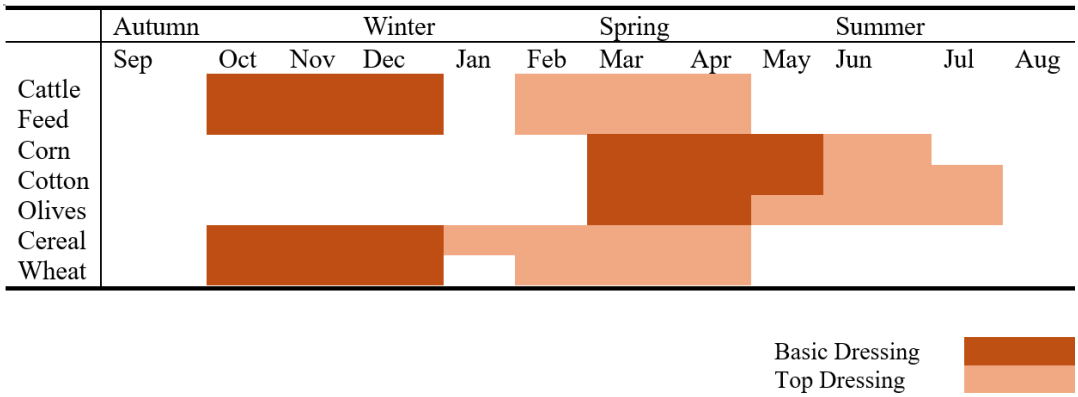
For each time step, the nutrient load in the study area is calculated as the weighted sum of the nutrient-loaded pixels (*step 4*). After estimating the nutrient load for each time step, a normalization equation is applied (*step 5*). The NNLI, calculated for each time step, provides an overview of the 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 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 (Bennett and Peters, 2004b).

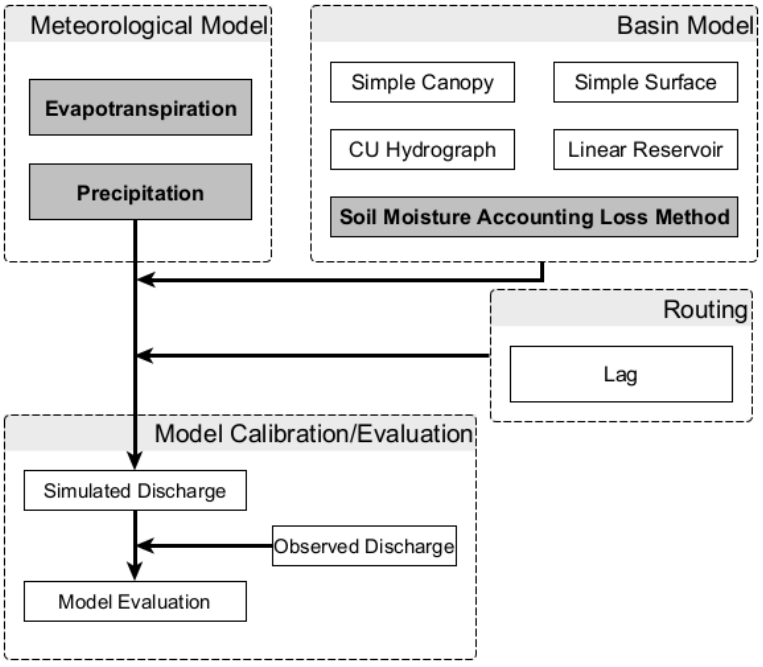


Figure 17: Flowchart of model description

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 (Véliz-Chávez et al., 2014), 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 (Bennett and Peters, 2004b) 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 (Clark, 1945). 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 (Kirkham, 2014). 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 storage were made in proportion to their soil storage. For the reasons mentioned above, these parameters underwent further refinement during the calibration procedure.

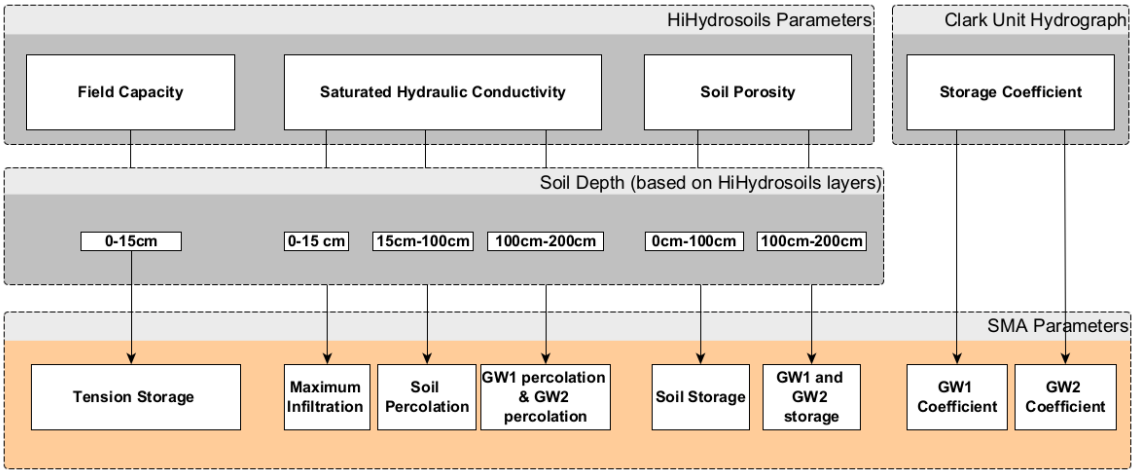


Figure 18: Parameters involved in the hydrological model

4.2.4 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.
- p_e is the expected agreement by chance, calculated as the sum of the product of the marginal probabilities for each class:

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

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

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).

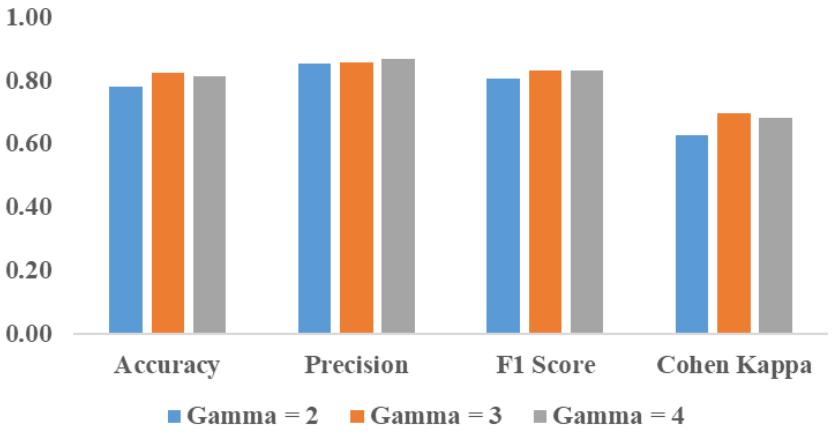


Figure 19: Baseline U-Net results

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).

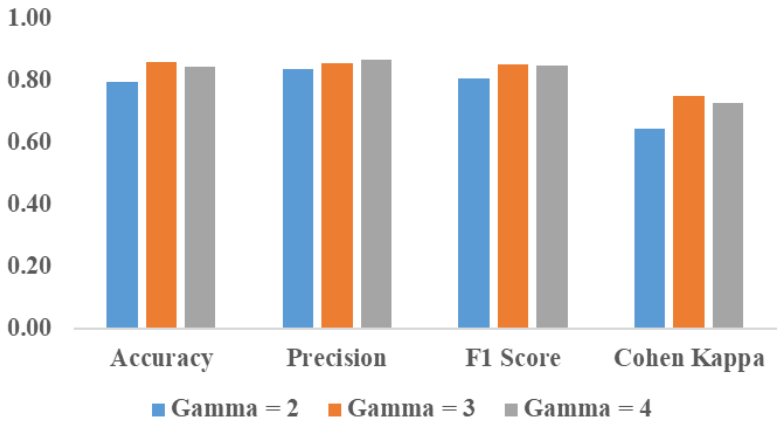


Figure 20: Time Compression U-Net Results

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).

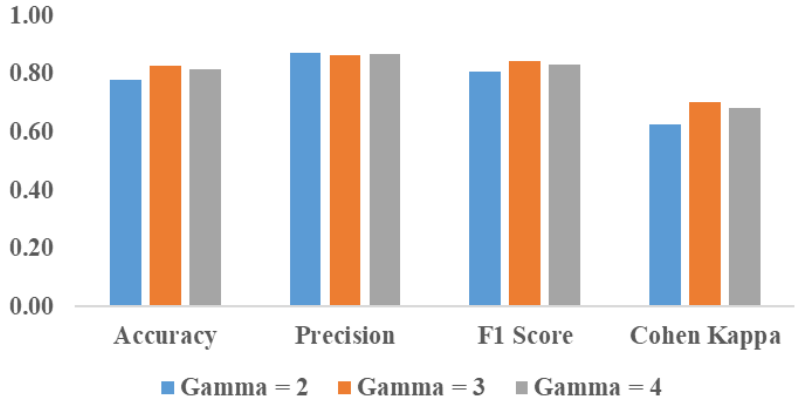


Figure 21: U-Net with attention at the skip-connections results

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.

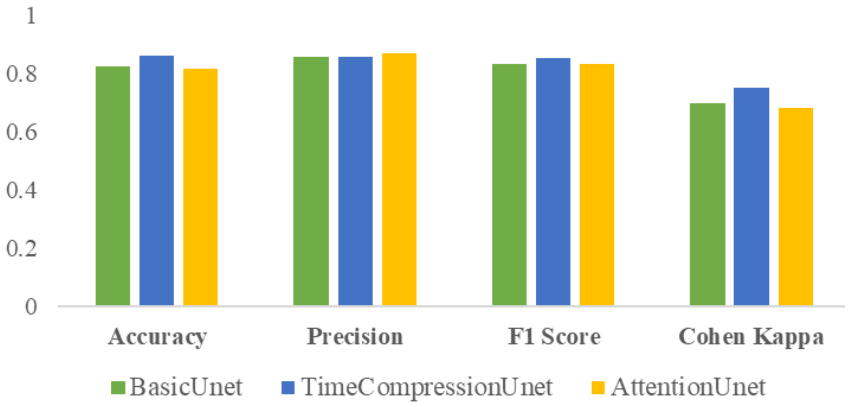


Figure 22: Comparison of the best performing models of each tested architecture

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 23).

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.

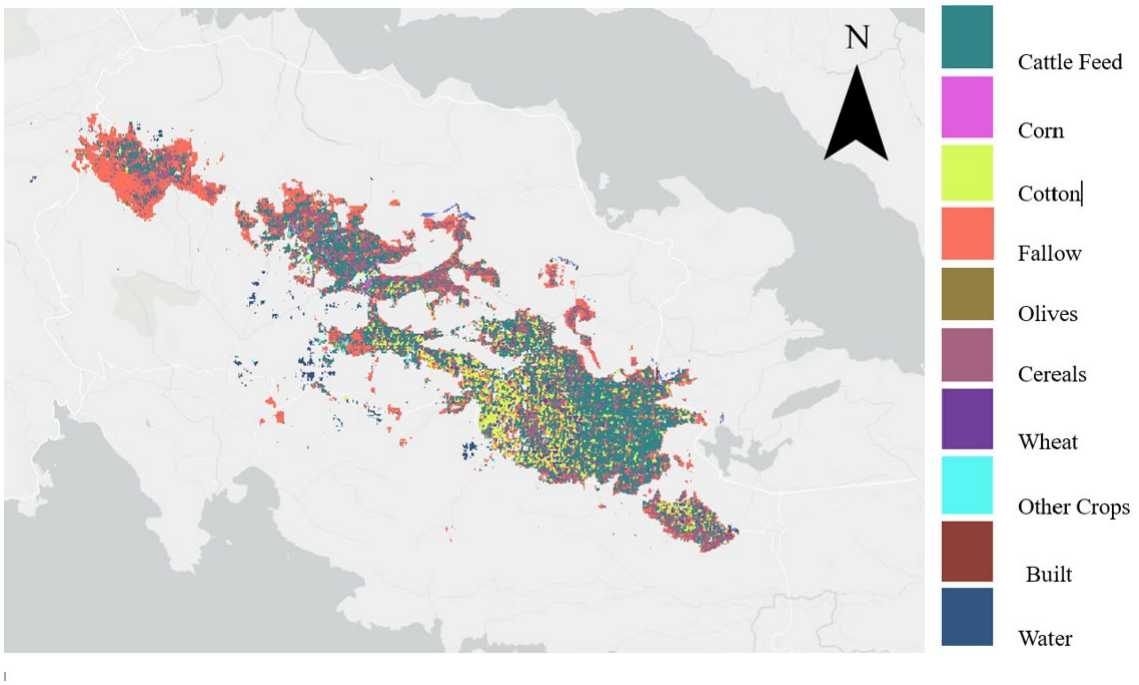


Figure 23: Segmented map of the crop types in study area

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 of the basin is presented in image below (see Figure 24).

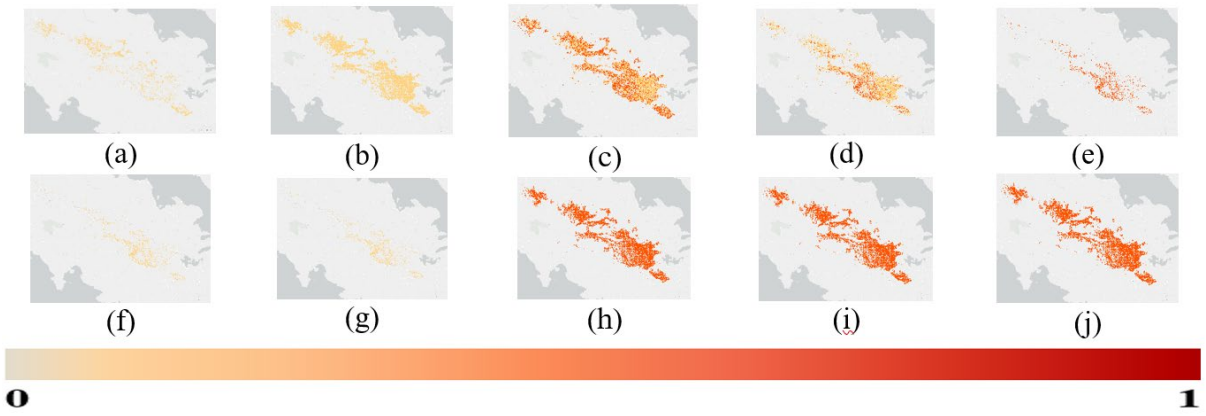


Figure 24: Monthly NNLI at the basin

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.

4.2.5 Conclusion and next steps

The proposed early warning system for the development of the first soft sensor incorporates various models that describe the factors contributing to eutrophication in the lake. Starting with the estimation of nutrient load, a supervised crop-type classification model was developed and integrated with agricultural practices to quantify the nutrient load in the basin. This approach provides stakeholders with valuable insights into when and where nutrient loads are located. Concurrently, the development of a hydrological model, based on Earth Observation (EO) data and precipitation forecasts, serves as an essential tool for discharge estimation at the operational level. Combining these two components enhances understanding and enables stakeholders to anticipate when nutrient runoff will occur, allowing them to prepare their treatment processes accordingly.

The next step should involve mapping crop types from previous years using the model described above. This will help examine how changes in crop patterns impact nutrient loads downstream and provide a larger dataset to test the early warning system more effectively. Additionally, incorporating temporal downscaling of nutrient load estimates from monthly to bi-weekly intervals, based on literature related to agricultural practices, could further refine predictions and improve the system's responsiveness.

4.2.6 State-of-the-art and replicability

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.

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 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

In-situ measurements: In-situ Chl-a measurements are obtained from the ISA USV Boat 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 (-PO4) and nitrate (NO3).

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

11-February-2019	24-July-2019
19-February-2019	25-Sep-2019
06-March-2019	28-May-2020
19-March-2019	05-August-2020
02-April-2019	24-September-2020
22-April-2019	24-May-2021
03-May-2019	03-November-2022
29-May-2019	

EO data: The EO data used for the soft sensor development include data from the Sentinel-1 and Sentinel-2 missions. Sentinel-1 is a satellite mission designed to provide all-weather radar imaging of Earth's surface. The first satellite, Sentinel-1 A, was launched in 2014 while Sentinel-1B was launched in 2016. Both satellites carry Synthetic Aperture Radar (SAR) instruments, operating in the C-band. Sentinel-1 data are used to dynamically extract the shape of the lake. The Sentinel-2 mission also consists of two satellites (Sentinel-2A and Sentinel-2B). 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.

