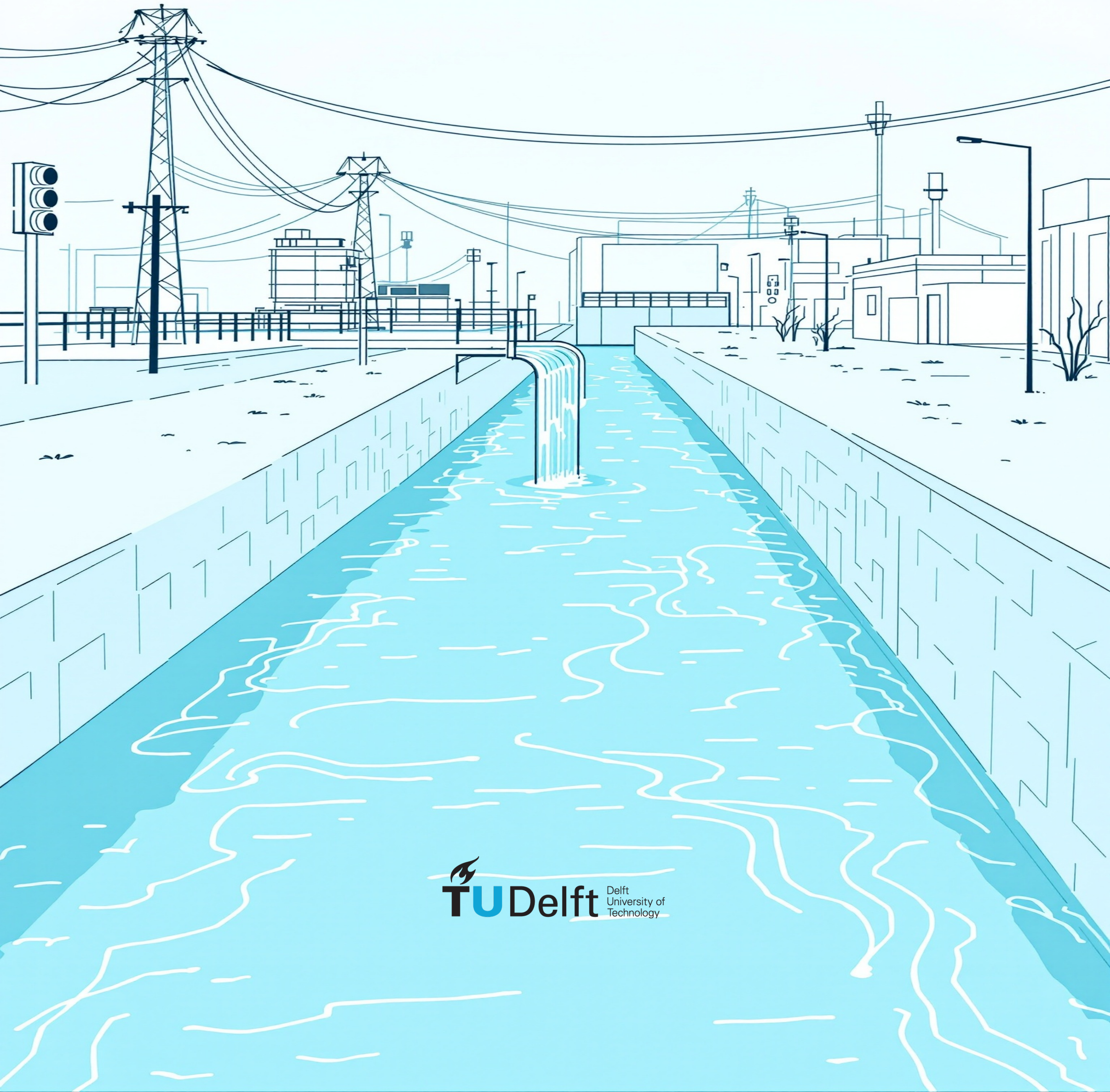
SHORT-TERM WATER QUALITY FORECAST IN THE DWTP INLET

Hang Long



SHORT-TERM WATER QUALITY FORECAST IN THE DWTP INLET

Hang Long

5743702

MSc Thesis
Department of Environmental Engineering

Thesis committee:

Prof.dr.ir. L.C. Rietveld (chair)
Dr. G. Kyritsakas
Dr.ir. R.W. Hut

Delft University of Technology
To be defended publicly at 08:30 on Friday, August 29, 2025

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Preface

I have always been fascinated by new technologies and their potential to solve real-world problems. This curiosity led me to choose the Data Science and Artificial Intelligence course, after which I was eager to gain hands-on experience applying these tools to practical challenges. I am especially grateful to my supervisor, Luuk, for offering me this fascinating thesis topic on short-term turbidity prediction at the Lekkanaal DWTP. Your guidance made it possible for me to explore an area where machine learning meets environmental engineering, and I have learned far more than I initially imagined.

I also want to sincerely thank Greg for supporting me throughout the writing process. I greatly enjoyed our weekly discussions in the office exploring new ideas. These conversations were a highlight of my thesis journey.

I am grateful to Rolf for reminding me during my Green Light Meeting that this thesis is not merely a mathematical problem of fitting data, but an engineering problem that requires a solid understanding of environmental engineering. On a personal level, I came to see it as a “life problem” as well—a challenge that pushed me to confront my weaknesses and grow into a better version of myself.

Last year, while working on this thesis, the Nobel Prize was awarded to Geoffrey Hinton and Demis Hassabis for their contribution in AI. I thought, “Great—I just need to do slightly better than they did.” Luckily, my supervisors reminded me that graduating with slightly worse work is entirely acceptable. So here I am, with a work that may not be perfect, but it is something I can be proud of—a testament to what I have learned, explored, and accomplished throughout this journey.

Hang Long
Delft, August 2025

Abstract

Access to clean and safe drinking water is essential for public health and sustainable development. Drinking water treatment plants (DWTPs) ensure water quality, but fluctuating raw water characteristics, particularly turbidity, challenge efficient coagulation and dosing control. Traditional strategies like jar tests and feed-backward control are limited by delayed results, making timely adjustments difficult. Short-term predictive tools based on machine learning (ML) offer a solution by forecasting water quality variations and enabling proactive control. This study develops several different ML models for short-term turbidity prediction at the Lekkanaal DWTP, addressing three questions: (1) which parameters influence turbidity, (2) which feature combinations yield optimal predictions, and (3) how far in advance turbidity can be reliably forecasted.

Historical water quality and hydrological data were collected from Waternet, KNMI, and Rijkswaterstaat, followed by preprocessing for reliable inputs. Candidate features were selected using Spearman correlation and Self-Organizing Maps (SOMs). Three regression models—AutoRegressive Integrated Moving Average (ARIMA), Random Forest (RF), and Long Short-Term Memory (LSTM)—were trained for different horizons, and feature importance analyzed using greedy selection and visualization tools. An RF classifier evaluated the feasibility of predicting peak turbidity events.

Results showed turbidity was driven by hydrological and physicochemical factors. Upstream discharge and turbidity strongly correlated with local measurements, highlighting the Lek River as a primary contributor, while EC and temperature showed negative correlations, reflecting dilution and seasonal sediment mobilization. SOMs confirmed high turbidity coincides with northward flows from the Lek River into the Amsterdam-Rhine Canal.

Feature analysis indicated univariate models using recent sensor_turbidity outperformed multivariate models; additional features introduced noise. The last three hours of turbidity dominated predictions across ARIMA, RF, and LSTM.

All models provided reliable short-term forecasts, with RF outperforming ARIMA and LSTM for 3- and 6-hour horizons. Extreme peaks were systematically underestimated, and RF classification detected fewer than 16% of peak events. Short-term forecasts up to six hours are feasible, but high-magnitude events remain challenging, emphasizing the need for enhanced monitoring and tailored strategies.

Keywords: Turbidity prediction, Drinking Water Treatment Plant, Short-term forecasting, Machine learning, Random Forest, LSTM, ARIMA, Water quality monitoring

Table of Contents

1 Introduction . . . 1

2 Literature review . . . 5

2.1 Coagulation and turbidity . . . 5

Coagulation and turbidity . . . 5

2.2 Traditional water quality prediction strategies . . . 7

Traditional water quality prediction strategies . . . 7

2.3 Machine learning model . . . 9

Machine learning model . . . 9

2.4 Feature Selection Techniques . . . 11

Feature Selection Techniques . . . 11

3 Material and Methodology . . . 15

3.1 Research area . . . 15

Research area . . . 15

3.2 Data collection . . . 17

Data collection . . . 17

3.3 Methodology . . . 18

Methodology . . . 18

3.3.1 Data quality assessment . . . 18

3.3.2 Runoff Time Analysis . . . 20

3.3.3 Feature Selection . . . 21

3.3.4 Model Training . . . 22

3.3.5 Evaluation Metrics . . . 26

3.3.6 Feature Attribution Analysis . . . 27

4 Results . . . 29

4.1 Data quality assessment . . . 29

Data quality assessment . . . 29

4.2 Runoff time analysis . . . 33

Runoff time analysis . . . 33

4.3 Feature selection . . . 36

Feature selection . . . 36

4.3.1 Spearman correlation . . . 36

4.3.2 SOMs . . . 39

4.4 Model comparison per forecast horizon . . . 41

Model Comparison per Forecast Horizon 41
4.4.1 1-hour Ahead Forecast 42
4.4.2 3-hour Ahead Forecast 46
4.4.3 6-hour Ahead Forecast 48
4.5 Feature attribution analysis 50
Feature attribution analysis 50
4.6 RF classification model 55
RF classification model 55

5 Conclusion **59**
5.1 Conclusion 59

References **61**

6 Appendix **65**
6.1 Complete Spearman Cross_Correlation Result 65
6.2 Model Configurations 68

List of Figures

Figure 1 Factors affecting coagulation and flocculation processes [9] 5

Figure 2 Phases through the coagulation–flocculation process [9] 6

Figure 3 Typical Examples of ANN, RNN, and LSTM [26] Red arrows: Hidden State. Green arrows: Forget Gate, Input Gate, and Output Gate . . . 9

Figure 4 Typical examples of ANN RNN and LSTM [28] 11

Figure 5 Pre-purification of source water at Lekkanaal [3] 15

Figure 6 Annual average discharge of Rhine and Maas 2000-2011 16

Figure 7 Locations of the data sources 17

Figure 8 TimeSeriesSplit 24

Figure 9 Data quality overview 29

Figure 10 Timestamp error in sensor_turbidity data 29

Figure 11 Timestamp error in Hag_dis data 30

Figure 12 Timestamp error in Nieu_EC data 30

Figure 13 Timestamp error in Lob_pH data 30

Figure 14 Nieu_wt with and without invalid data 31

Figure 15 Single point outliners of Lob_EC 32

Figure 16 Abnormal period of sensor_turbidity 32

Figure 17 Peaks of Lob_dis and Hag_dis 34

Figure 18 Nieu_dis 35

Figure 19 Spearman Cross_Correlation between sensor_turbidity and discharge . 35

Figure 20 Spearman Cross_Correlation between sensor_turbidity and EC 38

Figure 21 Spearman Cross_Correlation between sensor_turbidity and Lobith pH & turbidity 38

Figure 22 Spearman Cross_Correlation between sensor_turbidity and weather data 39

Figure 23 SOMs of 10 most relevant parameters 40

Figure 24 Categorical SOMs 41

Figure 25 Comparison of model performance across three prediction horizons (t+1, t+3, t+6) 42

Figure 26 ARIMA predictions (Forecast window = 1 hour) 42

Figure 27 Predictions of three model on test period (Forecast window = 1 hour) . 43

Figure 28 Scatterplot of predictions versus true values of three models (Forecast window = 1 hour) 44

Figure 29 Error density of three models (Forecast window = 1 hour) 45

Figure 30 Predictions of three model on test period (Forecast window = 3 hour) . 46

Figure 31 Scatterplot of predictions versus true values of three models (Forecast window = 3 hour) 47
Figure 32 Error density of three models (Forecast window = 3 hour) 47
Figure 33 Predictions of three model on test period (Forecast window = 6 hour) . 48
Figure 34 Scatterplot of predictions versus true values of three models (Forecast window = 6 hour) 49
Figure 35 Error density of three models (Forecast window = 6 hour) 50
Figure 36 Feature attribution analysis experiments 51
Figure 37 RF model greedy feature selection results (Forecast window = 1 hour) 51
Figure 38 RF feature importance 52
Figure 39 RF SHAP value 53
Figure 40 LSTM model greedy feature selection results (Forecast window = 1 hour) 54
Figure 41 LSTM feature importance 54
Figure 42 RF classifier experiments 55
Figure 43 Confusion Matrix 56
Figure 44 Spearman Cross_Correlation between sensor_turbidity and discharge . 65
Figure 45 Spearman Cross_Correlation between sensor_turbidity and water level 65
Figure 46 Spearman Cross_Correlation between sensor_turbidity and water temperature 66
Figure 47 Spearman Cross_Correlation between sensor_turbidity and EC 66
Figure 48 Spearman Cross_Correlation between sensor_turbidity and Lob_pH and Lob_turbidity 66
Figure 49 Spearman Cross_Correlation between sensor_turbidity and weather data 67

List of Tables

Table 1 Data source overview . . . 18
Table 2 Parameter Boundaries . . . 19
Table 3 Data overview . . . 33
Table 4 Average travel time from Lobith by peak match method. . . . 34
Table 5 Spearman correlation overview . . . 37
Table 6 Forecast results of three models (Forecast window = 1 hour) . . . 43
Table 7 Forecast results of three models (Forecast window = 3 hour) . . . 46
Table 8 Forecast results of three models (Forecast window = 6 hour) . . . 48
Table 9 Maximum importance of different features . . . 52
Table 10 Forecast performance metrics for single- and multi-feature LSTM models. 55
Table 11 Performance metrics for single- and multi-feature RF classifiers in peak event detection. . . . 56
Table 12 Model Configurations of section 4.4 . . . 68
Table 13 Model Configurations of section 4.5 . . . 68
Table 14 Model Configurations of section 4.6 . . . 68

1 Introduction

Background

Access to clean and safe drinking water is a fundamental necessity for human health and well-being. Recognized as one of the 17 Sustainable Development Goals, clean water and sanitation are critical not only for health but also for addressing poverty reduction, food security, education, ecosystems, peace, and human rights. However, reliable access to clean water is increasingly under pressure due to population growth and climate change, which threaten water availability and quality in many regions. [1] DWTPs constitute critical components within drinking water systems, ensuring that water provided to consumers is devoid of harmful microorganisms and hazardous substances.

In Amsterdam and surrounding areas, more than 1,4 million customers use tap water from the water-cycle company Waternet, with an average daily water consumption of 141 liter per person in 2021 [2]. There are two sources of drinking water in Amsterdam. The main source is from river Rhine, where water is pumped from the Lekkanaal near Nieuwegein. The other source is from seepage water rising from the ground in the Bethunepolder. [3].

Problem description

In a DWTP, treatment usually starts with coagulation. Coagulation effectively promotes the reduction of turbidity, heavy metals and pathogens. As a complex chemical process, the efficacy of coagulation is influenced by many factors, such as raw water turbidity, temperature, pH, mixing intensity and the dose of coagulant [4]. Achieving the optimal balance is challenging since there is no comprehensive or universally recognized mathematical description of the process [5].

Traditionally, operators of DWTPs determine the correct coagulant dose using jar tests, a method that simulates coagulation under different chemical conditions. However, these tests are costly, time-consuming, and offer delayed results, making it difficult to respond quickly to sudden water rapidly changes [6]. This delay hinders the timely implementation of corrective measures, which are essential for maintaining the efficacy of the treatment process and ensuring the safety of the drinking water supply.

Therefore, to cope with the fluctuating nature of influent water quality, more optimization methods for dosage control based on online-measurable water quality parameters are emerging [5]. These methods use physical sensors monitoring water quality parameters as feedback to control the dosing. The three common approaches are Feed-forward control, Feed-backward, and Feedforward-feedback control [4].

Feed-forward control continuously monitors raw water quality and respond to measured disturbances by adjusting coagulant dosing accordingly. However, its success heavily depends on the availability of robust models that establish the relationship between raw water characteristics and the ideal coagulant dose. Without extensive and representative data collection, this approach may struggle with accuracy, particularly during periods of high variability in water quality [4].

Feed-backward control, in contrast, uses treated water quality (after coagulation and/or whole treatment process) as the basis for dosing adjustments. While this approach provides a direct connection to the desired water quality, it is limited by system delays, which prevent real-time control. Such delays can lead to suboptimal dosing, including under-dosing or over-dosing, especially during seasons when raw water quality changes rapidly [4].

Feedforward-feedback control combines the two strategies but still suffers from their limitations, such as reliance on extensive data for calibration and the impact of system delays [4].

To further address the difficulties posed by fluctuating water quality and system delay, predictive tools that provide short-term forecasts of water quality are recognized as promising solutions. In particular, predicting water quality variations at the DWTP intake can serve as an important input for feed-forward control models, enabling more proactive and adaptive coagulant dosing strategies. By anticipating changes before they affect treatment performance, predictive models can provide DWTP operators with sufficient response time to prepare for and mitigate potential issues, ensuring more efficient and reliable treatment processes [7].

Building on this need for predictive tools, ML has emerged as a powerful approach for short-term water quality forecasting. Unlike traditional approaches, ML techniques harness large datasets and computational power to model complex, non-linear relationships that were once difficult to capture. Some ML architectures like LSTM networks have shown particular promise in time series forecasting tasks, as they are explicitly designed to capture temporal dependencies in sequential data [8]. In this context, developing an ML model for the short-term prediction of turbidity is particularly valuable. Turbidity is a key water quality parameter influencing the coagulation process, and sudden fluctuations can lead to dosing inefficiencies, higher chemical consumption, or even risks to treated water quality. A reliable turbidity forecasting model could therefore serve as an essential component of feed-forward control in DWTP operations, enabling operators to make more timely and precise dosing adjustments and ultimately improving treatment efficiency and reliability.

Research objectives and research questions

Although previous studies have explored predictive tools for water quality, most of them have been conducted at the daily scale, which limits their usefulness for operational decision-making in DWTPs where water quality can fluctuate within hours. As a result, there is still limited understanding of how influent turbidity can be forecasted on the short-term (hourly) scale, which is essential for timely coagulant dosing adjustments.

In addition, the Lekkanaal DWTP presents a particularly challenging case. The intake is located at the junction between the Lek River and the Amsterdam-Rhine Canal, two water bodies with distinctly different flow regimes. While the Rhine is dominated by natural river discharge dynamics, the canal is primarily regulated by navigation demands and hydraulic structures. The interaction of these contrasting flow patterns results in complex and highly variable conditions at the DWTP intake, which complicates the prediction of turbidity levels. This setting makes it especially relevant to examine which factors most strongly influence turbidity variability at this location.

To address these gaps, this thesis aims to develop an ML model for the short-term prediction of turbidity, a key water quality parameter influencing the coagulation process at the Lekkanaal DWTP. By leveraging historical data and advanced ML techniques, the study seeks to enhance turbidity forecasting accuracy and provide actionable insights for water treatment optimization. To guide this investigation, the following research questions will be addressed:

- What parameters influence turbidity levels at the Lekkanaal DWTP?

- Which combination of features yields the most effective turbidity predictions?

- How far in advance can turbidity levels be reliably predicted?

2 Literature review

2.1 Coagulation and turbidity

Coagulation is an essential process for the removal of suspended and colloidal material from raw water. Figure 1 shows the main factors affecting the coagulation and flocculation processes in water treatment.

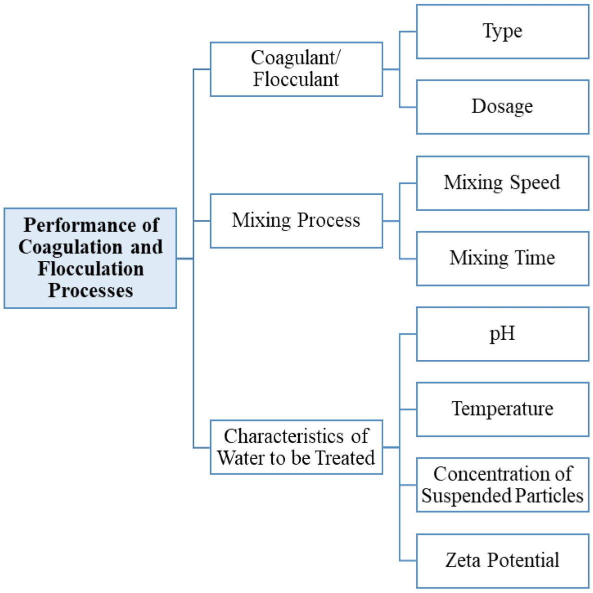

Performance of Coagulation and Flocculation Processes
Coagulant/ Flocculant
Type
Dosage
Mixing Process
Mixing Speed
Mixing Time
Characteristics of Water to be Treated
pH
Temperature
Concentration of Suspended Particles
Zeta Potential

Figure 1: Factors affecting coagulation and flocculation processes [9]

At Nieuwegein DWTP, ferric chloride has been chosen as the coagulant. Mixing is critical in coagulation (and flocculation), with fast mixing aiding coagulant-particle interaction and microfloc formation, and slow mixing supporting large floc aggregation. Improper mixing speeds or durations can reduce efficiency or cause floc breakdown. These processes are highly pH-dependent, as pH influences the formation of polymeric species from metal-based coagulants. Temperature changes also impact aggregation, while the concentration and characteristics of suspended particles, including zeta potential, play significant roles. Coagulants generally perform better in high-turbidity water, but extreme turbidity levels may require coagulant aids for effective treatment. An optimum dosage of coagulants/flocculants can be obtained by plotting the measured turbidity (or any other pollutant parameter) versus the applied dosage (Figure 2).

There is a sensor continuously monitoring four parameters of the raw water from Lekkanaal, including water temperature, discharge, pH and turbidity. Discharge is monitored to ensure sufficient drinking water supply and the other three water quality parameters are closely relevant for coagulation. In which, turbidity is an important indicator of the physical and chemical characteristics of aquatic systems.

Turbidity is a physical property of fluids, reflecting the presence of suspended particles

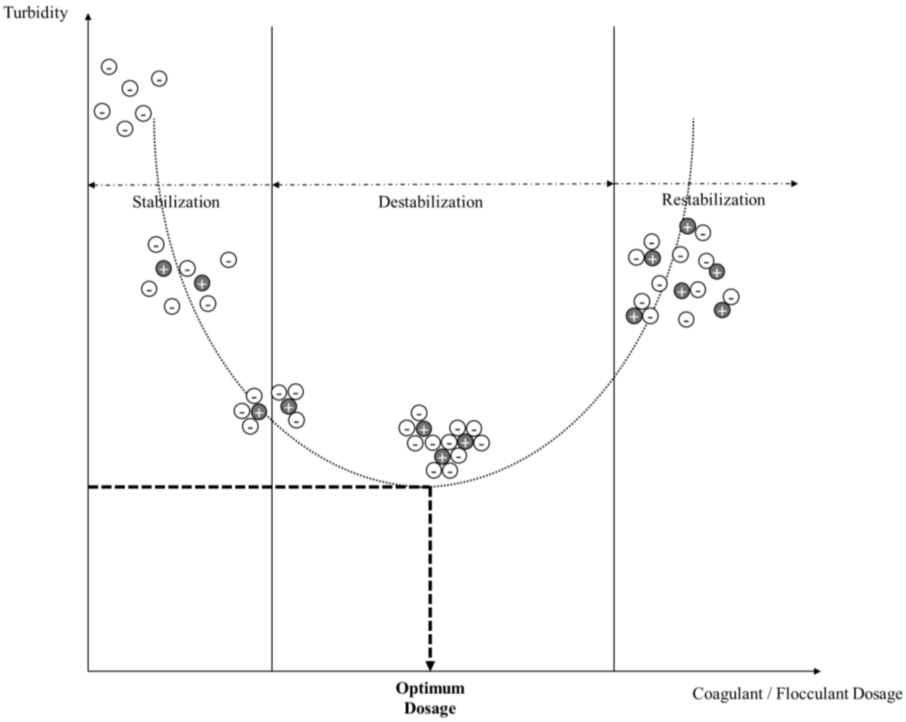


Figure 2: Phases through the coagulation–flocculation process [9]

in water [10]. The typical unites for turbidity are nephelometric turbidity unit (NTU), formazine turbidity unit (FTU), and formazine attenuation unit (FAU). Essentially, these are the same in value, but different methods are used to determine these values. In this work, turbidity is determined using a turbidimeter in FTU, which works by shining infrared light through a water sample and measuring the amount of light scattered by the particles at a 90-degree angle. The higher the degree of light scattering, the higher the turbidity and FTU value [10]. According to the World Health Organization's standards, the turbidity of drinking water should be below 1 FTU before disinfection, otherwise, the effectiveness of chlorination significantly decreases. In areas where fewer resources are available, the turbidity should be below 5 FTU.