4.3.3 Material 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

Based on the literature, many water quality satellite reflectance algorithms have been used for retrieving Chl-a concentration (Buma and Lee, 2020), (Johansen et al., 2018), (Ogashawara et al., 2021). For the development of the soft-sensor 4 remote sensing were selected, and 4 spectral bands of the Sentinel-2 satellite. The indices and the bands selected are the following:

- **Band 4 (Red Band, 665 nm):** Because Chl-a absorbs light strongly in the red region, it makes B4 of the Sentinel-2 satellite useful for detection Chl-a levels.
- **Band 8 (Near Infrared Band, 842 nm):** The reflectance of the NIR part of the spectrum increases with dense phytoplankton blooms, which are indicative of Chl-a presence.
- **Band 3 (Green Band, 560 nm):** The B3 reflectance helps distinguish between clear water and algal blooms, as water reflects more radiation in the green part of the spectrum as the Chl-a levels rise.
- **Band 2 (Blue Band, 490 nm):** In the blue band, the absorption of the sunlight rises in phytoplankton-rich waters offering another perspective for detecting changes in the water quality.
- **Enhanced Vegetation Index (EVI)**: EVI, is characterized by high sensitivity to high biomass areas, as it reduces the atmospheric influences and enhances the signal quality over water.

$$EVI = 2.5 * \frac{(B8-B4)}{(1+B8+6*B4-7.5*B2)} \quad \text{Eq. 10}$$

- **Normalized Difference Chlorophyll Index (NDCI)**: NDCI is specifically developed to detect Chl-a concentration in water, and characterized by high sensitivity to changes in water quality and algal concentration.

$$NDCI = \frac{(B5-B4)}{(B5+B4)} \quad \text{Eq. 11}$$

- **Red-Edge Chlorophyll Index (ReCl)**: ReCl was selected because the red-edge bands are sensitive to high concentration. The index leverages reflectance differences in red-edge and the near-infrared bands.

$$ReCl = \frac{(B6-B5)}{(B6+B5)} \quad \text{Eq. 12}$$

- **Normalized Difference Turbidity Index (NDTI):** NDTI is used to measure turbidity, which often correlates with phytoplankton and suspended matter in water.

$$NDTI = \frac{(B4-B3)}{(B4+B3)} \quad \text{Eq. 13}$$

- **Normalized Difference Vegetation Index (NDVI):** NDVI is widely used for the monitoring of vegetation health as it is sensitive to Chl-a absorption in the red and reflectance in the near-infrared band.

$$NDVI = \frac{(B8-B4)}{(B8+B4)} \quad \text{Eq. 14}$$

- **Green Normalized Difference Vegetation Index (GNDVI):** GNDVI utilizes the green band which makes it helpful for detecting the green biomass in water.

$$GNDVI = \frac{(B8-B3)}{(B8+B3)} \quad \text{Eq. 15}$$

- **Normalized Difference Red-Edge Index (NDRE):** NDRE is designed to capture chlorophyll content through the red-edge band (B5).

$$NDRE = \frac{(B8-B5)}{(B8+B5)} \quad \text{Eq. 16}$$

Variable Selection

For the Chl-a concentration estimation, it is essential to choose the variables that are most predictive for the model development. This approach helps reduce model complexity, improves computational efficiency and enhances generalization. By selecting the most relevant variables, the model is less prone to overfitting and eliminates noise from the irrelevant predictors. For the variable selection process, correlation analysis was conducted, the results of which are presented in the following table 10.

Table 10: Correlation analysis results of the Chl-a concentration estimation

	EVI	NDCI	ReCl	NDTI	Band 2	Band 3	Band 4	Band 8	NDVI	GNDVI	NDRE
Chl-a	0.104	0.016	0.138	0.140	0.203	0.225	0.189	0.175	0.016	0.005	0.077

The selected variables were those whose correlation coefficient was higher than 0.10 thus the variables used as input to the model were: EVI, NDCI, ReCl, NDTI, B2, B3, B4 and B8.

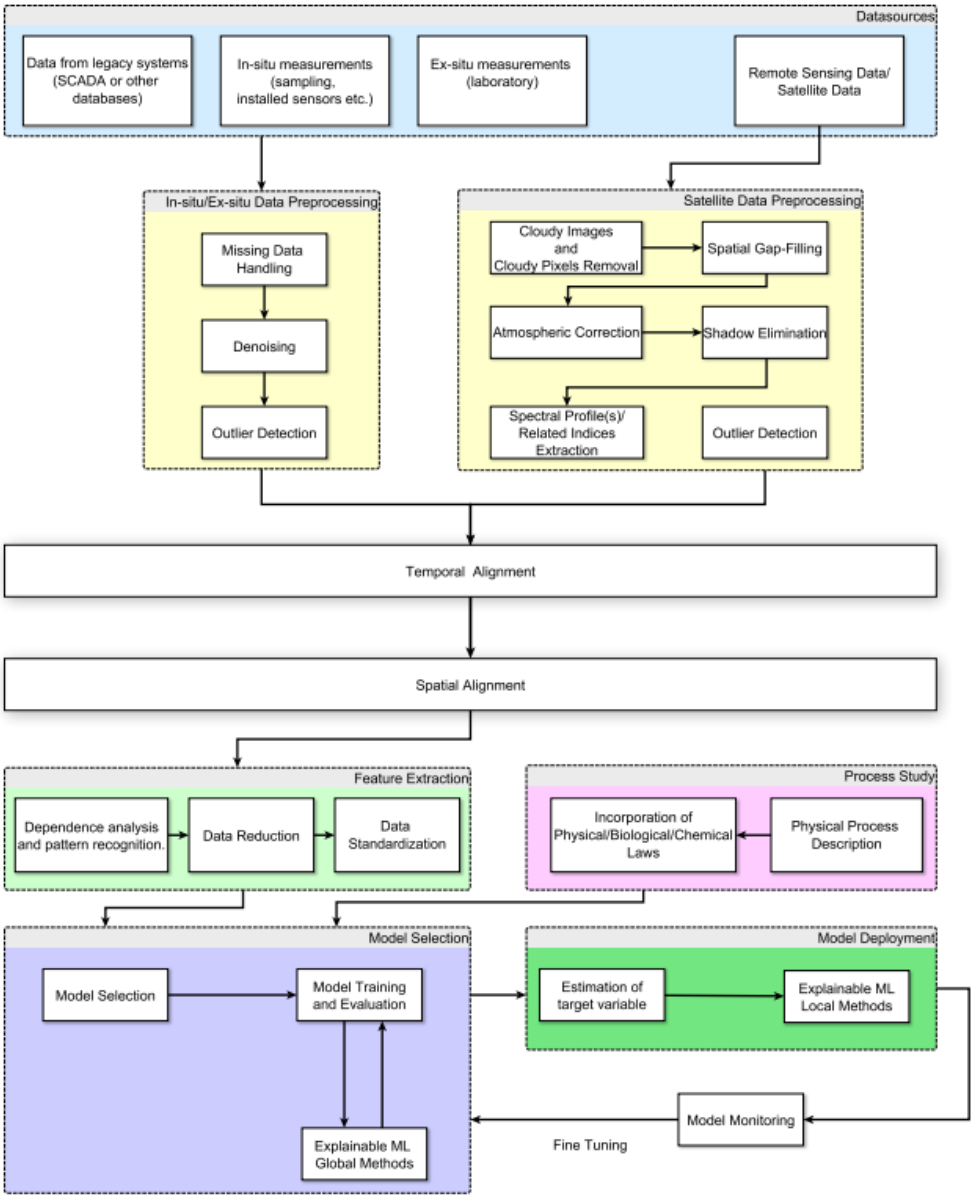


Figure 25: Chl-a concentration estimation flowchart

Model Development

For the model development a Multi-layer Perceptron architecture was chosen because of its ability to handle non-linearity and capture complex relationships. By learning patterns from the inputs, the MLP can provide accurate and general predictions. The MLP developed consists of 7 fully connected hidden layers, two of 96 neurons, three layers of 64 neurons and 2 layers of 32 two neurons. For the model training the dataset was divided into training validation and testing dataset and each input feature was normalized as presented in the following formula:

$$x' = \frac{x^i - x_{min}}{x_{max} - x_{min}} \qquad \text{Eq. 17}$$

Imbalanced Data Handling

The challenge encountered during model training was the imbalance in the dataset. Specifically, in Yliki Lake, chlorophyll-a (Chl-a) concentrations were predominantly observed in the medium concentration range, with significantly fewer instances of extremely low or high concentrations. This imbalance is better illustrated in Table 11 and Figure 26.

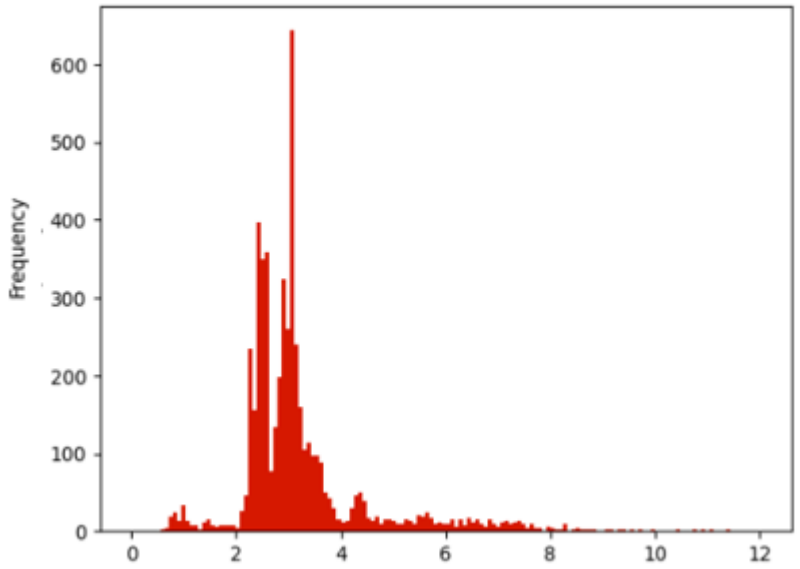


Figure 26: Histogram of Chl-a concentration

Table 11: 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.

4.3.4 Results

The model results indicate a relatively strong performance in terms of prediction accuracy. The training Mean Absolute Error (MAE) for the training dataset is 0.295 μg/l which suggests that the model is making relatively small errors on the training data. The testing MAE is slightly higher (0.348 μg/l), but still relatively low indicating that the model generalizes well. Finally, the R-squared value between the observed and predicted Chl-a concentration indicates that 87.1% of the variance in the Chl-a concentrations can be explained by the model and thus the model adequately captures the underlying relationships in the data.

Overall, the model performs well with low error rates on both training and testing sets, while the R-squared indicates its strong predictive capabilities. The map of the Chl-a concentration for the 23rd of October, 2023 are presented in Figure 27 below.

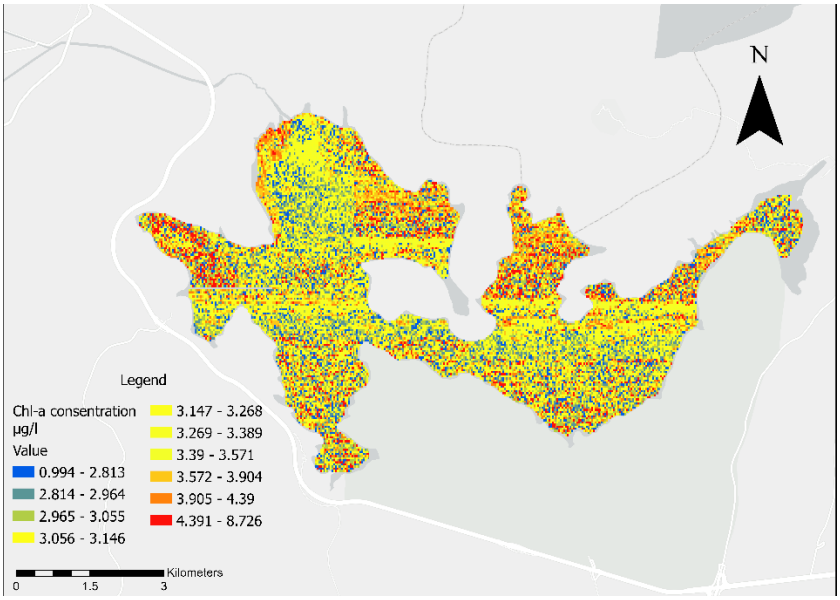


Figure 27: Chl-a concentration at 23-October-2023

4.3.5 Conclusion and next steps

While the model demonstrates strong performance, cross-validation across different regions should be explored. Testing the model on satellite data from various lakes or regions will help assess its robustness in diverse geographical and environmental contexts. Additionally, the incorporation of convolutional layers, which are effective at handling spatial information, could be considered. Evaluating and comparing a Convolutional Regression model against the current approach may provide further improvements in performance.

4.3.6 State-of-the-art and replicability

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 main novelty of the proposed soft sensor lies in the weighted sampling method utilized for training the neural network. Since Lake Yliki is an oligotrophic water body, most in-situ measurements show moderate Chl-a concentrations (2–8 µg/L, based on the literature). While this indicates good overall water quality, as the lake is not heavily eutrophic, it does not rule out the presence of higher Chl-a concentrations. Elevated Chl-a levels near the water abstraction site pose a significant threat to the water treatment process. By developing a model that accounts for the extreme Chl-a values, the proposed

approach ensures more accurate predictions in critical areas, thereby enhancing the reliability of water quality monitoring.

4.4 Soft Sensor 3 - Bloom Occurrence Probability Estimation (Floating Algal Index)

4.4.1 Problem statement

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.

4.4.2 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).

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

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

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 and Discharge 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 2012 to 2023. These measurements represent forecasted values on a daily time scale. The discharge data have also been incorporated into the NB model, and they were obtained from the results of the hydrological simulation presented in the soft-sensor #1. The variables used for the bloom occurrence probability estimation are presented in Table 12.

Table 12: Variables used for the bloom occurrence probability estimation

Dataset/Model	Variable
NOAA GFS	Meteorological Variable Forecasts
MODIS Reflectance Data	Floating Algal Index calculation
Nutrient Load at the Basin	Normalized Nutrient Load Index (NNLI)

4.4.3 Material and methods

The objective of this section is to establish a predictive model for algal bloom occurrences in Lake Yliki. The model will employ the method Naïve Bayes, considering both the meteorological conditions and the eutrophic status of the lake, as well as the spatial variability of algal accumulations in Yliki based on historical data. This approach will enable the estimation of bloom occurrence probabilities in Lake Yliki.

The adopted method for the algal bloom occurrence probability prediction consists of two main parts. To calculate the bloom occurrence posterior probability and the prediction of bloom occurrence probability in the future. Regarding the former, it is obtained by combining the conditional and the prior probability. The conditional probability is calculated by utilizing the FAI values and the meteorological data while the prior probability is calculated at pixels scale based on the MODIS FAI time series. The combination of the prior and the conditional probability results in posterior probability look-up tables. The latter, the prediction of bloom occurrence probability in the future is based on the combination of the posterior probability look-up tables with new satellite, meteorological and runoff data. The flowchart for the methodology is presented in Figure 28.

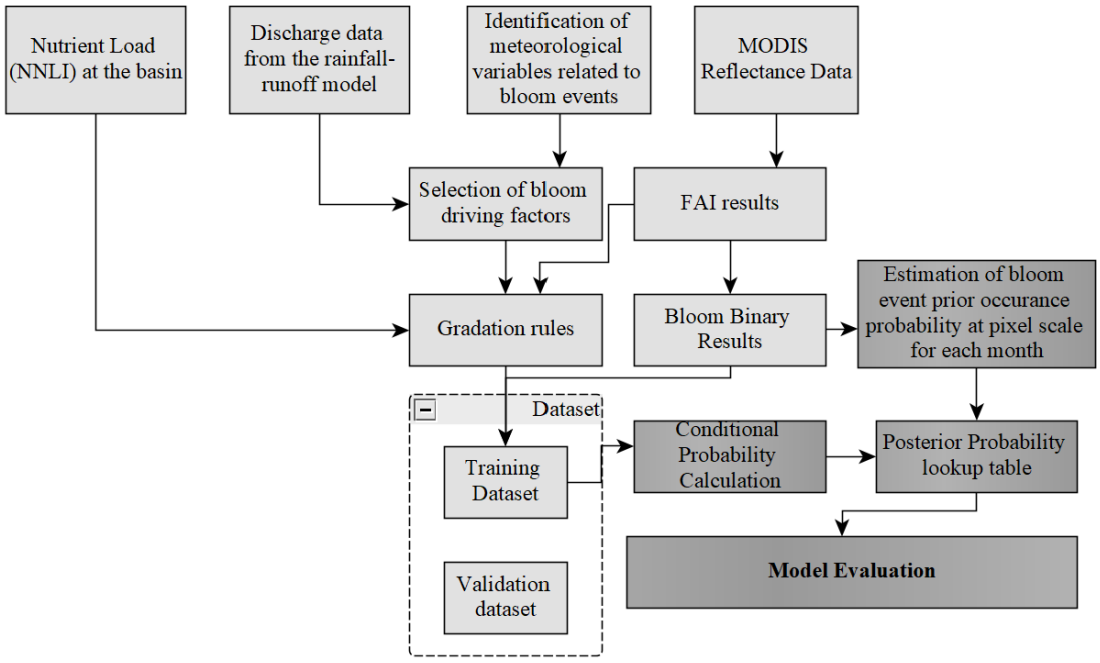


Figure 28: Flowchart of the bloom occurrence probability estimation

Predicting algal blooms using the NB model: Basic Principles

The posterior probability of bloom occurrence is estimated at pixel scale by integrating conditional information as follows:

$$p(C|F) = \frac{p(F|C) * p(C)}{p(F)} \qquad \text{Eq. 20}$$

In the equation, C represents the class variable, which represents the bloom or non-bloom state (C = {c1, c2}), F represents the future variable, which represents the joint effect of parameters related to bloom events (F = {f_1, f_2, f_3,f_n}), $p(F|C)$ represents the conditional probability under the state bloom and non-bloom and finally, $p(C)$ is the algal bloom occurrence frequency, as derived from the MODIS time series at pixel scale.

In the equation of the posterior probability estimation, the term $p(F|C)$ is high dimensional and thus difficult to estimate because it is calculated as $p(f_1|C)p(f_2|C,f_1)p(f_3|C,f_1,f_2)p(f_n|C,f_1,f_2,f_{n-1})$. At this point the NB theory comes into play, and it assumes that the conditional features are independent. With this assumption, the conditional probability under the bloom state is calculated as follows:

$$p(F|C=c_1)=\prod_{j=1}^{d}p(f_j|C=c_1) \qquad \text{Eq. 21}$$

such that:

$$p(C|F)=\frac{p(C)\prod_{j=1}^{d}p(f_j|C=c_1)}{\sum_C p(C)\prod_{j=1}^{d}p(f_j|C=c_1)} \qquad \text{Eq. 22}$$

A specialized variant of the NB algorithm in the Categorical Naïve Bayes (CNB), which is designed for categorical features, and is suited for features that can be distinctly separated into multiple categories. CNB is built around this idea (that the features fall into discrete category) and can handle both categorical and binary data, which it treats as a special case of categorical data with only two available labels. The categorical NB will be used to estimate the bloom occurrence probability in the lake. The establishment of the dataset, the feature selection and the gradation assignment methodology as well as the probability estimation methodology are presented in the following paragraphs.

Variable Selection

To identify the most influential factors for model development, a two-step process was employed. Initially, key variables were selected through an extensive review of relevant literature. Each variable was then subdivided into categories representing its behaviour over different time intervals. For each variable, calculations were made for the same-day value, as well as 2-day, 3-day, 4-day, and 5-day averages, along with the maximum values for these periods. This resulted in sub-categories such as 2-day, 3-day, and 5-day averages, in addition to the corresponding maximum values. The variable selection was guided by the Random Forest classifier. FAI values for the entire lake were calculated for the period from 2016 to 2023, synchronized with the available meteorological data.

In the Random Forest (RF) algorithm, an ensemble of decision trees is constructed. Each tree is trained on a random subset of data, with branches formed by splitting based on the feature that provides the best division at each decision point. The importance of each feature in the RF classifier is determined by how much it improves the purity of the splits across all trees in the forest. The quality of each split is measured by impurity reduction, with features that cause a significant reduction in impurity being considered more important. The results of the RF feature importance are provided in Figure 29.

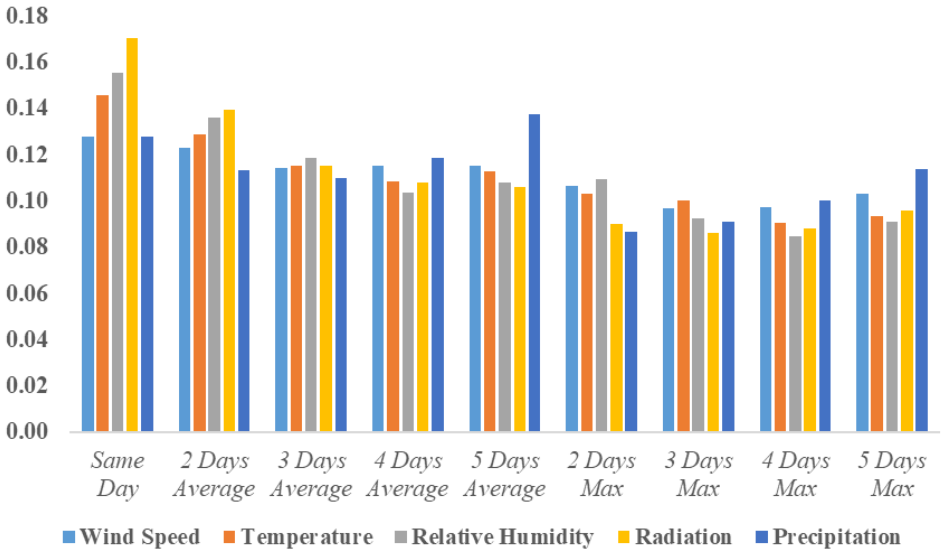


Figure 29: Results of the feature importance analysis

As seen in Table 13, the same day forecast is the most important feature and better captures the FAI variability for all variables, except for the precipitation, for which the most important feature is the 5 days average. The table below summarizes the selected meteorological sub-categories.

Table 13: Summary table of the meteorological variables used for the FAI estimation

Meteorological Variable	Selected sub-category
Wind Speed	Same-day forecast
Temperature	Same-day forecast
Relative Humidity	Same-day forecast
Radiation	Same-day forecast
Precipitation	5-day average

The discharge data used in the NB model are the results of the hydrological simulation described in the early warning system (Soft-sensor #1). The most important features for FAI predictions were selected the same way as the meteorological factors (RF classifier). Again, the discharge data were grouped in 2-day, 3-day, 4-day, and 5-day averages, along with the maximum values for these periods. The result of this analysis is provided in the table below. The 5-day average is the most important feature, as it is concluded by the analysis (see Figure 30).

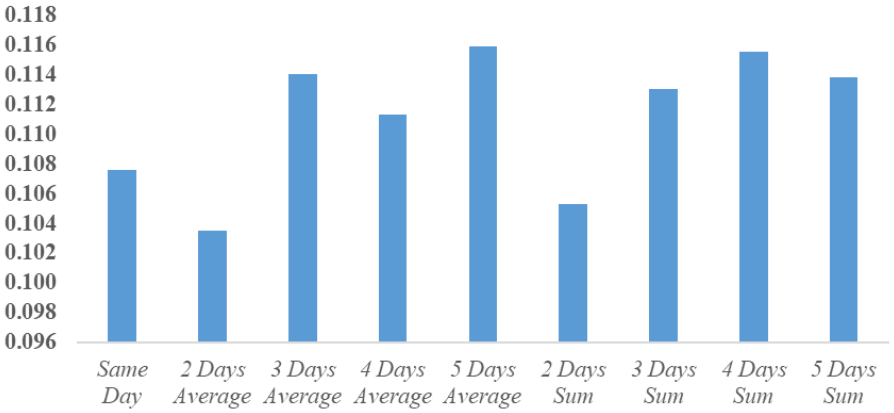


Figure 30: Summary table of the discharge variables used for the FAI estimation

Data Preparation for the NB algorithm

To prepare the inputs for categorical Naïve Bayes, data needs to be broken down into categories. Each input variable, including meteorological data, discharge, and nutrient load, was assessed based on its distribution during bloom and non-bloom days. If a variable had higher values during bloom events, it was marked as "positive," suggesting a supportive role in bloom occurrences. In contrast, variables showing higher values on non-bloom days were labelled as "negative."

Meteorological variables were further classified into three levels, referred to as g1, g2, and g3, using two key percentiles. For variables with a positive impact, the thresholds included the 25th percentile on bloom days (B0.25) and the 75th percentile on non-bloom days (N0.75). Negative-impact variables were classified using the 75th percentile on bloom days (B0.75) and the 25th percentile on non-bloom days (N0.25). Initially, an analysis was conducted to filter days based on bloom occurrence, and then percentiles were calculated to categorize environmental factors into low, medium, or high levels according to their relevance to bloom or non-bloom thresholds. Each variable was grouped as follows: g1 represents values ≤ L1 (small threshold), g2 for values between L1 and L2, and g3 for values > L2 (large threshold).

Similar distribution characteristics were presented by almost all the input parameters including average wind speed and relative humidity on the same day, 5-day average precipitation and discharge, and the nutrient load in the basin (NNLI), indicating they have a positive effect on bloom occurrence. However, air temperature on the prediction day was lower on bloom days than on non-bloom days and has a negative effect on algal bloom occurrence (see Figure 31).

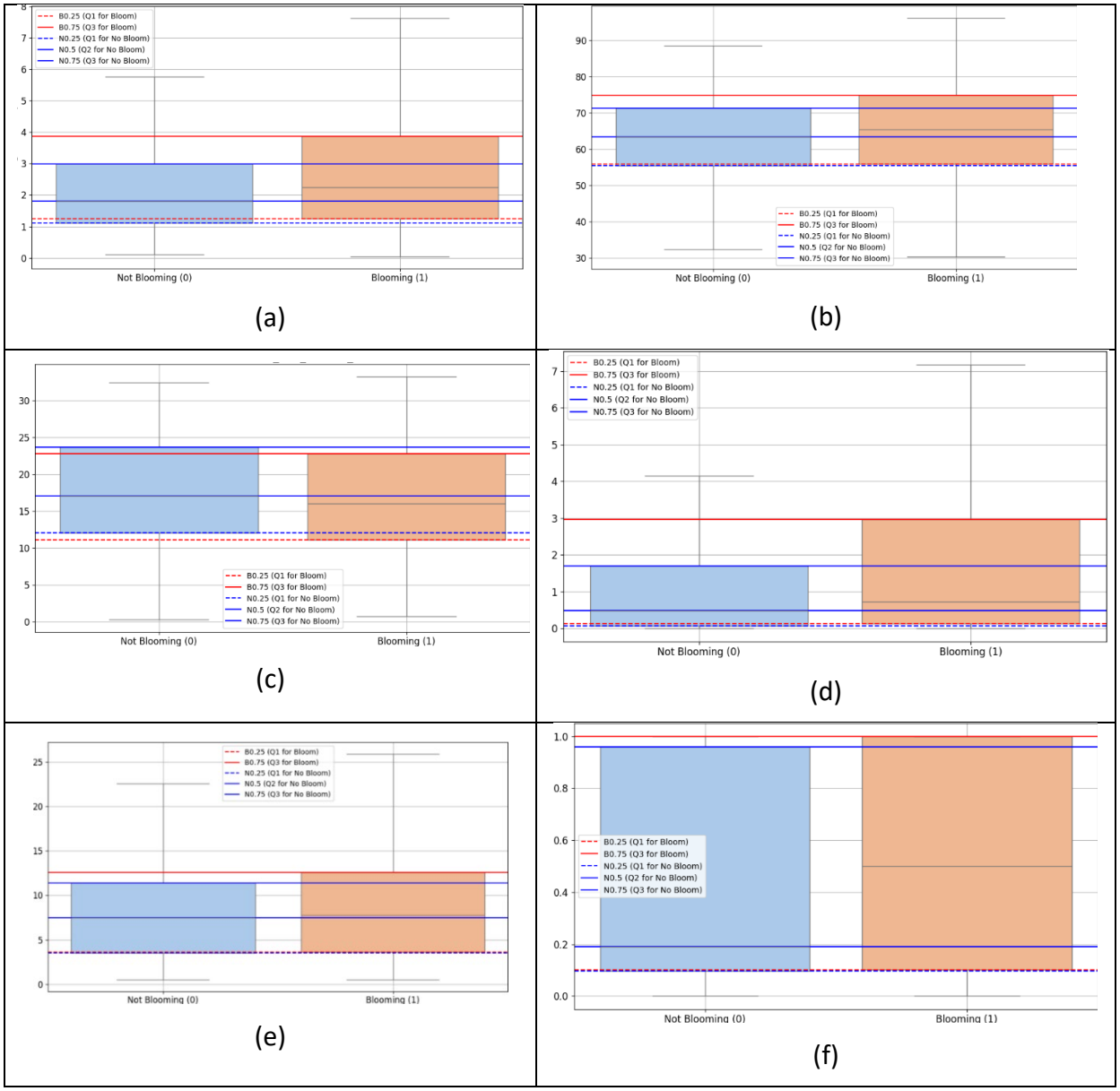


Figure 31: Grading method for the input variables

After the gradation of the input variables as described above, the quantile of each variable is differentiated as small and large, are denoted as L1 and L2 respectively and are available at the Table 14.

Table 14: Grading thresholds

Grading threshold	L1	L2
Average Wind Speed on the Prediction Date	1.25	2.98
Relative Humidity on the Prediction Date	55.85	71.31
Temperature on the Prediction date	12.06	22.77
5-days Average Precipitation	0.12	1.70
5-days Average Discharge	3.59	11.40
NNLI	0.10	0.96

Prior Probability Estimation

In a Naïve Bayes model, it's crucial for prior probability to be both credible and realistic. While ground observation records are the typical way to gather this prior information, algal blooms often appear unexpectedly and sporadically across large regions, which makes traditional methods unreliable. To overcome this, using long-term satellite imagery provides a unique and statistically sound approach to track bloom occurrences. In this study, prior probabilities were calculated by evaluating satellite images to represent the likelihood of blooms at each pixel location, using the following formula:

$$p_m(c1) = \frac{number\ of\ c1 + 1}{N + 2} \qquad \text{Eq. 23}$$

In this equation, m represents the pixel number, c1 represents the bloom state, and N represents the number of observations. The results of the prior probability estimation at the pixel scale for each month are presented in Figure 32:

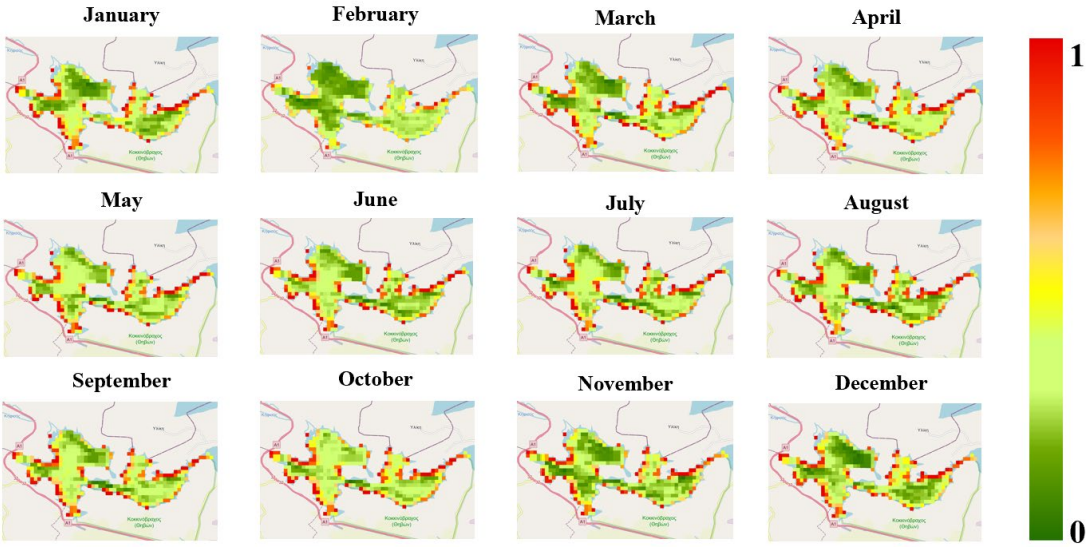


Figure 32: Monthly Bloom Occurrence Probability at pixel scale (250m)

Conditional Probability Calculation of Algal Bloom Occurrence

Each complete observation in the NB models was composed of the bloom state variable (0 or 1), the selected meteorological, discharge and nutrient load variables. The conditional probability of bloom and non-bloom in the following day for each pixel was calculated using the following equation.

$$p_m(F|c_i) = \prod_{j=1}^{d} p_m(f_j|c_i) \qquad \text{Eq. 24}$$

Where m represents the pixel number, f_j represents the selected features where j = 1…6, and c_i represents the bloom or non-boom state, were i = 1 (bloom) or 2 (non-bloom).

4.4.4 Results

For the model development, the dataset was split into the training and the validation dataset in a 0.8/0.2 ratio. The training dataset is used for the development of the look-up tables allowing for quick look-up during prediction, making the model efficient. In the prediction process, the posterior probability is

calculated by using the prior and the likelihood calculated during training. Finally, the class with the highest posterior probability is chosen as the predicted class for the instance.