Coagulation can be described as a processing to remove turbidity, color and natural organic matter from raw water [11]. Turbidity is considered to be as an indirect parameter in determining the removal efficiency of coagulation [11, 12]. Also, turbidity can be used to determine the optimal dosage as shown in Figure 2.

The sources of turbidity are diverse, ranging from natural processes such as erosion, particle transport and sedimentation to anthropogenic activities like urban runoff, industrial discharges, pesticides and microplastics [11]. Turbidity in rivers is primarily influenced by environmental factors such as precipitation, discharge, temperature, and upstream turbidity levels.

Many studies have shown correlation between turbidity and precipitation. Theoretically, rainfall intensity may directly induce erosion resulting in high turbidity runoff flowing into rivers.[13].

Similarly, river discharge directly affects sediment transport, with high discharge events resuspending particles and increasing turbidity levels [14]. Temperature impacts the physical and chemical properties of water, influencing sediment settling rates and biological activity, which can alter turbidity [15]. Lastly, upstream turbidity contributes to the downstream conditions, as suspended particles and pollutants transported from upstream sources accumulate and affect water quality further downstream. Monitoring these factors is essential for understanding and managing turbidity in DWTP, ensuring effective coagulation and water treatment.

Rainfall can increase turbidity by causing soil erosion and surface runoff. Intense rainfall washes loose sediments, organic matter, and pollutants into rivers, leading to higher turbidity levels [13]. During storms, the rapid influx of suspended particles can significantly reduce water clarity [14].

Higher discharge rates, often due to heavy rainfall or snowmelt, increase the force of water flow, which can resuspend settled sediments from the riverbed. This leads to elevated turbidity levels, as more particles remain suspended in the water column [14]. Conversely, lower discharge allows sediments to settle, reducing turbidity.

Water temperature influences the growth of algae and microbial activity, which can contribute to turbidity [15]. Some studies also report that temperature may affect the photoelectric components inside turbidity sensors, such as optical emitting and receiving electronic components. As temperature increases, the sensor's performance may be altered, potentially leading to a decrease in measured turbidity [16].

Turbidity levels in a river are influenced by conditions upstream. If upstream sources contribute large amounts of suspended particles—whether from natural sediment transport, industrial discharges, or urban runoff—these particles will be carried downstream, maintaining or increasing turbidity levels [13, 14].

2.2 Traditional water quality prediction strategies

Traditionally, water quality prediction is typically approached through two main strategies: mechanistic models and data-driven models. While both strategies have been widely used, they each have significant limitations that make them less suitable for addressing the complexities of modern water quality management.

Mechanistic models are built on physical, chemical, and biological principles to simulate the behavior of water bodies. These models solve differential equations to describe processes like pollutant transport, chemical reactions, and biological activities. Examples of mechanistic models include the QUAL2K model and the WASP (Water Quality Analysis Simulation

Program) [17, 18], both of which are used to simulate water quality dynamics in various aquatic environments.

However, mechanistic models have several critical drawbacks:

- They are computationally expensive, especially when simulating large, complex water systems, requiring substantial computational resources. WASP uses compartment modeling approach and simulates spatial and temporal conservation of mass implementing a finite-difference equation for each compartment or segment [17, 19].

- Detailed input data is often required. QUAL2K requires flow and concentrations for headwater, discharges and withdrawals; reach segment lengths, elevations, hydraulic geometry and weather data parameters. WASP requires simulation and output control, model segmentation, advective and dispersive transport, boundary concentrations, point and diffuse source waste loads, initial concentrations and kinetic parameters, constants and time functions [17, 18]. These may not always be readily available, particularly in regions with limited monitoring infrastructure.

- These models are based on strict assumptions about the behavior of water systems (e.g., linear relationships and predefined processes), making them less effective in capturing non-linear dynamics and uncertainties that are common in real-world water systems. QUAL2K can only simulates the main stem of a river and does not simulate branches of the river system [17, 18].

- Mechanistic models are often rigid and cannot easily incorporate new data or adjust to unexpected changes in environmental conditions, making them less adaptable to evolving water quality challenges. QUAL2K can only incorporate basic climate and meteorological data and has limited dynamic simulation capabilities. [19, 20].

These limitations make mechanistic models less practical for use in environments where data is limited, or the water quality dynamics are complex and non-linear.

Data-driven models, such as Multiple Linear Regression (MLR) and ARIMA, offer an alternative to mechanistic models [21]. These models predict water quality based on historical data, without requiring detailed knowledge of the underlying physical processes. For instance, MLR uses linear relationships between predictors (e.g., temperature, rainfall) and a dependent variable (e.g., water quality parameters), while ARIMA models temporal dependencies in time series data to forecast future values.

Despite their advantages, data-driven models also have notable limitations:

- MLR assumes linear relationships between variables, which may not capture complex, non-linear interactions present in water quality data. This makes MLR unsuitable for capturing more intricate dependencies [22].

- External factors (such as sudden changes in land use, pollutant discharge, or extreme weather events) that affect water quality may not be captured in the historical data, limiting the models' ability to predict water quality under dynamic conditions.

The main drawbacks of traditional data-driven models lie in their inability to adapt to new data quickly, their reliance on specific assumptions (like linearity), and their need for substantial, high-quality datasets [23, 24].

2.3 Machine learning model

In recent years, the field of water quality prediction has seen significant advancements with the application of ML models. Unlike traditional models, ML approaches are capable of capturing complex patterns and relationships within large datasets without the need for explicit physical modeling. ML models are particularly useful in addressing the limitations of traditional methods, such as handling non-linear dynamics, adapting to new data, and making predictions in data-scarce environments. Among the various ML models, Artificial Neural Networks (ANNs) has emerged as powerful tools for water quality prediction [21].

ANNs

ANNs are a class of ML models inspired by the structure and function of the human brain. ANN was firstly introduced in the 1970s, but its importance wasn't fully appreciated until a famous paper by David Rumelhart, Geoffrey Hinton and Ronald Williams [25]. ANNs consist of an input layer, hidden layers, and an output layer as shown in Figure 3 (a). The interconnected nodes (neurons) in each layers can process input data and passing the output to the next layer. These networks can learn complex, non-linear relationships between input variables and output predictions by adjusting the weights of the connections through training.

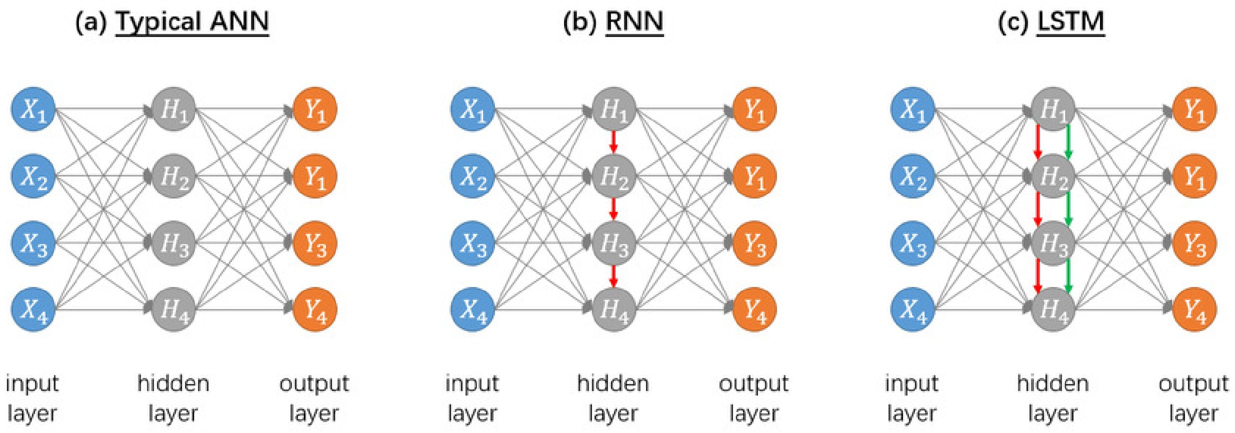


Figure 3: Typical Examples of ANN, RNN, and LSTM [26]
Red arrows: Hidden State. Green arrows: Forget Gate, Input Gate, and Output Gate

ANNs are widely recognized due to their flexibility and ability to model complex patterns in the data. From 2008, the use of the ANN technique has been booming in the field of water quality prediction. ANNs have been used to predict various key water quality parameters by many researchers. Most used target paramters are dissolved oxygen (DO), biological dissolved oxygen (BOD), Chemical Oxygen Demand (COD), water temperature and pH [4, 8, 21].

However, while ANNs are effective at capturing non-linear relationships, they may struggle to model temporal dependencies in time series data. Water quality parameters often exhibit time-dependent behaviors, such as seasonal trends and periodic fluctuations, which are not easily captured by standard feedforward ANN models. This limitation led to the development of more specialized models, such as Recurrent Neural Networks (RNNs).

RNNs

Unlike traditional ANNs, RNNs are designed to handle sequential data, making them ideal for time series forecasting tasks. RNNs have the unique ability to maintain a "memory" of previous time steps through feedback connections between neurons. It takes the output of the previous moment as the input of the next moment to affect the weights at the next moment as shown in Figure 3 (b). This allows RNNs to model temporal dependencies and predict future values based on past observations [24].

In the context of water quality prediction, RNNs have been applied to forecast time-dependent water quality parameters by learning from past measurements and environmental factors. These models are particularly useful for predicting changes in water quality over time, especially in cases where long-term dependencies exist, such as seasonal variations and trends driven by climatic events.

However, traditional RNNs suffer from issues like vanishing gradients during training, which can make it difficult for the network to learn long-term dependencies. To address this limitation, the more advanced LSTM networks were introduced [27].

LSTM Networks

LSTM networks, a specialized type of RNN, are specifically designed to capture long-term dependencies in sequential data. LSTMs address the vanishing gradient problem by using a unique gating mechanism that regulates the flow of information through the network as shown in Figure 3 (c) in green arrows. This enables LSTMs to maintain and update long-term memory, making them highly effective for time series prediction tasks where long-term relationships exist [24]. The relation between ANN, RNN and LSTM can be seen in Figure 3.

In water quality prediction, LSTMs have shown great promise in forecasting parameters that exhibit long-term trends and periodic fluctuations, such as nutrient levels, turbidity, and dissolved oxygen. LSTMs are particularly useful when dealing with time series data that includes seasonal patterns, trends, and cyclical fluctuations that traditional methods,

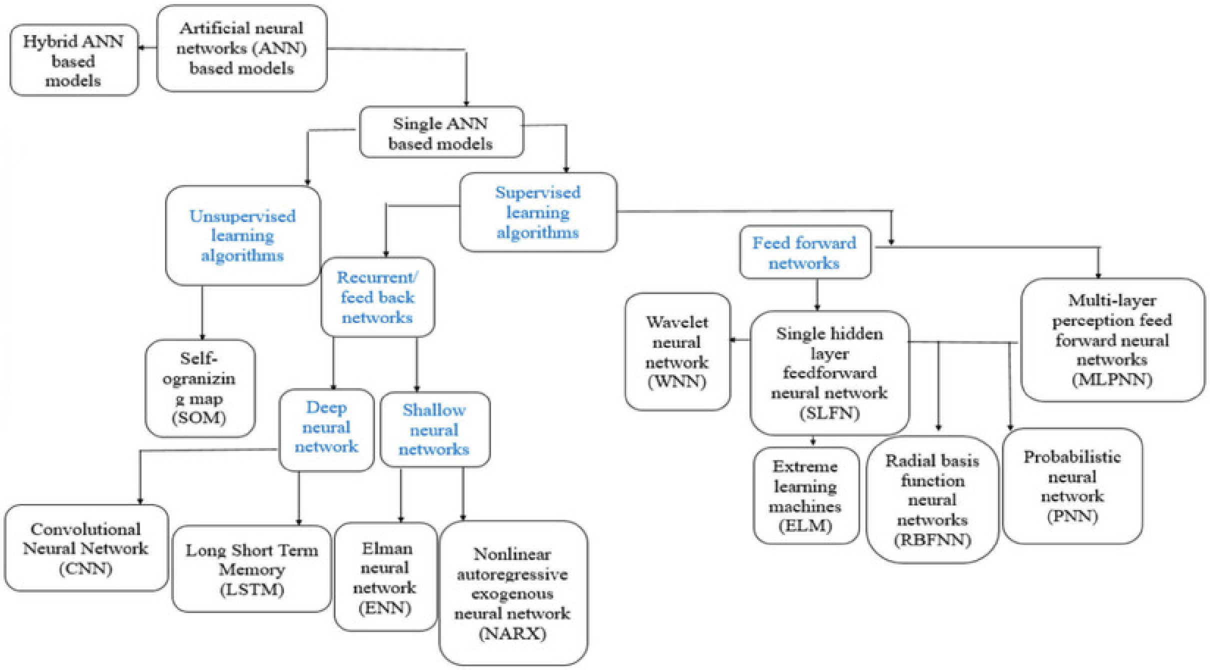

Hybrid ANN based models
Artificial neural networks (ANN) based models
Single ANN based models
Unsupervised learning algorithms
Supervised learning algorithms
Feed forward networks
Recurrent/ feed back networks
Self-ogranizin g map (SOM)
Deep neural network
Shallow neural networks
Wavelet neural network (WNN)
Single hidden layer feedforward neural network (SLFN)
Multi-layer perception feed forward neural networks (MLPNN)
Convolutional Neural Network (CNN)
Long Short Term Memory (LSTM)
Elman neural network (ENN)
Nonlinear autoregressive exogenous neural network (NARX)
Extreme learning machines (ELM)
Radial basis function neural networks (RBFNN)
Probabilistic neural network (PNN)

Figure 4: Typical examples of ANN RNN and LSTM [28]

like ARIMA or simple ANNs, may struggle to capture [24, 29].

The application of LSTMs in water quality prediction has been demonstrated in various studies, where they have been used to predict parameters like pH, temperature, and chemical oxygen demand (COD) by learning from historical time series data. LSTMs have the advantage of being able to incorporate past water quality data, meteorological conditions, and hydrological factors to generate accurate short- and long-term predictions [24, 29].

LSTMs have proven to be an effective tool in water quality prediction, offering superior performance in modeling temporal dependencies and non-linear relationships compared to traditional methods. Their ability to process sequential data makes them particularly suitable for applications in water systems where time-based patterns are crucial for accurate forecasting.

2.4 Feature Selection Techniques

Feature selection plays a pivotal role in ML applications for water quality prediction. It involves identifying the most relevant input variables that significantly influence the target outputs while reducing redundancy and irrelevant information. The benefit of applying feature selection techniques includes:

- Improved model accuracy: By removing irrelevant or redundant features, models are less prone to overfitting and can generalize better to unseen data.

- Increased computational efficiency: Reducing the feature set decreases training time and resource requirements.

- Enhanced interpretability: Focusing on the most important features aids in understanding the relationships and dynamics within water systems.

Broadly, feature selection techniques can be categorized into three main approaches [30]:

1. **Filter Methods**: These methods evaluate the statistical properties of each feature with respect to the target variable independently of the ML model. Examples include techniques like correlation analysis, mutual information, and variance thresholding. Filter methods are computationally efficient and are often used as a preprocessing step.

2. **Wrapper Methods**: These methods involve training and evaluating an ML model iteratively with different subsets of features to determine the combination that provides the best performance. Techniques such as recursive feature elimination (RFE) and forward or backward feature selection fall under this category. Although wrapper methods provide tailored results, they are computationally intensive, especially for large datasets.

3. **Embedded Methods**: These methods integrate feature selection directly into the ML model training process. Regularization techniques, such as Lasso (L1 regularization) and Elastic Net, are commonly used embedded methods that automatically penalize irrelevant or redundant features.

In this research, two filter-based techniques were employed for feature selection: correlation analysis and SOMs. These methods were selected for their complementary strengths in detecting linear and non-linear relationships in multivariate water quality data.

Correlation analysis is a classical filter method that evaluates the strength and direction of association between variables. There are three most commonly used correlation coefficients:

- **Pearson correlation** measures the strength of a linear relationship between two continuous variables. It assumes normally distributed data and is sensitive to outliers, making it less suitable when the relationship is non-linear or the data is skewed.

- **Spearman correlation** is a rank-based method that assesses how well the relationship between two variables can be described by a monotonic function. It does not require the assumption of normality and is more robust to outliers and non-linear trends.

- **Kendall's Tau** is also a rank-based measure but uses concordant and discordant pairs to assess association. It is generally more conservative than Spearman and provides better estimates for smaller datasets, though it is less commonly used due to higher computational cost.

In this study, Spearman correlation was chosen over other commonly used methods due to its suitability for capturing monotonic but non-linear relationships, which are often observed in environmental and hydrological datasets [31]. Features showing consistently low Spearman correlation were eliminated early in the training data preparation process. This preliminary step helped reduce dimensionality while preserving variables with strong predictive potential.

SOMs, first introduced by Kohonen [32], are a type of unsupervised artificial neural network that project high-dimensional input data onto a lower-dimensional (typically 2D) grid while preserving the topological relationships among the data points.

A SOM consists of two main layers:

- **Input layer:** Each input node corresponds to one feature or variable in the dataset (e.g., turbidity, pH, discharge).
- **Output layer (Kohonen map):** A two-dimensional grid of neurons (also called units), typically arranged in a rectangular or hexagonal layout. Each neuron is associated with a weight vector of the same dimension as the input data.

During training, the SOM algorithm organizes the neurons in the output grid such that similar input patterns are mapped to nearby neurons, preserving the topological relationships of the original data. The training process includes the following steps:

1. **Initialization:** Each neuron in the grid is initialized with a randomly generated weight vector.
2. **Input presentation:** A data point (i.e., an input vector) is selected and presented to the map.
3. **Best Matching Unit (BMU) selection:** The neuron whose weight vector is closest to the input vector—usually based on Euclidean distance—is identified as the BMU.
4. **Weight update:** The BMU and its neighboring neurons are updated to make their weights more similar to the input vector. The size of the neighborhood and the learning rate decrease over time.
5. **Iteration:** Steps 2 to 4 are repeated for multiple epochs, allowing the neurons to become specialized and organized based on the input data.

After training, similar inputs activate neurons located near each other on the map. This self-organization results in a structured, low-dimensional representation of the original high-dimensional data. The trained SOM can be visualized using different output maps that help interpret the structure and relationships within the data:

- **Component planes:** These visualizations show the distribution of each input variable across the map. Each component plane corresponds to a single variable and is typically color-coded (e.g., dark for low values, bright for high values). By comparing multiple component planes, one can detect patterns, correlations, and variable groupings. For example, if the regions of high turbidity and high discharge align spatially across their respective component planes, it may indicate a strong relationship between those variables.

- **U-matrix (Unified Distance Matrix):** The U-matrix displays the distance between the weight vectors of neighboring neurons. Areas with small distances (lighter colors) indicate clusters of similar inputs, while areas with large distances (darker colors) mark boundaries between different clusters. This makes the U-matrix a powerful tool for identifying natural groupings or cluster structures in the data.

Together, these visualization tools enable the user to explore complex datasets in an intuitive way, discover patterns, identify redundant or highly correlated features, and gain a deeper understanding of the data structure. SOMs are particularly effective for clustering, dimensionality reduction, and visual pattern recognition in environmental datasets [33]. In the context of this study, SOMs were applied to identify clusters of similar features and detect non-linear dependencies that may not be captured by traditional correlation analysis.

3 Material and Methodology

3.1 Research area

In the early years, surface water in river Amstel and city's canals were drinkable for citizens in Amsterdam. As surface water became more polluted and increasing demand with population growth, the dune water supply system is built from 1853. And in 1957, a pipeline connecting Lekkanaal and dune was built to recharge the dune and maintain the groundwater level [34]. Now surface water from Lekkanaal contributes about two thirds of tap water in Amsterdam.

Water extracted from the channel is not suitable for consumption in its raw form. Waternet's multi-step treatment process transforms this raw water into safe drinking water. This process begins with water extraction, followed by pre-purification, where larger pieces of dirt is removed by coagulation and sedimentation. Subsequently, the water undergoes sand and gravel filtration, removing finer particles and harmful substances such as ammonia, with assistance from naturally occurring bacteria. The filtered water is then subjected to natural purification by infiltration into the sand dunes of the Amsterdam Water Supply area, where harmful bacteria and viruses are further broken down over a three-month period. Afterward, the water is transported to the Leiduin treatment facility, where it undergoes a series of final purification steps: sand filtration, ozonation, lime removal, carbon filtering, and slow sand filtration. This thorough process ensures that the water distributed to homes is of high quality, safe for consumption, and meets regulatory standards.

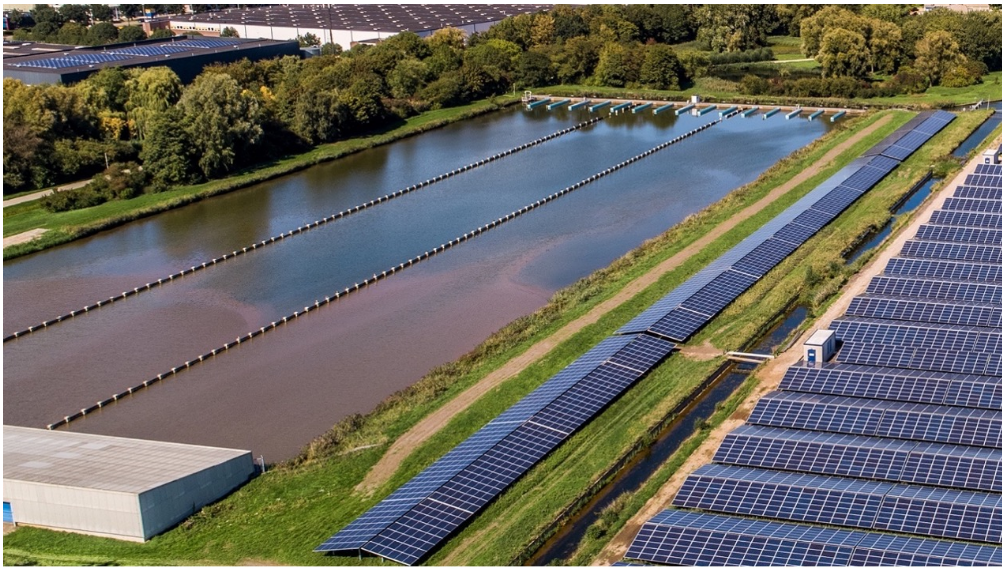


Figure 5: Pre-purification of source water at Lekkanaal [3]

For water from Lekkanaal, the initial pre-purification step is carried out at the DWTP in Nieuwegein, the largest DWTP in the Netherlands, where coagulation and sedimentation happen. Coagulation in water treatment is defined as the process of adding a chemical coagulant or coagulants to suspended, colloidal, and dissolved matter for subsequent processing by flocculation or to produce conditions that will allow the particulate and dissolved matter to be removed later [3]. At Nieuwegein, ferric chloride is added to the raw water as a coagulant. It neutralizes the electrical charges on colloidal particles, causing them to destabilize and aggregate into larger, settleable flocs. Basic stoichiometric reactions occurring during the coagulation process for ferric chloride is given below:

$$FeCl_3 + 3\,HCO_3^- \longrightarrow Fe(OH)_3(s) + 3\,Cl^- + 3\,CO_2 \tag{3.1}$$

These flocs sink to the bottom and are removed from the water. After pre-purification, there are much less harmful substances such as bacteria and viruses. Water becomes crystally clear instead of brown-green.