The overall accuracy of the Categorical Naïve Bayes (CNB) model is 0.5524. The components of the classification report are presented in the following Table 15:

Table 15: Accuracy of the CNB model

	Precision	Recall	F1-score
0	0.51	0.73	0.60
1	0.65	0.42	0.51

For class 0, the model has a precision of 0.51, which means it's only correct 51% of the time when it predicts this class, indicating a high rate of false positives. On the other hand, it has a recall of 0.73 for class 0, so it successfully catches 73% of the actual instances. Class 1 tells a different story: here, precision is higher at 0.65, meaning 65% of the predictions for this class are accurate, but recall drops to 0.42, showing the model misses quite a few true instances of class 1. The F1-scores—0.60 for class 0 and 0.51 for class 1—reveal a clear trade-off between precision and recall, especially for class 1, where the model struggles to detect positive cases.

In general, these results suggest the model is somewhat better at correctly predicting class 1, though its sensitivity is limited, resulting in missed detections. The outputs of the proposed tool have been produced displaying the probability for the weekly bloom occurrence between 1st and 7th September 2021 as a set of maps. Each map presents the probability of the blooming event; in fact, each pixel gives the probability of the same event: from light to dark green, ranging from 0 to 1.

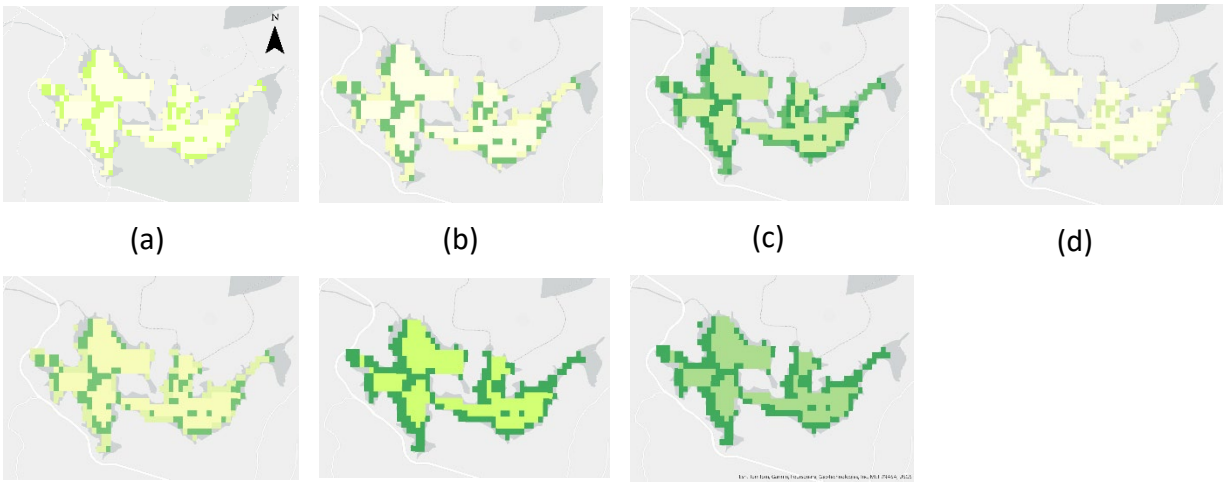


Figure 33: Results of the NB model from 1-9-2021 (a) to 7-9-2021 (g)

4.4.5 Conclusion and next steps

The soft sensor for bloom occurrence probability offers a comprehensive approach to estimating bloom events in the lake at the pixel level. By integrating the functions of the early warning system, it provides key information on discharge and nutrient loads within the basin. Additionally, incorporating forecasts of meteorological variables allows the model to give the stakeholders timely and valuable insights.

Although the current model's results are acceptable in terms of the F1 score and recall for the non-bloom class, there is significant potential for improvement. To enhance performance, the next steps will involve exploring alternative models and architectures, such as Bayesian Neural Networks (BNNs), to leverage their advanced computational capabilities.

4.4.6 State-of-the-art and replicability

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, including discharge data, nutrient loads, 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 Soft Sensor 4 – Estimation of Water Quality Index in the Yliki Lake at pixel level

4.5.1 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.

4.5.1 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

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, EC, 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 Table 7.

EO-data: For the development of this soft sensor both radar and multispectral data are utilized. The radar data includes the data from the Sentinel-1 mission for the dynamic extraction of the lake's shape and the multispectral data are those of the Sentinel-2 mission, the same way as for the development of the Chl-a soft sensor.

Preprocessing

The preprocessing of the in-situ measurements follows the preprocessing steps as described in the Chl-a soft sensor. For the outlier removal it was conducted based on expert judgment and the lower and upper accepted limits were determined based on the recommendations provided in the campaign report. The water body extraction, satellite data pre-processing, and the spatiotemporal-alignment processes are the same as those described for the Chl-a development.

4.5.2 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 34. 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.

The values of Chl-a, DO and pH that are estimated following the flowchart outlined for the second soft sensor. Specifically, before estimating the Water Quality Index (WQI) as shown in the flowchart, each of the input water quality variables is mapped. This means that the variables used for the WQI, such as Chl-a, are first estimated using a model like the Chl-a soft sensor, and spatial results for these variables across the entire lake are obtained. Additionally, the imbalanced data present in the DO and pH measurements will be addressed using techniques designed for handling imbalanced data.

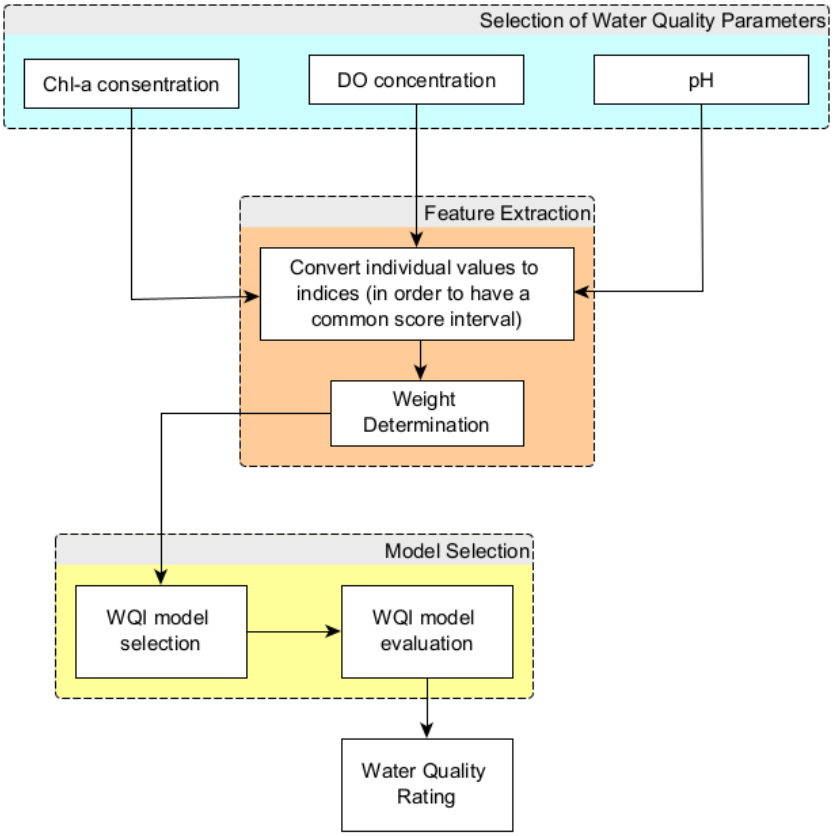


Figure 34: *WQI estimation flowchart*

Regarding the modeling of the Water Quality Index (WQI), as outlined in the flowchart, the process begins with modeling the input variables across Lake Yliki. Once this is completed, the individual values are converted to 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.

Finally, various aggregation functions can be used, as documented in the literature. These include additive, multiplicative, and combined aggregation functions, as well as the minimum operator and unique linear or non-linear aggregation methods.

4.5.3 Results

In the development of the Chl-a soft sensor, it was demonstrated that combining satellite data with in-situ water quality measurements can effectively model the target parameters. This finding establishes a strong foundation for creating a satellite-based soft sensor for water quality monitoring.

Preliminary analysis of available data and literature suggests that modelling the remaining parameters, dissolved oxygen (DO) and pH, using remote sensing is feasible. Notably, similar studies have achieved this with fewer in-situ measurements. In the case of Yliki Lake, data collection and pre-processing are ongoing, and complete results are pending. The final outcomes are expected to capture the dynamics of individual variables and provide a comprehensive view of water quality at the pixel scale on the lake's surface.

4.5.4 Conclusion and next steps

The anticipated results of the WQI soft sensor are expected to validate the hypothesis that overall water quality in the lake can be effectively modelled at the pixel level. This will provide a valuable tool for high-resolution water quality monitoring.

Moving forward, the priority is to complete data collection and pre-processing. Following this, the development of the WQI soft sensor will proceed using a Multi-Layer Perceptron (MLP) Neural Network, leveraging its capability to model complex relationships between satellite data and water quality parameters.

4.5.5 State-of-the-art and replicability

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.

The novelty of the proposed soft sensor has to do with the training of separate models for each parameter and then adapting the WQI function based on the characteristics of Lake Yliki. Incorporating weighting in the index, based on the relevance to the water body characteristics, provides a potentially more accurate representation of the overall quality of the water body.

5. Soft Sensors for Val de Bagnes demo case

5.1 Introduction/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 Ruinettɛ. 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 5 - Early Warning System for Bacteriological Contamination in Sarayer

5.2.1 Problem statement and soft-sensor development flow-chart

The early warning system for the Val de Bagnes demo case will follow a similar approach to the early warning system developed for nutrient runoff in the Athens demo case. It will integrate satellite data with in-situ measurements to provide a comprehensive monitoring solution. Satellite data will be utilized to estimate the contaminant load from bacteriological pollutants, by mapping grazing patterns.

5.2.2 Data sources and data preprocessing

The available data sources include both in-situ and satellite data and are presented in the following table (see Table 16):

Table 16: Datasets for the early warning system for the Val de Bagnes demo case

1.	**BactoSense Measurements (from 1-10-2023 to 29-07-2024)**
2.	Historical Data from Mayentset and Combavuatsi
3.	NOAA GFS Meteorological Variables (as hydrological simulation variables)
4.	Multispectral and radar satellite imagery (e.g., Sentinel-1, Sentinel-2, Landsat-8) for the mapping of the pasture area

5.2.3 Material and methods

As mentioned above, the approach is similar to the development of the early warning system for nutrient runoff in the Athens demo case. This approach differs in some respects from the one described above, as the in-situ data (the in-situ crop-type information in the early warning system for the Athens demo case) will not be used to map the non-point pollutant source but rather to calibrate and evaluate the early warning system (the in-situ measurements refer to the bacteriological contamination in Mayentset).

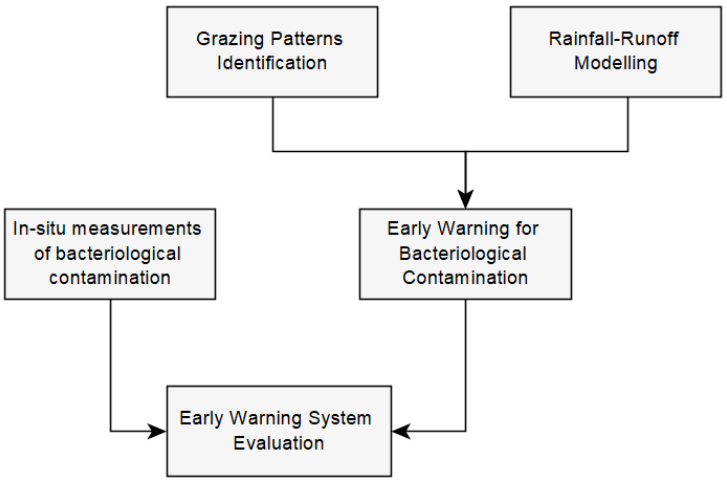


Figure 35: Proposed flowchart for the early warning system for bacteriological contamination

In-situ measurements will capture concentrations of bacteriological parameters, which will be used to evaluate the system's effectiveness. Additionally, a rainfall-runoff model will be developed to simulate the transport of bacteriological contaminants within the study area, enhancing the system's predictive capabilities. The flowchart of the early warning system for bacteriological contamination is presented at the following figure (see Figure 35):

5.2.4 Conclusion and next steps

The preparatory work for the ALTIS demo case has highlighted its significant potential to be fully developed across all stages. This includes comprehensive efforts on problem identification, the development of a flowchart, and the implementation of soft sensors. By the end of the next reporting period, it is expected that a final version of the functional soft sensors will be developed, alongside progress on other demo cases in Amsterdam and Athens.

The anticipated outcomes for the early warning system addressing bacteriological contamination in the ALTIS demo case are projected to align closely with those of the Athens demo case. Specifically, the integration of non-point pollutant source identification and pollutant quantification with rainfall-to-runoff transformation shows promise. This approach is expected to yield valuable results for notifying stakeholders of potential bacteriological contamination risks.

5.2.5 State-of-the-art and replicability

The proposed methodology, although slightly different from the Early Warning System (soft-sensor #1) proposed for the Athens demo case for nutrient runoff, leverage similar principles of non-point source identification through EO data and in-situ measurements. The inclusion of in-situ measurements for bacteriological contamination would be critical for calibrating and evaluating the system's predictions.

6. Soft sensors for Amsterdam demo case

6.1 Introduction/Demo case description

Description of the demo case

Waternet supplies more than 90 million cubic meters of drinking water annually to consumers in the Amsterdam area. The water is produced at two different drinking water treatment plants (DWTP), Leiduin and Weesperkarspel.

Figure 36: Overview of Amsterdam demo case area

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 37.

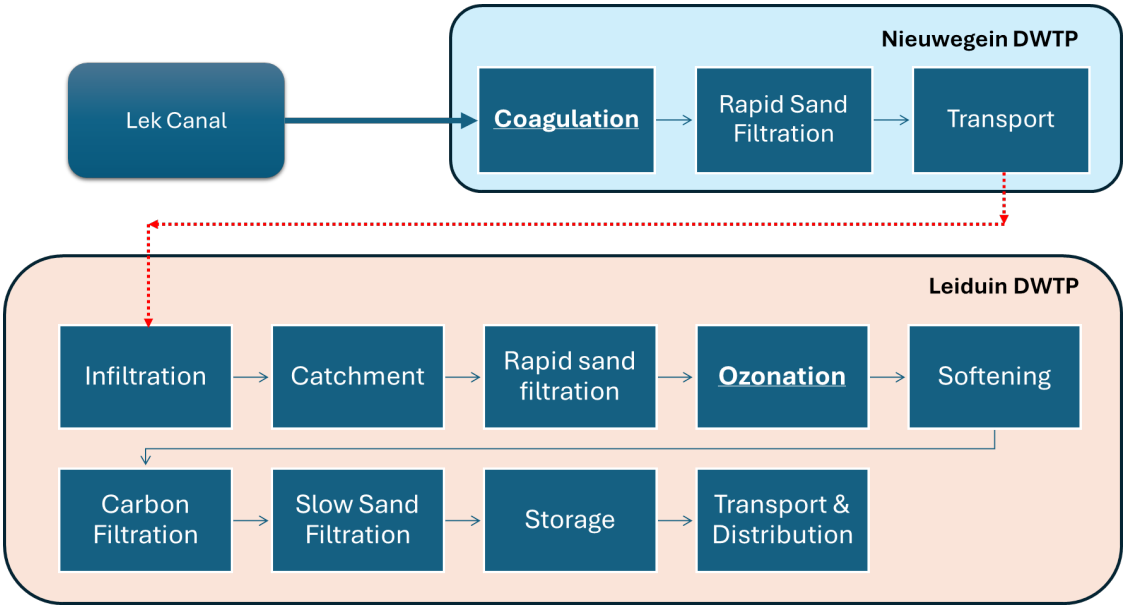


Figure 37: schematic of the Leiduin drinking water treatment plant

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:

1. **Soft Sensor #6 - Estimation of turbidity of the coagulation – flocculation process (KWR).**

   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.

2. **Soft Sensor #7 – Early prediction of turbidity in the DWTP inlet (TUD).**

   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

3. **Soft Sensor #8 – 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 6 - Estimation of turbidity of the coagulation-flocculation process

6.2.1 Problem statement and soft-sensor development flow-chart

Waternet is a company that operates a treatment plant at Nieuwegein, the Netherlands. This treatment plant comprises a coagulation-flocculation treatment technique, which relies on the addition of ferric chloride to remove contaminants like organic carbon and phosphates. The operation of this technique and the amount of ferric chloride added rely on measurement data and expert knowledge of operators and process engineers. To support the decision making by these employees and minimize the use of ferric chloride, we have developed a soft sensor that can help interpret the status of the coagulation-flocculation technique and give advice regarding its operation.

In the soft sensor developed in this project, we build a model that predicts the turbidity at the outlet of the treatment plant, based on readily available measured input parameters. In the soft sensor developed in this project, we build a model that predicts the turbidity at the outlet of the treatment plant, based on readily available measured input parameters. This model allows operators to have a good indication of this important outlet parameter at the start of the treatment, getting an indication of future outcomes. The turbidity at outlet 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 technique. 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 turbidity at outlet 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.

6.2.2 Data sources and data preprocessing

A large dataset was provided by Waternet. The different types of data that are used by the soft sensor are summarized in Table 17.

The first six parameters in Table 17 are model input or *feature* parameters, and the last one is the model output or *target* parameter. Turbidity, temperature, pH and flow are measured by hard sensors at the inlet or outlet, and the amount of ferric chloride ($FeCl_3$) added for treatment and the lanes used are actively managed. The lanes used refers to which separate reactor paths are used: in Nieuwegein there are three reactor paths, or 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 17: 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
Flow (influent)	Floating point number	$1 \cdot 10^3$-$7 \cdot 10^3$	m^3 per hour	Amount of water taken 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
Lanes used	Set of Booleans	[0-1, 0-1, 0-1]	-	Which of the three lanes are used at this time.	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 six 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 Material 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,[4] 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,[5] 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,*[6] 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[4].

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 as described in Table 3. This overall model was developed as graphically depicted in 38.

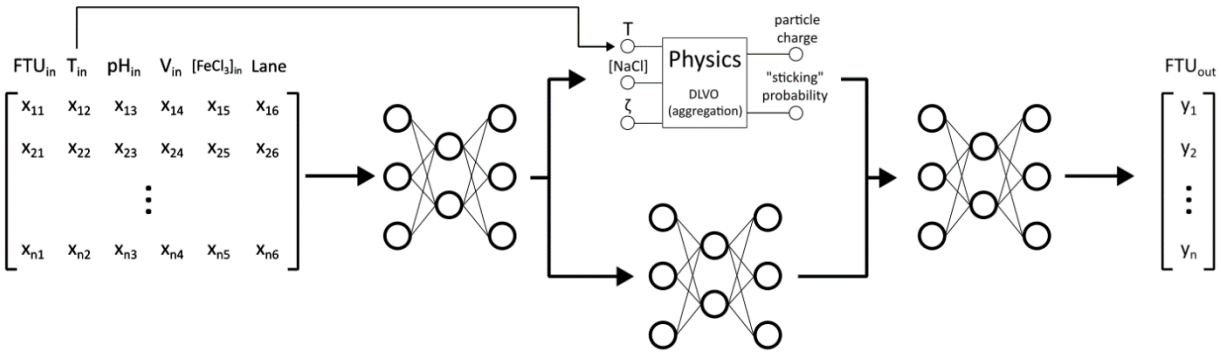


Figure 38: Geometry of the model. The data enters into a neural network, which pipes its outputs into two branches. The first branch is the physics model, the second branch is another neural network. The outputs of both branches are piped into a final neural network, and its outputs are the predictions of the model.

The input data (Figure 38, 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

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

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 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 38) 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 18: 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	90	75	50
Learning rate	$5 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	$5 \cdot 10^{-3}$
Neurons per hidden layer, first neural network subsection	75	120	150
Number of hidden layers, first neural network subsection	5	10	5
Neurons per hidden layer, second neural network subsection	75	120	150
Number of hidden layers, second neural network subsection	5	10	5
Neurons per hidden layer, last neural network subsection	20	20	60
Number of hidden layers, last neural network subsection	15	10	7

The performances of the final models can be found in Table 4. 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 39. The performances on unseen data (testing set) are between 0.4-0.5, which has a slightly better performance than decision tree-based models developed by Waternet given the same input data. It is notable that the multi-branch model has the highest performance, even though the difference is 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 and yield the most accurate predictions. 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 19: Performances of the final model. For reference: the R^2

	R^2 training set	R^2 testing set
Physics-piped model	0.72	0.42
Purely neural-network model	0.77	0.44
Multi-branch model	0.74	0.47

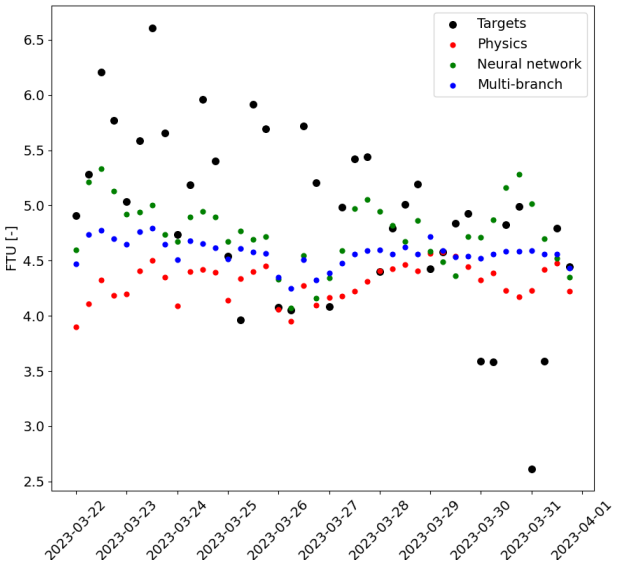


Figure 39: Example data points reflecting the original data targets (effluent turbidity) and the predictions by the three models

6.2.5 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 7 - 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 reduces turbidity, removes pathogens, and improves the overall clarity of the water. Optimizing dosing and optimal conditions is crucial for efficient coagulation. However, fluctuating influent water quality poses significant challenges to the task of achieving and maintaining these optimal conditions. Direct measurement of water-quality parameters can be challenging, often requiring laboratory experiments with delays ranging from days to weeks. Moreover, most treatment processes exhibit slowly changing behaviour, influenced by rapid fluctuations in water quality and control actions. Traditional methods, 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. 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 impending water quality events.

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) have been extensively employed in water quality prediction. Since 2016, the number of research publications concerning water quality management using AI or ML has steadily increased. These studies aim to estimate various water quality parameters across different segments of the water supply chain, from source to tap. For example, (Ortiz-Lopez et al., 2023)used rainfall and discharge data as inputs for two ML algorithms—support vector machines and ANNs—while (Lu and Ma, 2020) applied a hybrid approach combining random forest (RF) with XGBoost to predict turbidity at the inlet of DWTPs. However, these studies did not delve into the relationships between multiple WQ variables measured at DWTPs or identify key factors driving turbidity changes. Furthermore, they largely overlooked the potential of state-of-the-art deep learning approaches, which could leverage time-series data to uncover hidden patterns and provide more accurate predictions.

This study seeks to address these limitations by exploring the applicability of a data-driven model based on ML algorithms for short-term turbidity prediction in the Nieuwegein DWTP. The methodology begins with the application of techniques such as time-lag cross-correlation, self-organizing maps (SOMs) (Kohonen Teuvo, 1990), and Spearman's rank cross-correlation to identify time-lagged relationships between input variables and inlet turbidity using hourly timesteps. A classification model employing the long short-term memory (LSTM) algorithm (Hochreiter and Urgen Schmidhuber, 1997) is then developed to detect turbidity events in the DWTP inlet. This methodology will empower Waternet operators to be able to enhance their coagulation processing management strategies and prevent the complete shut-

down of the DWTP. While the data used in this study is not openly accessible due to confidentiality constraints, the methodology itself is designed to be replicable in similar settings with comparable datasets. By tailoring the approach to locally available data, researchers and water utility operators of other DWTPS can adapt this model to enhance coagulation management strategies, reduce operational risks, and prevent costly shutdowns of DWTPs.

6.3.2 Data sources and data preprocessing

Data Collection

The Nieuwegein DWTP takes water from the Lek Canal. The water that feds the canal is directed there from river Rhine. 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 Lobith. Other stations that are located upstream of the DWTP are the Hagestein boven river station and the Nieuwegein station in the Lek canal. 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. Regarding 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.

Table 20: Datasets used for the soft sensor development

Variables	Location	Data availability period	Data Frequency	Data Source
River discharge (m^3/s)	Lobith	Jan 2020-Dec 2023	10-min	RIWA
River discharge (m^3/s)	Hagestein boven	Jan 2020-Dec 2023	10-min	RIWA
Canal flow (m^3/s)	Nieuwegein	Jan 2020-Dec 2023	10-min	RIWA
Water flow(m^3/h)	DWTP inlet	Jan 2020-June 2024	5-min	Waternet
Water temperature (°C)	DWTP inlet	Jan 2020-June 2024	5-min	Waternet
pH	DWTP inlet	Jan 2020-June 2024	5-min	Waternet
Turbidity (FTU)	DWTP inlet	Jan 2020-June 2024	5-min	Waternet

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:

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. 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.

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. **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.
5. **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 Material and methods

Calculating water travel time

The water travel time was calculated using the average flow rate at each station and the distance between them. The results are presented in the following diagram and show that the overall travel time from Lobith to Lek canal is roughly 41 hours with Hagestein boven being roughly in the middle. Thus, the final dataset was reformatted to capture this time lags.

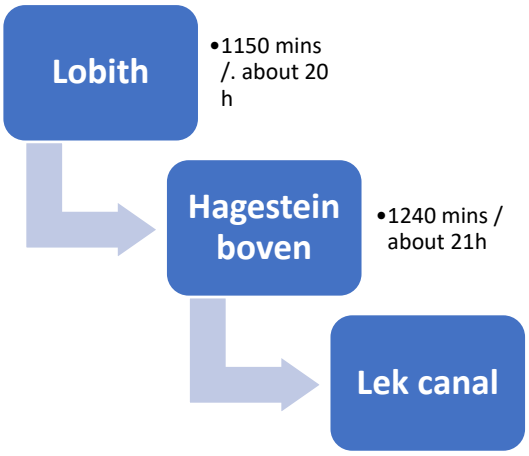


Figure 40: Schematic of the water travel time from Lobith river station to Nieuwegein DWTP

Turbidity event identification and forecasting horizon

The aim of this work is to identify turbidity events and not predict the actual turbidity values though time. Therefore, this is a classification problem, a threshold to define turbidity is required. This threshold was set to a period of at least 4 hours where the turbidity is more than 15 NTU. The inputs for this classification were, the river discharge in the two river stations and the Lek canal, pH of the raw water in the DWTP inlet, temperature of the raw water in the DWTP inlet and the DWTP intake water flow. The forecasting horizon was set equal to 3 and 6 hours ahead.

Correlation identification

The correlation identification included the use of self-organising maps (SOMs), the time-lag cross correlation and the Spearman cross correlations with a time lag. The forecasting horizon for of the predictive model was set to 3 and 6 hours, thus the Spearman cross correlation was tested with the use of the time-lag between the inputs and the outputs.

Self-organising Maps: A SOM 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,

Time-lag cross correlation: Time-lagged cross-correlation was performed to identify the temporal dependencies between input parameters (e.g., river discharge, temperature, pH, and flow) and the DWTP inlet turbidity. The time lag used was from 0 up to 25 hours.

Spearman Cross-Correlation Analysis: Spearman cross-correlation analysis was conducted to quantify the non-linear correlations between the input dataset and the DWTP inlet turbidity. Unlike time-lagged cross-correlation, Spearman correlation evaluates rank-based correlations, providing insights into non-linear and complex dependencies like the ones that occur in water environments.

Turbidity event predictive model using LSTM algorithms

This data-driven methodology outlines the development of an LSTM classification model to predict turbidity events based on water quality parameters with a forecasting horizon of 3 and 6 hours ahead. The LSTM model was initially run using all the available variables and then using the most important variables (predictors) as identified by the correlation process. The data-driven model architecture was set was set as follows:

- **Data preparation**: The time-lagged input dataset was standardized.
- **Input and Predictive Horizon Setup:** Each sample was represented as a sequence of 24 hours of input data, paired with a target turbidity class - event or non-event class. As the dataset was heavily imbalanced towards the non-event class, the random under sampling technique was used to reduce the number of non-event samples in the training dataset.
- **Architecture Design:** The LSTM model architecture was structured to capture temporal dependencies effectively for classification purposes. The model included two hidden LSTM layers, followed by a dense layer with a single neuron and a sigmoid activation function to produce a binary classification output. Two dropout layers with were applied between each layer to prevent overfitting.

- **Input Shape and Output:** The model's input shape was set based on the sequence length of 24 hours and the number of the input variables. The output was a binary class indicating the presence or absence of a turbidity event within the defined predictive horizon.
- **Hyperparameters:** The hyperparameter tuning was conducted using KerasTuner's Random Search. The number of units per LSTM layer and dense layer, the learning rate, and the dropout rate were tuned for optimal performance based on the accuracy of the model. The model was trained for a maximum of 10 epochs over a trial period of 10 repetitions and the number of batches per epoch was set equal to 36.
- **Splitting the data:** Data was split into training and validation sets, with 80% used for training and 20% for validation. Within the training set, an additional 20% was reserved to evaluate model performance during hyperparameter tuning.

Evaluation metrics

The classification performance of the model was evaluated using 3 different performance metrics, recall, precision and F1-score. The formulas for these metrics are as follows:

$$Recall = \frac{TP}{TP + FN} \quad \text{Eq. 26}$$

$$Precision = \frac{TP}{TP + FP} \quad \text{Eq. 27}$$

$$F1 - score = 2\frac{Recall * Precision}{Recall + Precision} \quad \text{Eq. 28}$$

Where:

TP: True positives / FN: False negatives / FP: False positives

6.3.4 Results

Initial analysis of turbidity data

The descriptive statistics of the turbidity dataset are presented in the following table and a group distribution of the turbidity measurements is shown in the following figure. These statistics indicate that most of the turbidity measurements are below the threshold of 20NTU. This is a

Table 21: Descriptive statistics of the turbidity dataset

Mean	**16.05**
Standard Deviation	7.7
Minimum	0
Maximum	167.33
80th percentile	19.66

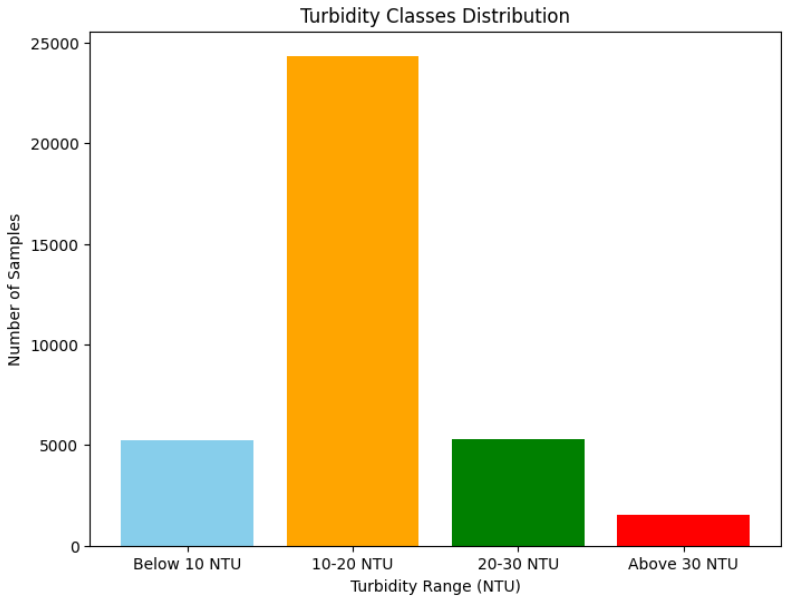


Figure 41: Turbidity distribution classes

Correlation identification

A SOM was generated using as inputs 6 different variables, the intake water flow (label Flowrawwater), the water temperature (label Temperaturerawwater), pH of the raw water (label pHrawwater), turbidity in the raw water(label FTUraw water), Hagestein boven (label Hag_dis), and Lobith discharge flow (Lob_dis). The output SOM is presented in the following figure. The key correlations identified in this SOM are as follows:

- There is a clear correlation between high intake water flow and high temperature.
- As expected, Hagestein boven discharge is perfectly correlated with Lobith discharge.
- There is an inverse correlation between high intake water flow and high turbidity.
- There is an inverse correlation between high temperature and high turbidity.
- There is no clear correlation between pH and turbidity.