Lekkanaal is a canal that connects the River Lek to the Amsterdam-Rhine Canal at Nieuwegein. Located within the Rhine–Meuse–Scheldt delta, this area has a complex water network interconnected by rivers and canals. Although, the flow direction in Lekkanaal can be bidirectional. The source of the water can be traced back to the River Rhine, which enters the Netherlands at Lobith. The Rhine is charged by a mixture of snowmelt, rainwater and groundwater with mean discharge of 2200 m3/s [35].

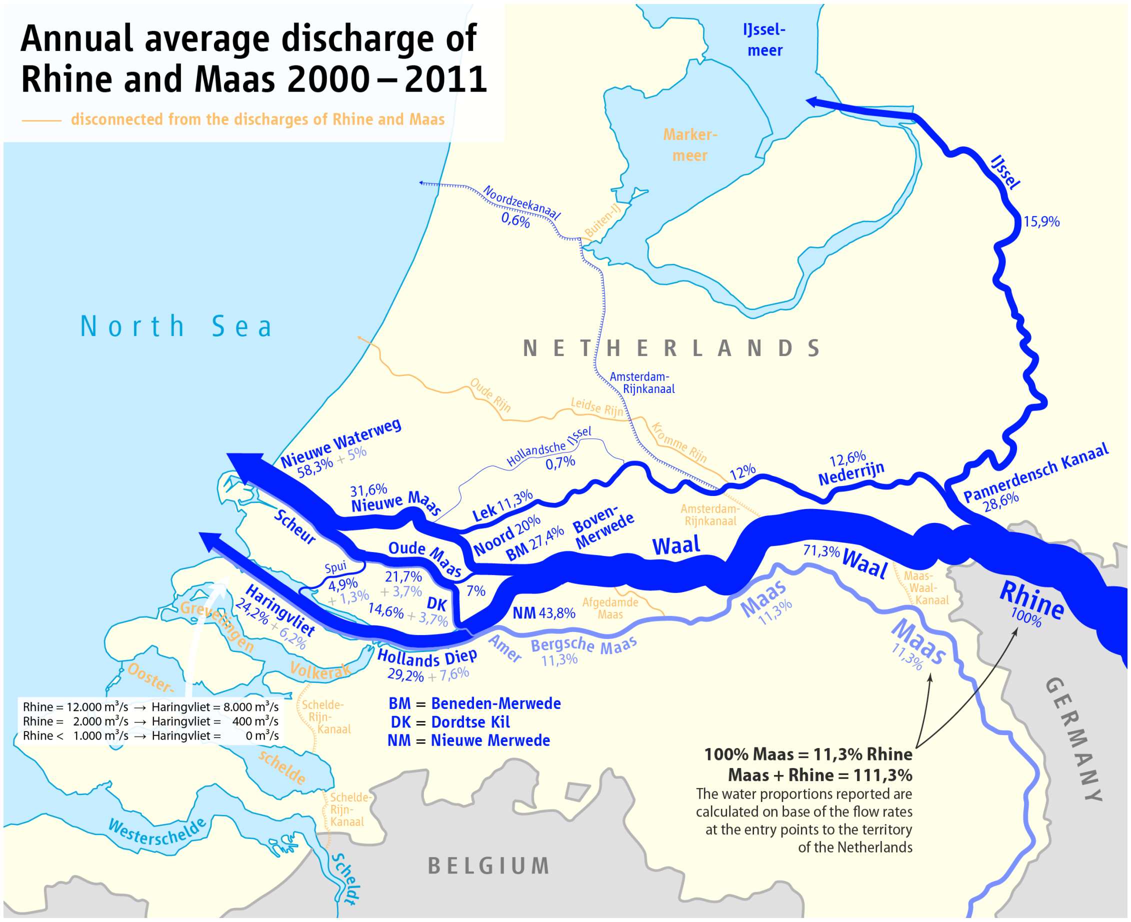


Figure 6: Annual average discharge of Rhine and Maas 2000-2011

The Netherlands has a temperate maritime climate with mild winters and cool summers. Mean winter temperatures are about 3°C and mean summer temperatures are around 17°C. Annual precipitation averages 85 cm and is fairly evenly distributed throughout the year. Evaporation exceeds rainfall between April and August [35].

3.2 Data collection

There are three data sources for this study.

- **Waternet** operates the DWTP at Nieuwegein, where key water quality parameters, including turbidity, pH, flow, and water temperature, are monitored using sensors at the inlet.

- **Rijkswaterstaat**, a Directorate-General of the Ministry of Infrastructure and Water Management in the Netherlands, is responsible for the design, construction, management, and maintenance of the country's primary infrastructure. It collects extensive water data from sensors nationwide. In this study, historical discharge, water level, water temperature, and turbidity data were obtained from its Waterbericht website.

- **The Royal Netherlands Meteorological Institute (KNMI)** is the Dutch national weather service, responsible for weather forecasting and monitoring climate, air quality, and seismic activity. It also serves as the national research and information center for these fields. In this study, historical precipitation, air pressure, and temperature data were obtained from KNMI's weather station at Deleen.

The locations of the data sources is plotted in Figure 7:

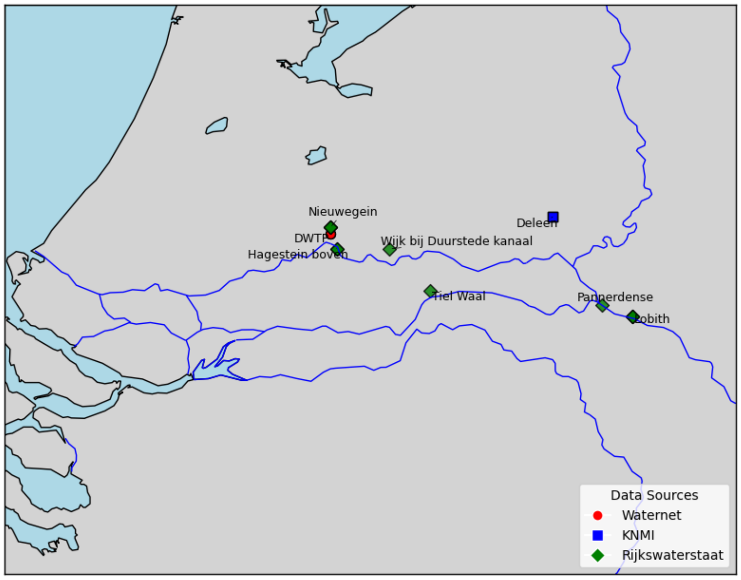


Figure 7: Locations of the data sources

The overview of data is listed in Table 1:

Table 1: Data source overview

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

3.3 Methodology

3.3.1 Data quality assessment

Time series data have automatically been recorded by various sensors at multiple locations. The raw data may contain several types of errors, such as outliers, missing values, flat-lining data, and invalid measurements [36]. To ensure data reliability, we followed the data quality assessment framework proposed by Gleeson et al. [37], which addresses the following data issues.

Timestamp errors

Timestamp errors occur when the time intervals between consecutive data points deviate from the expected interval, leading to irregularities in time-series analysis. These errors can result from data logging failures, synchronization issues, or missing records.

To detect timestamp errors, time intervals between consecutive timestamps were calculated, and deviations from the expected interval (determined as the most common interval) were identified. The cumulative time error was also tracked to assess long-term discrepancies.

Missing data

Missing data can arise due to sensor failures, data transmission errors, or incomplete records. Identifying and addressing missing values is crucial for maintaining the reliability of data-driven models and analyses.

To detect missing data, each column was scanned for NaN (Not a Number) values. The total count of missing values per column was summarized. Linear interpolation was then applied to ensure data competences.

Invalid data

Invalid data points refer to values that fall outside predefined acceptable boundaries, which can distort analysis and lead to inaccurate conclusions. These boundaries may be based on physical limitations, sensor specifications, or domain knowledge.

To detect invalid data, each column was checked against lower and upper boundaries. If a value fell below or exceeded these limits, it was marked as invalid. This approach ensures a comprehensive understanding of data integrity and aids in necessary corrections. Table 2 shows the lower and upper boundaries for each time series.

Table 2: Parameter Boundaries

Name	Lower boundary	Upper boundary	Unit
sensor_turbidity	1	1e3	FTU
sensor_flow	-1e4	1e4	m^3/h
sensor_tempearture	3	27	°C
sensor_pH	7.5	8.6	
Lob_dis	-1e4	1e4	m^3/s
Hag_dis	-1e4	1e4	m^3/s
Nieu_dis	-1e4	1e4	m^3/s
Pan_dis	-1e4	1e4	m^3/s
Tiel_dis	-1e4	1e4	m^3/s
Nieu_wl	-100	100	cm
Wijk_wl	-100	100	cm
Nieu_wt	0	50	mS/m
Hag_wt	0	50	mS/m
Lob_wt	0	50	°C
Nieu_EC	1e5	1e10	°C
Lob_EC	1e5	1e10	°C
Lob_pH	7	9	
Lob_turbidity	1	1e3	FTU
P	9000	11000	Pa
RH	0	500	0.1mm
T	-50	50	°C

Single point outliers

Single point outliers are data points that significantly deviate from surrounding values, often caused by temporary sensor errors or sudden environmental disturbances. These anomalies can introduce bias in the analysis if not properly identified and handled.

In this study, single point outliers were detected using a pre- and post-window z-score approach. The method involves computing rolling means and standard deviations for 6 data points before and after each observation. The z-score measures how far a data point deviates from the mean in terms of standard deviations. It is calculated using the formula:

$$z = \frac{x - \mu}{\sigma}$$

Where x is the data point, μ is the mean of the surrounding values (computed from the pre- or post-window), and σ is the standard deviation of those values. A data point is considered an outlier if both its pre- and post-z-scores exceed z-score threshold. In this work, a z-score threshold of 60 with a 36-point window was applied to detect only significant single-point outliers in the dataset. This strategy ensures that only genuinely anomalous values are identified, distinguishing them from natural variability.

Flat-lining data

Flat-lining data occurs when a sensor repeatedly records the same value over an extended period, potentially indicating sensor malfunction, data transmission failure, or a lack of variability in the measured parameter. Such anomalies can distort data analysis and lead to misleading conclusions.

To detect flat-lining, all time series were first resampled to a common hourly resolution by computing the mean within each time interval. Consecutive identical values were then identified, and their durations were calculated. If a flat-lining period exceeded a predefined threshold of 24 hours, it was flagged as an anomaly. This rule was not applied to rainfall data, as rainfall values could remian at zero for extended periods due to the absence of precipitation.

Since turbidity is the target variable of the ML model, any anomalous periods in the turbidity data were removed entirely to avoid introducing misleading patterns. For all other input parameters, identified anomalies were corrected using linear interpolation. This approach ensures a complete and continuous input dataset for training the ML model.

3.3.2 Runoff Time Analysis

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

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

However, the Lekkanaal and Amsterdam-Rhine Canal exhibit different flow characteristics due to hydraulic regulation. Water in these canals can flow in both directions, and discharge peaks are often unclear. In such cases, correlation analysis with time lag was applied instead. Specifically, Spearman correlation was calculated between turbidity and other water quality time series with time lags ranging from 0 to 15 days in 1-hour increments. Spearman correlation is preferred over the commonly used Pearson correlation because it is rank-based and better suited for detecting non-linear relationships. The Spearman correlation coefficient (ρ) is calculated as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3.2}$$

where:

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

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

3.3.3 Feature Selection

Feature selection is performed using correlation analysis and SOMs to identify the most relevant parameters for turbidity prediction.

First, Spearman correlation obtained from runoff time analysis is used to evaluate the relationship between turbidity and other parameters. Time series with a Spearman correlation coefficient larger than 0.3 were considered significant and selected as candidate features.

Next, the selected features from the correlation analysis were further analyzed using SOMs to explore non-linear relationships and potential feature redundancy. Numerical features were first clipped between the 5th and 95th percentiles to reduce the influence of outliers. The values were then normalized using the default range normalization provided by the SOM Toolbox. A SOM was trained on the normalized numerical data using the default 'imp' (importance-based) initialization method. After training, component planes were visually inspected to identify variables with similar patterns or cluster behaviors. Features that showed consistent, unique patterns in the SOM were considered important and retained for LSTM modeling.

This two-step process ensured that only meaningful and non-redundant variables were retained, effectively reducing the input feature space and model complexity. By minimizing redundancy, it also helped prevent overfitting and improved overall predictive performance. The importance of the selected features was further evaluated during model development through greedy feature selection.

3.3.4 Model Training

In this study, four models were developed to address two complementary tasks at the DWTP intake:

1. Short-term turbidity forecasting (regression)

2. Turbidity state identification (classification into normal or peak)

For the regression task, three models representing distinct modeling paradigms were employed: ARIMA, a classical statistical model for univariate time series forecasting; RF, a non-linear, multivariate model based on decision tree ensembles; and LSTM, a deep learning model designed for sequential data. These models were trained and evaluated using consistent data splits and forecast horizons (1 hour, 3 hours, and 6 hours ahead). While ARIMA provides a baseline from traditional statistical modeling, RF introduces a non-linear multivariate benchmark, and LSTM represents a more advanced approach capable of capturing complex temporal dependencies and feature interactions in dynamic water quality conditions.

In addition to continuous value forecasting, a second RF model was trained for a binary classification task: detecting whether upcoming turbidity levels fall into a normal or peak state. This classification task is motivated by the operational need for timely alerts during abnormal turbidity events, enabling rapid adjustments to treatment processes. The classification model uses the same predictor variables and temporal splitting strategy as the regression RF model but is trained on binary labels derived from a predefined turbidity threshold.

ARIMA (regression)

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

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

Given its simplicity and interpretability, ARIMA is often used as a baseline in time series prediction tasks.

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

- Stationarity Check: The augmented Dickey-Fuller (ADF) test is used to assess whether the series is stationary. If necessary, differencing is applied to remove trends.
- Hyperparameter Selection: The optimal values for p, d, and q are chosen using a grid search guided by the Bayesian Information Criterion (BIC). Compared to the Akaike Information Criterion (AIC), BIC penalizes model complexity more strongly. This ensures a more parsimonious model, especially important for large datasets like this work.
- Model Fitting and Forecasting: The final ARIMA model is fit to the training data. To evaluate the ARIMA model at different forecast horizons (e.g., 1-step, 3-step, 6-step ahead), a fast rolling forecast method is used. For each time step t in the test set, the trained model generates a multi-step forecast using get_forecast(steps=h), where h is the desired forecast horizon. The last forecast in the multi-step forecast sequence is extracted for comparison with the actual observed value. The model is then updated incrementally using model_fitted.append([new_observation], refit=False) to incorporate the next true observation, avoiding costly retraining. This process is repeated iteratively, simulating how the model would perform in a real-time deployment setting.

RF (regression)

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

Before training, all continuous variables were normalized to the range $[0, 1]$ using `MinMaxScaler` for consistency across modeling pipelines. The scaler was fitted on the training data only, and the same transformation was applied to the test set using stored parameters to avoid information leakage. Additionally, the target variable, sensor_turbidity, was log-transformed

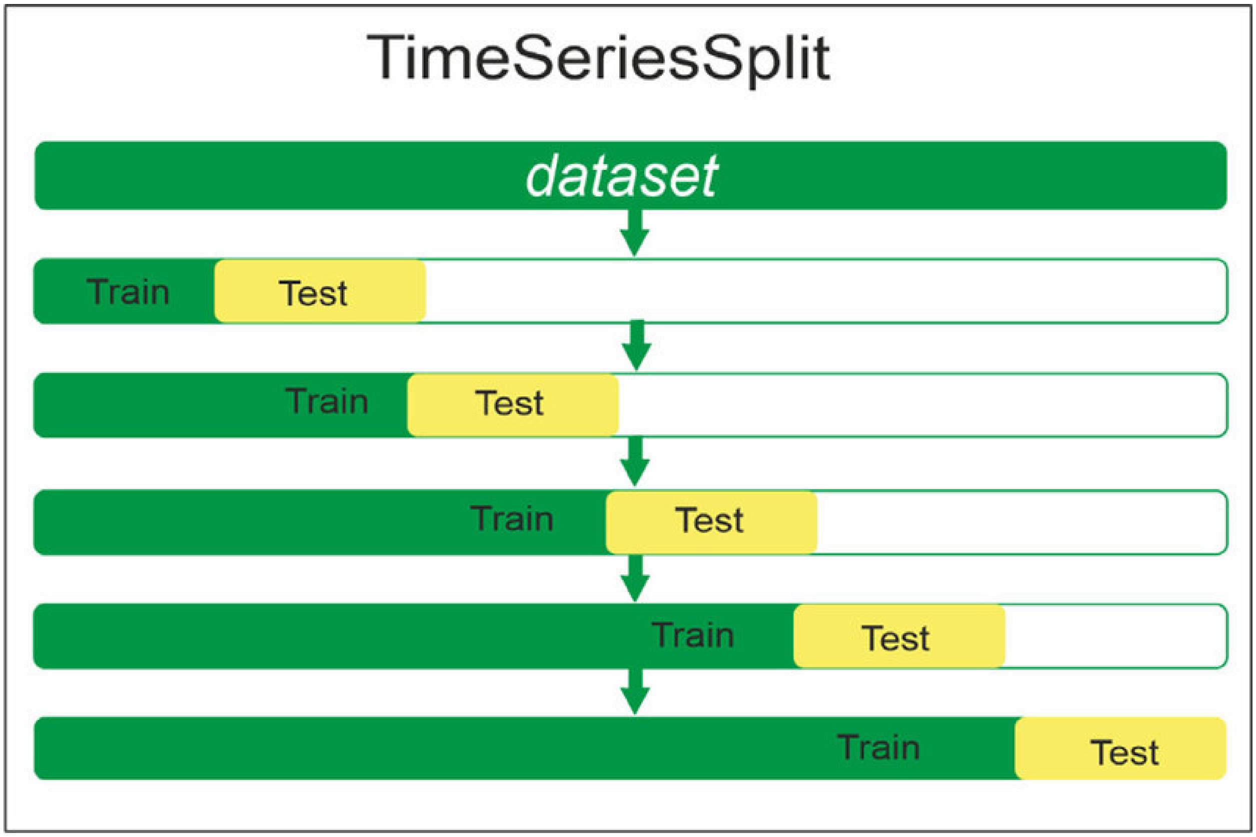


Figure 8: TimeSeriesSplit
[38]

using `np.log1p` to assess its impact on RF model performance, potentially reducing skewness, stabilizing variance, and enhancing the model's ability to capture non-linear patterns.

Like ARIMA model training, the RF model is trained on the first 85% of the data, while the remaining 15% is reserved for testing. K-fold cross-validation is used to assess model performance by splitting the training data into multiple subsets, ensuring reliable hyperparameter tuning and reducing overfitting. TimeSeriesSplit, with 5 folds in this work, is employed to preserve temporal order and prevent data leakage, ensuring that validation data always follows training data chronologically, as shown in Figure 8.

For each forecast horizon, hyperparameter tuning is performed using Optuna, an efficient automated hyperparameter optimization framework. The training and tuning process with Optuna involves:

- Using the training folds generated by TimeSeriesSplit to train and validate RF models with the proposed hyperparameters;
- Calculating the mean MSE across folds as the objective metric for optimization;

- Conducting multiple trials to identify the hyperparameter combination that minimizes the validation MSE;
- Retraining the RF model on the full training set using the best-found hyperparameters before evaluating on the test set.

The hyperparameter search space includes:

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

LSTM (regression)

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

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

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

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

- Target variable: sensor_turbidity or sensor_turbidity_log

- Hidden size: 16 to 256
- Number of LSTM layers (num_layers): 1 to 2
- Learning rate: 10^{-5} to 10^{-2}

RF (classification)

In addition to the regression-based turbidity forecasting models, a RF classifier was implemented to detect whether upcoming turbidity at the DWTP intake would be in a normal or peak state.

The feature set, data scaling, and temporal splitting strategy were identical to the RF regression model. In this case, however, the continuous turbidity series was converted into a binary target variable by applying a percentile-based threshold T_{peak} to the sensor_turbidity values, labeling observations above the threshold as peak and the remainder as normal. To avoid information leakage, T_{peak} was computed using only the training data.

Hyperparameter tuning was again performed using Optuna with TimeSeriesSplit cross-validation, but the mean F_2-score ($\beta = 2$) across folds was used as the optimization objective. The final model was retrained on the full training set with the best-found hyperparameters before evaluation on the test set.

3.3.5 Evaluation Metrics

The performance of the trained models was evaluated using two distinct sets of metrics, corresponding to the regression and classification tasks.

Regression models (ARIMA, RF, LSTM)

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

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

- MAE: Computes the average absolute difference between predicted and actual values; lower values indicate better performance.

- Relative Absolute Error: Computes the mean of absolute errors divided by their corresponding absolute observed values, expressed as a percentage; lower values indicate better accuracy.

Classification model (RF)

Performance was assessed using the confusion matrix, Precision, Recall, F_1-score, F_2-score ($\beta = 2$), and the Area Under the Receiver Operating Characteristic Curve (ROC–AUC).

- Confusion matrix: A tabular summary of predicted versus actual class labels, showing counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

- Precision: The proportion of predicted peak events that were actually peak in the observed data.

- Recall: The proportion of actual peak events that were correctly identified by the model.

- F_1-score: The harmonic mean of Precision and Recall, balancing the two metrics equally.

- F_2-score ($\beta = 2$): Similar to the F_1-score, but giving Recall twice as much weight as Precision.

- ROC–AUC: Measures the model's ability to distinguish between peak and normal states across all classification thresholds, with values closer to 1 indicating better discrimination.

All metrics were computed on the held-out test set to ensure consistent and fair model comparison.

3.3.6 Feature Attribution Analysis

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

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

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

- Built-in feature importance scores were extracted, which reflect the average contribution of each variable in reducing prediction error across the ensemble.

- SHAP (SHapley Additive exPlanations) was used to provide more nuanced local and global attributions. Summary and beeswarm plots were generated using TreeExplainer to visualize how each feature influenced predictions, including the direction and magnitude of its effect.

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

Together, these tools provided a transparent and model-specific understanding of how input variables contributed the predictions.

4 Results

4.1 Data quality assessment

Based on the data quality assessment framework. Figure 9 shows overall quality of the original data collected from Waternet, Rijkswaterstaat, and KNMI.

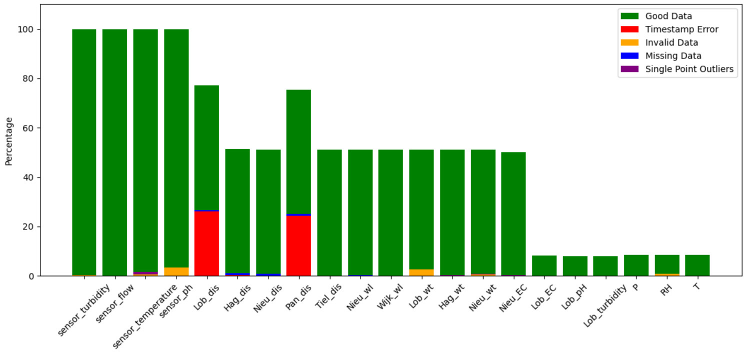


Figure 9: Data quality overview

Timestamp Errors

Sensor data show annual timestamp errors due to daylight saving time adjustments at the end of October, as shown in Figure 10. These errors were corrected by shifting the data forward to fill the missing hour.

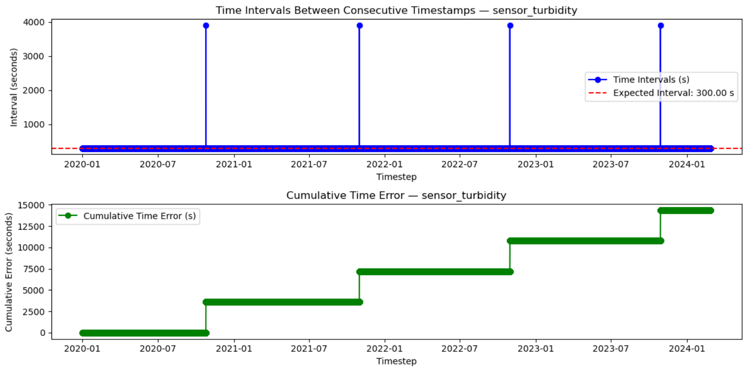


Figure 10: Timestamp error in sensor_turbidity data

Discharge data from Lobith, Hagestein, and Pannerdense occasionally contain repeated timestamps with few measurement values at the same time, as shown in Figure 11. To resolve this, mean values of these measurements are calculated for repeated timestamps.

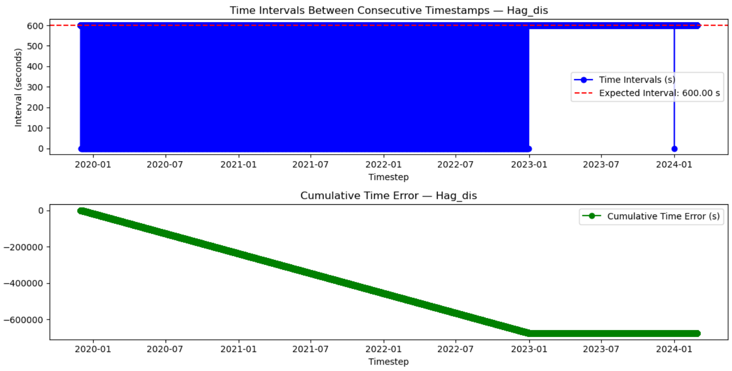


Figure 11: Timestamp error in Hag_dis data

Water temperature and EC data have inconsistent temporal resolutions, as shown in Figure 12. These inconsistencies are corrected by generating a uniform timestamp series and aligning each measurement to the nearest expected timestamp.

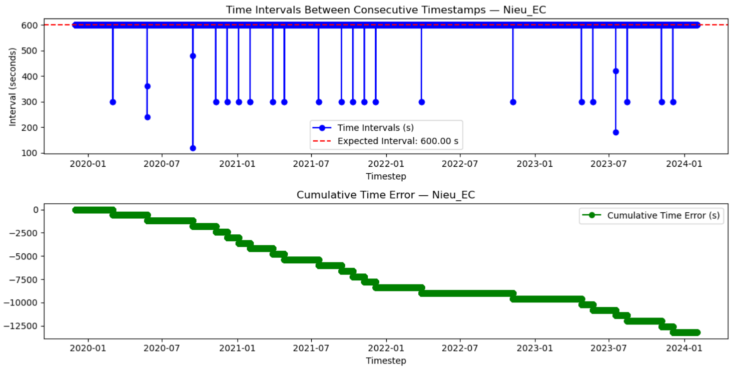


Figure 12: Timestamp error in Nieu_EC data

The pH and turbidity data from Lobith contained missing timestamps, which were filled based on the expected hourly temporal resolution.

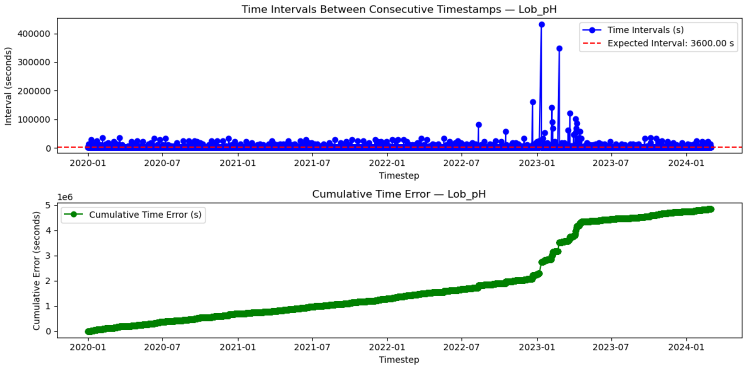


Figure 13: Timestamp error in Lob_pH data

Missing data

Missing data occurs only in discharge and water level measurements, accounting for less than 1% of the time series.

Invalid data

For sensor data, water temperature and pH values have predefined valid ranges, as determined by the DWTP operators.

For the Rijkswaterstaat dataset, records containing extreme outliers, such as temperature of 1e12 oC, were removed to ensure data integrity, as illustrated in Figure 14. These values were deemed physically implausible and indicative of measurement errors, justifying their exclusion from the analysis.

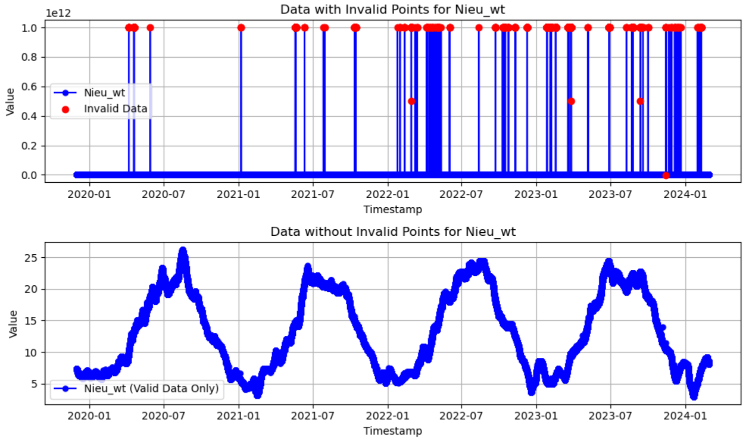


Figure 14: Nieu_wt with and without invalid data

In the original RH dataset, precipitation values below 0.05 mm are recorded as -1 mm. These values are replaced with 0 mm.

Single point outliers

Single point outliers occur more frequently in water temperature data due to the inherent nature of water temperature, where long periods of constant values can be followed by a small spike. This spike is often identified as an outlier, as shown in Figure 15. Removing these outliers and replacing them with linear interpolation did not affect the quality of the dataset much. In other cases, abnormal single point outliers can be correctly detected.

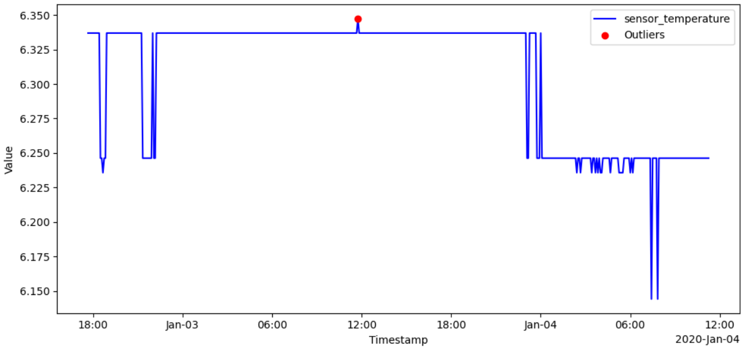


Figure 15: Single point outliners of Lob_EC

Flat-lining data

Flat-lining data were detected and interpolated linearly after all data was converted into hourly data. Furthermore, 3 weeks of turbidity data with abnormally low variance was found from 2020-11-16 14:00 to 2020-12-11 12:00, as shown in Figure 16. This period is manually removed.

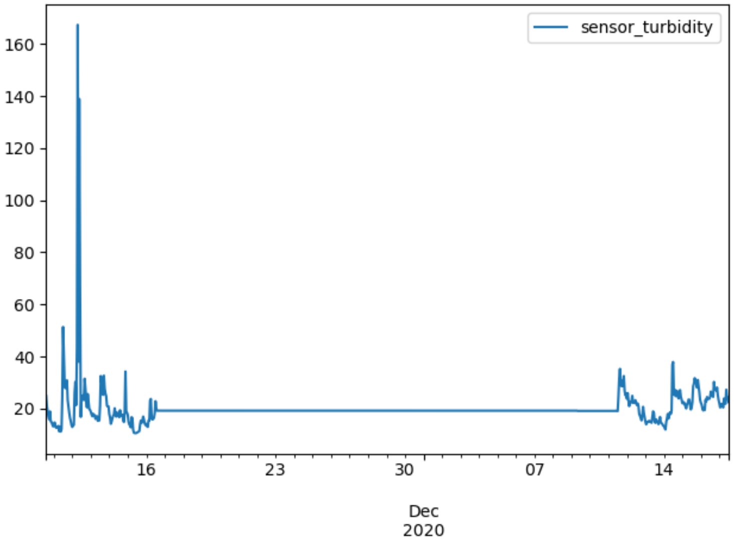


Figure 16: Abnormal period of sensor_turbidity

Table 3 shows the overview of all time series.

Table 3: Data overview

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

4.2 Runoff time analysis

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

First, the peak match method was applied to identify the runoff time between discharge stations. Figure 17 shows the peaks of Lob_dis and Hag_dis, where 13 common peaks were identified. The average travel time between these peaks was calculated to be around 28 hours.

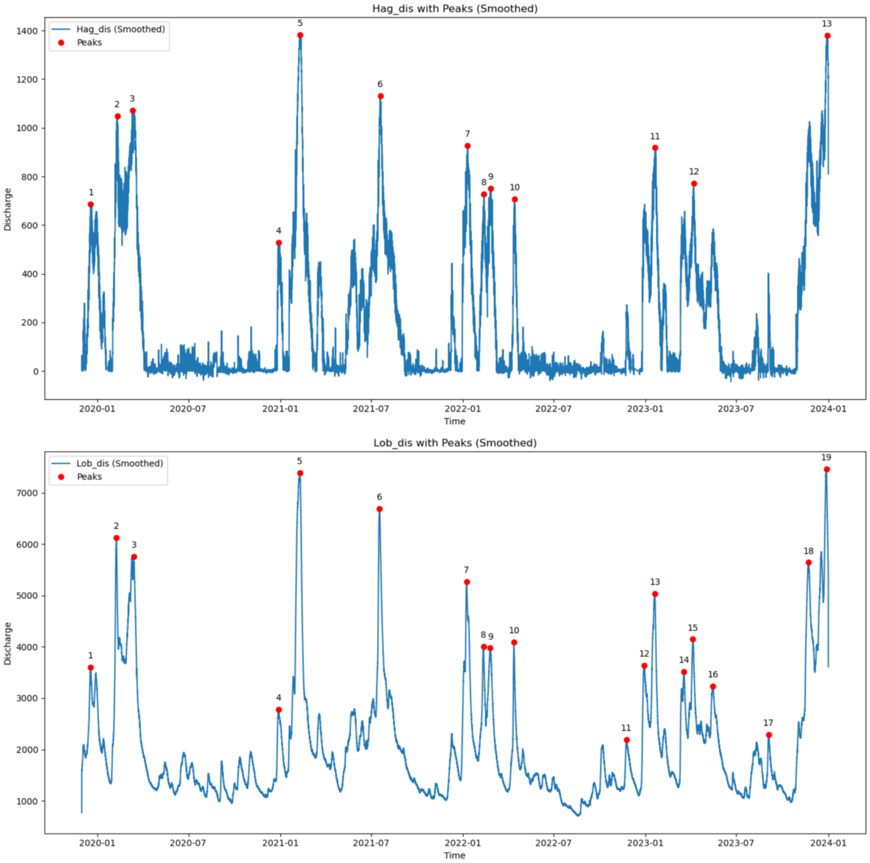


Figure 17: Peaks of Lob_dis and Hag_dis

The average travel time between corresponding peaks for each pair of discharge stations was calculated in the same manner, as shown in Table 4.

Table 4: Average travel time from Lobith by peak match method.

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

Figure 18 presents discharge data from Nieuwegein, located at the downstream end of the Lekkanaal. Unlike river discharge time series, the canal flow exhibits bidirectional behavior and lacks distinct peak patterns. As a result, the peak-matching method— which depends on clear maxima to estimate time lags—was not applicable. Instead, a correlation-based approach was adopted. Spearman correlation was calculated between each input time series and the target variable sensor_turbidity over time lags ranging from 0 to 15 days.

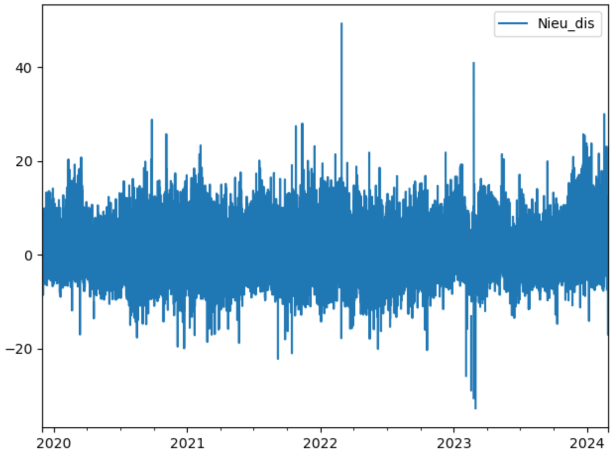


Figure 18: Nieu_dis

Figure 19 shows the result of the Spearman correlation between sensor_turbidity and discharge time series. Distinct correlation peaks can be observed from discharge measurements along the river. The time lag difference between the correlation peaks of Lob_dis and Pan_dis is 2 hours, which aligns with the travel time obtained from the peak match method, as well as for Tiel_dis (17 hours).

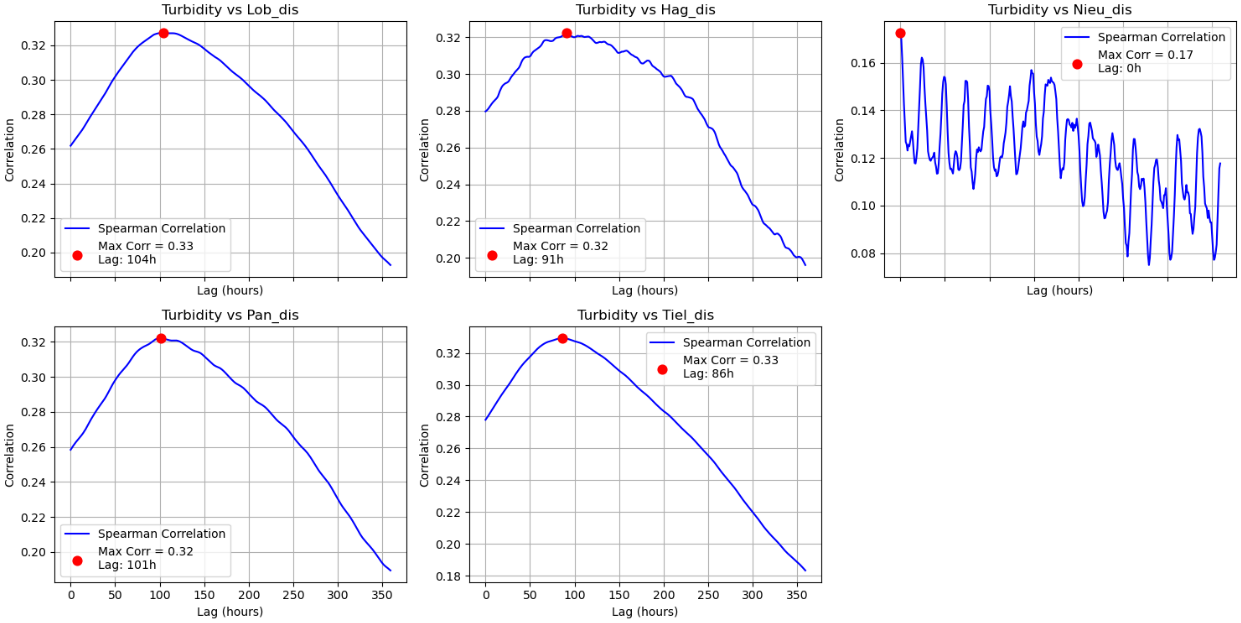


Figure 19: Spearman Cross_Correlation between sensor_turbidity and discharge

The time lag between the correlation peaks of Lob_dis and Hag_dis is 13 hours—shorter than the travel time estimated using the peak-matching method. This discrepancy arises because peak matching focuses on flood periods, when weirs and floodplains increase the runoff time [39]. In contrast, correlation analysis accounts for all flow conditions, capturing the generally faster and smoother water movement under normal, unobstructed conditions.

While the peak-matching method provided reasonable travel time estimates for river discharge, it could not be applied to the canal, where flow patterns lack distinct peaks. In contrast, Spearman correlation offered a more robust and universally applicable estimate of time lags across both river and canal conditions. Therefore, the correlation-based lag values were used to align all time series with sensor_turbidity, ensuring temporal consistency of the input data for subsequent ML modeling.

4.3 Feature selection

4.3.1 Spearman correlation

Table 5 provides an overview of the Spearman correlation results. The upstream discharge data show the strongest positive correlations with sensor turbidity, while temperature-related data and EC values in Nieuwegein exhibit the strongest negative correlations. The time lags indicated in the table will be used to align the sensor_turbidity time series with the other parameters.

Among collected time series, 11 of them were identified with maximum Spearman correlation coefficients exceeding 0.3 or falling below -0.3. These belong to four categories of parameters: upstream discharge and turbidity showed strong positive correlations with sensor_turbidity, while EC and water temperature exhibited strong negative correlations.

The strongest positive correlations with sensor_turbidity were observed for upstream discharge and turbidity time series. Increased discharge from upstream areas, especially during storm events or snowmelt, typically mobilizes sediments, organic matter, and other suspended solids from the riverbed and catchment surfaces, leading to elevated turbidity downstream [14, 40]. These suspended materials are transported with the flow, causing spikes in turbidity levels at downstream monitoring points such as the DWTP intake. In contrast, water temperature and EC measured near the DWTP showed strong negative correlations with sensor turbidity. For EC, this inverse relationship is often explained by the dilution effect: high turbidity events usually occur during periods of increased discharge, which introduce large volumes of sediment-rich but ion-poor runoff into the system. As a result, the concentration of dissolved ions—and thus EC—tends to decrease when turbidity rises [41, 42]. Water temperature also tends to be lower during high-discharge events, where rainfall or snowmelt contributes cooler water to the system [43]. Moreover, warmer temperatures generally promote particle settling due to decreased water viscosity, leading to lower turbidity under calm, stable conditions [44]. Therefore, both EC and temperature exhibit inverse relationships with turbidity and provide valuable complementary information for turbidity prediction.

Figure 20 shows the result of the Spearman correlation between sensor_turbidity and EC time series from Lobith (approximately 70 km upstream of the DWTP) and Nieuwegein

Table 5: Spearman correlation overview

Parameter	Max correlation	Time lag (h)
sensor_turbidity	1	0
Lob_turbidity	0.3391	118
Tiel_dis	0.3292	86
Lob_dis	0.3272	104
Hag_dis	0.3223	91
Pan_dis	0.322	101
Nieu_dis	0.1726	0
sensor_pH	0.1006	0
RH_daily_accum	0.0494	156
RH	0.0328	120
Nieu_wl	-0.0718	52
Wijk_wl	-0.074	52
Lob_EC	-0.1127	53
P	-0.1269	339
Lob_pH	-0.1639	152
sensor_flow	-0.2285	0
Nieu_EC	-0.3293	0
T	-0.3896	36
sensor_wt	-0.4252	0
Nieu_wt	-0.4256	9
Hag_wt	-0.4292	8
Lob_wt	-0.4343	8

(about 1.5 km away). Clear negative correlation can be observed between sensor_turbidity and Nieu_EC, while no significant correlation is found with Lob_EC. This difference can be attributed to the spatial distance: the water quality at Nieuwegein, being much closer to the DWTP, is more representative of the conditions affecting sensor_turbidity. In contrast, the lack of correlation with Lob_EC suggests that ion concentrations change along the river, reducing its predictive relevance for local turbidity levels near the DWTP.

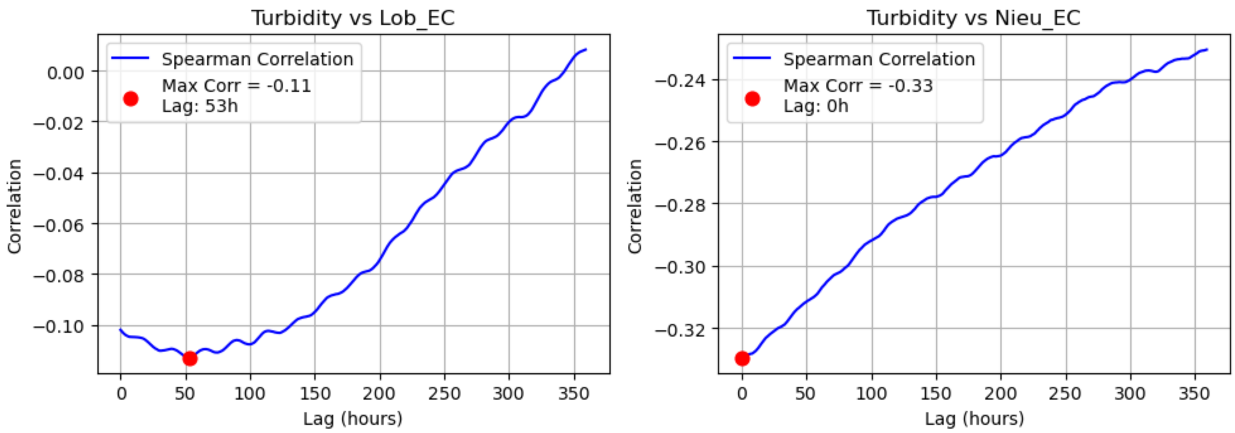


Figure 20: Spearman Cross_Correlation between sensor_turbidity and EC

Figure 21 shows the result of the Spearman correlation between sensor_turbidity and Lobith pH & turbidity time series. The correlation with pH exhibits a clear daily pattern, similar to that observed in water temperature, indicating a 24-hour cycle in pH values at Lobith. In contrast, the correlation with Lobith turbidity shows a positive peak at a time lag of approximately 104 hours—similar to the lag observed in the discharge time series—suggesting that turbidity is transported downstream from the Rhine and contributes to turbidity levels in the Lekkanaal.