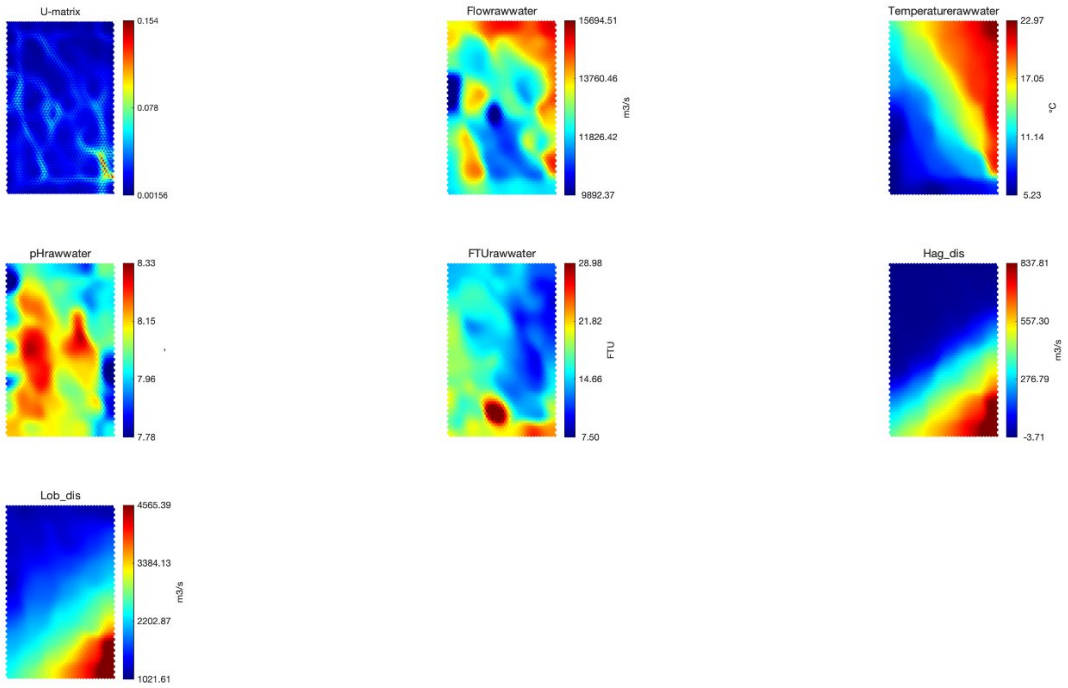


Figure 42: SOMs output using all the inputs parameters and turbidity

The time-lag cross correlation between the inputs and turbidity for a period between 0 hours to 25 hours was implemented and the results are presented in the following plot. This graph shows that there is no significant change of the linear correlations between input variables and turbidity over time. Moreover, it confirmed the visual corelations that were identified with the use of SOM. More specifically, the highest correlation, is the inverse correlation between water temperature and turbidity. Other important correlations were between turbidity and the river discharges and the reverse correlation between turbidity and the intake water flow.

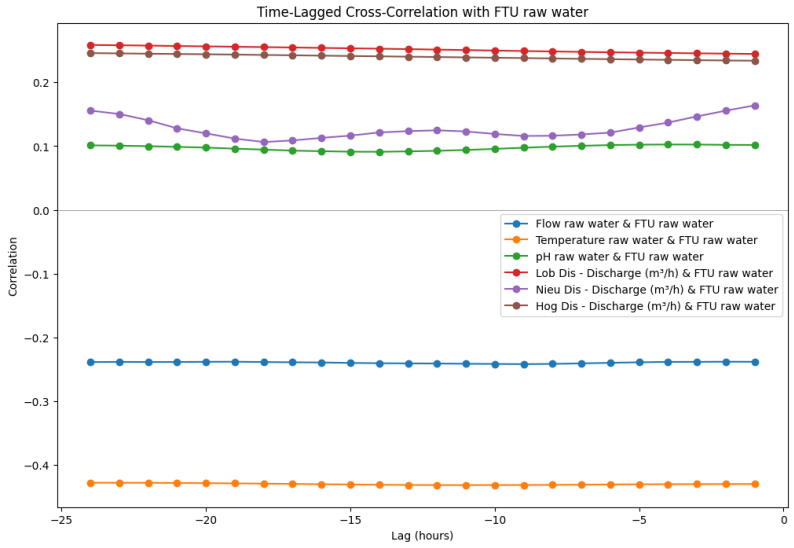


Figure 43: Time-lagged correlation between the inputs and turbidity

The final correlations investigation was the calculation of the time-lagged Spearman Correlation matrix (see figure 44 below). This output indicated a higher non-linear inverse correlation between water temperature and turbidity (0.53 Spearman vs 0.45 linear correlation) but a lower non-linear inverse correlation between discharges and (0.2 Spearman vs 0.24 linear correlation) turbidity.

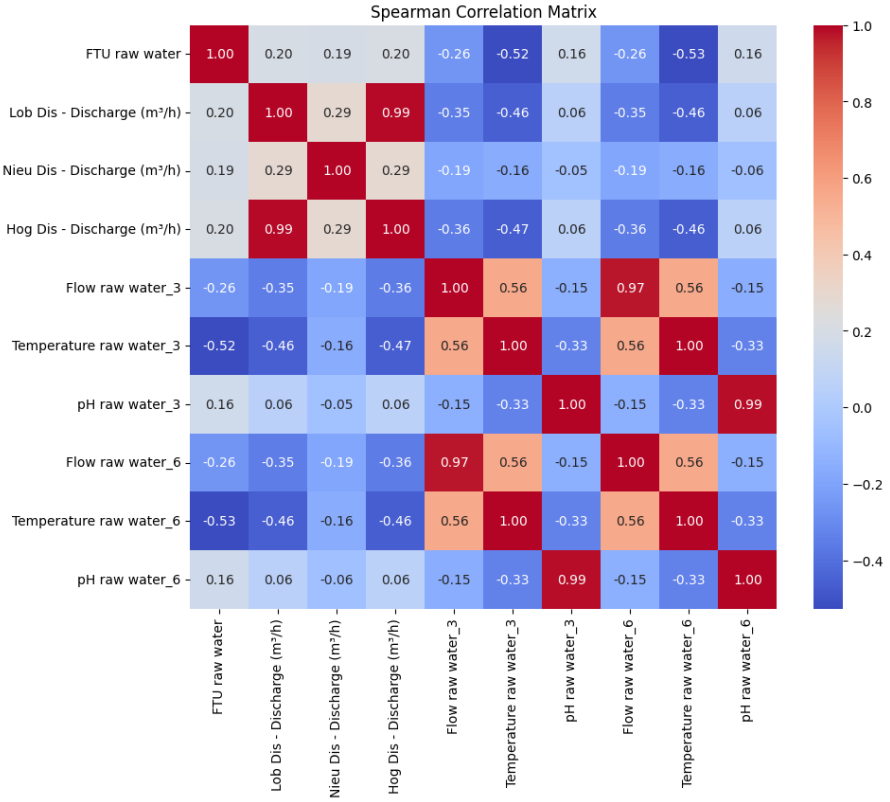


Figure 44: Spearman correlation between the inputs and turbidity

The results of this analysis suggest that there is a small influence of the pH in the turbidity of the raw water. Moreover, the significant correlation between the discharge flows in the 2 main river stations could have been a factor that influenced the training of the predictive model. Thus, it was decided that the model should be run also using 3 variables, the discharge in Lobith station, the intake water flow and the temperature.

Predictive model results and output

The LSTM predictive model was tested 4 different times, and the results are presented in the following table. The results indicate that besides the data unbalancing, the model was able to achieve highly accurate results. The best model was model number 3 with a forecasting horizon of 3 hours ahead when all 6 predictors were used as inputs. However, the prediction accuracy still remained high when the forecasting horizon was 6 hours ahead and when fewer but more important predictors were used. This finding also indicates the direct relationship between temperature changes and increased turbidity in the river water. This finding should be further explored in the following year of this project.

Table 22: Performance metrics of the turbidity predictive model

Model number	Variables	Predictive horizon	Precision	Recall	F1-Score
1	6 (all the parameters)	6	0.83	0.91	0.86
2	3 (Lobith discharge, water temperature, DWTP intake flow)	6	0.78	0.88	0.80
3	**6 (all the parameters)**	**3**	**0.85**	**0.93**	**0.88**
4	3 (Lobith discharge, water temperature, DWTP intake flow)	3	0.79	0.89	0.84

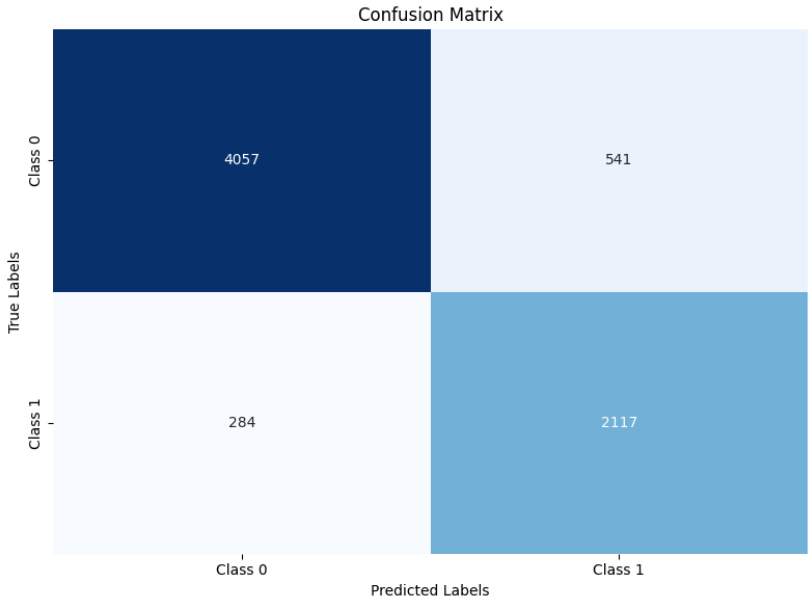


Figure 45: Correlation matrix of model number 4 (3 hours predictive horizon with three predictors)

6.3.5 Conclusion and next steps

A methodology was developed to predict high turbidity events in the drinking water treatment plant (DWTP) intake, combining correlation analysis for identifying influential factors with an LSTM model for event prediction. Currently, the model's predictive performance is good, but further improvements are necessary. The following actions are proposed to enhance the model:

- **Explore alternative balancing techniques:** The under-sampling technique produced satisfactory results as demonstrated above. However, future work should test other methods such as ADASYN class weight adjustment and balanced random forest and explore if these would improve the model performance.
- **Considering a regression approach** for predicting turbidity to provide continuous output of turbidity.
- **Rerun the model** with more data that will include the period after June 2024 and up to potentially September 2025.

- **Incorporate additional WQ parameters:** In the upstream river stations different water quality parameters are measured with different frequency. Conductivity is one of these parameters that should be investigated further as it is measured more frequently and may indicate early insights of the raw water quality condition.
- **Integrate rainfall data:** Rainfall heavily impacts the natural organic matter (NOM) in the rivers and consequently influences the turbidity of the water. Further work will be implemented to test the impact of rainfall events on the turbidity in the DWTP intake
- **Use of ensemble models:** Use of a combination of different fine-tuned LSTM models and combine their predicted outputs to generate a more robust result.
- Analyse grab samples and low-frequency water quality data: Further investigation in low-frequency data for data correlations that may indicate how turbidity is related to heavy metals and NOM concentration.
- **NOM and heavy metals soft sensor development:** The final goal of this work is to develop a soft sensor of heavy metals and NOM concentration. This step is heavily depended to the data availability for these water quality parameters and the predictive ability of the fine-tuned turbidity predictive model.

6.4 Soft Sensor 8 - 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 drinking water treatment plants (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 (NOM), and ultraviolet (UV) 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 (USEPA, 1991). 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 ozone concentration and CT values, using historical process data such as water temperature, flow rate, and ozone dosage. Three machine learning techniques have been tested and compared: Random Forest (Breiman, 2001) , Long Short-Term Memory (Hochreiter and Urgen Schmidhuber, 1997), networks, and a Physics-Informed Neural Network (PINN) (Karniadakis et al., 2021). 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.

A key advancement in this study is the implementation of a Physics-Informed Neural Network (PINN) for modeling the ozonation process. Unlike traditional machine learning models that rely solely on data, PINNs integrate physical laws—expressed through ordinary differential equations (ODEs)—directly into the learning framework. This hybrid approach allows the model to capture the complex interactions between key variables, such as ozone dosage, water temperature, and natural organic matter, while adhering to the established kinetics of ozone decomposition. Moreover, the ozonation prediction model is unique in its focus. By providing daily and in the future continuous estimates of ozone concentration and exposure (CT) of the ozonation process, which was not addressed in the DWTPs in the present time. This state-of-the-art approach bridges the gap between empirical methods and advanced predictive analytics, setting a new standard for ozonation process management in drinking water treatment. Finally, as this is a data-driven approach, the replicability of this approach is possible to other DWTPs that measure the same WQ parameters. The only requirement is the recalibration of the ozone ODE to site-specific conditions. This replicability allows a broader adoption of the model that consequently will ensure that other utilities will be able to achieve similar improvements in maintaining high disinfection efficiency with an optimized energy consumption.

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 from the rapid sand filtration outlet and the ozonation outlet, providing a snapshot of water quality parameters at these stages.

The study covers a 3.5-year period from January 2020 to June 2024, offering a robust dataset for analysis. A summary of the specific parameters used in this work is presented in the following Table 23.

Table 23: 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 chambers 1-5 - Lane 1 (mg/l)
- Water Temperature (°C) in the ozonation inlet	-Ozone concentration in chambers 1-5 - Lane 2 (mg/l)
- UV in the ozonation inlet (254nm)	-Ozone concentration in chambers 1-5 - Lane 3 (mg/l)
-UV in the ozonation outlet(254nm)	-Ozone concentration in chambers 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 Sample 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:

$c_{O3,i}$: the ozone concentration at chamber I (mg/l)

n: the total number of chambers

t: time in minutes

Combining the 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 modeling stages.

6.4.3 Material and methods

Soft Sensor development

The aim of this soft sensor is to predict the ozone concentration in the ozonation tank and estimate the CT daily. As the ozone concentration in each chamber was measured on average at 12:00 once per week, for the validation of this model, only the data of the sampling date and its previous one were used as inputs. The initial approach was to predict the ozone concentration in the last chamber of the ozonation tank. However, preliminary analysis showed that ozone concentration in the final two chambers consistently approached zero, suggesting that the third chamber being the last chamber that facilitates disinfection. In the following table, the inputs (sensor and grab samples variables) used for predicting the output at the day d at 12:00 are shown.

Table 24 The inputs and the outputs of the soft sensor.

Input Variables	Source	Output data
- Hourly ozone dosage at the day d for the hours 0:00 to 11:00 (12 data points)	Sensor Data from SCADA	C_{O3} / CT at day d and hour 12
- Hourly UV at the day d for the hours 0:00 to 11:00 (12 data points)	Sensor Data from SCADA	
-Hourly Turbidity at the day d for the hours 0:00 to 11:00 (12 data points)	Sensor Data from SCADA	
- Hourly water flow at the day d for the hours 0:00 to 11:00 (12 data points)	Sensor Data from SCADA	
-Hourly temperature at the day d for the hours 0:00 to 11:00 (12 data points)	Sensor Data from SCADA	
-Flow rate at the day d for the hours 0:00 to 11:00 (12 data points)	Calculated using sensor data	
-Average ozone dosage at the day d-1 (1 data point)	Calculated using sensor data	
- Average UV at the day d-1 (1 data point)	Calculated using sensor data	
-Average temperature at day d-1 (1 data point)	Calculated using sensor data	
-Average water flow at day d-1 (1 data point)	Grab Samples dataset	
-UV in the inlet at day d (1 data point)	Grab samples dataset	
-UV in the outlet at day d (1 data point)		

Machine learning algorithm selection

For the development of the soft sensor three different data-driven approaches were employed and evaluated: Random Forest (RF), Long Short-Term Memory (LSTM), and a Physics-Informed Neural Network (PINN).

Random Forrest (RF)

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 daily ozone concentrations and CT values based on selected input features, such as water temperature, flow rate, pH, UVA, and ozone dosage.

- **Hyperparameter Optimization**: Grid search was applied to optimize key hyperparameters of the RF model, including 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.

Long Short-Term Memory (LSTM)

The LSTM neural network was used to capture the sequential dependencies in the time series data for the ozonation process. LSTM is well-suited for time-dependent data, as it maintains information over longer time periods through a gated memory mechanism.

- **Model Architecture**: The LSTM model was designed with an input layer, one or more hidden LSTM layers, and a dense output layer. The model receives a sequence of past time steps as input and outputs a single predicted value for the ozone concentration or CT. We set the past time steps to be equal to 4 hours.
- **Hyperparameter Optimization**: Grid search was employed to optimize the LSTM hyperparameters, including the number of LSTM units (number of hidden layers), learning rate, batch size, and dropout rate. The optimal combination of hyperparameters was determined based on MSE values from cross-validation.

Physics informed Neural Network (PINN)

The Physics-Informed Neural Network (PINN) was developed to integrate physical knowledge of the ozonation process with the data-driven neural network approach. 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:

$$\frac{d_{C_{O_3}}}{dt} = -k_{UVA}(UV - UV_o)Y - k_{O_3}c_{O_3} \quad \text{Eq. 30}$$

Where:

K_{UVA}: the UVA decay rate (min^{-1})

UV: The UV in the water (m^{-1})

UV_o: The UV after the completion of the ozonation process (m^{-1})

Y: the yield for ozone consumed per UV decrease (mg/l/ m^{-1})

K_{o3}: the first order kinetic of ozone decomposition (min^{-1})

C_{o3}: the concentration of ozone (mg/l).