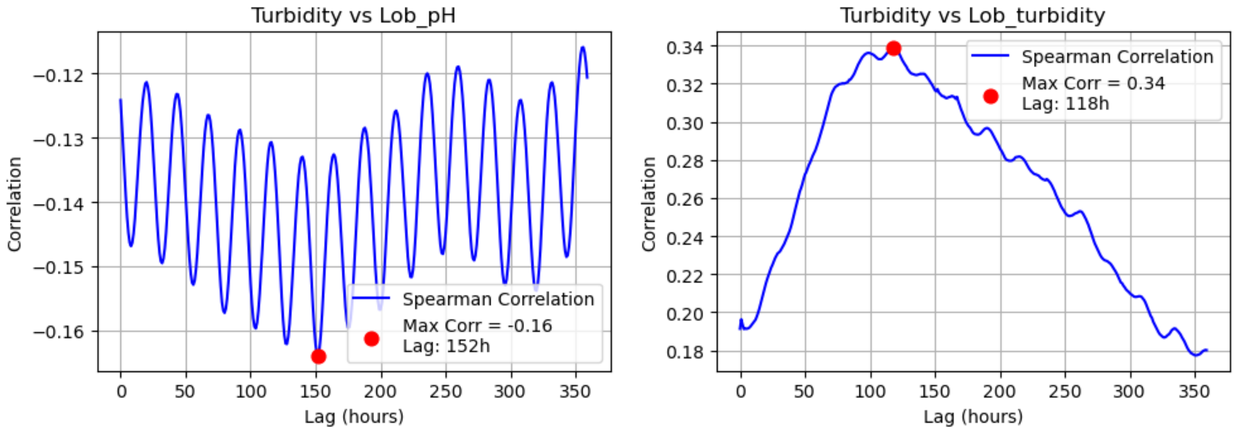


Figure 21: Spearman Cross_Correlation between sensor_turbidity and Lobith pH & turbidity

Figure 22 shows the result of the Spearman correlation between sensor_turbidity and weather data time series. No clear correlation is observed with air pressure. Air temperature exhibits a similar pattern to water temperature, but with a larger time lag, indicating that its effect on turbidity is less direct than that of water temperature.

Although accumulated rainfall is often considered a useful predictor for turbidity [31], the correlation between the two remains weak in this study. This may be attributed to the region's flat topography and highly regulated waterway system, which diminish the direct impact of rainfall-induced runoff on turbidity. Similar observations were reported by [14], who found that in the Göta Älv River in Sweden, the relationship between precipitation and turbidity was generally weak except during flood events. In such lowland and regulated systems, the influence of local rainfall is often buffered by upstream storage and flow management, leading to a decoupling of rainfall and turbidity under normal flow conditions.

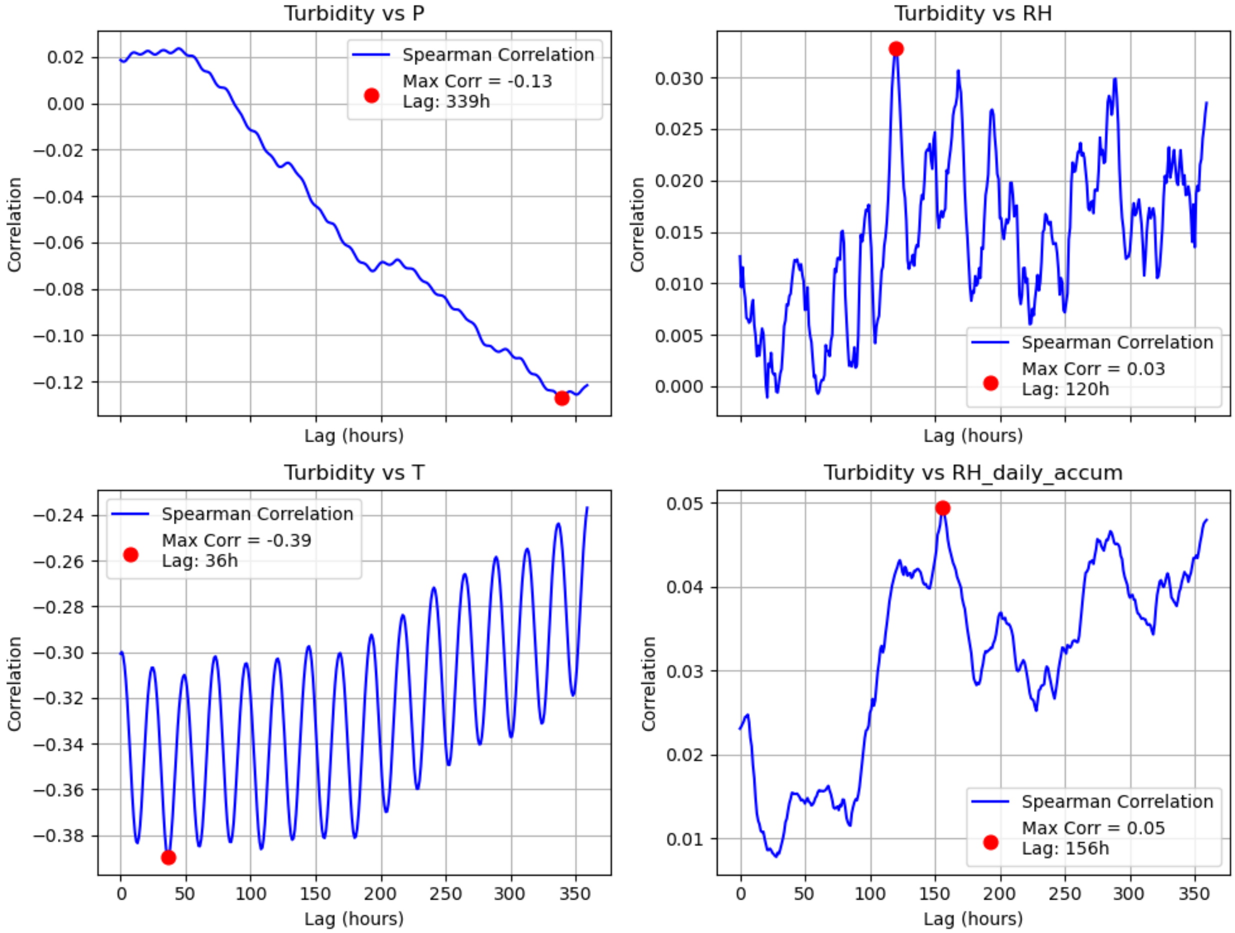


Figure 22: Spearman Cross_Correlation between sensor_turbidity and weather data

4.3.2 SOMs

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

These 11 time series and sensor_turbidty were used to generate SOMs, with the results shown in Figure 23. The first row of SOMs presents senor_turbidity and its positively correlated variables. High turbidity events at the DWTP (bottom-right corner of the SOM) are consistently associated with periods of elevated discharge in the Lower Rhine catchment. This pattern is also reflected in the turbidity measurements at Lobith, suggesting that increased upstream discharge enhances the transport of suspended matter into the downstream reaches and subsequently affects raw water quality at DWTP.

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

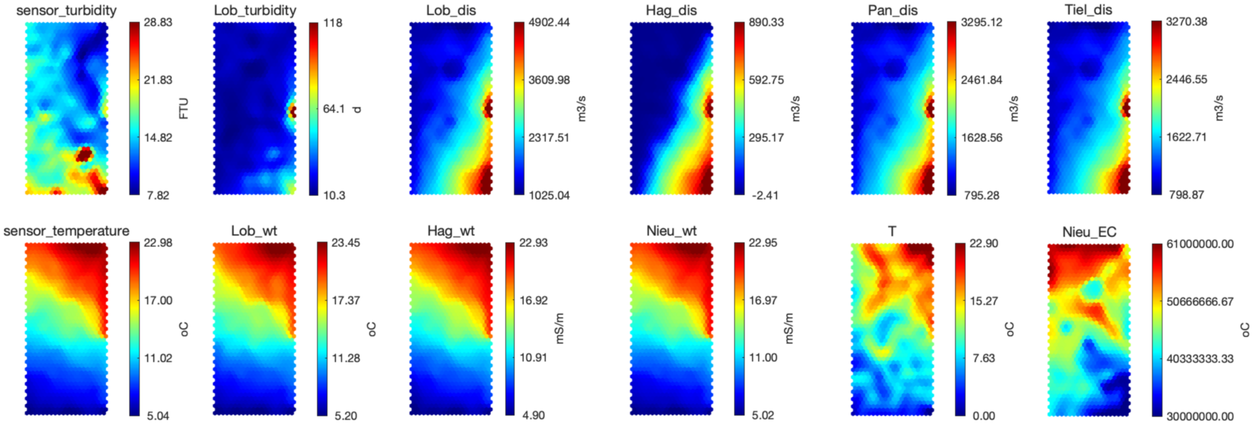


Figure 23: SOMs of 10 most relevant parameters

In addition to the numerical features, two categorical attributes, flow_direction and month, were included in the SOM analysis to provide contextual insight into the observed turbidity patterns. The flow_direction attribute represents the direction of water movement in the Lekkanaal: a value of 1 indicates flow from the Lek River toward the Amsterdam-Rhine Canal (northward), while -1 indicates the reverse (southward). Incorporating flow direction helps us to identify the primary source of turbidity in Lekkanaal since the flow there is bidirectional. Similarly, the month attribute adds seasonal context, helping to identify temporal patterns in turbidity related to climatic and hydrological cycles.

As shown in Figure 24, high turbidity events are predominantly associated with a positive flow direction, indicating that water flows from the Lek River toward the Amsterdam-Rhine Canal. This suggests that the Lek River is a major source of turbidity during these events. This observation aligns with the earlier correlation analysis, where upstream discharge showed a strong positive correlation with sensor turbidity. The Lek River generally has a higher discharge than the Amsterdam-Rhine Canal and therefore a greater capacity to transport suspended solids. The categorical SOM for the Month attribute shows that high turbidity events occur most frequently between December and February, corresponding to the winter season. This seasonal pattern supports the previous finding of a negative correlation between turbidity and water temperature, as colder months are associated with lower temperatures.

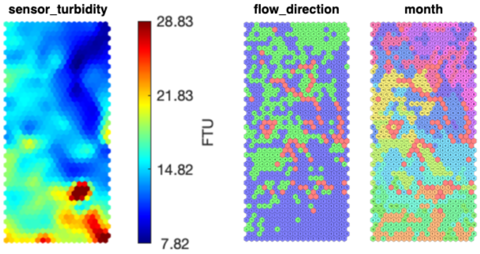


Figure 24: Categorical SOMs

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

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

4.4 Model comparison per forecast horizon

This section compares the forecasting performance of ARIMA, RF and LSTM—across different forecast horizons. To ensure a fair comparison, all models were trained using only sensor_turbidity as the input feature. This aligns with the ARIMA model's univariate nature and isolates the models' ability to learn temporal patterns from turbidity measurements alone. Figure 25 shows the schematic representation of the model setup.

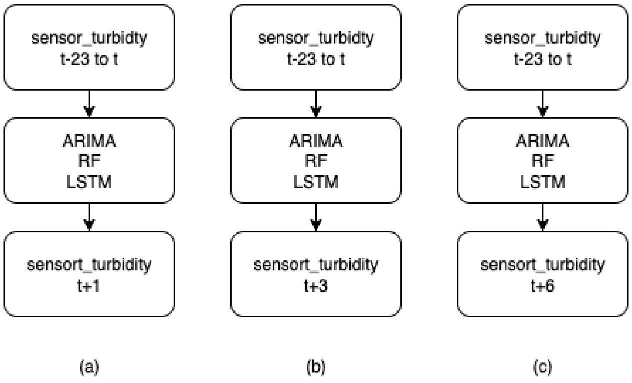


Figure 25: Comparison of model performance across three prediction horizons (t+1, t+3, t+6)

To determine the structure of ARIMA model, ADF test was applied to assess stationarity. The test yielded a statistic of -10.15 with a p-value < 0.01, providing strong evidence against the null hypothesis of non-stationarity. As a result, no differencing was required ($d = 0$). A grid search over the parameter space $p, q \in [0, 5]$ was then conducted, and the ARIMA(3, 0, 1) model was selected based on the lowest BIC.

4.4.1 1-hour Ahead Forecast

Figure 26 illustrates the data split used in the 1-hour ahead forecast experiment. Each model was trained on the first 85% of the time series, spanning from January 1, 2020, to July 22, 2023. The remaining 15%, from July 22, 2023, to February 27, 2024, was held out as the test set for evaluating predictive performance.

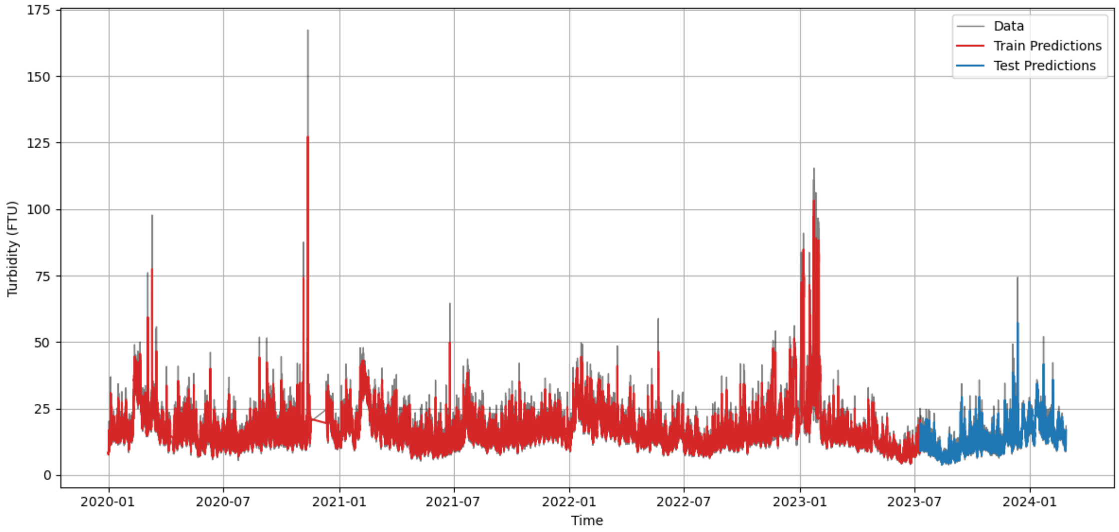


Figure 26: ARIMA predictions (Forecast window = 1 hour)

During the training period, two notable events occurred: the highest peak turbidity value of 167.33 FTU recorded on November 12, 2020, at 07:00:00, and a series of high-turbidity events in January and February 2023. In contrast, the test period exhibited peak events of lower magnitude compared to those in the training period. Additionally, a rare period of low turbidity occurred in August 2023, which was infrequently observed in the training set.

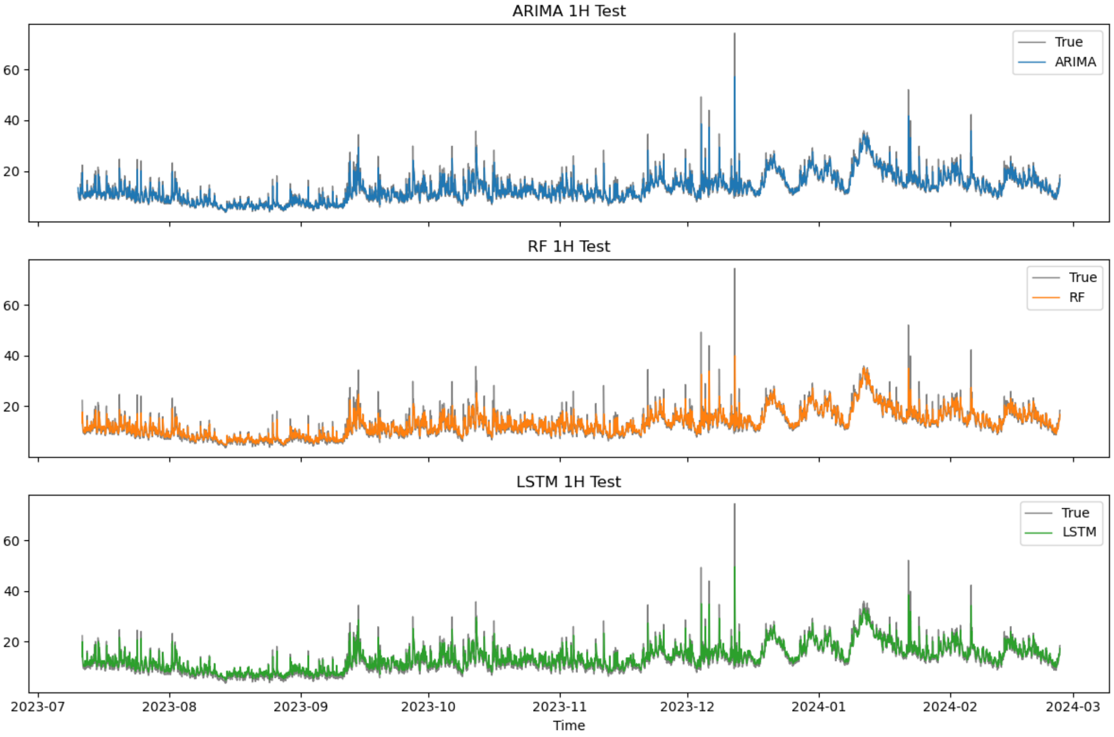


Figure 27: Predictions of three model on test period (Forecast window = 1 hour)

Table 6: Forecast results of three models (Forecast window = 1 hour)

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

Table 6 summarizes the performance of the three models on the test set for 1-Hour forecasting. All models achieve superior fit (NSE $>$ 0.8). LSTM performed slightly worse in terms of log NSE (0.85) compared to the other two models (0.87), indicating LSTM performed worse in low-magnitude turbidity events. The RF model also outperformed others in terms of MSE (5.34) and RMSE (2.31), reflecting better handling of large errors. Conversely, ARIMA achieved the lowest MAE (1.25) and relative absolute error (9.19%), suggesting it performs well on typical values. LSTM's highest MSE (5.71) and MAE (1.49) indicate reduced accuracy, underperforming the other two models in terms of typical values and outliers.

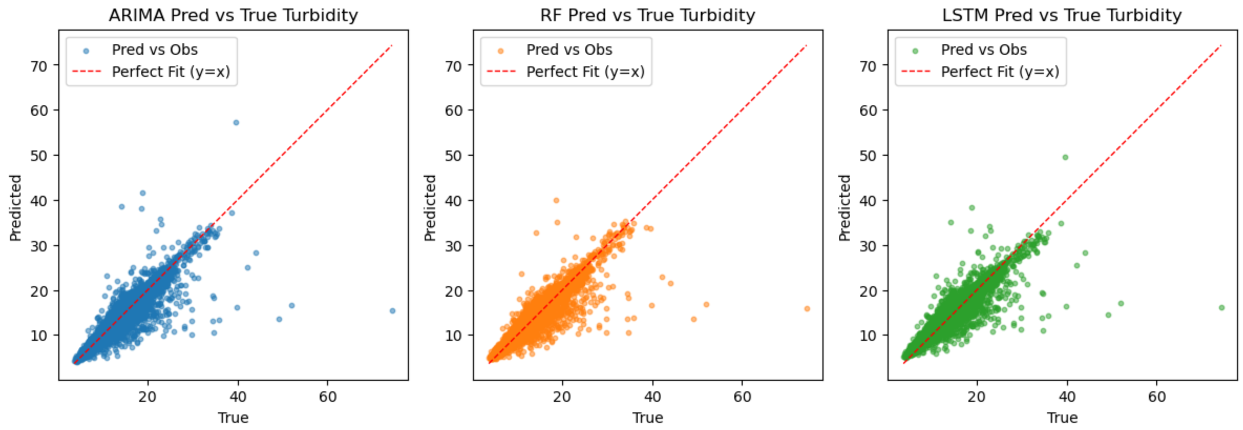


Figure 28: Scatterplot of predictions versus true values of three models (Forecast window = 1 hour)

Figure 28 shows error scatterplots. Most errors cluster along the zero-error line, indicating good overall fit for typical turbidity values. However, all three models consistently underestimate peak events, as shown by low prediction at high true turbidity values in the lower part of the plots. ARIMA exhibits the most symmetrical error pattern, with relatively same amount of overestimation (positive errors) and underestimation (negative errors). In contrast, RF shows less pronounced overestimation with fewer large positive errors, contributing to its lowest MSE and RMSE. LSTM displays distinct overestimation for low-turbidity events (e.g., August 2023), aligning with its lowest log NSE.

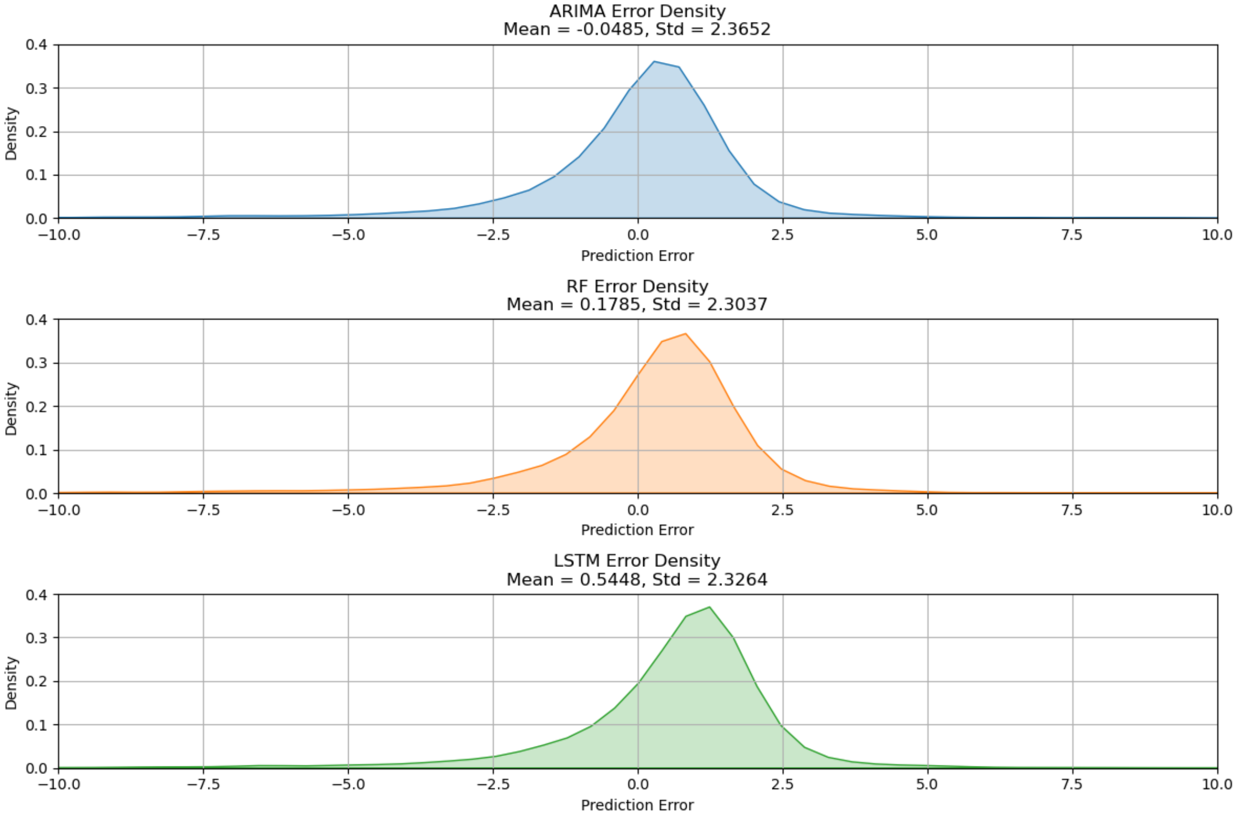


Figure 29: Error density of three models (Forecast window = 1 hour)

Figure 29 shows the density of errors from three models. All three curves exhibit positive skewness with longer tails on the left, indicating a tendency to overpredict during non-peak periods and occasional underprediction during peak events. Among the models, ARIMA demonstrates the least bias, with a mean error of -0.0485, aligning with its lowest MAE. In contrast, LSTM exhibits the highest bias (mean error of 0.5448), consistent with its highest MAE, suggesting systematic overprediction. RF, with a mean error of 0.1785 and the lowest standard deviation (2.3037), supports its stability across peak and non-peak events.

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

4.4.2 3-hour Ahead Forecast

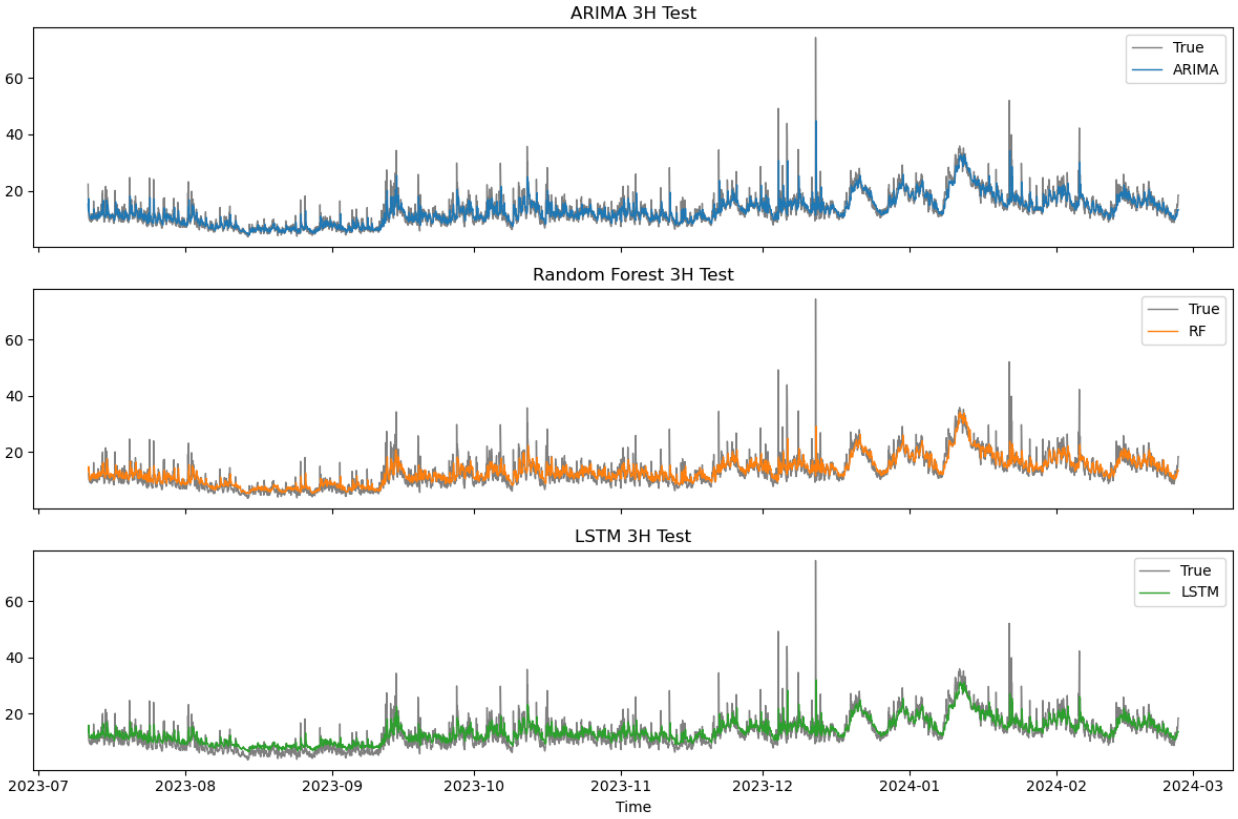


Figure 30: Predictions of three model on test period (Forecast window = 3 hour)

Table 7: Forecast results of three models (Forecast window = 3 hour)

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

Table 7 presents 3-hour forecast performance. Compared to 1-hour forecasts, all models show increased underfitting, with NSE dropping from 0.8 to around 0.69. RF maintains the highest NSE (0.70) and lowest MSE (8.46) and RMSE (2.91), while ARIMA achieves the lowest MAE (1.84) and relative absolute error (14.1%). LSTM's log NSE (0.69 vs. 0.76) indicates worsening performance for low-turbidity events, with the highest MSE (9.22) and MAE (2.10).

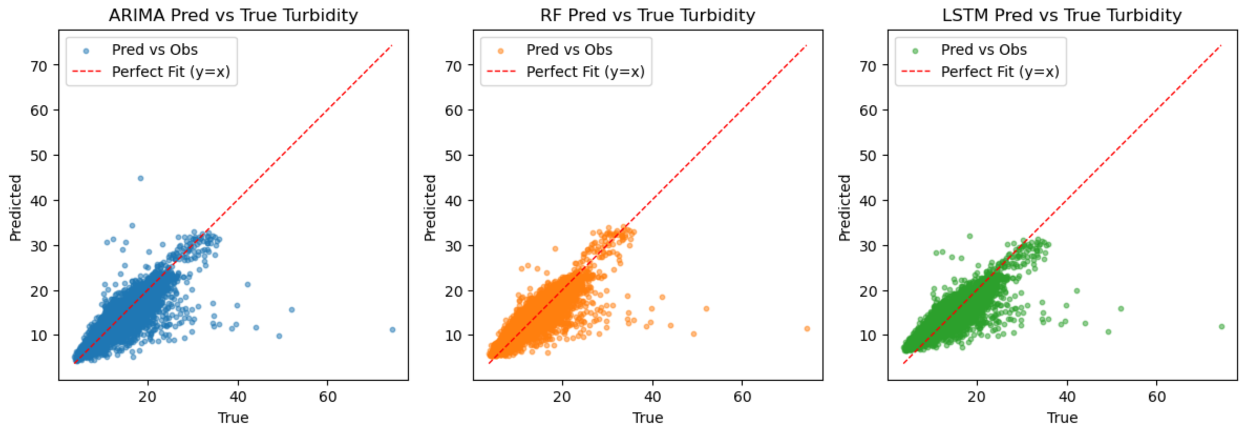


Figure 31: Scatterplot of predictions versus true values of three models (Forecast window = 3 hour)

Figure 31 confirms persistent peak underestimation. ARIMA's symmetrical errors contrast with RF's and LSTM's asymmetrical patterns, with latter two showing fewer large positive errors. LSTM's overestimation of low-turbidity events intensifies.

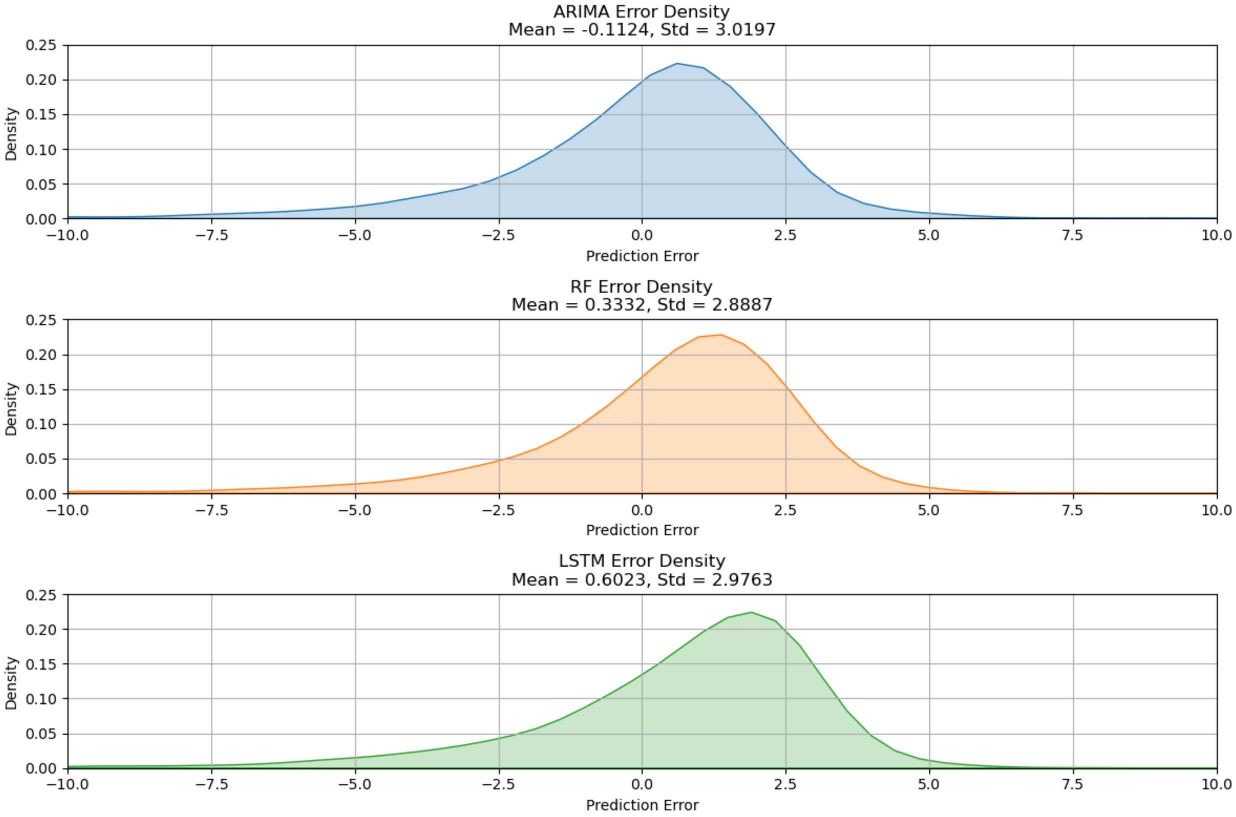


Figure 32: Error density of three models (Forecast window = 3 hour)

Figure 32 shows increased positive skewness and bias compared to 1-hour forecasts. ARIMA remains least biased (mean error: -0.112), while LSTM's bias is highest (0.545). RF's moderate bias and lowest standard deviation ensure stable error control.

Overall, 3-hour predictions mirror 1-hour trends but with reduced accuracy. RF remains best for dynamic conditions, ARIMA for non-peak events, and LSTM continues to underperform.

4.4.3 6-hour Ahead Forecast

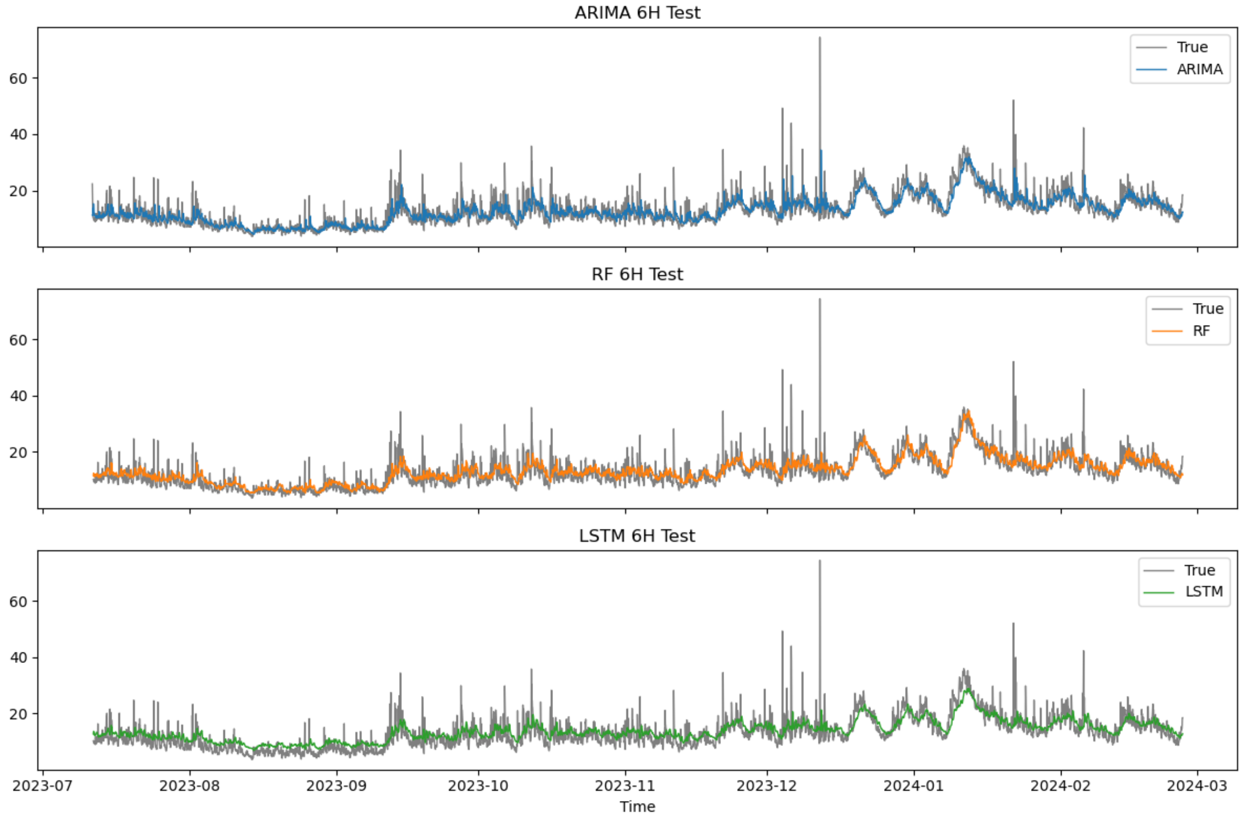


Figure 33: Predictions of three model on test period (Forecast window = 6 hour)

Table 8: Forecast results of three models (Forecast window = 6 hour)

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

Table 8 shows a further decline in forecast accuracy compared to 3-hour results. RF maintains the best overall performance with the highest NSE (0.66) and lowest MSE (9.76) and

MAE (2.16). ARIMA retains the lowest relative absolute error (16.7%), while LSTM's performance deteriorates further, with the lowest log NSE (0.61) and highest errors.

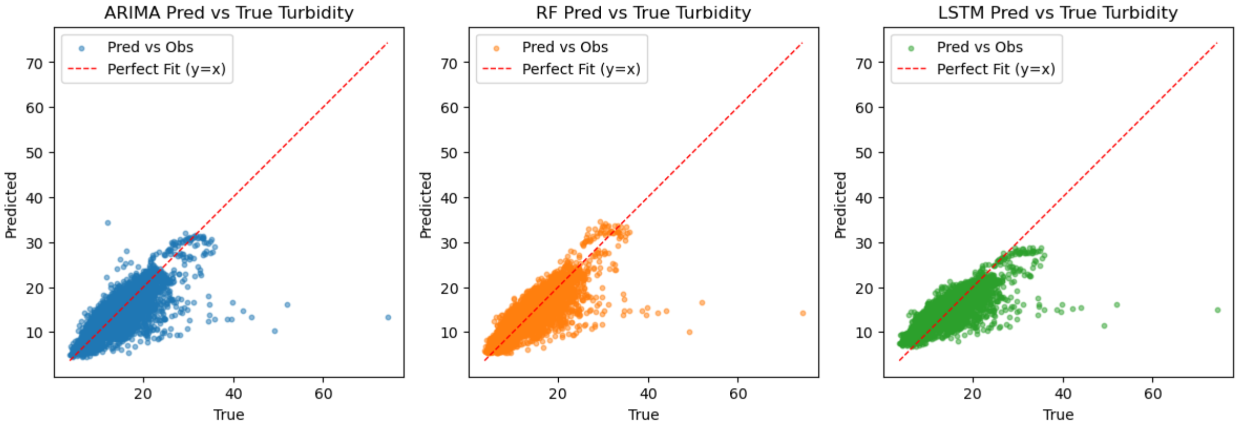


Figure 34: Scatterplot of predictions versus true values of three models (Forecast window = 6 hour)

Figure 34 confirms persistent peak underestimation across all models, with ARIMA showing some overestimation but RF and LSTM presenting highly asymmetrical errors, lacking major overpredictions. LSTM continues to overestimate low-turbidity events, consistent with previous findings.

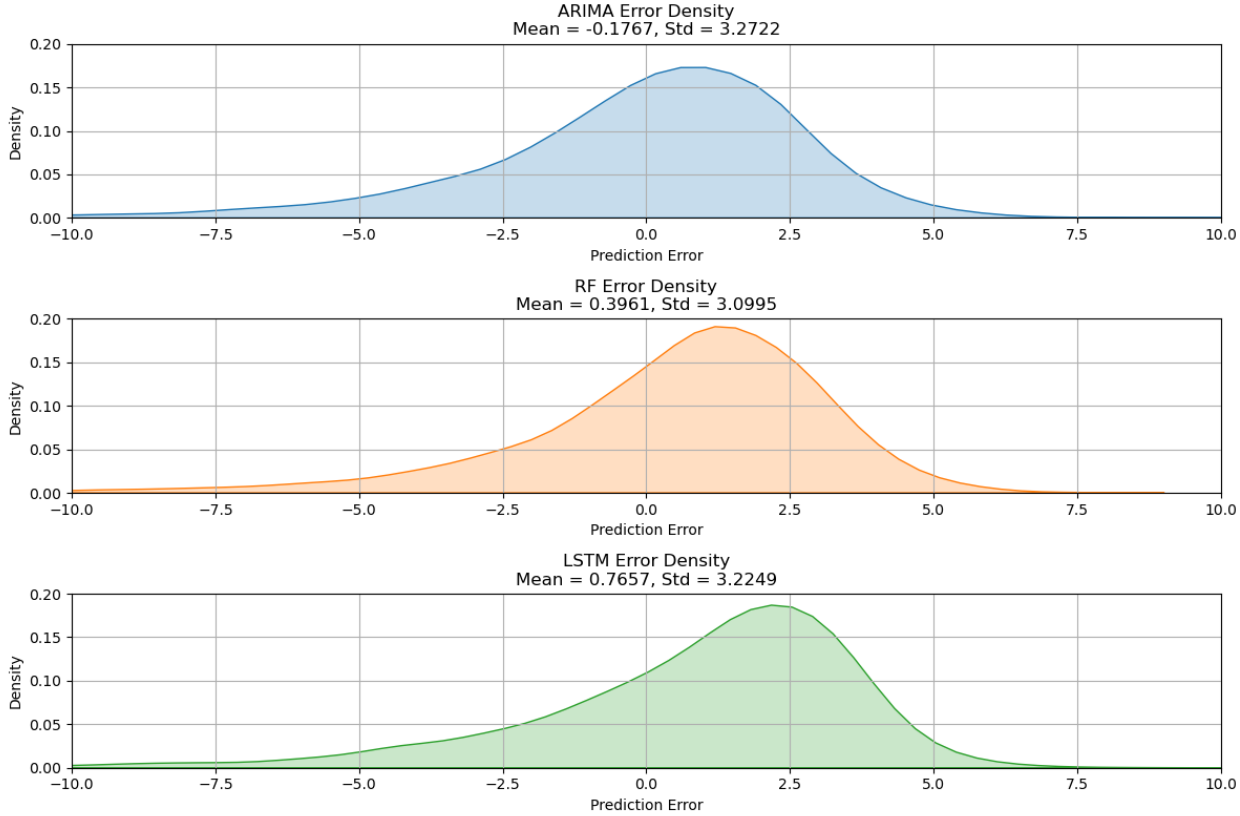


Figure 35: Error density of three models (Forecast window = 6 hour)

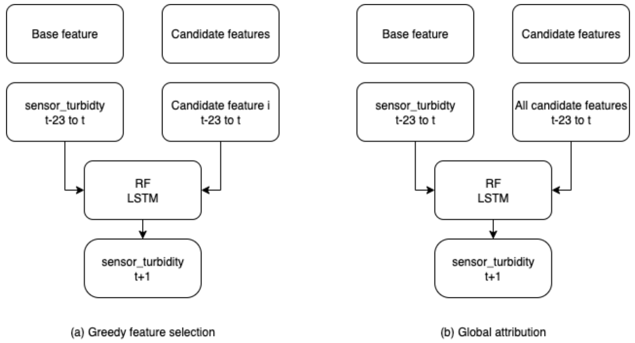
Error density plots (Figure 35) show increased positive bias and variance relative to 3-hour forecasts. ARIMA remains the least biased but with higher error spread, while RF exhibits a balanced bias-variance trade-off. LSTM shows the highest bias and error variability.

Overall, 6-hour forecasts demonstrate the expected decline in accuracy, with RF's relative stability reaffirming its suitability for longer-term operational forecasting.

Across all forecast horizons (1-, 3-, and 6-hour), ARIMA, RF, and LSTM consistently underestimate peak turbidity events, and accuracy diminishes as the lead time increases. RF consistently achieves the lowest MSE and RMSE, increasingly outperforming the other models at longer horizons. ARIMA performs best for short-term, non-peak conditions, delivering minimal bias and the lowest MAE. Meanwhile, LSTM underperforms overall, particularly struggling with low-turbidity events and rare extremes, exhibiting the highest bias and error variability.

4.5 Feature attribution analysis

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



(a) Greedy feature selection

(b) Global attribution

Figure 36: Feature attribution analysis experiments

Figure 36 illustrates the two experimental setups for feature attribution analysis.

(a) Greedy feature selection. Each candidate feature is individually added to a base feature and used to train a separate model. The change in performance relative to the base feature quantifies the marginal contribution of that candidate. This approach captures the isolated effect of each feature.

(b) Global attribution. All candidate features are combined with the base feature to train a single model. Feature importance is then computed for all features simultaneously using SHAP (RF model) and Captum (LSTM model). This provides a comparative "horizontal" view of feature contributions within the same model.

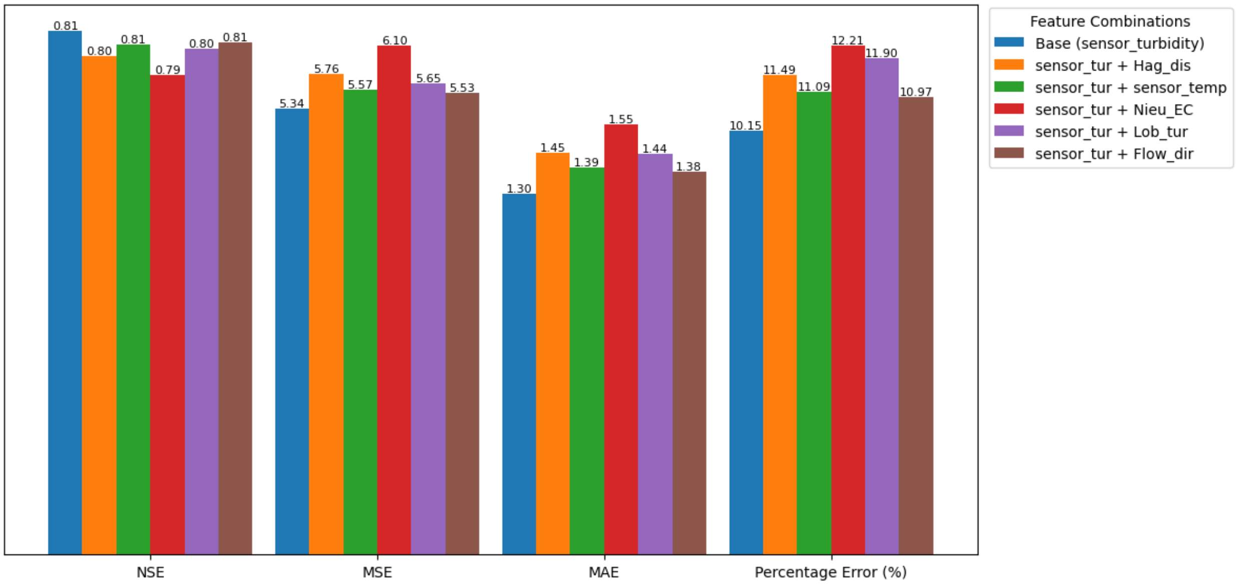


Figure 37: RF model greedy feature selection results (Forecast window = 1 hour)

Figure 37 presents the key performance metrics from the RF model's greedy forward feature selection. The base model, using only sensor_turbidity, outperforms all other feature combinations across all metrics, suggesting that additional features introduce noise and reduce RF model accuracy.