This ODE was used in the work of (Van Der Helm et al., 2008) to describe the ozone decomposition during the ozonation process in two steps the rapid ozone consumption step and the rather slow decay step. The rapid consumption is related to the reactions between ozone and NOM and the second part regarding the UV reaction is related to the slow decay part.

As regards the CT the above calculation was formed as follows:

$$\frac{dC}{dt} = \sum_{i}^{n} \frac{d_{C_{O_3}}}{dt} * t = \left(-k_{UVA}(UV - UV_o)Y - k_{O_3}c_{O_3}\right) * t \quad \text{Eq. 31}$$

- **Model Structure**: The PINN incorporates both measured data and the governing ODE. A feedforward neural network serves as the data-driven component, while the ODE regularizes the model output by penalizing deviations from the ozone decomposition dynamics.
- **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 (ODE loss component). The latter penalizes deviations from the

ODE 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 ODE loss function equally.

- **K constants estimation**: The constants estimations were calculated using the least square method for calibration in the training dataset.

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:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(pred_i - obs_i)^2 \quad \text{Eq. 32}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(pred_i - obs_i)^2} \quad \text{Eq. 33}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(pred_i - obs_i)^2}{\sum_{i=1}^{n}(pred_i - obs_{mean})^2} \quad \text{Eq. 34}$$

Where:

$pred_i$: the predicted value for observation i

obs_i: the actual value for observation i

n: the nmber of samples

obs_{mean}: the mean of the observed values

6.4.4 Results

The model was trained using 80% of the dataset and then it was tested in the remaining 20% which remained unseen. All 3 algorithms were tested to understand which one performs better for this dataset. Moreover, a further feature selection analysis was conducted for the identification of the most important variables that could reduce the overall model inputs, and consequently the computational requirements, without affecting the model performance. Two feature selection algorithms were used, the Pearson correlation and the RF feature importance. The results are presented in the rest of the session, firstly for the ozone concentration model and then for the CT.

Estimation of Ozone concentration

All models were initially run using all the input variables. Following these results, in the feature selection process both cross correlation and RF feature importance indicated that 4 were the key variables for the prediction of the ozone, hourly water temperature, daily average water temperature, daily average water dosage and daily average ozone flow. Further training of the RF and the PINN model was then conducted using only these 4 parameters. All the results are presented in the following table 25.

Table 25: Performance metrics of the model in the testing dataset for the ozone estimation model.

Model	Best hyperparameters	Variables	MSE	RMSE	R^2
RF	Max depth:2, min samples leaf 1, min samples per split: 2, number of trees: 200	All	0.008	0.087	0.64
LSTM	lstm units:2, lstm 1 neurons: 256, lstm 2 neurons: 50 dropout rate 0.3	All	0.024	0.171	0.43
PINN	Hidden layer size: 64, learning rate:0.001, num epochs: 1000	All	0.014	0.12	0.52
RF	Max depth:2, min samples leaf 1, min samples per split: 2, number of trees: 200	hourly water temperature, daily average water temperature, daily average ozone dosage and daily average water flow	**0.0044**	**0.0661**	**0.8**
PINN	Hidden layer size: 64, learning rate:0.001, num epochs: 1000		0.011	0.105	0.55

RF was the outperforming model in all the performing metrics and the LSTM model was the worst performing model. Moreover, after the feature selection process, RF performance was improved as the table shows. PINN was better performing than LSTM. In the following plot the predicted vs observed outputs of the best RF model is presented.

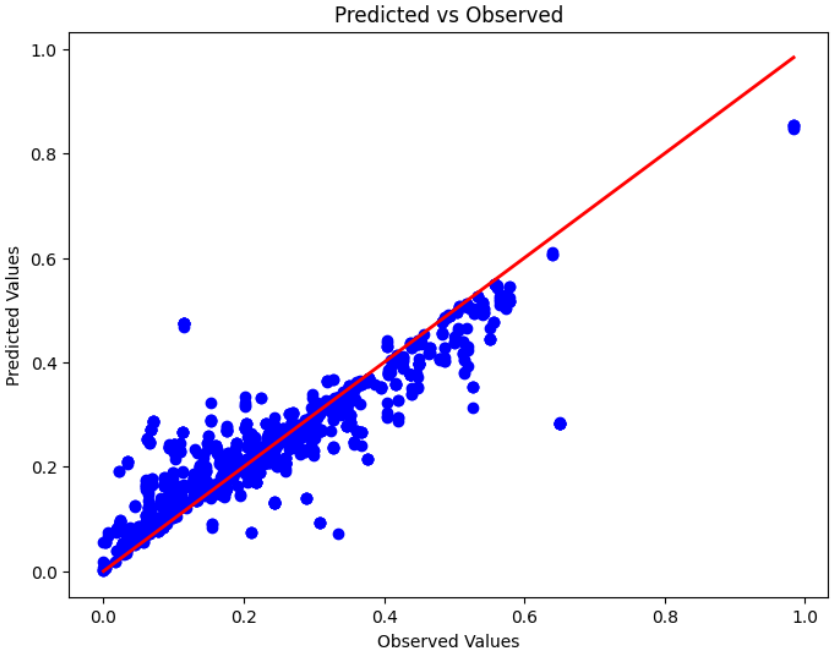


Figure 46: Predicted vs Observed ozone concentration values using RF model with 4 input variables

Estimation of CT

The same process as the one followed for the ozone estimation was followed for the CT estimation model. In the feature selection process for this model the key variables were daily and hourly water temperature, the hourly UV effluent in the rapid sand filtration, the daily ozone dosage and the ozone generation rate. A further training of the RF and the PINN model was then conducted using only these 5 parameters. All the results are presented in the following table 26.

Table 26: Performance metrics of the model in the testing dataset for the CT estimation model.

Model	Best hyperparameters	Variables	MSE	RMSE	R^2
RF	Max depth:2, min samples leaf 1, min samples per split: 2, number of trees: 200	All	0.06	0.245	0.801
LSTM	lstm units:2, lstm 1 neurons: 256, lstm 2 neurons: 50 dropout rate 0.3	All	0.1089	0.33	0.21
PINN	Hidden layer size: 64, learning rate:0.001, num epochs: 1000	All	0.095	0.31	0.45
RF	Max depth:2, min samples leaf 1, min samples per split: 2, number of trees: 200	daily and hourly water temperature, hourly UV effluent in the rapid sand filtration daily ozone dosage and hourly ozone generation rate	**0.052**	**0.229**	**0.826**
PINN	Hidden layer size: 64, learning rate:0.001, num epochs: 1000		0.082	0.28	0.52

RF was again the best performing model and as above its performance was improved after the feature selection process. PINN had better results than LSTM and has also improved its performance after the feature selection process. In the following plot the predicted vs observed outputs of the best RF model is presented.

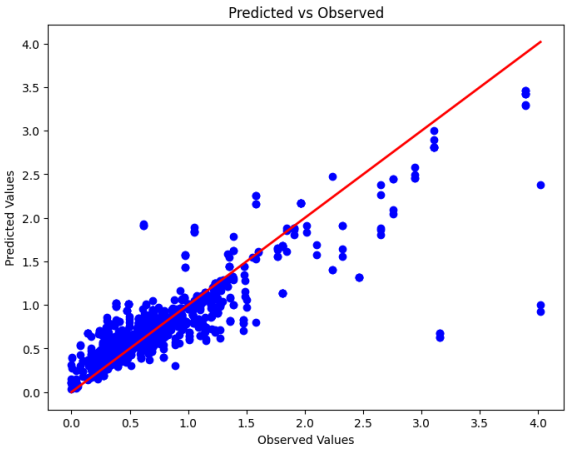


Figure 47: Predicted vs Observed CT values using RF model with 5 input variables

6.4.5 Conclusion and next steps

A data-driven soft sensor model has been developed for daily estimation of ozone concentration in the ozonation chamber and overall ozone exposure. After testing the model with various variable sets and three machine learning algorithms—Random Forest (RF), Long Short-Term Memory (LSTM), and Physics-Informed Neural Networks (PINNs)—the results were evaluated using RMSE, MSE, and R^2 as performance metrics. Among these, RF achieved the best performance (for ozone R^2=0.8 and for CT R^2=0.826), while LSTM yielded moderate results. The PINN model, though initially underperforming compared to RF, demonstrated potential for improvement.

Feature selection highlighted water temperature and ozone dosage as key variables influencing the model's accuracy, as expected. Moving forward in the third year of the project, several steps are planned to further refine the soft sensor's performance and enhance the PINN model. Upon finalizing the model, additional work will focus on developing soft sensors for bromate and assimilable organic carbon (AOC) prediction. The next steps include:

- **Incorporating hourly time-series data** from the previous day to explore potential impacts on model performance.
- **Applying advanced feature selection methods,** such as sensitivity analysis and unsupervised learning techniques (e.g., self-organizing maps), to identify and quantify the primary variables affecting ozone concentration and exposure.
- **Conducting sensitivity analyses on the PINN's total loss function,** exploring model enhancements through weighted adjustments of data loss and ODE loss within the loss function.
- **Further optimizing the k rate constants** in the PINN model, either through Adam-based optimization or a hybrid approach to optimize both constants.
- **Developing a bromate soft sensor** following the completion of the CT. This new soft sensor will utilize another ODE described in Van der Helm et al. (2008).
- **Developing an AOC soft sensor** following the completion of the ozone model. This new soft sensor will utilize another ODE described in Van der Helm et al. (2008).
- **Rerun the model** with more data that will include the period after June 2024 and up to potentially September 2025.

7. Upscaling and European added value (EAV)

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.

8. References

Ahearn, D.S., Sheibley, R.W., Dahlgren, R.A., Keller, K.E., 2004. Temporal dynamics of stream water chemistry in the last free-flowing river draining the western Sierra Nevada, California. J Hydrol (Amst) 295, 47–63. https://doi.org/10.1016/j.jhydrol.2004.02.016

Ayushi, Buttar, P.K., 2024. Satellite Imagery Analysis for Crop Type Segmentation Using U-Net Architecture. Procedia Comput Sci 235, 3418–3427. https://doi.org/10.1016/J.PROCS.2024.04.322

Balch, J.K., St. Denis, L.A., Mahood, A.L., Mietkiewicz, N.P., Williams, T.M., McGlinchy, J., Cook, M.C., 2020. 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 12, 3498. https://doi.org/10.3390/RS12213498

Barraza-Moraga, F., Alcayaga, H., Pizarro, A., Félez-Bernal, J., Urrutia, R., 2022. Estimation of Chlorophyll-a Concentrations in Lanalhue Lake Using Sentinel-2 MSI Satellite Images. Remote Sens (Basel) 14, 5647. https://doi.org/10.3390/RS14225647/S1

Bennett, T.H., Peters, J.C., 2004a. 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 104. https://doi.org/10.1061/40517(2000)149

Bennett, T.H., Peters, J.C., 2004b. 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 104, 1–10. https://doi.org/10.1061/40517(2000)149

Bitew, M.M., Gebremichael, M., 2011. Evaluation of satellite rainfall products through hydrologic simulation in a fully distributed hydrologic model. Water Resour Res 47. https://doi.org/10.1029/2010WR009917

Bitew, M.M., Gebremichael, M., Ghebremichael, L.T., Bayissa, Y.A., 2012. Evaluation of High-Resolution Satellite Rainfall Products through Streamflow Simulation in a Hydrological Modeling of a Small Mountainous Watershed in Ethiopia. J Hydrometeorol 13, 338–350. https://doi.org/10.1175/2011JHM1292.1

Breiman, L., 2001. Random Forests.

Bresciani, M., Giardino, C., Stroppiana, D., Dessena, M.A., Buscarinu, P., Cabras, L., Schenk, K., Heege, T., Bernet, H., Bazdanis, G., Tzimas, A., 2019. Monitoring water quality in two dammed reservoirs from multispectral satellite data. Eur J Remote Sens 52, 113–122. https://doi.org/10.1080/22797254.2019.1686956

Buma, W.G., Lee, S. Il, 2020. Evaluation of Sentinel-2 and Landsat 8 Images for Estimating Chlorophyll-a Concentrations in Lake Chad, Africa. Remote Sensing 2020, Vol. 12, Page 2437 12, 2437. https://doi.org/10.3390/RS12152437

Caballero, I., Roca, M., Santos-echeandía, J., Bernárdez, P., Navarro, G., 2022. 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) 14, 2744. https://doi.org/10.3390/RS14122744/S1

Cao, Z., Ma, R., Duan, H., Pahlevan, N., Melack, J., Shen, M., Xue, K., 2020. A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes. Remote Sens Environ 248. https://doi.org/10.1016/j.rse.2020.111974

Clark, C.O., 1945. Storage and the Unit Hydrograph. Transactions of the American Society of Civil Engineers 110, 1419–1446. https://doi.org/10.1061/TACEAT.0005800

Clough, S.A., Shephard, M.W., Mlawer, E.J., Delamere, J.S., Iacono, M.J., Cady-Pereira, K., Boukabara, S., Brown, P.D., 2005a. Atmospheric radiative transfer modeling: a summary of the AER codes. J Quant Spectrosc Radiat Transf 91, 233–244. https://doi.org/10.1016/J.JQSRT.2004.05.058

Clough, S.A., Shephard, M.W., Mlawer, E.J., Delamere, J.S., Iacono, M.J., Cady-Pereira, K., Boukabara, S., Brown, P.D., 2005b. Atmospheric radiative transfer modeling: a summary of the AER codes. J Quant Spectrosc Radiat Transf 91, 233–244. https://doi.org/10.1016/J.JQSRT.2004.05.058

De Sousa, L.M., Poggio, L., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., n.d. SoilGrids 2.0: producing quality-assessed soil information for the globe. https://doi.org/10.5194/soil-2020-65

Ding, X., Zhang, J., Jiang, G., Zhang, S., 2017. 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 9, 465. https://doi.org/10.3390/W9070465

Djerioui, M., Bouamar, M., Ladjal, M., Zerguine, A., 2019. Chlorine Soft Sensor Based on Extreme Learning Machine for Water Quality Monitoring. Arab J Sci Eng 44, 2033–2044. https://doi.org/10.1007/s13369-018-3253-8

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sens Environ 120, 25–36. https://doi.org/10.1016/J.RSE.2011.11.026

D'Silva, M.S., Anil, A.C., Naik, R.K., D'Costa, P.M., 2012. Algal blooms: A perspective from the coasts of India. Natural Hazards 63, 1225–1253. https://doi.org/10.1007/S11069-012-0190-9/FIGURES/3

Elhag, M., Gitas, I., Othman, A., Bahrawi, J., Gikas, P., 2019. Assessment of water quality parameters using temporal remote sensing spectral reflectance in arid environments, Saudi Arabia. Water (Switzerland) 11. https://doi.org/10.3390/w11030556

Fadillah Rahmat, R., Fadly Syahputra, M., Silvi Lydia, M., 2016. Real Time Monitoring System for Water Pollution in Lake Toba, in: 6 International Conference on Informatics and Computing (ICIC).

Feldman, A., 2016. Hydrologic modeling system HEC-HMS: technical reference manual.

Feng, J., Chen, H., Zhang, H., Li, Z., Yu, Y., Zhang, Y., Bilal, M., Qiu, Z., 2020. Turbidity estimation from GOCI satellite data in the turbid estuaries of China's coast. Remote Sens (Basel) 12, 1–20. https://doi.org/10.3390/rs12223770

Gensemer, R.W., Gondek, J.C., Rodriquez, P.H., Arbildua, J.J., Stubblefield, W.A., Cardwell, A.S., Santore, R.C., Ryan, A.C., Adams, W.J., Nordheim, E., 2018. Evaluating the effects of pH, hardness, and dissolved organic carbon on the toxicity of aluminum to freshwater aquatic organisms under circumneutral conditions. Environ Toxicol Chem 37, 49–60. https://doi.org/10.1002/etc.3920

Gower, J., Hu, C., Borstad, G., King, S., 2006. Ocean color satellites show extensive lines of floating sargassum in the gulf of Mexico. IEEE Transactions on Geoscience and Remote Sensing 44, 3619–3625. https://doi.org/10.1109/TGRS.2006.882258

Gower, J., King, S., 2008. Satellite Images Show the Movement of Floating Sargassum in the Gulf of Mexico and Atlantic Ocean. Nature Precedings 2008 1–1. https://doi.org/10.1038/npre.2008.1894.1

Hafeez, S., Wong, M.S., Ho, H.C., Nazeer, M., Nichol, J., Abbas, S., Tang, D., Lee, K.H., Pun, L., 2019. Comparison of machine learning algorithms for retrieval of water quality indicators in case-ii waters: A case study of hong kong. Remote Sens (Basel) 11. https://doi.org/10.3390/rs11060617

Hamid, A., Bhat, S.U., Jehangir, A., 2020. Local determinants influencing stream water quality. Appl Water Sci. https://doi.org/10.1007/s13201-019-1043-4

He, Y., Wang, X., Xu, F., 2022. How reliable is chlorophyll-a as algae proxy in lake environments? New insights from the perspective of n-alkanes. Science of the Total Environment 836. https://doi.org/10.1016/j.scitotenv.2022.155700

Herman, M.R., Nejadhashemi, A.P., Abouali, M., Hernandez-Suarez, J.S., Daneshvar, F., Zhang, Z., Anderson, M.C., Sadeghi, A.M., Hain, C.R., Sharifi, A., 2018. Evaluating the role of evapotranspiration remote sensing data in improving hydrological modeling predictability. J Hydrol (Amst) 556, 39–49. https://doi.org/10.1016/J.JHYDROL.2017.11.009

Hochreiter, S., Urgen Schmidhuber, J.¨, 1997. Long Short-Term Memory. Neural Comput 9.