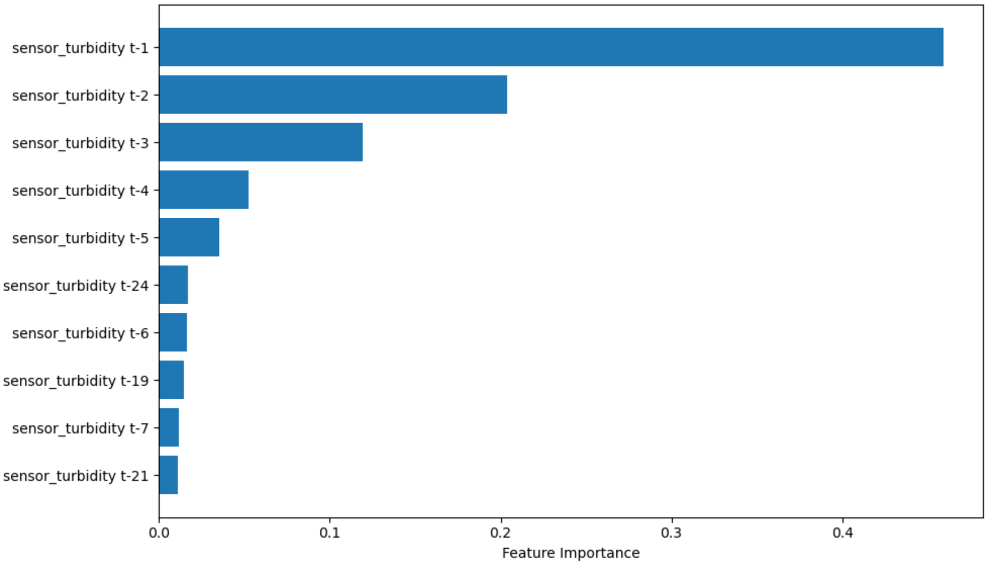


Figure 38: RF feature importance

All five additional features are used to train a new RF model for important analysis. Figure 38 displays the top 10 most important features for the RF model, all of which are sensor_turbidity time lags. The importance of sensor_turbidity t-1 to t-3 are substantially higher than others, aligning with the ARIMA BIC test, which selected ARIMA(3,0,1) as the optimal model, indicating that the three most recent data points are critical for 1-hour-ahead predictions.

Table 9: Maximum importance of different features

Feature	Maximum importance
sensor_turbidity	**0.459**
Lob_turbidity	9.34e-3
Nieu_EC	2.99e-3
Hag_dis	2.97e-3
sensor_temperature	1.21e-3
flow_direction	1.1e-4

Table 9 lists the maximum importance of each kind of input. sensor_turbidity is significantly more important than all other features.

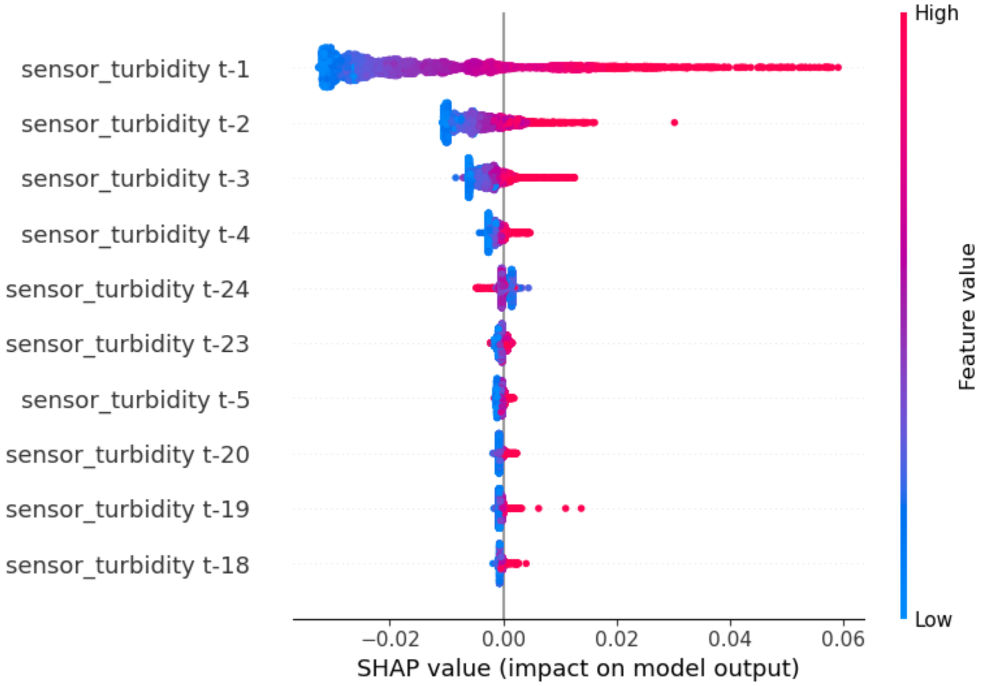


Figure 39: RF SHAP value

Figure 39 presents a SHAP summary plot, highlighting the top 10 features based on their mean absolute SHAP values. The results reveal that sensor_turbidity time lags, particularly t-1, t-2, and t-3, dominate the feature importance rankings. This indicates that recent turbidity measurements are the primary drivers of accurate predictions, where high values of sensor_turbidity t-1 to t-3 (red dots on the right) strongly increase predicted turbidity and low values (blue dots on the left) decrease it.

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

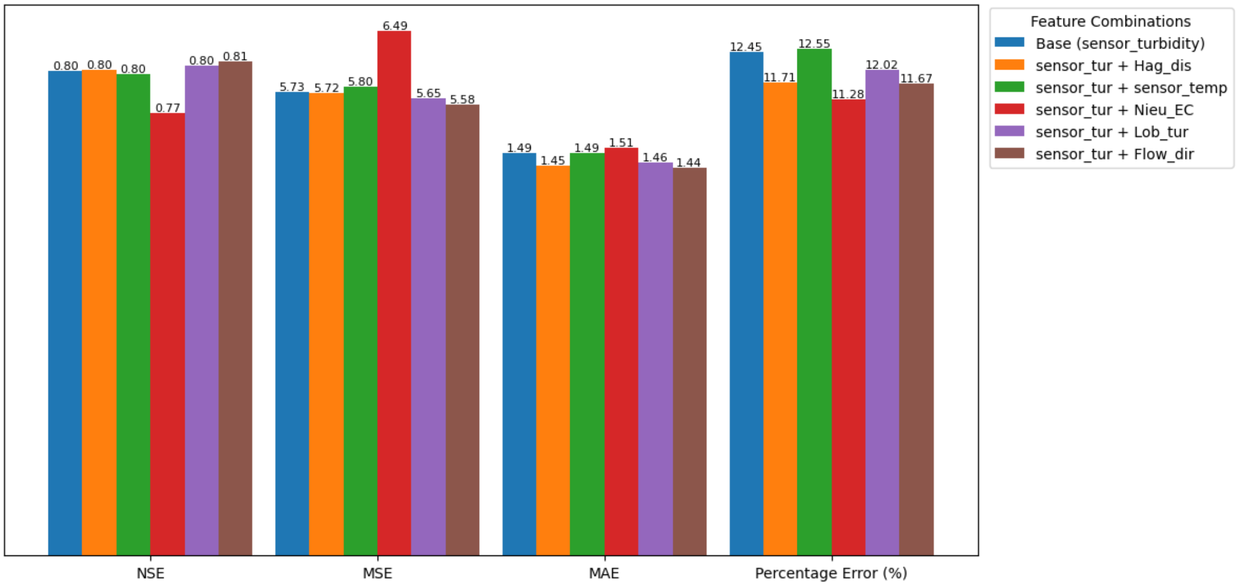


Figure 40: LSTM model greedy feature selection results (Forecast window = 1 hour)

Figure 40 presents the key performance metrics from the LSTM model's greedy forward feature selection. Incorporating Hag_dis, Lob_turbidity, and flow_direction results in only slight improvements in model performance.

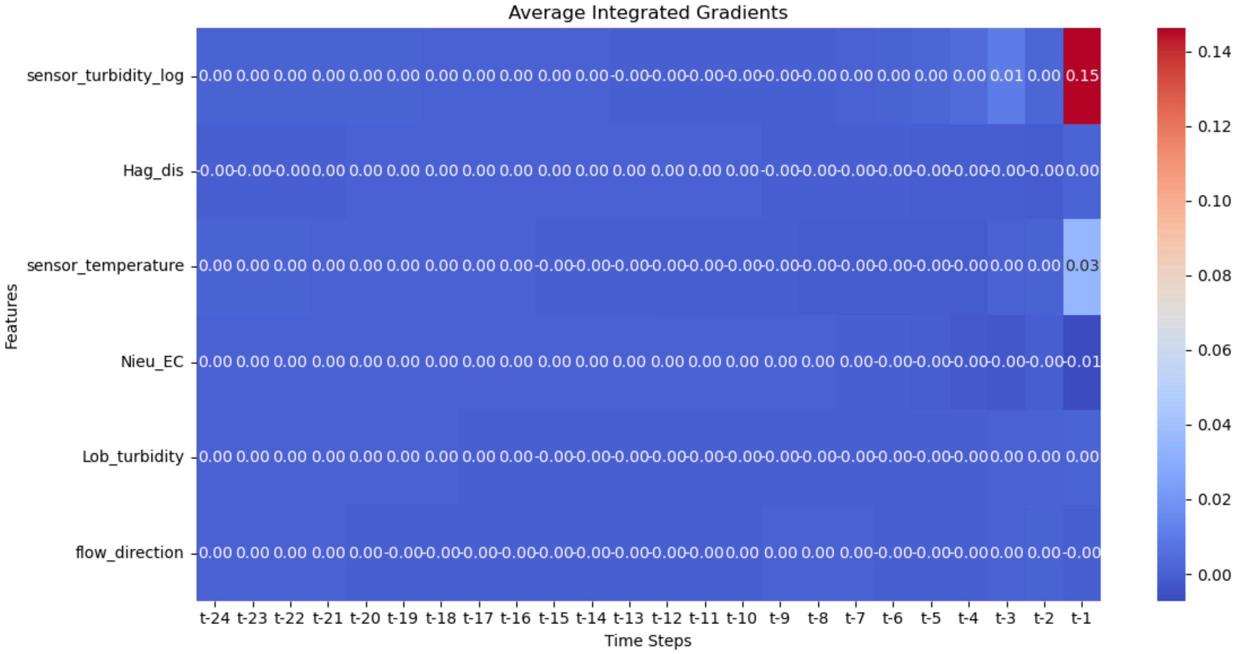


Figure 41: LSTM feature importance

These features were subsequently used to train a new LSTM model for global feature attribution analysis. Figure 41 presents the importance of input features for the LSTM model, showing results consistent with ARIMA and RF, where sensor_turbidity at lags t-1 is much more influential than other input. Similarly, the LSTM model trained with five features underperforms across all metrics compared to the LSTM model trained solely with sensor_turbidity time series, as shown in Table 10.

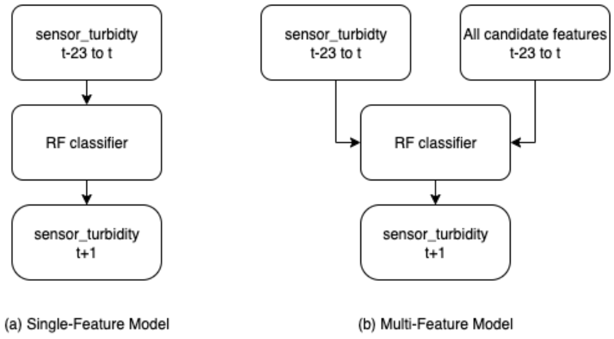
Table 10: Forecast performance metrics for single- and multi-feature LSTM models.

Model	NSE	log NSE	MSE	RMSE	MAE	Relative abs error
LSTM (1 feature)	**0.80**	**0.85**	**5.71**	**2.39**	**1.49**	**12.37%**
LSTM (5 features)	0.75	0.78	7.39	2.72	1.82	15.93%

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

4.6 RF classification model

In this section, peak turbidity events are defined as those where the sensor_turbidity value exceeds $T_{\text{peak}} = 28.21$ FTU, corresponding to the 95th percentile of the training set turbidity distribution. All other observations are classified as normal. As illustraed in Figure 42, the RF classifier was initially trained using only the sensor_turbidity feature. Subsequently, the classifier was trained with an expanded feature set to evaluate whether incorporating multiple predictors improves classification precision in distinguishing peak turbidity events.



(a) Single-Feature Model

(b) Multi-Feature Model

Figure 42: RF classifier experiments

Figure X presents the confusion matrices for the RF classifier trained with a single feature (left) and multiple features (right). Both models achieve very high accuracy in detecting normal events (Normal, class 0): the single-feature model misclassified only 8 normal events as peaks (FP), while the multi-feature model completely avoided such false positives (FP = 0). This explains why both models maintain overall accuracies above 95

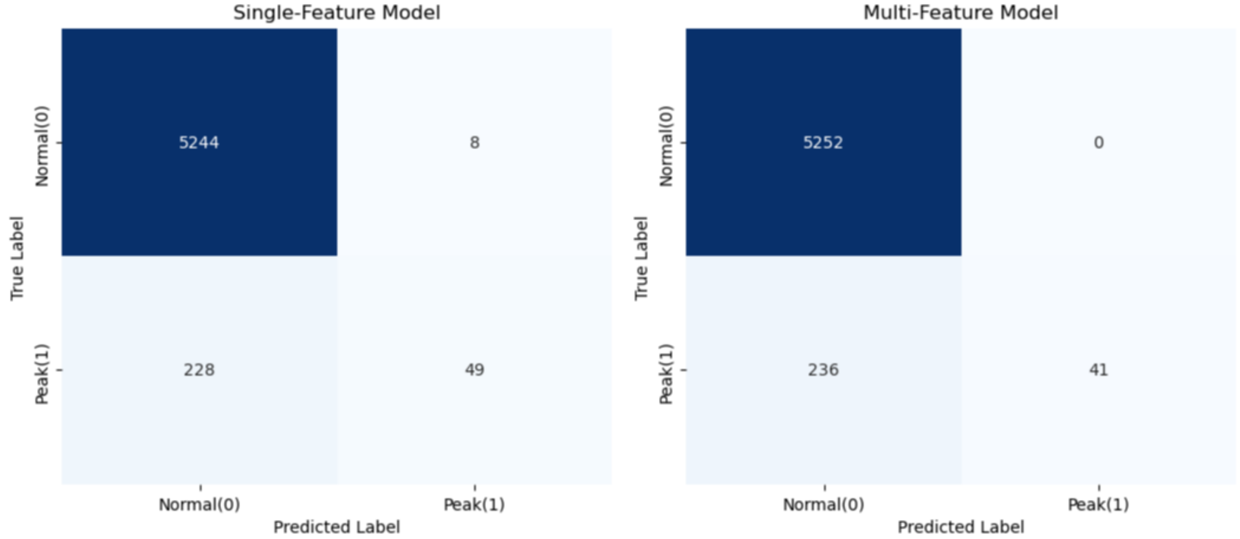


Figure 43: Confusion Matrix

Table 11: Performance metrics for single- and multi-feature RF classifiers in peak event detection.

Model	Accuracy	Precision	Recall	F_1-score	F_2-score	ROC–AUC
Single-feature RF	0.9577	1.0000	0.1552	0.2687	0.4688	0.8710
Multi-feature RF	0.9573	1.0000	0.1480	0.2579	0.4487	0.8641

Despite the high accuracy and perfect precision (1.0000) for both models—indicating no false alarms when a peak is predicted—both classifiers struggle significantly to detect peak events (Peak, class 1). The single-feature model correctly detected only 49 out of 277 peak events (true positives, TP = 49), missing 228 peaks (false negatives, FN = 228). Similarly, the multi-feature model detected even fewer peaks, with 41 true positives and 236 missed peaks. This results in very low recall values of 15.5% and 14.8% for the single- and multi-feature models, respectively. This evidence clearly shows that while the models are highly reliable when predicting peaks (high precision), they fail to capture most peak events (low recall). The multi-feature model's reduction of false positives to zero comes at the cost of missing even more peaks, further lowering recall. Consequently, both models suffer from the classic imbalanced classification issue, where rare but critical peak events are under-detected.

This challenge aligns with the regression analysis results, which also underestimated peak turbidity values across all forecasting approaches. The consistent underperformance in detecting high-magnitude peak events, both in classification and regression, highlights the fundamental difficulty of predicting rare and extreme events in this dataset.

The consistent difficulty in detecting peaks across both modeling approaches suggests that short-term turbidity forecasting in this context is dominated by statistical pattern recognition based on immediate past values, rather than being driven by identifiable physical processes. This also explains why simpler models such as ARIMA and RF outperformed the more complex LSTM, as there were no significant temporal lags or multivariate interactions for the latter to exploit.

To enhance model performance in future work, two key areas should be addressed:

1. Data Sufficiency and Representativeness

Accurate hourly prediction of turbidity is a challenging task. In this study, hourly water quality data is insufficient. Turbidity and EC time series were only available at DTWP and from Lobith station. However, Lobith is located too far from Nieuwegein to offer meaningful insight into upstream conditions, and on-site measurements do not capture inflows or disturbances occurring between these two locations. The lack of intermediate sensor data—such as at Hagestein—substantially limits the model's ability to detect upstream turbidity transfer dynamics that could improve prediction accuracy.

Further constraining model performance is the extremely low representation of peak turbidity events in our dataset (less than 5%). Because the model encounters very few instances of high turbidity during training, it tends to converge toward average behavior and cannot generalize well to rare, high-variance episodes. Imbalanced time-series regression problems are well documented to degrade predictive capacity. Extending the period of observation to include more extreme events would enable the model to better learn these critical dynamics.

Moreover, prior research shows that incorporating additional water-quality indicators—such as dissolved oxygen and ammonium—can enhance short-term turbidity forecasting. For example, Iglesias *et al.* (2014) developed an artificial neural network model for hourly turbidity prediction in northern Spain. The input data included measurements of turbidity, ammonium, EC, dissolved oxygen, pH, and temperature every 5 minutes from various sensors at the automated monitoring stations [45]. Installing additional sensors between Lobith and Nieuwegein—such as at Hagestein—would provide richer multivariate time-series inputs and likely yield more accurate and robust turbidity predictions.

2. Temporal Resolution and Noise

Using hourly data introduces a high level of random noise influenced by transient factors that our models and data sources do not capture—such as passing vessels stirring sediment in canals or localized disturbances [14, 46]. Therefore, many turbidity prediction research used daily turbidity as target [47, 48] even though they have higher temporal resolution data. Aggregating data at daily or multi-hour intervals (e.g. 4-hour or daily averages) may effectively suppress such noise, stabilize the underlying trends, and yield more robust predictions.

5 Conclusion

5.1 Conclusion

This thesis set out to answer three key research questions regarding short-term turbidity prediction at the DWTP:

- What parameters influence turbidity levels at the Lekkanaal DWTP?
- Which combination of features yields the most effective turbidity predictions?
- How far in advance can turbidity levels be reliably predicted?

1. Turbidity levels at the Lekkanaal DWTP are primarily influenced by a combination of hydrological and physicochemical parameters. Among these, upstream discharge and turbidity show the strongest positive correlations with local turbidity measurements. These relationships suggest that water flowing from the Lek River into the Lekkanaal carries significant amounts of suspended solids, especially during high-flow events caused by rainfall or snowmelt. The SOMs further support this, revealing that high turbidity events often coincide with a northward flow direction—from the Lek River into the Amsterdam-Rhine Canal—highlighting the Lek River as a primary contributor to elevated turbidity. Conversely, EC and water temperature exhibit strong negative correlations with turbidity. Lower EC during high turbidity events likely reflects dilution effects caused by increased surface runoff and suspended particles, which reduce the concentration of dissolved ions. Similarly, low water temperatures—typically observed in winter months—are associated with higher turbidity. This seasonal pattern aligns with increased discharge and sediment mobilization during colder months. These findings suggest that turbidity at the Lekkanaal DWTP is influenced not only by immediate upstream conditions but also by broader seasonal and hydrodynamic factors.

2. Feature selection analyses, including greedy feature selection and importance assessments via SHAP for RF and Captum for LSTM, demonstrated that a univariate model relying solely on sensor_turbidity outperforms multivariate models. Additional features, such as Lob_turbidity, Nieu_EC, Hag_dis, sensor_temperature, and flow_direction, despite exhibiting correlations with turbidity, introduced noise and reduced predictive accuracy across all models. The dominant importance of historic sensor_turbidity data, especially last three hours, confirmed by ARIMA's optimal configuration (ARIMA(3,0,1)), RF's SHAP analysis and LSTM's Captum analysis, underscores the sufficiency of recent turbidity measurements for robust predictions. This finding highlights the importance of model simplicity, as additional correlated features failed to enhance forecasting performance.

3. The models were evaluated across 1-, 3-, and 6-hour forecast horizons. All models exhibited robust performance for 1-hour-ahead predictions, with NSE exceeding 0.8. However, accuracy declined with longer horizons, with NSE dropping to approximately 0.69 at 3 hours and 0.62-0.66 at 6 hours. RF consistently outperformed ARIMA and LSTM, maintaining

the highest NSE (0.66 at 6 hours) and lowest RMSE across all horizons. ARIMA provided the least biased predictions (lowest MAE) for 1- and 3-hour forecasts, making it reliable for non-peak events. LSTM consistently underperformed, with the highest MAE and bias, particularly struggling with rare low-turbidity events. Taken together, these results indicate that the general turbidity pattern can be reliably forecast up to 6 hours in advance, with RF providing the most robust performance.

Despite these positive results for overall forecasting skill, all models systematically underestimated high turbidity peaks—the most critical events for DWTP operations. This limitation was confirmed by the RF classification experiment, which, while achieving perfect precision (1.0000), detected only a small fraction of peaks, with recall values below 16%. The consistent underperformance in capturing rare, high-magnitude events across both regression and classification frameworks highlights the fundamental challenge of peak prediction under the current data constraints. Consequently, while short-term turbidity dynamics can be reliably forecast up to 6 hours in advance, the reliable prediction of extreme turbidity episodes remains unresolved. Future research should focus on expanding monitoring coverage, improving data representativeness of peak events, and developing methods tailored to imbalanced time-series prediction.