Hu, C., 2009a. A novel ocean color index to detect floating algae in the global oceans. Remote Sens Environ 113, 2118–2129. https://doi.org/10.1016/J.RSE.2009.05.012

Hu, C., 2009b. A novel ocean color index to detect floating algae in the global oceans. Remote Sens Environ 113, 2118–2129. https://doi.org/10.1016/J.RSE.2009.05.012

Hydrologic Modeling System HEC-HMS Technical Reference Manual CPD-74B, 2000.

Immerzeel, W.W., Droogers, P., 2008. Calibration of a distributed hydrological model based on satellite evapotranspiration. J Hydrol (Amst) 349, 411–424. https://doi.org/10.1016/J.JHYDROL.2007.11.017

Jia, Y., Lan, H., Jia, R., Fu, K., Su, Z., 2024. ENHANCED U-NET ALGORITHM FOR TYPICAL CROP CLASSIFICATION USING GF-6 WFV REMOTE SENSING IMAGES. Engenharia Agrícola 44, e20230110. https://doi.org/10.1590/1809-4430-ENG.AGRIC.V44E20230110/2024

Jiang, L., Wu, H., Tao, J., Kimball, J.S., Alfieri, L., Chen, X., 2020. Satellite-Based Evapotranspiration in Hydrological Model Calibration. Remote Sensing 2020, Vol. 12, Page 428 12, 428. https://doi.org/10.3390/RS12030428

Johansen, R., Beck, R., Nowosad, J., Nietch, C., Xu, M., Shu, S., Yang, B., Liu, H., Emery, E., Reif, M., Harwood, J., Young, J., Macke, D., Martin, M., Stillings, G., Stumpf, R., Su, H., 2018. Evaluating the portability of satellite derived chlorophyll-a algorithms for temperate inland lakes using airborne hyperspectral imagery and dense surface observations. Harmful Algae 76, 35–46. https://doi.org/10.1016/J.HAL.2018.05.001

Juntunen, P., Liukkonen, M., Lehtola, M.J., Hiltunen, Y., 2013. Dynamic soft sensors for detecting factors affecting turbidity in drinking water. Journal of Hydroinformatics 15, 416–426. https://doi.org/10.2166/hydro.2012.052

Kapalanga, T.S., Hoko, Z., Gumindoga, W., Chikwiramakomo, L., 2021. Remote-sensing-based algorithms for water quality monitoring in olushandja dam, north-central namibia. Water Supply 21, 1878–1894. https://doi.org/10.2166/ws.2020.290

Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. Nature Reviews Physics. https://doi.org/10.1038/s42254-021-00314-5

Katsouras, G., Chalaris, M., Tsalas, N., Dosis, A., Samios, S., Lytras, E., Papadopoulos, K., Synodinou, A., n.d. INTEGRATED ECOSYSTEM ECOLOGY (CHLOROPHYLL-A) OF EYDAP'S RESERVOIRS PROFILES BY USING ROBOTIC BOATS | Enhanced Reader. 5th International Conference 'Water Resources and Wetlands,' Tulca, Romania, pp. 202–210.

Kaufman, Y.J., Justice, C.O., Flynn, L.P., Kendall, J.D., Prins, E.M., Giglio, L., Ward, D.E., Menzel, W.P., Setzer, A.W., 1998. Potential global fire monitoring from EOS-MODIS. Journal of Geophysical Research: Atmospheres 103, 32215–32238. https://doi.org/10.1029/98JD01644

Kaufman, Y.J., Justice2, C., Flynn, L., Kendall, J., Prins, E., Ward, D.E., Menzel, P., Setzer, A., n.d. Kaufman et al., Remote Sensing of Fires from EOS-MODIS Monitoring Global Fires from EOS-MODIS.

Khandelwal, A., Karpatne, A., Marlier, M.E., Kim, J., Lettenmaier, D.P., Kumar, V., 2017. An approach for global monitoring of surface water extent variations in reservoirs using MODIS data. Remote Sens Environ 202, 113–128. https://doi.org/10.1016/J.RSE.2017.05.039

Kirana, A.P., Ariyanto, R., Ririd, A.R.T.H., Amalia, E.L., 2020. Agricultural drought monitoring based on vegetation health index in East Java Indonesia using MODIS Satellite Data. IOP Conf Ser Mater Sci Eng 732, 012063. https://doi.org/10.1088/1757-899X/732/1/012063

Kirkham, M.B., 2014. Infiltration. Principles of Soil and Plant Water Relations 201–227. https://doi.org/10.1016/B978-0-12-420022-7.00013-6

Kite, G.W., Droogers, P., 2000. Comparing evapotranspiration estimates from satellites, hydrological models and field data. J Hydrol (Amst) 229, 3–18. https://doi.org/10.1016/S0022-1694(99)00195-X

Kloos, S., Yuan, Y., Castelli, M., Menzel, A., 2021. Agricultural drought detection with modis based vegetation health indices in southeast germany. Remote Sens (Basel) 13, 3907. https://doi.org/10.3390/RS13193907/S1

Kohonen Teuvo, 1990. The Self-organizing Map. Proceedings of the IEEE 78.

Leip, A., Billen, G., Garnier, J., Grizzetti, B., Lassaletta, L., Reis, S., Simpson, D., Sutton, M.A., De Vries, W., Weiss, F., Westhoek, H., 2015. Impacts of European livestock production: Nitrogen, sulphur, phosphorus and greenhouse gas emissions, land-use, water eutrophication and biodiversity. Environmental Research Letters 10. https://doi.org/10.1088/1748-9326/10/11/115004

Ling, F., Li, Xinyan, Foody, G.M., Boyd, D., Ge, Y., Li, Xiaodong, Du, Y., 2020. Monitoring surface water area variations of reservoirs using daily MODIS images by exploring sub-pixel information. ISPRS Journal of Photogrammetry and Remote Sensing 168, 141–152. https://doi.org/10.1016/J.ISPRSJPRS.2020.08.008

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

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

Luo, J., Ni, G., Zhang, Y., Wang, K., Shen, M., Cao, Z., Qi, T., Xiao, Q., Qiu, Y., Cai, Y., Duan, H., 2023. A new technique for quantifying algal bloom, floating/emergent and submerged vegetation in eutrophic shallow lakes using Landsat imagery. Remote Sens Environ 287, 113480. https://doi.org/10.1016/J.RSE.2023.113480

Meyers, G., Kapelan, Z., Keedwell, E., Randall-Smith, M., 2016. Short-term Forecasting of Turbidity in a UK Water Distribution System, in: Procedia Engineering. Elsevier Ltd, pp. 1140–1147. https://doi.org/10.1016/j.proeng.2016.07.534

MODIS Chlorophyll Fluorescence (MOD 20), n.d.

Mohammed, H., Hameed, I.A., Seidu, R., 2018. Comparative predictive modelling of the occurrence of faecal indicator bacteria in a drinking water source in Norway. Science of the Total Environment 628–629, 1178–1190. https://doi.org/10.1016/j.scitotenv.2018.02.140

Moses, W.J., Gitelson, A.A., Berdnikov, S., Povazhnyy, V., 2009. Satellite estimation of chlorophyll-a concentration using the red and NIR bands of MERISThe azov sea case study. IEEE Geoscience and Remote Sensing Letters 6, 845–849. https://doi.org/10.1109/LGRS.2009.2026657

Mu, M., Li, Y., Bi, S., Lyu, H., Xu, J., Lei, S., Miao, S., Zeng, S., Zheng, Z., Du, C., 2021a. Prediction of algal bloom occurrence based on the naive Bayesian model considering satellite image pixel differences. Ecol Indic 124, 107416. https://doi.org/10.1016/J.ECOLIND.2021.107416

Mu, M., Li, Y., Bi, S., Lyu, H., Xu, J., Lei, S., Miao, S., Zeng, S., Zheng, Z., Du, C., 2021b. Prediction of algal bloom occurrence based on the naive Bayesian model considering satellite image pixel differences. Ecol Indic 124, 107416. https://doi.org/10.1016/J.ECOLIND.2021.107416

Najafzadeh, M., Basirian, S., 2023. Evaluation of River Water Quality Index Using Remote Sensing and Artificial Intelligence Models. Remote Sens (Basel) 15, 2359. https://doi.org/10.3390/RS15092359/S1

Najafzadeh, M., Homaei, F., Farhadi, H., 2021. Reliability assessment of water quality index based on guidelines of national sanitation foundation in natural streams: integration of remote sensing and data-driven models. Artif Intell Rev 54, 4619–4651. https://doi.org/10.1007/S10462-021-10007-1/FIGURES/6

Nas, B., Karabork, H., Ekercin, S., Berktay, A., 2009. Mapping chlorophyll-a through in-situ measurements and Terra ASTER satellite data. Environ Monit Assess 157, 375–382. https://doi.org/10.1007/S10661-008-0542-9/METRICS

NRC Soil Science Division, U., n.d. Soil Survey Manual 2017; Chapter 4.

Ogashawara, I., Kiel, C., Jechow, A., Kohnert, K., Ruhtz, T., Grossart, H.P., Hölker, F., Nejstgaard, J.C., Berger, S.A., Wollrab, S., 2021. 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 13, 1542. https://doi.org/10.3390/RS13081542

Olivetti, D., Cicerelli, R., Martinez, J.M., Almeida, T., Casari, R., Borges, H., Roig, H., 2023. Comparing Unmanned Aerial Multispectral and Hyperspectral Imagery for Harmful Algal Bloom Monitoring in Artificial Ponds Used for Fish Farming. Drones 7. https://doi.org/10.3390/drones7070410

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, 2406–2426. https://doi.org/10.2166/hydro.2023.122

Ovakoglou, G., Alexandridis, T.K., Crisman, T.L., Skoulikaris, C., Vergos, G.S., 2016. 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 51, 37–46. https://doi.org/10.1016/J.JAG.2016.04.007

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

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

Pereira, J., Saraiva, F., 2020. A Comparative Analysis of Unbalanced Data Handling Techniques for Machine Learning Algorithms to Electricity Theft Detection. 2020 IEEE Congress on Evolutionary Computation, CEC 2020 - Conference Proceedings. https://doi.org/10.1109/CEC48606.2020.9185822

Pereira, L.S., 1998. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56.

Qi, L., Hu, C., Mikelsons, K., Wang, M., Lance, V., Sun, S., Barnes, B.B., Zhao, J., Van der Zande, D., 2020. In search of floating algae and other organisms in global oceans and lakes. Remote Sens Environ 239, 111659. https://doi.org/10.1016/J.RSE.2020.111659

Raffuse, S.M., McCarthy, M.C., Craig, K.J., Dewinter, J.L., Jumbam, L.K., Fruin, S., James Gauderman, W., Lurmann, F.W., 2013. High-resolution MODIS aerosol retrieval during wildfire events in California for use in exposure assessment. Journal of Geophysical Research: Atmospheres 118, 11,242-11,255. https://doi.org/10.1002/JGRD.50862

Rai, R., Sahu, C.K., 2020. Driven by Data or Derived through Physics? A Review of Hybrid Physics Guided Machine Learning Techniques with Cyber-Physical System (CPS) Focus. IEEE Access. https://doi.org/10.1109/ACCESS.2020.2987324

Rout, N., Mishra, D., Mallick, M.K., 2018. Handling imbalanced data: A survey. Advances in Intelligent Systems and Computing 628, 431–443. https://doi.org/10.1007/978-981-10-5272-9_39/TABLES/6

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

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

Schindler, D.W., 2012. The dilemma of controlling cultural eutrophication of lakes. Proceedings of the Royal Society B: Biological Sciences 279, 4322–4333. https://doi.org/10.1098/RSPB.2012.1032

Schindler, D.W., Vallentyne, J.R., 2008. The Algal Bowl. The Algal Bowl. https://doi.org/10.1515/9781772126341/HTML

Seyyedi, A., Bohlouli, M., Oskoee, S.N., 2023. Machine Learning and Physics: A Survey of Integrated Models. ACM Comput Surv 56. https://doi.org/10.1145/3611383

Shi, W., Wang, M., 2009. Green macroalgae blooms in the Yellow Sea during the spring and summer of 2008. J Geophys Res Oceans 114. https://doi.org/10.1029/2009JC005513

Silva, G.M.E., Campos, D.F., Brasil, J.A.T., Tremblay, M., Mendiondo, E.M., Ghiglieno, F., 2022. Advances in Technological Research for Online and In Situ Water Quality Monitoring—A Review. Sustainability (Switzerland). https://doi.org/10.3390/su14095059

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

Spelmen, V.S., Porkodi, R., 2018. A Review on Handling Imbalanced Data. Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies, ICCTCT 2018. https://doi.org/10.1109/ICCTCT.2018.8551020

Tian, R., Chen, J., Sun, X., Li, D., Liu, C., Weng, H., 2018. Algae explosive growth mechanism enabling weather-like forecast of harmful algal blooms. Scientific Reports 2018 8:1 8, 1–7. https://doi.org/10.1038/s41598-018-28104-7

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

USEPA, 1991. Guidance Manual for Compliance with the filtration and Disinfection requirements for public water systems using surface water bodies.

Usman, M., Liedl, R., Shahid, M.A., Abbas, A., 2015. Land use/land cover classification and its change detection using multi-temporal MODIS NDVI data. Journal of Geographical Sciences 25, 1479–1506. https://doi.org/10.1007/S11442-015-1247-Y/METRICS

Van Der Helm, A.W.C., Rietveld, L.C., Baars, E.T., Smeets, P.W.M.H., Van Dijk, J.C., 2008. 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 57, 435–449. https://doi.org/10.2166/aqua.2008.089

Véliz-Chávez, C., Mastachi-Loza, C.A., González-Sosa, E., Becerril-Piña, R., Ramos-Salinas, N.M., 2014. Canopy Storage Implications on Interception Loss Modeling. Am J Plant Sci 05, 3032–3048. https://doi.org/10.4236/AJPS.2014.520320

Wang, Y ; ;, Yang, L ;, Yue, J ;, Li, Q ;, Lin, J ;, Liu, Q, Guan, G., Wang, Yonggui, Yang, Ling, Yue, Jinzhao, Li, Qiang, Lin, Jianyun, Liu, Qiang, 2022. 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 19, 11818. https://doi.org/10.3390/IJERPH191811818

Weng, W., Zhu, X., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. IEEE Access 9, 16591–16603. https://doi.org/10.1109/ACCESS.2021.3053408

Yang, F., 2012. Water Leak Detection and Localization Using Multi-Sensor Data Fusion. California State University.

Yin, H., Pflugmacher, D., Kennedy, R.E., Sulla-Menashe, D., Hostert, P., 2014. Mapping annual land use and land cover changes using MODIS time series. IEEE J Sel Top Appl Earth Obs Remote Sens 7, 3421–3427. https://doi.org/10.1109/JSTARS.2014.2348411

Zhan, X., Sohlberg, R.A., Townshend, J.R.G., DiMiceli, C., Carroll, M.L., Eastman, J.C., Hansen, M.C., DeFries, R.S., 2002. Detection of land cover changes using MODIS 250 m data. Remote Sens Environ 83, 336–350. https://doi.org/10.1016/S0034-4257(02)00081-0

Zhang, Y., Shi, K., Liu, J., Deng, J., Qin, B., Zhu, G., Zhou, Y., 2016. 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 569–570, 1517–1529. https://doi.org/10.1016/J.SCITOTENV.2016.06.244

Zhao, H., Yang, S., Wang, Z., Zhou, X., Luo, Y., Wu, L., 2015. Evaluating the suitability of TRMM satellite rainfall data for hydrological simulation using a distributed hydrological model in the Weihe River catchment in China. Journal of Geographical Sciences 25, 177–195. https://doi.org/10.1007/S11442-015-1161-3/METRICS

Zheng, L., Wu, M., Cui, Y., Tian, L., Yang, P., Zhao, L., Xue, M., Liu, J., 2022. What causes the great green tide disaster in the South Yellow Sea of China in 2021? Ecol Indic 140, 108988. https://doi.org/10.1016/J.ECOLIND.2022.108988