References

1. United Nations. *Water and Sanitation - Sustainable Development Goals* Accessed: 2024-12-03. 2015.
2. Waternet. *Average water use* https://www.waternet.nl/en/service-and-contact/tap-water/average-water-use/.
3. Waternet. *Waar komt ons drinkwater vandaan?* 2.%09https://www.waternet.nl/service-en-contact/drinkwater/waar-komt-ons-drinkwater-vandaan/.
4. Sheng, D. P. W., Bilad, M. R. & Shamsuddin, N. Assessment and optimization of coagulation process in water treatment plant: a review. *Asean journal of science and engineering* **3,** 79–100 (2023).
5. Ratnaweera, H. & Fettig, J. State of the art of online monitoring and control of the coagulation process. *Water* **7,** 6574–6597 (2015).
6. Joo, D.-S., Choi, D.-J. & Park, H. Determination of optimal coagulant dosing rate using an artificial neural network. *Journal of water supply: research and technology—aqua* **49,** 49–55 (2000).
7. Li, L., Rong, S., Wang, R. & Yu, S. Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: a review. *Chemical engineering journal* **405,** 126673 (2021).
8. Jaffar, A., Thamrin, N. M., Amin, M. S., Ali, M., Misnan, M. F. & Yassin, A. I. M. Water quality prediction using lstm-rnn: a review. *Penerbit umt journal of sustainability science and management* **17,** 205–226 (2022).
9. Kurniawan, S. B., Abdullah, S. R. S., Imron, M. F., Said, N. S. M., Ismail, N. I., Hasan, H. A., Othman, A. R. & Purwanti, I. F. Challenges and opportunities of biocoagulant/bioflocculant application for drinking water and wastewater treatment and its potential for sludge recovery. *International journal of environmental research and public health* **17,** 9312 (2020).
10. Matos, T., Martins, M., Henriques, R. & Goncalves, L. A review of methods and instruments to monitor turbidity and suspended sediment concentration. *Journal of water process engineering* **64,** 105624 (2024).
11. Muthuraman, G. & Sasikala, S. Removal of turbidity from drinking water using natural coagulants. *Journal of industrial and engineering chemistry* **20,** 1727–1731 (2014).
12. Sohrabi, Y., Rahimi, S., Nafez, A. H., Mirzaei, N., Bagheri, A., Ghadiri, S. K., Rezaei, S. & Charganeh, S. S. Chemical coagulation efficiency in removal of water turbidity. *International journal of pharmaceutical research* **10,** 188–194 (2018).
13. Lee, C.-S., Lee, Y.-C. & Chiang, H.-M. Abrupt state change of river water quality (turbidity): effect of extreme rainfalls and typhoons. *Science of the total environment* **557,** 91–101 (2016).

14. Göransson, G., Larson, M. & Bendz, D. Variation in turbidity with precipitation and flow in a regulated river system–river göta älv, sw sweden. *Hydrology and earth system sciences* **17,** 2529–2542 (2013).

15. Roozen, F., Van Geest, G., Ibelings, B., Roijackers, R., Scheffer, M. & Buijse, A. Lake age and water level affect the turbidity of floodplain lakes along the lower rhine. *Freshwater biology* **48,** 519–531 (2003).

16. Shi, M., Ma, J. & Zhang, K. The impact of water temperature on in-line turbidity detection. *Water* **14,** 3720 (2022).

17. Mulla, N. H., Krishna, B. & Manoj Kumar, B. A review on water quality models: qual, wasp, basins, swat and agnps. *International journal of scientific research in civil engineering* **3,** 58–68 (2019).

18. Darji, J., Lodha, P. & Tyagi, S. Assimilative capacity and water quality modeling of rivers: a review. *Aqua—water infrastructure, ecosystems and society* **71,** 1127–1147 (2022).

19. Wai, K. P., Chia, M. Y., Koo, C. H., Huang, Y. F. & Chong, W. C. Applications of deep learning in water quality management: a state-of-the-art review. *Journal of hydrology* **613,** 128332 (2022).

20. Noor, S. S. M. & Saad, N. A. *A review on qual2k water quality model: comparative analysis with other models, recent advances and future directions* in *E3s web of conferences* **599** (2024), 02008.

21. Chen, Y., Song, L., Liu, Y., Yang, L. & Li, D. A review of the artificial neural network models for water quality prediction. *Applied sciences* **10,** 5776 (2020).

22. Chang, F.-J., Chen, P.-A., Chang, L.-C. & Tsai, Y.-H. Estimating spatio-temporal dynamics of stream total phosphate concentration by soft computing techniques. *Science of the total environment* **562,** 228–236 (2016).

23. Huo, S., He, Z., Su, J., Xi, B. & Zhu, C. Using artificial neural network models for eutrophication prediction. *Procedia environmental sciences* **18,** 310–316 (2013).

24. Choi, H., Suh, S.-I., Kim, S.-H., Han, E. J. & Ki, S. J. Assessing the performance of deep learning algorithms for short-term surface water quality prediction. *Sustainability* **13,** 10690 (2021).

25. Nielsen, A. *Neural networks and deep learning* 2015.

26. Ma, J., Ding, Y., Gan, V. J., Lin, C. & Wan, Z. Spatiotemporal prediction of pm2. 5 concentrations at different time granularities using idw-blstm. *Ieee access* **7,** 107897–107907 (2019).

27. Waqas, M. & Humphries, U. W. A critical review of rnn and lstm variants in hydrological time series predictions. *Methodsx,* 102946 (2024).

28. Zhang, F. & Fleyeh, H. *A review of single artificial neural network models for electricity spot price forecasting* in *2019 16th international conference on the european energy market (eem)* (2019), 1–6.

29. Song, C. & Zhang, H. Study on turbidity prediction method of reservoirs based on long short term memory neural network. *Ecological modelling* **432,** 109210 (2020).

30. Dhal, P. & Azad, C. A comprehensive survey on feature selection in the various fields of machine learning. *Applied intelligence* **52,** 4543–4581 (2022).

31. Ortiz-Lopez, C., Torres, A., Bouchard, C. & Rodriguez, M. 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 (2023).

32. Kohonen, T. The self-organizing map. *Proceedings of the ieee* **78,** 1464–1480 (1990).

33. Mounce, S., Blokker, E., Husband, S., Furnass, W., Schaap, P. & Boxall, J. Multivariate data mining for estimating the rate of discolouration material accumulation in drinking water distribution systems. *Journal of hydroinformatics* **18,** 96–114 (2015).

34. Peters, S., Ouboter, M., Lugt, K. v. d., Koop, S. & Leeuwen, K. v. Retrospective analysis of water management in amsterdam, the netherlands. *Water* **13,** 1099 (2021).

35. De Vriend, H., Wang, Z., vanMaren, B. & Peng, Z. in *Delta sustainability: a report to the mega-delta programme of the un ocean decade* 263–292 (Springer, 2024).

36. Teh, H. Y., Kempa-Liehr, A. W. & Wang, K. I.-K. Sensor data quality: a systematic review. *Journal of big data* **7,** 11 (2020).

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

38. Da Silva, D. G., Geller, M. T. B., dos Santos Moura, M. S. & de Moura Meneses, A. A. Performance evaluation of lstm neural networks for consumption prediction. *E-prime-advances in electrical engineering, electronics and energy* **2,** 100030 (2022).

39. Bod, J. *Spreading of the runoff times in the dutch rhine delta* B.S. thesis (University of Twente, 2021).

40. Lu, Y., Chen, J., Xu, Q., Han, Z., Peart, M., Ng, C.-N., Lee, F. Y., Hau, B. C. & Law, W. W. Spatiotemporal variations of river water turbidity in responding to rainstorm-streamflow processes and farming activities in a mountainous catchment, lai chi wo, hong kong, china. *Science of the total environment* **863,** 160759 (2023).

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

42. Wu, J., Stewart, T. W., Thompson, J. R., Kolka, R. K. & Franz, K. J. Watershed features and stream water quality: gaining insight through path analysis in a midwest urban landscape, usa. *Landscape and urban planning* **143,** 219–229 (2015).

43. Michel, A., Brauchli, T., Lehning, M., Schaefli, B. & Huwald, H. Stream temperature and discharge evolution in switzerland over the last 50 years: annual and seasonal behaviour. *Hydrology and earth system sciences* **24,** 115–142 (2020).

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

45. Iglesias, C., Martínez Torres, J., García Nieto, P. J., Alonso Fernández, J. R., Díaz Muñiz, C., Piñeiro, J. & Taboada, J. Turbidity prediction in a river basin by using artificial neural networks: a case study in northern spain. *Water resources management* **28,** 319–331 (2014).

46. Althage, J. Ship-induced waves and sediment transport in göta river, sweden. *Tvvr 10/5021* (2010).

47. Rele, B., Hogan, C., Kandanaarachchi, S. & Leigh, C. Short-term prediction of stream turbidity using surrogate data and a meta-model approach: a case study. *Hydrological processes* **37,** e14857 (2023).

48. Stevenson, M. & Bravo, C. Advanced turbidity prediction for operational water supply planning. *Decision support systems* **119,** 72–84 (2019).

6 Appendix

6.1 Complete Spearman Cross_Correlation Result

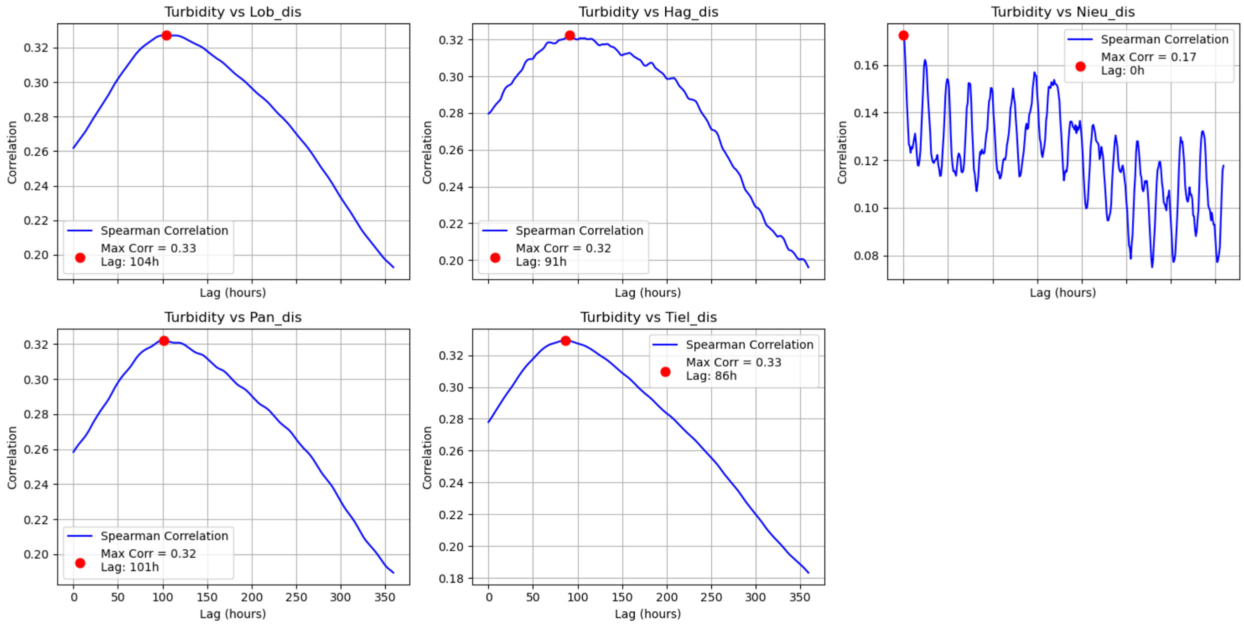


Figure 44: Spearman Cross_Correlation between sensor_turbidity and discharge

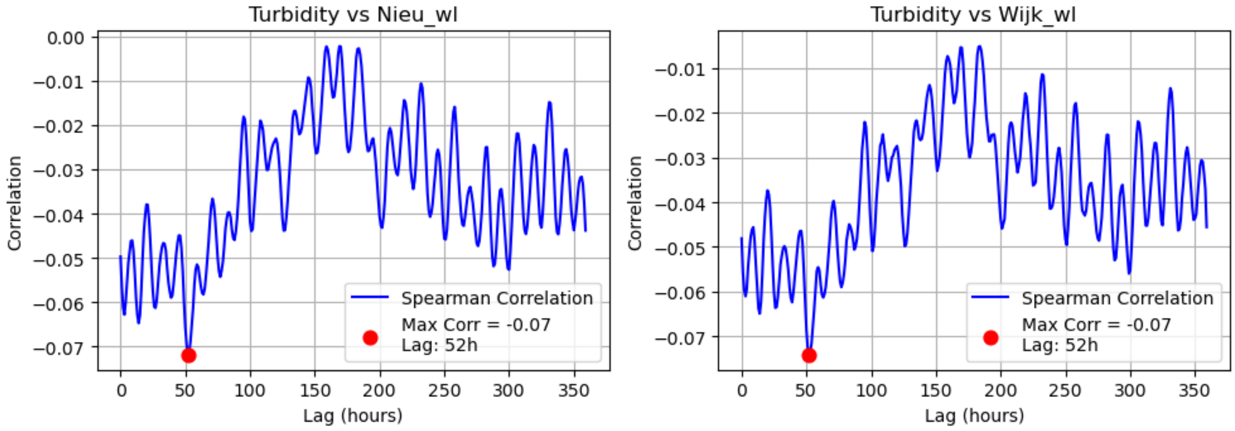


Figure 45: Spearman Cross_Correlation between sensor_turbidity and water level

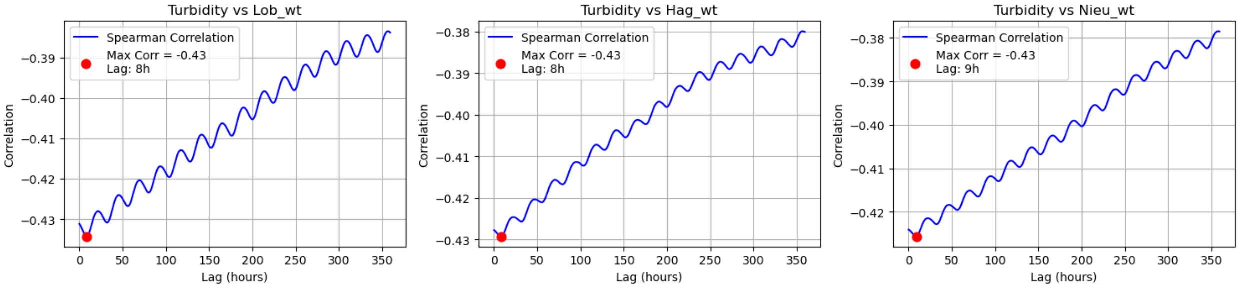


Figure 46: Spearman Cross_Correlation between sensor_turbidity and water temperature

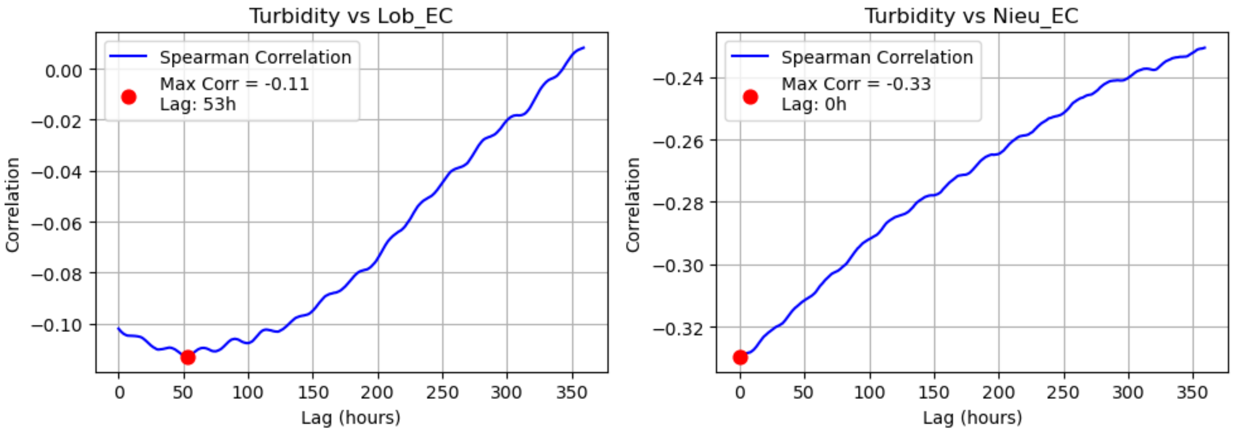


Figure 47: Spearman Cross_Correlation between sensor_turbidity and EC

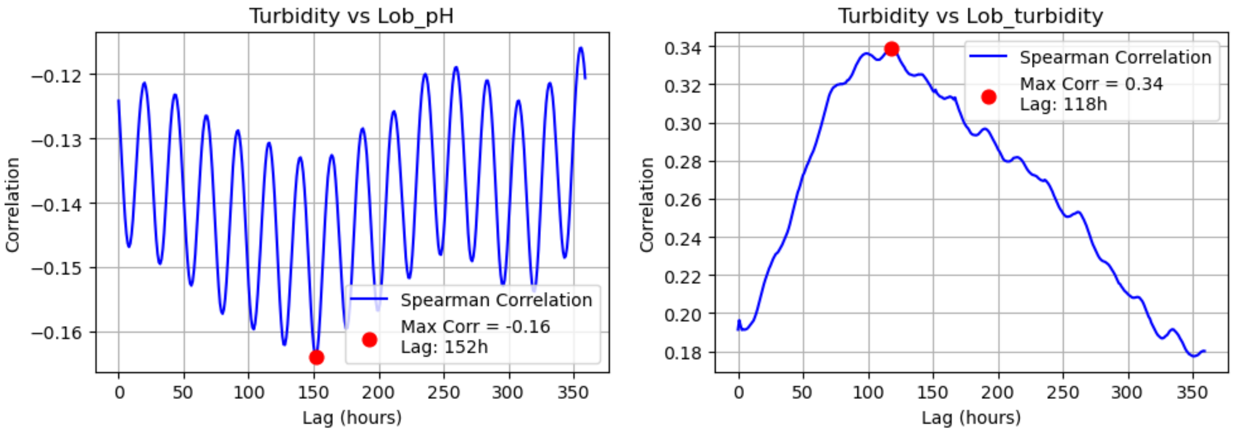


Figure 48: Spearman Cross_Correlation between sensor_turbidity and Lob_pH and Lob_turbidity

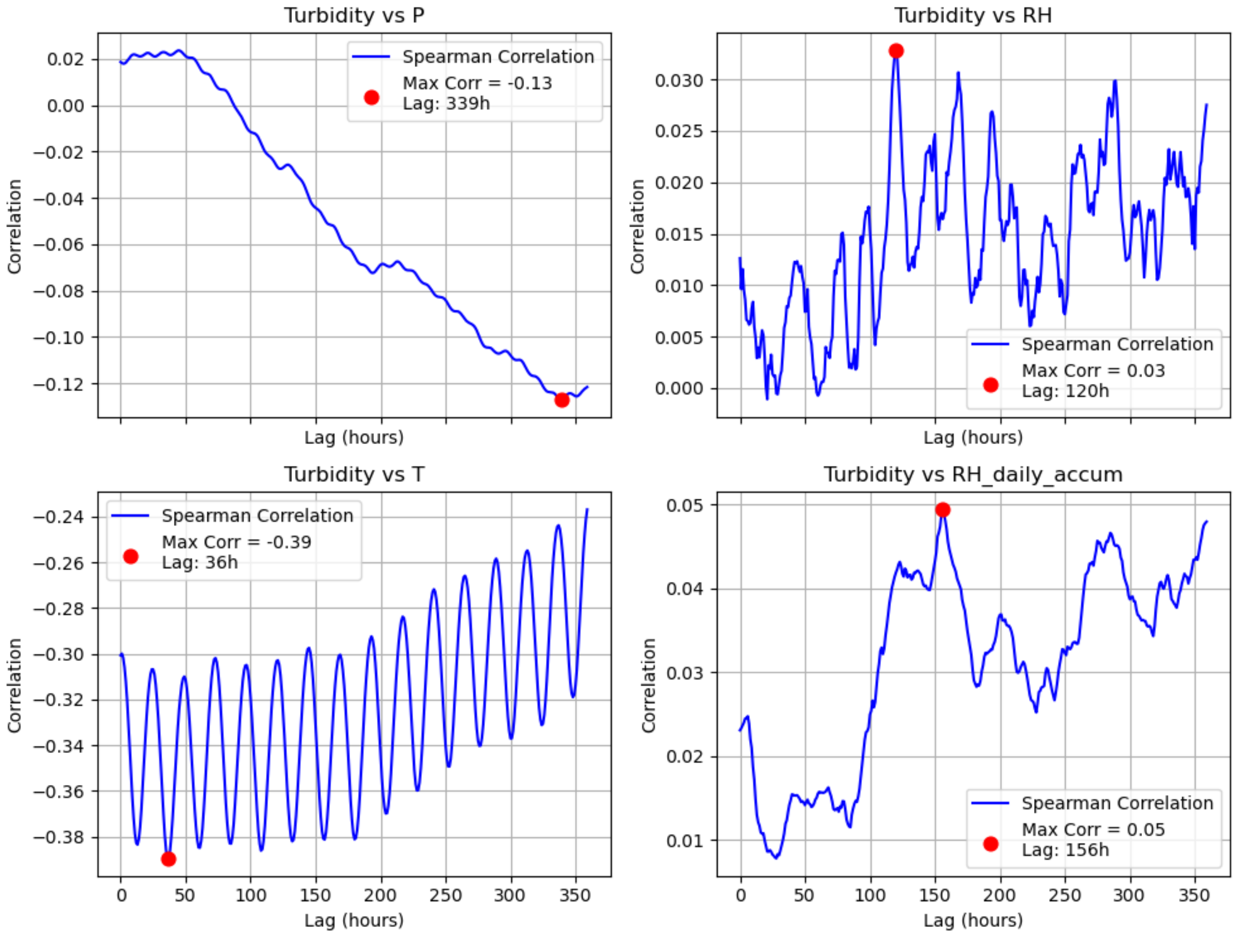


Figure 49: Spearman Cross_Correlation between sensor_turbidity and weather data

6.2 Model Configurations

Table 12: Model Configurations of section 4.4

Model Type	Forecast Horizon	Input Feature	Input Length	p	d	q		
ARIMA	1	sensot_turbidity	24	3	0	1		
ARIMA	3	sensot_turbidity	24	3	0	1		
ARIMA	6	sensot_turbidity	24	3	0	1		
Model Type	**Forecast Horizon**	**Input Feature**	**Input Length**	**n_estimators**	**max_depth**	**min_samples_split**	**min_samples_leaf**	**max_features**
RF	1	sensot_turbidity	24	103	8	8	1	10
RF	3	sensot_turbidity	24	654	7	11	1	9
RF	6	sensot_turbidity	24	330	7	11	20	9
Model Type	**Forecast Horizon**	**Input Feature**	**Input Length**	**hidden_size**	**num_layers**	**learning_rate**		
LSTM	1	sensot_turbidity_log	24	167	1	8.33E-05		
LSTM	3	sensot_turbidity_log	24	32	1	4.87E-05		
LSTM	6	sensot_turbidity_log	24	216	2	1.02E-05		

Table 13: Model Configurations of section 4.5

Model Type	Forecast Horizon	Input Feature	Input Length	n_estimators	max_depth	min_samples_split	min_samples_leaf	max_features
RF	1	sensot_turbidity + one candidate feature	24	103	8	8	1	10
RF	1	sensot_turbidity + all candidate features	24	425	6	8	1	50
Model Type	**Forecast Horizon**	**Input Feature**	**Input Length**	**hidden_size**	**num_layers**	**learning_rate**		
LSTM	1	sensot_turbidity + one candidate feature	24	167	1	8.33E-05		
LSTM	1	sensot_turbidity + all candidate features	24	62	1	1.29E-04		

Table 14: Model Configurations of section 4.6

Model Type	Forecast Horizon	Input Feature	Input Length	n_estimators	max_depth	min_samples_split	min_samples_leaf	max_features
RF	1	sensot_turbidity	24	777	1	13	14	18
RF	1	sensot_turbidity + all candidate features	24	814	7	11	2	48